A Large Linux-PC Farm for Online Event Reconstruction and Data Management at HERA-B *

A. Gellrich†

Department of Physics, Humboldt University Berlin, Germany

Abstract. HERA-B enters a new regime with respect to data rates and volumes in HEP pointing towards the LHC era. The experiment exploits a Linux-PC farm for full online event reconstruction with 200 CPUs which has been in operation in 2000. Conceptual aspects of the Farm and the data handling in HERA-B as well as installation and running experiences after 6 months of full-time operation are presented and discussed.

INTRODUCTION

HERA-B (1, 2) is a fixed target experiment at DESY with high sensitivity for rare heavy flavor decays and QCD-studies. The HERA-B detector is a forward spectrometer with roughly 600000 readout channels. Particles are produced by a thin wire target which is inserted into the beam halo of HERA's 920 GeV proton ring. The detector consists of a vertex detector, several tracking devices, and particle identification systems. It is optimized for tracking in a high occupancy environment.

Taking into account the small production rates for heavy flavors compared to inelastic events in hadroproduction, a very high initial event rate is needed to produce sufficient numbers of interesting events. The HERA-B wire target can be steered to produce events at a rate of 10 MHz with several superimposed interactions. HERA-B writes events of roughly 150 kB size at rates of typically 20 Hz to archive which leads to data volumes of a few ten TB per year. The high input rate and the small signal to background ratio require a trigger and data acquisition system which provides sufficient filtering with a background suppression of 10^6 and moreover is able to process event data as they come along online. Sufficient processing power is obtained by introducing PC-farms into the trigger and data acquisition system. HERA-B makes use of a level 2 trigger farm consisting of 240 CPUs and a level 4 online reconstruction farm, in the following simply called Farm, with 200 CPUs. Due to the still small fraction of $O(10\%)$ of interesting physics events in the finally recorded samples a clear concept for the data management is needed.

HERA-B has entered a new regime with respect to data rates and volumes in HEP approaching conditions which will be present in the LHC era.

FARM

As the last stage of HERA-B's four level trigger and data acquisition system (3, 4) a large PC farm was built. Its purpose is to carry out full event reconstruction online before data are logged on disk and archived on tape. Re-processing is severely restricted due to the resources which would be needed to read back data from archive (tape) and the CPU power for processing. Retrieving a data volume of 30 TB from tape at a typical bandwidth of 4 MB/s takes already more than 100 days.

The tasks of the Farm can be summarized as follows:

- Fully reconstructed events are provided online.
- Event classification is performed.
- A final event selection called level 4 trigger is done.
- Detailed data quality monitoring is carried out.
- Data to be used for calibration and alignment is derived and collected.
- Data logging and archiving is prepared.
- The vast computing power can be used for event data re-processing.

Detailed studies of various implementation scenarios had been carried out before a solution based on standard

* Talk presented at VII International Workshop on Advanced Computing and Analysis Techniques in Physics Research, ACAT 2000, 16-20 October 2000, Fermilab, Batavia, USA
† Supported by the German ministry of education and science.

off-the-shelf components was taken. For information on studies in this context see (5, 6).

The Farm exploits 200 Pentium-III/500 MHz processors housed in commodity dual-processor PCs with 256 MB SDRAM and 13.1 GB (E)IDE-disks. The PCs are grouped into so-called mini-farms with 15 nodes connected via switched Fast-Ethernet. Up-links are realized by GigaBit-Ethernet. The hardware was completed end of 1999 and was put into operation in 2000.

Linux was chosen as the operating system since it has become the main development platform in HEP as well as in HERA-B. The farm-PCs are set up workstation-like. With an average event processing time of 4 s a rate of 50 Hz can be processed with 200 CPUs.

Within the farm nodes tasks are realized as separate Unix processes, using Inter-Process Communication tools for data and message exchange. Between nodes home-made message passing systems based on the Internet protocols UDP and TCP/IP are applied. For archiving data to tape OSM[1] software is used.

Event reconstruction software is housed in a frame program (*ARTE*). This frame program is also used for offline purposes such as Monte Carlo simulation and physics analysis (7). Although C/C++ has become the standard programming language, parts of the software are still written in FORTRAN. Data structures are based on ZEBRA-tables which are defined by a data description language. Recorded events consist of such tables. The Linux operating system on the Farm allows to use *ARTE* without modifications. I/O is done by means of shared memory rather than files. Offline developed software can be used directly on the Farm. The traditional separation between online and offline software could be overcome which eases and speeds up developments considerably.

Data quality monitoring and collection of information for calibration and alignment are major benefits of online processing. To achieve the designed detector performance in particular of the trigger system a fast feedback system is mandatory which updates calibration and alignment constants online. To make use of the full statistics, which is distributed over 200 nodes, software called a Remote Histogramming Package was developed. It allows to gather locally filled histograms by central processes. The package provides HBOOK-like functions and can also be used offline.

