Data preservation efforts at LEP (ALEPH, DELPHI, L3)

First Workshop on Data Preservation and Long Term Analysis in HEP

26.-28. January 2009
DESY, Hamburg

André Holzner
CERN
and
PARSE.Insight (EU FP7)

Outline

- L3 issues
  - Analysis within the collaboration
  - Collaboration with other experiments
  - Data analysis model
  - Preservation scenarios / what was preserved
  - Documentation
  - State of preservation in 2009
  - A few not-so-technical remarks
- ALEPH specific issues
- DELPHI specific issues

(OPAL will be/has been presented by Matthias Schroeder in a seprate talk)

L3

contributions from: Salvatore Mele, Luca Malgeri, Luc Pape

Analysis within the collaboration

- Working groups:
  - Fermion-pair, two-photon, Higgs*, QCD, SUSY*, WW, ZZ* (* = shared common inter-group ntuples containing jets and leptons in predefined event-topologies)
- Number of active analyses:
  - In 2001: about 80
  - Today: ~6 electroweak, 1 QCD analysis,
  - Succesfully completed the analysis program in the first three years after shutdown
  - all key electroweak measurements published by 2006

- Shared MC/data production responsibilities:
  - centralized production up to clusters/tracks level managed by 2-3 FTE
  - Analysis-group specific data formats ('ntuples') done by typically one person per analysis group
- Analysis organisation within the collaboration:
  - Mostly one person working on one analysis
  - Analysis code private in most cases
  - Two analyses per channel for 'sensitive' analyses (Higgs, WW)

Collaboration with other epxeriments

- Extent of collaboration with sister experiments:
  - LEP Higgs working group:
    - See talk by Peter Igo-Kemenes
  - LEP SUSY working group:
    - Exchange of 2D Histograms (data,signal efficiencies, background) for combinations
  - LEP Electroweak working group:
    - systematic uncertainty matrices in numerical form
  - WW working group:
    - LEP-Wide ASCII files with MC four-vectors (before hadronization) e.g. for assessment of common systematics
- Participation required at least some level of openness and standardisation (which are prerequisites for preservation and open access).

Data analysis model

- Levels of abstraction:
  - 'DSU': tracker/calorimeter hits and tracks/cluster,
  - 'DVN': tracks and clusters only
- Calibration:
  - At the beginning of each data-taking year (LEP II): LEP delivered a few pb^{-1} at the Z resonance for calibration purposes.
  - Mapping of operational/non-operational periods of detector parts onto MC production after data taking ('RDVNs')
  - Further MC tuning necessary in some cases: e.g. determine tracking chamber efficiencies from data more accurately for B-Tagging using an iterative procedure
- Databases:
  - run information, file catalogs
  - Flat files and FATMEN
  - Note that this information needs to be archived as well !

Data analysis model

- Size of individual (RAW) events: ??
- Software releases:
  - Revision management: PATCHY
  - Frequency: 1-3 per year (in the final years)
    - Last releases: mainly bugfixes, improved tracking / b-tag
  - Validation: some distributions checked by eye for new releases
- Monte Carlo simulation strategy:
  - Detector simulation:
    - Geant3 based
    - No fast simulation available
  - Physics simulation: a variety of generators: PYTHIA, Excalibur, KORALW/Z, etc.
  - Production: Analysis groups prepare data cards, centralized production managed by 2-3 FTE

Data analysis model

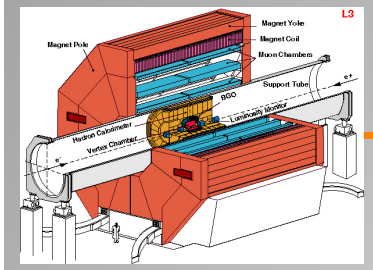
- Storage and access:
  - DSU/DVN stored in CASTOR at CERN
    - Developed a C++ framework to access DVN (tracks/clusters)
    - Wrote an application using the FORTRAN code to read the reconstructed data and store it in ROOT trees
  - It was not foreseen to give access to anybody outside the collaboration

