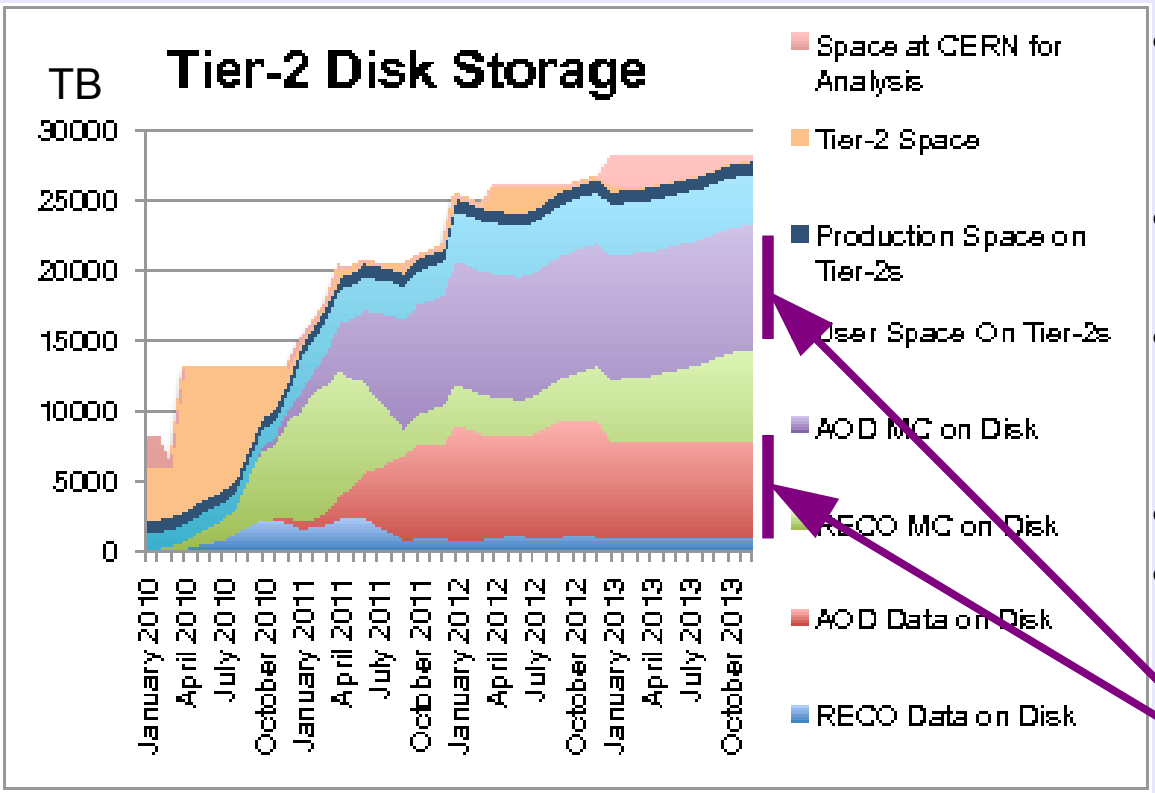
Mini-AOD: A New Analysis Data Format for CMS

