First Experience with Online Reconstruction in H1

Ralf Gerhards and Zbigniew Szkutnik
Deutsches Elektronen Synchrotron DESY
Notkestrasse 85, 2000 Hamburg 52

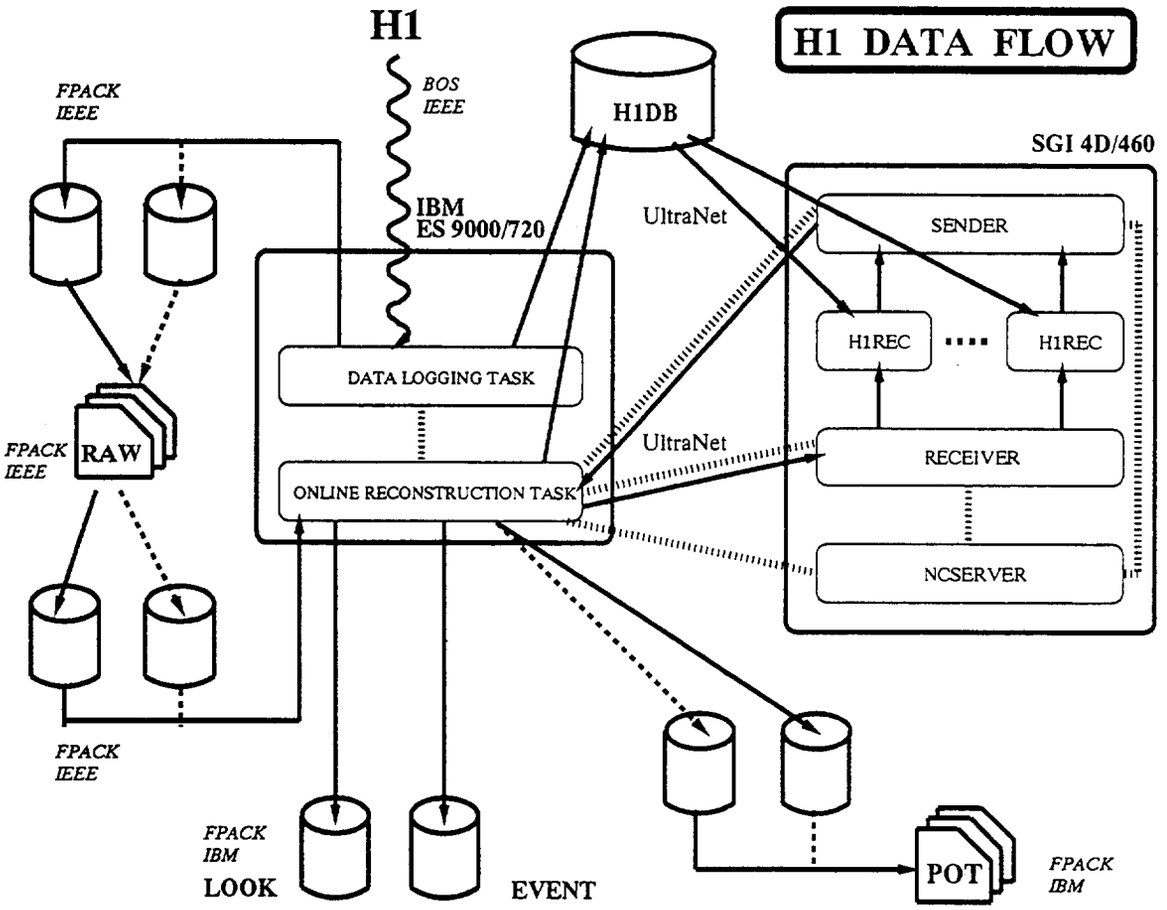
In spring 1992, the H1 detector at the HERA electron proton collider at DESY came into operation. The *ep* collisions recorded with an average rate of about 10 Hz are reconstructed almost parallel to the data taking with a delay of a few hours. This "online reconstruction" includes rejection of background events, downscaling of particular event classes, and event classification. It is performed on SGI workstations connected via fast UltraNet links to the IBM mainframe computer which is the file server for raw as well as for reconstructed data. The necessary software to operate the online reconstruction has been developed within the H1 collaboration together with the groups of the DESY computer centre. This report describes the experience with the online reconstruction during the first months of data taking at the HERA collider.

