HERA Data Preservation.

Report on behalf of the DESY Data Preservation Group

David South (DESY)

PRC 70 Open Session

DESY-Zeuthen, 14th October 2010

HELMHOLTZ ASSOCIATION

DESY

Contents

DPHEP: International Study Group on Data Preservation

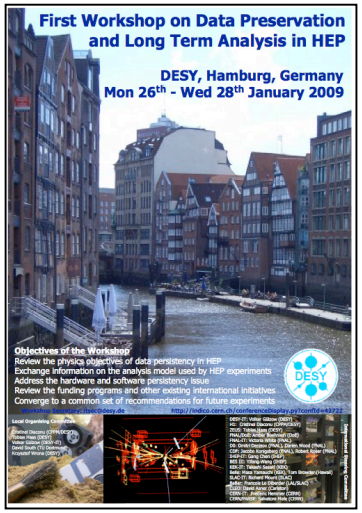
DPHEP Study Group for Data Preservation and Long Term Analysis in High Energy Physics

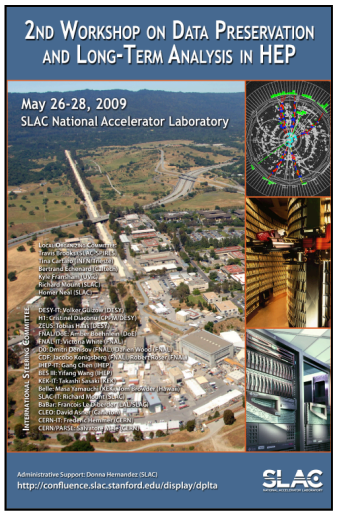
- Group has grown since 2008 to around 70 contact persons
- Endorsed by ICFA summer 2009

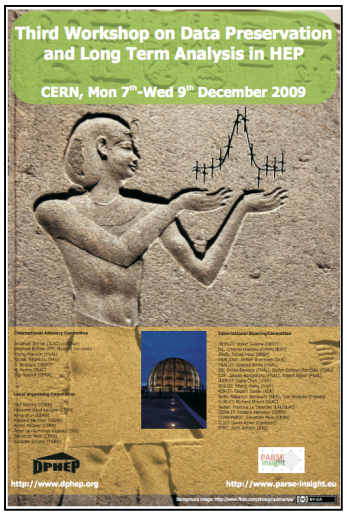
- Chair: Cristinel Diaconu (DESY/CPPM)
- Working Groups
  - Physics Cases: François Le Diberder (SLAC/LAL)
  - Preservation Models: D. South (DESY), Homer Neal (SLAC)
  - Technologies: Stephen Wolbers (FNAL), Yves Kemp (DESY)
  - Governance: Salvatore Mele (CERN)
- International Steering Committee
  - Participants from ee, ep and pp collider experiments
  - Associated computing centers at the labs
  - Some funding agencies
- International Advisory Committee
  - Chairs: Jonathan Dorfan (SLAC), Siegfried Bethke (MPIM)
  - Advisers: Gigi Rolandi (CERN), Michael Peskin (SLAC), Dominique Boutigny (IN2P3), Young-Kee Kim (FNAL), Hiroaki Aihara (IPMU/Tokyo), Alex Szalay (JHU)

DPHEP Activities

- First contacts established in September 2008
- Series of DPHEP workshops held since 2009
  - DESY (Jan 2009)
  - SLAC (May 2009)
  - CERN (Dec 2009)
    preceded by a public symposium
  - KEK (July 2010)



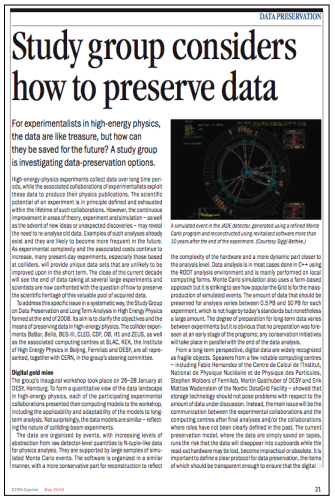




- Confront data models, clarify the concepts, set a common language, investigate technical aspects, compare with other fields such as astrophysics and others handling large data sets
- With the ultimate aim of providing a set of recommendations concerning data preservation for past, present and future HEP experiments

DPHEP Publications and Visibility

- First DPHEP publication in December 2009, **arXiv:0912.0255**
  - Initial survey of the HEP data landscape, and the physics case for preservation
  - Different models of preservation identified, as well as addressing future governance
- Second publication, “Blueprint for Data Preservation in HEP” *in preparation*
  - Executive summary written at KEK workshop, presented to ICFA at ICHEP 2010
  - More detailed description and comparison of analysis models of HEP experiments
  - Future organisation of the DPHEP Study Group, including a dedicated DPHEP Project Manager
  - Specific examples of projects developing between the experiments, with FTE estimates
- Visibility of the DPHEP Initiative increasing
  - Physics conferences Moriond QCD 2009, QCD 2010
  - Plenary talks at CHEP 2009 and 2010
  - Articles in physics standards and German national press


DATA PRESERVATION

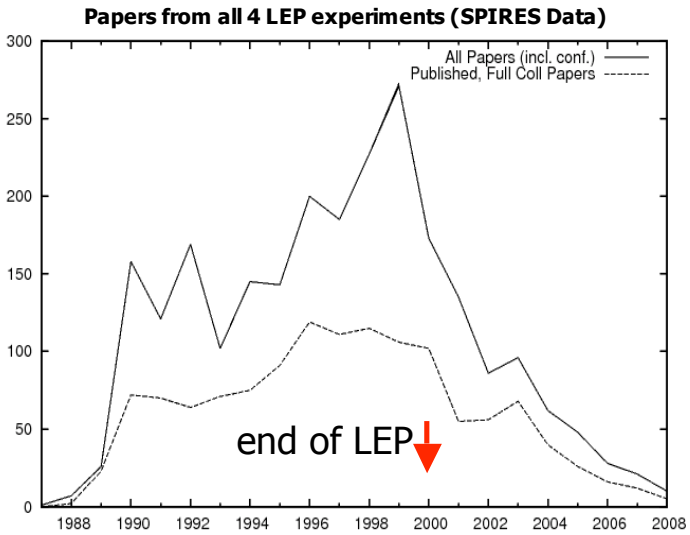
Study group considers how to preserve data

