The ZEUS and H1 long term data preservation projects in Max-Planck Institute für Physik

Andrii Verbytskyi

2nd DPHEP Collaboration Meeting
Geneve,
March 14, 2017

HERA data preservation motivation

- Future data (re-)analysis with new models and new approaches.
- Modelling for the future experiments.

HERA reminder:

- The only $e^{\pm}p$ collider, 1991-2007;
- $27.5\,GeV\ \ e^{\pm}$; $460, 575, 820, 920\,GeV\ \ p$;
- (Un)polarized $e^{\pm}$ collide with p;
- Polarised $e^{\pm}$ collide with $H/D/.../Xe$ targets (HERMES);
- p collide with nuclear targets (HERA-B).

Two complementary active sides for the HERA data preservation: DESY (covered by Achim) and MPP.

MPP model for DPHEP

Data preservation is about new and interested results with old data.
We describe ingredients and tools:

- Data bits
- Software
- Experiment documentation
- DP policies and documentation
- But in the end we are interested in **physics** .

Brief idea: enable physics and make it doable with modern methods in modern environments with minimal effort.

MPP model for bits preservation

MPP bits preservation is similar to approach from DESY.
The main differences comes from the ideas to

- Enable option for worldwide access via Grid.
- Study options to benefit from larger Data Preservation efforts.

MPP model for documentation preservation and policies

MPP data preservation relies on the documentation preserved by DESY/DESY library/InSpire.
The main idea is to provide the missing or DPHEP@MPP-specific documentation:

- Access to data in MPCDF (H1 and ZEUS).
- Monte Carlo generation procedures, event display usage, database of data samples, virtual machine usage (ZEUS).

The policies on the data/documentation access are same as in DESY: defined by collaboration spokespersons.

MPP model for software preservation

MPP software preservation is different from DESY.
Explicit effort put to make software it work in the next 10-15 years.

- Rely on industry, not HEP-only standards.
- Enable integration and compatibility with new physics software, e.g. data bases and Monte Carlo generators.

Main focus was on ZEUS software: Frozen environment that can be installed on virtual or real machine is provided. i.e. ISO image of full operating system relying on Intel $x86$ architecture.
Though not implemented, same approach is applicable to H1.

ZEUS

ZEUS analyses statistics

The results of recent yeas are produced in Data Preservation mode.

- 1X papers since 2014
- active analyses, more results on the way.

ZEUS data bits in MPP

- ZEUS data are stored in MPCDF on locally accessible tapes and in disk pool.
- Access via multiple protocols with grid tools worldwide to disk pools..
- Grid-enabled storage for new samples (Monte Carlo) and analysis is available.
- Straightforward procedure to update or add new samples.
- Stored part of private Ntuples/analyses directly in MPP.
- Future: keep only most important bits in disk pool and have a mechanism to get the other data if needed.

ZEUS data in MPP and DESY: Bits statistics

Data content is simple ROOT and PAW ntuples (see talk from Achim).
Logs and some inputs for Monte Carlo simulation are available.

	MPCDF	**DESY**
Files:	**1M+**	**1M+**
Volume:	$250\,T$	$250\,T$
Work area:	**yes**	**yes,limited**
Access:	**Worldwide**	**DESY**
Protocols:	**Multiple, see list**	**NFS mount**
Auth:	**Grid certificate**	**DESY account**

Available at:

- gsidcap://grid-srm.rzg.mpg.de:22128/pnfs/rzg.mpg.de/data/zeus
- grid-gftp2.rzg.mpg.de
- davs://grid-dav.rzg.mpg.de:2880//zeus
- ...

Documentation and documentation preservation in MPP

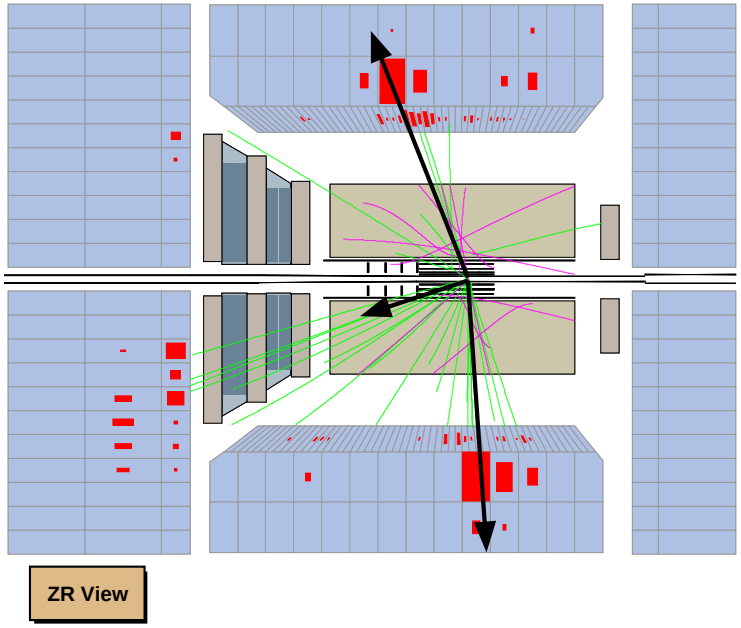
As most documentation is available in InSpire and DESY sites, MPP covers the missing parts:

