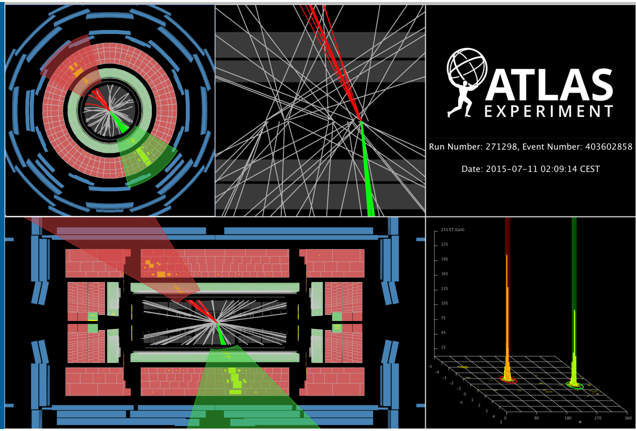
Argonne NATIONAL LABORATORY

FIRST EXPERIENCE WITH THE NEW ATLAS ANALYSIS MODEL



JACK CRANSHAW
Argonne National Laboratory
On behalf of the ATLAS Collaboration

ICHEP2016, Chicago, Illinois
August 4, 2016

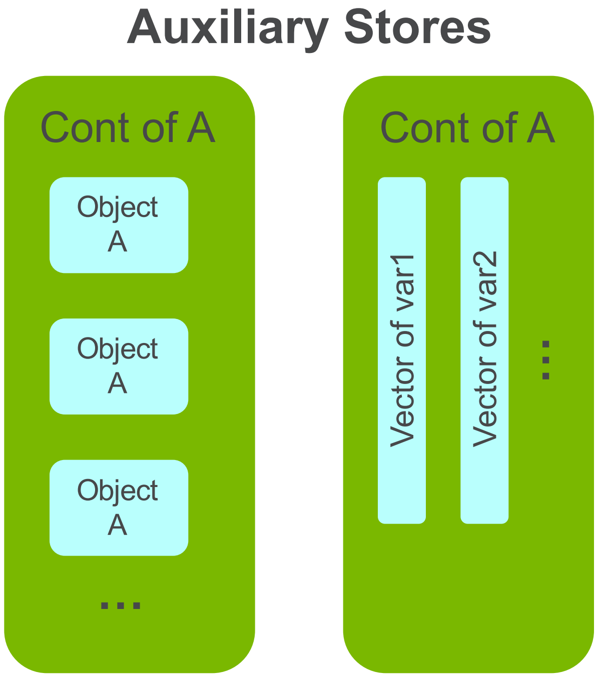
IMPROVING THE RUN 1 MODEL

- The Run 1 ATLAS data format was different for analysis and reconstruction.
  - Storage usage was sub-optimal as some content was duplicated in different formats.
  - Format conversions meant operational inefficiencies.
  - Multiple formats made
    - tool maintenance a problem as there were either multiple tools for efficiency corrections, etc. or the tools had to have different code for different inputs.
    - consistency checks more complicated.
- Clear opportunity for improvement in Run 2.
  - We developed a more flexible data format (xAOD) that can be configured in ways which work for both analysis and reconstruction.
  - We developed a system to produce and catalog samples that are prefiltered for analysis using tools which physicists can run themselves and feed any developments directly back up the processing chain.

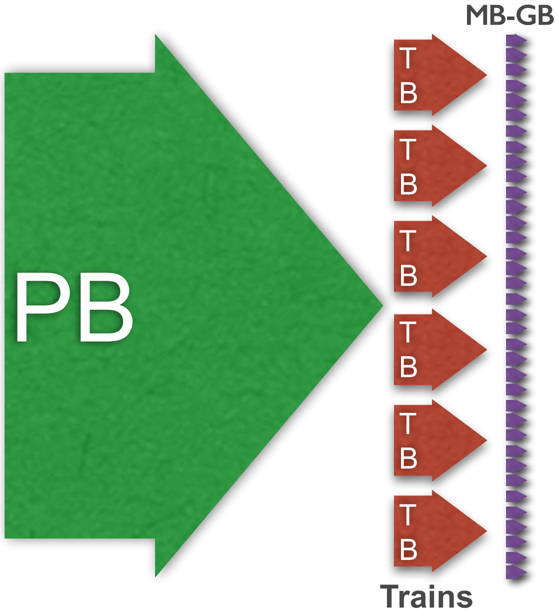
RUN 2 ANALYSIS OBJECT DATA (AOD)