Data Tier	Size (Data)	Size (MC)
DSU (hits)	1161 GB	6473 GB
DVN (tracks, clusters) ZEBRA	90 GB	2490 GB (RDVN) 1659 GB (DVN)
DVN (tracks,clusters) ROOT	106 GB	2104 GB
New particle ntuples (analysis-group specific)	23 GB	151 GB

Data analysis model

- Business model of data preservation:
  - Agreement with CERN/IT to keep tapes and support
  - "Agreement" to keep CERNLIB
  - CASTOR storage:
    - Paid for tapes once
    - Migration for free
      - negligible compared to cost of LHC tapes
      - Expect drop in price / TB with newer technologies
  - Relying on people still being around in other experiments in case of need to reanalyze the data

Preservation scenario A



Raw data

Tracks, Clusters, Jets, Leptons,

Analysis group wide data format

Histograms, matrices,…

Reco.

MC

Det.Sim.

Physics model

Analysis group wide code

Selec-tion

Inter-preta-tion

Analysis group wide code

Selec-tion

Inter-preta-tion

Selec-tion

Inter-preta-tion

g)

h)

e)

f)

b),c),d)

a)

arXiv

Preservation scenario A

- Scenario A:
  - Work your way from the publication towards the raw data (go as far as you can):
    a) Preserve publications electronically (arXiv)
    b) Preserve histograms, uncertainty matrices etc.
    c) Preserve personal ntuples
    d) Preserve tools ("macros") to go from c) to b)
    e) Preserve analysis groups data
    f) Preserve tools ("personal analysis code") to go from e) to d)
    g) Preserve reconstruction output (tracks, clusters)
    h) Preserve tools ("analysis group code") to go from g) to f)
    i) Etc.

- Unfortunately, b)-d) and f) did not happen
- e) happened for searches-related groups
- Further preservation (of raw data) is probably impossible

Example of scenario A

EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN-PH-EP/2004-024
June 8, 2004

Studies of Hadronic Event Structure in e^+e^- Annihilation from 30 GeV to 209 GeV with the L3 Detector

The L3 Collaboration

Abstract

In this Report, QCD results obtained from a study of hadronic event structure in high energy e^+e^- interactions with the L3 detector are presented. The operation of the LEP collider at many different collision energies from 91 GeV to 209 GeV offers a unique opportunity to test QCD by measuring the energy dependence of different observables. The main results concern the measurement of the strong coupling constant, α_s, from hadronic event shapes and the study of effects of soft gluon coherence in charged particle multiplicity and momentum distributions.

