Status of data archiving at ZEUS

Janusz Szuba (DESY)

Third Workshop on Data Preservation and Long Term Analysis in HEP

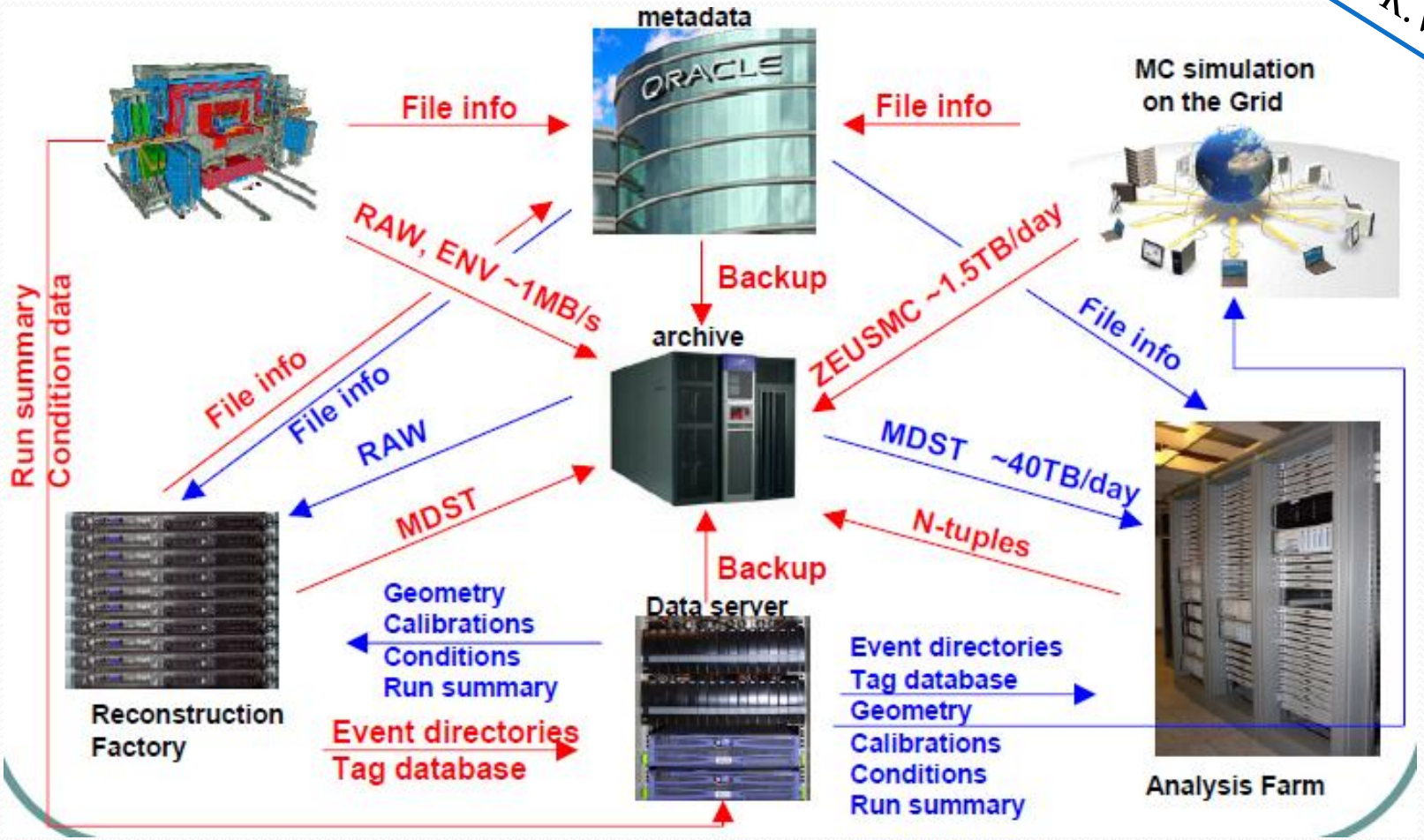
CERN, Mon 7th-Wed 9th December 2009

Outline

- ZEUS computing model and data format
- ZEUS computing strategy
- Common ntuples project
- MC simulation
- Summary

