ZEUS Analysis and Computing Model

Janusz Szuba

DESY

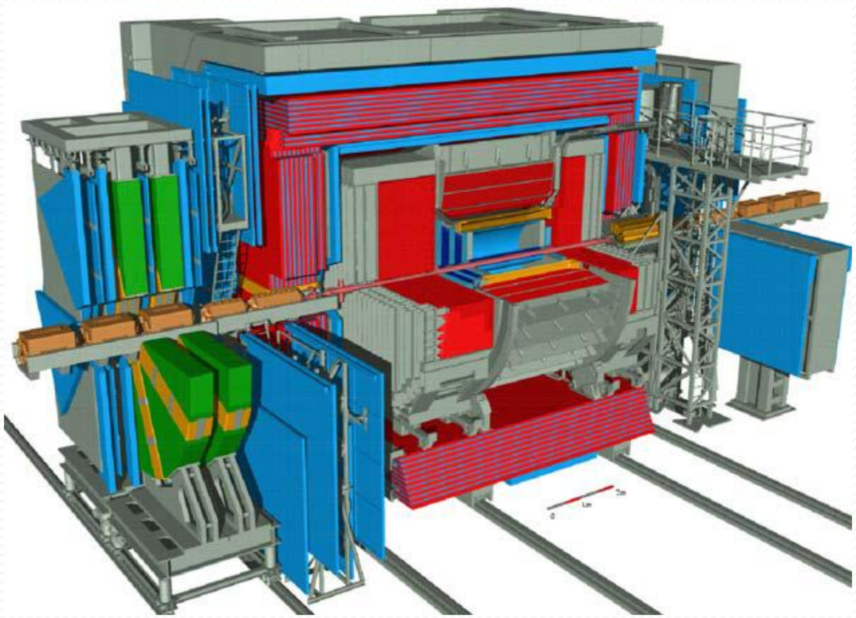
First Workshop on Data Preservation and Long Term Analysis in HEP

Outline

- Analysis within the ZEUS Collaboration
- Data reconstruction and analysis model
- Plans and prospects for short to long term data preservation

Overview of the ZEUS Experiment

- The ZEUS detector recorded ep collisions in two periods
  - HERA I (1992-2000) with collected 130 pb-1 (180 Mevents)
  - HERA II (2003-2007) with collected 380 pb-1 (410 Mevents)
- Upgrades in the second period
  - Luminosity upgrade and polarization of electron beam
  - Tracking upgrade in central and forward region
    - silicon microvertex detector MVD
    - forward straw tube tracker STT

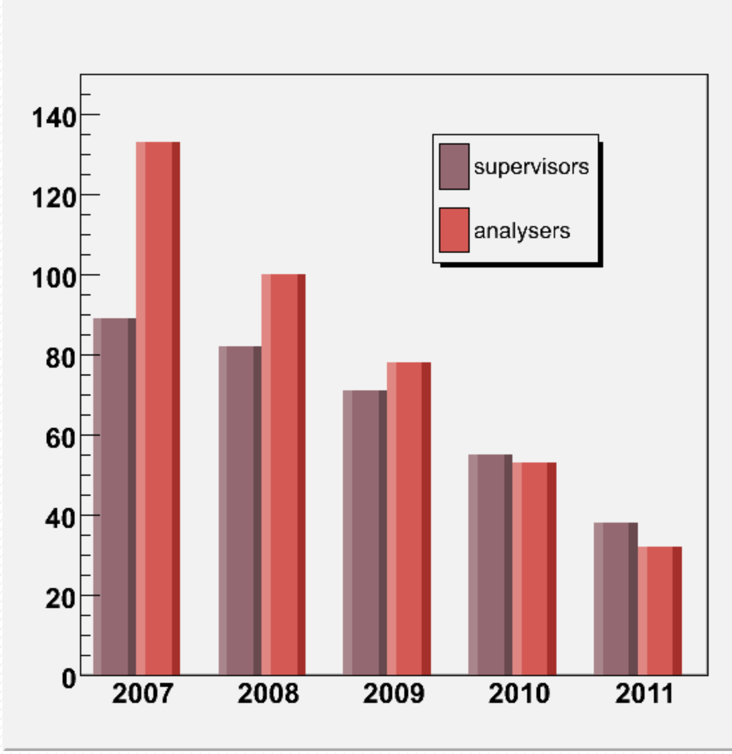


Physics Program

- The main physics goals in view of the HERA II upgrades :
  - Studies of EW physics with the polarized e+/e- beams (e.g. CC/NC cross sections)
  - Measurements of beauty and charm production rates with unprecedented precision thanks to upgraded tracking
  - Measurement of the proton longitudinal structure function F_L with help of the special two period at the end of the data taking with reduced proton beam energy
  - Precision tests of QCD with high statistics jet measurements