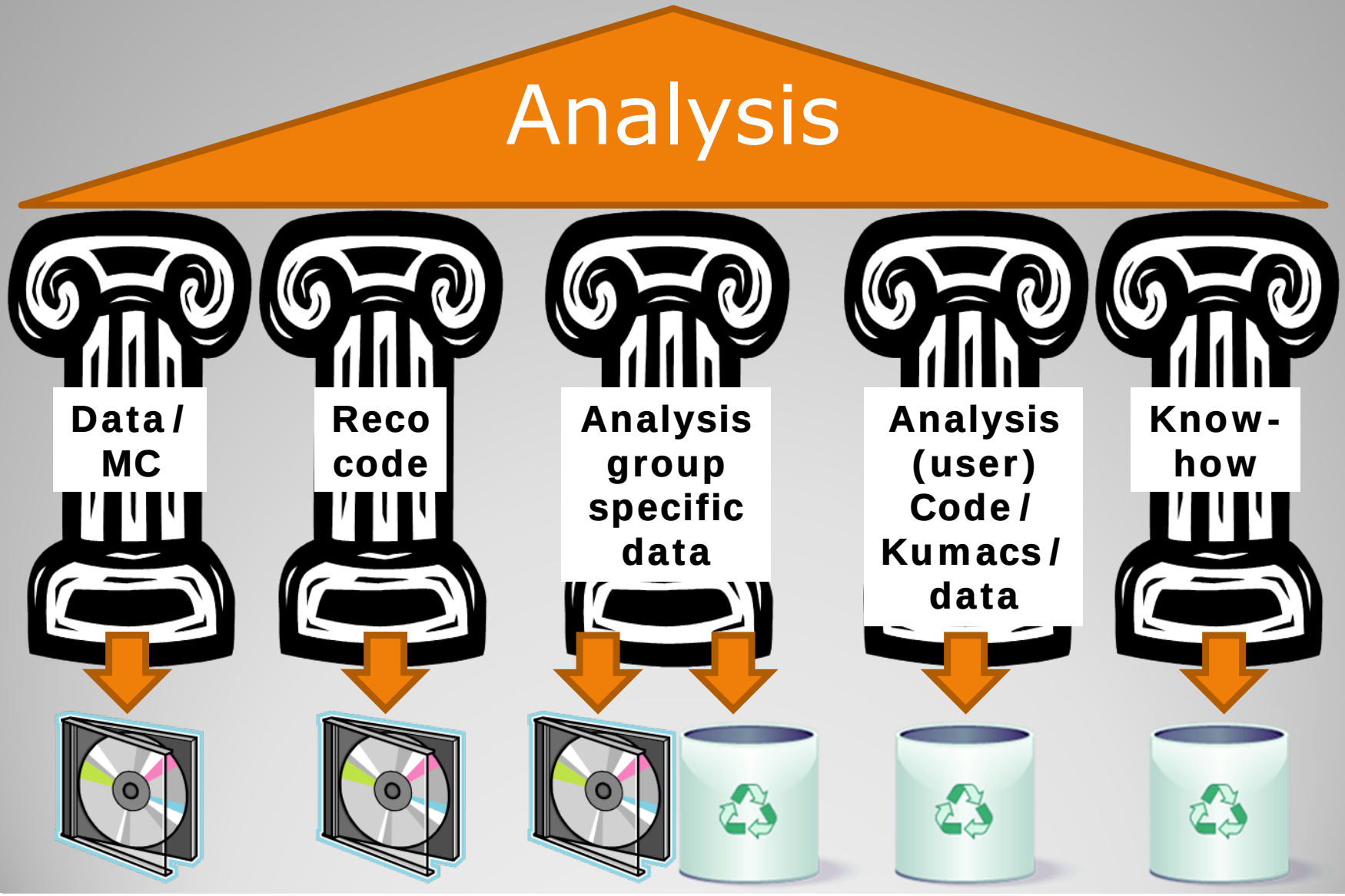
Submitted to *Physics Reports*

	$\frac{1}{\sigma}\cdot\frac{d\sigma}{dD}$		
D	at $\sqrt{s}=$ 194.4 GeV	at $\sqrt{s}=$ 200.2 GeV	at $\sqrt{s}=$ 206.2 GeV
0.000–0.016	34.894 ± 1.033 ± 0.772	33.486 ± 1.015 ± 0.586	33.974 ± 0.795 ± 0.595
0.016–0.032	9.195 ± 0.517 ± 0.296	9.230 ± 0.510 ± 0.537	8.981 ± 0.387 ± 0.218
0.032–0.048	4.744 ± 0.377 ± 0.218	4.744 ± 0.417 ± 0.340	4.963 ± 0.293 ± 0.128
0.048–0.064	3.568 ± 0.341 ± 0.159	3.175 ± 0.319 ± 0.137	2.909 ± 0.241 ± 0.132
0.064–0.080	1.975 ± 0.284 ± 0.158	2.321 ± 0.288 ± 0.147	2.618 ± 0.239 ± 0.096
0.080–0.096	1.683 ± 0.274 ± 0.327	2.172 ± 0.281 ± 0.108	1.530 ± 0.210 ± 0.130
0.096–0.112	1.633 ± 0.295 ± 0.228	1.350 ± 0.257 ± 0.120	1.647 ± 0.221 ± 0.137
0.112–0.128	1.086 ± 0.255 ± 0.099	0.860 ± 0.261 ± 0.302	0.923 ± 0.179 ± 0.107
0.128–0.144	0.961 ± 0.354 ± 0.181	0.742 ± 0.222 ± 0.152	1.093 ± 0.206 ± 0.154
0.144–0.160	0.469 ± 0.215 ± 0.232	0.855 ± 0.244 ± 0.175	0.326 ± 0.172 ± 0.091
0.160–0.176	0.437 ± 0.231 ± 0.175	0.712 ± 0.224 ± 0.145	0.499 ± 0.170 ± 0.088
0.176–0.192	0.295 ± 0.208 ± 0.123	0.855 ± 0.273 ± 0.138	0.667 ± 0.189 ± 0.145
0.192–0.208	0.171 ± 0.167 ± 0.057	0.418 ± 0.204 ± 0.146	0.234 ± 0.156 ± 0.080
0.208–0.224	0.288 ± 0.183 ± 0.071	0.308 ± 0.215 ± 0.101	0.547 ± 0.187 ± 0.163
0.224–0.240	0.418 ± 0.206 ± 0.125	0.215 ± 0.169 ± 0.141	0.337 ± 0.167 ± 0.085
0.240–0.256	0.045 ± 0.134 ± 0.212	0.461 ± 0.221 ± 0.170	0.301 ± 0.157 ± 0.137
