H1 Data Preservation Status
and Activities Since the Last Workshop

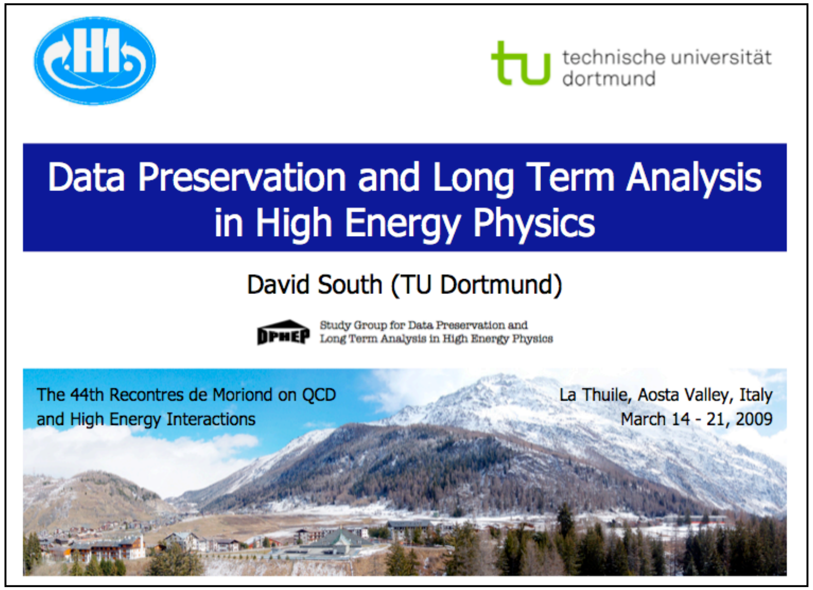
David South (TU Dortmund)

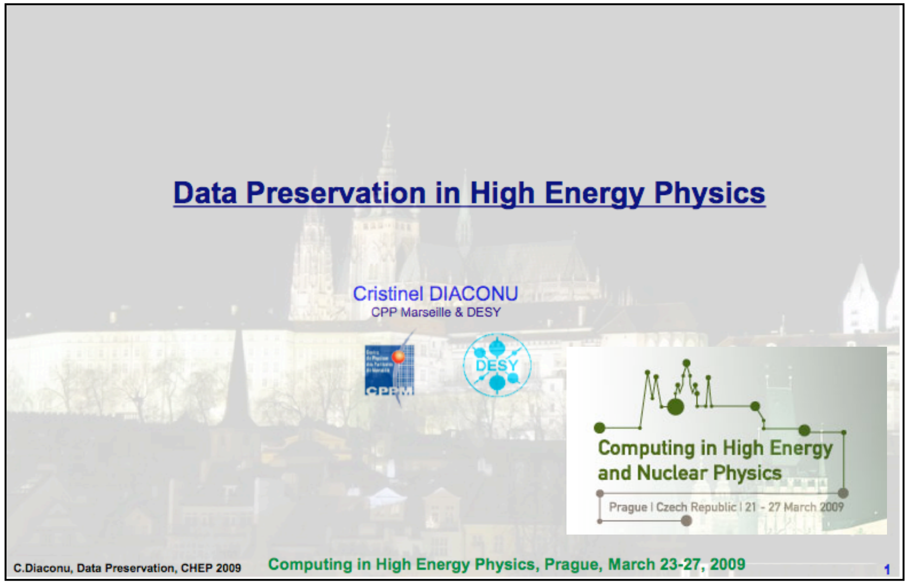
Cristinel Diaconu (CPPM), Roman Kogler (MPIM), Sergey Levonian (DESY),
Benno List (Univ. Hamburg), Bogdan Lobodzinski (DESY), Jan Olsson (DESY),
Dmitri Ozerov (DESY-IT), Daniel Pitzl (DESY), Michael Steder (DESY)

