PC Farms for Triggering and Online Reconstruction at HERA-B

J.M.Hernández (DESY) for the SLT, L4 and DAQ groups of the HERA-B Experiment

Abstract

The HERA-B data acquisition and triggering systems make use of Linux PC farms for triggering and online event reconstruction. We present in this paper the requirements, implementation and performance of both PC farms. They have been fully working during the year 2000 detector and trigger commissioning run.

Keywords: DAQ, Farm, trigger

1 Introduction

The HERA-B experiment [1] is a fixed target experiment at HERA, designed for the measurement of CP violation in the decay of B mesons. The data acquisition and triggering system must cope with more than half a million detector channels, a 40 MHz interaction rate and a signal to background ratio of around 10^{-10}. A highly selective multi-level trigger with a suppression factor of 10^{-6} , and a networked and high bandwidth data acquisition system, to handle the huge amount of data involved, have been implemented.

2 The HERA-B Trigger System

HERA-B implements a multi-level trigger to filter a few Hz of interesting events from a 40 MHz interaction rate. The main trigger signature is a pair of high p_T leptons which form a secondary vertex having the J/Ψ invariant mass.

In figure 1 the first table shows for each of the trigger levels the input rate, the input data volume and the time available for the trigger decision. The second table contains the data flow into the different hardware components of the DAQ. The number of detector channels is 550000 and the unsparcified (sparcified) event size is 500000 (120000) bytes respectively.

Level	Latency	Input	
		Trigger	Data
	sec	Hz	byte/s
Pre	10^{-6}	$10 \cdot 10^6$	$90 \cdot 10^9$
1	$12 \cdot 10^{-6}$	$10 \cdot 10^6$	$10 \cdot 10^9$
2	$7 \cdot 10^{-3}$	$50 \cdot 10^3$	$250 \cdot 10^6$
3	$100 \cdot 10^{-3}$	$500 \cdot 10^0$	$250 \cdot 10^6$
4	$4 \cdot 10^{0}$	$50 \cdot 10^0$	$6 \cdot 10^6$

Data flow (bytes/s)	
→L1 pipe	$5 \cdot 10^{12}$
→L2 buffers	$25 \cdot 10^9$
→L3 processors	$250 \cdot 10^6$
→L4 processors	$6 \cdot 10^6$
→tape	$2.4 \cdot 10^6$

Figure 1: Input rate, data flow and latency of the trigger levels

The large input event rate forces the first level trigger (FLT) to be a hardware trigger entirely build from specialized processors. It performs a Kalman filter based online tracking.

Due to the large data volume and the small latency, the Second Level Trigger (SLT) code only processes from the events regions of interest (RoI) developed by the FLT. The complete event information is stored in a distributed system of Second Level buffers (SLB) waiting for the SLT trigger decision. The SLT is a software trigger running on a PC farm of 240 processors. It refines the tracks found by the FLT adding more tracking super-layers and detector information, and applies cuts on secondary vertices. After passing the SLT, the events are built from the SLB's and processed at the same node using the Third Level Trigger code (TLT). The TLT looks at event properties beyond the triggered J/Ψ tracks. The events passing the TLT are sent to the online reconstruction Linux PC farm made up of additional 200 processors. The full event reconstruction is performed online to make most efficient use of computing resources and to minimize time delays between data taking and physics analysis. New calibration and alignment constants are derived online from the reconstructed events and if necessary fed back into the trigger and online reconstruction processors to keep the performance of the trigger and reconstruction under detector variations.