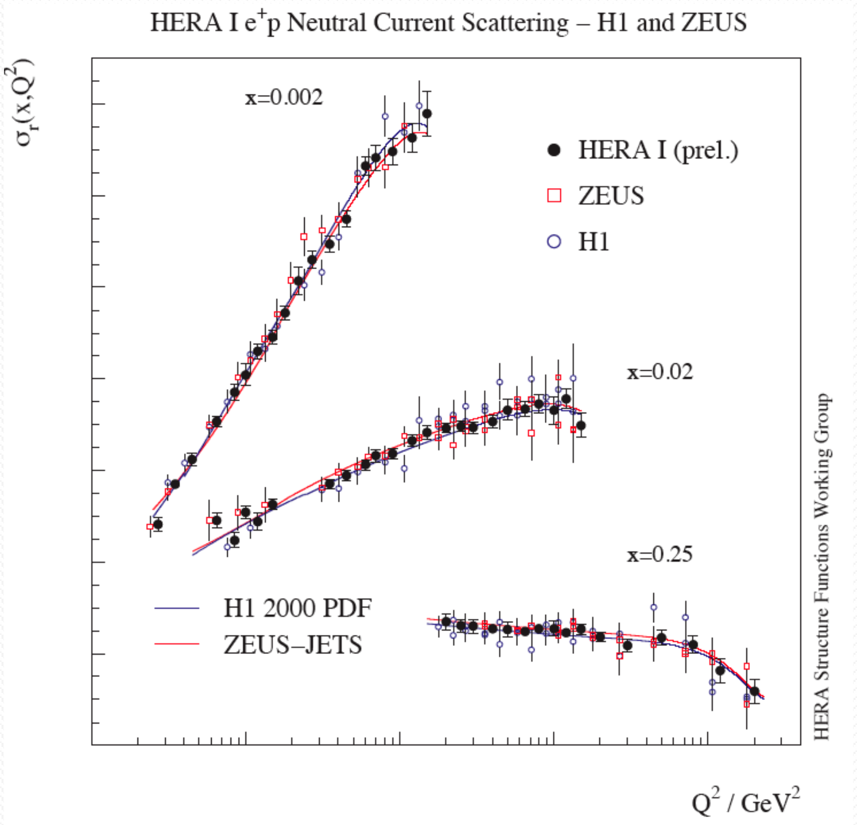
Analysis strategy in the ZEUS Collaboration

- There are currently about 70 analysis topics
- Analysis divided into several Physics Working Groups
  - High Q2 and Exotics
  - Longitudinal Structure Function
  - QCD and Hadronic Final States
  - Heavy Flavour Physics
  - Diffraction and Vector Mesons (now integrated into QCD and HFL)
- Require two independent analysis for a paper