Second Workshop on Data Preservation and Long Term Analysis

SLAC NATIONAL ACCELERATOR LABORATORY 26 - 28 May 2009

Data Preservation Conference Presentations

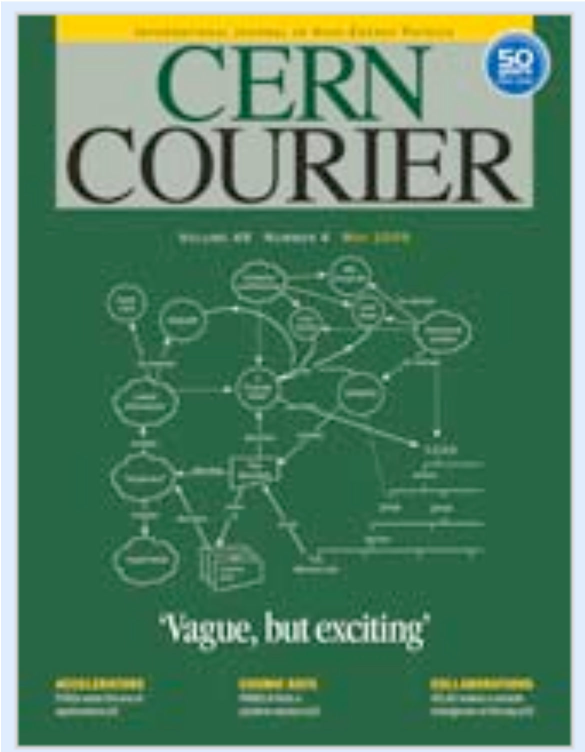




- Plenary presentations given this spring at *Recontres de Moriond QCD* (DS) and *Computing in High Energy Physics* (Cristi Diaconu) conferences
  - Broaden the visibility of the DPLTA project to a wider field
  - Both talks received positive feedback from an interested audience

- Hope for more coverage at the summer conferences: EPS in Krakow and of course Lepton Photon in Hamburg

CERN Courier Article



Study group considers how to preserve data

For experimentalists in high-energy physics, the data are like treasure, but how can they be saved for the future? A study group is investigating data-preservation options.

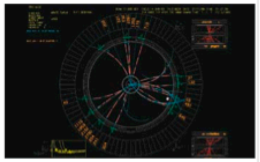
High-energy-physics experiments collect data over long time periods, while the associated collaborations of experimentalists exploit these data to produce their physics publications. The scientific potential of an experiment is in principle defined and exhausted within the lifetime of such collaborations. However, the continuous improvement in areas of theory, experiment and simulation – as well as the advent of new ideas or unexpected discoveries – may reveal the need to re-analyse old data. Examples of such analyses already exist and they are likely to become more frequent in the future. As experimental complexity and the associated costs continue to increase, many present-day experiments, especially those based at colliders, will provide unique data sets that are unlikely to be improved upon in the short term. The close of the current decade will see the end of data-taking at several large experiments and scientists are now confronted with the question of how to preserve the scientific heritage of this valuable pool of acquired data.

To address this specific issue in a systematic way, the Study Group on Data Preservation and Long Term Analysis in High Energy Physics formed at the end of 2008. Its aim is to clarify the objectives and the means of preserving data in high-energy physics. The collider experiments BaBar, Belle, BES-III, CLEO, CDF, DØ, H1 and ZEUS, as well as the associated computing centres at SLAC, KEK, the Institute of High Energy Physics in Beijing, Fermilab and DESY, are all represented, together with CERN, in the group's steering committee.

Digital gold mine

The group's inaugural workshop took place on 26–28 January at DESY, Hamburg. To form a quantitative view of the data landscape in high-energy physics, each of the participating experimental collaborations presented their computing models to the workshop, including the applicability and adaptability of the models to long-term analysis. Not surprisingly, the data models are similar – reflecting the nature of colliding-beam experiments.

The data are organized by events, with increasing levels of abstraction from raw detector-level quantities to N-tuple-like data for physics analysis. They are supported by large samples of simulated Monte Carlo events. The software is organized in a similar manner, with a more conservative part for reconstruction to reflect the complexity of the hardware and a more dynamic part closer to the analysis level. Data analysis is in most cases done in C++ using the ROOT analysis environment and is mainly performed on local computing farms. Monte Carlo simulation also uses a farm-based approach but it is striking to see how popular the Grid is for the mass-production of simulated events. The amount of data that should be preserved for analysis varies between 0.5 PB and 10 PB for each experiment, which is not huge by today's standards but nonetheless a large amount. The degree of preparation for long-term data varies between experiments but it is obvious that no preparation was foreseen at an early stage of the programs; any conservation initiatives will take place in parallel with the end of the data analysis.



A simulated event in the JADE detector, generated using a refined Monte Carlo program and reconstructed using revitalized software more than 10 years after the end of the experiment. (Courtesy Siggi Bethke.)