Reminder of the complexity of ZEUS computing model

Thanks to K. Wrona

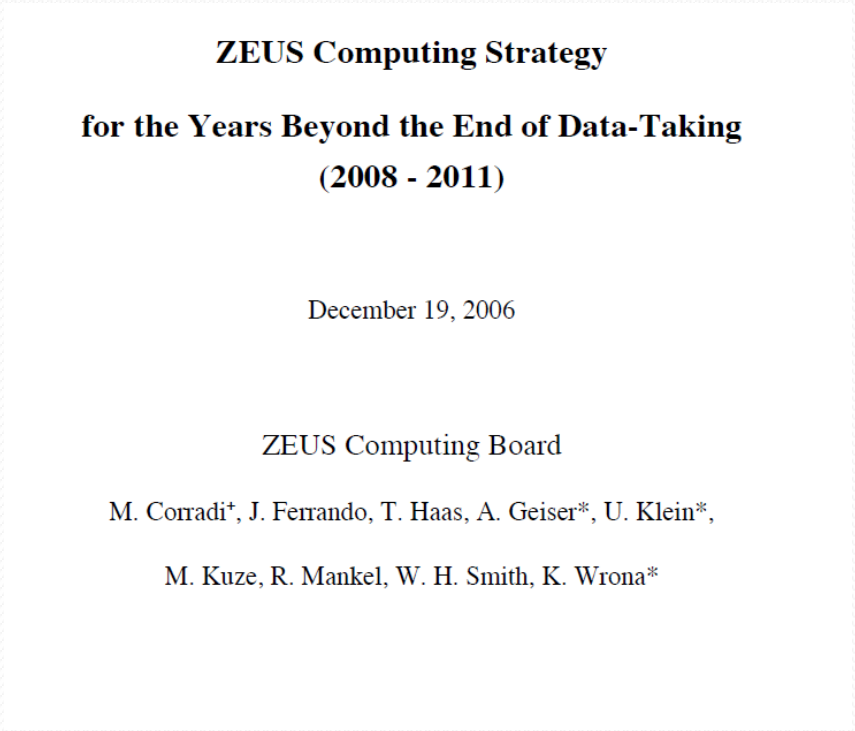


Data format and software

- RAW, MDST, MC
  - ADAMO (Entity-relationship model) structures written in ZEBRA files
  - Calibration, conditions, geometry and alignment are kept in database-like system called General ADAMO Files (GAFs)
  - Recontruction/simulation – Fortran, Geant 3, Cernlib and small fraction of C, C++
    - Tracking /vertexing in C++
- User analysis software
  - PAW and ROOT ntuples
  - FORTAN framework and physics analysis libraries (a bit of C)
- Event display – ROOT based

A bit of history - ZEUS Computing strategy document

- Defined the strategy of the ZEUS Collaboration for the following years in terms of:
  - Reconstruction activities
  - Analysis model
  - Monte Carlo production
  - Data preservation
  - Data storage and access
  - Costs estimation and funding
- Followed by Appendix which extends the funding till 2013



ZEUS Computing Strategy

for the Years Beyond the End of Data-Taking

(2008 - 2011)

December 19, 2006

ZEUS Computing Board

M. Corradi†, J. Ferrando, T. Haas, A. Geiser*, U. Klein*,

M. Kuze, R. Mankel, W. H. Smith, K. Wrona*

Extended ZEUS Computing Strategy for the Years 2012-2013 12 November, 2007

Appendix to the ZEUS Computing Strategy document for the years 2012-2013

ZEUS Computing strategy document – excerpts

- Analysis operation
  - Gradual move from mdst to common ntuples (2009-10)
  - Access to mdst and even raw data still foreseen
  - Central ZARAH operation till the end of 2011
- Long term conservation of ZEUS data
  - We are talking about ~>10 years
  - Not feasible with present analysis model (mdst)
  - Data formats and software can be unsupported in future
  - Simple data format based on physics quantities
  - Root common ntuples
- Access and storage
  - dCache with DESY tape library
  - Workgroup servers infrastructure for subsample local storage and analysis

- Analysis on the common ntuples since 2012
- Removal of RAW, MDST, MC tapes from robot
  - Reprocessing of ZEUS data and simulation of new MC is not foreseen after 2011
- Funding within the Collaboration planned till 2013

