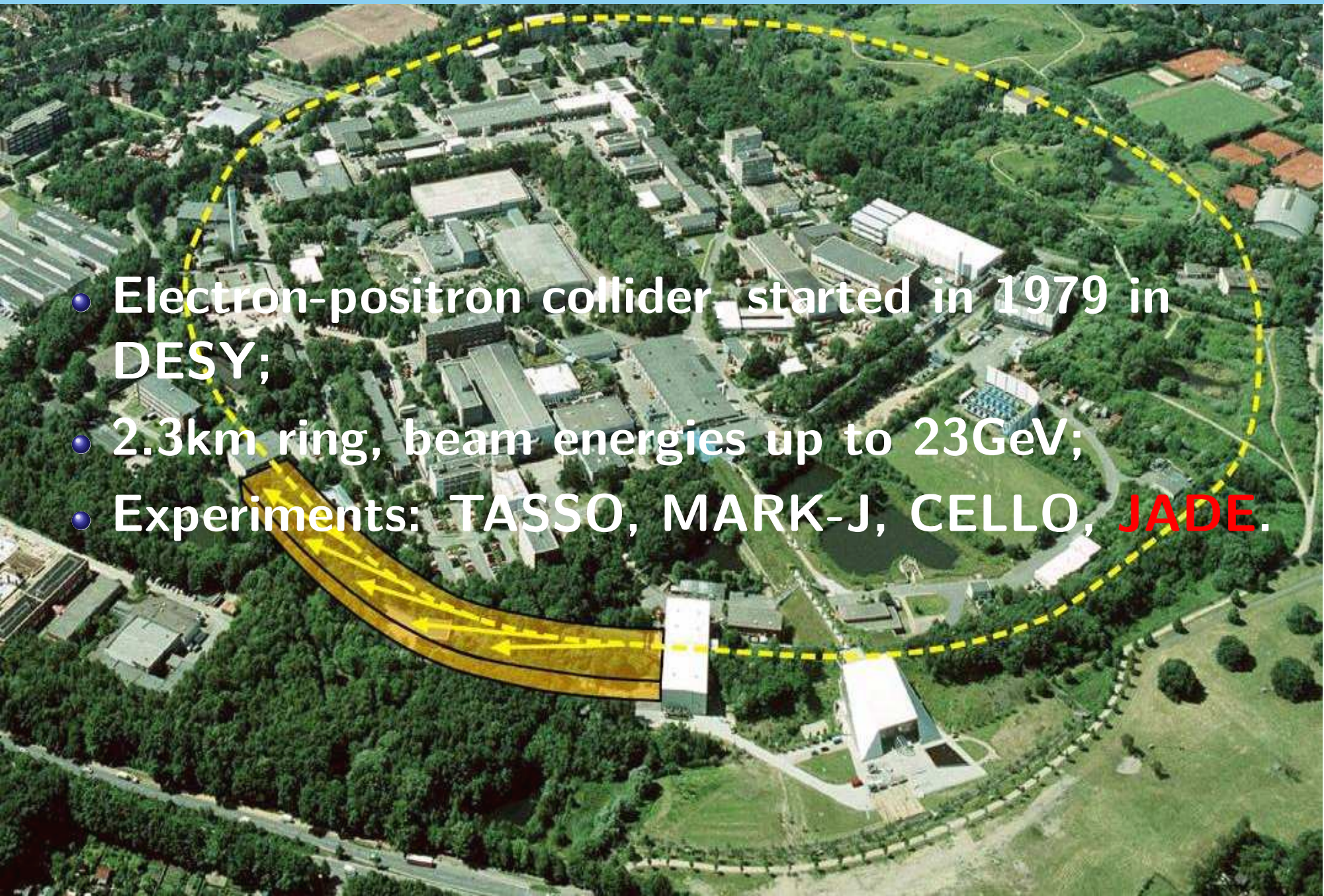
The JADE long term data preservation projects in Max-Planck Institute für Physik

