technische universität dortmund

H1 Analysis and Computing Model

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)

First Workshop on Data Preservation and Long Term Analysis

26 - 28 January 2009

Contents

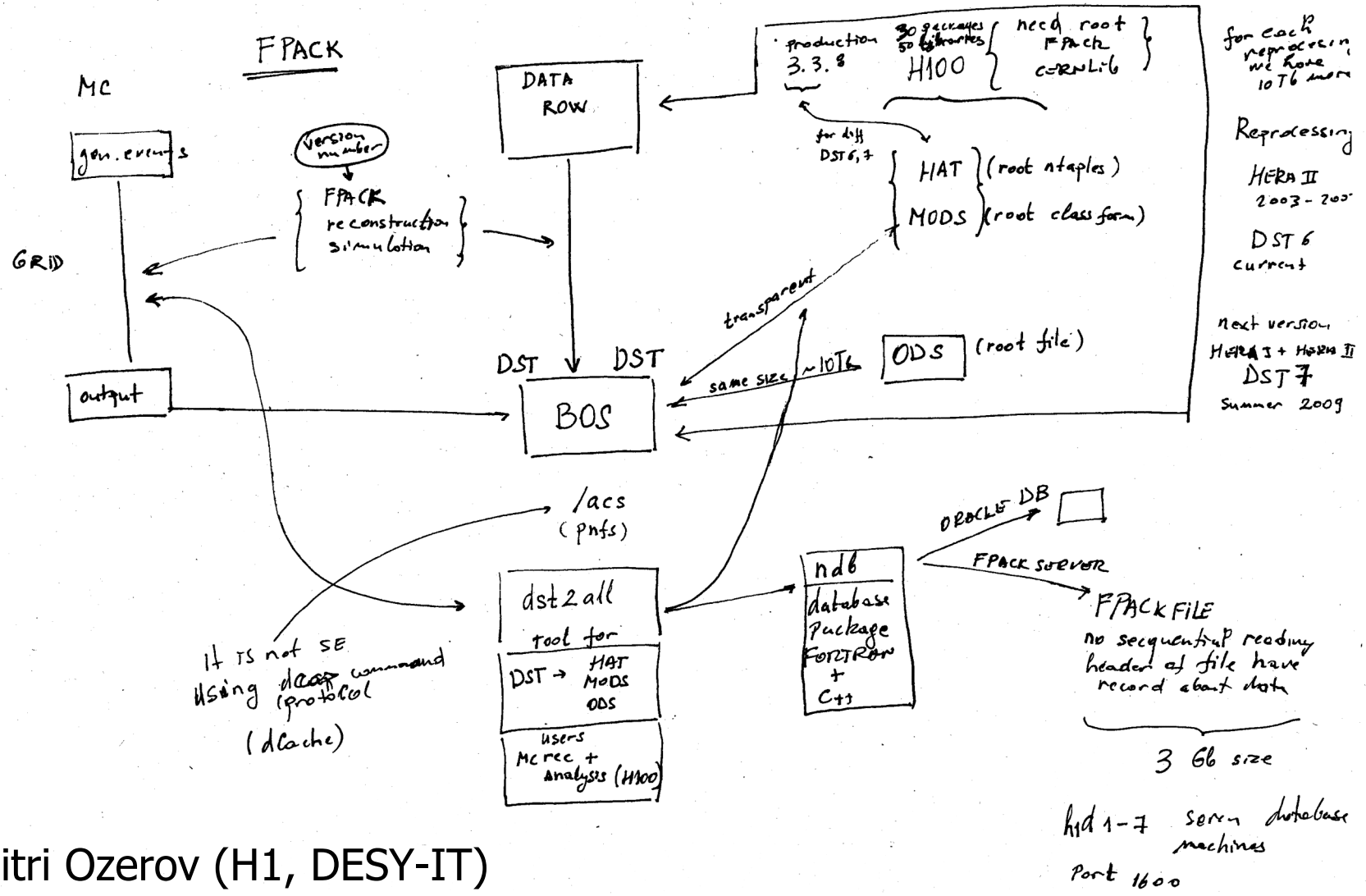
The H1 Data Analysis Model

H1 Data and MC Production

How Analysis is Organised within H1

Status of H1 Data Preservation

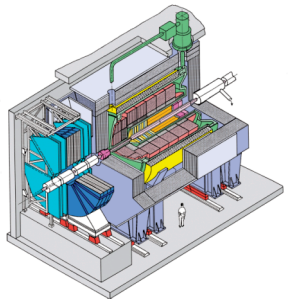
H1 Data Analysis Model



Dmitri Ozerov (H1, DESY-IT)

H1 Data Analysis Model

+

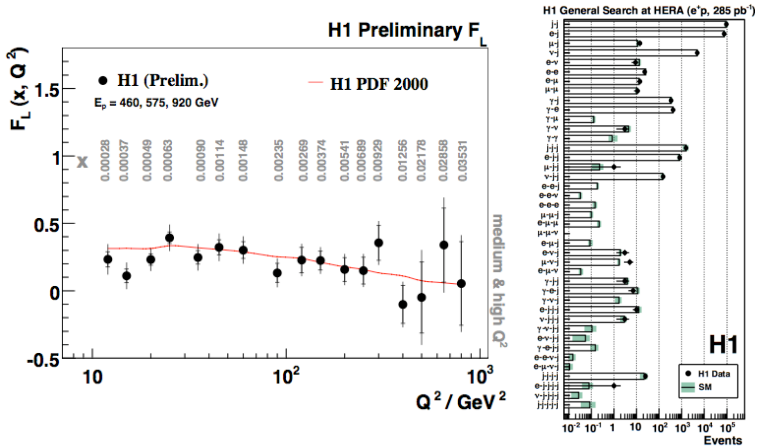


HERA delivered $e^{\pm}p$ collisions 1992-2007 and the H1 Collaboration collected 0.5 fb^{-1} of data, ~ 10^9 events