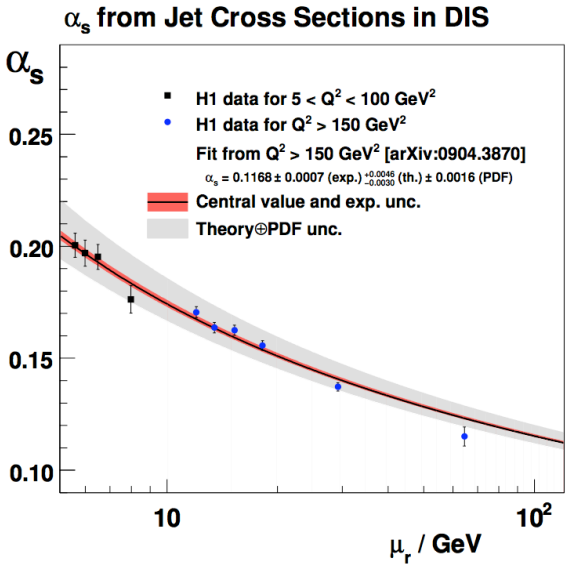

Wissenschaft

Die Hieroglyphen von morgen

The Physics Motivation for Data Preservation

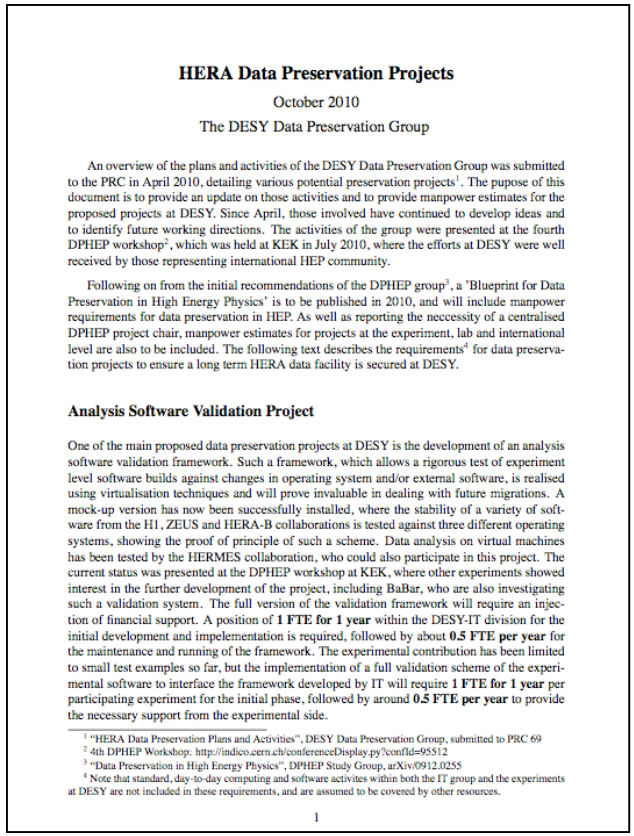
- Long term completion and extension of the existing physics program
  - Person power decreases rapidly towards end of experiment
- We may want to re-do previous measurements
  - Increased precision, reduced systematics
  - New and improved theoretical calculations / MC models
  - Newly developed analysis techniques
- We may want to perform new measurements
  - At energies and processes where no other data are available (or will become available in the future)
  - This is particularly relevant to HERA $e^{\pm}p$ data
- Investigate if new phenomena found today
  - Go back and check in the old data
  - *Recent interest in energy flow at low x - see A. Geiser @ TRENTO*





The DESY Data Preservation Group

- DESY presence within DPHEP effort strong from the start
  - HERA experiments and BaBar led the initial discussions
- Joint enterprise started soon after 1st DPHEP workshop between the relevant groups at DESY



HERA Data Preservation Projects

October 2010

The DESY Data Preservation Group

An overview of the plans and activities of the DESY Data Preservation Group was submitted to the PRC in April 2010, detailing various potential preservation projects[1]. The pupose of this document is to provide an update on those activities and to provide manpower estimates for the proposed projects at DESY. Since April, those involved have continued to develop ideas and to identify future working directions. The activities of the group were presented at the fourth DPHEP workshop[2], which was held at KEK in July 2010, where the efforts at DESY were well received by those representing international HEP community.

Following on from the initial recommendations of the DPHEP group[3], a 'Blueprint for Data Preservation in High Energy Physics' is to be published in 2010, and will include manpower requirements for data preservation in HEP. As well as reporting the neccessity of a centralised DPHEP project chair, manpower estimates for projects at the experiment, lab and international level are also to be included. The following text describes the requirements[4] for data preservation projects to ensure a long term HERA data facility is secured at DESY.

Analysis Software Validation Project

One of the main proposed data preservation projects at DESY is the development of an analysis software validation framework. Such a framework, which allows a rigorous test of experiment level software builds against changes in operating system and/or external software, is realised using virtualisation techniques and will prove invaluable in dealing with future migrations. A mock-up version has now been successfully installed, where the stability of a variety of software from the H1, ZEUS and HERA-B collaborations is tested against three different operating systems, showing the proof of principle of such a scheme. Data analysis on virtual machines has been tested by the HERMES collaboration, who could also participate in this project. The current status was presented at the DPHEP workshop at KEK, where other experiments showed interest in the further development of the project, including BaBar, who are also investigating such a validation system. The full version of the validation framework will require an injection of financial support. A position of **1 FTE for 1 year** within the DESY-IT division for the initial development and impelementation is required, followed by about **0.5 FTE per year** for the maintenance and running of the framework. The experimental contribution has been limited to small test examples so far, but the implementation of a full validation scheme of the experimental software to interface the framework developed by IT will require **1 FTE for 1 year** per participating experiment for the initial phase, followed by around **0.5 FTE per year** to provide the necessary support from the experimental side.

[1] "HERA Data Preservation Plans and Activities", DESY Data Preservation Group, submitted to PRC 69
[2] 4th DPHEP Workshop: http://indico.cern.ch/conferenceDisplay.py?confId=95512
[3] "Data Preservation in High Energy Physics", DPHEP Study Group, arXiv:0912.0255
[4] Note that standard, day-to-day computing and software activites within both the IT group and the experiments at DESY are not included in these requirements, and are assumed to be covered by other resources.

1

- Data preservation plans at DESY and status of individual contributors described in the document submitted to the PRC in April 2010
  - Contributions from H1, ZEUS, DESY-IT, DESY-Library
- Update submitted to this PRC meeting, a more detailed proposal for data preservation at DESY, including manpower requirements
- Such projects crucial to ensure a long term HERA data facility at DESY

Data Preservation Models Identified by DPHEP

Preservation Model	Use case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data

Cost, complexity, benefits

- Only with the full flexibility does the full potential of the data remain
  - Level 4 type programme was required by JADE and ALEPH re-analyses
- H1 and HERMES aim for DPHEP level 4, ZEUS between levels 3 and 4
  - Different approaches, can benefit from each other's experiences
  - A level 2 scheme for outreach using HERA data, collaborating via DPHEP, could also be pursued

ZEUS Future Analysis Model

- Maintain the ability of simulation of new MC after the end of the current analysis model and the current MC mass production system
- Standalone MC simulation package
  - Full chain from existing MC generators or generic MC interface (like LHA standard format) to simulation to Common Ntuple production
  - All dependencies included: calibration, condition, alignment, geometry, executables, steering card; unnecessary dependences removed
  - The core of the package is based on the current GRID production scheme
- Standalone MC simulation software package
  - Adapted makefiles scheme – one button recompilation
  - Serving as a source of new versions of executables for MC package
- Both packages can run either on real or virtual machines (tested on VirtualBox)
  - MC mass production on real machines (given necessary resources, GRID or local farm, available)
  - Once compilation breaks: MC package frozen: Virtual images with the last working OS
- Validation done with simple tool to compare different MC simulation releases
  - Basic set of histograms for comparison
  - Need to develop validation tool for MC and software packages
- Still a prototype system: for developers, not regular ZEUS members

