Reduced Formats Summary

Thanks to all speakers for their excellent talks, and everyone contributing to the discussions!

Thanks to our rapporteur Brian Cardwell!

Session convenors:
Allie Hall (United States Naval Academy)
Jana Schaarschmidt (University of Washington)
Loukas Gouskos (CERN)

Analysis Ecosystems Workshop II - May 23-25 2022 Orsay

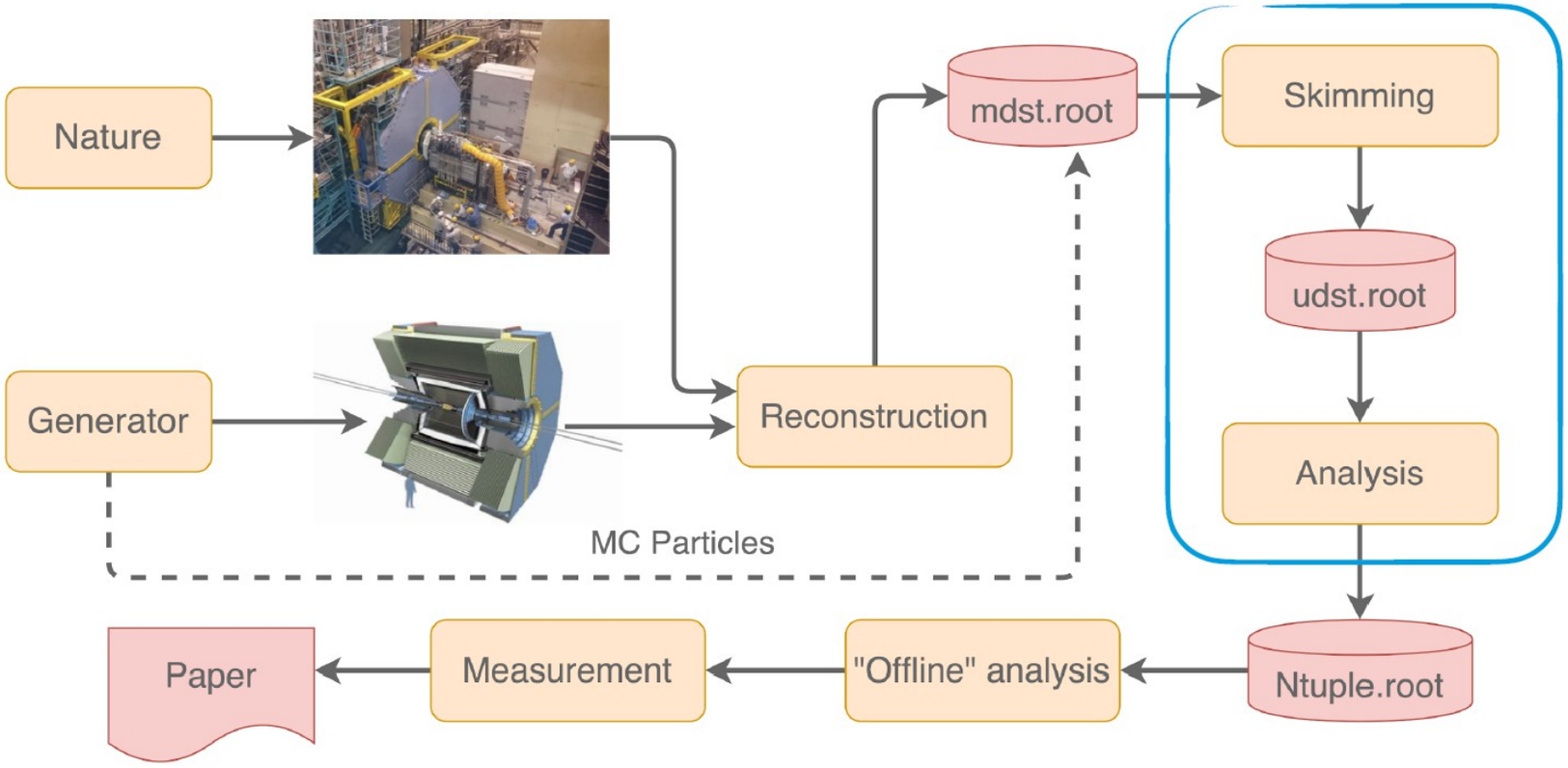
Reduced Formats in Belle II

Expected to collect a dataset of O(10) PB/year

Raw (~70 kB/event) → mDST (~15 kB/event) → uDST (~20 kB/event) (DST: Data Summary Table)

mDST contains tracks, clusters, MC information
uDST is a skimmed version of mDST but it holds also analysis objects (ie. particle candidates)