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

The raw data output from the detector is written to tape

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



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 Data

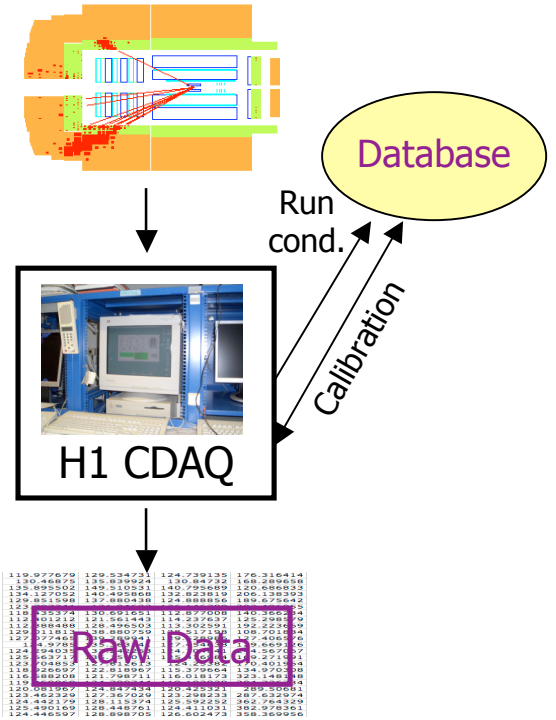
Collisions in the H1 detector every 96ns processed by H1 CDAQ

Physics events written out at 50 Hz

CDAQ also writes run conditions to and reads calibration constants from the H1 database

Raw data contains wire hits, channel numbers, collected charges; typical size 75 kB / event

Data events written as sets of BOS Banks combined and written out as FPACK records



BOS (Bank Operating System): Dynamic data and memory management system (1985)

FPACK: Machine-independent data handling I/O package (1991)

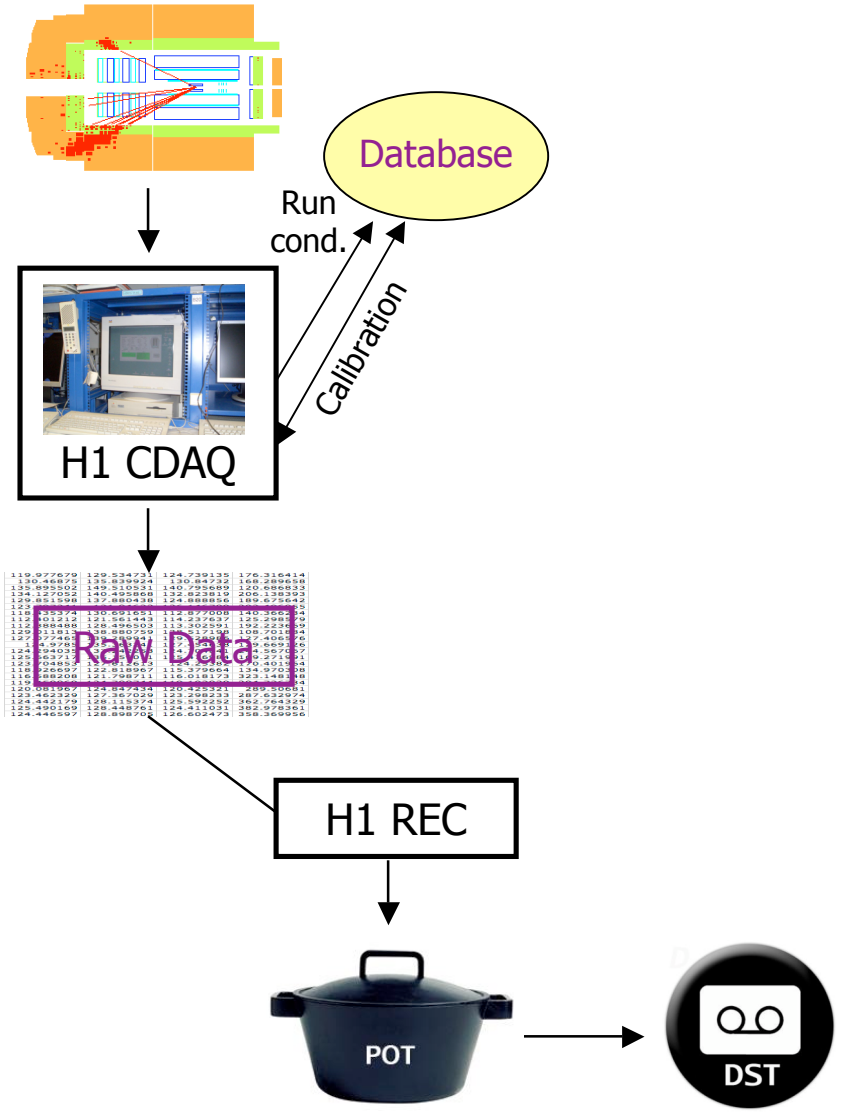
H1 Data Format: DST

Main reconstruction software written in Fortran since before data taking, modularised into several packages