0.256–0.272	0.266 ± 0.171 ± 0.109	0.018 ± 0.025 ± 0.188	0.000 ± 0.000 ± 0.000
0.272–0.288	0.000 ± 0.000 ± 0.000	0.000 ± 0.000 ± 0.000	0.057 ± 0.128 ± 0.133
0.288–0.304	0.000 ± 0.000 ± 0.000	0.098 ± 0.144 ± 0.032	0.045 ± 0.108 ± 0.044
0.304–0.320	0.036 ± 0.104 ± 0.026	0.000 ± 0.000 ± 0.000	0.124 ± 0.117 ± 0.078
0.320–0.336	0.022 ± 0.060 ± 0.059	0.149 ± 0.147 ± 0.102	0.148 ± 0.115 ± 0.100
0.336–0.352	0.074 ± 0.103 ± 0.024	0.054 ± 0.112 ± 0.113	0.018 ± 0.017 ± 0.059
0.352–0.368	0.000 ± 0.000 ± 0.000	0.048 ± 0.075 ± 0.038	0.000 ± 0.000 ± 0.000
0.368–0.384	0.000 ± 0.000 ± 0.000	0.025 ± 0.093 ± 0.010	0.000 ± 0.000 ± 0.000
0.384–0.400	0.039 ± 0.046 ± 0.086	0.000 ± 0.000 ± 0.000	0.000 ± 0.000 ± 0.000
First Moment	0.0387 ±0.0023± 0.0047	0.0435 ±0.0028± 0.0037	0.0429 ±0.0029± 0.0033
Second Moment	0.0056 ±0.0010± 0.0016	0.0064 ±0.0010± 0.0021	0.0064 ±0.0012± 0.0020

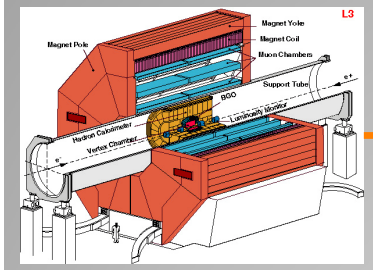
Table 52: Differential distribution for D-parameter at $\sqrt{s}$ = 194.4, 200.2 and 206.2 GeV. The first uncertainty is statistical, the second systematic.

140 pages of tables

What was preserved ?



Preservation scenario B



Raw data

Tracks, Clusters, Jets, Leptons,…

fourvectors

histograms

Physics model

Reco.

generic pre-selection

QUAERO

Inter-preta-tion

Inter-preta-tion

Inter-preta-tion

MC

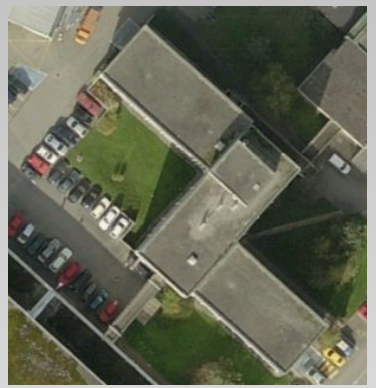
Det.Sim.