xAOD Format

- In Run 1 the AOD format had two aspects which made it unpopular with analyzers but popular with reconstruction.
  - It was fast to retrieve *groups of events*.
  - It wrote in a format *optimized for space* which required object reconstitution for some objects.
- For analysis what was wanted was different.
  - Fast retrieval of *individual variables*.
  - *Direct usability in ROOT* with no externals.
- Using an object called an auxiliary store, we are able to write data in *either format* simply by changing the ROOT settings. Some of the advantages of Run 1 were compromised, but the advantage of a single format outweighs them. (see poster on Saturday)



DATA REDUCTION:
A FEATURE COMMON TO MOST PHYSICS ANALYSES

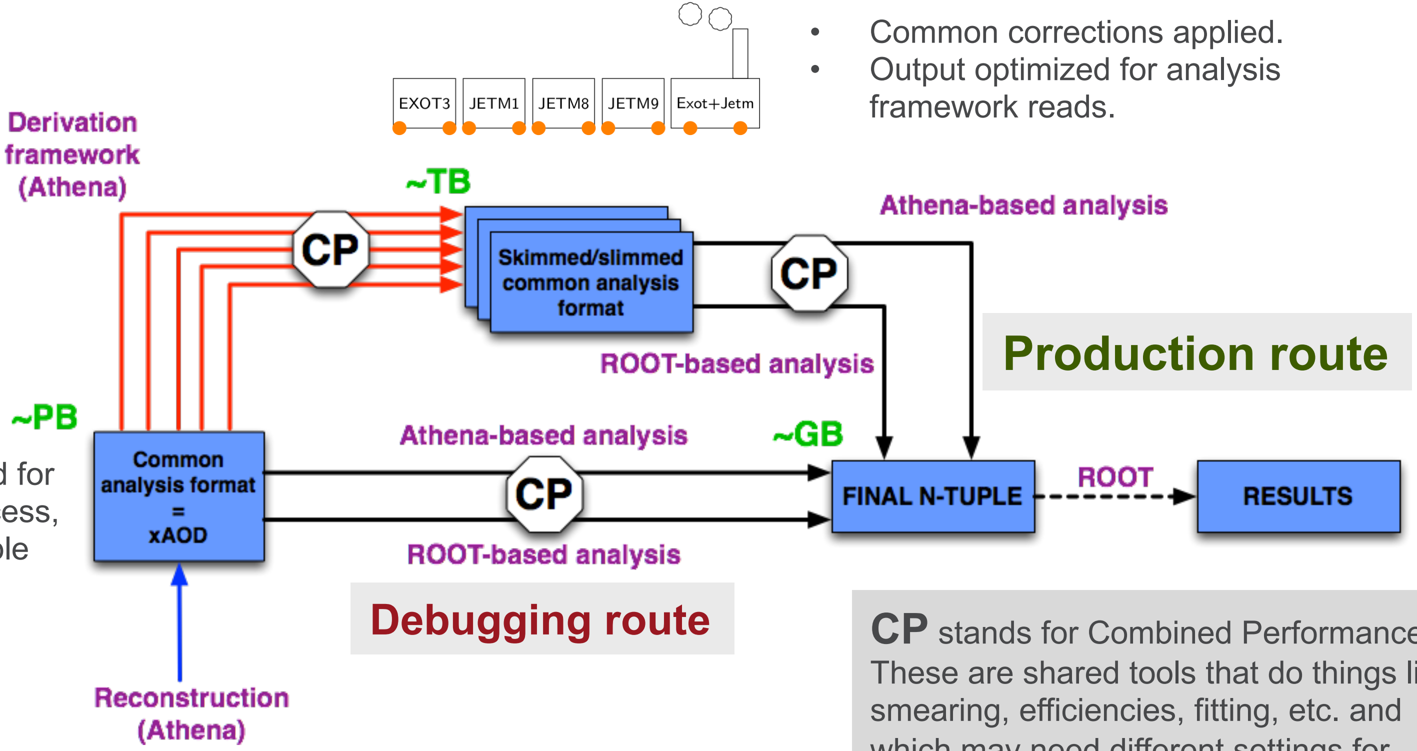


	Full output of reconstruction, ~PB size	One schema (xAOD event-wise)
	Derivations Intermediate analysis format ~TB size	~100 schemas (xAOD column-wise)