Many packages already frozen, but development continues in particular in H1REC

Output written to POT (Production Output Tape): includes raw data and first reconstruction: clusters of cells, energy sums, first track fits and vertices; typical size 200 kB / event

Most relevant information written from POT to DST (Data Summary Tape); typical size 18 kB / event



H1 Data Format: Monte Carlo

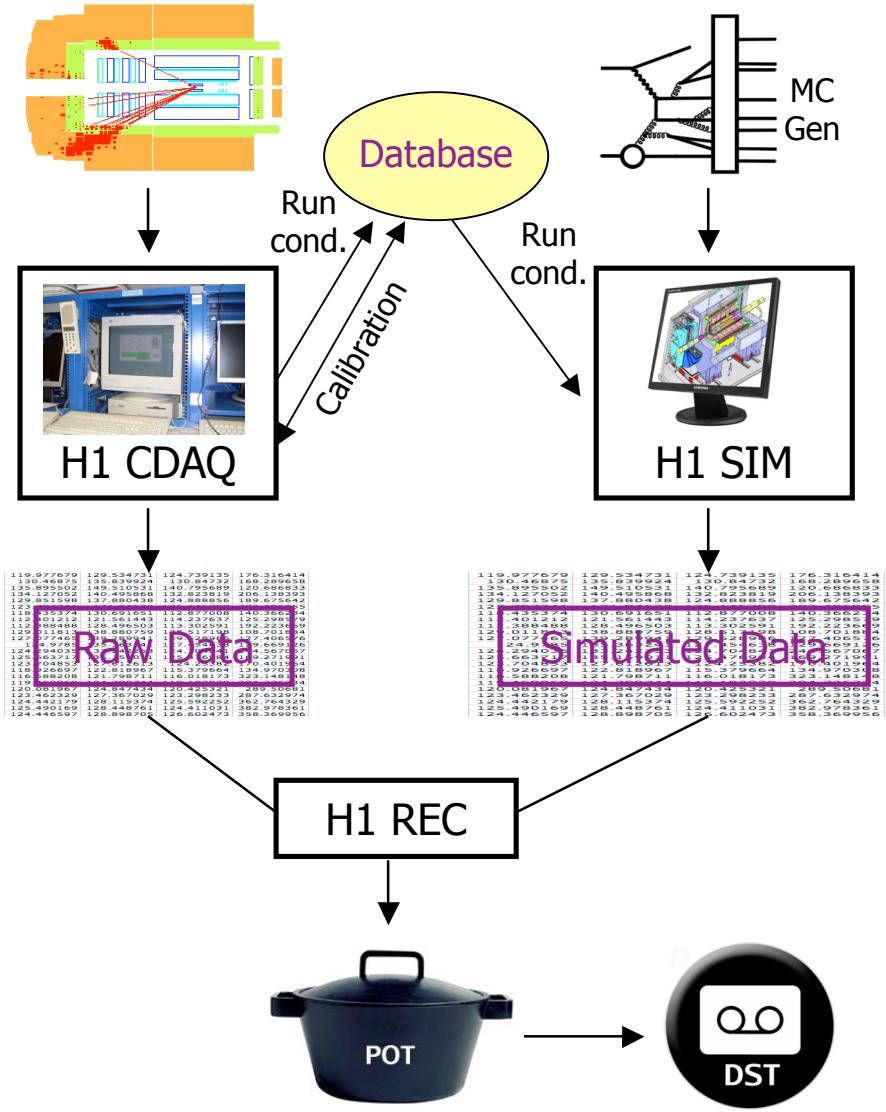
Monte Carlo events main package H1SIM, H1 detector simulation based on GEANT 3, run conditions read back from H1 database

Each event takes about 10 seconds for full simulation; typical size for MC DST of 40 kB / event

Simulated data in identical format to real data from H1 detector, with further generator level information

Same reconstruction software (H1REC version) used for data and MC ensuring equivalency

Large scale MC Production - *see later*



The H1OO Project

H1OO Project started in 2000, around the start of the HERA shutdown, with the aim of enabling physics analysis at H1 in one coherent framework

H1OO is built upon ROOT and uses its structure and basic functionality:

- TObject as base class provides object I/O, error handling, inspection.
- Inheriting from TObject provides the possibility to store objects in collections and to read and write them to files; inheritance extensively used in ROOT and H1OO

Code Organisation:

