CERN EP R&D Programme

Software Work Package

Graeme Stewart

CERN EP R&D Programme

- Wide ranging programme of R&D into fundamental detector technology for LS3 and beyond
  - Thus, primarily after the start of HL-LHC in Run4
  - Also targeting experiments at future colliders (FCC, CEPC, ILC)
- Challenges to be faced at FCCee
  - Very high tracking precision (few μm) and low material budgets
  - Highly granular calorimeters (particle flow reconstruction)
  - Timing at 1-20ns
- Challenges to be faced at FCChh
  - Pileup 1000 (30 x 10^{34} instantaneous luminosity)
  - 10-100 more radiation than HL-LHC
  - Track timing of 5-10ps
  - Improved magnet technology

EP R&D Phase 1 2020-2024

Overview of Work

WP1 - Silicon Detectors

WP2 - Gas Based Detectors

WP3 - Calorimetry and Light-based Detectors

WP4 - Detector Mechanics

WP5 - Integrated Circuit Technology

WP6 - High-Speed Links

WP7 - Software

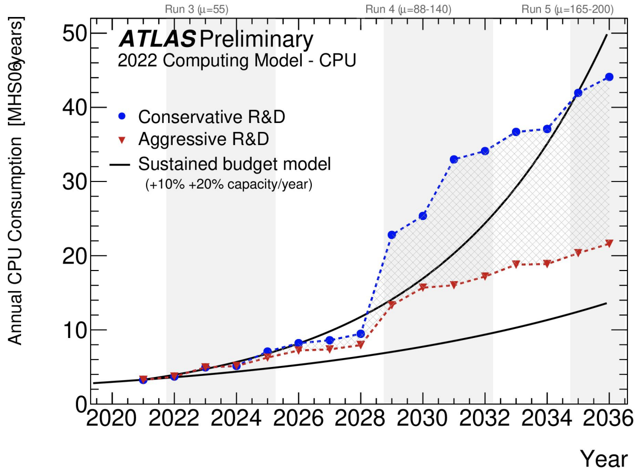
WP8 - Magnet Technologies

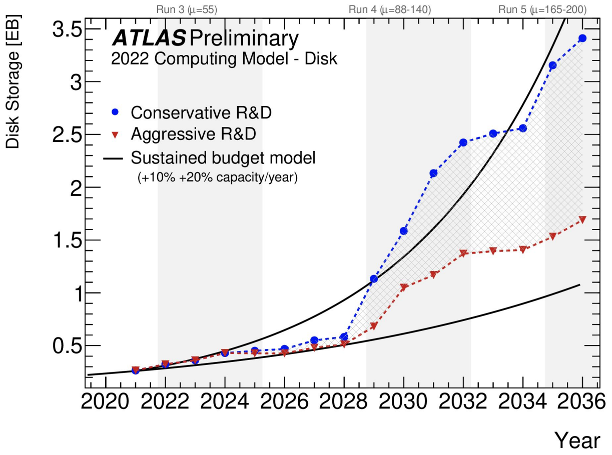
Wide ranging activities - well aligned with ECFA Detector R&D Roadmap

With strong cooperation with other groups and institutes

Software Work Package

- Software R&D needed to tackle significant challenges for future experiments
  - Differs from hardware projects in timescale
  - Therefore Run4 HL-LHC upgrades for ATLAS and CMS are also in scope
- Want to improve physics performance
  - Especially at high pile-up
- Control of resources is also a major factor
  - Projections for optimistic R&D programmes still not in alignment with hardware cost evolution
  - Efficient software for physics is well aligned with energy efficiency
- However, for far future experiments we also need software solutions today for design and performance studies