Introduction and requirements

In late spring 1992, the HERA electron proton collider at DESY provided first luminosity for the experiments H1 and ZEUS. The high collision rate of 10,41 MHz and a correspondingly high trigger rate are placing demanding requirements not only on the data acquisition but also on the event reconstruction.

The H1 detector and its data acquisition system are described elsewhere [1, 2]. Various levels of hardware triggering, software filtering, and digital compression are employed to reduce the final data sizes to acceptable recording rates. A parallel array ("farm") of MIPS R3000 RISC processors [3] is serving as a final software trigger to reject background events.

The raw data are finally received at a maximum rate of 1.2 MB/s on the DESY IBM ES 9000/720 VF, formatted, temporarily stored on IBM 3390-2 disks, and written to 3480 cartridges whenever blocks of 245 MB have been filled. It is expected that, during normal HERA operation, the data taking will proceed with an average rate of 10 Hz or about 500 KB/s, resulting in an estimated yearly data volume of about 5 TB. These data are supposed to pass the event reconstruction program with a delay of a few hours only, to be available for physics analysis. Assuming a processing time of about 1 second per event, the computing power needed for this *online reconstruction* is exceeding by far the capabilities of the mainframe computer. Both HERA experiments are therefore using SGI 4D/460 multiprocessor workstations based on the RISC technology [4]. They are not connected to the data acquisition of the experiment but directly to the IBM mainframe, because its mass storage devices (4 STK 4400 ACS) are still better suited to store the huge amount of data and because a considerable fraction of the offline analysis will be done there. In addition, a completely independent operation from the data logging was desired.



A fast data link is required connecting the mainframe and the workstations that became available by an UltraNet installation at DESY in summer 1991. A (memory-to-memory) transfer rate of 7.5 MB/s has been measured using *two-way stripe* mode and two IBM BMPX channels, at a reasonably small CPU load on both hosts due to protocol evaluation.

Data flow for online reconstruction

The figure shows a sketch of the H1 data flow. The input to the *online reconstruction* is provided by the *raw data tapes*. Several processes are used to manage this task, in particular the data exchange between the IBM mainframe and the SGI workstations, using the standard TCP/IP protocol. Being started, the *online reconstruction task* connects via a fixed TCP/IP port to a server process on the SGI *(ncserver)* which in turn starts two processes to receive data from the IBM *(receiver)* and to send data back to the IBM *(sender)* as well as a number of event reconstruction programs, independently processing different events. Thereby, the UNIX *fork* facility and a set of *shared memory areas* and *semaphores* to share data among processes are used. Each process is accessing the master H1 database on the IBM independently using a *database server*. The processes are reading events from an input buffer which has been previously filled by the *receiver* and writing accepted events into an output buffer which in turn will be send by the *sender* whenever it is filled completely. The number of input and output buffers can be defined at runtime.

At present, four input and two output buffers are used, each of about 3 MB size.

The data are received on the IBM by the *online reconstruction task* and written temporarily to IBM 3390-2 disks. Whenever about 260 MB of data have been received, a job is being submitted that sorts the data by run and event numbers and copies about 200 MB to a *production output tape* (POT, 3480 cartridge) and the rest to an intermediate file which will be included in the next sorting process. This ensures that data are sorted globally throughout all cartridges. Besides event data, *histogram data* produced by the reconstruction processes are sent using the same data path. In addition, every hundredth event is written to a single event file for online event display. It should be emphasized that any data transfer is proceeding via UltraNet. The complete system is being controlled and supervised on the IBM mainframe.

The whole SGI part of the system is written in the C language as is the network part while the reconstruction program and the major part of the supervising software are coded in FORTRAN 77.