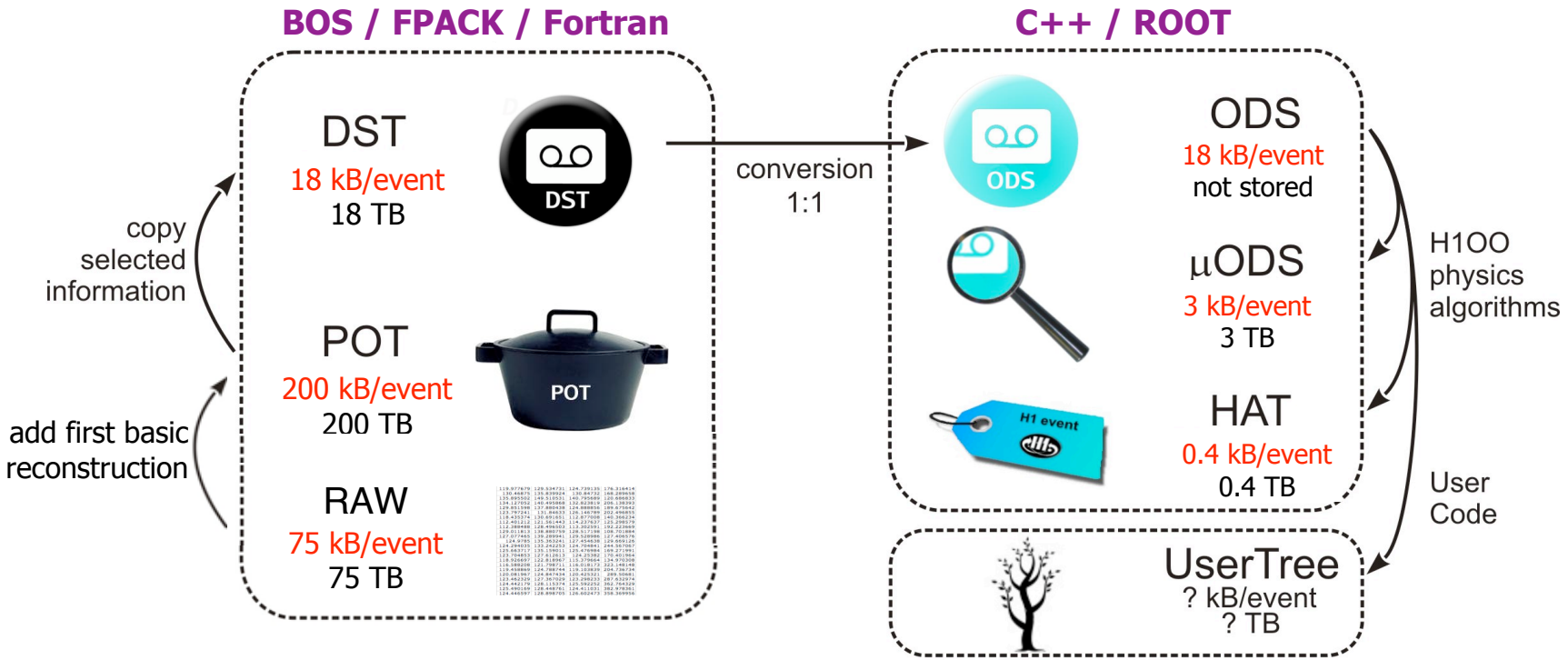
- More than 350,000 lines of code organised in about 600 classes contained in about 50 packages
- Every package is compiled to one (or more) shared libraries no circular dependence among core packages

Release Strategy:

- Release series and production releases linked to DST (reprocessing) versions
- Development H1OO releases every 2-3 weeks

Connection between H1OO and DST

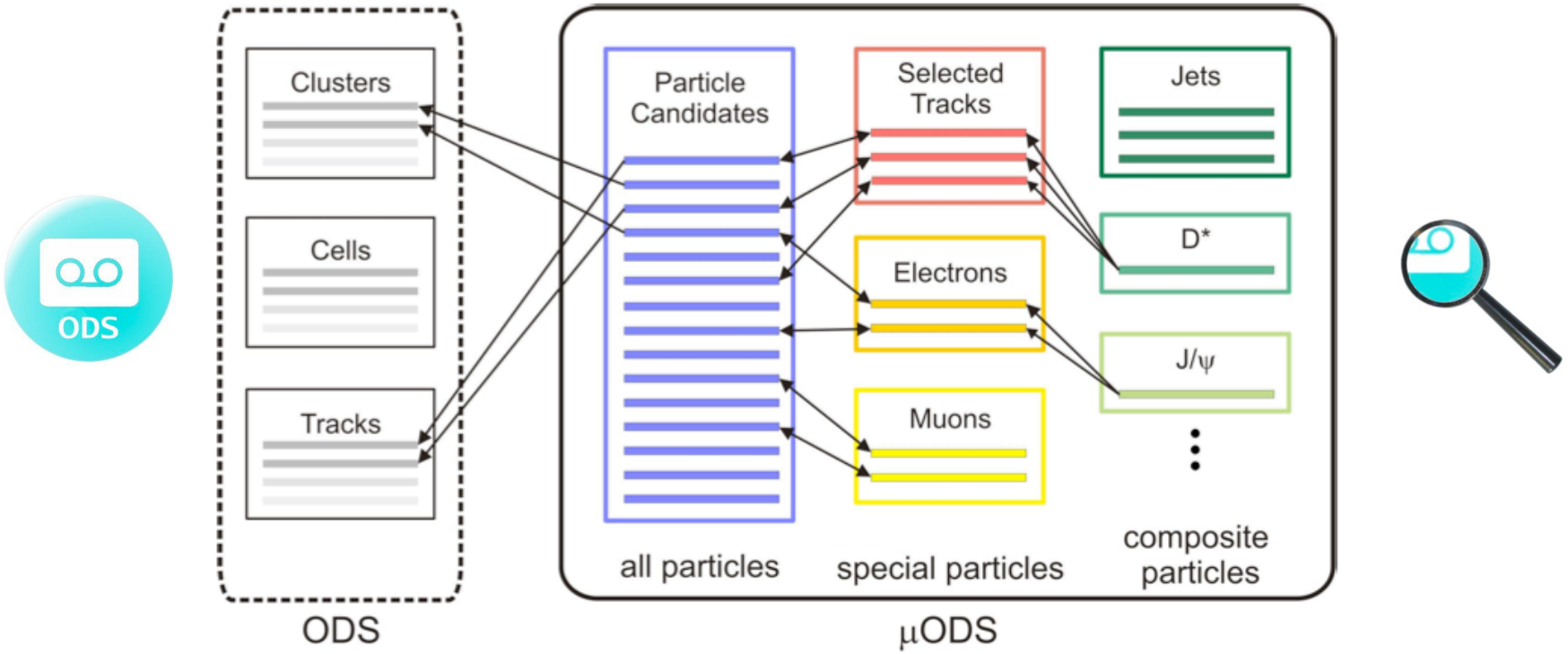
H1OO data composed of 3 layers of ROOT files: Object Data Store (ODS, dynamic access) , a smaller version (μODS) and H1 Analysis Tag (HAT)