Skims defined via python-based classes, currently 70-80 skims exist, production of skims is a bottleneck, with a huge load on i/o, not so much on CPU. Not every analysis however can use such a skim.



Long-term data preservation is an important topic, interest in central services to achieve this.

Details in Michels talk

Reduced Formats in CMS

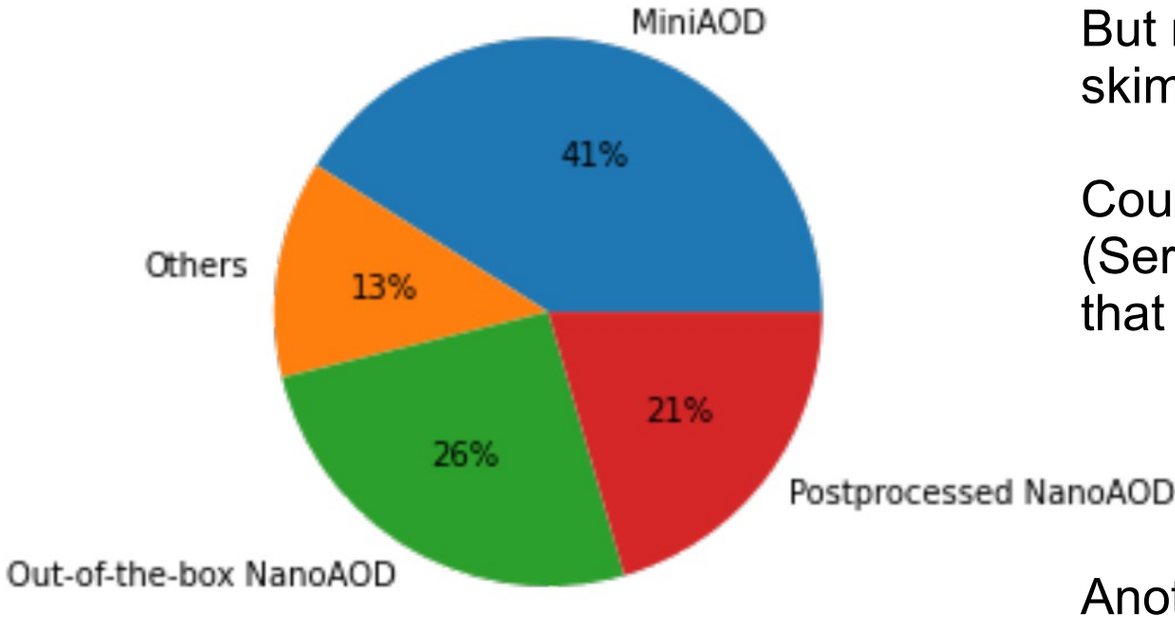
AOD (~500 kb/event) → **MiniAOD** (~50 kB/event) → **NanoAOD** (~2 kB/event)

MiniAOD format contains slimmed object collections, PFlow and tracks

NanoAOD is a flat ntuple, strictly controlled to keep size small, containing high-level objects

MiniAOD and NanoAOD serve 85% of all current analysis!