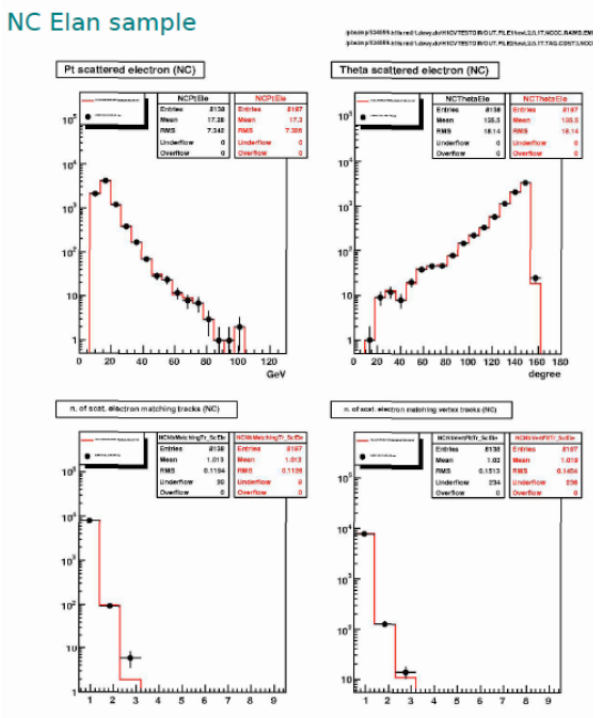
HERMES Future Analysis Model

- MC production requires a fully functional data production chain
  - Aiming to maintain the ability of new MC mass productions
  - Full software analysis chain (mDST based)
  - Complete data reproduction not foreseen, possibility not excluded, software-wise
- No external software dependencies
  - Database server part of local software (independent on network type or availability)
  - Local version of CERNLIB used
  - ROOT needed ONLY for certain analysis frameworks (not critical)
- Transition to SLD5 smooth on real and virtual machines
  - Static binaries compiled on SLD3 run flawlessly on SLD5
  - Recompilation on SLD5 requires gcc3/f77 (available on future OS, SLD6...?)
- Tests running on real SL5 batch nodes and a VirtualBox image
- Validation procedure in development phase

H1 Future Analysis Model

- H1 plans a *rolling model of preservation*, with a production timescale of say 3 months interval
  - Regular recompilation of analysis level software
  - Full data production of μODS/HAT (analysis level) data and MC
- Define a strategy for a rolling preservation model
  - Always use newest versions or freeze external software?
  - Aim to at least incorporate ROOT updates
  - Full level 4 version: Adopt changes in OS, include Fortran
  - Continue using the database / or maybe have a snapshot
- We will need good validation tools
  - Such a scheme already exists to validate the files content of the analysis level software between different releases
  - Expanding this validation to include full analysis selections, as well as the Fortran (simulation and reconstruction) code
- Aim to validate the **whole analysis chain** from RAW data to the publication plots

Some numbers from current productions:
** Read and copy 13.5 Tb of HERA II DST format data to Grid*
** 900 Grid jobs each running on average 20 hours*
** Produce 1.3 Tb of HERA II mODS/HAT format data*
** Ideally: 1 day to produce data, 1 day to download from the Grid*



Software Migrations and Dependencies

- All HERA experiments are currently moving to Scientific Linux (DESY) 5
- Start with the latest OS, rather than preserving something already outdated
  - Non-default (but better) compiler gcc-4.4 needed by H1 and ZEUS
  - Problems changing from g77 to gFortran, most software now compiled; HERMES still use gcc-3/g77
  - ZEUS problem with ADAMO tools for updating calibrations, constants (ok when final versions..)
- Identify external dependencies: also many common points between experiments
  - CERNLIB, ROOT, ORACLE, FastJet, Neurobayes, …
  - H1 problem with GKS (event display graphics), which no longer works in SLD5, no source available
- What about a unified validation suite, which can compare different DSTs, software releases and even analyses running under different operating systems, different external software versions, different ..?
  - *Could be used by all HERA experiments, ..and experiment X !*

Wait a moment, won't *Virtualisation* solve everything?

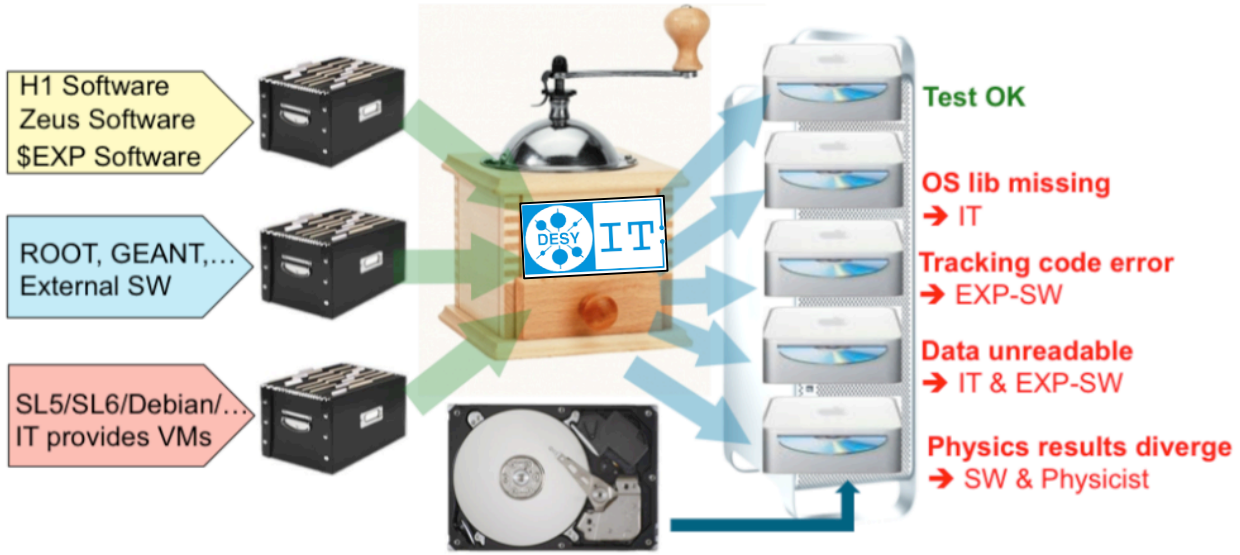
My first and very naïve ansatz

- OK, why don't we just put everything an a virtual machine?
  - Data archival is done elsewhere, just need "to plug that into the VM"
  - Your VM contains everything you need to develop and run code and analysis
- The problem would then be reduced to maintain virtual images, and maintain their ability to run. In the Cloud era, seems like a trivial task
- Problems: Everything in IT is a moving target:
  - Will your network always be the same?
  - Will your access protocol always be the same?
  - Are you sure you do not need new software (e.g. MC generators) that require a new OS?
  - Are you sure your i386/SL4 VM will produce the same results when emulated on a quantum computer in NN years?
  - What about service you need, like CondDB,…
- Naïve virtualization will not work… but still, virtualization can help

Yves Kemp (DESY-IT)

Towards a Generic Solution

- Validation and standards using virtualisation project by DESY-IT



- Detect incoherence in absence of intensive human survey
- Useful collaboration for future OS transitions and preservation
- Person power needs are being evaluated: Test with 5% pilot project

Yves Kemp, Dima Ozerov (DESY-IT)

Workflow: One test in detail (5% mock-up)

Experiment XYZ:
`application.sh`
`test.sh`
`files.tgz`
Put into VM image

`application.sh`
- Can be precompiled executable
- Better: Compile source code

`test.sh`
- Do something with the binaries obtained in `application.sh`

Experiment XYZ:
Check extracted logs and analyse resulting files

Prepare test VM

Run VM and perform tasks within it

Extract results from VM and analyze them

IT:
Raw VM image
system_rpm.sh
Preparation for automatic execution

IT:
Run VM
Monitor execution

IT: extract
`system.log`
`app.log`
`test.log`
`output.root`
Put on web-server;
Analyse errors