Preservation scenario B

- Scenario B:
  - We had an opportunity to implement QUAERO in L3
  - Produce four-vectors with a selection which is as loose as possible
  - Pros:
    - Easily allows for new future searches (e.g. testing future extensions of the Standard Model)
    - A whole lot of machinery is included:
      - Automatic checking of agreement data/MC
      - Search across hundreds of distributions for best sensitivity
      - Interface to MadGraph
      - Automatic procedure to tune a fast simulation
      - Web interface to submit new physics models
    - Don't have to run after each analysis person individually
  - Cons:
    - Need to understand your data **really** well. Such an "overall analysis" was only attempted once before in L3 and then abandoned
- We invested some effort (fraction of a person), based on "new particles" analysis group tuples with a somewhat loose selection
  - Tried to understand (and possibly fix) discrepancies in data
  - gave up at some point

Documentation

- A big part of the documentation was stored here:



- Documentation maintenance: 3 meals per day + Coffees (+ other costs)

- Very expensive solution in the long term (especially after data taking has stopped)

State of preservation in 2009

- A handful of collaboration members still around, scattered across the globe, if still in HEP…
  - Editorial board still working
  - Still publishing some papers
- A few simple tests done last week:
  - Reconstruction program at least produces a welcome message when run on lxplus

    Most of the libraries are linked in statically, some (e.g. libshift) however dynamically, may lead to problems on newer platforms
  - List of MC requests still readable
  - PAW seems to run (problems opening remote display though)
- However:
  - CERNLIB support not foreseen any more on SLC5: lack of manpower, especially for certification on a 64bit operating system.

    IT IS NOT CLEAR WHETHER WE WILL BE ABLE TO ACCESS LEP DATA ON SLC5 !

A few not-so-technical remarks

- Preservation effort started too late. We consider it failed.
- However, publication effort was a success !
- Among possible reasons for the failure of the preservation effort:
  - Effort started too late (after data taking was completed)
  - Based on 1-2 persons, not even working 100% on it
  - Everybody's analysis code was 'private' (stored in user's directory, not in central storage)
  - Inheriting of analysis typically by person-to-person oral training instead of providing documentation
  - Private corrections (e.g. additional smearing of MC) often did not go into central code
  - People left to other experiments quickly after end of data taking

A few not-so-technical remarks

- Among possible reasons for the failure of the preservation effort:
  - Reconstruction code hard to read: All function names six characters long (and the two first reserved for the subsystem…).
  - With the overlap of concepts of (open) access and reservation, and the difficulty of opening a debate at a later stage in the life of the collaboration, priority was given to complete data analysis AND publish multi-dimensional distributions allowing at least some re-interpretation of the most unique data

A few not-so-technical remarks

(my personal observations comparing the late LEP and the early LHC era)

- Things which potentially could have helped:
  - Twiki for easy documentation update
  - Person actively running after people to update the documentation (as e.g. in CMS)
  - Use of tools known outside HEP as opposed to HEP-specific solutions (e.g. CVS vs. PATCHY) flattens learning curve for outsiders/newcomers

    (note that some technological choices had historical and cultural reasons)
  - LEP-Wide combination of more analyses, at higher level of detail
  - Spread of collaboration across different time zones (like LHC experiments), encourages email exchanges and filling Wikis with documentation
  - (young) people who 'grew up' with the habit of looking for information/documentation on the internet and using keyboards

ALEPH

contributions from: Marcello Maggi, Roberto Tenchini

ALEPH specific information

- All analyses based on Mini-DST (one format for the whole collaboration)
- Databases: ORACLE based
- Using BOS instead of ZEBRA for data storage
  - Worked ok on SLC4, not yet clear on SLC5

- Number of active analyses: one, potentially a few more
- Policy that each ALEPH member can publish an analysis using ALEPH data (without collaboration approval) under certain conditions (see 'use of ALEPH data' on ALEPH home page):
  - Excludes certain cases (e.g. some precision measurements)
  - Also for pedagogical/teaching uses
  - Four papers published under this scheme

ALEPH specific information

- Data: stored on castor, in addition:
  - Each participating institute has a laptop with a frozen analysis system and 2TB of disk
    - Immune to operating system upgrades (don't connect it to the network for security reasons)
    - Geographically distributed backup !
- Business model of long term support: goodwill of experts still around
- List of long-term contact persons for analyses on the ALEPH web page
- Expert to interface to new physics generators available

