

CHEP 2019 Highlights

M.Potekhin

NPPS Weekly

December 20th, 2019

General notes

- Great conference, strong showing by NPPS members
- Full CHEP 2019 Timetable: <https://indico.cern.ch/event/773049/timetable/#20191104>
- Proceedings submission is open...

Links

- Contribution list: <https://indico.cern.ch/event/773049/contributions/>
- An excellent overview presented by Paul Gessinger-Befurt at the recent ATLAS S&C week:
https://indico.cern.ch/event/823341/contributions/3652854/attachments/1954400/3246425/2019-12-02_chep19-summary_v8.pdf

- NPPS - many thanks to NPPS members who provided input (Johannes, Paul N, Paul L)
- Paul Nilsson's notes:
<https://docs.google.com/document/d/1zTZrpwCgF2tYq9384Q2kcHLhk72ouqWBX27iGbU3Gks/edit?usp=sharing>
- Johannes' notes:
https://docs.google.com/document/d/14vl0qqNZOhkRIPeTxRw1OzpLTSv_KLmXqn6Um9cwsfM/edit?usp=sharing
- Maxim's notes:
https://docs.google.com/document/d/1C1hH7udplrinJgeP9ucqE2d1nOFYnGFwK_fwS7TFCq4/edit?usp=sharing
- Paul Laycock's summary of Track 2 (Offline Computing):
<https://indico.cern.ch/event/773049/contributions/3581347/>

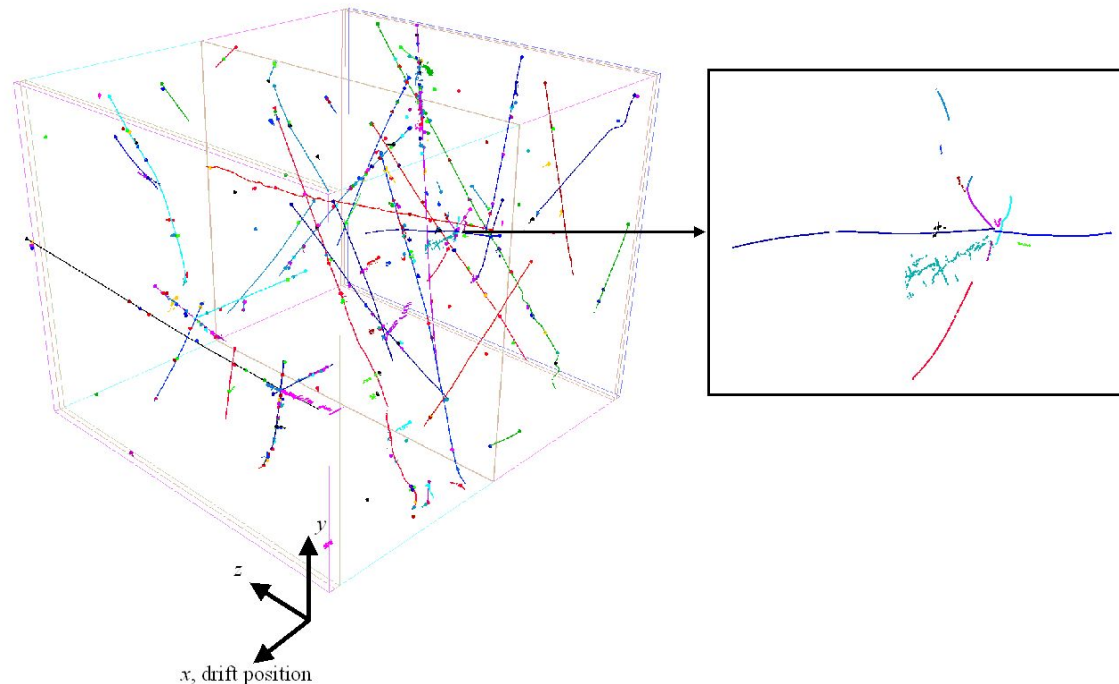
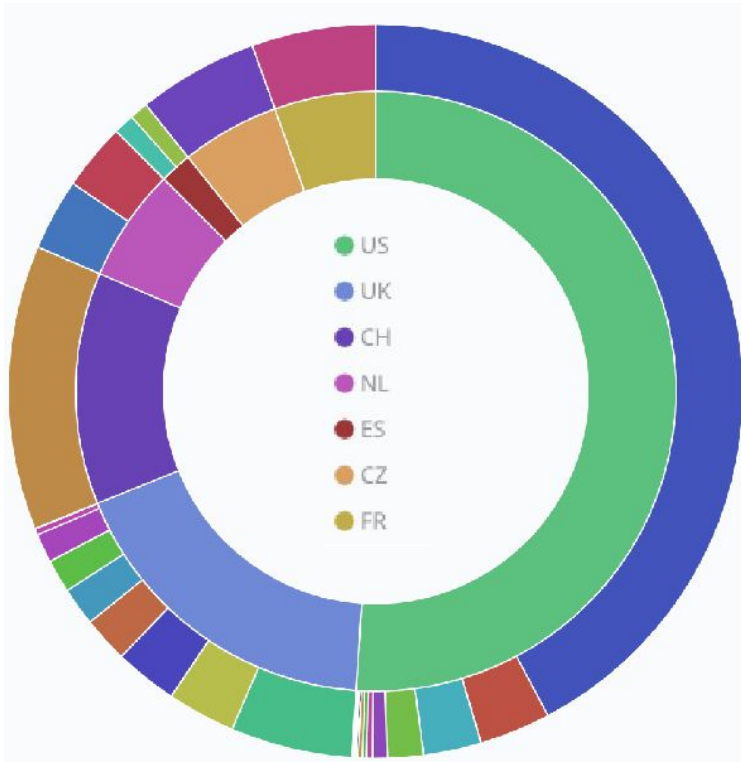
What's in these slides