3 The HERA-B Data Acquisition system

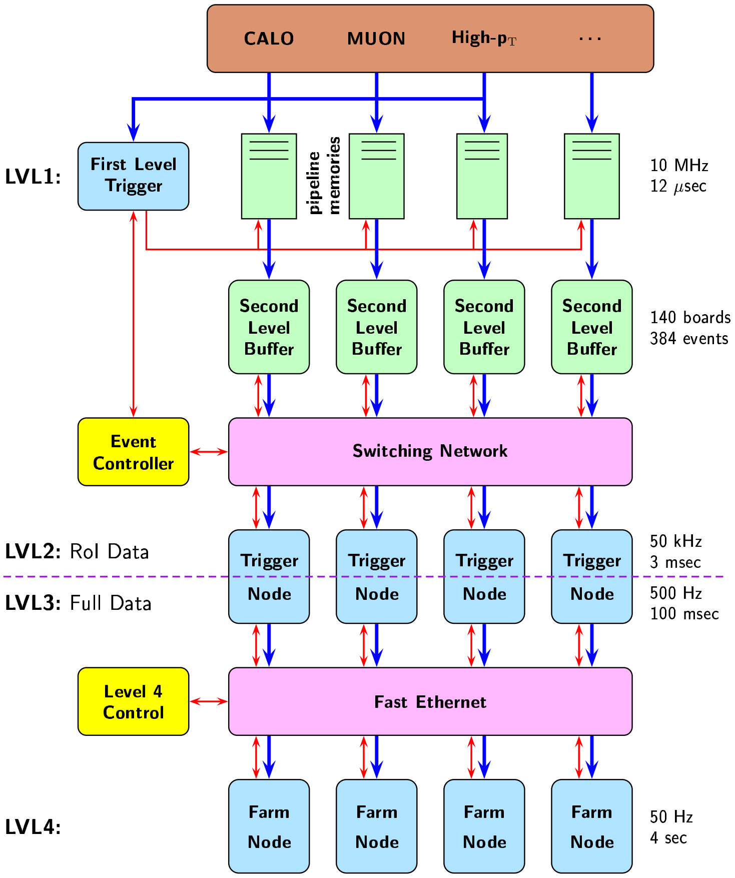

Figure 2: The HERA-B Data Acquisition and Trigger systems

Figure 2 depicts the HERA-B Data Acquisition System (DAQ) in connection with the different trigger levels. The HERA-B DAQ is a high bandwidth and high transaction rate

system to serve the needs of the HERA-B high level triggers.
The data path components in the DAQ are the following:

- ADSP2160 SHARC Processors: Digital Signal Processors used for data buffering in the SLB's and for the switching network connecting the SLB's and the Second Level Processors (SLP).
- Farm Processors: Intel Pentium PCs distributed in the SLT farm and the Online reconstruction farm.
- SHARC to PCI interface card [6] to connect the SLP's with the SHARC switch.
- Fast/Gigabit ethernet switches to connect the SLT farm with the Online reconstruction farm and that with the data logger node.

In the year 2000 data taking period, the SHARC switch was made of 264 interconnected DSPs. The bandwidth of the switching network was measured to be 1 Gbyte/sec with a maximum message rate of 2.6 MHz. This figures are well above the design values, a required bandwidth of 500 Mbyte/sec (250 MB/s for the RoI traffic and 250 MB/s for the event building) and a maximum message rate of 1.8 MHz.

4 The Second/Third Level Trigger PC Farm

Given the already reduced output rate of 50 KHz from the FLT, a PC FARM can be used to carry out the SLT algorithm [5]. The PCs have to have access to the detector data needed by the SLT algorithm. Each node takes care of processing a given event. Since the collection of the data of complete events at a rate of 50 KHz would require an unattainable bandwidth in the data transfer from the Front-End-Electronics of the detector to the farm, the SLT algorithm must work on small regions (Regions of interest) in the events, typically 1% of the whole event size. The raw detector data are stored in the distributed system of second level buffers during the execution of the SLT algorithm and accessed via the high bandwidth and high transaction rate sharc switching network.

The number of processors needed by the trigger farm is determined by the product of the input rate and the average latency of the trigger algorithm. For an average latency of around 3.8 msec per event and an input rate of 50 KHz, 190 processors are needed for the SLT. For the TLT, given the input rate of 500 Hz and an average latency of 100 msec, 50 processors are needed. In total, the Second/Third Level trigger farm is made of 240 nodes.

The size of the storage in the SLB's has to be at least large enough to store the number of events which simultaneously are being processed in the SLT given by the number of nodes in the SLT farm. To cope with rate fluctuations in the FLT, more events should be stored. There is room for 250 to 350 events in the SLB system.

Unlike the pipeline storage system of the detector readout electronics in which the events are overwritten after a given period of time, in the SLB system every single event can in principle stay arbitrarily long. Therefore, the maximum latency for a given event in the SLT is not limited by the storage size of the SLB but only the average latency. This way the system can cope with fluctuations in the latency of the SLT algorithm which depends heavily on the number of candidates found by the FLT.