Validation Suite Test-Run, July 2010

	SL4	SL5	Fedora 13	Type of Test
ROOT v5.26	- no F77 compiler gfortran found - libX11 MUST be installed	Estimated ROOTMARKS: 1534.29	Estimated ROOTMARKS: 1512.76	Compilation
H1 Data Analysis	Processed 47243 events with J/Psi candidates Histogram written to jpsi_mods.root	Processed 47243 events with J/Psi candidates Histogram written to jpsi_mods.root	Processed 47243 events with J/Psi candidates Histogram written to jpsi_mods.root	Run pre-compiled tgz using compat libs
ZEUS MC Production	> ls -lh ZEUSMC.HFSZ627.E8954.GRAPE.Z01 4.2 MByte	> ls -lh ZEUSMC.HFSZ627.E8954.GRAPE.Z01 4.2 MByte	> ls -lh ZEUSMC.HFSZ627.E8954.GRAPE.Z01 4.2 MByte	Run pre-compiled tgz using compat libs
HERA-B Software	Compilation OK DB connect fails	Compilation OK DB connect fails	Compilation failed – needs code change	Compilation

(HERMES to come)

- Proof of principle shown to work, such a system would provide a powerful tool
  - Requires development within DESY-IT and input preparation by each participating experiment

DESY-IT: 1 FTE for 1st year, then 0.5 / year
Per expt.: 1 FTE for 1st year, then 0.5 / year

HERA Data Formats for Preservation

> Final ZEUS data reprocessing to mDST completed in 2009

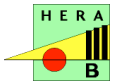
- Basic preserved data format: ROOT based “Common Ntuples” (CN) (two more iterations in 2011)
- Ultimately RAW, MDST data and MC removed from robots, keep only CN
- Reduces total amount to be preserved for ZEUS from the current 1 PB to ~ 100 TB

> Final H1 reprocessing of HERA II data 2009, equivalent HERA I repro ongoing

- Common analysis software H1OO started 2000, uses a ROOT based data format, used by all H1
- In addition, a monthly MC production of up to 1/4 billion events
- H1 to preserve RAW data, as well as at least one DST and analysis level versions
- Estimate total amount to be preserved for H1 to be ~ 200-500 TB

> Main format for HERMES analyses is the mDST

- New production planned before final freeze
- Last years of data taking with recoil detector, still need improved calibrations
- MC productions on Grid for ongoing analyses
- Total amount to preserve on tapes ~ 150-200 TB



> Preservation of HERA-B data under investigation within DESY-IT

- Total amount of data currently ~ 250 TB, will decrease once a preservation model is established

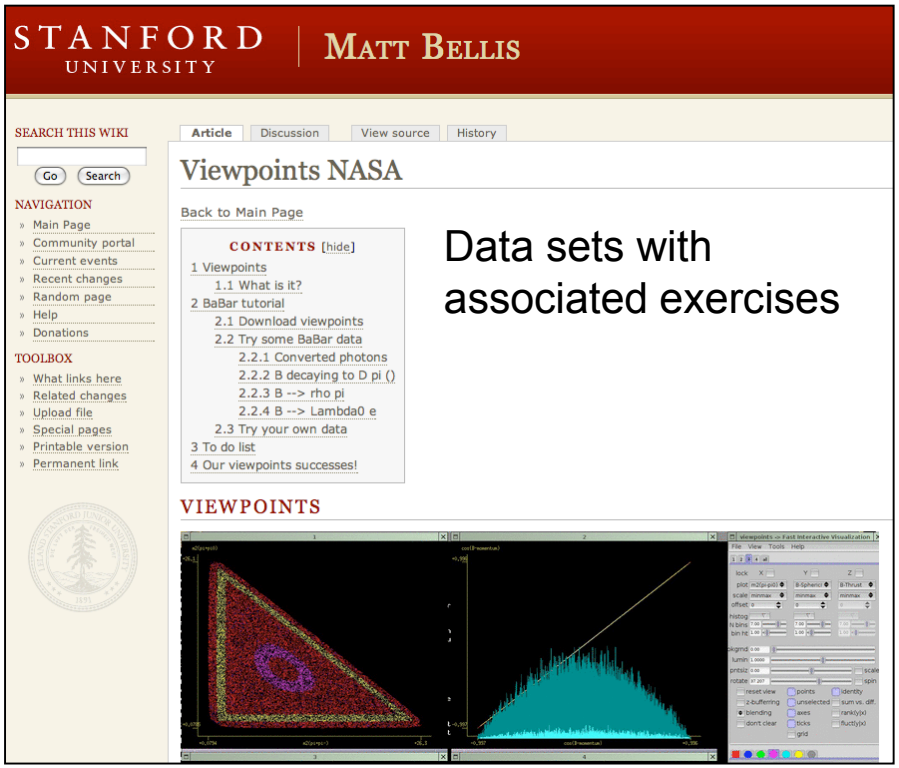
An Archival System for the HERA Data

- The scope of the validation framework does not foresee an examination of the condition of complete HERA data sets
- The present storage system at DESY-IT is not suitable for the archive storage of the HERA data
  - Discovery of problems with file integrity depends on user activity (only read files are checked..)
  - There is too much manual work involved (tape migrations, database consistency)
  - Weak connection between the user-end system and the storage back-end
- Proposal to develop a system needed for long term preservation at DESY
  - End-to-end data integrity checks, periodic inventory of whole archive with file fingerprint (checksum)
  - Verification that the file content is not damaged
  - File recovery in a reasonable amount of time (from second copy or via media recovery)
  - Possibility to store and retrieve large amounts of data
  - Ease the migration to newer storage technology