Carl Vuosalo
University of Wisconsin-Madison
on behalf of the
CMS Collaboration

Special Thanks to Giovanni Petrucciani and Andrea Rizzi

CMS Run 2: 10X More Data

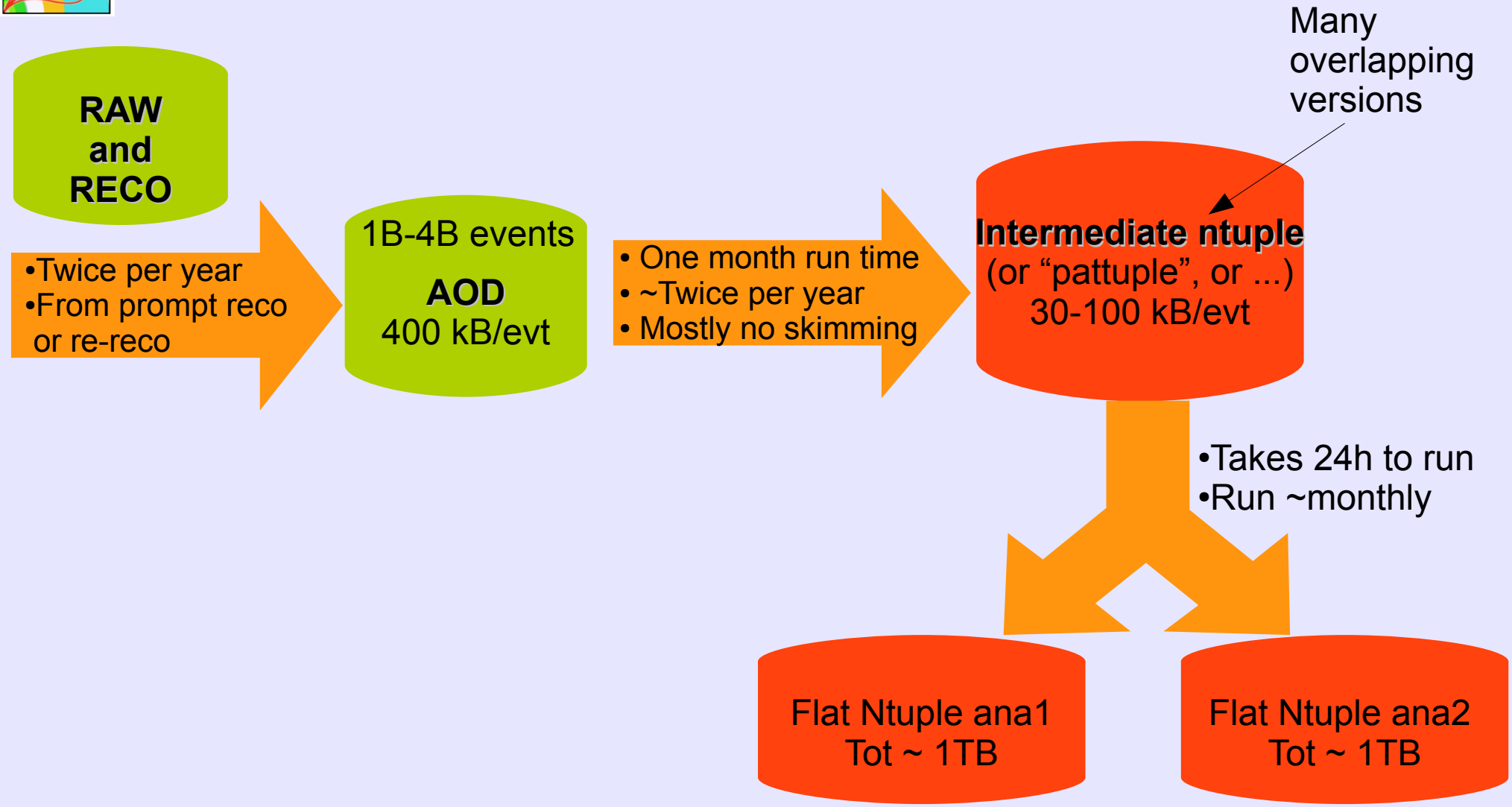
Run 1



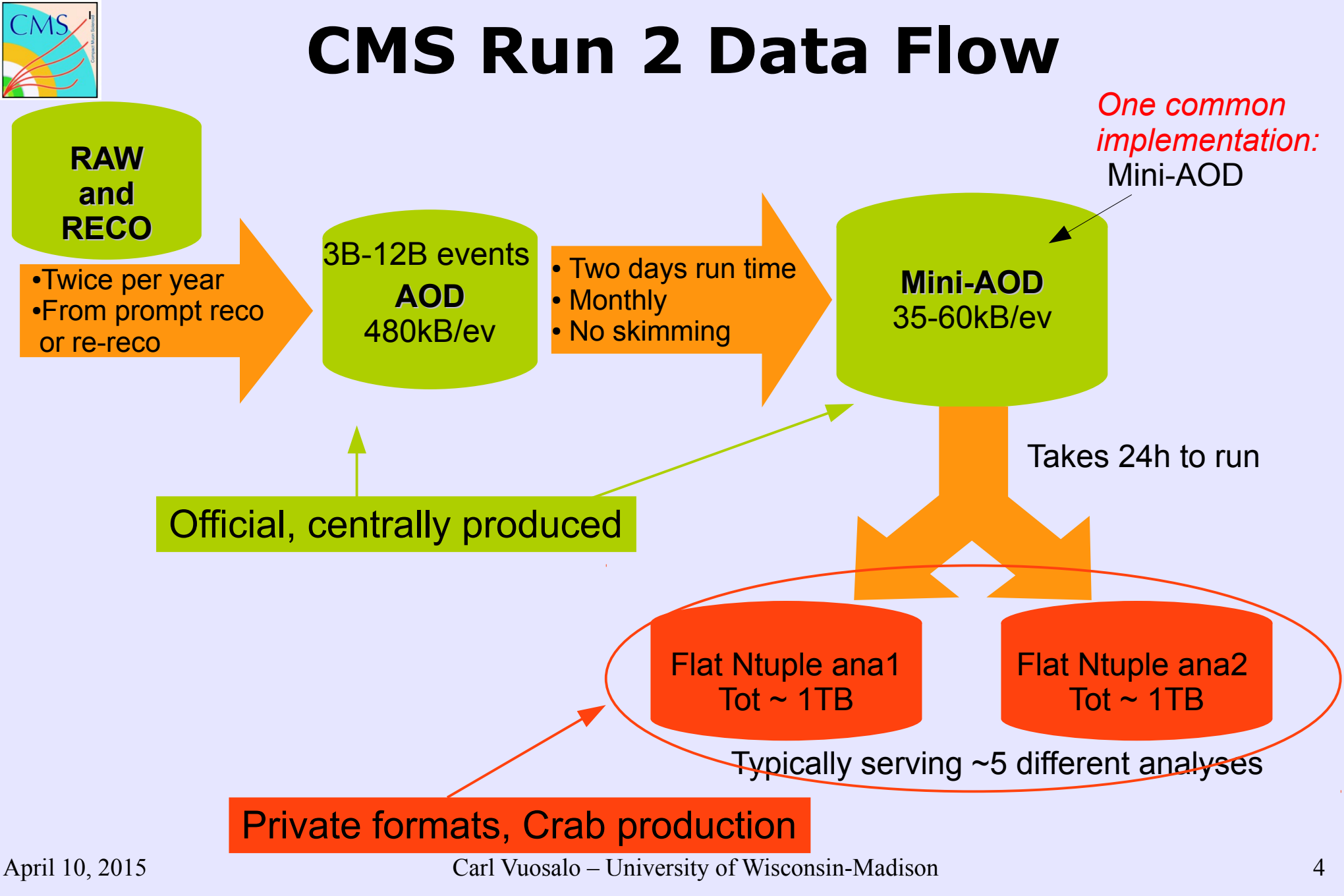
- CMS Run 2 expected to produce **ten times more data** compared to Run 1
- Disk storage resources only increasing marginally
- AOD (Analysis Object Data) used for analysis in Run 1, but it uses too much space for Run 2
- Need more compressed data format
- Mini-AOD provides solution

AOD total in 2013 was ~20 PB

Old CMS Run 1 Data Model



CMS Run 2 Data Flow



Mini-AOD Philosophy

- Use minimum amount of space
- Extract only minimum required data
- Re-use existing data formats and algorithms when appropriate
- Maintain flexibility for:
  - New analysis techniques
  - Re-tuning
  - Analysis-dependent options
- Don't over-optimize – Requirements are to:
  - Store 5 billion events on a Tier-2 site
  - Process those events in 1-2 days