- No need to precisely duplicate material in the links listed on the previous page
- However let's get inspiration from it and try to pick topics of common interest
 - based on the notes of NPPS members who attended the conference
 - disclaimer - a massive amount of material, impossible to cover all, even at the highlight level (cf. I have 24 items in my notes)
- Items are presented in no particular order



DUNE (Heidi Schellman)

- https://indico.cern.ch/event/773049/contributions/3581360/attachments/1937523/3211448/Schellman_CHEP2019_v4-wide.pdf
- A very broad overview of many DUNE topics from physics to the computing model, data volume, CPU budget etc; a good read for those unfamiliar with DUNE and protoDUNE
- However more interesting/challenging topics missing e.g. the DAQ architecture, details of reconstruction techniques



Rucio outside of ATLAS (Mario Lassnig)

- <https://indico.cern.ch/event/773049/contributions/3474416/>
- Belle II, CMS, DUNE, SKA, and LIGO... *and plenty of others...* community experience
- “Shared use of the global research infrastructures will become the norm”
- “Competing requests on a limited set of storage and network, data centres will be multi-experiment”

+feedback

- Easy to integrate into existing infrastructure and software
- Automation of dataflows
- Detailed monitoring
- Easy to contribute code/extensions

-feedback (being addressed)

- "Installation is only easy when you've done it before"
- "Configuration relies on too many ambiguous things"
- Documentation