Andrii Verbytskyi on behalf of JADE Resurrection Group

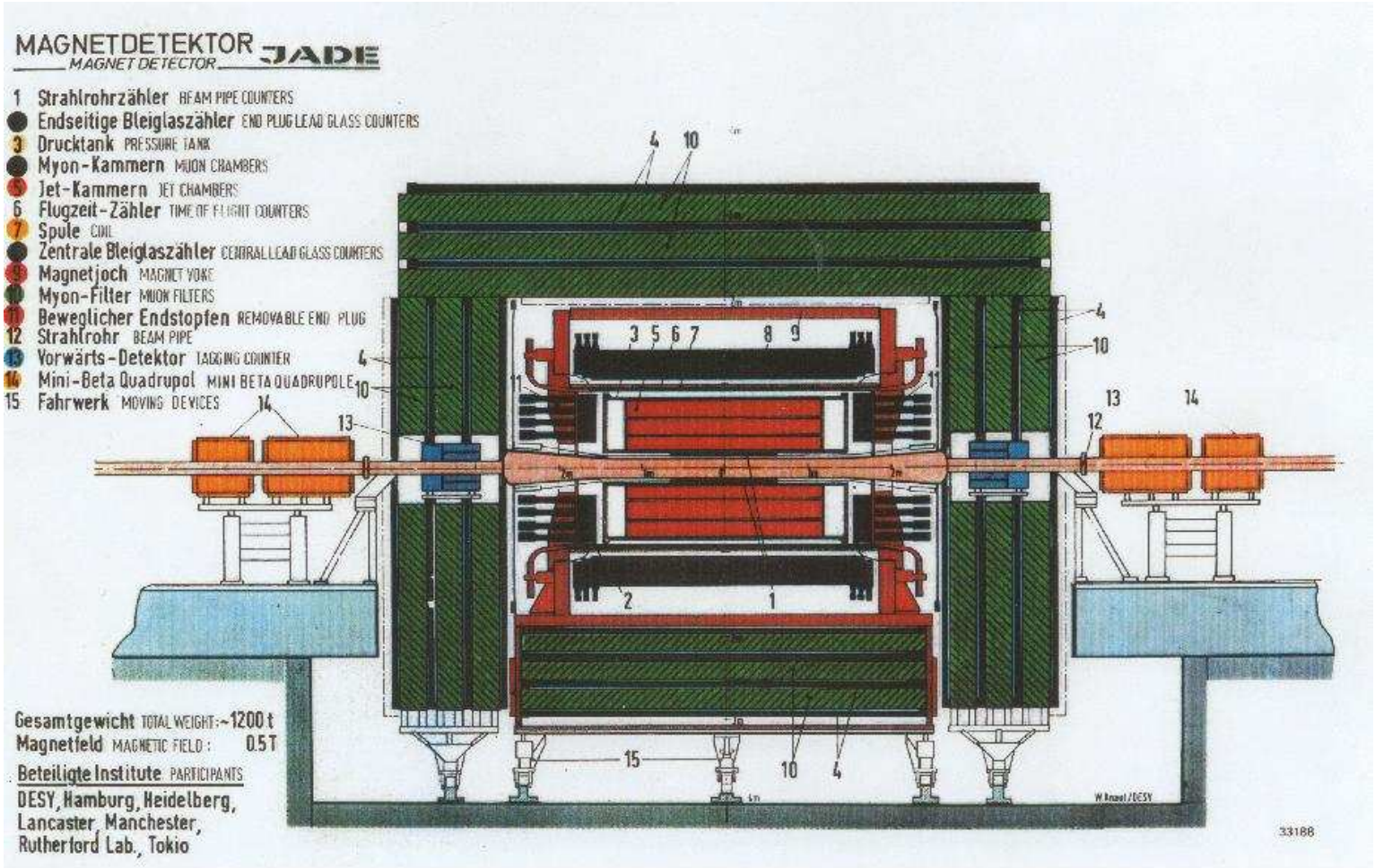
2nd DPHEP Collaboration Meeting
Geneva,
March 14, 2017

PETRA



- Electron-positron collider started in 1979 in DESY;
- 2.3km ring, beam energies up to 23GeV;
- Experiments: TASSO, MARK-J, CELLO, JADE.

JADE



- multipurpose detector
- large solid angle coverage
- digital readout
- advanced tracking system
- calorimeter
- muon chambers

35 years later: still unique energy range coverage!

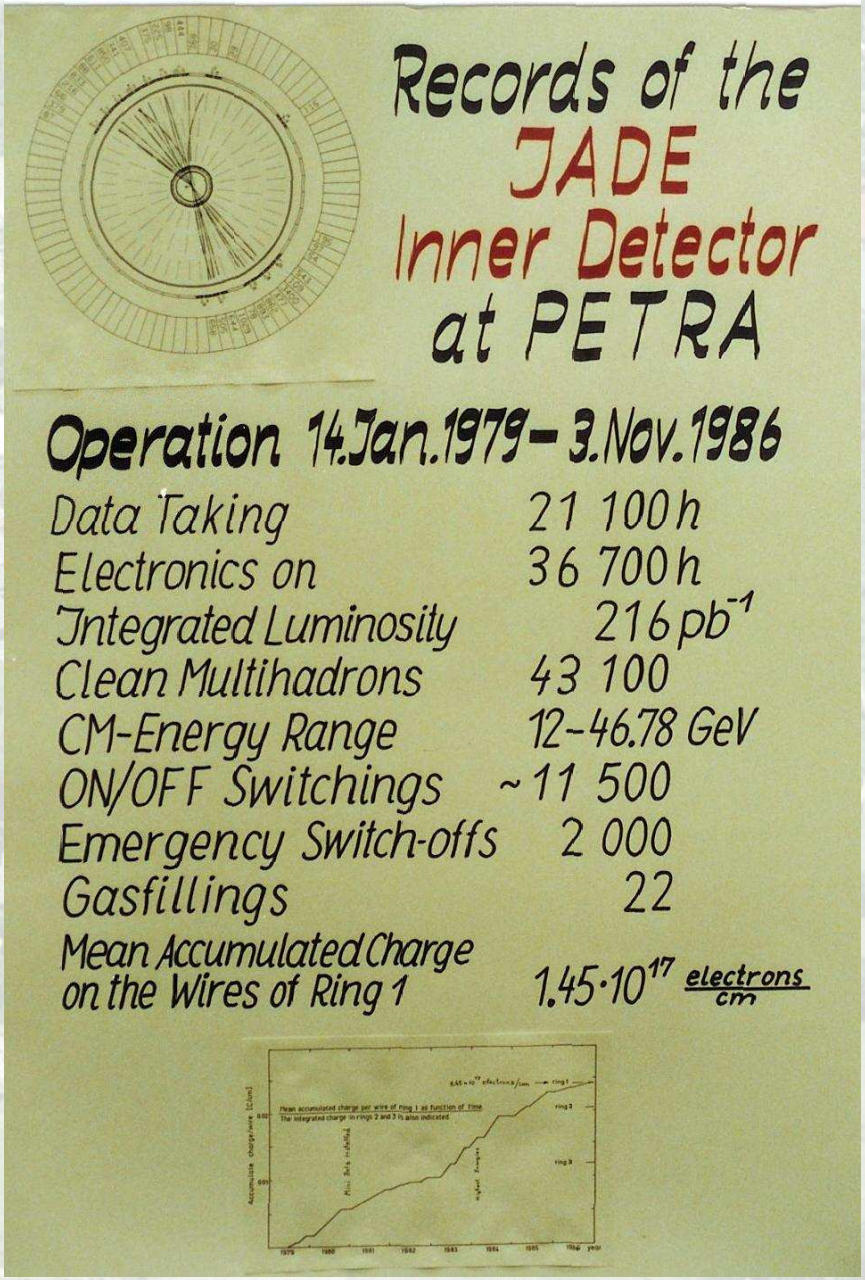
JADE and data preservation

JADE@PETRA reminder:

- One of key experiments for QCD: discovery of gluon, α_s measurements.
- **The oldest and most successful Data Preservation effort!**

Motivation for data preservation:

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



JADE data preservation history

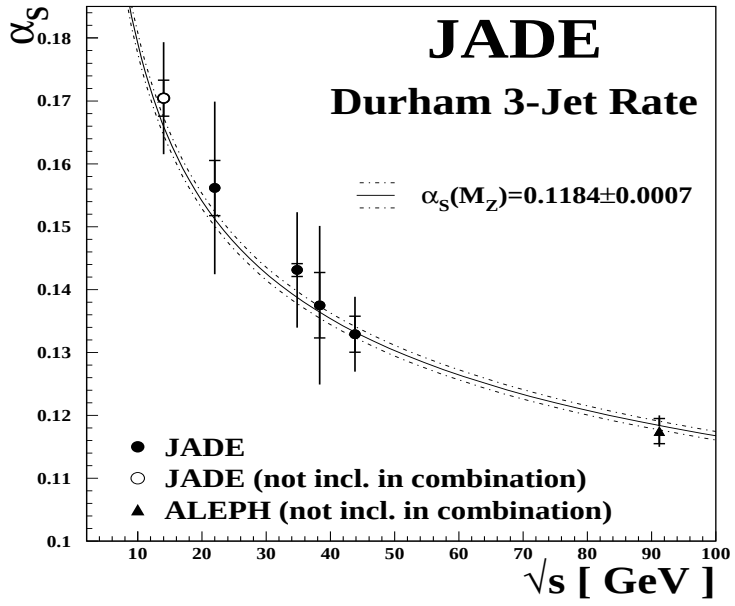
- 1986: End of data taking.
- 1995: Private initiative to rescue data (J. Olsson)
- 1995-2003: Preservation in MPP
  - Physical transfer of data to MPP;
  - Software port to AIX4.3@IBM RS6000;
  - Interface to Pythia6 and Herwig MC;
  - PAW output;
  - Preservation of paper documentation.
- 1996-2013: Physics
  - 11 papers, O(40) conference talks, thesis, JADE notes.
- 2016-2017: Update of preservation:
  - Data is available online;
  - Software port to Linux(Mac)@$x86_64$;
  - Virtualisation;
  - Interface to HepMC3 (enables most modern MC generator);
  - ROOT output;
  - Preparation of digital documentation on computing notes.

MPP DP model for JADE

Yes, 11 papers!

Data preservation is about new and interested results with old data. In addition the Data Preservation experience with JADE has an extreme importance on itself.

In out model we describe ingredients and tools:



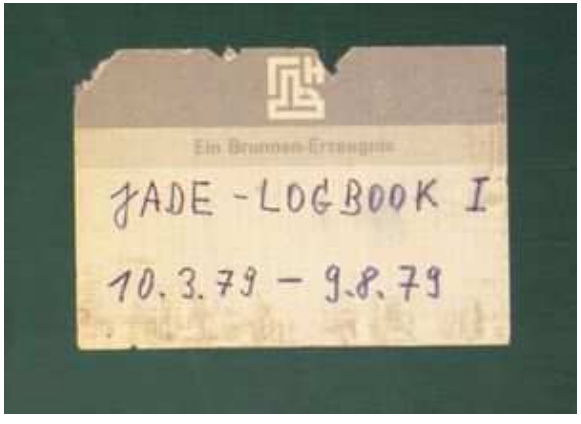
EPJ C73 (2013) 3, 2332

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

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

MPP DP model for documentation and policies

- JADE publications are available in InSpire, journals, arXiv or scanned by KEK.
- The non-digital documentation is preserved in DESY and MPP.
- Logbooks included!
- Some available online as well, see details at https://wwwjade.mpp.mpg.de/



MPP DP model for JADE data bits

- JADE data are stored in MPCDF on locally accessible tapes and in disk pool.
- Access via multiple protocols with grid tools worldwide to disk pools.
- Straightforward procedure to add new (MC) samples.
- In the end nowadays all the data from JADE can fit to a modern USB stick.

JADE data in MPP: Bits statistics

Data are PAW ntuples, ASCII files and FPACK compressed data.

	MPCDF
Files:	**5.4k**
Volume:	$600GB$ **(**645×10^9 **bytes)**
Work area:	**yes**
Access:	**Worldwide**
Protocols:	**Multiple, see list**
Auth:	**Grid certificate**

MAX PLANCK COMPUTING & DATA FACILITY
RECHENZENTRUM GARCHING DER MAX-PLANCK-GESELLSCHAFT
MPCDF

Available at:

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

MPP model for software preservation

Explicit effort put to make software it work in the next 10-15 years.
Previous efforts and high quality of code made it possible.
Key ideas:

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

JADE software in MPP

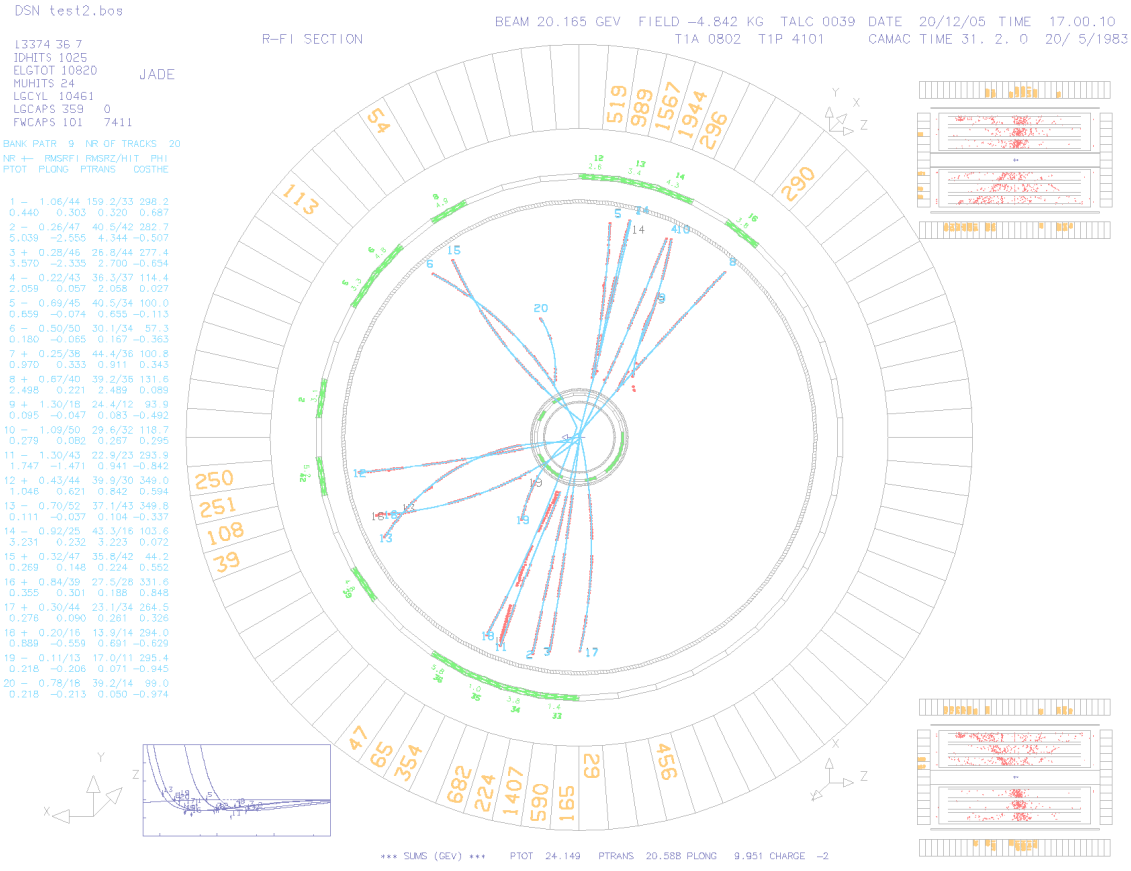
- Full chain for reconstruction of raw data to PAW/ROOT ntuples was resurrected.
- Main software for the analysis of early preserved and reconstructed data is vanilla PAW or ROOT(via h2root).
- Additional software includes:
  - event display;
  - Monte-Carlo generation packages;

(Original)Software is available at https://wwwjade.mpp.mpg.de

JADE software porting quest in detail

- Original sources in FORTRAN IV (1974), FORTRAN 77, Sheltran, Mortran, Assembler ported in 2003 to AIX4.3@RS6000, the old routines replaced with CERNLIB libraries and compiled with IBM FORTRAN.

  Huge work!



- In 2017 everything is problematic: AIX@ppc is Big Endian architecture, no CERNLIB in modern systems, IBM FORTRAN is not for free, 32 bit systems are dying out etc.

JADE software porting quest: step one

One can start from:

- PowerPC64 multi-arch CentOS7 Big Endian on QEMU2.7;
- CERNLIB compilation for PowerPC **Yes, for many flavours!**;
- Trial version of IBM xlf15 for Linux;
- Manual or semiautomatic fixes of differences between FORTRAN dialects (e.g. HOLLERITH, INTEGER vs. CHARECTER).

FYI: Some codes date back to 1974.

JADE software porting quest: step two

Once the software passes simplest tests:

- Switch to gfortran;
- Start to use cmake (best thing for FORTRAN!);
- Switch to Little Endian Linux and control I/O Endianess with GFORTRAN_CONVERT_UNIT ;
- Replace as many as possible CERNLIB routines with ROOT or dummy routines;
- The remaining routines just copy from CERNLIB (i.e. create "picocernlib");
- **Compile codes in native 64-bit mode.**

JADE software porting quest: step two

Once the code is compiled on a standard computer, one has to fix bugs in functionality or implement missing features.
Here the problems begin.

- Part of the code was lost long time ago: muon reconstruction is not available.
- Event display was ported to HIGZ/CERNLIB in 2003. To the version of CERNLIB that is hard to find. Most graphics primitives re-implemented in ROOT, but display is still not fully operational.
- The data is compressed with FPACK utility, but the way it was done is not known, same as the way to unpack it. A collaboration with J.Olsson might be helpful.

JADE software porting quest: step three

On the positive site:

- The key to productive usage of JADE data in the past was an option to generate MC with much newer generators. The option was implemented with integration of Pythia/Herwig/Ariadne codes into JADE software.
- Now it is simpler: Monte Carlo generators have standard output formats and only a converter from a standard format to JADE-specific input has to be developed.
- Done with HepMC3, very nice library.

JADE software environment/VM

A certain environment is needed for the analysis.

- 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.
- Unlimited number of installations $\rightarrow$ potentially usable on clouds;
- Usage not restricted to any laboratory or virtualisation software. Can run anywhere.
- Usage of ISO installation image assures vendor independence.