From a long-term perspective, digital data are widely recognized as fragile objects. Speakers from a few notable computing centres – including Fabio Hernandez of the Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules, Stephen Wolbers of Fermilab, Martin Gasthuber of DESY and Erik Mattias Wadenstein of the Nordic DataGrid Facility – showed that storage technology should not pose problems with respect to the amount of data under discussion. Instead, the main issue will be the communication between the experimental collaborations and the computing centres after final analyses and/or the collaborations where roles have not been clearly defined in the past. The current preservation model, where the data are simply saved on tapes, runs the risk that the data will disappear into cupboards while the read-out hardware may be lost, become impractical or obsolete. It is important to define a clear protocol for data preservation, the items of which should be transparent enough to ensure that the digital ▷

content of an experiment (data and software) remains accessible.

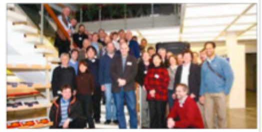
On the software side, the most popular analysis framework is ROOT, the object-oriented software and library that was originally developed at CERN. This offers many possibilities for storing and documenting high-energy-physics data and has the advantage of a large existing user community and a long-term commitment for support, as CERN's René Brun explained at the workshop. One example of software dependence is the use of inherited libraries (e.g. CERNLIB or GEANT3), and of commercial software and/or packages that are no longer officially maintained but remain crucial to most running experiments. It would be an advantageous first step towards long-term stability of any analysis framework if such vulnerabilities could be removed from the software model of the experiments. Modern techniques of software emulation, such as virtualization, may also offer promising features, as Yves Kemp of DESY explained. Exploring such solutions should be part of future investigations.

Examples of previous experience with data from old experiments show clearly that a complete re-analysis has only been possible when all of the ingredients could be accounted for. Siggi Bethke of the Max Planck Institute of Physics in Munich showed how a re-analysis of data from the JADE experiment (1979–1986), using refined theoretical input and a better simulation, led to a significant improvement in the determination of the strong coupling-constant as a function of energy. While the usual statement is that higher-energy experiments replace older, low-energy ones, this example shows that measurements at lower energies can play a unique role in a global physical picture.

The experience at the Large Electron-Positron (LEP) collider, which Peter Igo-Kemenes, André Holzner and Matthias Schroeder of CERN described, suggested once more that the definition of the preserved data should definitely include all of the tools necessary to retrieve and understand the information so as to be able to use it for new future analyses. The general status of the LEP data is of concern, and the recovery of the information – to cross-check a signal of new physics, for example – may become impossible within a few years if no effort is made to define a consistent and clear stewardship of the data. This demonstrates that both early preparation and sufficient resources are vital in maintaining the capability to reinvestigate older data samples.

The *modus operandi* in high-energy physics can also profit from the rich experience accumulated in other fields. Fabio Pasian of Trieste told the workshop how the European Virtual Observatory project has developed a framework for common data storage of astrophysical measurements. More general initiatives to investigate the persistency of digital data also exist and provide useful hints as to the critical points in the organization of such projects.

There is also an increasing awareness in funding agencies regarding the preservation of scientific data, as David Corney of the UK's Science and Technology Facilities Council, Salvatore Mele of CERN and Amber Boehnlein of the US Department of Energy described. In particular, the Alliance for Permanent Access and the EU-funded project in Framework Programme 7 on the Permanent Access to the Records of Science in Europe recently conducted a survey of the high-energy-physics community, which found that the majority of scientists strongly support the preservation of high-energy-physics data. One important aspect that was also positively appreciated in the survey answers was the question of open access to the data in conjunction with the organizational and technical matters, an issue that deserves careful consideration. The next-generation publications database, INSPIRE, offers extended data-storage capabilities that could be used immediately to enhance public or private information related to scientific articles, including tables, macros, explanatory notes and potentially even analysis software and data, as Travis Brooks of SLAC explained.



Participants of the first workshop on data preservation and long-term analysis in high-energy physics at DESY, Hamburg. (Courtesy DESY.)

While this first workshop compiled a great deal of information, the work to synthesize it remains to be completed and further input in many areas is still needed. In addition, the *raison d'être* for data preservation should be clearly and convincingly formulated, together with a viable economic model. All high-energy-physics experiments have the capability of taking some concrete action now to propose models for data preservation. A survey of technology is also important, because one of the crucial factors may indeed be the evolution of hardware. Moreover, the whole process must be supervised by well defined structures and steered by clear specifications that are endorsed by the major laboratories and computing centres. A second workshop is planned to take place at SLAC in summer 2009 with the aim of producing a preliminary report for further reference, so that the "future of the past" will become clearer in high-energy physics.

