A Prototype System for the Processor Farm of the HERA-B Experiment at HERA

A. Gellrich [a], R. Dippel [a], I.C. Legrand [a], H. Leich [a], S. Scharein [b], F. Sun [a], P. Wegner [a]

[a] *DESY-IfH Zeuthen, Germany*

[b] *Humboldt-University Berlin, Germany*

For the HERA-B experiment at HERA multi-processor systems (farms) for the last levels of the trigger and data acquisition system are planned. Based on implementation studies a concept for a prototype system was derived which will be set up in 1997. It will contain already the main components of the final system. The hardware decision for the processor nodes will be mainly influenced by the price/performance ration of the processor nodes. Currently under discussion are workstations and PCs.

Key words: farm; trigger; workstation; PC;

1 Introduction

The HERA-B experiment at HERA is starting its physics program in 1998. Its main goal besides various other B-physics topics is to study CP violation in decays of neutral B-mesons [1],[2]. To produce a sufficient number of b-quarks, a wire target is operated in the halo of the 820 GeV HERA proton beam. The initial event rate of 10 MHz with an average of four interactions per event corresponds to the HERA bunch crossing rate. A sophisticated four level trigger is required to filter interesting B-Physics events which appear at a rate of below 1 Hz. The entire trigger chain must obtain a reduction factor of 10^6 to allow for an output rate to mass storage of 20 Hz (see table 2).

2 Characteristics and Requirements

The trigger levels L2-L4 will have farm-like architectures.

L2: The main task of the L2 trigger is to perform a track re-fit in regions defined by the L1 trigger, to do tracking in the magnet, and to exploit information of the vertex detector. It will operate on parts of the data only. The processing time is limited by the buffer depth of the event data buffer (7.5 ms) from which the L2 nodes receive data via a switching network.
The L2 trigger system must be capable of handling high input bandwidths and providing fast response to control messages. See [3] for details.

L3: L3 processors are connected to the event data buffers by the same switching network as L2. They receive the entire data of an event, e.g. event building takes place just before L3. Contrary to L2, all data of an event are stored in the processor node's local memory. From the L3 entire events are routed to the L4 system. Algorithms for L3 under study incorporate particle identification, further tracking, and local pattern recognition methods.
The requirements to the L3 are similar to the L2.

L4: The main tasks of the L4 processor farm is to perform a full online event reconstruction and a final event selection before data are sent to mass storage. In addition, monitoring of data and calibration tasks will be carried out. The number of nodes (200) is determined by the processing time needed for the reconstruction program and the input rate. For such a large system, the price/performance

ratio is one of the most crucial items. The huge computing resources should be also usable for re-processing and Monte Carlo tasks in offline periods. The application software which is running on the L4 processor farm is developed and written by the offline software group of HERA-B. It is mainly written in Fortran and makes use of CERN packages.
The L4 system must provide an offline-like software environment which allows to run HEP application software. It has to be available all year long, also during shut-down periods.

Level	Input rate	Time	Reduction	Method of the trigger algorithm
L1	10 MHz	10 μs	200	tracking, high p_T, dilepton mass
L2	50 kHz	10-50 μs	100	track re-fit, magnet tracking, vertexing
L3	500 Hz	100 ms	10	further tracking, particle-ID
L4	50 Hz	4 s	2	full event reconstruction