...community-driven development

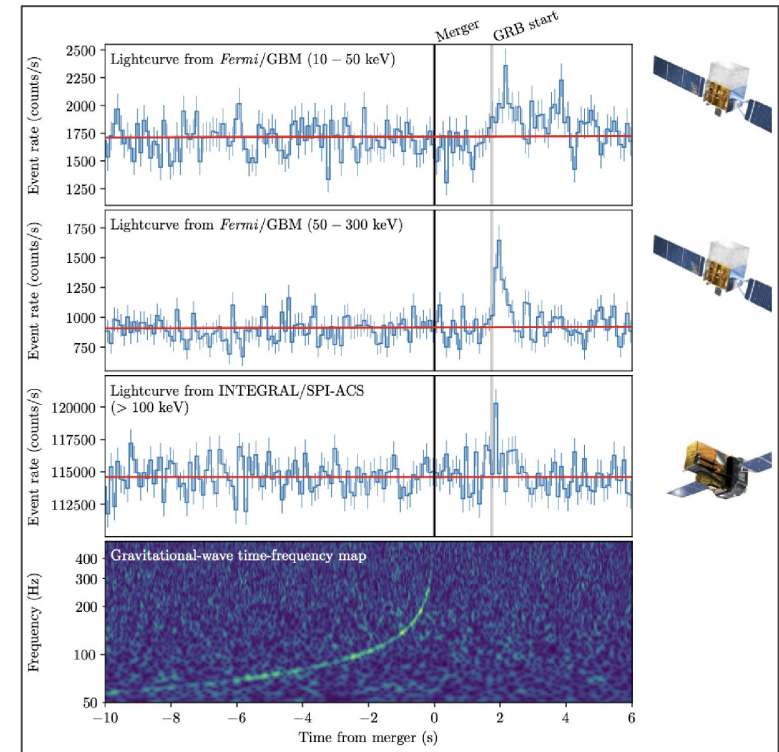
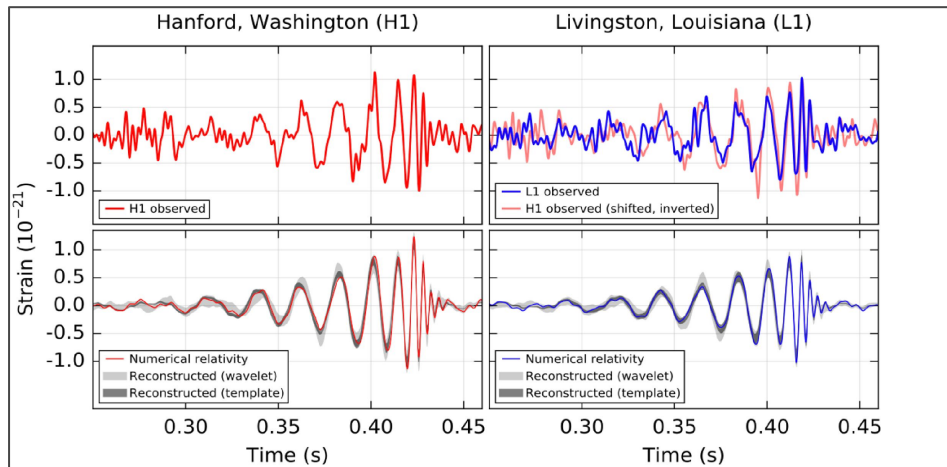
A growing community

Logos displayed include: Aeneas, LCLS, Belle II, Science & Technology Facilities Council, COMPASS, CERN, ARC, ATLAS EXPERIMENT, ESCAPE, AMS-02, XENON Dark Matter Project, CMS, DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT, EGI, EISCAT, VIRGO, cta, LSST, ICECUBE SOUTH POLE NEUTRINO OBSERVATORY, FTS, eXtreme DataCloud, Fermilab, SKA SQUARE KILOMETRE ARRAY, LIGO, Open Science Grid, and EUXDAT.

