Full Online Event Reconstruction at HERA-B

A. Gellrich[1,2], *H. Leich*[1], *U. Schwanke*[1], *F. Sun*[1], *P. Wegner*[1]

[1] Deutsches Elektronen-Synchrotron, DESY Zeuthen, Platanenallee 6, 15738 Zeuthen, Germany
[2] mailto:Andreas.Gellrich@desy.de, http://www-hera-b.desy.de/subgroup/farm

Abstract

We present and discuss concept and implementation as well as first performance results of HERA-B's online reconstruction farm. The system was completed recently and is taking part in the first physics run of HERA-B in 2000. It consists of 200 Intel PentiumIII/$500\,MHz$ CPUs which are housed in standard off-the-shelf PCs and are connected by Fast-Ethernet.

Keywords: large Linux-PC farm, commodity hardware, online calibration and alignment

1 Introduction

The HERA-B experiment [1, 2] at DESY in Hamburg/Germany is currently preparing for its first physics run with the full detector installed. The main goal is to study CP violation in so-called golden decays of neutral B-mesons into $J/\Psi K_s^0$ and subsequently into two leptons and two pions. To collect $O(1000)$ fully reconstructed golden decays per year, an event rate of $10\,MHz$ with 4 superimposed interactions is required. b-quarks are produced in collisions of the halo of HERA's $920\,GeV$ proton beam with an internal wire target which can be steered accordingly. HERA-B's DAQ and trigger system [3] exploits four levels to select a rate of less than $1\,Hz$ of interesting physics events from the initial rate. The designed background suppression is $O(10^{\,6})$. To allow for immediate data analysis and to avoid time-consuming re-processing of data, events will be reconstructed online before being logged to tape. Archiving of event data is planned at a rate of $20\,Hz$ which leads to a data volume of $20\,TB/year$ assuming $100\,kB/event$ and $10^{\,7}\,sec/year$. Since trigger efficiency and background suppression depend heavily on the quality of calibration and alignment of the detector components, online monitoring and updating of the constants database is implemented. During event reconstruction, quantities which are needed for calibration and alignment are derived and sent to a central destination. By making use of the data coming from the reconstruction, constants are updated. New constants are distributed to the reconstruction processes and to the trigger system. HERA-B's large data volume of $20\,TB/year$ and processing times for event reconstruction of several seconds require a clear strategy for data handling. The paradigm is to bring the application to the data rather than performing offline event reconstruction from archive. Processing must be performed in the data path as far as possible. Moreover, fully reconstructed events can be handed to the user immediately. A standard (Unix-like) environment in the online system, allows to use HEP offline software developments directly, e.g event reconstruction programs on the farm. Hence, the separation between online and offline software can be given up, which allows for efficient work and easy implementations.

HERA-B's first level trigger (FLT) is realized in hardware. By using information from the calorimeter or the muon system which is combined with data of tracking devices, the event rate is reduced to $50\,kHz$. The second and third level trigger steps (SLT and TLT) are performed in a multi-processor system (PC-farm) with dedicated data links using DSPs to an intermediate buffer

[4]. From then on events are transfered by way of standard network components to the fourth level (4LT) for online reconstruction, since the event rate is as moderate as $50\,Hz$. The SLT/TLT farm is a trigger system with real-time requirements to bandwidth and latency whereas 4LT is dominated by processing power needs.

2 Concept and Architecture

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 calibration and alignment,
- event data re-processing in shutdown periods.

The system must guarantee event data transfer. Therefore a push architecture exploiting a safe message passing protocol is used for the event data stream, whereas monitoring information is collect from the nodes by a pull architecture.

The requirements to 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,
- stay within cost limits.

For a farm-like architecture building blocks could be identified:

- Processing nodes,
- data links,
- switching network,
- event flow control,
- file and data server.

To process events at a rate of $50\,Hz$ with processing times of $4\,sec$ on a modern CPU (Intel PentiumIII/500 MHz), 200 farm nodes are needed. The network must be capable of routing $5\,MB/sec$ to the farm nodes and $2\,MB/sec$ from the system to archive.

3 Implementation

Concept and architecture of the 4LT farm allow to use commodity hardware for costs reasons, e.g. off-the-shelf components. See also earlier studies in [5]. The processing power can be provided by modern PC-CPUs. The moderate bandwidth requirements can be met by Fast-Ethernet.

3.1 Hardware

Processing nodes: As farm nodes Intel-CPUs were chosen which are housed in dual-CPU PCs. The PCs are equipped with PentiumIII/500 MHz processors, $256\,MB$ SDRAM, and $13.1\,GB$ (E)IDE-disks. Each node can buffer $O(10)$ events in its local memory.

Network: The network is built of 24-port CISCO-switches. The farm nodes are grouped into eight so-called mini-farms. The data link to mass storage media is realized by means of GigaBit-Ethernet. Event data is stored intermediately on large disks before being copied to tape.

Services: Three PCs are used to provide NFS-service for executables, collect and display slow control data by way of an http-server, and to allow for local event data logging. The logging node houses large SCSI-disks which can keep event data for several hours of standard running at $2\,MB/sec$. In addition, a local tape drive (DLT7000) is planned to run independent of DESY's central archiving facilities.

3.2 Software

Operating system: The farm-PCs run Linux which has become a standard software platform in HERA-B as well as in HEP in general. On each farm node a multi-process environment is available. Tasks are implemented as separate processes which communicate by means of Unix IPC-tools such as message queues, shared memory, and semaphores.

