H1oo - A Centralised Analysis Framework
for the H1 Experiment

Michael Steder on behalf of the H1 collaboration
DESY, Notkestr. 85, 22607 Hamburg, D
E-mail: michael.steder@desy.de

Abstract. The H1 experiment recorded an integrated luminosity of almost $0.5\,pb^{-1}$ ($0.3G$ events in HERA-I, $0.8G$ events in HERA-II) at the $ep$ collider HERA in the years 1992 through 2007. The goal of the H1oo project is to provide a data storage and analysis environment and in particular standardised and fast event selection facilities that improve the overall H1 physics analysis efficiency and performance. The software is based on and embedded in the ROOT framework, which provides a generic collection of tools for high-performance I/O jobs, as well as for the interactive analysis environment and for the event display. In this coherent approach all different physics analyses employ the same data and MC samples, which provide the users with the best possible knowledge of the H1 collaboration. The analysis files are produced centrally in a very efficient semi-automatic procedure, allowing a very short turn-around time when installing new reconstruction and analysis software versions.

1. Introduction

The Analysis Framework H1oo revises previous analysis paradigms in various respects. It provides analysis-ready data rather than mere collections of functions and algorithms. The H1 collaboration has continuously improved a common, extendable and re-usable framework in which the latest alignment and calibration parameters as well as standardised physics algorithms are accessible for all users. All data files for testing and analysis are produced centrally in a semi-automatic procedure, the production automatically includes the application of alignment and calibration as derived from the detector experts. This effort substantially enhanced the physics capabilities of official H1 software and - at the same time - reduced the turn-around time of physics analyses. For transparent exchange of algorithms between different working groups and portability of code between the different stages of data production and physics analysis, it is vital to have a homogeneous framework in which one common programming language and coding convention is used.

2. The H1 Detector

Charged particles emerging from the $ep$ interaction region are measured by the central tracking detector (CTD). The CTD consists of two large cylindrical central jet drift chambers (CJC) which are interleaved by a $z$-chamber and arranged concentrically around the beam-line in a magnetic field of 1.16T. A 5-layer multi-wire proportional chamber (CIP) is situated inside the inner CJC. The central silicon tracker (CST) is installed close to the interaction point and consists of two layers of double sided silicon strip sensors.
| Level   | Size per Event | Total Size  |
|---------|----------------|-------------|
| POT     | 200.0 kB       | 200.0 TB    |
| DST     | 15.0 kB        | 15.0 TB     |
| ODS     | 0.5 kB         | not stored, interface to DST |
| µODS    | 3.0 kB         | 1.5 TB      |
| HAT     | 0.5 kB         | 0.2 TB      |

Table 1. File size on the different levels of abstraction. The physics analyses are based on the condensed information in the HAT and µODS files, however pointers to the full detector information are implemented on all levels.

Figure 1. A map of the H1 detector as obtained from a nuclear interaction (NI) finder. Visible are the elliptic beam-pipe, the φ-segmented forward silicon tracker and the five layers of the CIP.

Figure 2. DST7 achieves a systematic uncertainty on the measurement on Θ_{Trk} of well below 1 mrad, indicated as solid line as function of the difference of the tangent of the dip angle λ = π/2 – Θ.

Charged and neutral particles are measured in the liquid argon calorimeter (LAr) which surrounds the tracking chambers and in a lead/scintillating-fibre calorimeter (SpaCal) in the backward region. The calorimeters are surrounded by the solenoidal magnet and the iron return yoke. The yoke is instrumented with 16 layers of limited streamer tubes, forming the central muon detector (CMD). A detailed description of the full H1 detector can be found in [1].