2019-11-04 Mario Lassnig :: Rucio :: CHEP'19 6

Gravitational Waves (Paul Lasky)

- <https://indico.cern.ch/event/773049/contributions/3581363/>
- The original black hole merge event
- Neutron star merge (most studied event in history)
- **Data-rich science** nowadays due to increased sensitivity of the instruments - new candidate events almost daily
- Neutron star merger provides a probe of the **nuclear matter equation of state** (oscillations)



Kubernetes

- Just two references here, out of ~10 talks at CHEP
- Observation - compared to previous CHEP there is more focus on integration of concrete experiments' frameworks and systems
- Using Kubernetes as an ATLAS computing site (F. Megino)
 - <https://indico.cern.ch/event/773049/contributions/3473808/>
 - CERN IT has integrated Kubernetes into their cloud
 - PanDA integration (cf. technical details of the pilot deployment)/lessons learned
- Science Box (Enrico Bocchi)
 - <https://indico.cern.ch/event/773049/contributions/3473819/>
 - investigating storage services at CERN running in containers
 - EOS deployed in containers on hybrid clouds

Deployment of containers on the diverse ATLAS infrastructure (A.Forti)

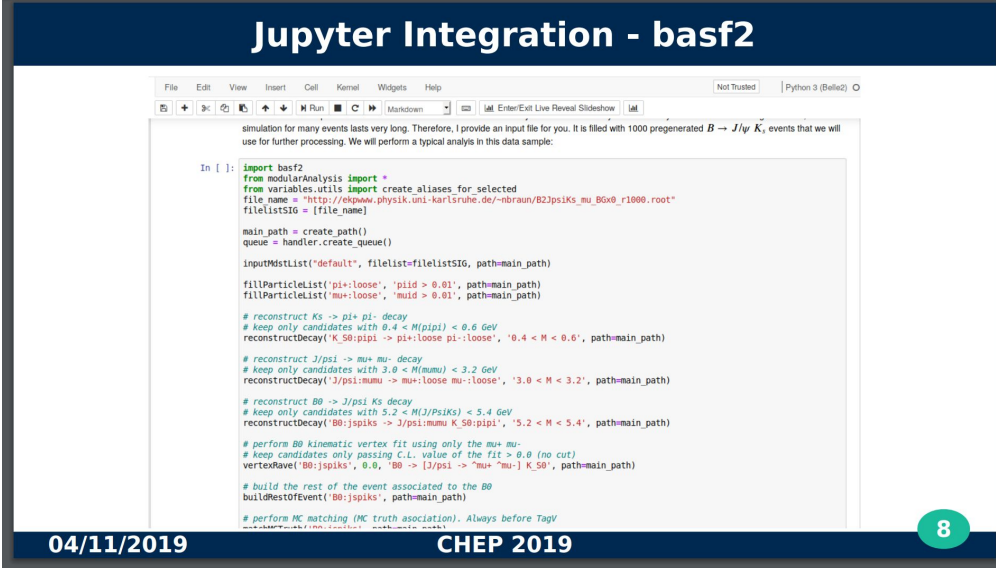
- <https://indico.cern.ch/event/773049/contributions/3473816/>
- Layered containers
- Container transform rather than a payload transform
 - Payload has all it needs in the container
 - Containers get downloaded from the registries
- Different types of images - CVMFS
 - “standard images”
 - user standalone images
 - “fat HPC images”

WLCG Authorisation; from X.509 to Tokens (Andrea Ceccanti)

- <https://indico.cern.ch/event/773049/contributions/3473383/>
- Free users from complexities of certificate management duties
- Token-based authorisation widely adopted in commercial services
- Problem: current grid middleware does not support token based authorisation
- OAuth2
- Rucio integration

Belle II (D.Dossett)

- A good general overview
- <https://indico.cern.ch/event/773049/contributions/3581361/>
- Overview, physics
- Basf2 - the framework
- Converters for the “old” Belle
- Jupyter integration
- **CDB supported by BNL**
- Calibration and alignment framework
- Scale challenges in Belle 2
- Common mDST format for reco and MC
- Details of data flow
- Raw data replication to BNL
- Rucio (also reference to Sergei’s talk)
- Dataset search in Belle 2 (GUI)
- Q/A:
 - Converters for the old Belle 2 - QA - pre-calibrated data only
 - how do they do online calibrations? Not too clear



