HERA-B Data Acquisition System

J.M.Hernández[1] ,D.Ressing,V.Rybnikov, F.Sánchez[2], G.Wagner
DESY, Notkestrasse 85, D-22603 Hamburg, Germany.

Abstract

The HERA-B experiment is dedicated to the measurement of CP-violation in decays of neutral B-mesons. The B-mesons are produced in a formidable background of inelastic proton-nucleon interactions. The challenge for the DAQ system is a dead-timeless readout which requires unprecedent speed of storing and processing the data.

The Data Acquisition system has been installed in summer 1999 and it was providing data for detector and trigger commissioning runs in summer 2000.

I. Introduction

HERA-B was proposed in 1994 [1] as a fixed target hadronic B-factory to measure the CP violation decay asymmetry of neutral B's decaying to the CP eigenstates $J/\Psi K_s$. The detector and the trigger system had been designed to work at event rates of about 40 MHz. The fraction of interesting events to the minimum bias events is estimated to be around 10^{-12} [1]. A highly selective trigger is designed to extract the J/Ψ signal from the large amount of minimum bias events.

A. HERA-B trigger

The HERA-B trigger system consists of several levels. The First Level Trigger (FLT) [2] combines the lepton candidates from the muon detector and the electromagnetic calorimeter with hits in four of the tracking layers and determines the momenta by imposing a rough vertex constrain. The FLT cuts finally on the invariant masses of the two lepton candidates.

The Second Level Trigger (SLT) [3] is a software trigger running on a Farm of 240 Linux PC's. The SLT is able to digest input rates of about 50 KHZ and it is used to refine the FLT tracks and to apply cuts on the secondary vertices.

After passing the SLT, the events are build and processed at the same node using the Third Level Trigger (TLT) code. The TLT looks at event properties beyond the triggered J/Ψ tracks (i.e. detached multiprong vertices for B meson tagging).

On a positive answer the event is sent to a second PC farm (Fourth Level Trigger) to be recontructed [4]. The events are reconstructed online providing information for ONLINE calibration and alignment. This information is basic to keep the performance of the triggers under detector variations. The final logging rate to tape is about 50 Hz.

[1]Now at DESY Institut für Hochenenergiephysik, Platanenallee 6, D-15738 Zeuthen (Germany).

[2]Now at Max Planck Institut für Kernphysik, Postfach 103980, D-69029 Heidelberg (Germany).

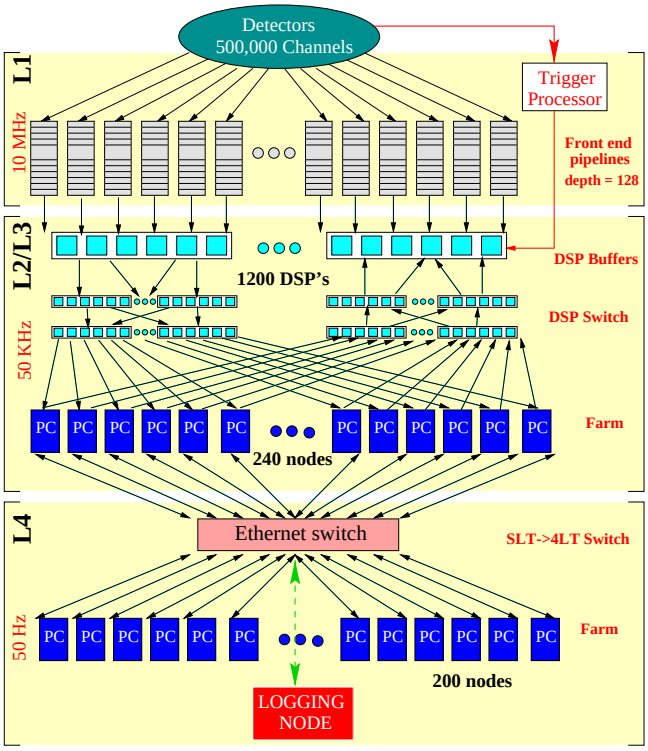

Figure 1: HERA-B Data Acquisition system scheme and correspondence to the HERA-B trigger levels.