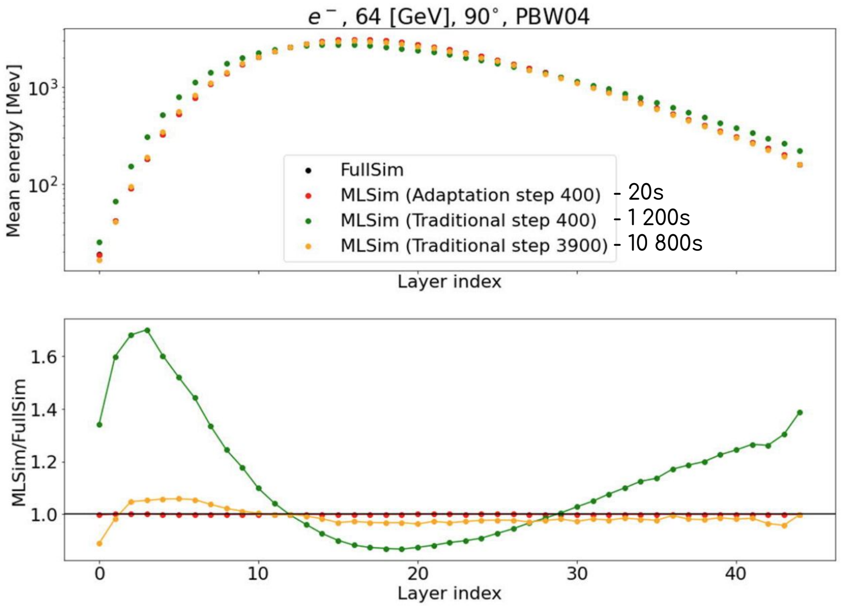
Work Package Overview

- Faster Simulation
  - Machine Learning models for fast simulation
  - Starting work now on GPU particle transport in AdePT
- Reconstruction
  - Tracking in high pileup, working with the Acts project
  - High granularity calorimetry at high pileup
- Analysis
  - New ROOT I/O format, RNTuple
  - Scale-out for analysis via Distributed RDataFrame
- Turnkey Software
  - Flexible software stack for future experiments at new facilities (FCC, ILC, CLIC, CEPC, EIC)
  - Data model generation for modern hardware, PODIO
  - Starting work now in heterogeneous framework support

EP R&D Website has contact information for various tasks

Faster Simulation Highlights I

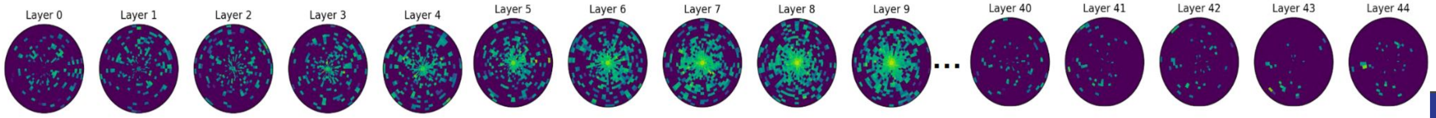
- Improved integration of machine learning simulation into Geant4
  - Par04 example, showing how to use ONNX/PyTorch model integration
  - This reduced memory consumption for ATLAS FastCaloSim models by 66x (!) over LWTNN
- Development of meta-learning models for calorimeters
  - Greatly improve training times for new calorimeter layouts by starting with an already trained model
  - This is a very useful approach for the design phase of new calorimeters