The screenshot shows a Jupyter Notebook interface with a dark blue header titled "Jupyter Integration - basf2". The notebook content includes a text block explaining that a simulation for many events is long, so an input file with 1000 pregenerated $B \rightarrow J/\psi K_s$ events is provided. Below this is a code cell with the following Python code:

```
In [ ]: import basf2
from modularAnalysis import *
from variables.utils import create_aliases_for_selected
file_name = "https://indico.physik.uni-karlsruhe.de/~nbraun/B2jpsiKs_mu_B0x_r1000.root"
fileListSIG = [file_name]

main_path = create_path()
queue = handler.create_queue()

inputMdstList("default", fileList=fileListSIG, path=main_path)

fillParticleList('pi+loose', 'pid > 0.01', path=main_path)
fillParticleList('mu+loose', 'muid > 0.01', path=main_path)

# reconstruct Ks -> pi+ pi- decay
# keep only candidates with 0.4 < M(pipi) < 0.6 GeV
reconstructDecay('K_S0:pi+pi- -> pi+loose pi-loose', '0.4 < M < 0.6', path=main_path)

# reconstruct J/psi -> mu+ mu- decay
# keep only candidates with 3.0 < M(mu+mu-) < 3.2 GeV
reconstructDecay('J/psi:mu+mu- -> mu+loose mu-loose', '3.0 < M < 3.2', path=main_path)

# reconstruct B0 -> J/psi Ks decay
# keep only candidates with 5.2 < M(J/psiKs) < 5.4 GeV
reconstructDecay('B0:jpsiks -> J/psi:mu+mu- K_S0:pi+pi-', '5.2 < M < 5.4', path=main_path)

# perform B0 kinematic vertex fit using only the mu+ mu-
# keep candidates only passing C.L. value of the fit > 0.0 (no cut)
vertexRave('B0:jpsiks', 0.0, 'B0 -> J/psi -> mu+ mu- K_S0', path=main_path)

# build the rest of the event associated to the B0
buildMdstOfEvent('B0:jpsiks', path=main_path)

# perform MC matching (MC truth association). Always before TagV
```

At the bottom of the notebook, there is a footer with the date "04/11/2019" and the event name "CHEP 2019". A small green circle with the number "8" is located in the bottom right corner of the notebook frame.