	Final n-tuple ~MB-GB size	~1000 schemas (ROOT various)

- Centrally produced for both data and Monte Carlo.
- These formats tend to be specific to a single analysis or group of analyses.
  - *One can think of these as physics group 'software experiments' that can be repeated monthly as necessary.*
- Calibrations and common object selections are often applied as derivations are made.
- They generally need to contain all variables needed for calculating systematics.
- Stringent limits on size and close coordination with physics groups who share resources.
  - ~100 derivations are grouped into of order 20 trains which run 2-10 derivations (*carriages*).

THE RUN 2 ANALYSIS MODEL FOR ATLAS

- Common corrections applied.
- Output optimized for analysis framework reads.



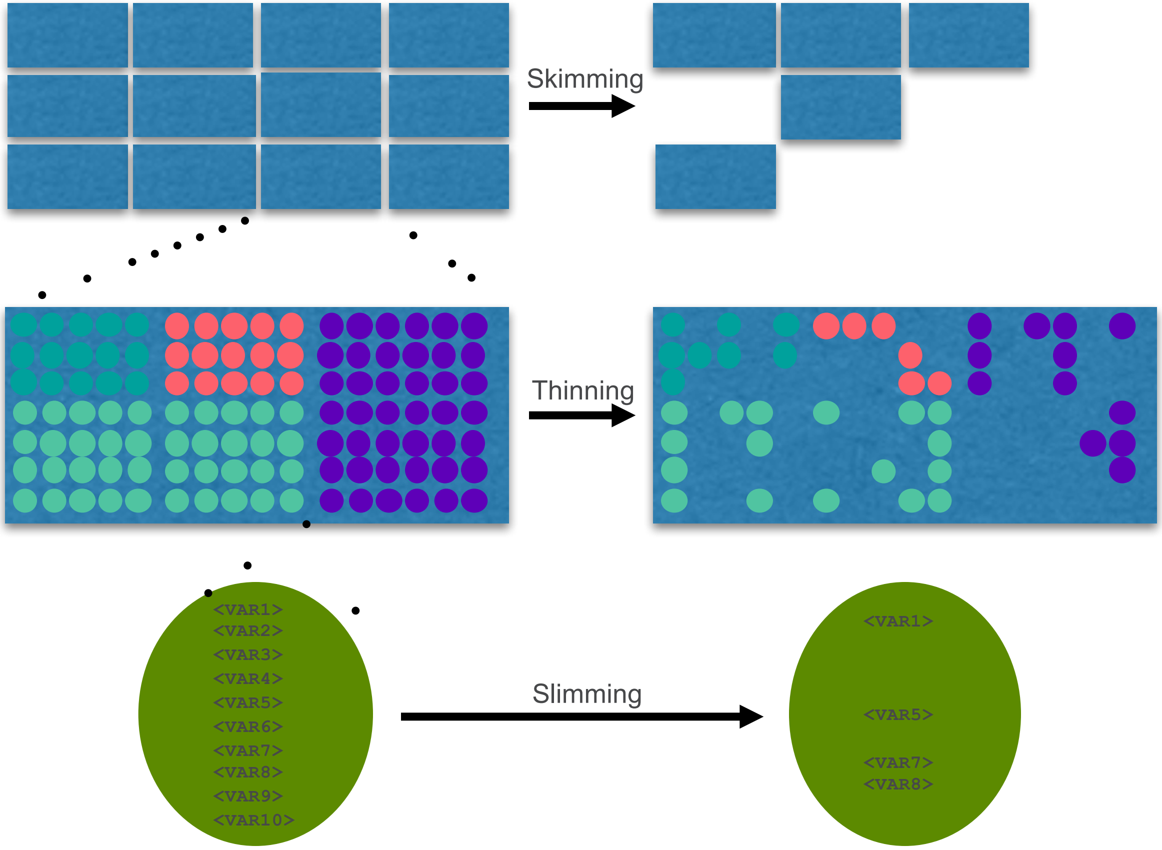
Optimized for event access, not variable access