Mini-AOD Event Content

- **High-level physics objects** in PAT (Physics Analysis Tools) format, including detailed information
- **All Particle Flow (PF) Candidates**, in packed format with only basic kinematic information
- **Trigger info:** Bits, 4-vectors of objects, prescales
- **MC truth:** selected generator-level particles (including all the final-state ones); GenJets; generator, LHE, and PDF info
- Other analysis-level information: primary and secondary vertices, E_T^{miss} cleaning filters

Physics Objects in Mini-AOD

Electrons:

- Keep high-quality electrons
- Detailed info for $p_T > 5$ GeV

Muons:

- Keep all with $p_T > 5$ GeV, or that pass loose ID
- All information saved

Taus:

- Keep those with $p_T > 18$ GeV
- Save IDs & links to PF candidates

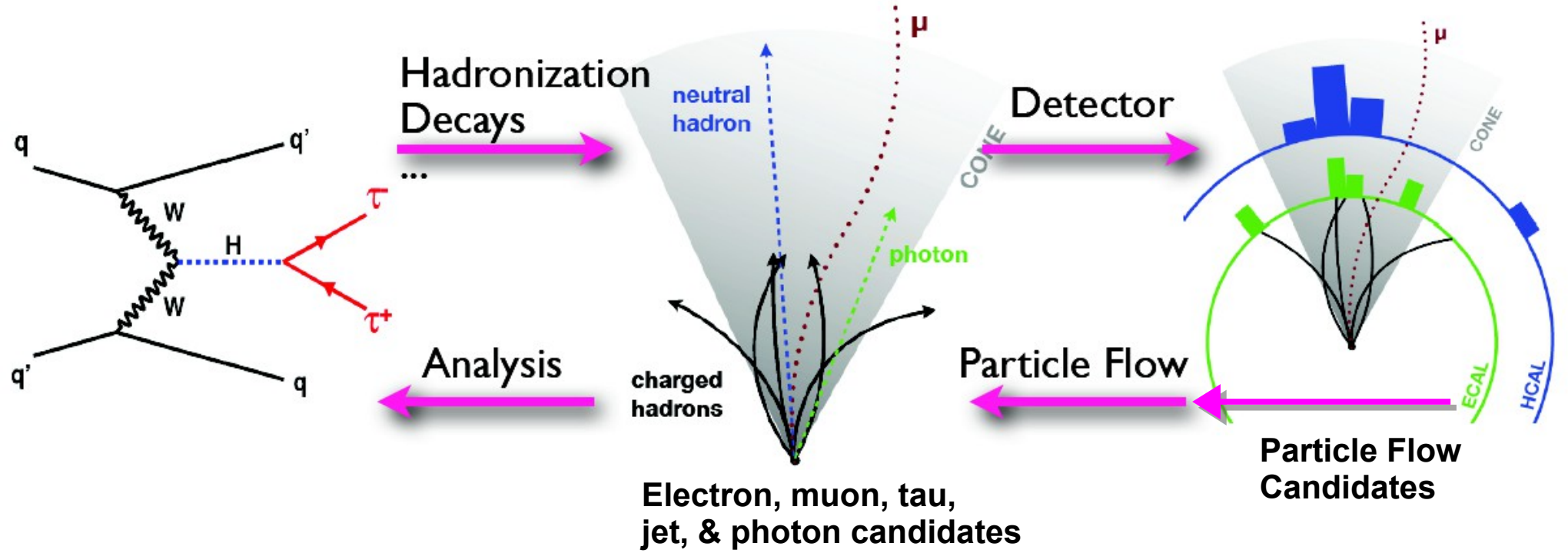
Photons:

- Keep those with $p_T > 14$ GeV
- Detailed info for high-quality photons

Jets (ak4PFchs, ak8PFchs):

- Keep those with $p_T > 10$ GeV ($p_T > 100$ for ak8 jets)
- Note: Jet energy corrections are applied
- Keep daughters, ID info, b-tag discriminators (ak4), substructure info (ak8).