Joint H1/ZEUS analyses

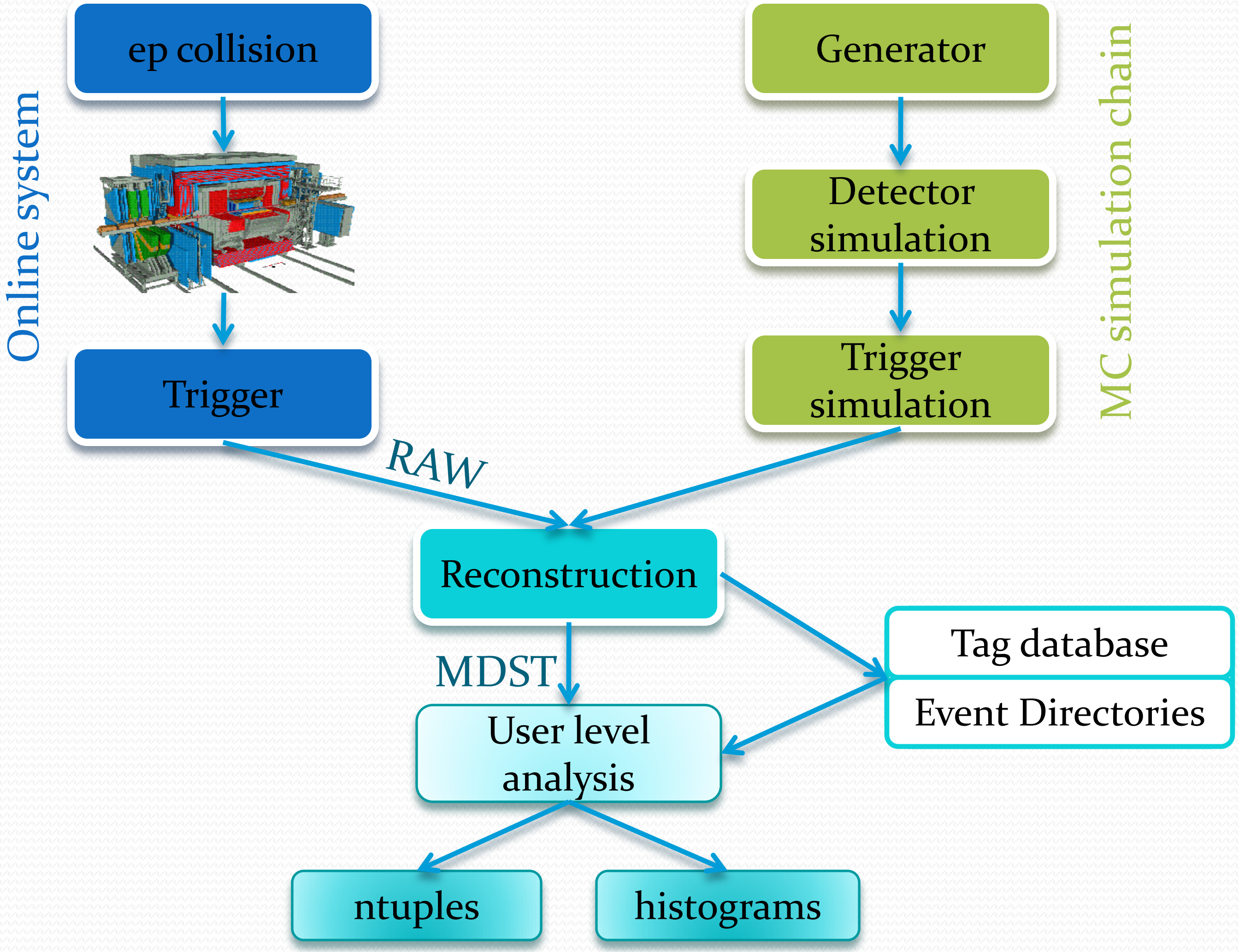
- Combine ZEUS and H1 statistics
- Cross calibrate the detectors
- Achieve better precision and sensitivity to rare processes
- Joint physics workgroups
  - Structure functions
  - Diffraction
  - Multileptons
  - Jets
  - …



Offline Group

- The Offline Group has the following tasks
  - Storing and processing of constantly growing data sample
  - Provide computing infrastructure and services for data reconstruction, simulation and analysis
  - Maintain efficient, transparent and scalable access methods to archived data
- Divided into subgroups
  - Reconstruction and reprocessing
  - Analysis and reconstruction computing farm
  - Monte Carlo production
  - Software maintenance
- Contains about 12 people
- Cooperates with physics groups, detector component experts and DESY IT division