II. HERA-B DAQ

The HERA-B data acquisition system is made of all high level trigger levels (SLT, TLT and 4LT) and the logging protocol plus the interconnections (Hardware and Software) between them.

The technology applied to each level had been selected to accomplish the requirements of the trigger protocols, see Fig. 1. The DAQ system is built from 4 components:

- The Fast Control System (FCS) [5] was designed by the Electronic Department at DESY to synchronize the subdetector Front End pipelines readout, distribute the trigger decision to the Front End pipelines and interface the trigger to the HERA-B DAQ system.
- ADSP21060 SHARC Processors: Digital Signal Processors (DSP) from Analog Devices which are used for data buffering for the SLT and switching connection from SLT Buffers to SLT decision units.
- A farm of Pentium PC's for Second and Third Level Triggers. A SHARC-link to PCI-bus interface allows the communication between the data buffers in the DSP cluster and the decision units.

- A farm of Dual Pentium III PC's (500 MHz), for the online reconstruction. The data transfer between SLT and 4LT nodes is done via Fast Ethernet network.

III. HERA-B MESSAGING SYSTEM

The HERA-B DAQ system is based on the communication between 2000 processes distributed among several operative systems: UNIX, LINUX and LYNX and several platforms: SGI, PC, VME-CETIA and Digital Signal Processors. The messaging system [8] is based in a central name server with a directory structure to allow grouping of processes under subcomponent branches (domain). The name server associates the domain and process name to the IP addres and Port of the corresponding process. The interprocess communication is done via a messaging protocol designed by the HERA-B DAQ system runnning on UDP. This messaging system translates from Big to Little Indian data formats to allow communication between different hardware platforms.

An additional identifier was added to the name server to allow the comunication to the DSP cluster. This number, so called HERA-B channel, identifies each DSP in the cluster and allows the intercomunication between them and the UNIX processes. The DSP processes are controlled via VME controller, that acts as a routing of messages from Ethernet (RPM) to the DSP Protocol (RPS). The name server identifies each DSP in the cluster with the IP address and port of the controller running on VME and the HERA-B channel that identifies the chip in the DSP network.

IV. THE SHARC BOARD

The SHARC board, see Fig.2, is a 6U VME card designed by MSC in Stutensee (Germany) in collaboration with HERA-B. Each board holds 6 ADSP-21060(SHARC) chips by Analog Devices running at 40 MHZ.

The ADSP-21060(SHARC) chip has a 512 kBytes of on-chip memory to hold event data and software code. The 6 ADSPs share a global memory bus for inboard communication, with a maximum bandwidth of 240 MBytes/s in 48 bit word mode. Each SHARC chip on the board drives 6 parallel transceiver link ports for off-board communication, capable to transmit 40 MBytes/s per link. The SHARC links are used to connect SHARCs in different boards of the switch architecture. The ADSPs have 10 independent DMA controllers, 6 of the for SHARC link data transfer and 4 for the global memory bus communication.

The communication to the VME controller is done in two ways:

- The VME controller is able to map the 6 SHARC's memory in the board. The controller is able to read/write in any of the SHARC local memory locations.
- The SHARC board has two, see Fig.2, independent FIFO's for input and output data transfer.