Since the SLP's have access to the whole data of a given event, the SLT algorithm can in principle make use of the information of regions of the events arbitrarily large. The only limitation is the bandwidth and the maximum transaction rate between SLB's and SLP's. The size of the RoI's can be changed at any time. In the same way, additional sub-detectors or detector channels outside the RoI's can be used.

4.1 The SLT/TLT Farm Hardware

The SLT/TLT nodes are standard PCs equipped with both off-the-shelf and custom-made components. Among the standard components are the Intel Pentium III 400 MHz processor, 64 MB SDRAM memory, floppy drive and a 100 Mbit/s Fast Ethernet adapter. Specially developed for the connection to the SLB system was the SHARC to PCI interface adapter together with the driver code. The non-standard I/O driver copies data directly between the SHARC link interface and the user space. The throughput of the sharc-to-pci interface is 40 MB/s in both directions. Another remarkable feature of this link is the very small latency, around 1 μsec. There is another special interface card in the nodes for slow control using the CAN protocol. This serial line is used for powering up/down and reseting the nodes as well as for measuring temperatures and voltages in the mother board and the fan speed.

The nodes are disk-less. A reduced linux operating system is installed in a floppy disk from which they boot. At booting time the nodes issue a bootp request to the farm server to have an IP address assigned. The nodes NFS-mount a file-system from the server to have access to the executables of the trigger code. Only essential processes run in the nodes to avoid as much as possible process scheduling by the operating system. The trigger process should stay active as much as possible. No real time extensions of linux were necessary to ensure a good performance of the trigger code.

4.2 Data transfer to the Online Reconstruction Farm

The data transfer from the SLT/TLT farm to the online reconstruction farm is managed by a controller process running in one of the nodes of the SLT/TLT farm. This controller keeps internally a list of free online reconstruction nodes. When a SLT/TLT node is ready to send one event to the reconstruction farm it requests via the sharc link from the controller the address (IP number and port) of one free online reconstruction node. The events are sent via a standard Fast Ethernet (100 Mbit/s) network with Gigabit Ethernet (1000 Mbit/s) uplinks.

5 The Fourth Level Online Reconstruction PC Farm

Due to the vast amount of event data (around 20 TB/year) and the large event reconstruction time (around 4 sec/event), immediate data analysis can only be ensured by performing event reconstruction online before event data are archived. In addition event classification and event selection to further reduce the output rate from 50 Hz to 20 Hz are done in the fourth level farm (4LT) [7]. The SLT/TLT farm is a trigger system dominated by bandwidth and latency whereas the 4LT farm is dominated by processing power needs. The main tasks of the 4LT system are:

- Full online event reconstruction,
- event classification,
- final event selection (4LT trigger step),
- data logging,
- data quality monitoring,
- preparation of data for online calibration and alignment,
- event re-processing during shutdown periods.

The system must guarantee event data transfer. Therefore a push architecture is used for the event data stream, whereas monitoring information is collected from the nodes by a pull architecture. To process events at a rate of 50 Hz with processing times of 4 sec on a modern CPU (Intel Pentium III 500 MHz), 200 farm nodes are needed. The network must be capable of routing 5 MB/sec to the L4 farm nodes and a similar amount of data from the system to archive.

5.1 The 4LT Farm Hardware

The requirements of the system can be summarized as follows: It needs to provide sufficient processing power and bandwidth, be flexible to handle widely spread processing times and data rates, be scalable to easily increase the processing power and stay within cost limits.

For a farm architecture, the following building blocks can be identified: processing nodes, data links, switching network, event flow control, file and data server.

The concept and architecture of the 4LT farm allow to use off-the-shelf components. The processing power can be provided by modern PC-CPUs. The moderate bandwidth requirements can be met by Fast Ethernet. As farm nodes, Intel CPUs were chosen which are housed in dual CPU PCs. The PCs are equipped with Intel Pentium III 500 MHz processors, 256 MB SDRAM. Each node can buffer of the order of 10 events in its local memory.

The network is built of 24-port CISCO Fast Ethernet switches. In addition to the data stream which goes from TLT nodes to 4LT nodes and after processing to a central logger, control messages and monitoring data must be routed through the system. The data link to the mass storage media is realized by means of Gigabit Ethernet. Event data are buffered on large disks in the logger machine before being copied to tape.