We don't want to re-write all the reconstruction software, instead: a 1:1 conversion from DST to ODS: H1Track, H1Cluster objects generic reading of BOS banks

H1Tree (based on ROOT TTree) links all layers together in one event loop

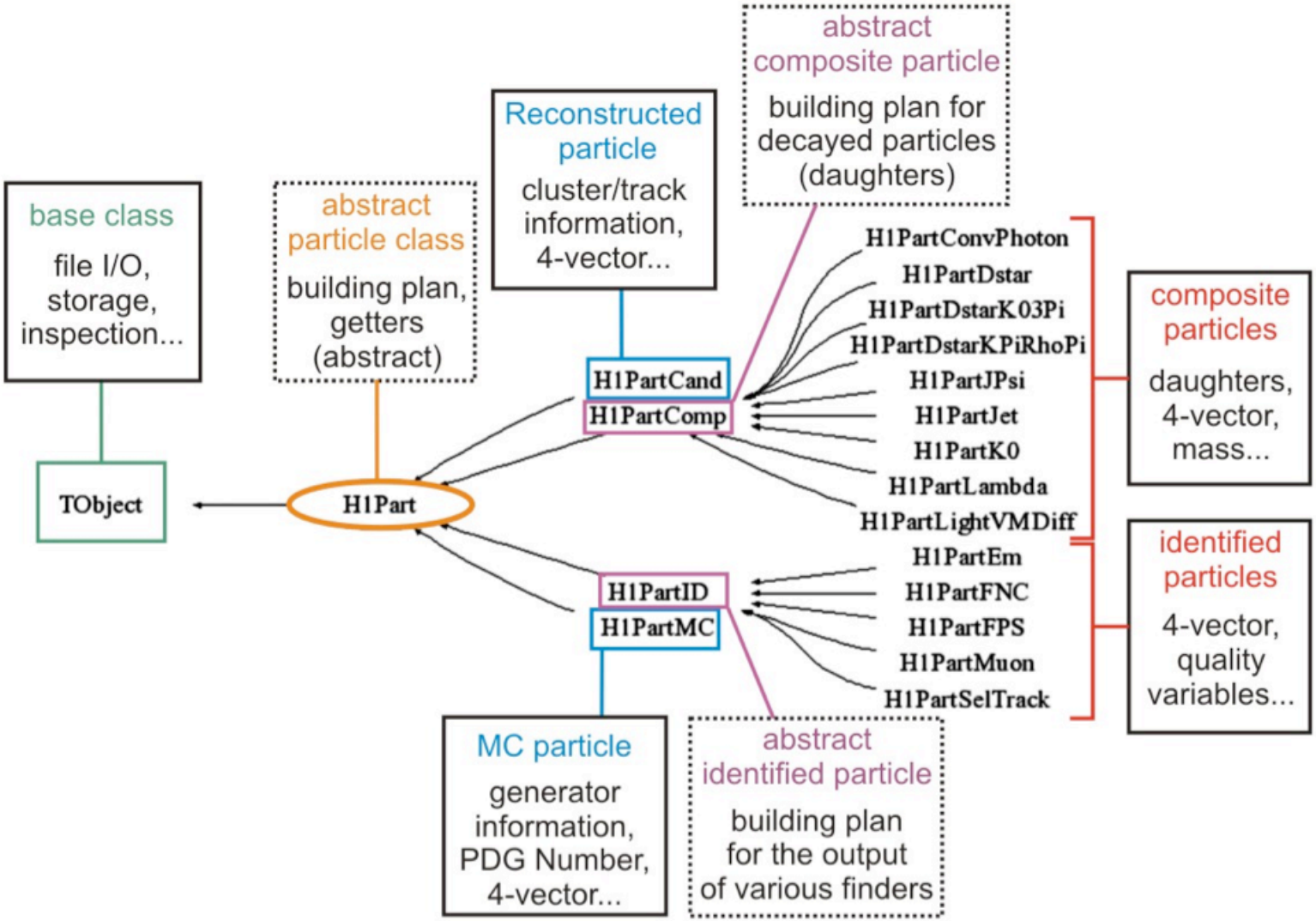
H1OO: micro Object Data Store (μODS)



Not just a selection of ODS information but rather particle candidate finders, written as classes with much use of inheritance

Results of particle finders stored on μODS, with pointers back to ODS (DST) information (original tracks, clusters, cells..)