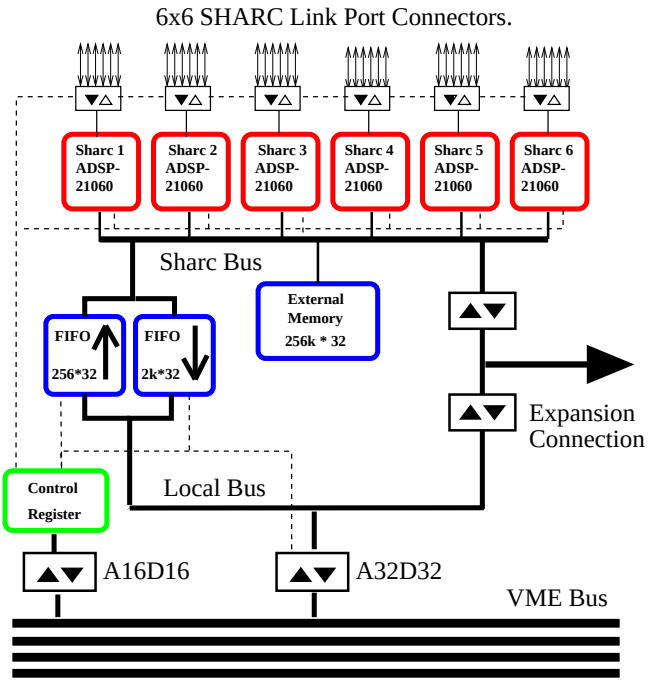

Figure 2: Schematics of the SHARC board with the connection to the VME BUS interfaces. There are two interfaces: A16D16 to set local control registers and A32D32 that maps the memory of the SHARC DSP's.

V. THE SLT PROTOCOL

The Second Level Trigger [3] has been designed to reduce the event rate by two orders of magnitude from an FLT output rate of 50 kHz and to provide high efficiencies on the interesting physics triggers. The algorithm is based exclusively on data from within Regions-of-Interest (RoI) pointed to by the previous trigger step. During second level processing, event data resides in the distributed Second Level Buffer (SLB) system build on SHARC DSP. The SLT processing is performed on a large computer farm, where events are handled in parallel by the farm nodes. Detector data from within RoIs is fetched on demand over the low-latency switching network which provides full connectivity between the SLB boards and the processing nodes, see Fig.3. Processing is continued through multiple steps until the track candidate either is rejected or has passed all requirements. For each step, additional data is pulled from the SLB system. Surviving track candidates form the basis of the event level decision. In case of an accept, the full event data is finally collected in the trigger node for event assembly and served to the next trigger level.

During the Second Level Trigger decision the DSP switch and data buffers has to deal with two type of messages. The ROI protocol requires small messages of around 140 Bytes (the size of 1 standard detector Front End Driver) while the Event Building requires messages sizes of about 800 Bytes (the size of 6 Tracker Front End Drivers connected to a Single DSP chip). Early Monte Carlo studies predicts a total of 250 MBytes/s of ROI messages (1.8×10^6 messages per second) and 250 MBytes/s of Event Building. These numbers are used below to estimate the DAQ performance.

VI. Digital Processor Switch

The SLT requires a low-latency, high bandwidth switching network to route data requests from the trigger processors to the buffer nodes. The switch will also carry the supervisor and buffer manager traffic as well as event data into the Second Level Trigger farm for event building.

The 140 SHARC cluster boards which comprise the Second Level Buffer (SLB) system are grouped into 10 buffer blocks of 14 boards, each as it is shown in Fig.3. Each SLB receives messages from the SHARC board through the request line. Outgoing messages are routed through the board along the reply line.

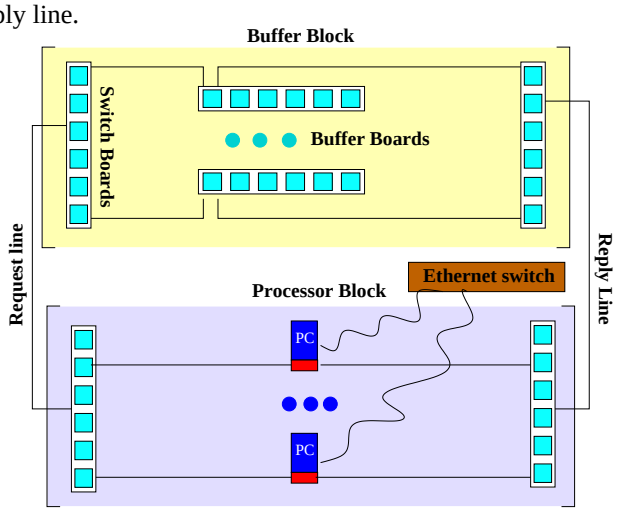

Figure 3: Scheme of the switch architecture with one Buffer Block and one Processor Block. The rectangular boxes represent SHARC boards with 6 DSP each (shadowed squares).