Faster Simulation Highlights II

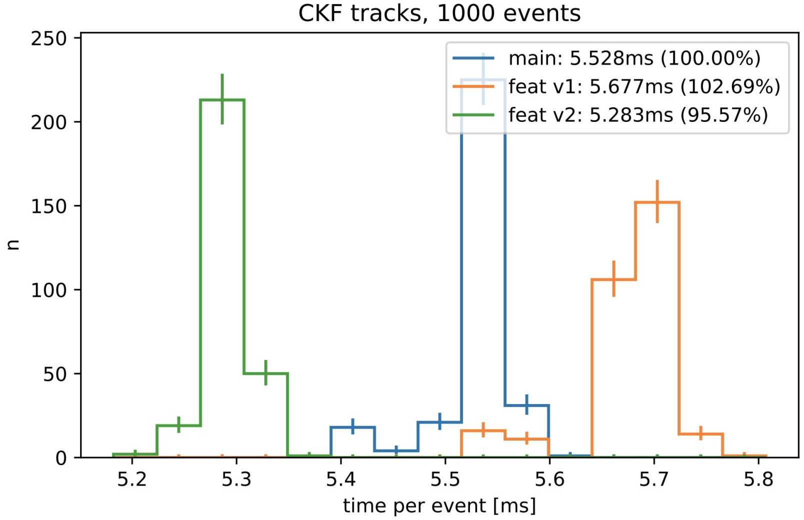
- Publication of datasets as part of the CaloChallenge (cf. TrackML)
  - 3 Datasets of increasing complexity/granularity
  - Workshop in Rome to assess the results in May
  - Some promising results shown at the ML4jets workshop last year
    - E.g., normalising flow models
- Starting investigation into Foundational and Transformer Models now
  - Foundation models are large and adaptable to many tasks
  - Transformer models work on sequences and handle long range dependencies (attention mechanism)
    - They are a hot topic now (ask ChatGPT!)
    - See shower development as a sequence
  - Models investigated that use a classifier, VQ-VAE to generate "tokens"
    - These are automatically learned features
  - Then apply a transformer to the initial set of tokens to generate new ones, the next step in the sequence

View a shower as a sequence



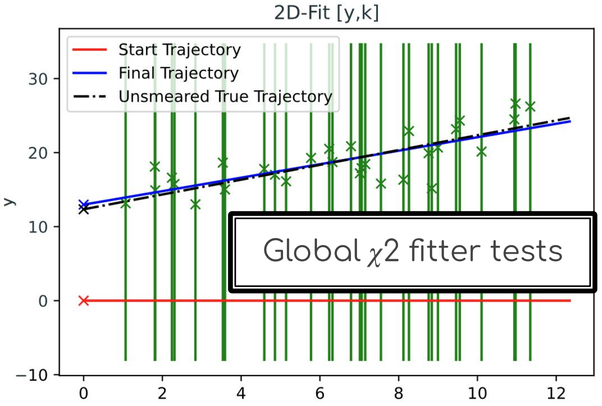
Reconstruction Highlights - Tracking I

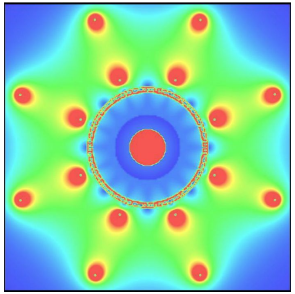
- Track reconstruction work in EP R&D takes place in the context of Acts
  - Experiment independent track reconstruction toolkit, derived from ATLAS tracking
  - Now in production for ATLAS and sPHENIX
- Improvements in the core of the project
  - Python testing framework, reducing build time memory consumption, DD4hep integration, etc.
- Latest updates have been on reworking the Acts Track EDM to allow direct use of experiment code
  - Achieved this without loss of performance with ATLAS xAOD



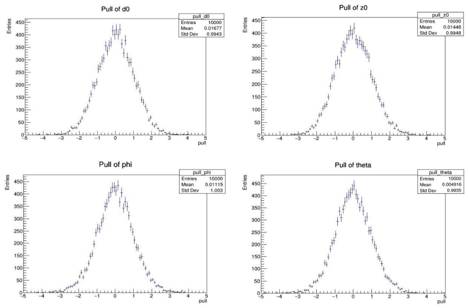
Reconstruction Highlights - Tracking II

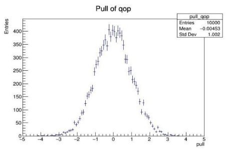
- New robust fitters:
  - Kalman filter, Gaussian sum filter, Global χ^2 fitter
- GPU tracking chain
  - A lot of R&D on fundamental tools to support an application as complex as tracking on GPUs
    - vecmem - low level uniform ergonomic data representation for CPUs and GPUs
    - detray - detector geometry on GPU, avoiding polymorphism and virtual calls
    - algebra-plugins - generalised linear algebra library (wrap low level Eigen, Vc, SMatrix, Fastor (new!), etc)
    - cofvie - vector field interpolation library for GPUs
  - Traccc - put all of the components together into a tracking demonstrator