The *shared memory* concept turned out to be particularly valuable for managing several simultaneous processes on a single multiprocessor computer. An independent network access to the data by each process, as it was embedded in a first version of the program, led to much lower efficiencies.

The use of a recently developed package for machine independent I/O *(F-package)* [5] is of great advantage in managing the heavy data exchange between different computers. In particular, the data consist of fixed length physical records with record headers that contain information about the format, the representation (IEEE, IBM, VAX), and the type (BOS, LOOK) of its data.

Experience during first data taking

After half a year of coding, a first version of the program was released in autumn 1991 and used during the H1 test run with cosmics in October/November. Despite important drawbacks found in the networking part and the sorting of events (for that purpose Fortran keyed access had been used), this test was successful in showing that the concept worked technically.

After improving the limiting parts, a revised version was available early 1992. It was proved in another test run in April/May that the program could stand the requirements of online reconstruction. The program was then used for online event reconstruction from the very beginning of the *ep* data taking in June/July.

The average reconstruction time was about 1.25 seconds per event, after partial optimization of the reconstruction program code, which results in a maximum processing rate of 4.5 Hz, by making full use of one SGI workstation. The CPU load was 600 % (6 processors). The event size increased by about 50 %. The reconstruction program has proven to be quite stable. To catch occasional failures, a signal-handler was implemented. Installations on different platforms helped to find quite a few inconsistencies and coding problems. To cope with the data taking rate and to keep the amount of reconstructed data reasonably small, a filter step has been included in the reconstruction program which is presently based on tracking information only. The resulting POT volume was of about the same size as the raw data sample.

No effort is being made to order events before writing the POT cartridges. Sorting about 250 MB of data and writing a POT cartridge takes 8 to 10 minutes real time which is roughly twice the time needed to produce a *raw data cartridge.*

Because of long breaks between luminosity periods it is presently no problem to catch up the data taking and all data acquired by the H1 detector have been reconstructed and made available for physics analysis within a few hours. At higher luminosity, however, a further data reduction will be needed, for example by using more sophisticated background rejection and downscaling particular (physics) event classes. On the other hand, an upgrade of the processing power is envisaged for early spring 1993. For the time being, a second SGI workstation can be included into the operation which has already been successfully tested. It should be stressed that there are clearly no limitations in the network part.

The reliability and performance of the SGI systems is impressing. Some imperfections in UltraNet resulted from this project being the very first attempt to write a socket-based application using UltraNet under IBM/MVS and could be continuously removed in close cooperation with UltraNet staff on site.

Summary

During the first period of HERA operation in july 1992, the *online reconstruction* of H1 data was successfully tested. This concerns the use of a multiprocessor RISC based workstation as well as a fast data link between the IBM mainframe and the workstation. It was shown that the proposed scheme is able to cope with the rate of H1 data taking, provided a filter step is included in the reconstruction. This filter keeps, in addition, the output data volume within reasonable limits. Looking forward to upgrades in the areas both of workstation computing power and mass storage facilities, the challenging requirements of data processing at HERA will most probably be fulfilled by the present configuration.

Acknowledgement. The authors gratefully acknowledge the help of many colleagues from the DESY computer center in operating both the DESY IBM and the SGI and in developing the necessary software. The authors would also like to express their gratitude to their colleagues in the H1 collaboration for the many fruitful discussions to develop and improve this challenging computing project.

References

[1] H1 Collaboration, *Technical proposal for the H1 detector*, DESY, March 1986

[2] W.J.Haynes, *Bus based architectures in the H1 data acquisition system*, Int. Conf. on Open Bus Systems, Zurich, Switzerland, RAL 92-048, August 1992

[3] A.J.Campbell, *A RISC multiprocessor event trigger for the data acquisition system of the H1 experiment at HERA*, RAL 91-060, September 1991

[4] M.Ernst, K.Künne, *A new mass-storage system at DESY*, these proceedings

[5] V.Blobel, *The F-package for input-output*, these proceedings