Similarly the 240 SLT nodes are grouped into 12 processor blocks, with one SHARC board for incoming messages and one for outgoing messages. The Second Level Trigger nodes communicates with the DSP switch via a dedicated SHARC2PCI interface card [7], this way the SLT nodes are able to send and receive messages using the local DSP message Protocol (RPS).

The switch is completed connecting the outputs of each of the SLB blocks to each input of the processor blocks and vice versa. Two additional SHARC cluster boards, the event controller, is added to the system as master buffer manager and to interface the Fast Control System to the SLT nodes.

A. DSP switching implementation.

Each DSP has up to 12 communication channels: 5 to send/received from the other DSP in the Board, 6 SHARC links to communicate to DSP in other boads and one to send/received messages from the VME controller. Each SHARC, see Fig.4, has an identification number and a routing table stored in memory[3]. Every message received by the DSP is routed to one of the 12 ports according to the routing table. This process is repeated until the destination channel corresponds to the local DSP identifier, then the mesaage is received and process by the

[3]The routing table is computed at booting time based on a SHARC connection Data Base

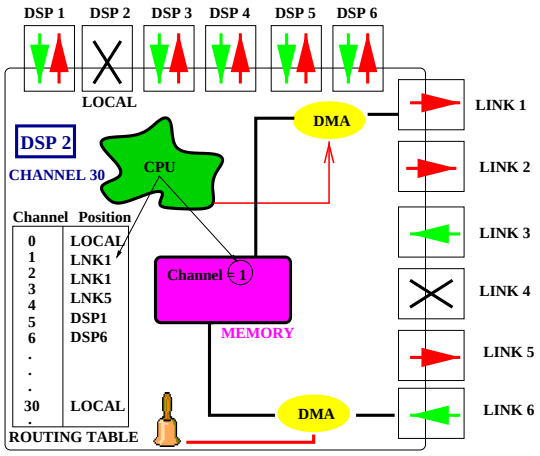

Figure 4: Scheme of the switching protocol on the SHARC Digital Signal Processor.

locally [6]. The messages are ansynchronously received and send using asynchronous DMA, leaving the SHARC CPU the possibility of routing or responding other messages.

A multicast protocol has been implemented using the same DSP switch network. The multicast is an additional DSP identifier, HERA-B channel, with an mask of ports associated in the routing table. The DSP copies the received messages as many times as the corresponding bit in the mask is set and enables independent DMA channels for each of the messages. This way the message is multiplied at different levels of the DSP network. The multicast is used to distribute the same information to the SLT nodes (i.e. calibration data) and to the SLT buffers (buffer updates and control messages) minimizing the load of the switch network.

B. Event Controller.

The Event Controller [6, 9] is the core of the Second Level Trigger traffic over the switch. It is in charge of receiving the trigger from the Fast Control System and associate a free Second Level Trigger Node to the corresponding buffer index where data is stored. It is at the same time the responsible of freeing Second Level Buffers when they are freed by the nodes (Master Buffer Manager). The Event Controller is implemented as a multiprocessor task where tasks are distributed between several SHARC chips in a board using the multiprocessing capabilities of the SHARC board. The actual implementation of the event controller is able to handle up to 70 kHz input rate from the Fast Control System [9].

VII. Logging protocol

The Fourth Level Trigger (4LT) [4] can be characterized as offline-like with respect to real-time requirements. The total rate from the SLT/TLT to the 4LT farm has been estimated to be larger than 12 MB/s with an average event size of 150 kBytes. SLT/TLT accepted events are routed to the 4LT farm nodes by a switching network, see Fig. 1, with a dynamic association of SLT/TLT nodes (sources) to 4LT nodes (destination). Reconstructed events are routed afterwards to a logging machine (e.g. IRIX) through a Giga-Bit line. The

logging machine buffers the acquired events in a 106 GByte disk before data is stored on tape. The performances of both protocols are shown in Table 1 and are compared with the design values.

Table 1
Performance of the SLT/4LT and 4LT/Disk ethernet communication and comparison to the design values.