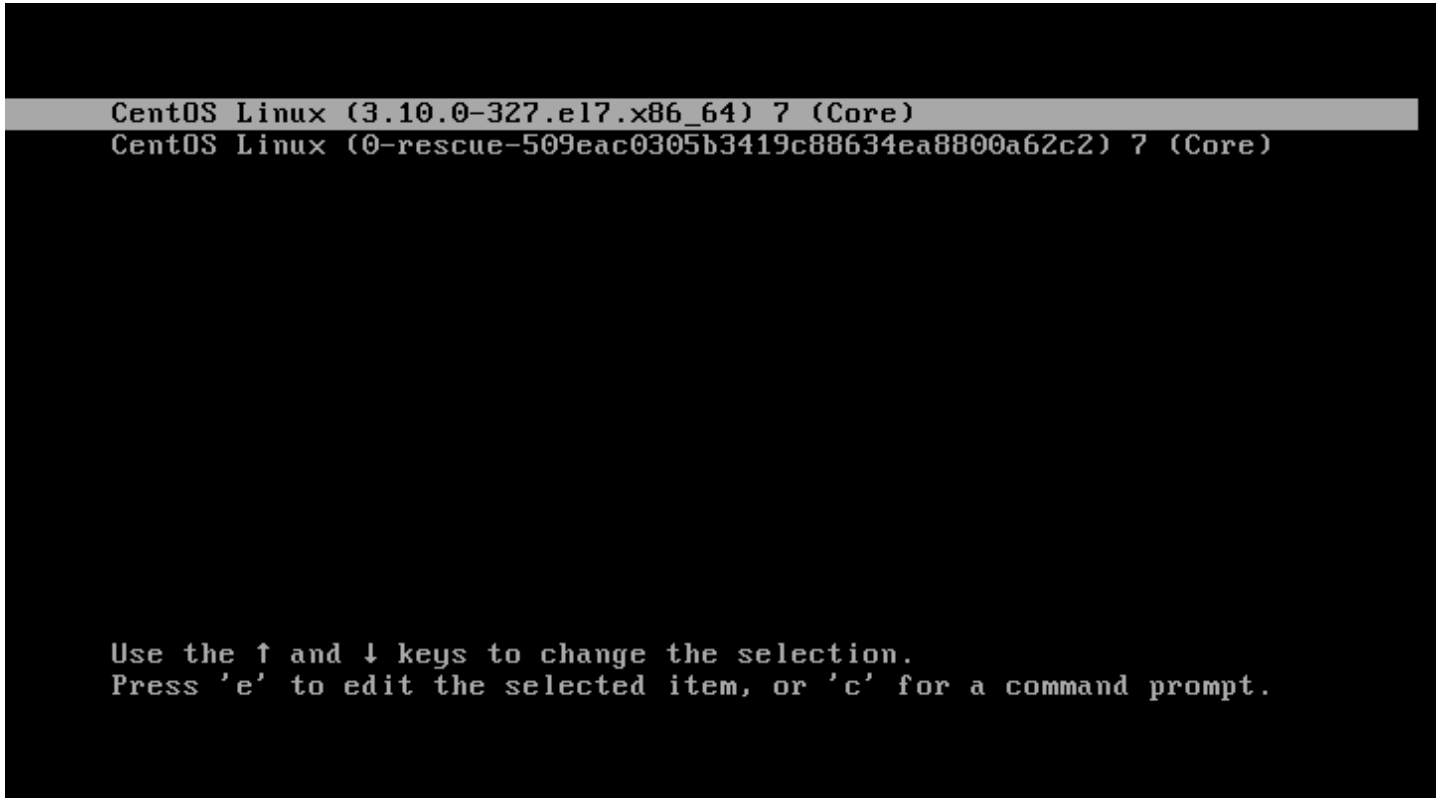
VM is available at: https://wwwjade.mpp.mpg.de

JADE software environment/VM

VM includes:

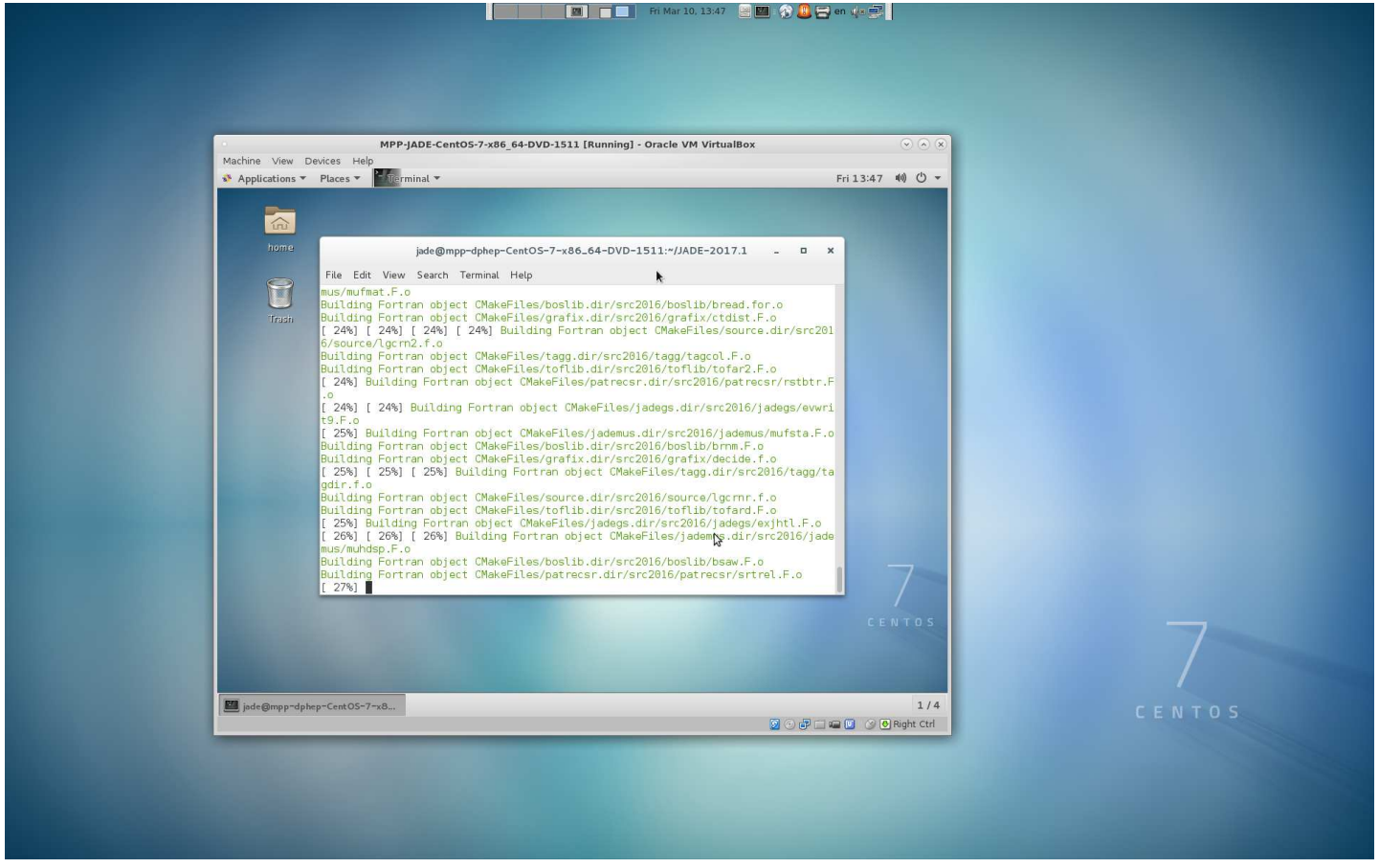
- JADE software: ROOT, MC simulation, event display, file catalogue, setup scripts etc. See a detailed list in backup slides.
- Modern MC generators, FastJet, CERNLIB, PAW and other popular and “not really” packages.
- Anything you will want to install...
- Agree access and download it.
- VirtualBox images are provided as well.