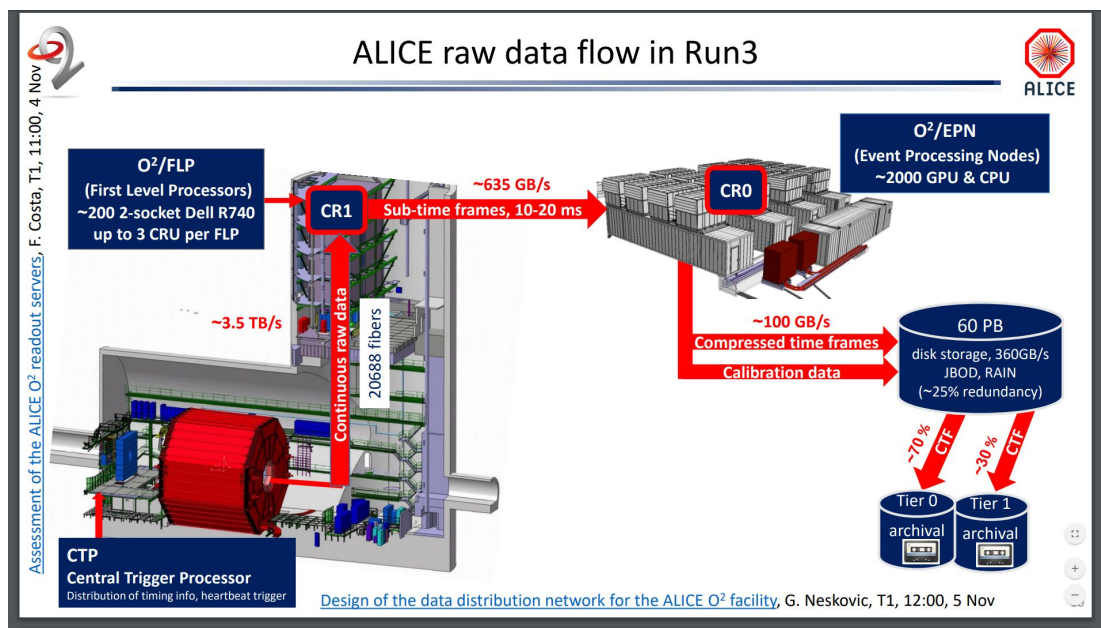
Distributed data management in Belle II

(S.Padolski)

- Summary of current work on DDM
- DDM Missing functionality (select items)
 - Based on the outdated File Catalog (LFC)
 - No data retention time principle
 - No automatic SE occupancy rebalancing
 - Monitoring
 - etc etc etc
 - ...the team decided to evaluate Rucio
- Keeping the pre-existing API for the data management
- Staged approach
- Monitoring challenges/updates
- Q/A
 - What's the plan for migration? Intermediate database?

ALICE readout and data reduction for Run3 (R.Shahoyan)

- <https://indico.cern.ch/event/773049/contributions/3581368/>
- Cardinal reworking for Run 3 and 4
 - aim to boost the number of collected events by a factor of 100
 - requires combination of data reduction and compression by factor >35
- Latency of the detector itself is also a limitation
- First level processor - 200 CPUs
- 3.5GB/s -> 600GB/s reduction
- Event processing nodes - 2000GPU and CPU
- Continuous data stream
- Heartbeats
- ZMQ
- FairMQ
- FairRoot
- O2 framework
 - Based on ALFA platform
 - derived from FairRoot
- DPL - data processing layers
- TPC tracking is redone
 - Cellular Automaton



Data Analysis in ALICE Run3 framework (Giulio Eulisse)

- <https://indico.cern.ch/event/773049/contributions/3476164/>
- Reduce costs of data read: Runs 1-2: Trains (collection of “wagons”)
 - data read once per train
 - Run 3 - improve trains
- Validated set of wagons runs on the Grid
- Can only do AOD in run 3 (too much data!)
- Goal: each Analysis Facility go through the equivalent of 5PB of AODs every 12 hours (~100GB/s)
- Recompute some data as opposed to preservation - to save space
- Personal comment:
 - the large volume of data dictates fundamental change of architecture
- Column oriented analysis
- Shared memory
- Crucial - message passing, subscription
- Plan to use Apache Arrow
- Complex material, please see slides...

ALICE at CHEP



ALICE Upgrade presentations at CHEP 2019



November 4:

[ALFA: A framework for building distributed applications](#), M. Al-Turany, T5, 11:30

[Jiskefet, a bookkeeping application for ALICE](#), M.Teitsma, T4, 11:45

[AliECS: a New Experiment Control System for the ALICE Experiment](#), T. Mrnjavac, T1, 14:00

[The ALICE data quality control system](#), P. Konopka, T1, 15:15

November 5:

[Assessment of the ALICE O² readout servers](#), F. Costa, T1, 11:00

[A VecGeom navigator plugin for Geant4](#), S. Wenzel, T2, 11:30

[Design of the data distribution network for the ALICE Online-Offline \(O²\) facility](#), G. Neskovic, T1, 12:00

[Data Analysis using ALICE Run3 Framework](#), G.Eulisse, T6, 11:45

[System simulations for the ALICE ITS detector upgrade](#), S. Nesbo, T2, 12:15

[GPU-based reconstruction and data compression at ALICE during LHC Run3](#), D.Rohr, TX, 14:15