3. Simulation and Reconstruction
The raw data recorded with the H1 detector at HERA were written to POT (Physics On Tape) files (cf. tab. 1), containing among other things wire hits, channel numbers and cell energies as well as a first reconstruction of cells and tracks. In a next step the events are fully reconstructed with the H1 reconstruction software h1rec and information relevant for analysis is copied to the DST (Data Summary Tape) files. Simulated physics events are generated using various Monte Carlo (MC) programs and passed to the Geant 3 based H1 detector simulation. After the simulation step, the MC events are reconstructed using the same reconstruction software as for data. Over the years, the H1 reconstruction software has been improved continuously. With the latest data reprocessing ‘DST 7’ the H1 software has reached its final precision. Tracks in the central tracking detector are now searched for using a broken line fit treating the transition region between inner and outer jet chambers as thick scatterer, the silicon tracking detectors are...
Figure 3. Correlation plot of generated and reconstructed electromagnetic fraction of a cluster ($f_{em,jet}$) as obtained from the new cluster separation algorithm [3].

included in the vertex reconstruction and dE/dx information is used as mass hypothesis during track reconstruction. The dead material description in the MC simulation has been tuned using the output of a nuclear interaction (NI) finder, which identifies secondary vertices originating from the interaction of particles with the detector material. A detector map created with the well performing NI finder is shown in figure 1. After all the z-vertex resolution is heavily improved, a systematic uncertainty of 1% per track and of less than 1 mrad on $\Theta_{Trk}$ (cf. fig. 2) is achieved for the CTD, allowing to employ a very high precision for the final H1 publications.

4. Production of H100 analysis files
4.1. Data Structure
The H100 data is organized in a three-layer structure, corresponding to different ROOT[2] trees. The data stored in the different layers are produced by so-called filling code, which is organized in modular units called finders. The algorithms and the data are organized in about 600 classes in around 50 packages all inherited from TObject (ROOT) and implemented in C++. With this H100 provides a collaboration wide standard of doing event and particle reconstruction and selection. The base layer, ODS (Object Data Store), is an ROOT-format interface to the full information stored in the DST files. The middle layer, $\mu$ODS (micro ODS), consists of particle level four-vector information and contains all information needed by physics analyses. The uppermost layer, HAT (H1 Analysis Tag), contains event level information like particle numbers and event kinematics for fast selection of events.

4.2. Cluster Separation, Alignment, and Calibration
In order to improve the software compensation of the LAr calorimeter, a neural network has been trained to improve the separation of electromagnetic and hadronic clusters. This allows an almost perfect reconstruction of the electromagnetic fraction as shown in figure 3. A detailed description of this procedure can be found in [3]. Applying the cluster separation yields a better starting point for the subsequent calibration, since the measured energies are already close to the true ones.

In the latest H100 release, a new calibration scheme has been implemented unifying different calibration methods and making the application to data and MC fully transparent for all users and very easy to access. After calibration the electron energy uncertainty is well below 0.5% and a jet energy uncertainty of less than 1% has been achieved as shown in figure 4, where the (double) ratios of the transverse momentum of the hadronic final state ($P_{T,h}$) and of $P_{T}$ calculated from the double-angle method\(^1\) are plotted. In order to achieve such a precise jet

\[^1\] The reconstruction of kinematic variables in the double-angle method employs the scattering angle of the electron $\theta_{e}$ and the inclusive hadronic angle $\gamma_{h}$. A more detailed description of the method is available in [3].
energy reconstruction, a new hadronic calibration package has been developed. In a first step, all clusters are calibrated, and in a second step, all clusters inside jets are fine-tuned as function of $\Theta_{\text{jet}}$ and $\eta_{\text{jet}}$ using an unbinned $\chi^2$ method in order to avoid steps at bin edges.

4.3. Particle Finders and Physics Algorithms

During the production of the analysis files, particle finders run on the reconstruction output creating objects for identified particles (e.g. electrons, muons) and composed particles (e.g. D*, Jets, J/$\psi$, K0) in the $\mu$ODS files. The created objects allow a fast, efficient, and user-friendly access to all relevant information of the particular particle. The structure of these objects is illustrated in figure 5. The code of the individual particle finders is developed and maintained by the experts of the collaboration and is subject of regular enhancements.