ATLAS magnetic field on GPU

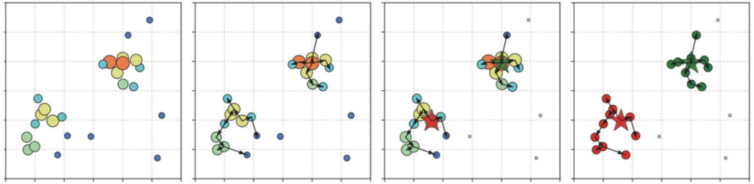




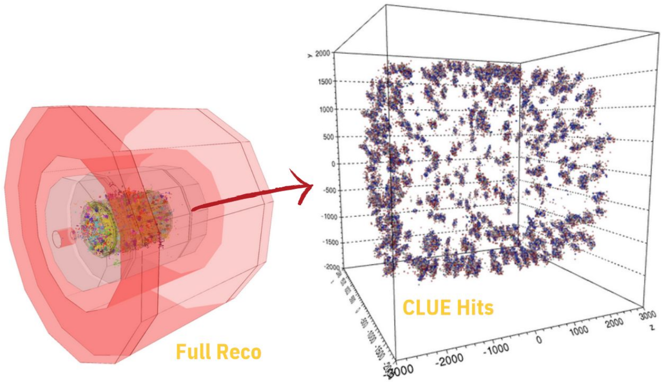
Traccc pulls (i.e., yes it works!)

Reconstruction Highlights - Calorimetry I

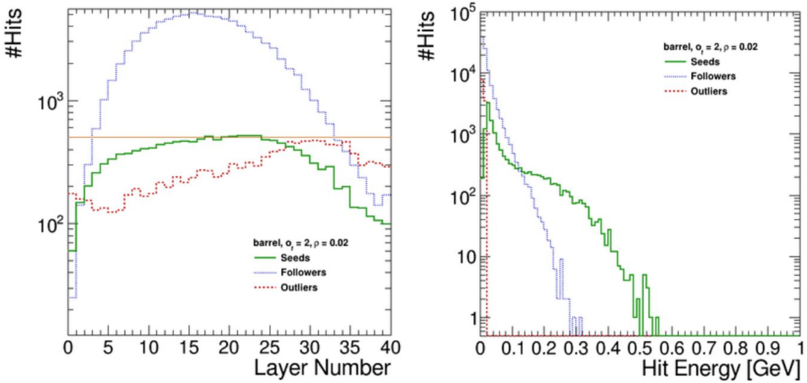
- Reconstruction using a novel density based algorithm, CLUstering of Energy (CLUE)
  - Close collaboration with CMS on HGCAL reconstruction



- Develop k4CLUE as a standalone package that can be integrated for other experiments
  - In particular the Key4hep stack (part of the Turnkey software)
  - Used to reconstruct events using the CLIC detector concept, which has high granularity calorimeter

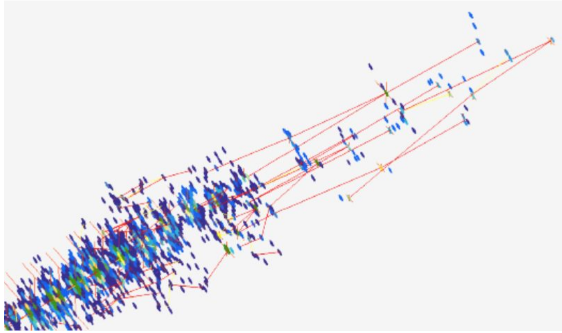


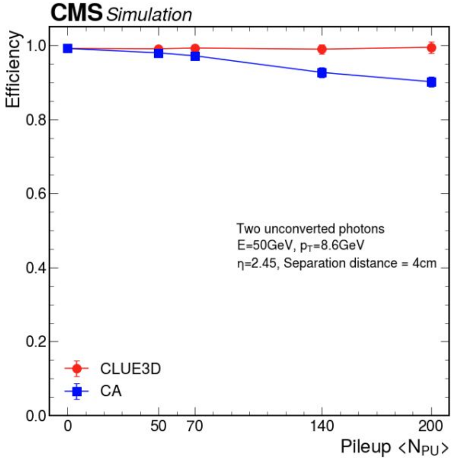
Sample: 1 event of 500 gammas generated with 10 GeV energy in the CLIC detector.