[Running synchronous detector reconstruction in ALICE using declarative workflows](#), M. Richter, TX, 16:30

[Running ALICE Grid Jobs in Containers - A new approach to job execution for the next generation ALICE](#), M.Melnik, T7, 17:45

[Using multiple engines in the Virtual Monte Carlo package](#), B.Volkel, T2, 17:45

Posters:

[Fast and Efficient Entropy Compression of ALICE Data using ANS Coding](#), M.Lettrich, T1

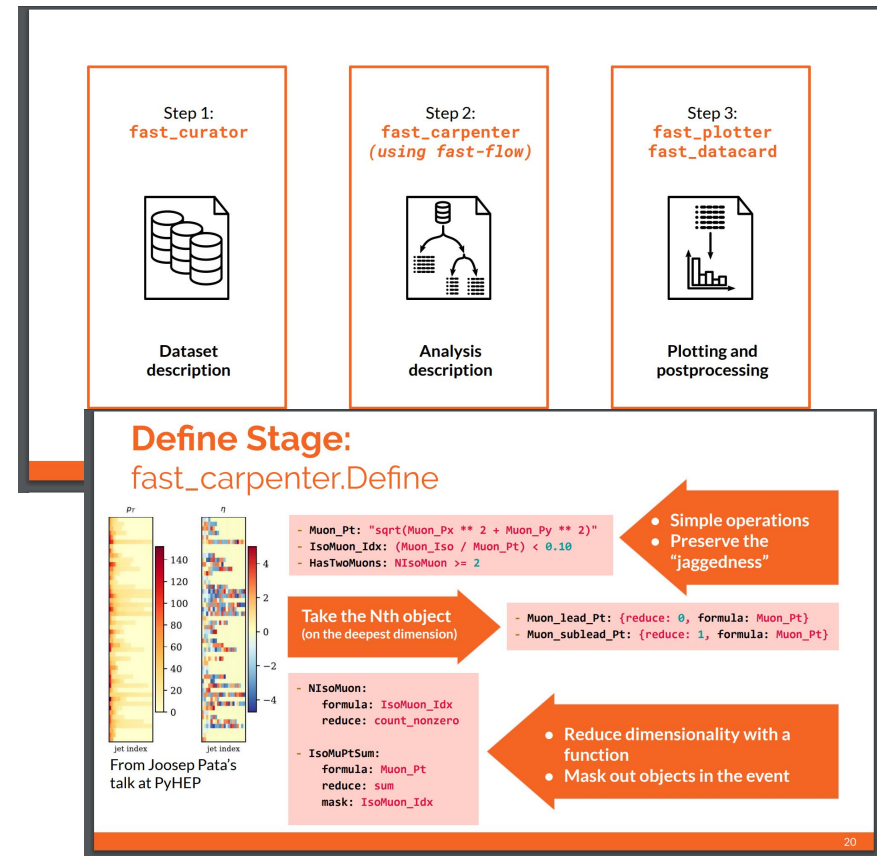
[Space point calibration of the ALICE TPC with track residuals](#), O. Schmidt, T1

[The evolution of the ALICE O² monitoring system](#), A. Wegrzynek, T1

33

The F.A.S.T. toolset (B.Krikler)

- <https://indico.cern.ch/event/773049/contributions/3476191/>
- “Using YAML to make tables out of trees”
 - an example of declarative approach to analysis
- “Your analysis repository is your analysis”
- Declarative languages the user says WHAT, the interpretation decides HOW
- Trees -> DataFrames
- User gives up flow control:
 - Cannot do
 - Loop over each event
 - add this to that if something is true
 - etc
- Allows:
 - More concise description
 - Fewer bugs
 - Easier to reproduce and share
 - Optimisation behind the scenes
- Ingredients
 - numpy/pandas/NumExpr/awkwardArray
 - See slides for YAML examples
- Integration with RUCIO TBD



Scikit-HEP kit (Eduardo Rodrigues)