Use of Inheritance in H1OO: H1Part



H1OO: H1 Analysis Tag (HAT)

0.4 kB/event
0.4 TB in total

Contains around 200 selected basic event variables, stored as flat ntuple format

Kinematics, trigger information, HFS sums, number of tracks, clusters, particles, diffractive variables, as well as generator quantities in the case of MC

Allows a fast event selection: 1.2s for 10^6 events

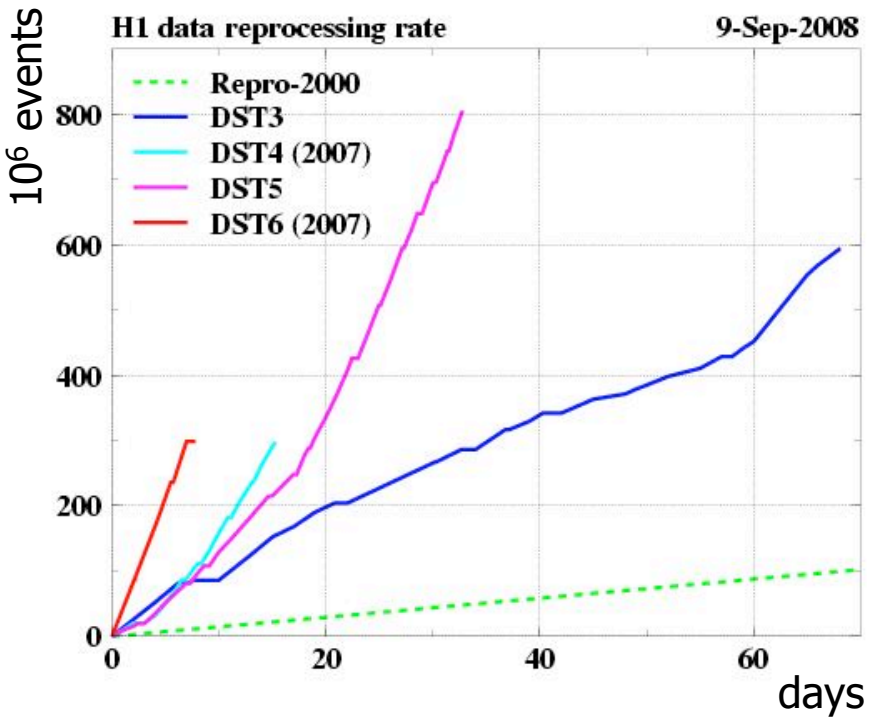
H1OO analysis model:

- Perform HAT selection first, run/event numbers are stored in memory µODS information is read only for events passing the HAT selection

- Access to ODS level done "on the fly", dynamically accessing the DST

- H1 batch farm used by all collaboration for analysis: about 600 CPUs

H1 Data Production: DST Level



Reprocessing of DST level done often during the last few years, made possible due to improved computing power and resources (batch farm also used for reprocessing)

Start from RAW data of *good* and *medium* runs + random events ("GM-cut" files, now stored on disk) and no longer make POT

Now possible to reprocess complete HERA II data (14 TB, 800 million events) in a few weeks

DST data stored on large *acs* and *acsdisk* resources (dCache)

Final reprocessing of HERA I+II data, planned for 2009 (DST 7)

H1 Data Production: H1OO Level

Regular central production of HAT and μODS files of complete dataset, to be used by all H1 analyses

H1OO executable transforms DST to root files and, similarly to DST production, uses FPACK to access the H1 database

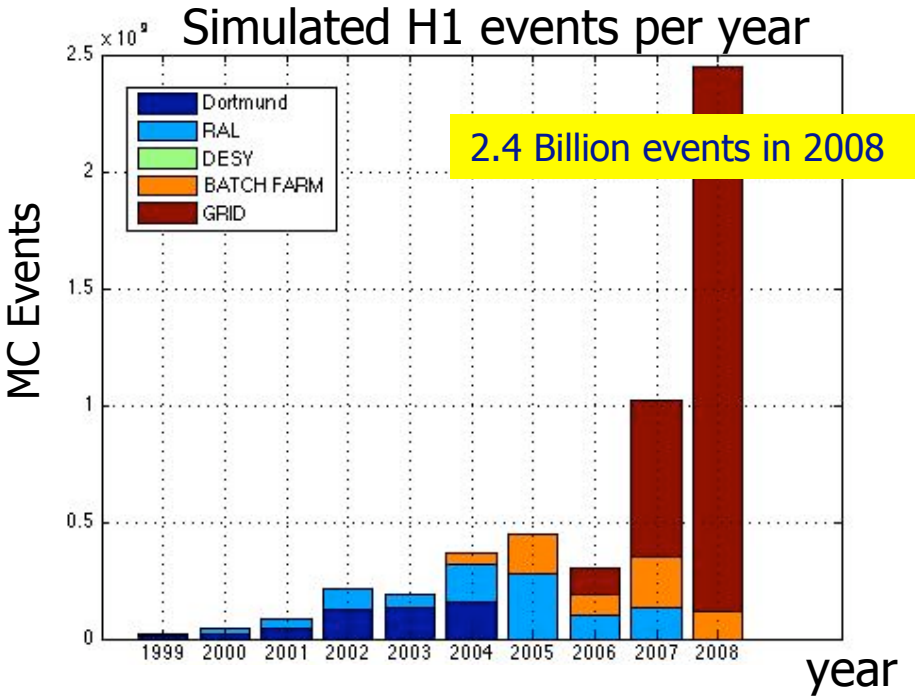
Production done on GRID, all HERA II data can be produced in 1 week, files again stored on *acs* and *acsdisk*