Data Processing Model



Online system

ep collision

Trigger

RAW

MC simulation chain

Generator

Detector simulation

Trigger simulation

Reconstruction

MDST

Tag database

Event Directories

User level analysis

ntuples

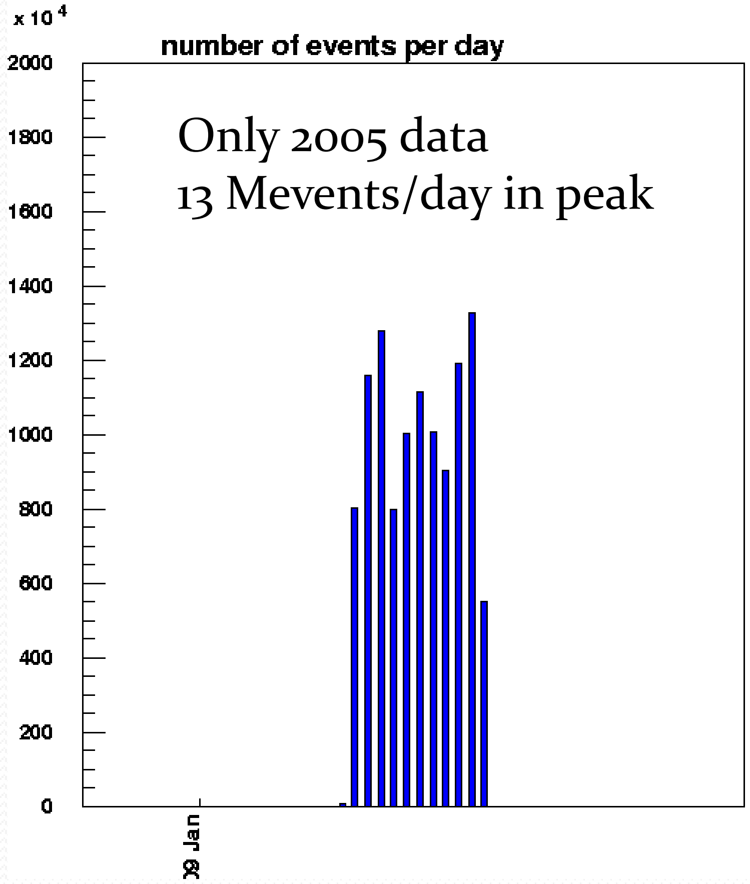
histograms

Reconstruction

- RAW data and reconstructed MDST (mini Data Summary Tapes) are kept in Entity-Relationship Model (ADAMO) structures based on ZEBRA file
  - Average sizes 125kB/event (RAW) and 75kB/event (MDST)
- Calibration, conditions, geometry and alignment are kept in database-like system called General ADAMO Files (GAFs)
- In the reconstruction process also produced are:
  - Events collections (Event Directories) allowing fast trigger selection
  - Event tag database (zesLite) based on ntuples with physics quantities for fast event selection
- Reconstruction and reprocessing system is centrally operated
  - Automated batch job submission, version control, web interface

Reprocessing and Grand Reprocessing

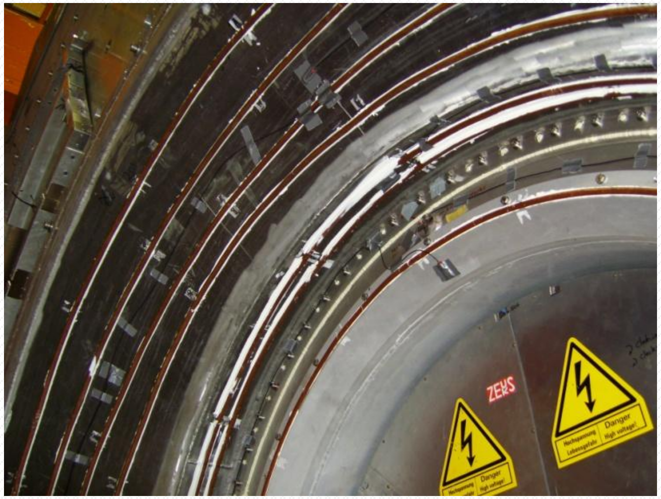
- Reasons for RAW data reprocessing
  - Tuning of calibrations and alignments
  - Reconstruction and analysis software development
- Each data taking period is reprocessed several times to account for the above improvements
- Grand Reprocessing after the end of data taking
  - Best knowledge and understanding of the detector incorporated into reconstruction software
  - Preceded with several test and validation procedures
  - All data periods are reprocessed to the same quality
- The final Grand Reprocessing of HERA II data is now almost finished
  - Processing time about 80 days
  - No plans to reprocess data again if validation is ok