- <https://indico.cern.ch/event/773049/contributions/3476182/>
- “Scikit-HEP is a community-driven and community-oriented project with the goal of providing an ecosystem for particle physics data analysis in Python”
- BNL participation
- Once again,
 - `uproot`
 - `awkward array`
- New Python bindings for the performant C++14 Boost::Histogram library
- “iminuit” (fitting)
- “Particle” package
- “scikit-stats” (statistics)
- HepMC3 (event record for MC generators)
- Visualisation
- FAST HEP affiliated (see above)
- We need to maintain a forward looking posture toward these developments

Analysis Declarative Languages for the HL-LHC (G.Watts)

- <https://indico.cern.ch/event/773049/contributions/3476174/>
- New analysis tools not necessarily based on ROOT
- Jupyter, numpy (declarative features)
 - immediate execution
 - DASK solves this problem <https://dask.org/>
- Belle II Decay Reconstruction Declarative Language
- Coffea
 - uproot
 - awkward array

Workshops

Two large workshops:

- ADL, CutLang, lhada2rivit, LINQ, Yaml as an ADL, NAIL, TTreeFormual, AEACUS and RHADAMANTUS

Here at CHEP:

- [F.A.S.T. \(yaml\)](#) – Monday, Nov 4th
- [COFFEA - Columnar Object Framework For Effective Analysis](#) – This session
- [A Functional Declarative Analysis Language in Python](#) – Tuesday Poster
- [HEP Data Query Challenges](#) – Thursday Poster
- [Striped Data Analysis Framework](#) – Thursday Poster
- [See my talk at the WLCG/HSF workshop on Analysis Eco-systems challenges](#)

The screenshot shows the Indico page for the 'HEP analysis ecosystem workshop' held from May 22-24, 2017, in Amsterdam. The page includes a search bar, a navigation menu with options like 'Overview', 'Timetable', and 'Contribution List', and a main content area with text describing the workshop's focus on the ROOT ecosystem and the evolution of analysis tools.

The screenshot shows the Indico page for 'Analysis Description Languages for the LHC' held from May 6-8, 2019, at Fermilab, Wilson Hall. It features a search bar and the Indico logo.



CERN analysis preservation framework: FAIR research data services for LHC experiments (P.Fokianos)

- <https://indico.cern.ch/event/773049/contributions/3476165/>
- FAIR (Findable, Accessible, Interoperable and Reusable)
- Stated goal (ambitious!):
 - Make the analysis components (metadata, files, tables, plots, likelihoods, wikis, etc) easily reusable - ex. in workflow engines, scripts, publication writing tools, push to other services (ex. HEPData, Inspire, Zenodo)
- Persistent data identifiers (i.e. “findable”)
- CERN SIS
- CAP
 - references to Web tools, screenshots
- Preservation <-> reusability
- Personal comment: seems like a lot of effort is still required to follow the process
- Invenio
- CAP: JSON and YAML uploader, Yadage
- REANA preserves workflows
- ...wealth of information in this presentation

The CMS approach to analysis preservation (Lara Lloret Iglesias)

- <https://indico.cern.ch/event/773049/contributions/3476181/>
- The CMS Collaboration is making an effort to make analyses reinterpretable
- Agencies demand *data preservation* plans from experiments
- Analysis details for each publication are preserved
- “Datacards” are preserved
- Links to Wikis
- This is internal, ideally all is to be migrated to the CERN analysis preservation portal
 - cf. the challenges in PHENIX
- <https://analysispreservation.cern.ch/>
- Import from CADI (CMS analysis management framework) to CAP
 - reference to datasets, workflow description, statistical treatment etc
 - CAP screenshots
- <http://www.reanahub.io/>
- Four people (!) are working on the REANA integration
 - more or less confirmed during the CERN workshop in November
- “The next big step will be preserving the implementation (CAP-Reana integration)”