# DATA STORAGE AND DATA ACCESS AT H1

R. GERHARDS, C. KLEINWORT, U. KRÜNER-MARQUIS

*Deutsches Elektronen Synchrotron, Notkestrasse 85, D-22603 Hamburg, Germany*

F. NIEBERGALL

*II. Institut für Experimentalphysik, Universität Hamburg,*
*Luruper Chaussee 149, D-22761 Hamburg, Germany*

The electron proton collider HERA at the DESY laboratory in Hamburg and the H1 experiment are now in successful operation for more than three years. The H1 experiment is logging data at an average rate of 500 KB/s which results in a yearly raw data volume of several Terabytes. The data are reconstructed with a delay of only a few hours, also yielding several Terabytes of reconstructed data after physics oriented event classification. Physics analysis is performed on an SGI Challenge computer, equipped with about 500 GB of disk and, since a couple of months, direct access to a StorageTek ACS 4400 silo. The disk space is mainly devoted to store the reconstructed data in very compressed format (typically 5 to 10 KB per event). This allows for very efficient and fast physics analysis. Monte Carlo data, on the other hand, are kept in the ACS silo and staged to disk on demand.

## 1 Introduction

At the electron proton collider HERA at the DESY laboratory in Hamburg, electrons of 27.6 GeV momentum are collided with 820 GeV protons at a bunch crossing rate of 10.41 MHz, delivering (design) luminosities of $1.5 \times 10^{31} cm^{-2} s^{-1}$ (peak luminosity) and $50 pb^{-1} y^{-1}$ (integrated luminosity). The luminosity values reached up to now are $5.0 \times 10^{30} cm^{-2} s^{-1}$ and $5 pb^{-1} y^{-1}$, respectively. The large bunch crossing rate poses particularly challenging requirements on the triggering and data acquisition facilities of the experiments.

Cross sections for various physics processes at HERA and the resulting data rates are presented in table 1, assuming an instantaneous luminosity of $10^{31} cm^{-2} s^{-1}$. It can be seen that the background rate is about three orders of magnitudes higher

Table 1: Visible cross sections and data rates for physics processes at HERA

| process | cross section | rate |
|---|---|---|
| $\gamma p$ | $20\,\mu b$ | $\mathcal{O}(100\,Hz)$ |
| $c\bar{c}$ | $1\,\mu b$ | $\mathcal{O}(10\,Hz)$ |
| $NC$ | $100\,nb$ | $\mathcal{O}(1\,Hz)$ |
| $b\bar{b}$ | $5\,nb$ | $\mathcal{O}(0.1\,Hz)$ |
| $CC$ | $0.1\,nb$ | $\mathcal{O}(0.001\,Hz)$ |
| p induced background | | $\mathcal{O}(100\,kHz)$ |

than the physics rate. Secondly, the rate of the more interesting channels is at least another three orders of magnitude lower. Thirdly, not all physics events can be logged to tape if a reasonable logging rate of 10 Hz is aimed for. Assuming effectively $10^7 s$ of HERA operation, this results in a yearly data volume of $10^8$ events. This is a new challenge for experiments at HERA, in the area of triggering, data acquisition, data recording and data processing, even if these values are still orders of magnitudes lower than expected in future experiments, in particular at LHC.