- Instructions on data access
- Up-to dated database of available samples
- Instructions for Monte Carlo generation and reconstruction

Available at: https://wwwzeus.mpp.mpg.de/dphep.html

ZEUS software in MPP

- Main software for the analysis is vanilla ROOT.
- Additional software includes:
  - ZEVIS, the event display based on ROOT;
  - CNINFO, the event data base, based on ROOT and SQLite3;
  - ZMCSP Monte-Carlo standalone generation packages – see next slides.
- +any ROOT extension that will work for you...



ROOT
Data Analysis Framework

SQLite + your favourite software e.g. FastJet

ZEUS software environment/VM

A certain environment is needed for the analysis, i.e. DESY NAF or MPP/MPCDF machines are working well now. In parallel:

- Virtual machines(VM) are a very attractive **long-term** solution;
- The way other experiments (LEP/LHC) are going;
- The solution has very generic requirements, will survive for a long time.

ZEUS software environment/VM

Vitalisation for ZEUS has a two-fold purpose: it provides **benchmark** environment that suppose to work for a long time and it can be used, if desired, as a **super-portable production environment**.
It is based on DVD ISO image with CentOS7 and all software. It has options for:

- Automatic install on virtual or real hardware;
- Unlimited number of installations $\rightarrow$ potentially usable on clouds;
- Usage not restricted to any laboratory or virtualization software. Can run anywhere.
- **Tested by ≈ 5 users, found to be stable and easier to use than other options.**

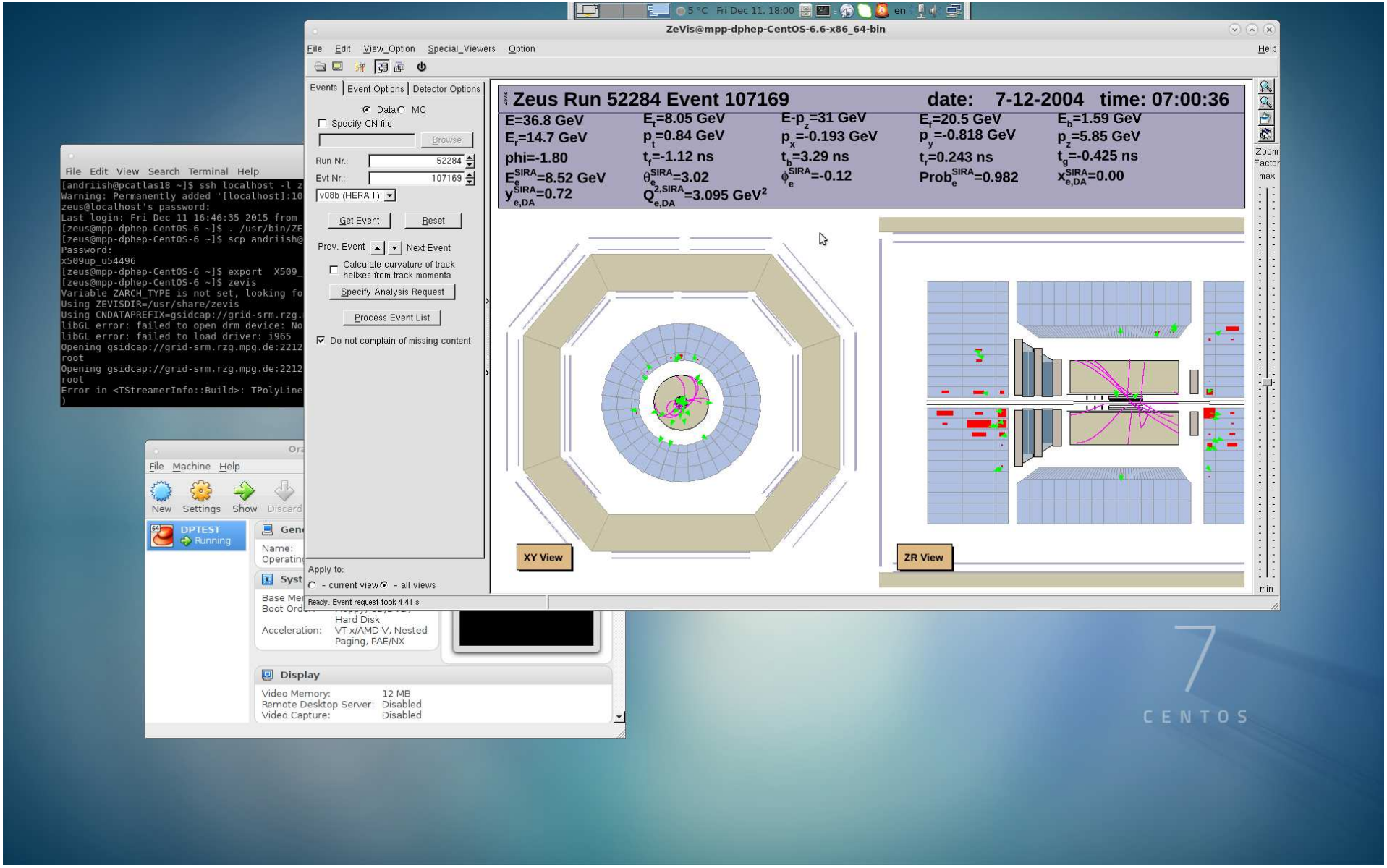
Note the system uses no specific "building" software. All software is in RPMs and installation disk is build with normal RedHat tools.