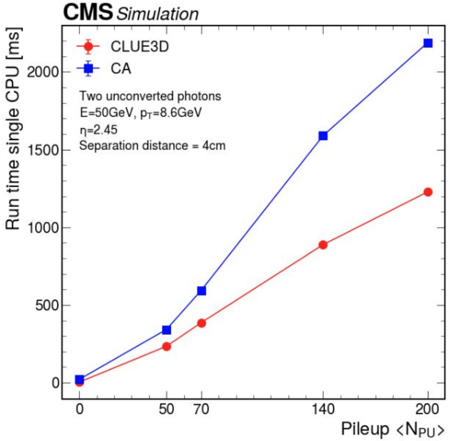


Reconstruction Highlights - Calorimetry II

- CLUE3D algorithm takes the output of layer-by-layer clustering and assembles 3D clusters in (z, η, φ)
  - Performance tuned significantly in 2022
  - Now outperforming cellular automata clustering for CMS
    - Adopted as the baseline for HGCAL reconstruction
  - CLUE3D is a parallelisable and GPU friendly algorithm

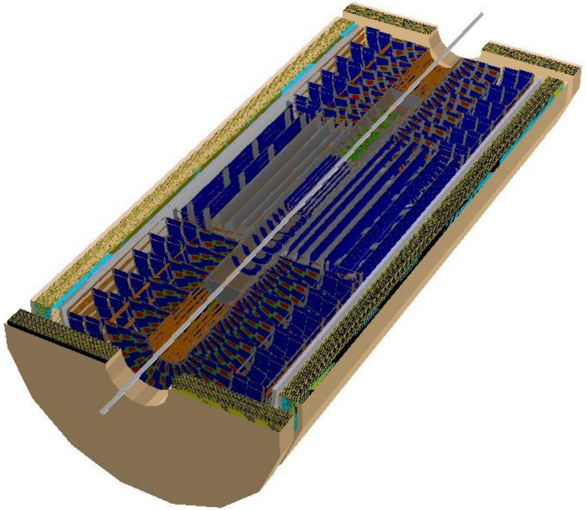






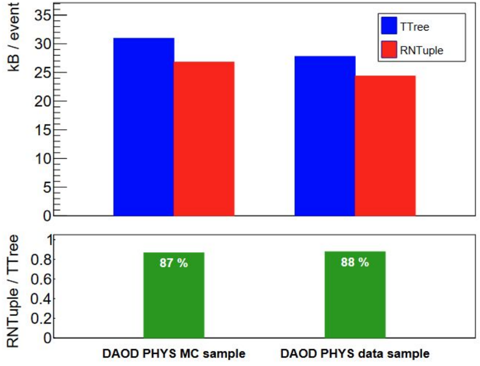
Open Data Detector

- Open Data Detector is a development of the TrackML detector
  - Pseudo HL-LHC detector
  - Much more realistic passive material map
- Began with tracker
  - Now added an EM calorimeter
- Will be very useful to provide benchmark datasets that are realistic, but experiment neutral
  - ML simulation developments
  - Tracking algorithm development and comparisons
- HCAL and muon system would be great to add