CP stands for Combined Performance. These are shared tools that do things like smearing, efficiencies, fitting, etc. and which may need different settings for different physics applications.

USER INTERFACE

- Derivation developers (physicists) use Athena but interact through interfaces provided by the derivation framework.
  - Athena advantages
    - Full access to Athena I/O infrastructure for streaming and configuration.
    - Full access to reconstruction algorithms when needed.
  - Derivation framework features
    - Interfaces for users to implement tools for skimming, thinning, and augmenting their data. (next slides)
    - A text-based event/object selector to minimize user-developed C++.
    - List of variables needed by the CP tools, allowing ‘smart’ slimming.
    - Monitoring of multi-carriage/train performance.

DATA REDUCTION OPERATIONS (100% -> 1%)



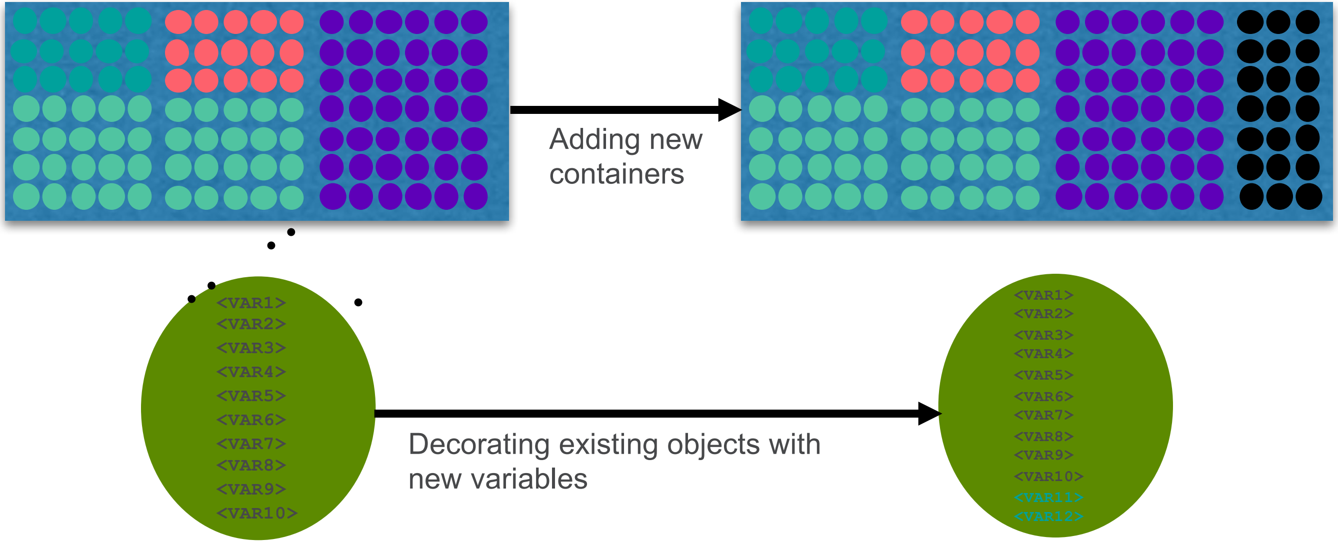
Skimming: removal of whole events based on pre-set criteria