ZEUS software environment/VM

VM includes:

- ZEUS software: ROOT, MC simulation, event display, file catalogue, setup scripts etc.
- Modern MC generators, FastJet, cernlib, PAW, Rivet and other popular and "not really" packages.
- Anything you will want to install...
- Available at: https://wwwzeus.mpp.mpg.de/dphep.html together with documentation and video tutorials.
- Agree access and download it.

Data preservation for ZEUS: Software environment/VM



Monte Carlo production

Motivation:

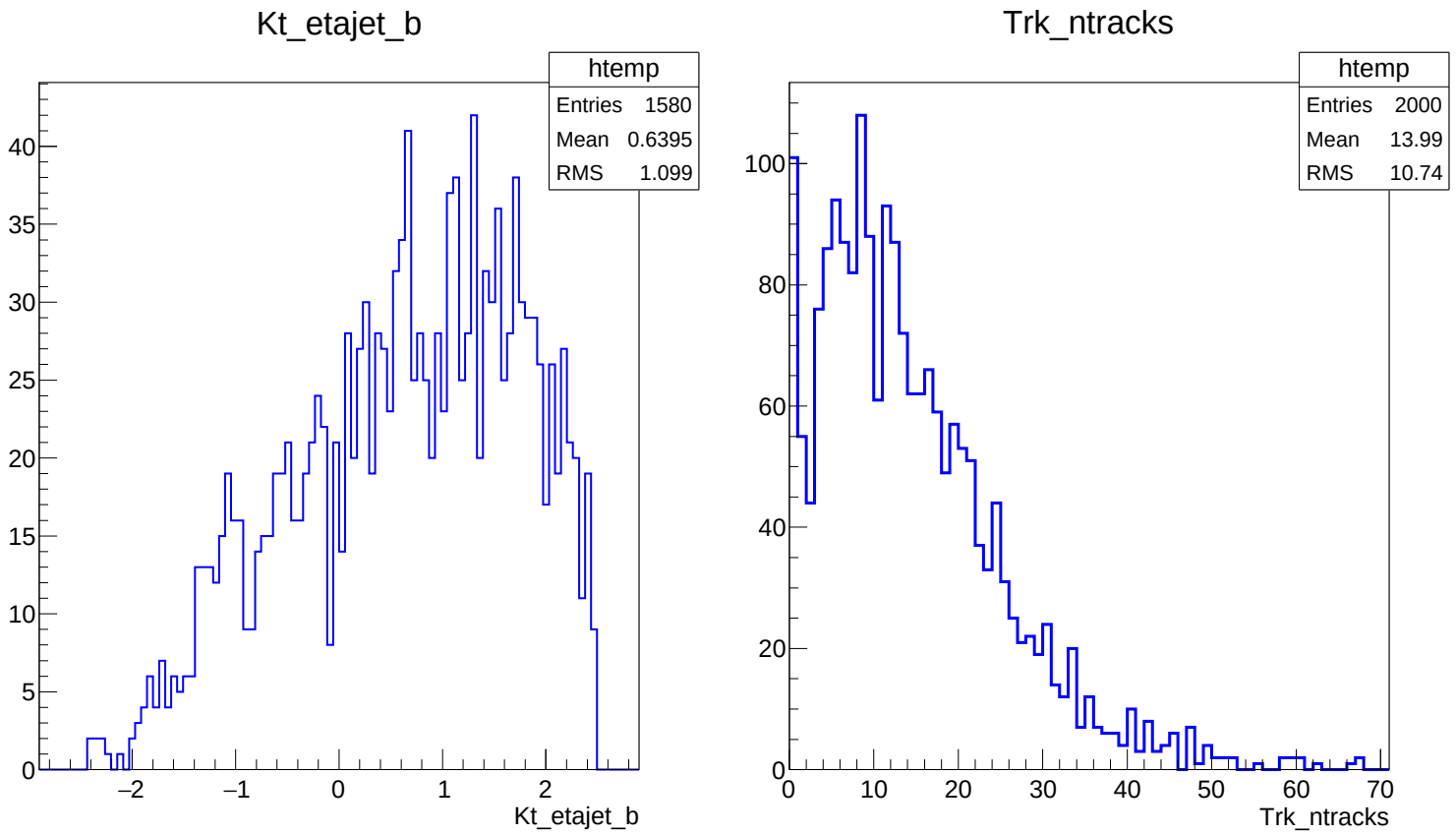
- Some analyses require more MC that is not available;
- Some analyses can be significantly improved with new MC;
- New MC generators/models can be tuned;
- New experiments can use it for the technical studies.