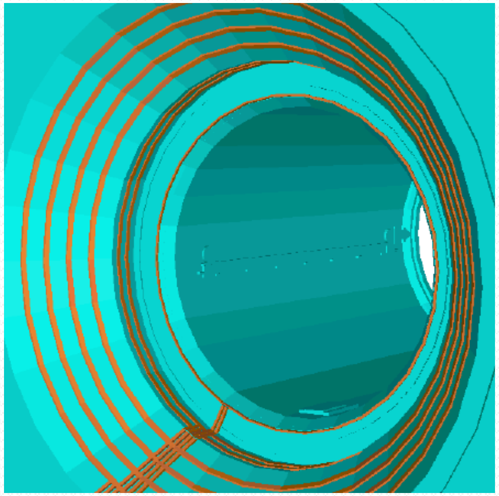


Only 2005 data
13 Mevents/day in peak

Monte Carlo Simulation

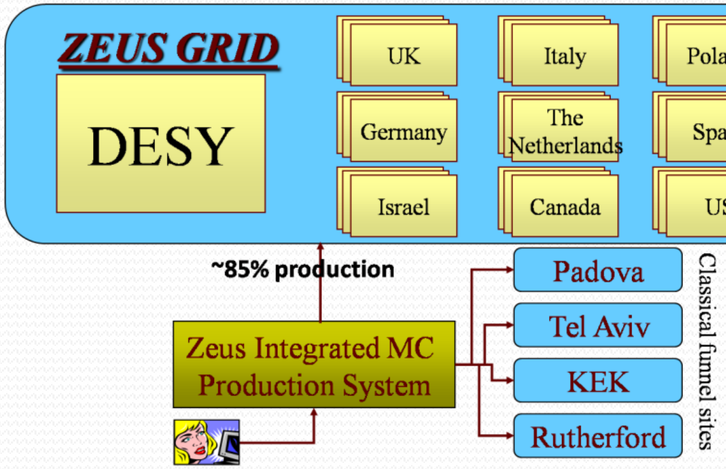
- All MC generators used in analysis are incorporated into the single common interface
- GEANT 3.21 for detector simulation
- Output format the same as MDST - ADAMO
  - Size per event 2-3 times larger than data due to simulation information
- Decommissioning of the detector gave some insights into more precise simulation
  - Alignment measurements
  - Dead material simulation
- The output from Grand Reprocessing will require further fine tuning of the simulation
  - Hit resolutions in tracking detectors
  - Calorimeter energy scale

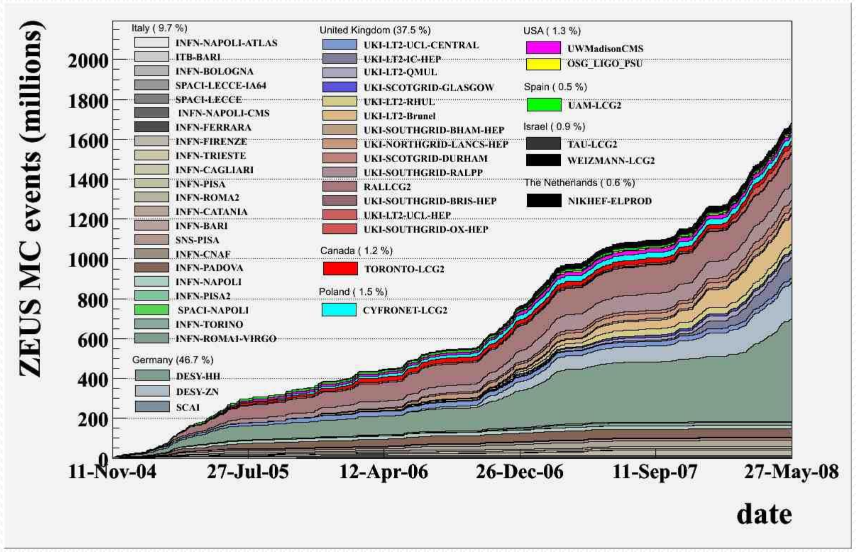




Monte Carlo Production

- ZEUS integrated Monte Carlo production system:
  - Grid based system
  - Classical distributed system called Funnel
- Grid covers more then 85% of the total production
- We are assuming preservation of Grid computing resources in the near future
  - Our use of GRID is parasitical comparing with future use of LHC experiments





Size of the data under analysis

- Size of the real reconstructed data and Monte Carlo simulation available for user analysis
  - the latest version of MDST and corresponding MC samples