Floats stored with limited precision (based on detector resolution)



But nearly half of NanoAODs are customized (either skimmed or extended with extra info)

Could avoid the „full-copy“ overlaps by central service (ServiceX, Crab, Dask, regular Batch, …)
that allows people to write extra columns („LegoAOD“)

Details in Lindsays talk

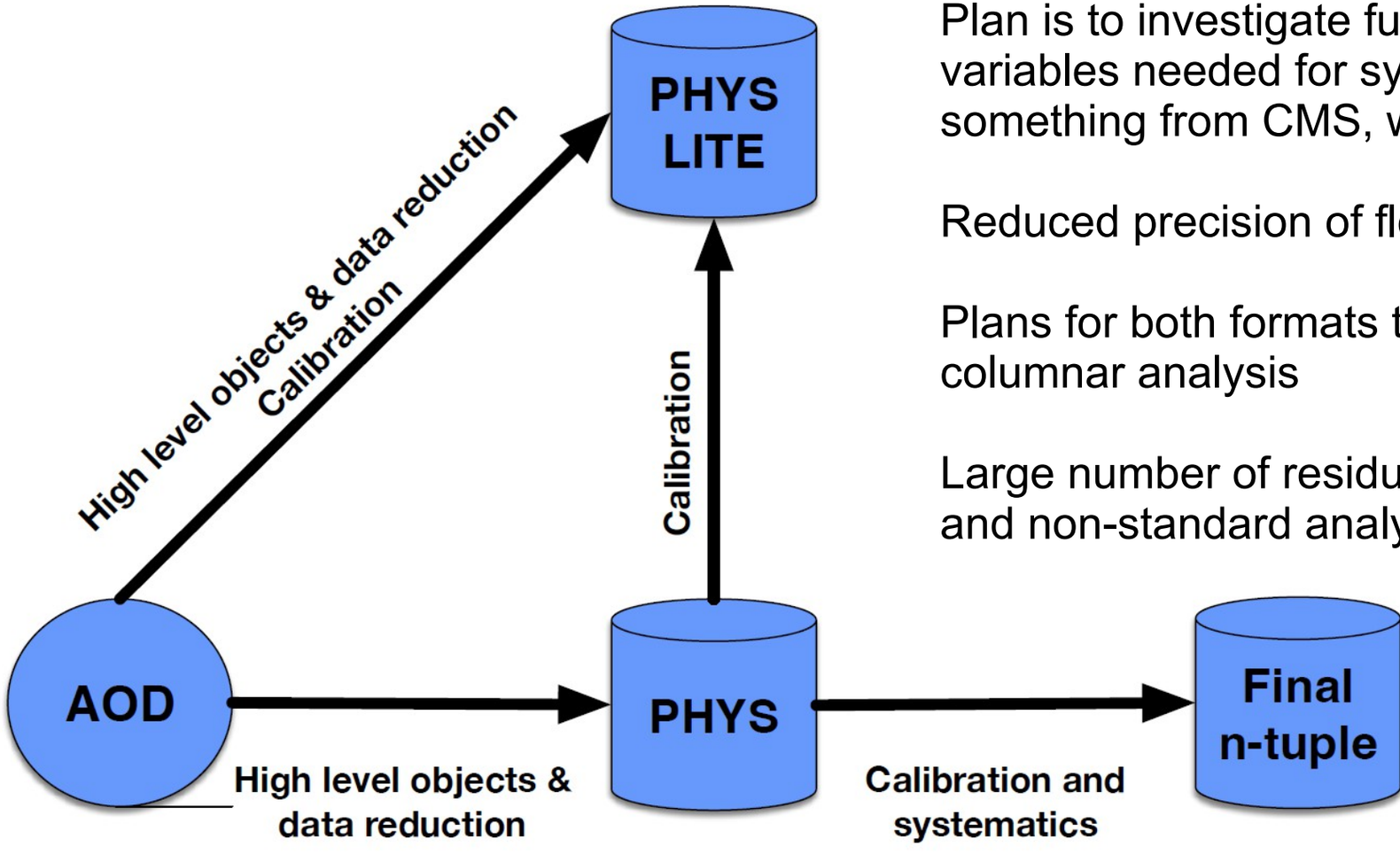
Another possibility: Object stores, eg. to avoid copying columns across processing tiers (→ see talk by Nick)

Reduced Formats in ATLAS

AOD (300-600 kB/event) → PHYS (30-50 kB/event) → PHYSLITE (10-15 kB/event)

Common formats aiming to be used by 80% of the analysis (PHYS in run-3, PHYSLITE in run-4)

PHYSLITE will be frequently produced using latest recommnedations for calibrations etc.