JADE software environment/VM



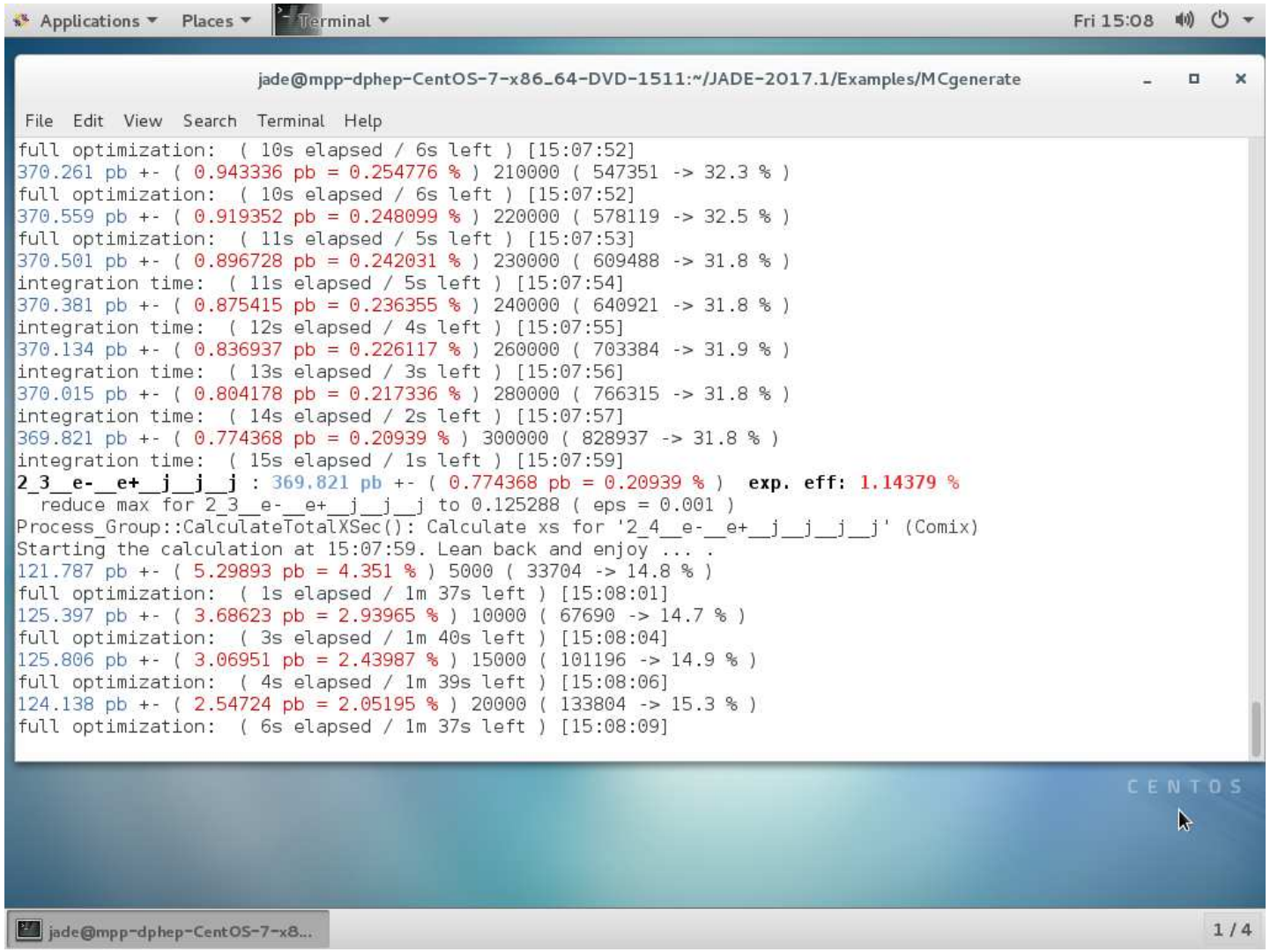
Generic start of VM.