The MC recipe ingredients

- Instruction to generate events with old ZEUS generators are prepared.
- An interface to modern HEP event records based on HepMC3 library is developed. ZEUS MC can be produced from with modern MC generator, e.g. NLO capable SHERPA+BlackHat.
- ZMCSP (ZEUS Monte Carlo Standalone Package) is a tarball with all the software needed for the reconstruction of MC simulated events. It has no external dependencies, runs on modern Grid clusters, virtual machine, a laptop. On the Grid it can produce 50-100M events[1] per week. Supplemented with example of scripts and documentation.

[1] ZEUS has 360M of data events

The MC generation: SHERPA+BlackHat multijet setup



Pseudorapidity distribution of jets with $k_T(R = 1.0)$ algorithm applied in in lab. frame (left), track multiplicity (right) with SHERPA+BlackHat MC simulated sample.

The MC generation

- So far two analyses used the available option of MC generation at least to some extend.
- No modern generators at the moment, but the option is attractive, especially for precision measurements.

H1

H1 statistics

The results of recent yeas are produced in Data Preservation mode.

- 1X papers since 2014
- active analyses, more results on the way.

H1 data in MPP and DESY: Bits statistics

	MPCDF	DESY
Files:	**0.6M**	**0.6M**
Volume:	$280\,T$	$280\,T$
Work area:	**yes**	**yes,limited**
Access:	**Worldwide**	**DESY**
Protocols:	**Multiple, see list**	**NFS mount**
Auth:	**Grid certificate**	**DESY account**

Available at:

- gsidcap://grid-srm.rzg.mpg.de:22128/pnfs/rzg.mpg.de/data/hone
- grid-gftp2.rzg.mpg.de
- davs://grid-dav.rzg.mpg.de:2880//hone
- ...

H1 software in MPP

Don't have specific setup, just same as in DESY (see Achim's talk). Briefly:

- Main software for the analysis is ROOT with many H1 specific classes. Large part of analysis information is available with vanilla ROOT.
- Huge effort by H1 Collaboration to ensure software can be compiled relatively easy on new systems.

H1 documentation in MPP

As most documentation is available in InSpire and DESY sites, MPP covers the missing parts:

- Instructions on data access in MPP, which is the same as for ZEUS

Available at: https://wwwzeus.mpp.mpg.de/dphep.html

Option for H1 software environment/VM

The approach of virtualization can be extended to H1, as the basic software it the same.

- ROOT, compilers, libraries, etc.
- Modern MC generators, FastJet, cernlib, PAW, other HEP packages.

MPP Data Preservation summary

- Data is accessible in DESY and MPCDF for H1 and ZEUS.
- An option for MC production with new and old MC generators exists for ZEUS and tested. Potentialy useful for H1 as well.
- Virtualization option is implemented for ZEUS. Potentialy useful for H1 as well.

Use cases for HERA data

- Something that now we are not aware about.
- QCD:
  - Proton structure, e.g. F_2 and F_L, strangeness in the proton;
  - Diffraction, e.g. combination of measurements;
  - Jets and event shapes with NNLO;
  - Photon structure, instantons, pentaquarks, etc.
- EW physics:
  - Prompt photons;
  - Electroweak couplings.

See arXiv:1601.01499 and arXiv:1512.03624 for details.