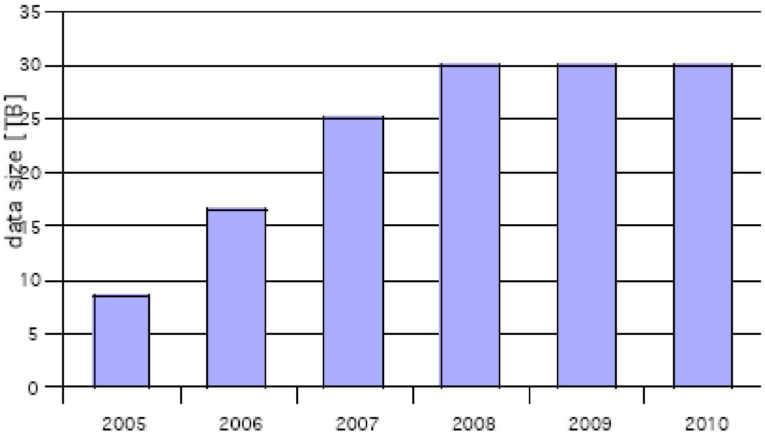


Fig.1: Cumulative amount of real-data under analysis in given year

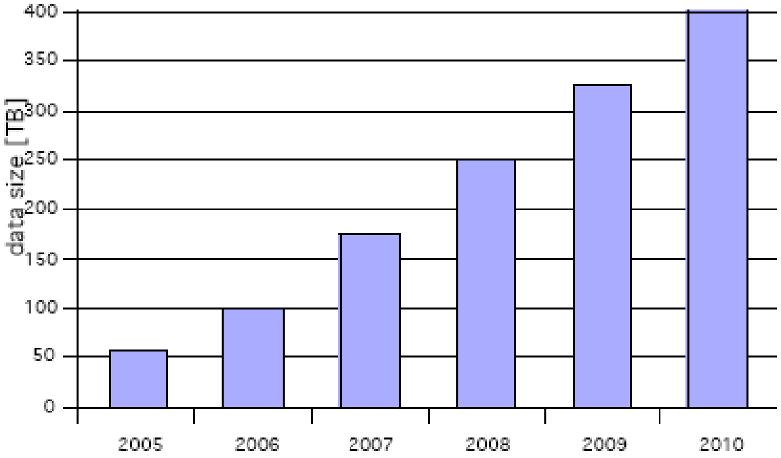
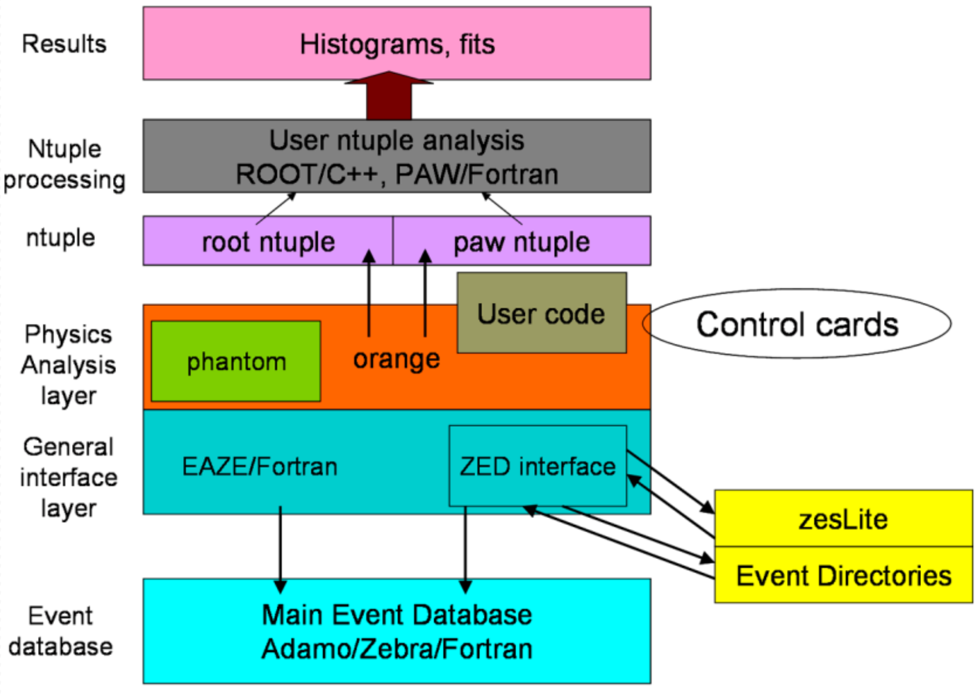