DESY-IT: 1 FTE for 1 year of development, future maintenance also necessary

Using HERA Data for Educational Purposes

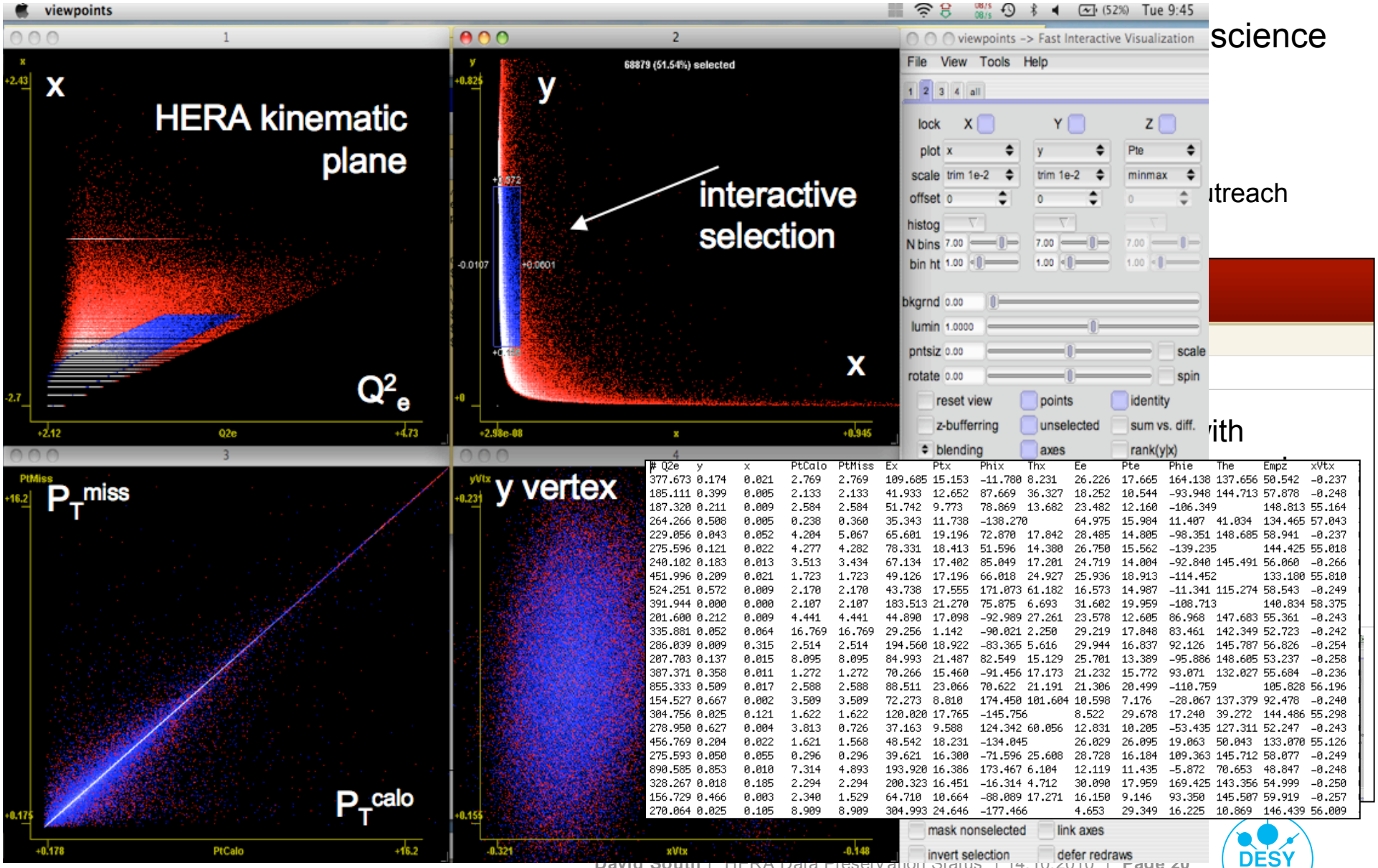
- Websites like www.teilchenwelt.de help further the public understanding of science
- Tutorials and exercises using real HERA data would be the next step
  - A HERA data outreach format is technically within reach and several ideas already exist
  - Such a scheme has started at BaBar, a wiki that could develop a true HEP data portal for outreach
  - Technical support for *media-Wiki* from DESY-IT



Data sets with associated exercises

http://www.stanford.edu/group/burchat/cgi-bin/bellis_mediawiki/index.php/Viewpoints_NASA

Using HERA Data for Educational Purposes



science

utreach

with

1 FTE for 1 year of development, between DESY-IT and all experiments

ZEUS and HERMES Documentation Efforts

- ZEUS non-digital documentation: notes, transparencies, technical drawings
  - Stored in documentation room in basement of building 1 – will be moved due to renovation
  - Consolidation, creation of electronic catalogue, handing over custody to DESY library is planned
  - Also digitalisation of old notes, theses is considered
- ZEUS digital documentation: mostly reside on the main ZEUS web server
  - Specific technical documentation (detectors, trigger) and electronic log book distributed over different machines
  - Consolidation of all relevant digital documentation on the main web server is planned
  - Migration of the main web server to newer hardware foreseen
  - Revision of personal web pages containing analysis details
- HERMES documentation wiki
  - Large effort to move all important documentation, technical notes to a wikipedia structure
  - Revisions performed by corresponding experts
  - Old web server now redundant, long preservation of wiki only (via INSPIRE?)
  - Electronic logbook running on separate box; image created, virtualisation possible future option

H1 Documentation Efforts

- Non-digital documentation initiative begun at DESY with dedicated manpower
  - Cataloguing, organisation and digitisation where appropriate of H1 papers, notes, drawings, talks..
  - Particularly timely due to building one renovation and relocation of documentation room
  - The DESY Library will eventually take over

- Digital documentation also investigated, but further resources are needed
  - Old online shift tools, detector files may be vulnerable, mostly not updated since July 2007
  - Electronic logbooks: H1, trigger, components, detailed run information
  - Migration of H1 web-server to DESY-IT virtual environment (migration to SLD5 completed)
  - Move the H1 documentation to external resources like INSPIRE
  - Streamline the content of the H1 web, rescue dead links, increase performance and attractiveness

H1 dig.docu.: 1 FTE for 1 year

Project between H1 and INSPIRE

- Start test project with INSPIRE to host H1 paper histories
  - INSPIRE beta launched: http://inspirebeta.net/
  - In discussions with Zaven Akopov, DESY-Library/INSPIRE, nice collaboration
  - We try extreme H1 example (Isolated leptons: 12 preliminaries!)

INSPIRE structure for publication history

- The preliminary reports will have each their own record, since they have information associated with them (varying figures, varying abstract, mostly varying presentation)
- Another record will be dedicated to the T0 stage (pre-T0, T0 and possibly addendum)
- Each drafting stage has it's own record (1st, 2nd, …) with corresponding figures and answers to draft
- Referee's report – a presentation of paper with summary of changes done to reflect the comments made by the collaboration
- Final version – the one that will be directly linked to the published paper (and is probably identical with it, unless there have been revisions submitted afterwards). In case of the mentioned revisions, they should be listed in this record as well.

  - Some other test ideas H1 notes, CB or other meetings, H1-wiki, …?

Example INSPIRE Record for an H1 Paper

Welcome to INSPIRE β. Please go to SPIRES if you are here by mistake.
Please send feedback on INSPIRE to feedback@inspire-hep.net

HEP :: HELP ..... SPIRES HEPNAMES :: INST :: CONF :: EXP :: JOBS

Home > Events with Isolated Leptons and Missing Transverse Momentum and Measurement of W Production at HERA

Information | References (52) | Citations (8) | H1 internal

Events with Isolated Leptons and Missing Transverse Momentum and Measurement of W Production at HERA.

H1 Collaboration (F.D. Aaron (Bucharest, IFIN-HH & Bucharest U.) *et al.*) *Show all 256 authors*.
2009

Eur.Phys.J. C64 (2009) 251-271
e-Print: **arXiv:0901.0488 [hep-ex]**

Abstract: Events with high energy isolated electrons, muons or tau leptons and missing transverse momentum are studied using the full e^\pm p data sample collected by the H1 experiment at HERA, corresponding to an integrated luminosity of 474 pb^{-1}. Within the Standard Model, events with isolated leptons and missing transverse momentum mainly originate from the production of single W bosons. The total single W boson production cross section is measured as 1.14 \pm 0.25 (stat.) \pm 0.14 (sys.) pb, in agreement with the Standard Model expectation. The data are also used to establish limits on the WW\gamma gauge couplings and for a measurement of the W boson polarisation.