JADE software environment/VM test



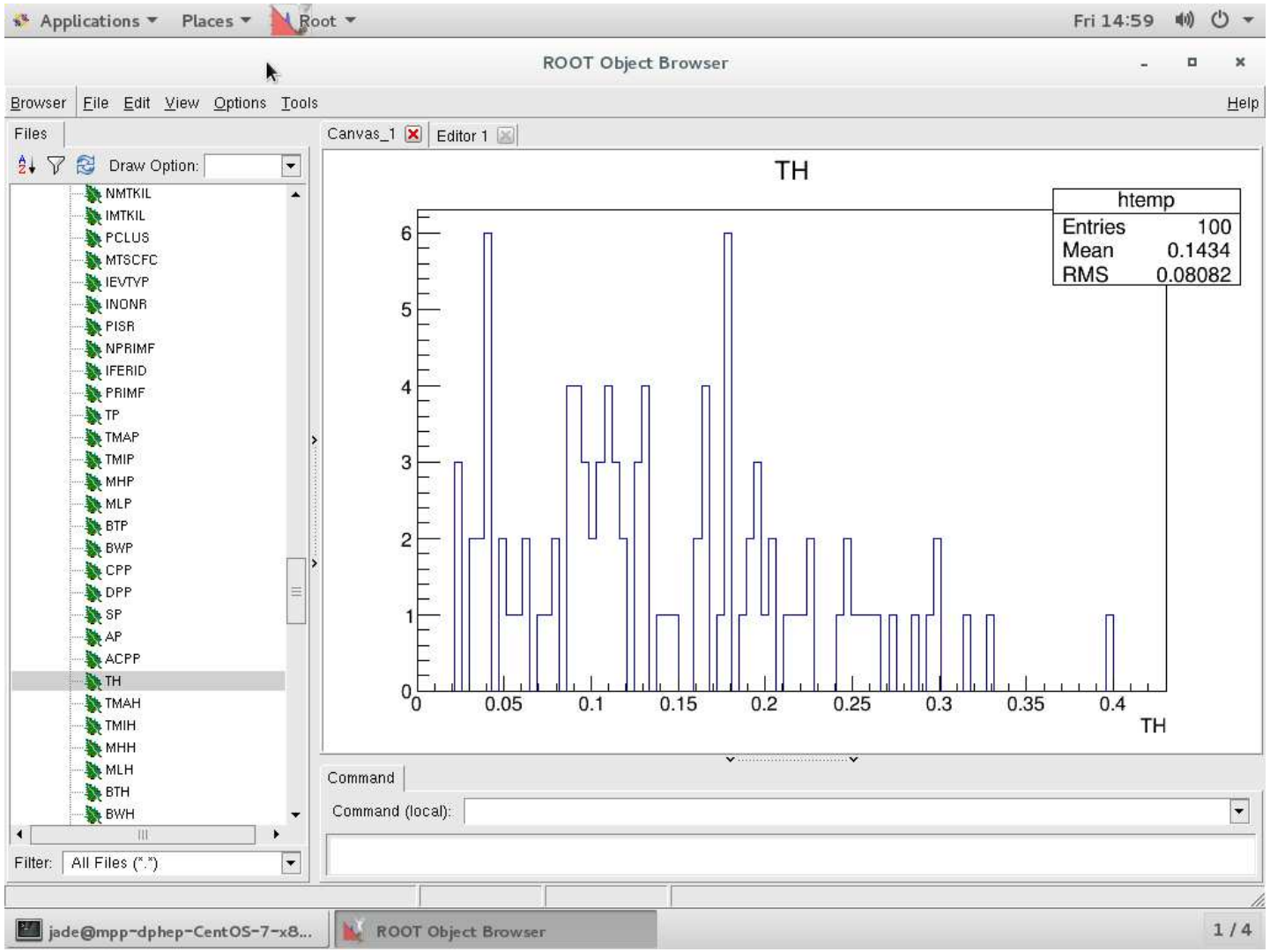
JADE software compilation in VM: takes about 4 minutes.

JADE software environment/VM test



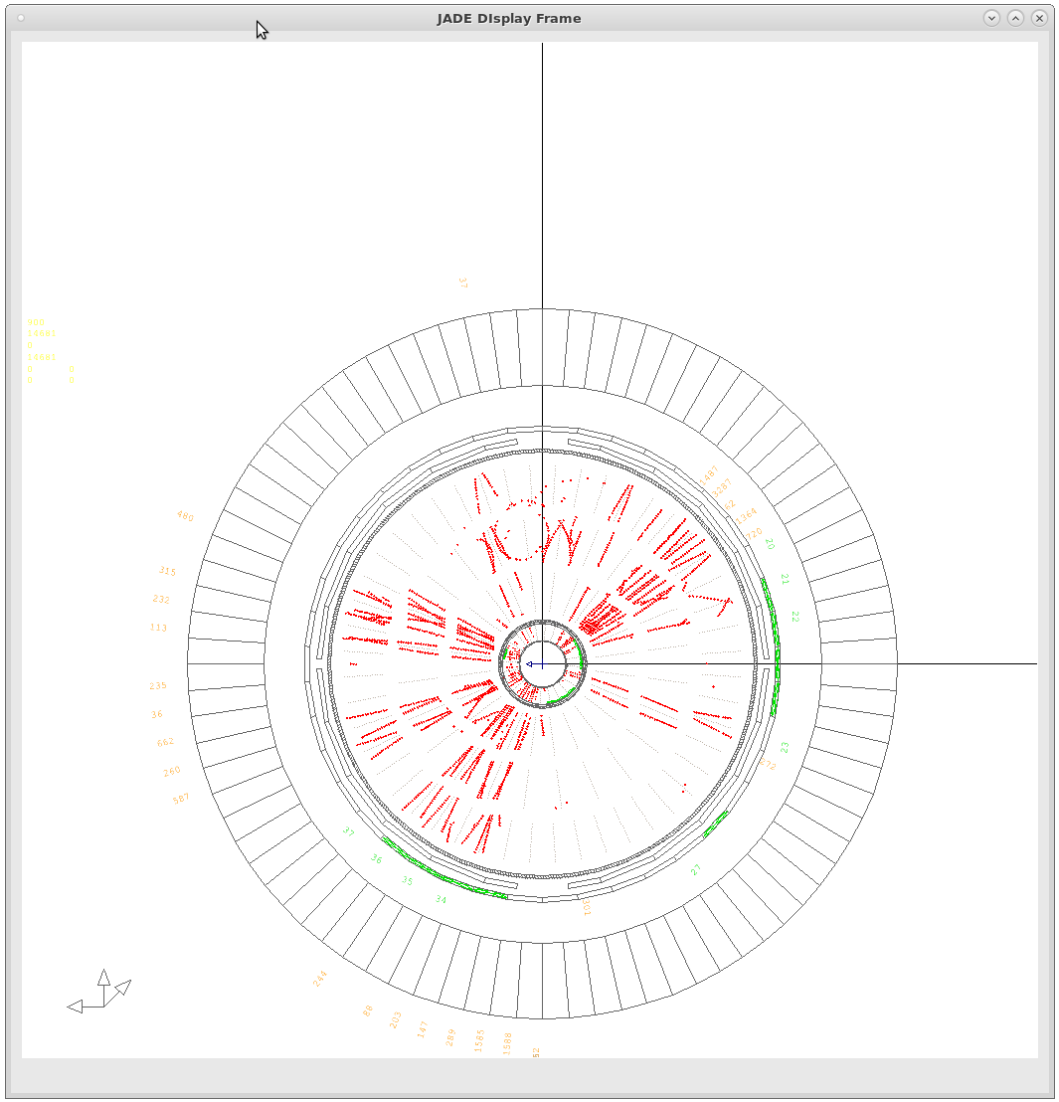
MC for JADE can be generated in VM or elsewhere. In this case SHERPA is running in VM.