The H1 computing environment is shown in figure 1. The main computing

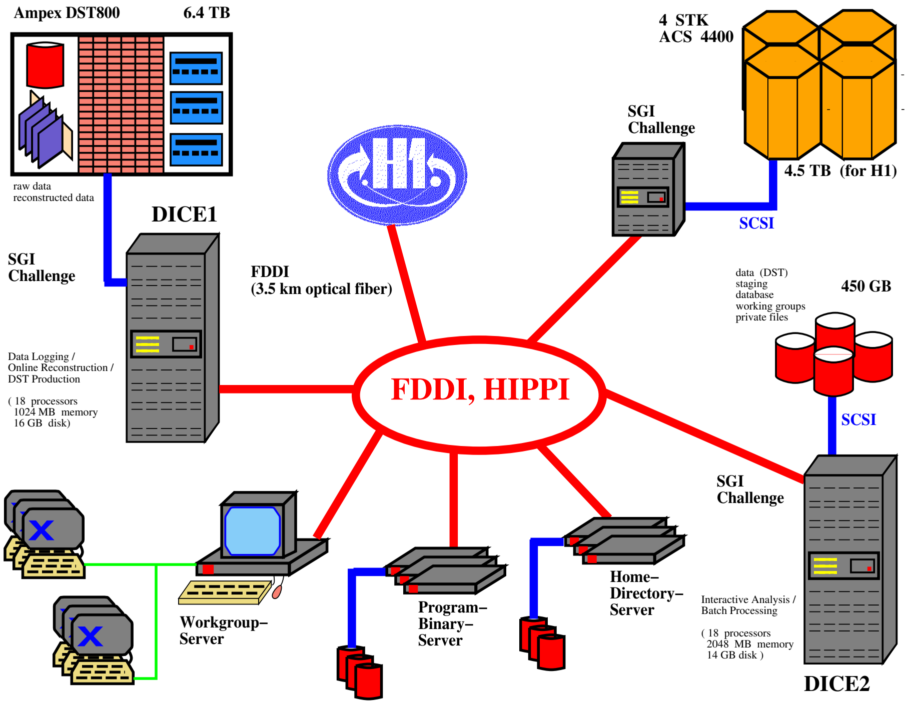


Figure 1: Computing environment of the H1 experiment

facilities are provided by two SGI Challenge XL computers with 18 MIPS R4400 processors each, used for *online* purposes (data logging, online reconstruction, reprocessing), and *analysis* purposes (batch and interactive), respectively. Data that need frequent and simultaneous access by many users is stored on about 500 GB of disk space, connected to the Challenge via SCSI interfaces. Tape service is provided by an Ampex DST800 TeraStore mass storage system with 3 DST600 recorders, based on helical scan D2 technology, and 4 StorageTek ACS 4400 tape libraries, with about 15 % of its storage capacity and 25 % of its I/O capacity reserved for the H1 experiment. Ampex tapes can be written and read with more than 10 MB/s from memory and about 3 MB/s including disk access. The capacity per tape amounts to 25 GB, the total robot capacity is 6.2 TB. Direct access to the StorageTek silos is provided by the Open Storage Manager (OSM) software[1]. The network infrastructure consists of FDDI and HIPPI with a Netstar Gigarouter as the central router. The H1 experiment is connected via a direct FDDI link to the Gigarouter as well[2]. Recently, the H1 collaboration joint the DESY workgroup server project[3,4], aiming at separated services for batch and interactive processing, home directory and program binary services.

## 2 Data Sources

The H1 detector and its data acquisition system are described elsewhere[5,6]. The data flow in the H1 experiment is shown in figure 2. Various levels of hardware