Particle Flow Algorithm



- Particle Flow (PF) algorithm uses PF candidates to reconstruct particle candidates from raw data or simulation
- Preserving PF candidates in Mini-AOD enables re-reconstruction of particle candidates with new techniques

Packed PF Candidates in Mini-AOD

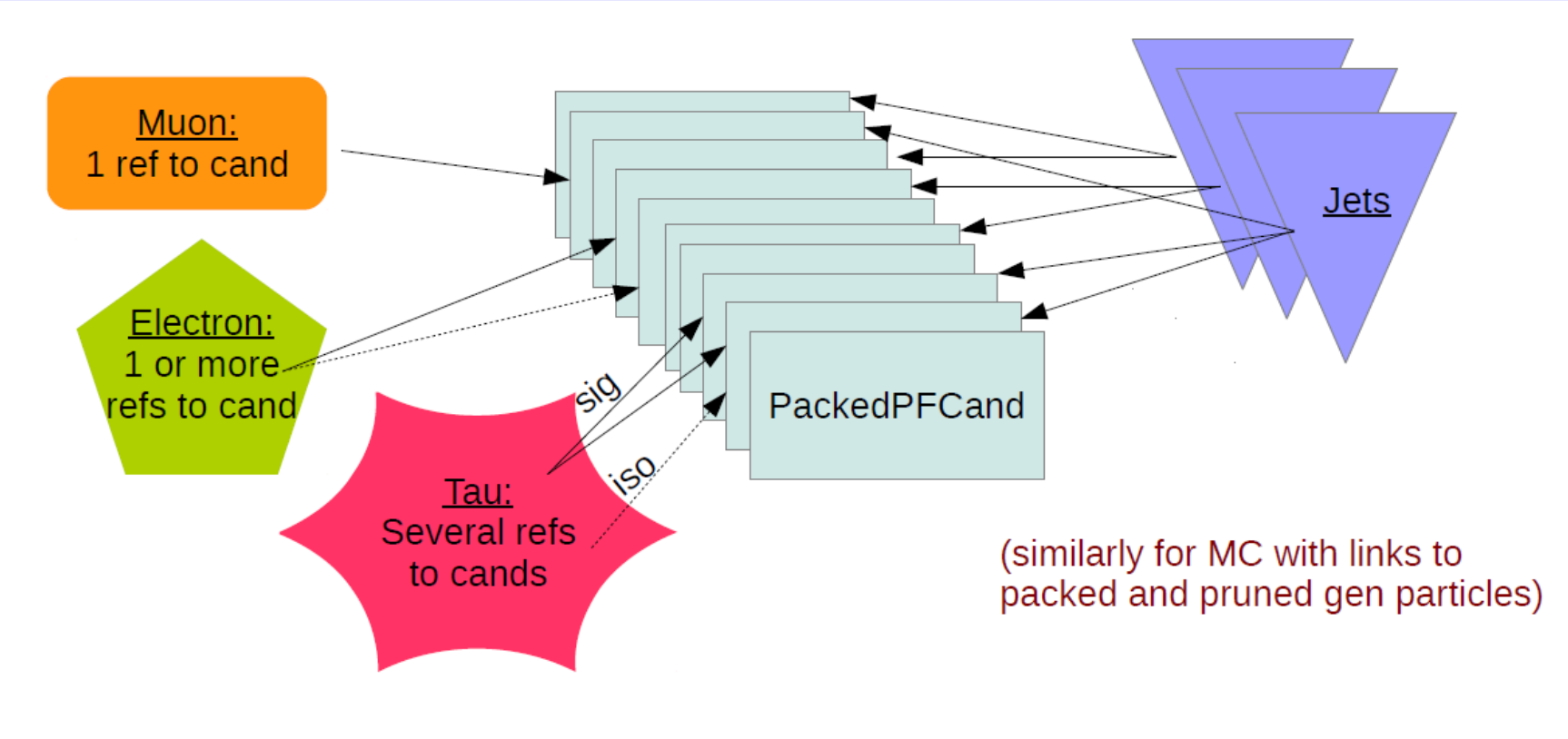
- For all packed candidates some basic info is saved:
  - PDG ID, 4-vector, charge, impact parameters
  - Lossy compression applied on variables, with precision of ~0.1%
  - Compression facilitated by sorting of PF candidates, conversion of double → float, reduction of covariance matrix precision from $10^{-7} \rightarrow 10^{-4}$
- Some extra quality flags are provided:
  - Association to primary vertex
  - Found or lost hits in innermost tracker layers
  - Track 'highPurity' flag