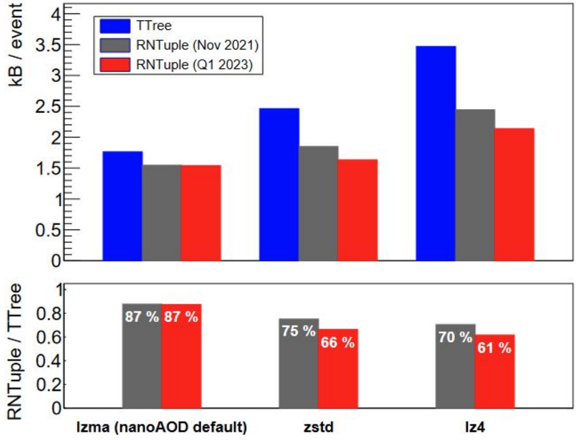


Faster Analysis - Highlights I

- Core building block of HEP data taking is ROOT I/O, but current TTree I/O is ageing
- New **RNTuple** I/O aims to offer state-of-the-art I/O performance and hardware support
  - Key features needed in HEP, e.g., flexible event structures, deeply nested data models, automatic streamers, schema evolution
- Recent work with the experiments
- Ensuring RNTuple supports the full set of needed features, e.g., late model extension in ATLAS xAOD
  - Optimisation of data compressibility via pre-processing fundamental types
  - Particularly good results for faster compression algorithms like zstd



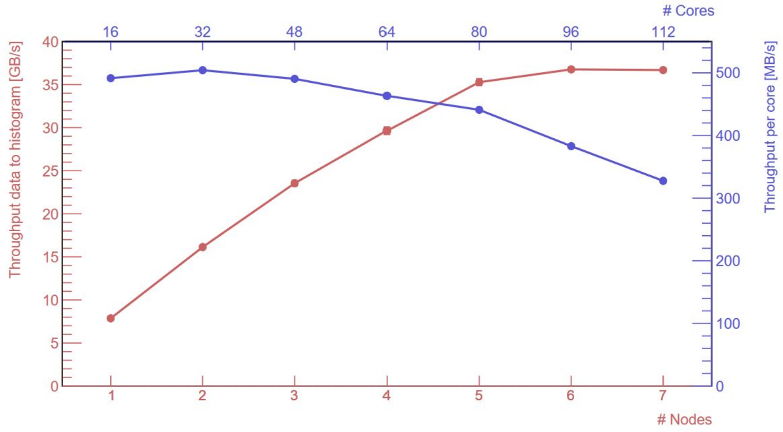
ATLAS xAOD in RNTuple



Compression tests for CMS nanoAOD

Faster Analysis - Highlights II

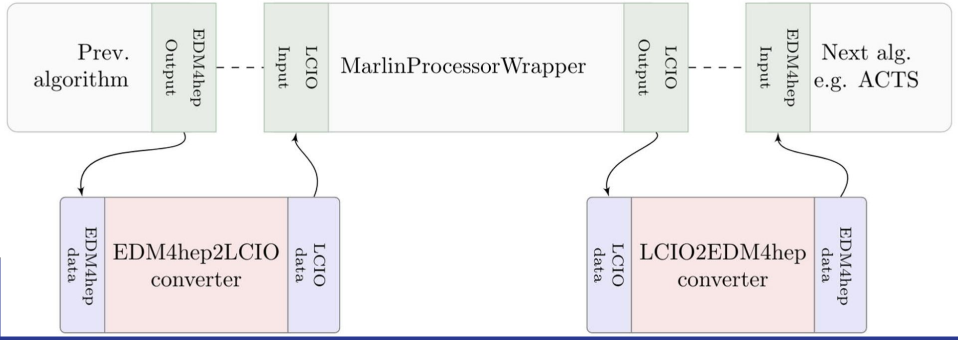
- Object stores increasingly used in HPCs and other data centres
  - Better scaling than file systems
- RNTuple object store performance has been improved
  - Target of 500MB/s/core and 10GB/s/node looks achievable
  - N.B. POSIX files still supported!
- Now planning to test RNTuple at large scale as part of preparations for HL-LHC
  - Strategy endorsed by the LHCC