Further reading

For more information about the Study Group for Data Preservation and Long Term Analysis in HEP, see www.dphep.org.

Résumé

Les données à l'épreuve du temps

En physique des hautes énergies, l'amélioration continue de la théorie, des expériences et des simulations, l'éclosion de nouvelles idées ainsi que des découvertes inattendues peuvent faire naître le besoin de réanalyser d'anciennes données. Cela s'est déjà fait et pourrait devenir plus fréquent à l'avenir. Afin de faire le tour de la question, un groupe d'étude sur la préservation et l'analyse à long terme des données de physique des hautes énergies a été constitué à la fin 2008, composé de représentants des grandes collaborations travaillant sur des collisionneurs de particules et des centres informatiques associés. Le but est de définir les objectifs, ainsi que les moyens de préserver les données de physique des hautes énergies.

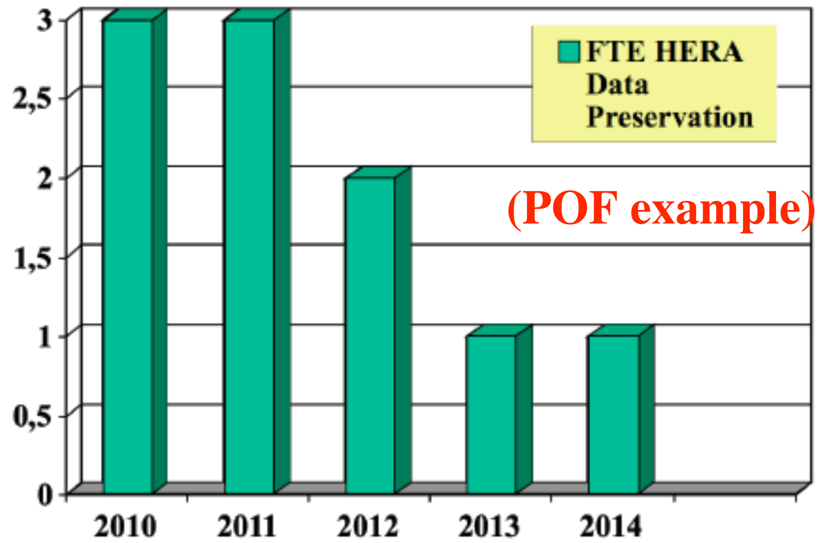
Cristinel Diaconu, *CPP Marseille and DESY Hamburg, and* **David South**, *Technische Universität Dortmund.*

- Two page article by CD and DS published in May edition of the CERN Courier, summarising the main points of the workshop in Hamburg

Securing the Financial Resources

- It was clear before and after the first workshop that such initiatives will require additional funding, in particular for manpower

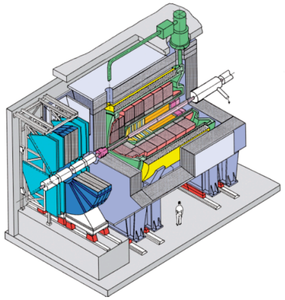
- The DESY "Program Oriented Funding" (POF) Committee fully supports the project and recommends 10 FTEs over a period of 5 years from 2010
  - to be decided in Autumn 2009



- DESY-IT have also applied for funding for a 3 year position, to begin immediately