Plan is to investigate further reduction by dropping variables needed for systematics → ATLAS could learn something from CMS, where systematics are parametrized

Reduced precision of floats is available but not yet applied

Plans for both formats to move to RNTuple to facilitate columnar analysis

Large number of residual formats needed for CP activities and non-standard analyses

Details in James talk

Exotic Signatures vs. Reduced Formats

Most analysis can use reduced formats, but what about the rest?

Exotic signatures present additional challenges since they rely on non-standard objects

First step is to be able to run on MiniAOD or PHYS, this alone would be great.

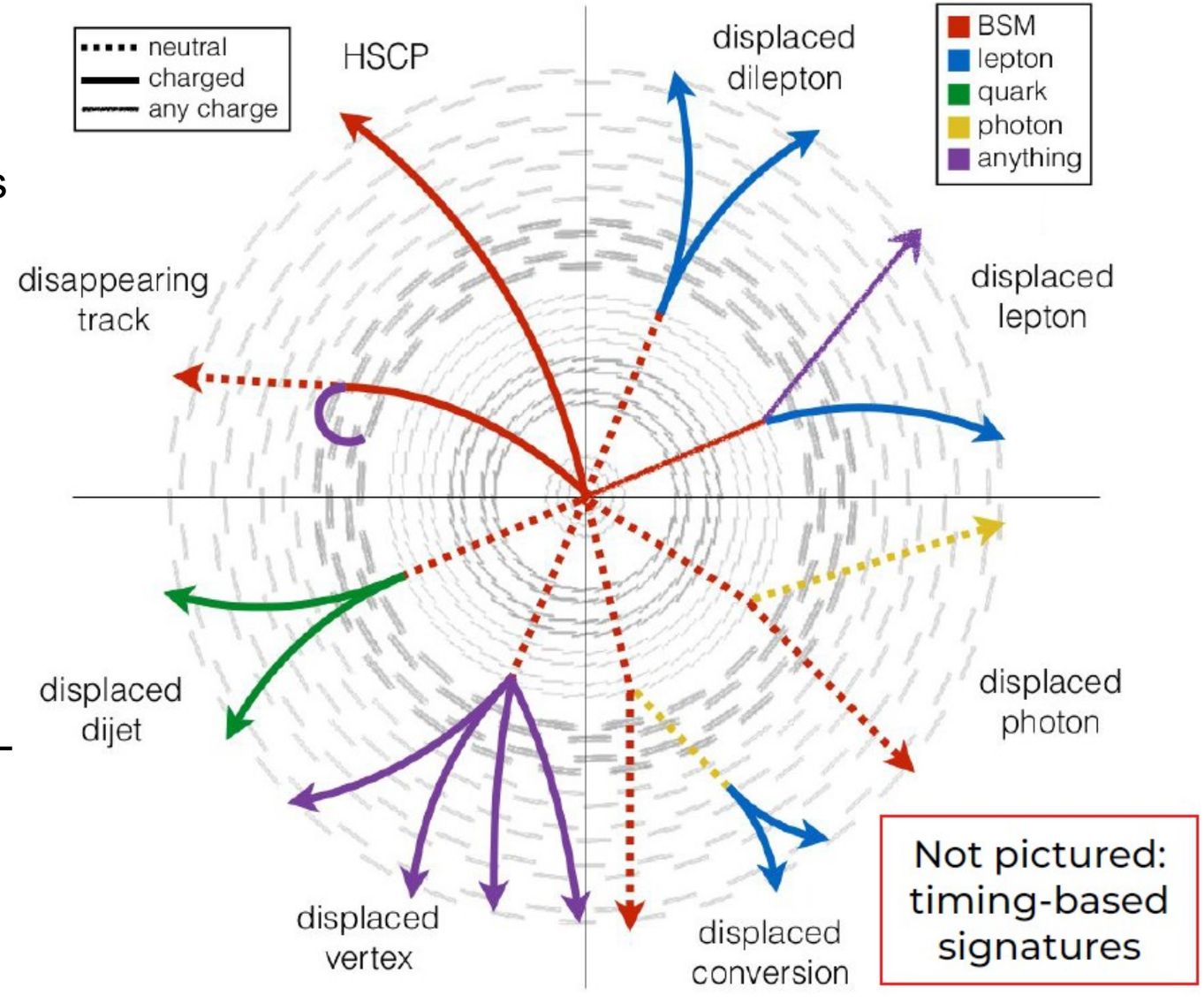
Some case studies:

- Displaced muons → can be added if filtering out those that overlap prompt muons

- Disappearing tracks → need ECAL and HCAL rechits → can be added but skim required

- Magnetic monopoles → unique signature, requires extra tracker and ECAL info, difficult

Custom formats needed, with dedicated skims



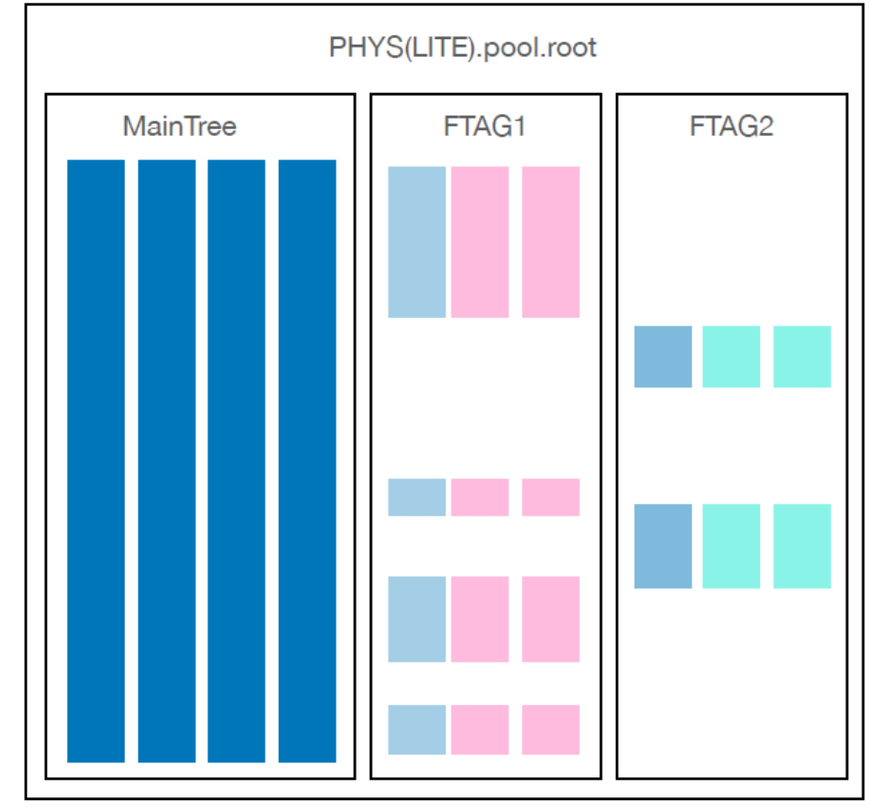
Details in Bryans talk

Augmenting PHYSLITE

PHYSLITE and PHYS are unskimmed fomats, adding more variables or objects is therefore expensive

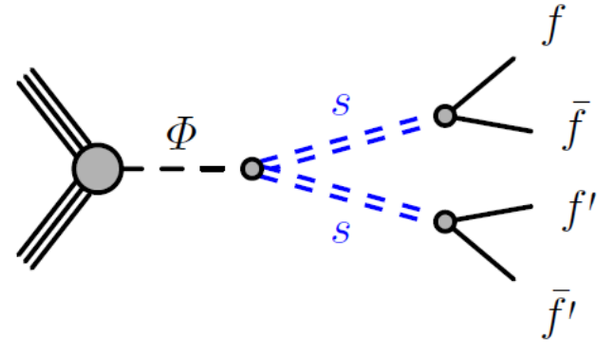
Idea: Add information only for a subset of events, in form of **friend trees** with a common index

Implementation works for storing the friend tree in the same file as the main tree, It gets more complicated (esp. book-keeping) when trees are stored across several files



Case study:

Displaced jets in the calorimeter



Requires that topoclusters are added to PHYS, which increases size by 140% (for ttbar).

Adding them only for events that pass the trigger leads to 2% increase. Very encouraging!

Details in Lukas' talk and Jackson's talk