JADE software environment/VM test



In the end of reconstruction a ROOT file with simple tree is delivered. Thrust distribution from the generated evens is shown.

JADE software environment/VM test



Test of the reconstructed MC files with event display. **Not all functionality is restored. To be fixed.**.

MPP Data Preservation summary

- Huge work was done since 1995!
- Many high quality, important physics results delivered.
- Data is accessible in MPCDF and software is **analysis-ready**.

Potential improvements are obvious: enable data reading, fixes to event display, improving documentation.

JADE Data Preservation applications

- Hadronisation effects.
- QCD with b-quarks and modern theory predictions, e.g. tests of future fully differential NNLO predictions for $e^+e^- \to \textit{hadrons}$ with heavy quarks.
- ...

Software

- Generic technologies for hardware and OS is an advantage.
- High code quality, clarity and stability is a key to success. Note: some routines are from 1974, i.e. 43 y.o. Few modern experiments can be sure that more than tiny fraction of their code will behave like that.
- Design of software is important.
- Good build system makes porting much easier
- Standardisation of I/O data formats is important.
- Documentation on simple things is important, i.e. meaning of arguments for utilities.

Data

- Compressing the data might look like a good solution, but it is bad. Saving 1Tb of disk space today mean lose of days to read the data back tomorrow. Do not compress your data!
- Keeping reference data/MC/results is important.
- Standardisation of I/O data formats is important.

What other experiments can learn from JADE DP: Some anecdotes

- one "calibration" file, with luminosities for each run and fill, was stored on a private account and therefore lost when DESY archive was cleaned up;
- Jan Olsson, when cleaning up his office in 1997, found an old ASCII-printout of the JADE luminosity file. Unfortunately, it was printed on green recycling paper - not suitable for scanning and OCR-ing. A secretary at Aachen re-typed it within 4 weeks. A checksum routine found (and recovered) only 4 typos.
- an old version of the original BOSlib 1979 version was found, on request at the Tokyo computer centre.
- Peter Bock, when cleaning out an old lab at the Physics Institute at Heidelberg University, found a few 9-track tapes containing original JADE MC files which were very valuable for validating results of first re-analyses in 1997

What other experiments can learn from JADE DP: Some anecdotes

- First port attempt in 2016 was done with AIX 4.3 machine used as router in MPP, but turns out VirtualBox is more useful.
- Many problems in software porting were solved in the industry: byte ordering problem in the modern gfortran, availability of Linux on PowerPCs, good FORTRAN build system (cmake) is available now, xlf compiler is available on Linux now.
- Googling the arxiver used fro the data, FPACK is a hard task.

List of custom/extra RPM packages on JADE VM

Some subpackages omitted.

blackhat-0.9.9-1.el7.centos.x86_64.rpm
blas-devel-3.4.2-5.el7.x86_64.rpm
cernlib-2006-36.el7.centos.i686.rpm
cernlib-devel-2006-36.el7.centos.i686.rpm
cernlib-packlib-gfortran-2006-36.el7.centos.i686.rpm
cernlib-static-2006-36.el7.centos.i686.rpm
cernlib-utils-2006-36.el7.centos.i686.rpm
epel-release-7-5.noarch.rpm
fastjet-3.1.2-1.el7.centos.x86_64.rpm
form-4.1-1.el7.centos.x86_64.rpm
geant321-2006-36.el7.centos.i686.rpm
gosam-2.0.3-1.el7.centos.x86_64.rpm
gosam-contrib-2.0-1.el7.centos.x86_64.rpm
Herwig-7.0.2-2.el7.centos.x86_64.rpm
lapack-devel-3.4.2-5.el7.x86_64.rpm
LHAPDF-6.1.6-6.el7.centos.x86_64.rpm
osg-ca-certs-1.55-1.el7.centos.noarch.rpm
osg-release-3.3-5.el7.centos.noarch.rpm
patchy-gfortran-2006-36.el7.centos.i686.rpm
paw-gfortran-2006-36.el7.centos.i686.rpm
PHOTOS-3.61-1.el7.centos.x86_64.rpm
pythia8-8.2.15-102.el7.centos.x86_64.rpm
qd-2.3.15-100.el7.centos.x86_64.rpm
qgraf-3.1.4-1.el7.centos.x86_64.rpm
root-5.34.36-1.el7.centos.x86_64.rpm
SHERPA-MC-2.2.0-3.el7.centos.x86_64.rpm
TAUOLA-1.1.5-1.el7.centos.x86_64.rpm
ThePEG-2.0.2-1.el7.centos.x86_64.rpm
vincia-1.2.02-1.el7.centos.x86_64.rpm
xbae-4.60.4-12.el7.centos.i686.rpm

All JADE software is packed as one rpm that is installed as is in /opt