Packed PF Candidate Capabilities

- **Packed PF candidates support:**
  - ✔Computation of arbitrary lepton and photon isolation algorithms, with any pile-up mitigation scheme
  - ✔Study of pile-up mitigation algorithms for jets and E_T^{miss}
  - ✔Jet re-clustering for substructure studies or event interpretation, including re-running of b tagging
  - ✔Other candidate-based analysis algorithms like isolated track veto or soft FSR photon recovery

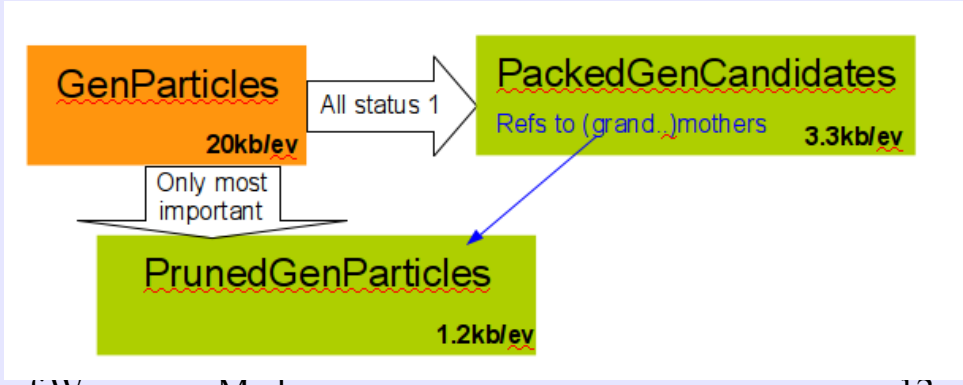
Cross-Referencing from Physics Objects

- Physics objects in Mini-AOD contain references to packed PF candidates corresponding to original PF candidates they came from
  - Useful for footprint removal in isolation, event interpretation