H1 Data Analysis Model

+



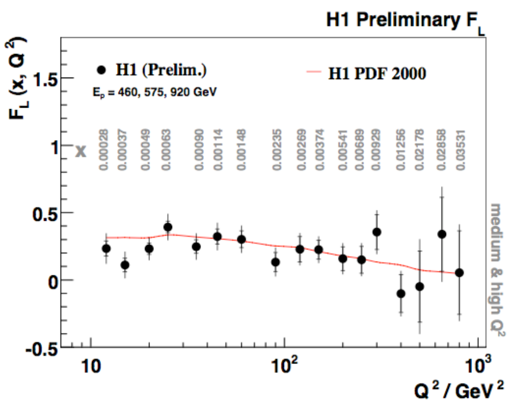
119.977679	129.534731	124.739135	176.316414
130.46875	135.839924	130.84732	168.289658
135.895502	149.510531	140.795689	120.686833
134.127052	140.495868	132.823819	206.138393
129.851598	137.880438	124.888856	189.675642
123.797241	131.84633	126.146789	202.496855
118.435374	130.691651	112.877008	140.366234
112.401212	121.561443	114.237637	125.298579
112.388488	128.496503	113.302591	192.223669
129.011813	138.880759	128.517198	108.701884
127.077465	139.289941	129.528986	127.406576
124.9785	135.363241	127.454638	129.669126
124.294035	133.242253	124.704841	244.567067
125.663717	135.159011	125.476984	169.271991
123.704853	127.612613	124.25382	170.401964
118.926697	122.818967	115.379664	134.970308
116.588208	121.798711	116.018173	323.148148
119.458869	124.788744	119.103839	204.736734
120.081967	124.847434	120.425321	289.50681
123.462329	127.367029	123.298233	287.632974
124.442179	128.115374	125.592252	362.764329
125.490169	128.448761	124.411031	382.978361
124.446597	128.898705	126.602473	358.369956

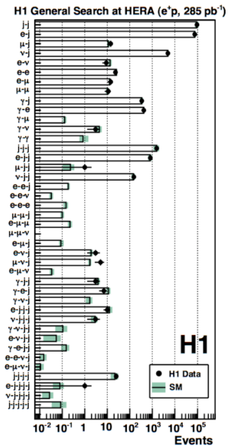
HERA delivered $e^{\pm}p$ collisions 1992-2007 and the H1 Collaboration collected 0.5 fb^{-1} of data, ~ 10^9 events

The RAW data output from the detector is written to tape

Raw data transformed into DST format using Fortran based software, regular re-processing