Thinning: removal of whole objects within events based on pre-set criteria

Slimming: removal of variables within objects uniformly across events

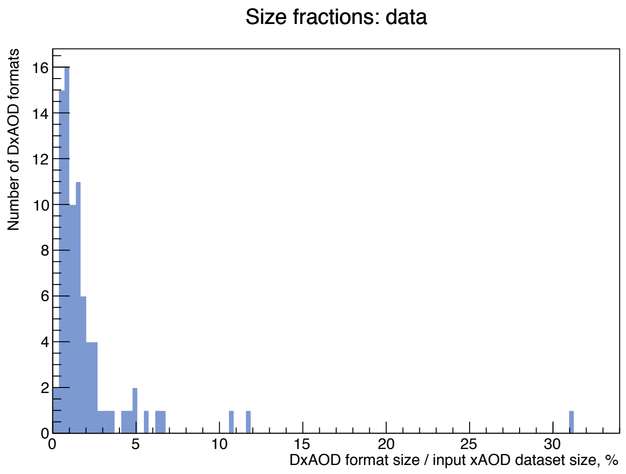
AUGMENTATION OPERATIONS

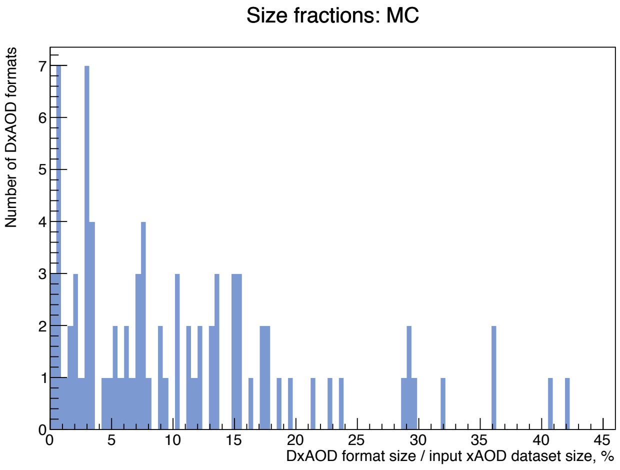
- Addition of new information (augmentation) is typically done in two ways:
  - Adding new reconstructed object containers: typically jets made with a modified algorithm.
  - Decorating existing objects with extra variables: typically the results of object selection by combined performance tools (e.g. “this is a good muon”)
- Augmentation can be shared across a train, saving CPU



IMPLEMENTED DERIVATIONS

- Can scale to more derivations simply by defining new carriages and new trains.
- Users can generally run over an entire derived dataset in one day.
- Derivations limited to 1% of AOD/carriage, 4% of AOD/physics group.





Group	Number of Carriages
B Physics	2
Egamma	8
Flav. Tag	4
Inner Det.	1
Muon	5
Tau	2
Exotics	18
Higgs	20
Jet	11
SM	5
SUSY	13
Tile	1
Top	4
Total	**94**

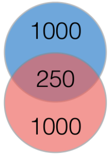
OVERLAPS

- We monitor both content and event-wise overlaps of the various derivations when data or derivations change.
  - Overlap is calculated as intersection/non-intersecting.
  - Most of the time there is not a problem.
    - Below is a recent figure for overlaps for the 14 derivations which had > 70% overlap with at least 1 other derivation.
  - When large event overlaps *are* detected, they are investigated, but they must also have large variable overlap and have similar development schedules.