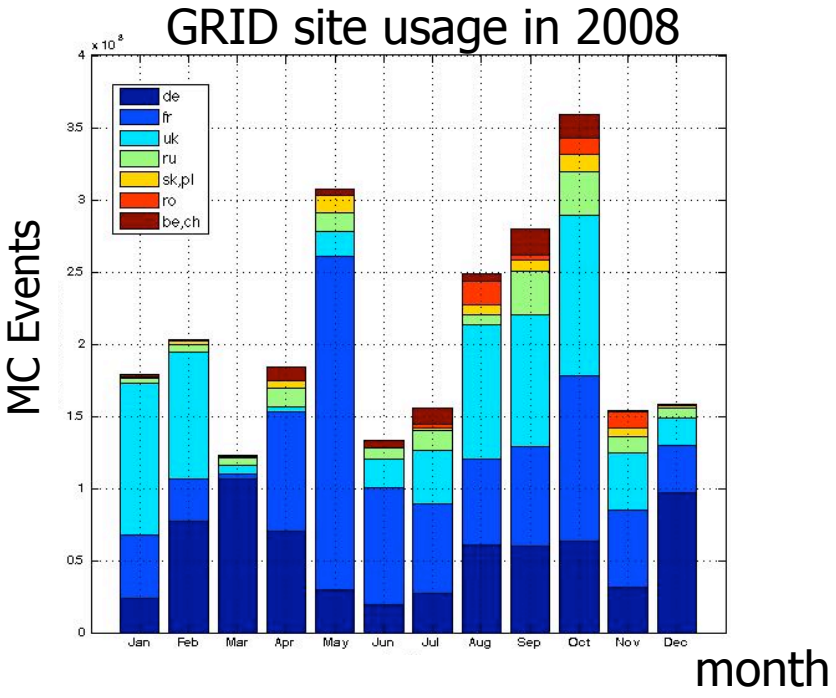
Last major release series 3.4, linked to DST 7, to be produced this year

Analysis level calibrations also done in H1OO, valid for all analyses

Development at analysis level in H1OO continues beyond 2009

H1 MC Production





Recent success story is large MC production thanks to (many) GRID resources: now crucial to analysis at H1

Similarly to data, centralised production of DST files and corresponding H1OO files (exclusively on DESY-GRID) done by team of experts

Organisation of Analysis in H1

H1 Management structure includes two physics coordinators

Five Physics working groups:

ELAN (Structure functions, NC and CC, electron analysis)

HAQ (Hadronic final states, jets, measurement of α_s, QCD analysis)

Diffraction (Rapidity gap events, leading baryons, vector meson production)

Heavy Flavour (Events with charm and beauty)

REX (Searches, rare and exotic processes)

Around 50 analyses active or planned for publication, majority within 2 years

Data and analysis software standard between groups, much sharing of centrally produced MC (representative in each physics working group)

Also several technical groups concerning:

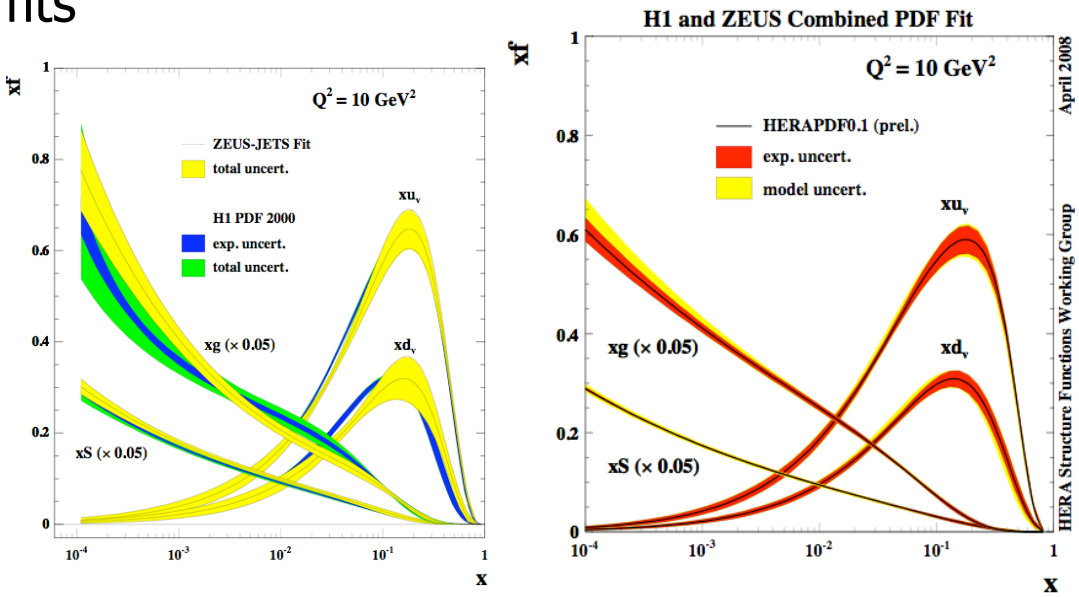
Tracking, Calibration, Fits and of course H1OO and H1 MC Production

H1+ZEUS Combined Analysis

Many combined H1 and ZEUS analyses and groups now active, including:

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 compared to separate results from H1 and ZEUS



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?

Status of Data Preservation at H1

Group of interested people formed within the collaboration to address such issues, the "H1 Data Preservation Task Force"

Which data should be preserved?

- RAW data of GM-cut files, total for HERA: 75 TB
- (Probably) one full set of DST, total for HERA: 18 TB
- A version of μODS and HAT as well (< 4 TB)? *See next page*
- In addition to calibration and cosmic runs, total data about 100 TB
- Amount of MC to be decided, but will be of same order (large set of DST 7)

- Conservatively estimate total amount to preserve at 500 TB

Do not expect to be limited by CPU or disk space in the future (data should be copied on to new media at regular intervals, say every 2 years)

Status of Data Preservation at H1

What about the software?