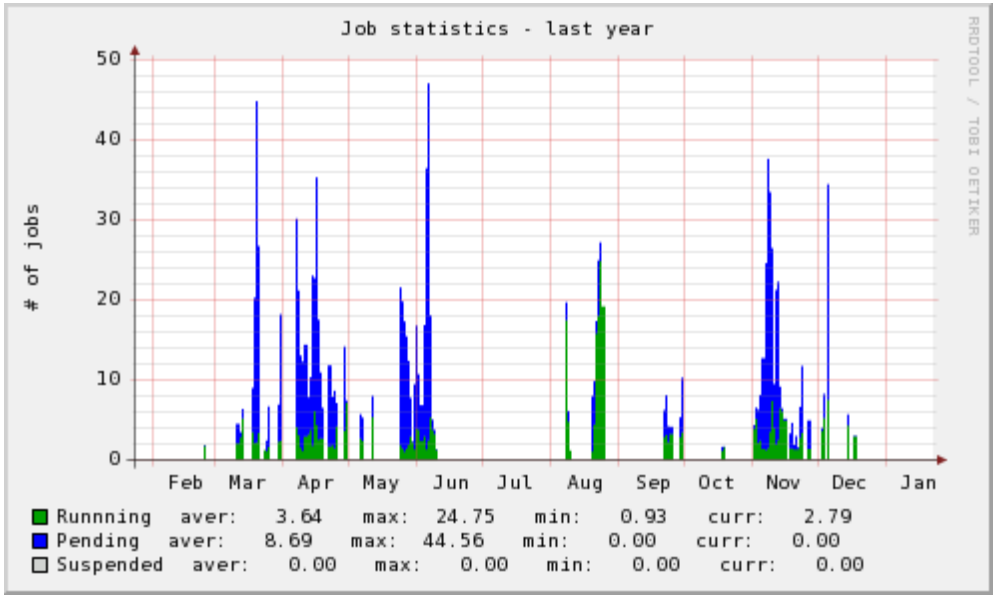
DELPHI

Original slides by Ulrich Schwickerath

Additional contributions from:
Ryszard Gokieli, Jan Timmermans

DELPHI status

- 10 papers still to be published, for 5 of them analysis probably still ongoing
- 165 registered users (not blocked accounts) at CERN
- all data (raw, short dst and MC) stored on CASTOR
- Fatmen database replaced by flat files (ASCII) years ago
- recently migrated to CASTOR2
- experiment software/libraries on AFS



DELPHI activity during 2008 at CERN

Raw Data (total): 6224 GByte
(includes baba+cosmics)
~600 GByte e+e- data

MC data (total): 11955.8 GByte

Detector state date: 2 GByte

Software archiving

Software – CD project:

- includes all DELPHI software in source code
  - includes all required build scripts to boot strap a working environment
  - does not rely on AFS
  - contains everything needed to run simulations and data reprocessing
    - CERNLIB
    - delsim (detector simulation code)
    - delana (event reconstruction code)
    - dstana (user analysis frame work)
    - delgra (event display, broken since some time due to GPHIGS)
    - idea (C++ user analysis framework, not possible to generate new MC !)
  - comes as a tar ball
- ... but no data and no generators included

Last updated: Oct. 2004

Issues

- broken event display
  - delgra is broken since we went to SLC4 because of changes of some symbols in glibc
  - no man power to port it to openGL or similar
- no 64bit support (on linux)
  - no reliable 64bit version of CERNLIB available
  - experts left, so in case of problems it will be difficult to debug
- impossible to migrate to SLC5 and beyond
  - CERNLIB is unavailable
  - CERN IT will move to SLC5 within the time scale of MONTHS.

Issues

- Tape loss discovered last weekend:
  - bad news about a tape which stored some of the DELPHI raw data files and MC used for analyses in the past, which apparently is LOST now.
  - DELPHI still has the hope that they can restore at least part of this data
- Clearly raises the question if it were possible to have replicas of all the LEP data eg. at the LHC Tier 1 centers.
  - This would be around 100TByte in total for LEP, including all real data, MC and raw data all years included.
  - DELPHI strongly supports this idea !