In periods without data taking, in particular in long shutdown periods, the system can be used for data reprocessing. From the Farm and logging point of view there is no principal difference between online running and re-processing. The involved software is the same: Instead of reading out the detector, raw event data are retrieved from tape and distributed to the Farm.

[1] Open Storage Manager

DATA MANAGEMENT

HERA-B must deal with a large data volume of 30 TB per year which needs to be inspected accumulatively over the lifetime of the experiment. Only a small fraction of events and/or information per event can be stored on random access media. Data management starts with the logging of event data to disk coming from the data acquisition system through the Farm. It incorporates consecutive writing of events to files, archiving of files, copying of files to commonly available disks, and retrieving of files from tape libraries (*staging*).

Normal user's physics analysis must be restricted to the reconstruction information of the events. Only a small fraction of $O(1\%)$ of raw data can be provided on disk. Three main types of event data files are distinguished:

- DST[2]: Reconstruction output plus raw data.
- MINI: Reconstruction output only (5% of DST).
- micro: Not yet defined subset based on ROOT (8).

It is planned to keep all MINI-files on disk which leads to a data volume of 1.5 TB per year. In addition, raw data information (DST) of roughly 1% of events can be stored on disk per year (300 GB/y).

At HERA-B calibration and alignment constants as well as slow control information are stored in a home-made database system based on Berkeley-DB (9). For event data management no commercial database is used since developments were not far enough advanced when the implementation was planned in 1995. HERA-B uses so-called index files as event directories. To keep data manageable, unique files names, convenient files sizes (400 MB per DST-file), and well-defined directory structures are exploited. Naming conventions allow to identify the source of data and reflect the logging mode. In normal data taking events are subsequently stored run-wise. Re-processing can be done run-by-run or for selected data sets.

Each event contains a classification bit mask which is assigned in the last part of the reconstruction. It distinguishes up to 32 categories (classes) of events. For each run an index file is written which contains one entity per event, consisting of the event number, date and time of data taking, and the classification. All index files are collected in a run catalogue. Index files and run catalogue allow for direct event access and are therefore crucial tools

[2] DST: Data Summary Tape

for efficient analyses which are usually based on small subsets of events. Special selections, in particular at the DST-level, are obtained by means of the event classification.

In order to make some DST-files temporarily available on disk, a so-called *staging* tool has been developed and included into the analysis frame program. If a user request a file which is not already in a dedicated disk pool (currently 206 GB), data will automatically be retrieved from tape. To make most efficient use of resources, files are kept on disk as long as space is available for further copies. The oldest file is deleted first.

SUMMARY

Installation, commissioning, and running of a large multi-processor system, exploiting standard components and Linux, turned out to be easy and efficient with respect to costs and manpower. The system was running stably and reliably during the run period 2000 and has been used for re-processing in the following shutdown. Designed performance values have been clearly exceeded in all fields. Input and output data rates up to 10 MB/s were reached to disk (logging). Archiving to tape was done at a sustained rate $4-5$ MB/s.

Compared to pseudo-online or offline approaches HERA-B benefits from immediately available fully reconstructed events for physics analysis, online event classification, and in particular from a comprehensive data quality monitoring system. Moreover, calibration and alignment constants are already derived online for some detector components which allowed for an online feedback. More detectors will be included in the coming months.

In 2000 a total data volume of more than 8 TB was logged and archived. Although this is only a fraction of the planned data volume of 30 TB per year, data management has become a major issue. Due to the small signal to background ratio of $O(10\%)$ even at the end of the trigger chain, HERA-B must cope with large event samples which need to be carefully inspected during physics analysis. Since typically only 1 TB can be stored on random access media, only parts of the full information of all events will be available on disk. Due to rather conservative approaches in condensing information, e.g., track selection, MINI-files still make up roughly 30% of DST-files.

The early phase of understanding the detector performance required rather frequent access to raw data (DST-files) in 2000. *Staging* has been a valuable tool to make DST-files temporarily available on disk. It turned out that in particular studies to understand detector hardware are based on small subsets of the data. Certain runs are favored by many detector groups. Physics analysis has started to move towards MINI-files which allows for immediate access to all events. First results were obtained.

HERA-B looks forward to the running period 2001 which will start after the luminosity upgrade shutdown of HERA end of summer.

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.
   Full Online Event Reconstruction at HERA-B, CHEP 2000, Padova, Italy, 2000.
6. http://www-hera-b.desy.de/subgroup/farm/welcome.html
7. H. Albrecht, *The Computing Model for HERA-B* CHEP '97, Berlin, Germany, 1997.
8. http://root.cern.ch/
9. A. Amorim et al., *The HERA-B database management for detector configuration, calibration, alignment, slow control, and data classification*, CHEP 2000, Padova, Italy, 2000.