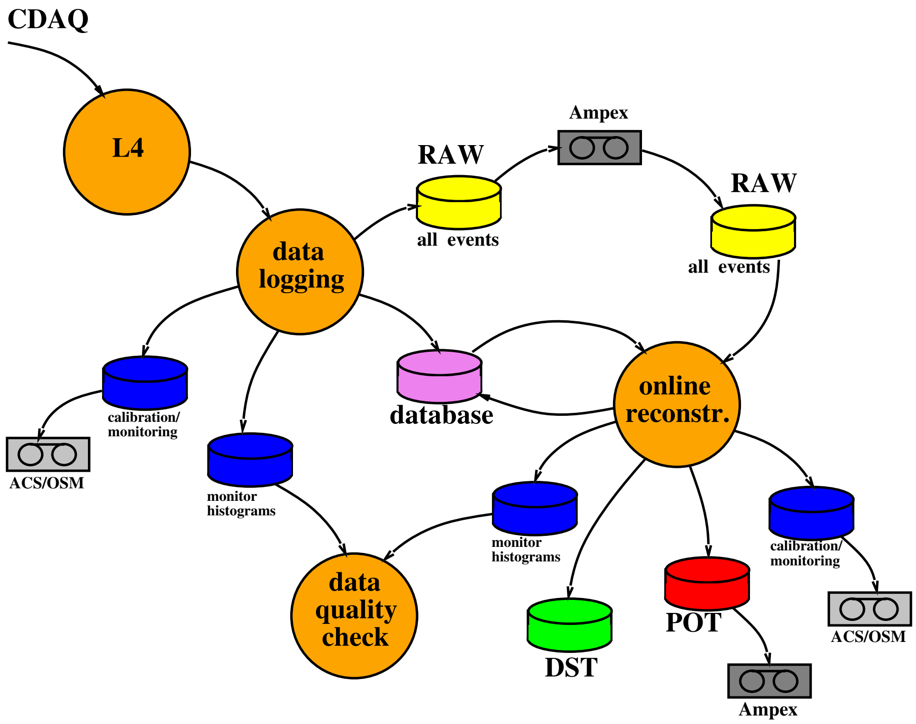


Figure 2: Data Flow in the H1 experiment

triggering, software filtering, and digital suppression are used to reduce about 3 MB of raw digitized information per event to acceptable recording rates. A parallel array ("filter farm") of 34 MIPS RISC processor based monoboard computers[7] is serving as a final online software trigger. Its main tasks are the recognition of, in particular proton-beam related, background, the determination of calibration parameters relevant for later steps in data processing, and the monitoring of the detector status online. The maximum input rate is 18 MB/s but presently limited to about 5 MB/s due to the mean processing time of approximately 0.5 s per event. The filter algorithms reject typically 80 to 90 percent of the events, yielding a reasonable data logging rate of 0.5 MB/s on average.

The data logging is performed using a dedicated SUN SparcStation server, integrated in VME. Data is read from L4 memory and sent to an SGI Challenge XL

Table 2: Data volumes

| | 1994 | nominal operation |
|---|---|---|
| RAW | | |
| number of events | $7.0 \times 10^7$ | $10^8$ |
| event size | $\sim 40\, KB$ | $\sim 50\, KB$ |
| data volume | $\sim 2500\, GB$ | $\sim 5000\, GB$ |
| POT | | |
| number of events | $1.2 \times 10^7$ | $2.5 \times 10^7$ |
| event size | $\sim 120\, KB$ | $\sim 120\, KB$ |
| data volume | $\sim 1500\, GB$ | $\sim 3000\, GB$ |
| DST | | |
| number of events | $1.2 \times 10^7$ | $2.5 \times 10^7$ |
| event size | $\sim 6\, KB$ | $\sim 6\, KB$ |
| data volume | $\sim 75\, GB$ | $\sim 150\, GB$ |

computer (dice1) via an FDDI link, using standard TCP/IP protocols[2,8]. Data is checked for internal consistency, converted to the standard H1 offline data format[9], and written to helical scan D2 tapes in the Ampex storage system *(raw data)*. Selected events, used for monitoring and calibration purposes, are copied to 3490 cartridges in the StorageTek silo. Calibration information, determined previously on the filter farm, is stored into the H1 database (on the DESY IBM, access via RPC).

The *transfer rate* using FDDI between the SUN SparcStation and the SGI Challenge, including TCP/IP protocol evaluation, is about 8 MB/s. Access to the VME memory on the filter farm side reduces the rate to about 5 MB/s which is the maximum possible data logging rate. Presently, about 2.0 MB/s are reached including the data logging task, i.e. data formatting, disk and tape access etc. The total number of events and the corresponding raw data volumes are presented in table 2 for the 1994 data taking period and nominal operation ($10^7\, s$ at 10Hz).

The *online reconstruction*[8] is performed on an SGI Challenge XL computer with 18 R4400 processors (150 MHz), starting from the *raw data tapes*. All H1 data is reconstructed within a few hours after data taking, making use of the calibration constants determined on the filter farm. The processing includes event classification and background rejection steps, based on reconstructed quantities. Classified events are written to tapes in the Ampex mass storage device *(POT data)*. DST events are written to disk for easy and parallel access for physics analysis *(DST data)*.

At a processing time of 2 s per event, a processing rate of 8 Hz is achieved which enables the reconstruction to keep up with the data taking rate of 10 Hz within a few hours. The number of events and the respective data volumes for 1994 data taking and for nominal operation are summarized in table 2.

The H1 detector simulation program is assembled within the GEANT framework. Monte Carlo production is mostly performed on workstation clusters outside DESY. Data transport is organized via DLT and Exabyte. All data is collected centrally at DESY in the StorageTek silo and on disk in form of DST data. The bookkeeping of all files is performed centrally at DESY, an implementation into the Oracle framework is being worked upon. A Monte Carlo production system with fully automized event distribution in a heterogeneous environment, based on PVM, is presently being developed, similar to Funnel[10]. The yearly data volume, assuming effectively 25 weeks of production, amounts to $2.5 \times 10^6$ ($12.5 \times 10^6$) events or about $500\,GB$ ($2.5\,TB$) presently (planned for end of 1995).