	EXOT0	EXOT4	EXOT10	HIGG1D1	HIGG5D1	JETM1	JETM3	JETM8	JETM9	JETM11	STDM4	SUSY5	SUSY6	SUSY8
EXOT0	502891 (100%)													
EXOT4	427343 (3.99%)	10647773 (100%)												
EXOT10	128808 (13.63%)	170178 (1.54%)	570837 (100%)											
HIGG1D1	169929 (14.94%)	226424 (2.02%)	570774 (70.95%)	804424 (100%)										
HIGG5D1	18459 (0.2%)	1371482 (7.61%)	6108 (0.07%)	6982 (0.07%)	8740180 (100%)									
JETM1	2938 (0.11%)	225187 (1.79%)	4599 (0.17%)	4624 (0.16%)	159209 (1.48%)	2145394 (100%)								
JETM3	386241 (73.2%)	357159 (3.34%)	94981 (10.71%)	121017 (11.06%)	13162 (0.14%)	1013 (0.04%)	411007 (100%)							
JETM8	2930 (0.14%)	212144 (1.75%)	4588 (0.21%)	4604 (0.19%)	158577 (1.55%)	1660093 (77.38%)	1007 (0.05%)	1660093 (100%)						
JETM9	2938 (0.11%)	225187 (1.79%)	4599 (0.17%)	4624 (0.16%)	159209 (1.48%)	2145394 (100%)	1013 (0.04%)	1660093 (77.38%)	2145394 (100%)					
JETM11	488068 (3.98%)	8371354 (57.67%)	179864 (1.42%)	244306 (1.91%)	760522 (3.76%)	141477 (0.99%)	410983 (3.36%)	136318 (0.99%)	141477 (0.99%)	12238918 (100%)				
STDM4	491629 (3.84%)	10457192 (80.49%)	195011 (1.48%)	263457 (1.97%)	1384221 (6.87%)	231596 (1.57%)	411007 (3.21%)	217481 (1.53%)	231596 (1.57%)	10367259 (70.65%)	12801582 (100%)			
SUSY5	500891 (2.92%)	8948443 (47.46%)	282859 (1.62%)	390570 (2.22%)	1515843 (6.22%)	353323 (1.86%)	411007 (2.4%)	328423 (1.78%)	353323 (1.86%)	12238918 (71.34%)	10936535 (57.5%)	17155997 (100%)		
SUSY6	389660 (3.23%)	4024321 (21.65%)	116448 (0.94%)	147819 (1.17%)	8724403 (72.82%)	160646 (1.15%)	369639 (3.08%)	159825 (1.19%)	160646 (1.15%)	3951791 (19.51%)	4561559 (22.58%)	4708007 (19.29%)	11964696 (100%)	
SUSY8	263590 (2.04%)	4383318 (23.13%)	6065 (0.05%)	6995 (0.05%)	8724282 (68.68%)	166070 (1.13%)	265958 (2.07%)	165262 (1.17%)	166070 (1.13%)	4389894 (21.38%)	5119163 (25.13%)	5180025 (21%)	10489639 (74.07%)	12687666 (100%)

100%	> 90%	> 80%	> 70%	> 60%	> 50%	> 40%	> 30%	> 20%	> 10%	> 0.01%	> 0%	== 0

Example:



$$100 \cdot \frac{250}{2000-250} = 14\%$$

MANAGING THE SOFTWARE

- ***Reconstruction (AtlasProduction)***
  - Releases used in Tier0 and on grid for reconstruction.
  - Releases cut roughly *twice a year*.
- ***Derivations (AtlasDerivation)***
  - Based on a stable AtlasProduction release.
  - Extended with derivation packages and some updates which would be disruptive to include in a current AtlasProduction release.
  - Releases cut roughly *monthly*.
- ***Physics (AthAnalysisBase)***
  - Based on a stable AtlasProduction release, Athena usable.
  - Slimmed down release with many packages used for reconstruction omitted, e.g. RAW data reading.
  - Releases roughly *every two weeks*.
- ***Physics (AnalysisBase)***
  - Like AthAnalysisBase, but with the reconstruction framework (athena) also dropped.

Both Athena *and* ROOT *analyses read at kHz event rates.*

CONCLUSION

- The Run 2 ATLAS analysis model has been a success and has proven more scalable than the Run 1 version.
  - We are able to run of order 100 trains.
    - Trains and carriages have been rearranged as well, which is possible because of nightly train testing.
  - Weekly coordination meetings provide regular feedback from physics groups. Well attended and short. Production of both data and monte carlo derivations have worked well.
  - Physics groups have successfully managed their own derivation sizes to stay within the resource limits.
  - The release schedule has been maintained.
  - There have been a drop in the number of cases of people trying to access the primary AOD directly. (discouraged by grid management as well).
  - The derivation framework has placed no serious constraints on development of physics analysis frameworks.
- It is foreseen that this system will work successfully for the rest of Run 2.