Keyword(s): INSPIRE: W: production | transverse momentum: missing-energy | DESY HERA Stor | H1

Record created 2009-01-05, last modified 2010-04-11 Similar records

Abstract and Postscript and PDF from arXiv.org
Journal Server
Reaction Data (Durham)

Export
BibTeX, EndNote, LaTeX(US), LaTeX(EU), NLM, DC

Envisage an additional link for H1 members only

Example

Welcome to INSPIRE ?. Please go to SPIRES if you are here by mistake.
Please send feedback on INSPIRE to feedback@inspire-hep.net

HEP :: HELP ..::.. SPIRES HEPNAMES :: INST :: CONF :: EXP :: JOBS

Home > > Search Results

Events with Isolated Leptons and Missing Transverse Momentum and Measurement of W Production at HERA

PUBLICATION HISTORY

Preliminary Results

HEP-EPS 2007 conference paper. | July 2007
Prepared for Deep Inelastic Scattering 2007 | April 2007
Prepared for 42nd Rencontres de Moriond (Electroweak) | January 2007
Prepared for the 62nd DESY PRC | October 2006
ICHEP 2006 conference paper |July 2006
Prepared for the 60th DESY PRC | November 2005
HEP-EPS 2005 conference paper | July 2005
Lepton Photon 2005 conference paper | June 2005
Prepared for Deep Inelastic Scattering 2005 | April 2005
Prepared for the 58th DESY PRC |October 2004
Analysis of High Pt HERA II Data | ICHEP 2004 conference paper | August 2004
High Pt Analysis of the HERA II Data |Prepared for Deep Inelastic Scattering 2004 |April 2004

T0 talks

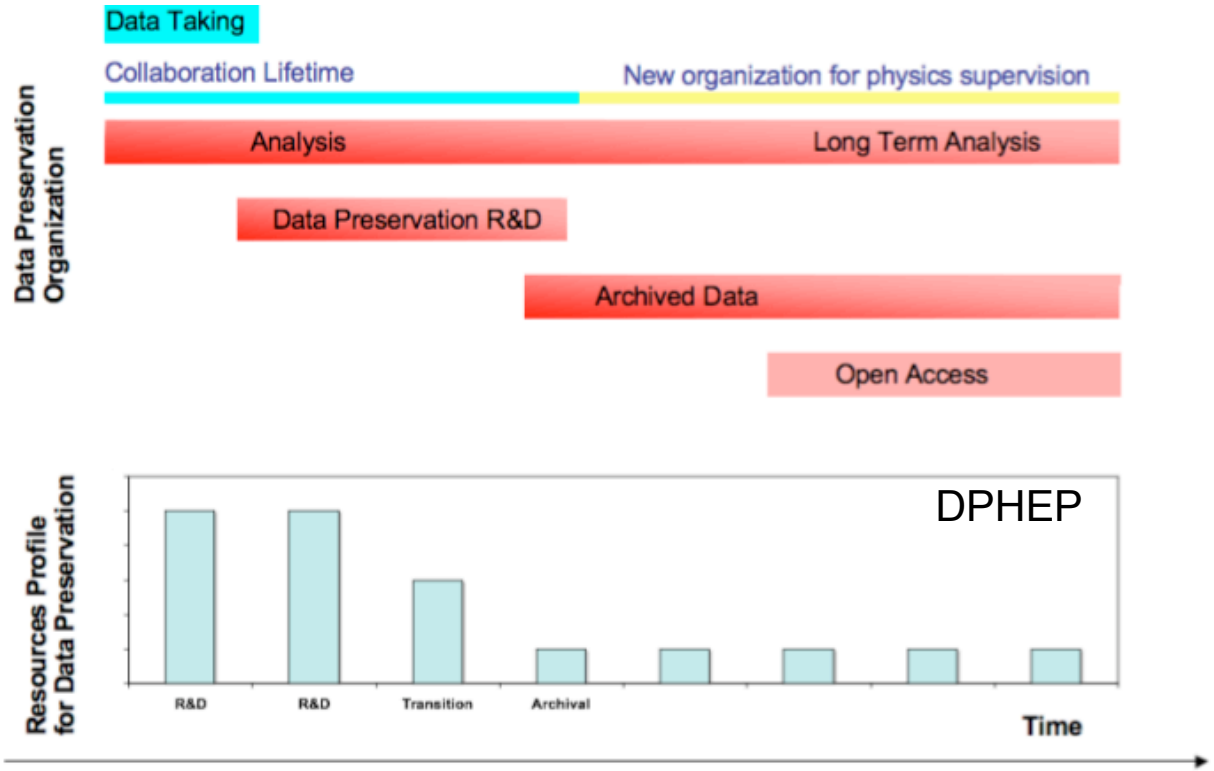
Pre-T0 Talk | 08.02.2008
T0 Talk | 24.07.2008
T0 Addendum | 14.08.2008

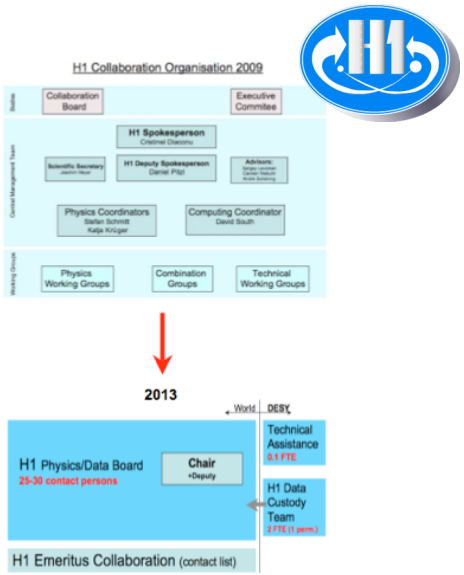
Paper Drafts

First Draft | Answers to Draft |15.08.2008
Second Draft | Answers to Draft | 19.11.2008
Referee Report | 20.11.2008
Final Version | 06.01.2009

Long Term Supervision of the HERA Data

- Requires good links between the host lab and the experiments
  - Realised using such projects as the joint validation and archival systems presented today
  - Sustained, reduced but non-zero manpower required after the initial development phase





- Future structure of the collaborations also under consideration by all experiments

Summary

- The $e^{\pm}p$ collisions collected at HERA are a unique data set!
  - Physics motivation has been detailed, full flexibility in the preservation model is desirable
- There is a strong participation of the DESY groups in the DPHEP initiative, guiding the future direction of data preservation in HEP
  - Data preservation effort at DESY unified between the different contributors
- Data preservation projects now identified, including:
  - In collaboration with DESY-IT: Validation and Data Archival, safeguarding the future of the HERA data
  - Together with the DESY Library: INSPIRE and future electronic documentation
  - Global initiatives via DPHEP: Education and Outreach
- FTE estimates provided for the realisation of such projects
  - More details in the document submitted to the PRC
  - Projects and FTE requirements also to be contained in the next DPHEP publication
- Good DPHEP visibility at CHEP conference in Taiwan next week
  - Plenary DPHEP talk + parallel talks on activity at DESY, BaBar, as well as HERA-B

EXTRA SLIDES

DPHEP Recommendations from the First Publication

- *An urgent and vigorous action is needed to ensure data preservation in HEP*
- *The preservation of the full analysis capability of experiments is recommended, including the preservation of reconstruction and simulation software*
- *An interface to the experiment know-how should be introduced: data archivist position in the computing centres*
- *The preservation of HEP data requires a synergic action of all stakeholders: experimental collaborations, laboratories and funding agencies*
- *An International Data Preservation Forum is proposed as a reference organisation. The Forum should represent experimental collaborations, laboratories and computing centres*