ZEUS Computing Strategy - revisited

- ZEUS reviewed on the last Collaboration Meeting its Computing Strategy document
  - The collaboration fully supports and commits itself to the concept of Common Ntuples
  - The current analysis model based on MDST will be no longer supported after 2011
  - The possibility of MC simulation will be revised and proposed solution presented

Common Ntuple project - history

Thanks to A.Geiser
CN Project Coordinator

General goals of common ntuple

- reduce use of computing resources
- improve efficiency of analyses
- in the "long" term (2010): fully replace MDST's

implementation

Project defined	January 2006
Prototypes at physics group level	October 2006
1st iteration (2005 RP data + some MC)	June 2007
2nd iteration (2006/7p RP data + full MC)	September 2008
3rd iteration (GR test samples)	February 2009
4th iteration (full HERA II GR data, ~350 M ev.	April 2009
+ "full" MC, ~300 M events)	September 2009

The current status and developement of Common Ntuples

- The full set of Common Ntuples for HERA II data and for almost all relevant MC samples is ready - and in use by analyzers
  - The physics preliminary results exist based on Common Ntuples - Measurement of beauty photoproduction from inclusive secondary vertexing at HERA-II (presented on DIS09)
- The outcome from the user (content revision, fixes) goes into the test iteration for one period of HERA II data and selected MC samples
- The next full iteration will start still this year, in view for usage for physics results in summer 2010
- By the end of 2011 2-3 more iteration foreseen
- Common ntuples are straightforward to use or be simplified for Open Access/Outreach/Education

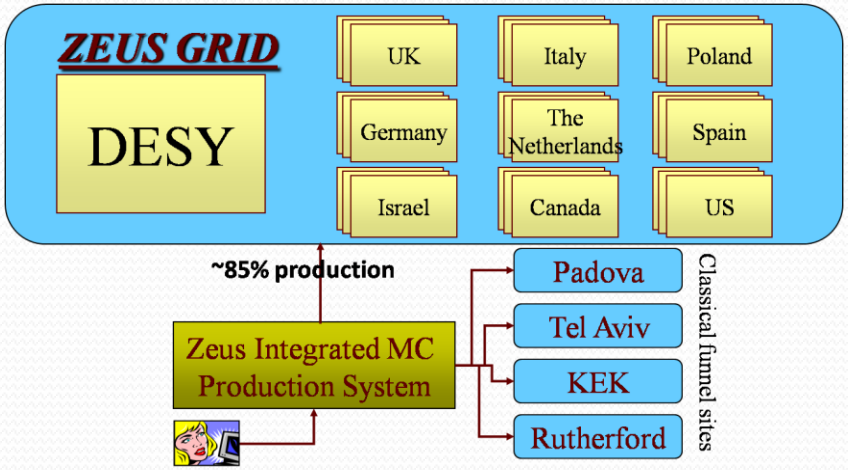
MC simulation status

- The current development and production will continue till the end of support of the current analysis/reconstruction model - or sufficiently earlier to accommodate Common Ntuple production (see also the next remark)
- Ntuplizing included in MC production will be developed
- Possible scenarios for maintaining possibility of MC simulation after 2011
  - Preservation of the current analysis/reconstruction model - not feasible
  - Simplistic detector/trigger/reconstruction simulation - hard to assess
  - Simplified production chain with current executables/calibration etc put into a virtual machine – we try to go for it

MC simulation preservation

- The current production with
  - Classical distributed computing system
  - GRID based system
  - Bookkeeping (databases)

  is hard to be any use without expertise



- The simplified production chain (from generation to simulation to common ntuple production) will be developed using virtualisation techniques
  - based entirely on GRID concepts and follow accepted standards within HEP in order to make maintenance easier
  - Virtualization has already been tried with success within the current production system

OS migration, storage etc

- Currently supported SL3 and SL4
- Deployment of SL5 well advanced
- Will there be SL6 within our timescale?

- Storage and access
  - dCache/tapes
  - workgroup servers for skiming/subsample storage and analysis
- Possibilities beyond 2011
  - keep dCache/tapes ?
  - nfs servers (maybe with http access to root files), with tape backup?
  - subsample generation and analysis on notebooks?

Instead of summary

Common ntuples

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

Virtualised MC simulation