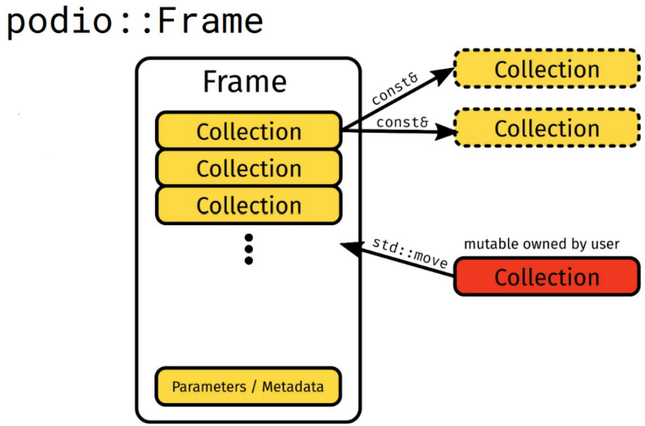
First scaling results on a 7 node DAOS cluster with 100GBit interconnects, using a 1TB data set of LHCb final-stage ntuples. Throughput is measured in uncompressed bytes and includes histogramming.

Turnkey Software - Highlights I

- Key4hep software stack encapsulates all software needed for building an experiment's offline software
  - cf. LCG releases from SFT, or CMS externals
- Key4hep is now adopted by CLIC, FCC, CEPC, ILD and (partly) by EIC
  - Used in the ECFA study on physics and experiments at e+e- Higgs/EW/Top Factories
- Adapting iLCsoft algorithms can be done with a special wrapper that converts from EDM4hep (Key4hep) to/from LCIO (iLCsoft)
  - This turns out to be very efficient, although we also have the first examples of algorithm rewrites from iLCsoft to Gaudi
  - Updated wrapper in WIP



Turnkey Software - Highlights II



- Modern hardware places demands on data layout
  - Need to support CPUs and GPUs at minimum
- Plain Old Data IO (PODIO) solves this problem by being a data model generator
  - YAML description of data objects and relations
  - C++ and Python code then generated automatically, along with ROOT and LCIO bindings
- Recently the concept of *Frames* was added to PODIO
  - Container for Run, LumiBlock, Readout Frame, etc.
  - Helps to define and implement thread safety
- Essential component of schema evolution, which is developed and now being finalised
- RNTuple support is now also very close

EP R&D Phase I and Phase II

- Original funding from 2020 to 2024
  - However, this was always intended to be an ongoing programme
  - Especially important now that we have significant alignment with the ECFA Detector R&D Roadmap
- Now making the case for EP R&D II, from 2024 to 2028
  - Evolve the programme, learning from the current results and bringing in new lines of R&D
  - Proposal made to the CERN directorate

Software Phase II Plans

- We continue and evolve many lines of development
  - Generative ML Simulation - Continuing to develop new models and strengthening connection to the experiments
  - Reconstruction tasks (tracking and calorimetry) - Refocus on *timing* as being a first class measurement in new detectors
  - Analysis - Testing of RNTuple at scale (analysis facilities!) and refining and finalising the format
- Propose additional lines
  - Simulation - particle tracking on GPUs (AdePT prototype and demonstrator)
  - Frameworks - rethinking frameworks and developing multi-node and heterogeneous computing support in Gaudi
  - Green computing - potential contribution of alternative languages (Rust, Julia) to software efficiency
  - HEP Core Libraries:
    - Re-engineered PyHEP interface, with modern cppyy and clang-repl
    - Heterogeneous computing support for 4-vectors, linear algebra (small/medium sizes) and histograms

Outlook…

- EP R&D for software has provided valuable support to many *high impact* projects in HEP computing and software
- Bearing fruits already, with early adoption of outputs by the experiments
- Almost all projects happen in *cooperation* with experiments and other institutes
  - Very important to coordinate efforts and broaden the base of R&D
  - Some very clear areas of common interest with Swift-HEP
- EP R&D II would give a great boost to long term prospects for software R&D

We are always very happy to have new lines of common interest