H100



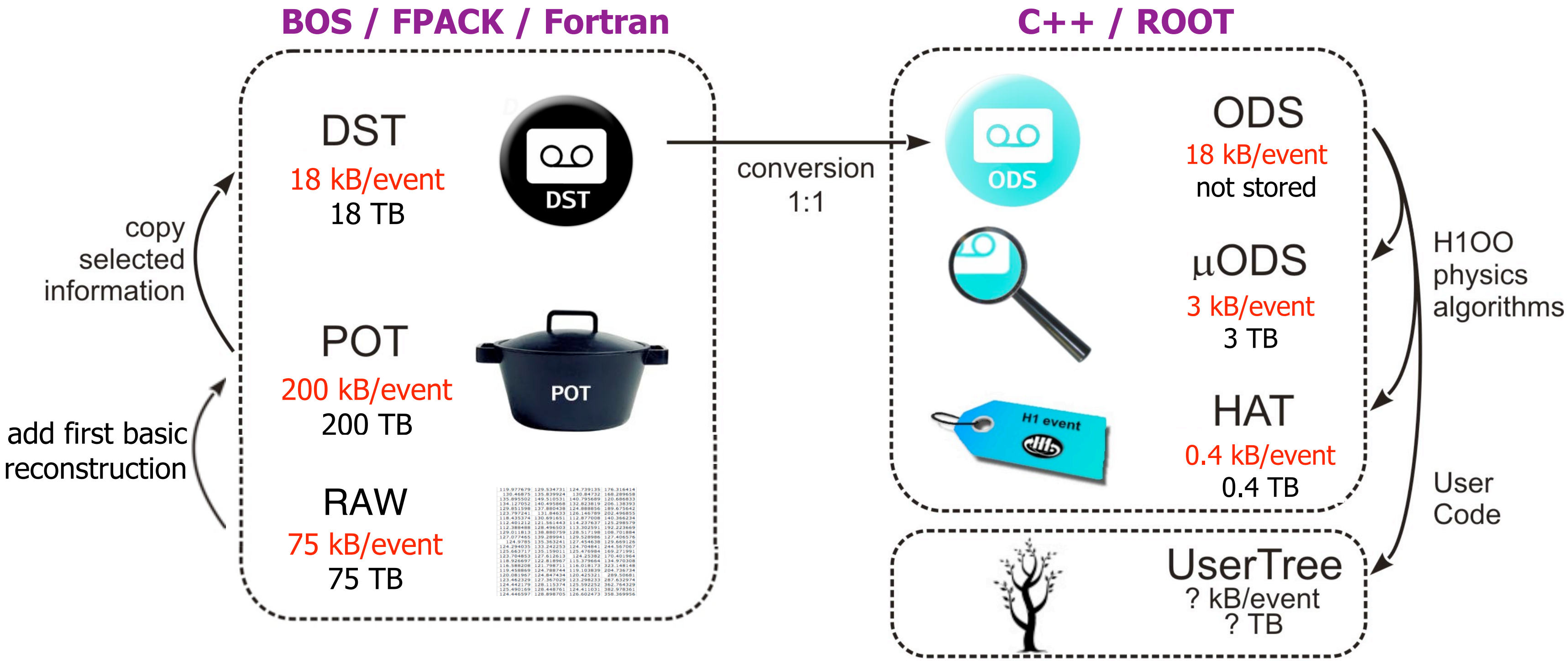


H1 publishes physics results

Regular common data and MC production, calibrations and analysis performed using central computing resources

Analysis level data format and software written in C++ and based on ROOT

H1 Data Format: RAW to the Analysis Level



Fortran based reconstruction software converts RAW data wire hits, channel numbers and collected charges into clusters of cells, energy sums, first track fits and vertices on the DST

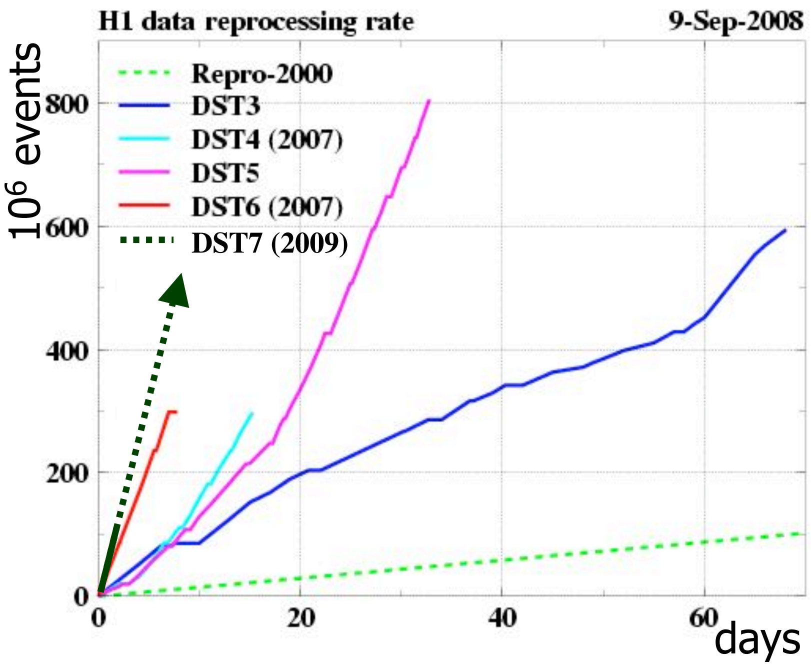
H1OO Analysis level data composed of 3 layers of ROOT files: Object Data Store (ODS, dynamic access of a 1:1 conversion of DST to C++ objects), a smaller version (µODS) and the H1 Analysis Tag (HAT)

DST 7 : The Final* H1 Reprocessing

*probably

- The preparation of DST 7 has been the main computing and software project at H1 since the first DPLTA workshop

- Many improvements prepared for use for the final precision H1 analyses; additional improvements at the analysis level



- DST 7 reprocessing of HERA II data begun **May 23rd**: 500M high energy + 300M low and medium energy run events

- Expect full HERA II data set to be available in about 3 weeks

- HERA I (250M events) to follow later this year

Plan for Data Preservation at H1

- **Data formats to be preserved**
  - RAW data of GM-cut files, total for HERA: 75 TB
  - At least one full set of data DSTs, total for HERA I+II: 18 TB
  - A version of μODS and HAT as well (< 4 TB)
  - In addition to calibration and cosmic runs, total data about 100 TB
  - Amount of MC to be decided, but will be of the same order

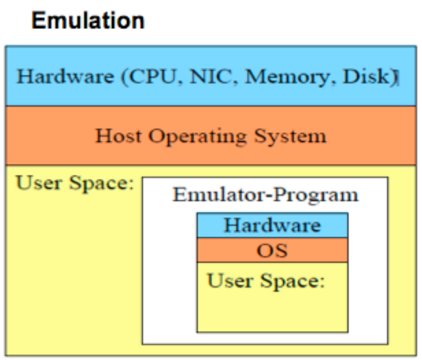
- Conservatively (x2) estimate total amount to preserve at 500 TB