5.2 The 4LT Farm Software

Major tasks are housed in separate processes, using the HERA-B message passing system between nodes and Unix standard interprocess communication within nodes. Reconstruction, event classification and selection packages are contained in a frame program called ARTE. It provides I/O and memory management for event data. ARTE is also used for offline analysis, including Monte Carlo event generation and detector simulation. ARTE contains interfaces to event data files for offline purposes as well as to shared memory segments for online usage on the 4LT farm.

5.3 Online Calibration and Alignment

During the reconstruction procedure, data which are needed to align and calibrate the detector are derived. To make use of the large statistics of up to 200 nodes providing such data in parallel, a scheme was developed to collect data in a central place (gatherer). Gathered data are then used to compute updated alignment and calibration constants which are saved in a central database.

In Hera-B all trigger levels and the online reconstruction depend on and have to access the most recent Calibration and Alignment (CnA) data. Updated CnA constants are produced online not only in the reconstruction farm but also in dedicated second level nodes. Those dedicated CnA nodes get calibration triggers from the DAQ at a sufficient rate.

The set of CnA constants used during data-taking at any time is identified by the so-called *CnA key*. This key is written in the event header of all events. It allows an event to be associated with all the calibration and alignment data used in its trigger process and online reconstruction.

A pushed architecture has been implemented for a low latency distribution of the CNA data to the Second level trigger processors. The updated CNA constants are multicasted using the fast and reliable SHARC links. On the other hand, the larger processing time of the online reconstruction nodes (secs), allows a slower distribution using a pull architecture. The reconstruction nodes fetch the updated CNA data from fast Database memory caches via fast ethernet.

5.4 Event data Re-processing

As the detector is better understood, the reconstruction packages further developed and improved CnA constants are produced, the event data need to be reprocessed. A system to exploit the CPU power of the 4LT farm for event data re-processing has been setup. It works similarly to the usual online processing scheme and makes use of the online protocols, in particular the logging and archiving facilities. Only the data source is different. Instead of passing raw event data from the DAQ system (via the SLT/TLT farm) to the 4LT farm nodes, a process retrieves data files from tape and distributes the events to the 4LT farm nodes. The event re-processing rate is only limited by the farm processing power, 50 Hz at a event reconstruction time of 4 secs.

6 Summary

HERA-B has successfully implemented trigger and online reconstruction farms in the DAQ. A low latency and high bandwidth switch for the Second Level Trigger system based on SHARC Digital Signal Processors has been built. It ensures full connectivity between the second level buffers and the second level processors. The SLT/TLT trigger and the online reconstruction farms are connected via a standard Fast/Gigabit Ethernet switched network. The trigger farm allows the implementation of flexible second and third level trigger codes. The online reconstruction farm allows immediate data analysis, data quality monitoring based on reconstructed data and online calibration and alignment. Updated CnA constants are fed back online into the trigger and online reconstruction farms. A system has been setup to exploit the CPU power of the farms to carry out data re-processing during shutdown periods.

In summary, PC farms running Linux provide a flexible, scalable and low cost solution for triggering in HEP experiments.

References

[1] HERA-B Collaboration. "HERA-B, and experiment to study CP violation in the B system using and internal target at the HERA proton ring", *DESY-PRC 95/01*

[2] J.M.Hernández, D.Ressing, V.Rybnikov, F.Sánchez, G.Wagner. "HERA-B Data Acquisition System", *Proceedings of the IEEE NPSS conference, Lyon, October 2000.*

[3] J.M.Hernández et al. "Farming in HERA-B", *Proceedings of the DAQ 2000 workshop at the IEEE NPSS conference, Lyon, October 2000.*

[4] G.Wagner. "Aufbau und Test der mit Digitalen Signal Prozessoren realisierten Komponenten des Datennahmesystem von HERA-B.". Ph.D.Dissertation. *Universitaet Hamburg, 2000*

[5] M.Dam. "Second and Third Level Trigger Systems for the HERA-B experiment." *Proceedings of the CHEP 98 conference* Chicago, September 1998.

[6] K.H.Sulanke and I.C.Legrand. "Fast PCI Communication Interfaces for Online Distributed Processing Systems." *Proceedings of the CHEP 98 conference* Chicago, September 1998.

[7] A.Gellrich et al. "The Fourth Level Trigger Online Reconstruction Farm for HERA-B." *Proceedings of the CHEP 98 conference* Chicago, September 1998.