5. Analysis Level Software

5.1. Equality in Diversity

The common software development in the H1 collaboration does not end at analysis level, though all users can write their own code to access data and MC events in the HAT and $\mu$ODS files. Many aspects are similar or even identical for all high energy physics analyses - e.g. event selection, filling and binning of histograms. Within H1oo exists a framework which decomposes all these aspects of physics analyses into dedicated classes like a ‘histogram manager’ or an ‘event selector’. This makes it easily possible to migrate (parts of) one analysis into another one and to reuse existing code.

5.2. H1Calculator

Another H1oo package, the H1Calculator, provides access to many event and particle quantities. Its modular design makes it easy to extend the H1Calculator functionality by adding new classes. The H1Calculator is implemented as singleton to ensure self-consistency for all access to HAT/$\mu$ODS variables or to variables composed of these. This is in particular important for systematic studies, where for example a shift in the electron energy results in a corresponding change of the total calorimetric transverse momentum.
6. MC Production on the Grid

After final data reprocessing (DST 7), 2.8 billion MC events for physics analyses have been simulated and reconstructed by the H1 MC team within few months. This number could only be achieved by a very fast and efficient production scheme employing the Grid infrastructure in Europe and Russia. In addition, the powerful H1 batch farm plays an important role for the production of the numerous small requests with less than $10^6$ events, since the overhead on the batch system to produce these small requests is small compared to the grid. All MC requests are registered in a central data base and the DST and H1oo files are produced centrally by the MC team. This scheme allows a production of up to 0.5 billion MC events per month, the steadily increasing average number of MC events per month is shown in figure 6.

7. Data Preservation

The H1 collaboration is very active in the field of data preservation within the DPHEP initiative. H1 is aiming for ‘level 4’ of the Data Preservation Models[4]: preserving reconstruction and simulation software as well as basic level data to allow full data analyses in the future. In this context the full software chain will be put into a validation scheme suggested by DESY-IT (cf. fig. 7), in which the software is automatically recompiled on a regular basis. The reconstruction and H1oo file production is run and an automatised validation script checks all links of the software chain.

8. Summary

In the past decade the H1oo analysis framework has become the standard analysis software framework within the H1 collaboration. Its coherent design makes it easy to employ the most performing algorithms and allows a rather short turn-around time for physics analyses and data production. The common H1oo data format and analysis framework will serve as firm footing for the data preservation project in which the H1 collaboration plays a very active role.

References

[1] I. Abt et al. [H1 Collaboration], Nucl. Instrum. Meth. A 386 (1997) 310 and 348.
[2] R. Brun and F. Rademakers, Nucl. Instrum. Meth. A 389 (1997) 81.
[3] R. Kogler, PhD thesis, University of Hamburg (2010), H1 thesis 590 (available at http://www-h1.desy.de/psfiles/theses/h1th-590.pdf).
[4] D. M. South (on behalf of the ICFA DPHEP Study Group), arXiv:1101.3186.
**H1oo**

A Centralised Analysis Framework for the H1 Experiment

Michael Steder, DESY

H1oo revises previous analysis paradigms in various respects. It provides analysis-ready data rather than mere collections of functions and algorithms. The H1 collaboration has continuously improved a common, extendable and re-usable framework in which both expert knowledge and standardised physics algorithms are accessible for all users. All data files for testing and analysing are produced centrally in a semi-automatic procedure, the production includes the application of the latest alignment and calibration constants derived from the experts of the particular subdetector.

This effort substantially enhanced the physics capabilities of official H1 software and - at the same time - reduced the turn-around time of physics analyses. For transparent exchange between different working groups and portability of code between different stages of data production and physics analysis, it is vital to have a homogeneous framework in which one common programming language and coding convention is used.

### Production of H1oo Analysis Files

#### Data Structure

Data from reconstruction is transferred to a three-layer structure, as shown in figure 1. The different layers are separated row-wise and also depend on different ROOT maps. The data stored in the different layers are produced by several filtering codes which are organised in modules with called filters. The data is reconstructed in about 600 classes in around 50 packages, all derived from TObject (RooT), and implemented in C++. With this, the H1oo provides a collaboration-wide standard of doing event and particle reconstruction and selection.