3 Implementation Studies

In 1996 a VME-based system with two nodes was set up and tested [4],[5]. The aim of tests were to study the feasibility to build a common L3/L4 farm for HERA-B. On one hand side the requirements of a online trigger system (L3) has to be met which are mainly high bandwidths and low latencies. One the other hand a software environment to run complex HEP application programs is needed.
As processor nodes Motorola boards with PowerPC 604/100 MHz CPUs and PCI-bus were chosen. The processors nodes ran an embedded version of the IBM operating system AIX which was loaded by means of a local net (Ethernet) from one node connected to a disk. The data were transferred to the PCI-bus via high speed links, using the Data-Strobe (DS) protocol. To connect the DS-links to the PCI-bus, interface cards were developed in house and brought into operation. The switching was done with a STC104 non-blocking crossbar switch, which is based on DS-links. To control the switch, a GUI making use of client/server concepts was written in Java [6]. The operating system AIX offers an standard software environment for application software. It was easily possible to adopt HERA-B software to it. The VME-system is a standard online solution. The bandwidth of the DS-links were measured to be up to 6.2 MB/s with a latency of 46 μs. These numbers would be sufficient for HERA-B purposes. The online capabilities of AIX are considered to be not sufficient for the L3 trigger. Developing device driver software for the interface cards turned out to be a difficult and time consuming task. It was also doubted that VME-based processor nodes offer the best price/performance ratio as compared to workstations and PCs.

4 Prototype Concept

L2 and L3 are pure online systems, whereas L4 will be used offline-like without hard real-time requirements. L2 and L3 must be connected to the event data buffer by means of a switching network, using high speed data links. The HERA-B DAQ-system will be realized with digital signal processor (SHARC) which allow to transfer data at a rate of up to 40 MB/s.
It was decided to use the same hardware for the L2 and the L3 trigger farm. Both systems will be highly scalable and will contain of up to 150 nodes, depending on the actually achieved reduction rates and the time consumption. The first choice will be PCs running the operating system Linux. It offers sufficient real-time capabilities, e.g. low latency, to avoid blocking of the switching network during event building. The interfacing of the SHARC-links to the PCI-bus will be done similar to the implementation studies described above. Driver software under Linux has been written already.
To meet the requirements of a high performance reconstruction farm, the L4 system will be separated from the DAQ-system. The anticipated input rate to the L4 farm is around 50 Hz and the event size is 100 kB. This rate and the estimated processing time of 4 s per event lead to a farm size of 200 nodes, which seems to be feasible to build. Due to the rather moderate bandwidth in the system a network, using a standard link protocol, e.g. Fast-Ethernet, can be build. As farm nodes either AIX-workstations with PowerPC 604e/200 MHz or Linux-PCs with PentiumPro/200 MHz are under discussion. The decision will be made on basis of realistic benchmark tests by evaluating the price/performance ratios. The

usage of commercially available standard of-the-shelf components is considered to be advantageous, concerning costs and manpower.
For 1997 a small size prototype system to demonstrate the functionality of the complete data chain (L1-L4) is foreseen.

5 Summary

The requirements to the HERA-B DAQ and trigger system open a field of a new generation of HEP experiments and are leading directly to LHC experiments. The tight time schedule of HERA-B requires concepts and implementations which are based on products available now. The entire system must be operational in 1998.
For the three higher level trigger systems farm-like architectures were planned and are currently under construction. As far as possible commercially available standard components will be used. Linux-PC seem to offer sufficient real-time capabilities to be used in the trigger area. The final decision for L4 farm nodes, which is due mid 1997, will be based on the price/performance ratios of PentiumPro/200 MHz PCs compared to PowerPC 604e/200 MHz workstations. Benchmarks are currently carried out.

Acknowledgment

We would like to thank the electronics group, especially K.-H. Sulanke and U. Schwendicke, and the computer center of DESY-IfH Zeuthen for their support.

References

[1] T. Lohse et al., Proposal, **DESY-PRC 94/02** (1994)

[2] E. Hartouni et al., Design Report, **DESY-PRC 95/01** (1995)

[3] M. Medinnis, NIM **A368**:161 (1995)

[4] A. Gellrich et al., CHEP '95, Rio de Janeiro, Brazil, World Scientific (1996)

[5] A. Gellrich et al., DAQ96, RCNP Osaka, Japan, World Scientific (1997)

[6] I.C. Legrand et al., AIHENP96-Lausanne, Switzerland, NIM (1996)