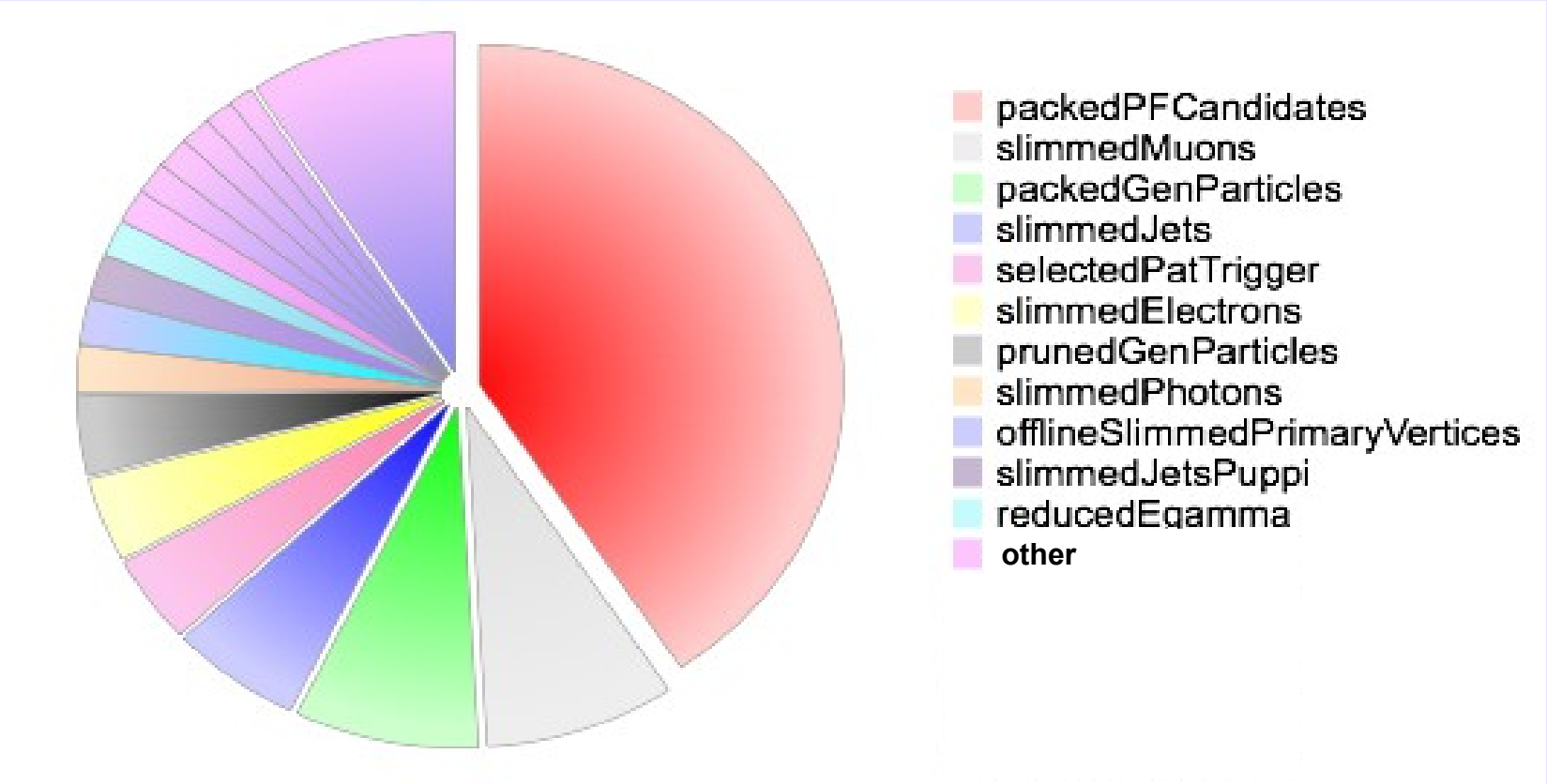
MC

- Small components like LHE records are kept
- GenParticles are large, so only stored as follows:
  - Packed GenParticles: Only status = 1 (hard interaction), with only 4-vector and PDG ID
  - Pruned GenParticles: standard GenParticles but only key ones:
    - Initial partons, heavy flavor, EWK bosons, leptons
- Packed GenParticles facilitate remaking GenJets with different algorithms
- Pruned GenParticles enable event classification, flavor definition, and matching of physics objects
- Each packed GenParticle linked to last surviving ancestor in pruned GenParticle collection