Configuration	Bandwidth (MBytes/s)	Design (MBytes/s)
SLT to 4LT	$\geq$ 12.0	2.0-4.0
4LT to Buffer DISK	12.0	2.0

VIII. SWITCH PERFORMANCE

The HERA-B DAQ system was completed during the last year. The switch topology can be summarized as follows:

- 12 switch Processor blocks, each one connecting 20 SLT nodes.
- 10 switch Buffer blocks, each one connecting 14 Buffer Boards.
- 2 SHARC boards act as the EVC and interface to the Fast Control System.

The maximum transaction rate in the switch is a function of the data transaction capabilities of the hardware and the software protocol overhead for the message routing. The SLT protocol is made of two level of message sizes as it was described in section V. The RoI protocol is mainly limited by the transaction rate (1/overhead) since the message size is small. The Event Building is limited by the total switch bandwidth, since it is optimized to the largest possible message size.

The limitations of the switch performance were measured using a test-bench where a series of simple topologies with controlled data flux were investigated. A basic transaction performance was meassured in two different cases [9]:

- Routing from SHARC link to SHARC board bus and vice versa: Overhead = 5.6 μs. That is 31.5 MBytes/s for a typical event building message size (0.8 kBytes) and 11. MBytes/s for a RoI message length (0.14 kBytes).
- Routing through the SHARC board bus. The typical performance is shown in Fig.5. The very simple topology used in this measurement, 3 senders and 3 independent receivers where used at the same time, kept the total number of message collisions in the SHARC bus at a very low level. The results are overstimated with respect of the meassurements in the real environment.

In addition to the test bench studies a set of measurements was done with the installed DSP switch network:

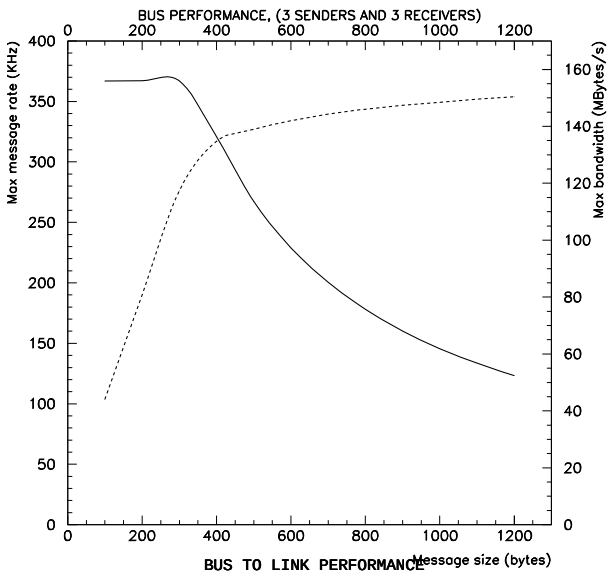

Figure 5: Maximum message rate (solid line) and bandwidth (dashed line) in the routing of messages through the SHARC Board BUS as a function of message size.

- The maximum switch throughput was measured building the complete detector at the highest speed. For 450 kBytes detector size a maximum of 2.2 kHz was measured. The limitation came from three of the switch blocks that were moving up to 120 MBytes/s each for messages of approximately 600 Bytes. This number is slightly smaller than the one quoted in the test-bench studies ($\approx$ 140 MBytes/s) where less message collisions are expected. The maximum message rate can be as well estimated to be $\approx$ 320 kHz from the test bench results ($\approx$ 370 kHz) and the measured inefficiencies at the real setup (120/140) due to message collisions on the switch board.
- The Bus to Link performance was measured in an special readout configuration where the transaction was limited by a single link connection. In this case a maximum of 32 MBytes/s was obtained for 800 Bytes message length, in agreement with the test bench results.

Table 2
Expected performance for different switch topology scenarios. 10+12 blocks was the topology during the last RUN and 15+15 blocks the expected topology for next year run. The numbers are compared with the SLT reference requirements as described in section V.