## 3 Data Access

Data access is implemented in a general package for machine independent input and output, F-package[9]. It provides a simple and unique system for data transfer with a common user interface for all operating systems used within the experiment. Among the important features are *i)* automatic wordformat conversion between different machine representations, *ii)* record selection options, *iii)* a network server for remote file access, *iv)* index file (event directory) based fast access to a subset of large datasets, *v)* ordered (simulated keyed) access for database applications, and *vi)* interfaces to FATMEN (used as a machine independent file catalogue), SHIFT[11] (for access to data on disk), to stage data from Ampex, ACS/UNIX (OSM), and ACS/MVS tapes[1], to access DLT and Exabytes, and to PVM for Monte Carlo mass production. All event (BOS) data and LOOK histogram data are stored in F-package format.

To ensure an efficient environment for physics analysis, data must be accessible by many users simultaneously, i.e. on random access devices. As a further requirement, the data of several years should be kept on disk for combined analyses. The SHIFT software is used to manage the disk space and to access disk files. The disk space is divided into a large pool used for general datasets (260 GB), dedicated pools for the different physics working groups (50 GB), and a pool for tape staging (70 GB). To save disk space, software compression algorithms are used, using apriori knowledge of significant number of bits based on detector resolutions.

Mass data (raw, reconstructed, Monte Carlo data), however, are only available on tape. Multi-user access to these data is only possible via staging onto disk. The StorageTek silos can be accessed both from the IBM/MVS mainframe computer and the SGI/UNIX cluster. Access to tapes in the MVS catalogue from the UNIX side needs (batch) job submission on the mainframe which is relatively slow and phasing out. (Direct) access to tapes in the silo is performed using the Open Storage Manager Software. Raw and reconstructed data, which are stored in the Ampex mass storage system, are accessible as well.

All H1 data is stored at DESY. External sites keep private copies of (subsets) of data. The tape media used for export and import are 3480/3490 cartridges (not recommended any longer, phasing out), Exabytes (main medium, recommended for small and medium size data), and DLT (starting, recommended for medium and large data samples). The expected data volumes for export are 500 GB per month (mainly DST and subsamples of data and Monte Carlo files) and for import 200 GB per month (Monte Carlo files). An external reprocessing of (raw) data is being investigated presently and will require a strong increase of the export/import facilities.

## 4 Conclusion

The three years since its commisioning have shown that the H1 experiment is able to cope successfully with the challenging requirements for computing at HERA. A computing environment for physics analysis has been established which allows H1 physicists to access all levels of H1 data in a convenient way.

A few topics are worth to be addressed when talking about computing for the next millennium. A unique solution to all mass storage problems, using a hierarchical storage system, is desirable. Users should only specify what data they need, no matter where the data is stored. A solution to manage maybe Petabytes of data in a database environment, allowing finally access to each individual byte in the data, is a challenge for the next millennium.


## Acknowledgments

The authors gratefully acknowledge the tremendous efforts by the DESY computer center groups in providing a convenient and usable computing facility for H1. The authors also express their gratitude to their colleagues in the H1 collaboration for many fruitful discussions. The authors particularly wish to thank Vicky White and Alberto Santoro and all their co-workers from Lafex and Fermilab for organizing an exciting conference at such a splendid place as Rio de Janeiro.


## References


1. M. Gasthuber, "Distributed Mass Storage and Management Systems at DESY", Proc. Int. Conf. on Comp. in High Energy Physics, San Francisco, USA, 1994
2. P. Fuhrmann, "HERA Data Acquisition over the Internet", these proceedings
3. K. Künne, "Interactive and Batch Analysis at DESY", these proceedings
4. F. Niebergall *et al.*, "Offline Computing at H1 – Experience and Future", these proceedings
5. I. Abt *et al.,* "The H1 Detector at HERA," DESY preprint 93-103 (1993), to be published in Nucl. Instr. Meth.
6. W.J. Haynes, "Experience at HERA with the H1 data acquisition system," Proc. Int. Conf. Comp. in High Energy Physics, Annecy, France (1992)
7. A. Campbell, "A RISC multiprocessor event trigger for the data acquisition system of the H1 experiment at HERA," Proc. Int. Conf. Real Time 1991, Jülich, Germany (1991)
8. P. Fuhrmann *et al.,* "Data Logging and Online Reconstruction in H1", Proc. Int. Conf. on Comp. in High Energy Physics, San Francisco, USA, 1994
9. V. Blobel, "The F-package for Input/Output", Proc. Int. Conf. Comp. in High Energy Physics, Annecy, France (1992)
10. B. Burow, "FUNNEL: Towards Comfortable Event Processing", these proceedings
11. J. Baud *et al.,* "SHIFT - User Guide and Reference Manual", CERN (1992)