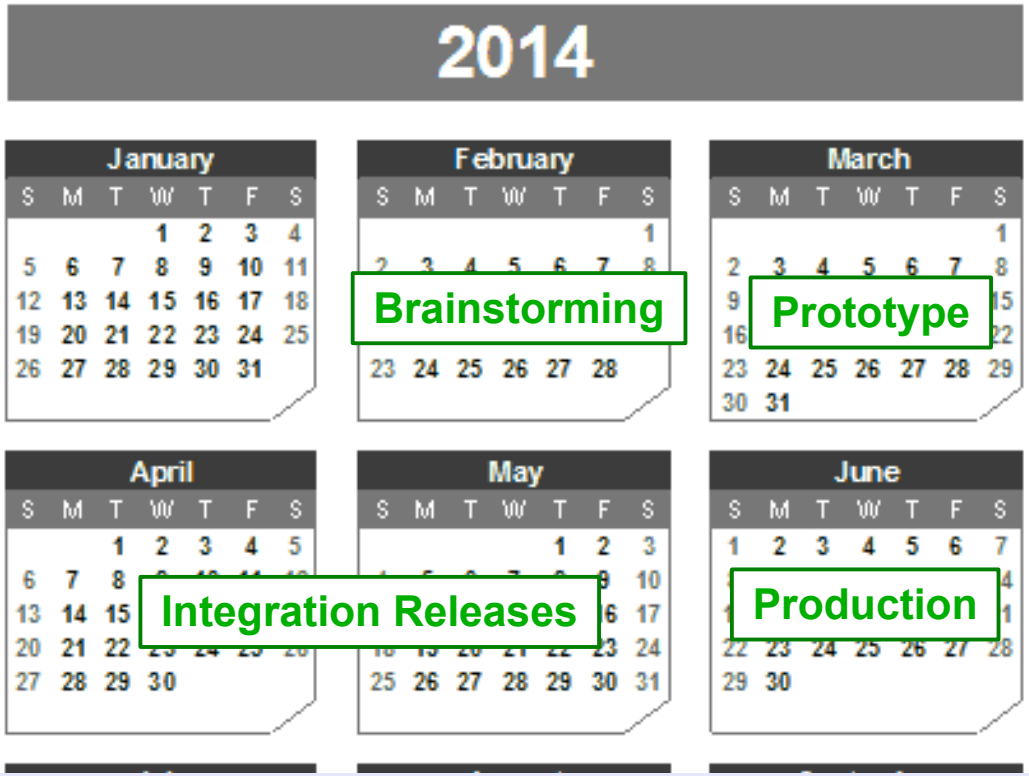


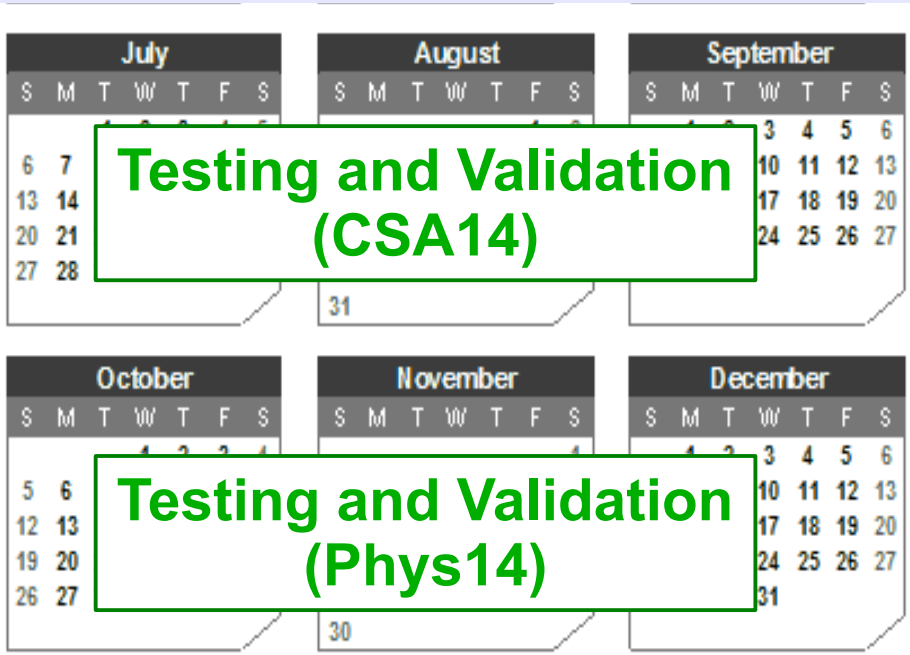
Composition of Mini-AOD

Overall size for ttbar MC: ~40 kB/event



Mini-AOD Development History





- Bugs and issues were discovered and fixed
- Some components added at user request, but only within size budget
- Mini-AOD now validated and ready as data taking begins

Summary

- CMS faces challenge of analyzing ten times more data with Run 2
- Previous method of storing AOD and many versions of intermediate ntuples would overflow storage capacity many times over
- Solution is compressed Mini-AOD
  - 10% of size of AOD
  - Replaces intermediate ntuples with standard format for most analyses
- Optimized collections storing only minimum amount of required information
- Still maintains flexibility for new analysis techniques and re-tuning
- Mini-AOD has been validated and is ready for use in Run 2