System software: Event data transfer is based on the Internet protocol TCP/IP. For the collection/gathering of information for data quality monitoring as well as for calibration and alignment, a UDP-based in-house development is used. Tape access is done via OSM.

Application software: HERA-B's main development platform has become Linux (S.u.S.E.). Most of the computing from Monte Carlo generation to analysis work is done on PCs. Application software such as event reconstruction and analysis packages are housed in a frame program which handles I/O and memory management. The frame program can be directly used on the 4LT farm without modifications. In the online case I/O is done from/to shared memory rather than files.

4 Status and Results

The HERA-B detector has been completed in the Christmas shutdown of 1999/2000. The 4LT farm is equipped with 200 CPUs in 100 dual-CPU PCs (see figure 1). All machines are installed and running. The price per node was $1.5\,kDM$. In the last running periods 1998 and 1999 the 4LT was always included in the data path. Usually a handful up to 45 nodes where used to receive event data from the SLT/TLT farm, to partly perform event reconstruction and data quality checks, and to send events to the mass storage media. The designed event and data rates could be exceeded. For event sizes above $O(10\,kB)$ up to $7.5\,MB/sec$ have been reached. It turned out that the network can be efficiently used even with small events by using a large number (> 20) of receiving/sending nodes. Parallel to normal data taking the complete 4LT farm has been extensively used for data re-processing. Tools were developed which efficiently use the network bandwidth by copying entire event data files to the local disks of the nodes. Current physics analyzes results such as the J/Ψ search are based on this re-processing. It could be shown that one dual-CPU PCs is equivalent to two single-CPU machines with respect to the time needed for event reconstruction. The reconstruction time itself is in the expected region of $O(1\,sec)$ although the detector has not been completely read out yet.

5 Summary

We described and discussed the Linux-PC farm for online event reconstruction at HERA-B. The system is complete, up, and running. It consists of 100 dual-CPU PCs which are integrated in a Fast-Ethernet network. The design parameters for event and data rates could be clearly exceeded. HERA-B's event reconstruction program is running on the farm and shows the expected timing. The usage of commodity off-the-shelf components as processing nodes and for the network allowed to easily meet the time, costs, and manpower estimates. The application of Linux as a software platform for online and offline turned out to be advantageous. It is possible to use the typically offline developed HEP application software without any modifications.

The farm is ready to perform full online event reconstruction in the upcoming first physics run with the complete HERA-B detector.

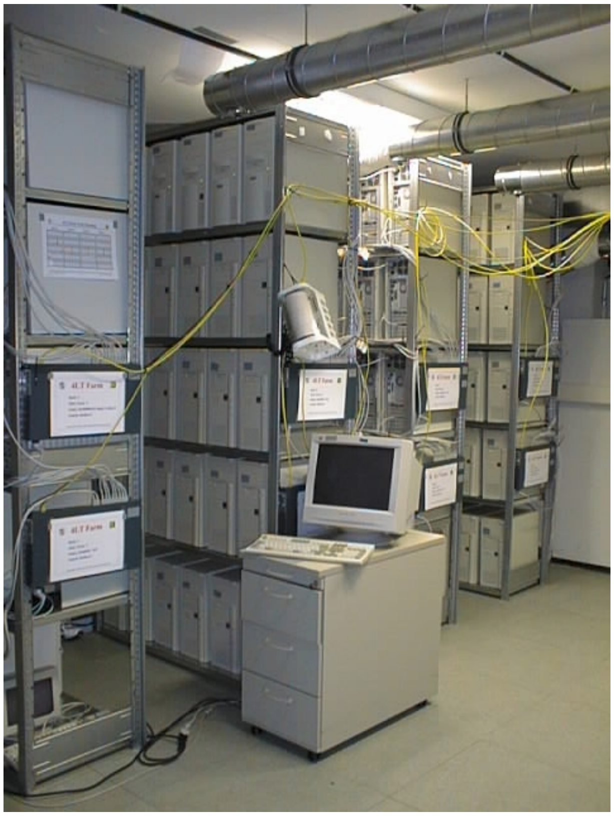

Figure 1: HERA-B's Linux-PC farm for online event reconstruction.

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. Dam et al., "Higher Level Trigger Systems for the HERA-B Experiment", IEEE Transactions on Nuclear Science Vol. **45**, No. 4 (1998).

4 A. Gellrich and M. Medinnis, "Higher Level Triggering Software", Nucl. Instr. Meth. **A408** (1998) 173-180.

5 A. Gellrich et al., "The Processor Farm for Online Triggering and Full Event Reconstruction of the HERA-B Experiment at HERA", *CHEP '95*, Rio de Janeiro, Brazil, 1995, "A Test System for the HERA-B Online Trigger and Reconstruction Farm", *DAQ 96*, Osaka, Japan, 1996, "A Prototype System for the Farm of the HERA-B Experiment at HERA", *CHEP '97*, Berlin, Germany, 1997, "The Fourth Level Trigger Online Reconstruction Farm of HERA-B", *CHEP '98*, Chicago, USA, 1998, "A Linux-PC Farm for Online Event Reconstruction at HERA-B", *11th IEEE NPSS Real Time Conference*, Santa Fe, USA, 1999, "The Linux-PC Farm for Online Event Reconstruction of HERA-B", *1st LCB Event Filter Farms Workshop at 3rd LHC Computing Workshop*, Marseille, France, 1999.