Fig.2: Cumulative amount of Monte Carlo data under analysis in given year

Current user level data analysis

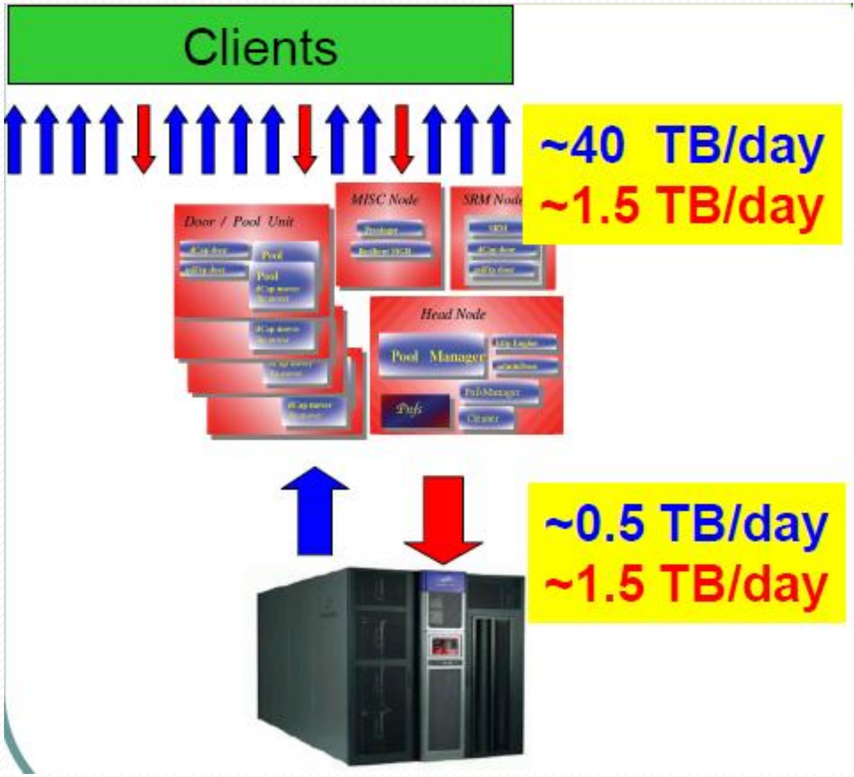


- All data are in one big MDST pool (no skims)
- Access to MDST files through general interface
- Events selection based on trigger or physics quantities in tag database
- Physics analysis layer includes all necessary reconstruction and analysis libraries and has hooks for a user code
- Steering provided by text control files
- The output – ROOT or PAW ntuples with all necessary physics quantities for particular analysis needs
- User ntuple analysis in ROOT or PAW provides final results

Analysis and Reconstruction Software

- The ZEUS repository contains more than 100 software packages
- The source code is maintained in CVS repository
- Every software package is developed according to predefined rules, using unified project structure and versioning scheme
  - Build environment based on a set of makefiles
  - Multi-platform support (suse8.2, sl3, sl4)
- Mixture of C, C++ and FORTRAN
- Legacy software (ADAMO, ZEBRA)
- Global software releases up to 2-3 times per year
  - Driven by reconstruction development and reprocessing cycles
  - After Grand Reprocessing mainly analysis libraries are developed