Configuration	Bandwidth (GBytes/s)	Message rate (MHz)	Performance to reference
10+12 blocks	1.2	2.6	110 %
15+15 blocks	1.8	4.8	180 %

IX. HERA-B DAQ BOOTING PROCEDURE.

As it was mentioned in the previous sections, the HERA-B DAQ system contains more than 2000 processes running during

the data taking. Every data taking session begins with the booting all the processes and establishing a tree like structure control infrastructure to operate the system according to the phase of the data taking (READY, PAUSE, RUN, etc.) [8].

The system boot-up is based on a process service. The service is provided by special processes running on every HERA-B machine called “process server” (proserv). A uniform interface based on HERA-B DAQ messaging system lets one to start-up and terminate processes remotely.

The boot-up procedure is controlled by a number of special processes called “processes manager”(prcmanag). During the booting procedure, they require “proservs” to start processes according to the information stored in the HERA-B DAQ data base. One or more process managers are involved in booting every subcomponent, i.e. all subcomponents are boot up simultaneously.

Once all the processes have been launched, the control infrastructure is established by means of state machine protocol applied by a half of processes in the system. The protocol provides the control and monitor of the subcomponents state with respect to their readiness to take data.

To get ready to take data different subcomponents may require different number of internal states of initialization. The task to bring all the states in the system to well defined one is fulfilled by special processes called “state machine branches”. Every branch process possesses the information about the state transition tables and state dependence tables of all the processes under its control. By means of these tables it can determine the overall state of the controlled processes as well as define the command to send to the processes to bring them to a desired state.

The branches obtain the state machine information during the boot-up procedure and therefore it can be dynamically changed by means of the DAQ data base.

The procedure to boot up the system and to bring it to the first well defined state takes less than 3 minutes for 2050 processes.

X. Summary

The HERA-B DAQ system is a high bandwidth and high transaction rate system to serve the needs of the HERA-B high Level triggers. The HERA-B high level triggers are software triggers running in two independent LINUX PC farms. A low latency switch fot the Second Level Trigger based on SHARC Digital Signal Processors has been built and it performs under specifications. A maximum of 1.0 GBytes/s has been meassured during Event Building in the Second Level Trigger Farm. The DSP switch has been integrated under the HERA-B messaging system allowing comunication between DSP and UNIX processes.

The communication from the Second and Fourth Level Farms have been implemented on a classical Ethernet switching with a meassured performance of 12 MBytes/s, well above the design needs.

The system had been running since one year in different configurations. 8 TBytes of data were recorded during the run from May to September 2000. The DSP switch performace will be increased to 1.8 GBytes/s (15+15 block topology) to cope with possible trigger ineficiencies.

XI. References

[1] HERA-B Collaboration, “HERA-B, An experiment to Study CP Violation in the B system Using an Internal Target at the HERA Proton Ring”. DESY-PRC 95/01.

[2] T.Fuljahn et al., “ Concept of the first Level Trigger for HERA-B experiment”. Proceedings Xth IEEE Real Time Conference,1997, pp.77.

[3] Mogens Dam, “Second and Third Level Trigger Systems for the HERA-B Experiment”. Proceedings CHEP 98 conference, CHICAGO September 98.

[4] A.Gellrich et al. “The Fourth level Trigger Online Reconstruction Farm for HERA-B”. Proceedings CHEP 98 conference, CHICAGO September 98.

[5] T.Fuljahn, “Aufbau und Charakterisierung des schnellen Kontrollsystems fuer das Experiment HERA-B”. Ph.D.Dissertation, Universität Hamburg,1999.

[6] F.Sánchez,“Digital Signal Processor software for the Hera-B Second Level Trigger”. Proceedings CHEP 98 conference, CHICAGO September 98.

[7] K.H.Sulanke and I.C.Legrand,“Fast PCI Communication Interfaces for On-Line Distributed processinf Systems”. Proceedings CHEP 98 conference, CHICAGO September 98.

[8] HERA-B Collaboration, “DAQ Online Software and the Run Control System in the HERA-B Experiment”. Proceedings CHEP 98 conference, CHICAGO September 98.

[9] G.Wagner, “Aufbau und Test der mit Digitalen Signal Prozessoren realisierten Komponenten des Datennahmesystems von HERA-B”. Ph.D.Dissertation, Universität Hamburg,2000.