NANOAOD:
a new compact event data format in CMS

Karl Ehatäht (NICPB, Tallinn)
on behalf of the CMS collaboration
karl.ehataht@cern.ch

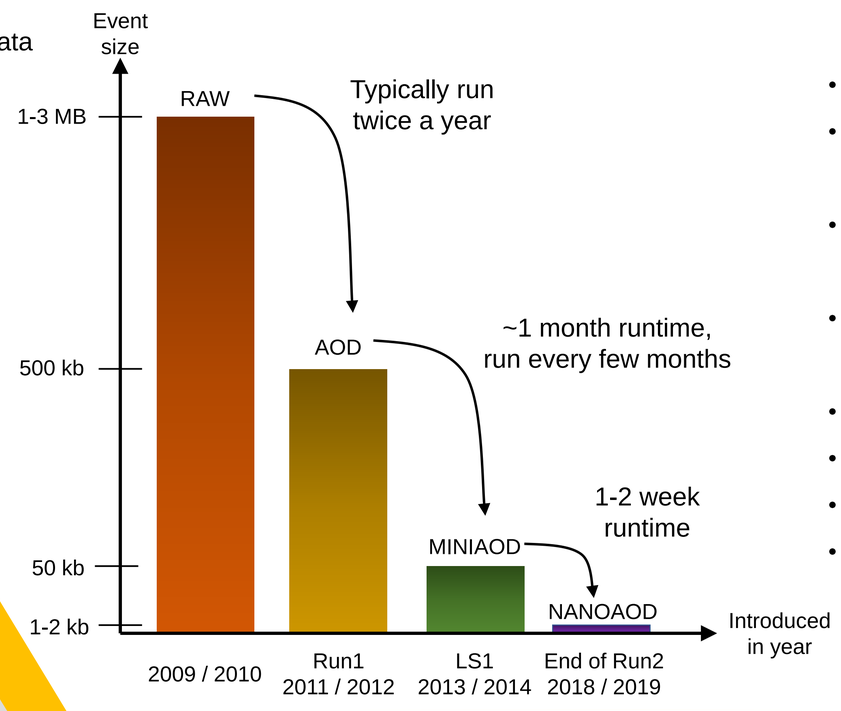

Figure 1. Evolution of data tiers used in CMS.

- Typical thresholds used in physics analyses applied to high level physics objects (μ, e/γ, τ_h, jets)
- Storing the output of e/γ, τ_h and b-jet identification algorithms, rather than the input variables to these algorithms
- Not storing redundant information that can be later derived from the event content (such as systematic variations of jet energy scale)
- Not storing 32-bit floating point numbers at full precision because experimental measurements are not performed at this high precision
- File size further reduced with LZMA compression algorithm
- Production fast at 10 events / second / core
- Reading faster than kHz with decompression
- Required storage in data tiers projected to reduce by a factor of two (more than 2 exabytes) if 50% of the analyses switch from MINIAOD to NANOAOD during HL-LHC [1]

Need to store more and more data as LHC continues to operate

Motivation

- No need for CMS software to perform data analysis
- Lower barrier of entry for new students
- Enables open data access for people outside CMS
- Documentation of physics objects embedded in NANOAOD Ntuples
- Increases integrity and quality of physics analyses
- Modular structure of the NANOAOD format promotes parallel and continuous development

Perform data analysis with plain ROOT

Need to manage reconstruction and identification “recipes” as the detector conditions evolve over time

- No tracks or individual particle candidates, store jets instead
- No details about detector configuration
- No object cleaning and skimming applied to maintain flexibility on an analysis level
- Standardized access to complex high level physics objects – simplifies automation of physics analyses
- No need for custom Ntuple production by individual analysis groups
- Time between data recording and publication shortened

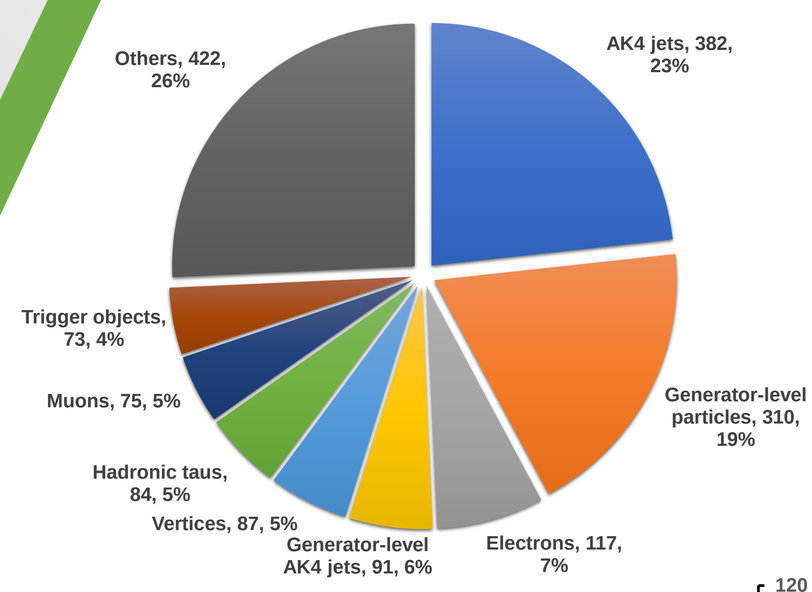

Figure 2. Breakdown of a simulated $t\bar{t}$ event by average storage size of its high level objects (in bytes).

Figure 4. Structure of NANOAOD Ntuple [2].

Centrally maintained companion tool available to perform analysis-specific steps after NANOAOD production:

- Computation of systematic variations of physics observables, data to simulation corrections, scale factors for identification efficiency
- Slimming and merging of object collections
- Calculation of more complex physics observables
- Event skimming

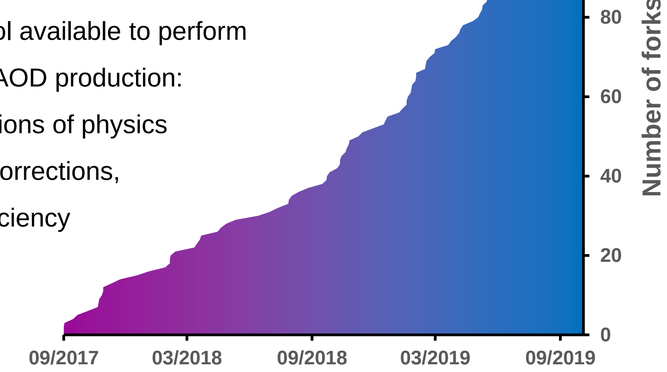

Figure 3. Adoption rate of the NANOAOD companion tool (github.com/cms-nanoAOD/nanoAOD-tools).

References

[1] "The NanoAOD event data format in CMS", M. Peruzzi, G. Petrucciani, A. Rizzi, ACAT 2019

[2] "A further reduction in CMS event data for analysis: the NANOAOD format", M. Peruzzi, G. Petrucciani, A. Rizzi, EPJ Web Conf. Volume 214, 2019

Future prospects

- Promote the format to achieve target adoption of 50-70% of all CMS physics analyses
- Develop customized NANOAOD formats for specialized tasks such as automation of object calibration workflow
- Explore the possibility to add support for alternate choices of generator information and parton distribution functions which are currently expressed as event weights but do not fit within the typical NANOAOD event size
- Increase the frequency of producing new NANOAOD Ntuples