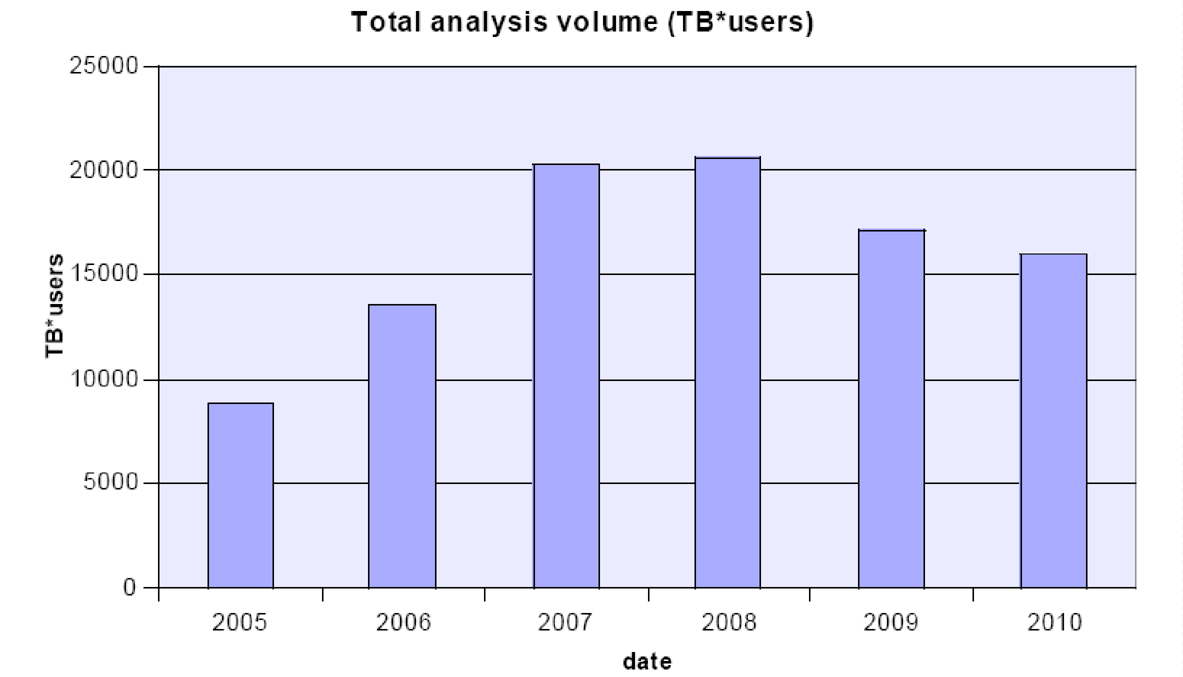
Storage and Data Access



- ZEUS data is archived on tapes
- RAW data is duplicated and stored in a separate robotic system
- The current data volume about 0.6 PB
  - Several generation of reprocessing and equivalent MC production
- Data access is optimized using dCache system
  - Significantly improves the overall transfer rate
  - Reduces the latency of accessing files
- Currently about 240 TB disc space

Analysis prospects

- The amount of resources needed for analyzing data depends on the total data size and the number of concurrent analyzers
- Analysis activities slowly decrease, with substantial residual in 2010
- The survey made in ZEUS collaboration and the consecutive addendum led to support for current analysis model up to 2013



Perspectives of the current analysis model

- There are several problems for the current analysis model to be preserved for a longer time
  - Legacy software (ADAMO, ZEBRA)
  - Code maintenance with OS upgrades
  - Expert knowledge of the detector and reconstruction algorithms – personal or documented
  - Funding for data storage (RAW, MDST, MC) and manpower to maintain the complex system
- Zeus started to define a way for medium term data preservation (up to 2013)

Common Ntuple Project

- The current analysis software is used to create common usage ntuples (real and MC data) with content wide enough to incorporate all possible physics analyses
- A simple ROOT ntuple format is used
- The resulting total ntuples size is expected to be between 10-20% of the size of data in MDST format
- The storage and access is unchanged with respect to the current model (tapes and dCache)
- The generation of new MC samples can only be done till the end of current analysis model support
- This strategy is seen as an intermediate step to define ultimate data format and content

Long term conservation of ZEUS data

- The HERA collider is a machine with unique physics capabilities, no comparable facility in the foreseeable future
- HERA data may provide answers which will possibly arise in the future experimental program at LHC or ILC
- The possibility to re-analyze HERA data over the time scale of about the next 10-20 years requires
  - Relatively simple data format
  - High abstraction level based on physics quantities rather than hardware/detector related
  - Encoding should ensure long term preservation

Open Access to ZEUS data

- In our view long term preservation is equivalent to offering the ZEUS data publicly available for any physicist or student
- The data could be used to educational, scientific or outreach purposes
- Checking new theoretical ideas require to maintain ability to simulate data at the appropriate abstraction level
  - Development of a new tool must be based on the present knowledge of the simulation of the detector
  - Possible parameterization must ensure adequate accuracy of the simulation
- MC simulation tool is seen as a 2-3 years project

Summary

- We believe that it will be difficult to preserve the current ZEUS analysis model beyond 2013
- Long term data preservation and open access to ZEUS data require higher level of abstraction
- Experience from common ntuples definition and joint H1/ZEUS analyses could help in defining final abstraction level
- New method of Monte Carlo simulation based on parameterized detector response is required