- FPACK/BOS designed as machine independent, IBM to UNIX conversion already done in 1996 - perhaps problems with 64 bit architecture?
- Also expect few headaches with Fortran code (H1REC, H1SIM..), some already frozen, define what else could be frozen and otherwise static

No further major development of Fortran after DST 7 - but still possible!

- H1OO analysis software is less clear: model is heavily reliant on ROOT framework, in particular I/O, TTree and TChain
- Try to incorporate ROOT updates: ROOT developers to patch H1OO?
- Try to remove as much ROOT as possible from H1OO, leaving only the crucial dependencies to be patched (H1Skeleton package..)?
- Perhaps a similar use of ROOT by many experiments could make this job easier on ROOT developers?

Foresee a "rolling preservation model", with regular recompilation of H1OO software and μODS/HAT file production, say every 3 months

Status of Data Preservation at H1

And finally the documentation?

- Much already in place, thanks to diligent authors
- Optimal use of ROOT provides much html documentation already

Concentrated effort still needed in the coming years

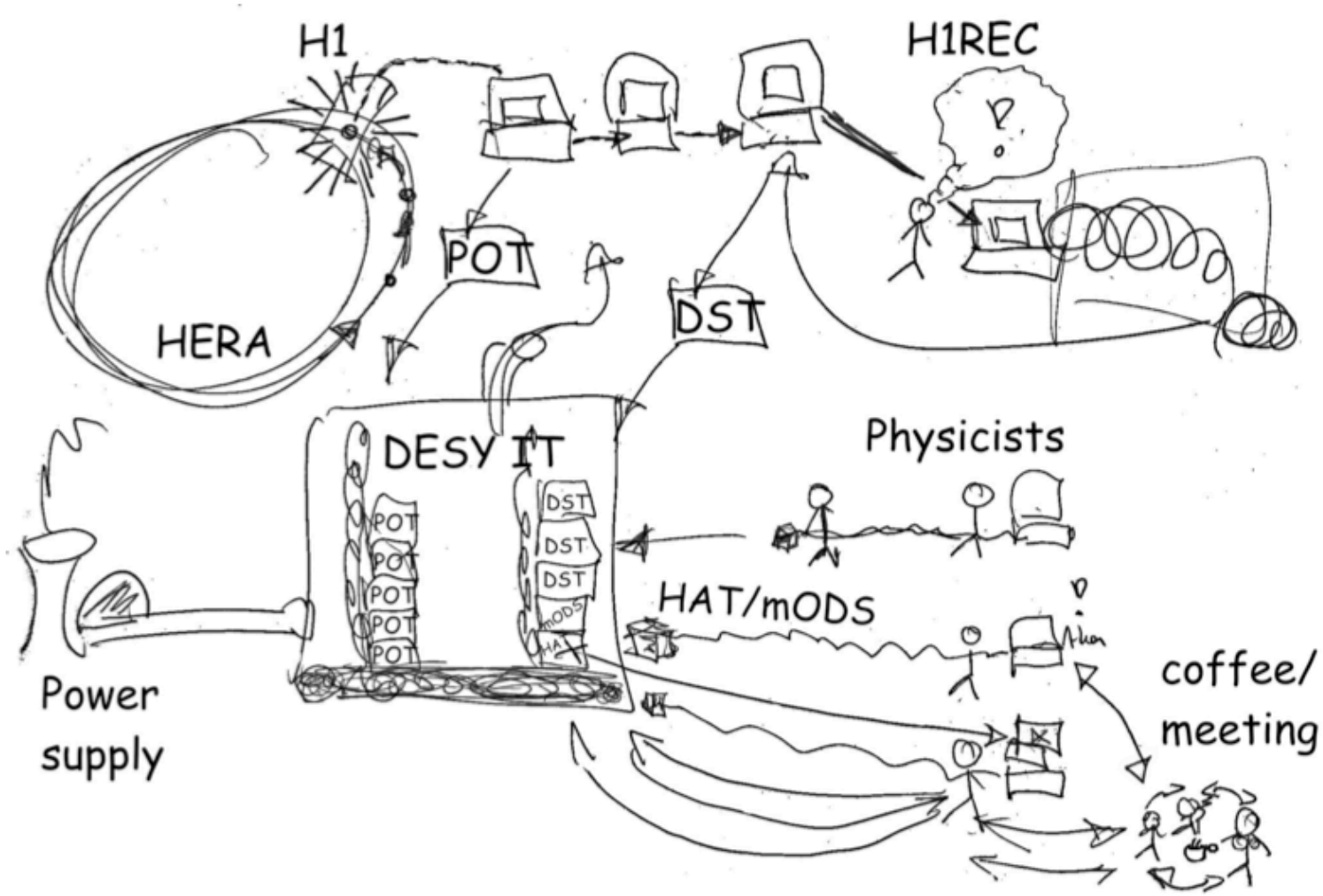
Other issues?

- Possible new flat format in H1OO possible made up of HAT and four vectors from finders in μODS
    - Could be independent of ROOT
    - Could be candidate for a common format with ZEUS?

- Future format of the H1 Collaboration itself, beyond 2013

- Open access of H1 data at some point in the future also under discussion

Summary



Roman Kogler, H1OO Release Coordinator

Summary

H1 reconstruction software approaching final version, including 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

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