The Long Tail of LEP

	All	ALEPH	DELPHI	L3	Opal
All physics	345	65	114	85	81
Electroweak	89	17	26	22	24
QCD	85	19	25	19	22
Higgs searches	37	6	14	8	9
SUSY searches	25	4	7	5	9
Exotica search	34	5	12	10	7
Flavor physics	30	6	15	4	5
Exclusive channels	21	3	8	8	2
Cosmo-LEP	12	3	3	6	-
Other	13	2	4	3	3

LEP Publications after 2004

- Large number of publications well after data taking stopped
- Large variety of topics
- Legacy publications (full data, combined results) came later

ZEUS Data Format: Common Ntuple Project

- The archival system deployed by ZEUS for long term preservation is based on Common Ntuple project
  - Wide content to allow full physics analysis
  - Only community tools are needed for analysis: ROOT, KTjets etc.
- Preparatory phase is based on iterative production incorporating new content and many improvements/additions in physics analysis tools
  - Constant transition from MDST level analysis to CN level analysis
  - Second analysis check on CN level
- This preparatory phase is the most important stage to make ZEUS data in CN format good enough to be preserved
  - We are about halfway in this stage
  - Not all physics analyses are fully checked to be able to be done on CN level
  - Evaluation of possible additional manpower requirements within ZEUS or DP project ongoing
- Currently available iterations:
  - 4^{th} (full) iteration of HERA II data and corresponding MC samples already available last year and used for several preliminary results presented at DIS 2010
  - 5^{th} (partial) iteration for content revision – end of 2009
  - 6^{th} (full data, partial MC) iteration – used already by several analysis for ICHEP10
- 7^{th} full data and MC iteration planed for November
  - Two more iterations (partial and full) planned for the next year

ZEUS Future Analysis Model: Virtualisation

- Prototypes tested on real workgroup server and virtual machine using VirtualBox
- MC package
  - Selected executable versions, predefined input files, local output storage
  - Input and output schemes still to be developed: book-keeping for mass production
- MC Software package
  - Automated recompilation of libraries and executables
  - Needed addition of external libraries (CERNLIB, ROOT, CLHEP)
  - Automated tools for a new OS/architecture adjustment of makefiles

DESY

ZEUS Data Formats for Preservation

- The Grand Reprocessing of ZEUS data finalised in 2009 – no more reprocessing foreseen in the current plan
- The basic preserved data format: Common Ntuples (CN)
  - Full data samples for HERA I and HERA II and a wide spectrum of Monte Carlo samples
  - Storage currently on dCache/tapes, in future depending on development in IT
  - ROOT based analysis, sub-sample storage on local workgroup servers
  - Two more iterations foreseen in 2011
- Current ZEUS data sizes:
  - RAW 68 TB, mDST 41TB
  - MC: all 730TB, last versions only 250 TB
  - CN all 3 iterations ~50TB; estimated size for 2 last iterations ~ 100-120 TB
- mDSTs for data and MC will be kept while the production of CN is still active
- Ultimately RAW, MDST data and MC removed from robots, and only the Common Ntuples kept on tape
- Total amount to be preserved for ZEUS reduces from current ~1 PB to 100 TB

H1 Data Formats for Preservation

- Final reprocessing of HERA II data done 2009, equivalent HERA I repro ongoing
- Common analysis software H1OO started 2000, uses a ROOT based data format
  - Used by most H1 analyses
  - Target calibrations reached in latest version using final reprocessing (1.0% HFS, 0.5% EM)
- In addition, a monthly MC production of up to 1/4 billion events
- Current H1 data sizes:
  - RAW data of good and medium runs: 75 TB: **the basic format for H1**
  - One full set of DSTs, total for HERA I+II: 18 TB
  - One version of common analysis level format, μODS and HAT (< 3 TB)
  - In addition to calibration and cosmic runs, total data about 100 TB
  - MC at least of the same order, MC sets for long term now being defined
- Estimate total amount to be preserved for H1 to be 200-500 TB

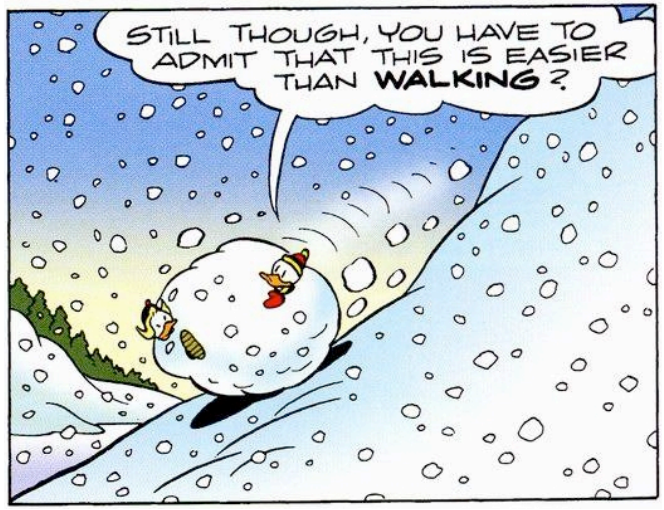
HERMES Data Formats for Preservation

- Main format for data analyses: micro DST (mDST)
- Current HERMES real data sizes
  - RAW data 110 TB
  - Main (track) 45 TB
  - mDST 2 TB
  - Latest and best production in total 4 TB
- Monte Carlo
  - mDST 25 TB on tape
- New production planned before final freeze
  - Last years of data taking with recoil detector, still need improved calibrations
  - MC productions on Grid for ongoing analyses
- Total amount to preserve on tapes: 150-200 TB

HERA-B Data Preservation

- HERA-B was an active collaboration until 2007 (last PRC talk 2006)
  - 33 institutes, 250 collaborators
  - Data taking in 2000 and 11.2002 - 02.2003
  - More than 20 physics publications, 4 most recent in 2009!
- Status of software
  - Last official release in 2003, runs on SuSe-7.2 (Linux release from 2001) (DL4)
  - Until 2010: a few old computers still within DESY-IT
  - From 2010: migration of data reading part of the software to newer operating system
- Next steps towards preservation
  - Freeze on a virtual machine the last official collaboration software version as default
  - Production model: migrate rest of software, write verification tests
  - Inspect data, total around 250 TB, identify and save only the essential part for preservation
- Manpower mostly from DESY-IT/Physics Computing Group, with connection to the former HERA-B software responsible

Freezing vs Rolling (or “Test-driven migration”)



- **Pro Freezing**
  - One-time effort, very small maintenance outside of analysis phase
  - Also allows software w/o code (but might fail with DRM / licensing issues)
- **Pro Test-driven migration**
  - Usability and correctness of code is guaranteed at every moment
  - Data accessibility and integrity can be checked as well
  - Fast reaction to standard/protocol changes
  - General code quality can improve, as designed for portability and migration
- **Cons Freezing**
  - Rely on certain standards and protocols that may evolve
  - Potential performance problems
- **Cons Test-driven migration**
  - Needs long-time intervention, more man-power and resources needed
  - Some knowledge of the frameworks must be passed to maintainers

H1 Virtualisation on the CERN VM

H1OO Virtualization Using CernVM Software Appliance.

How to get H1 Collaboration OO analysis framework and the grid User Interface on your laptop.