#### Cluster Separation, Alignment, and Calibration

To improve the software framework of the LAr calorimeter even more, a neural network has been trained to improve the separation of electromagnetic and hadronic clusters. This allows for a correct reconstruction of the electromagnetic fraction as shown in figure 3. Applying the cluster separation also helps to remove events, since those clusters are not always close to the true events.

In the latest H1oo release, a new calibration scheme has been implemented using different calibration methods and making the application of data and MC file transparent for all users and easy to access.

#### Particle Finders and Physics Algorithms

During the production of the analysis files, particle finders run on the reconstruction output creating objects for identified particles (e.g. electrons, muons) and composed particles (e.g. D*, Jets, J/ψ, K∗), which can be used to reconstruct jet energy and make the application to data and MC fully transparent for all users and very easy to access.

#### Analysis Level Software

Data preservation in the H1oo collaboration does not end at analysis level, though. All users are able to write their own codes to access data and MC events in the DST and other files. Many experts are using Extraction H1oo which decouples all aspects of physics analysis into dedicated packages (e.g. "h1Rec", "h1Rec", "h1Rec") in the DST files. These codes allow a fast, efficient, and user-friendly access to all relevant information of the particle part. The physics algorithm is the knowledge of all experts of the collaboration and is subject to regular improvements.

### Simulation and Reconstruction (DST 7)

The raw data recorded at the electron-positron collider HERA are reconstructed on tape files, containing among other things, the primary vertices, the electrons, muons and photons. In a first step, the events are centrally reconstructed using the H1 reconstruction software h1rec and information relevant for analysis is copied to the DST (Data Summary Tape) file.

**Simulated physics events are generated using various Monte Carlo (MC) processes and passed to the Green (based H1 detector simulation).** After the simulation step, the MC events are reconstructed using the same reconstruction software as for the real data.

On average, the H1 reconstruction software has been continuously improved. With the latest data representation DST 7 the H1oo team has invested a lot of time and progress. This version of the central data production framework allows for using central file systems, making the transition between working groups and detectors a smooth and efficient process. The data produced in the simulation is efficiently stored in the DST files, and is also available for testing.

After reconstruction the reconstructed files are written to POT (Physics On Tape) files, containing among other things, the primary vertices, the electrons, muons and photons. During the production of the analysis files, particle finders run on the reconstruction output creating objects for identified particles (e.g. electrons, muons) and composed particles (e.g. D*, Jets, J/ψ, K∗), which can be used to reconstruct jet energy and make the application to data and MC fully transparent for all users and very easy to access.

#### Analysis Level Software

In the latest H1oo release, a new calibration scheme has been implemented using different calibration methods and making the application of data and MC file transparent for all users and easy to access.

#### Particle Finders and Physics Algorithms

During the production of the analysis files, particle finders run on the reconstruction output creating objects for identified particles (e.g. electrons, muons) and composed particles (e.g. D*, Jets, J/ψ, K∗), which can be used to reconstruct jet energy and make the application to data and MC fully transparent for all users and very easy to access.

**MC Production on the Grid**

After the data reconstruction (DST 7), the DST files are passed to a central software called h1Rec, which allows for the production of DST files on grid computing resources. This allows for distributed and efficient production for large datasets and reduces the storage and computing costs.

**Data Preservation**

The central file system in the H1oo collaboration is called the DST file system. The data produced in the simulation is efficiently stored in the DST files, and is also available for testing.

In the latest H1oo release, a new calibration scheme has been implemented using different calibration methods and making the application of data and MC file transparent for all users and easy to access.

The reconstruction and analysis production is run and the automated validation script checks all links of the software chain.

Figure 9: Illustration of the data preservation infrastructure as suggested by DESY-IT.

**Figure 8.** Thumbnail of poster - pdf file available at [http://www-h1.desy.de/posters/](http://www-h1.desy.de/posters/)