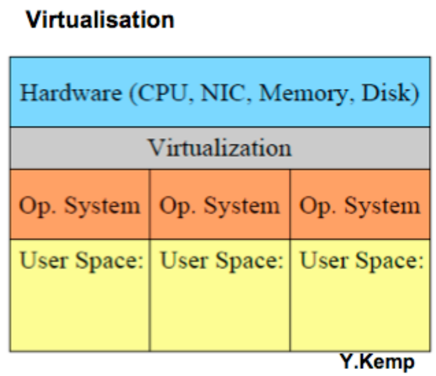
- Do not expect to be limited by CPU or disk space in the future
  - Preserved data/MC should be copied on to new media at regular intervals, say every 2 years
  - Expect cost of data migration to be double current costs:

    1 + 1/2 + 1/4 + 1/8 + .. = 2

Plan for Data Preservation at H1

- **Reconstruction / Simulation Software: latest DST version**
  - Mostly written in Fortran (some C, some C++), generally easier
  - Some parts already frozen since a good few years
  - Basic data format FPACK/BOS designed as machine independent
  - Database access via Oracle, could be fazed out in preservation model
  - No further major development after DST 7: *but should still be possible*
  - **Migration of OS:**
    - IBM to UNIX conversion already done in 1996
    - Since then a few Linux conversions, SL4-5 transition at DESY now
    - Efforts now ongoing to investigate the impact of (64 bit) SL5, including possible use of virtualisation techniques and adaptation of existing (SL4) executables





Plan for Data Preservation at H1

- **Physics Analysis Software**
  - Written entirely in C++ language
  - Unlike reconstruction software, further development is planned for the coming years, with DST 7 as the input
  - Model heavily reliant on ROOT framework, in particular I/O, TTree
  - Could try to remove as much ROOT as possible from H1OO, leaving only the crucial dependencies (H1Skeleton package..)
  - But most classes (usefully) inherit from TObject, not a good solution
  - Try to incorporate ROOT updates: ROOT developers to patch H1OO?

- SL5 compatibility with H1OO to be studied in coming months (stricter C++ compiler usually causes significant problems)

- *After development, foresee a "rolling preservation model" for the analysis software, with regular recompilation of H1OO software and μODS/HAT file production, say every 3 months*

Data Conservation Levels

	Minimum Level of Preservation
0	RAW data
1	Reconstruction Simulation Database considerations? Commercial software?
2	DST
3	Ntuple / analysis level data (and MC?) ***production***
4	Existing ntuple / analysis level
5	Combined analysis with a (for example) H1+ZEUS "ep ntuple"
6	Outreach : very simple format

The basic level to conserve

Essentially frozen, but ensure reconstruction software still compiles, so changes still possible. A new simulation: can it use old reconstruction (issue of F vs C++)?

Essentially frozen, DST 7 the "final" reconstruction software version

Rolling model, fluid preservation from here up: gives regular verification of full chain

Fixed ntuple, "all" analysis level info

Common format ntuple (repository?)

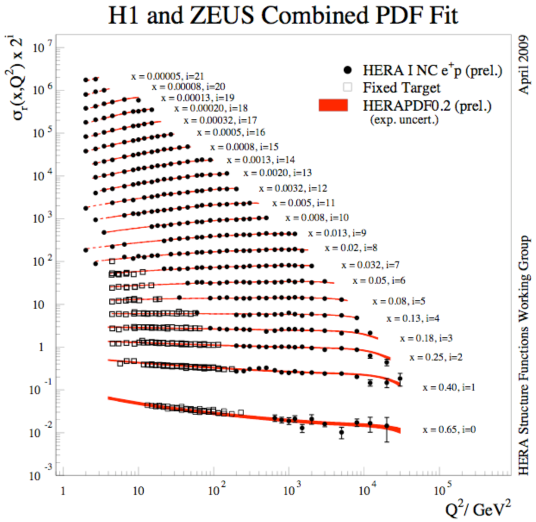
Not enough for full analysis(?), but rather for open access / outreach