Mihajlo Mudrinic (H1)

Step-by-step Instruction

- Install VMware Player or VirtualBox
- Download Latest CernVM
- Untar the file and open it with your VM Software.
- Play your CernVM image and wait until the end of boot process.
- Read out your IP Address.Fig1
- Open a web browser on your computer, and point to the IP address.Fig2
- Type user: admin password: password.
- Change the admin password.Fig3
- Setup an local user (Group Must be hone!!).Fig4
- Click on preference, "VO set to hone", open advance option and choose "enable grid user interface".Fig5
- Wait until CernVM rebootsFig6
- Login and type: source /opt/hone/etc/login.Fig7
- Good Luck!Fig8

Special Note for VirtualBox Software

Our suggestion is to use "Bridged Networking". The guest will obtain its IP address in the same way that the host does.

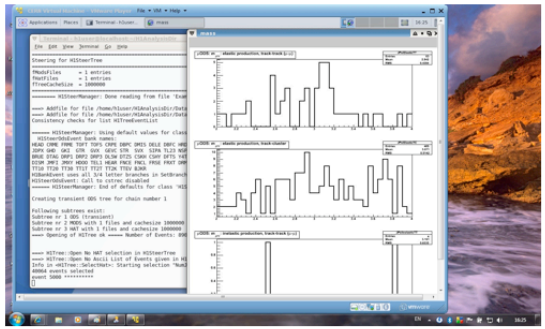
You can find on CERN wiki page special instructions for VirtualBox Users.

Introduction

Virtual machine software (VMware, VitrualBox, ...) is software allowing one to run two operating systems simultaniously on a single machine (you laptop). To promote the idea of using HEP data in scientific training, education and outreach we build "virtual" H1OO Linux image using CernVM Software Applience. CernVM is a CERN based R&D project which delivers a thin Virtual Software Appliance already used by LHC experiments (ATLAS,ALICE,CMS,LHCb). We would also like to thank the members of CernVM R&D project for dedicating one VM on the CERN domain to The H1 Collaboration on which we can build and publish "virtual" Linux images with the preinstal H1OO analysis framework (release 3.4.14).

System minimum requirements

- **Windows users** : desktop or laptop PC running Windows with the VMware Player/VirtualBox software installed (free software)
- **Macintosh users** : desktop or laptop Macintosh running Mac OS with the WMware Fusion/VirtualBox software installed (shareware/free software)
- at least a **1 GHz** processor
- at least **1 Gb** of RAM for the PC or Macintosh
- **2 Gb** of free disk space available



CernVM
SoFTware Appliance

The H1 Collaboration CernVM

Last updated 12.06.2010

- Studies of virtualisation ideas using H1OO within the CERN VM
  - Nice example of running an analysis using a virtual image of the H1 environment
  - New form of simple laptop installation of H1 software
  - Access to the [*large scale*] data remains an outstanding problem

Example: Enhanced Presentation of H1 Results

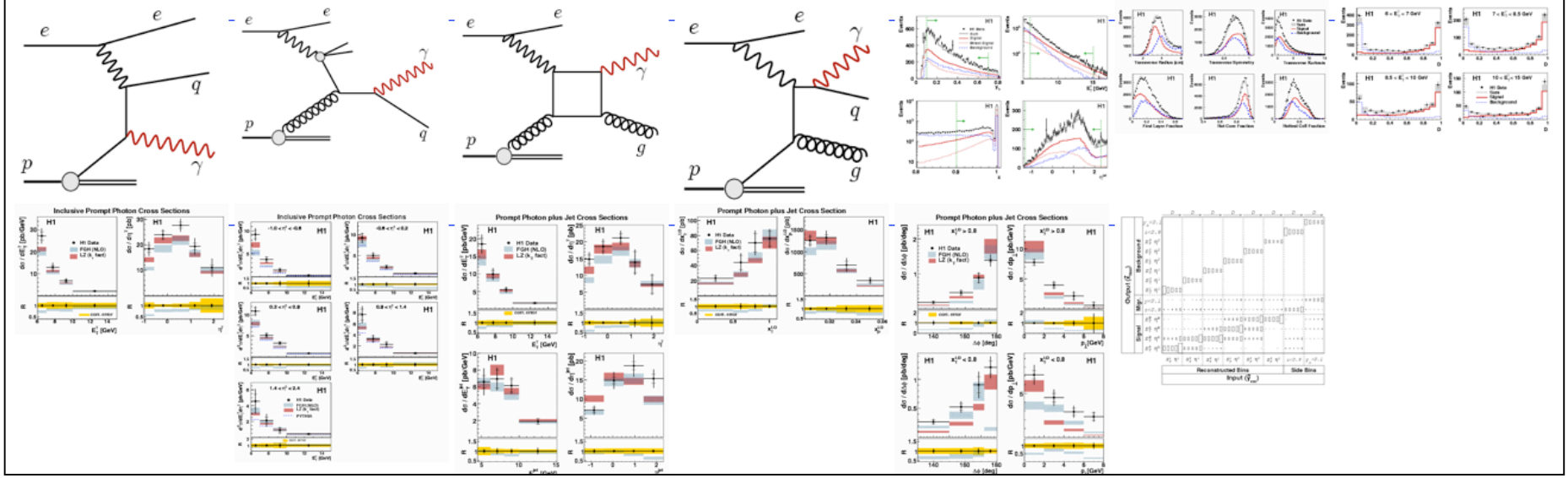
New Publication of the H1 Collaboration

DESY-09-135 **Prompt Photons in Photoproduction at HERA**

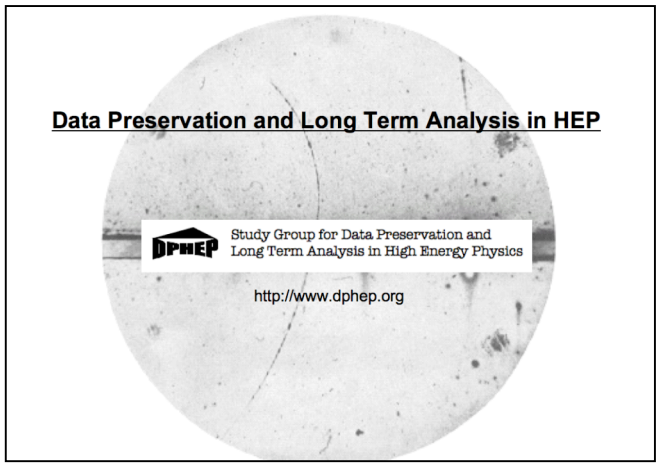
arxiv:0910.5631
H1-187

Reference	H1 Collab., F.D. Aaron et al., Submitted to EPJC (10/09) , 10/09
Figures	(1a) (1b) (1c) (1d) (2) (3) (4) (5) (6) (7) (8) (9) (10)
Links	back to overview Abstract from hep-ex Spires pdf version
Comments	

Gallery



DPHEP Seminars: Increasing the Awareness within HEP



DESY

technische universität dortmund

universitätbonn

IN2P3

INSTITUT NATIONAL DE PHYSIQUE NUCLÉAIRE ET DE PHYSIQUE DES PARTICULES

BROOKHAVEN NATIONAL LABORATORY

"Very interesting, good to know someone is thinking about this"

"I'm not sure you will be able to do level 4, it seems like quite a task" [people mostly won over a little with arguments about validation procedures]

"No hope already for ATLAS software to be aligned and unified as you suggest"

"You should rather turn the argument round and say *BECAUSE* it costs so little in FTE with respect to initial outlay, that it would be wrong *NOT* to do it"

DESY