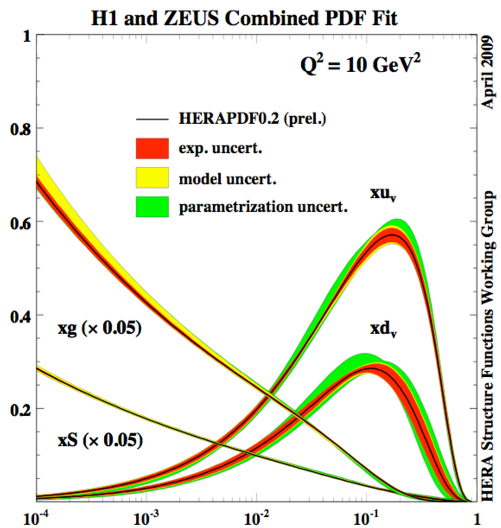
Use of Virtualisation / Emulation techniques ?

H1+ZEUS Combined Analysis

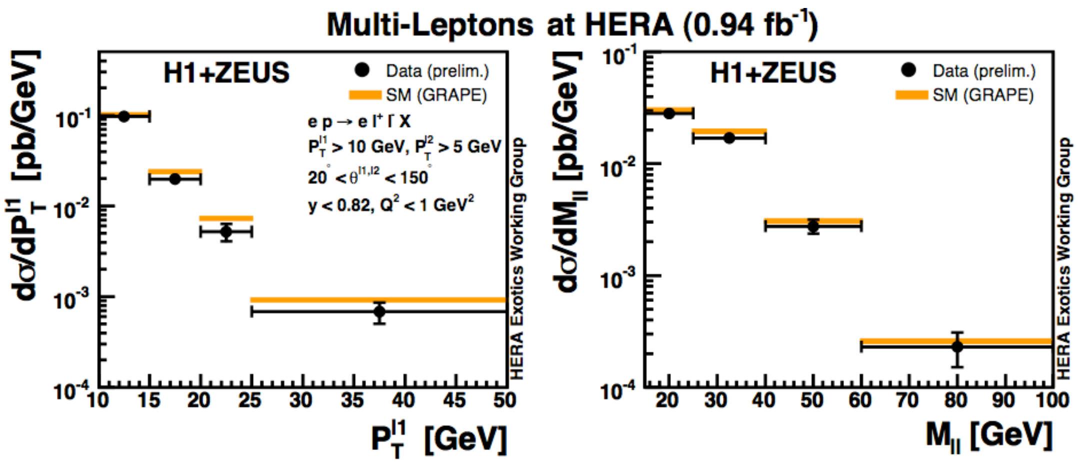
- Many combined H1 and ZEUS analyses and groups now active
  - Searches (high P_T lepton events); jets and α_s, diffraction, D^* events and structure functions and PDF fits

- Improvement from combined data set seen in much reduced uncertainty on PDF fits





- So far combined analyses performed by combining H1 and ZEUS histograms or even numbers rather then running a single true combined analysis



- Fully coherent combined analysis possible through common data format: Made up of HAT and four vectors from finders in μODS, also could be independent of ROOT. Start with ZEUS ntuple?

Planning the Future of the H1 Collaboration

	A	B	C
Spirit	It will work by itself	Plan it while we are around	After some point there will be no interest anyway
Governance	Collaboration	Expert committee	H1 Stops
Structures	Spokesperson, Physics Coordinators Collaboration Board Conveners	H1 Physics Committee Chair of H1 Phys. Co.	-
Data	Preserved	Preserved	"Stored"
Data Users	H1 members, as usual	Individuals and groups (including at least an H1 expert)	None (random)
Functioning	Regular meetings of all structures (slower rate)	Annual Physics Review	-
Author list	H1 members as usual	Editors + some H1 members (no default list)	-
Risk Analysis	Resources?	Reactivation resources?	Physics loss?

Summary

H1 reconstruction software "final" version achieved this month, DST 7, incorporating the best knowledge from over 20 years of development in a stable modular Fortran structure

H1OO analysis framework and data format based on ROOT used by over 90% of H1 analyses, resulting in better efficiency for physics results: development to continue for a while

Common, coherent data files and coordinated large scale MC production on the GRID contributes to a successful analysis model at H1

H1 Data Preservation Task Force set up to address the issue of H1 data and software preservation, first ideas of which presented today

> Need the dedicated manpower to oversee data preservation project, including a big boost in the documentation, which is sometimes in good shape thanks to diligent authors and the Optimal use of ROOT provides much html documentation

Future format of the H1 Collaboration itself, beyond 2013, also to be decided as well as open access to the H1 data