ATLAS Workflows: Derivations

Nurcan Ozturk

for the Derivation Production Team

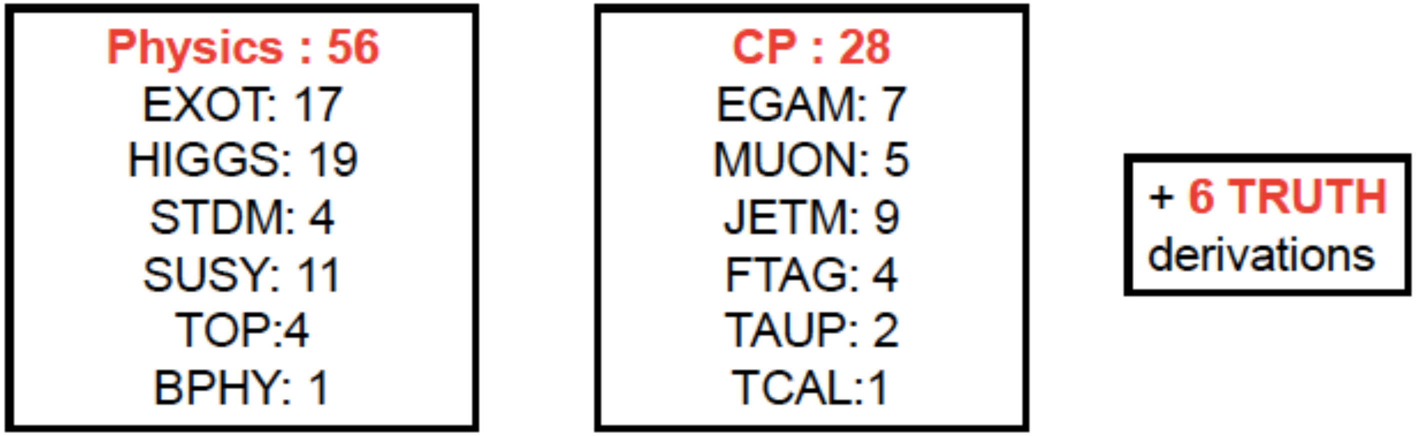
ATLAS Sites Jamboree, January 27th 2016

https://twiki.cern.ch/twiki/bin/view/AtlasProtected/DerivationProductionTeam

The University of Texas ARLINGTON

Derivation Production

- Two types of productions: AOD->DAOD_X and EVNT->DAOD_TRUTHX with internal merging done by JEDI.
- Multiple outputs are produced in a single job, namely train production with carriages; 18 trains on data, 19 trains on MC, 4 of the trains run in single-carriage mode (group trains, CPU-intensive carriages separated from others).
- Number of derivations continously produced is 84.



Physics : 56
EXOT: 17
HIGGS: 19
STDM: 4
SUSY: 11
TOP:4
BPHY: 1

CP : 28
EGAM: 7
MUON: 5
JETM: 9
FTAG: 4
TAUP: 2
TCAL:1

+ 6 TRUTH
derivations

- Open-ended production during data taking.
- MC derivations are run by the groups themselves, however not in train mode but by producing single derivations (except for a few shorter trains).

Task Parameters, Merging

- nGBPerJob=10 and nFilesPerJob=100 are used as default for all tasks after EVNT->DAOD_TRUTH production was introduced to avoid very short jobs.

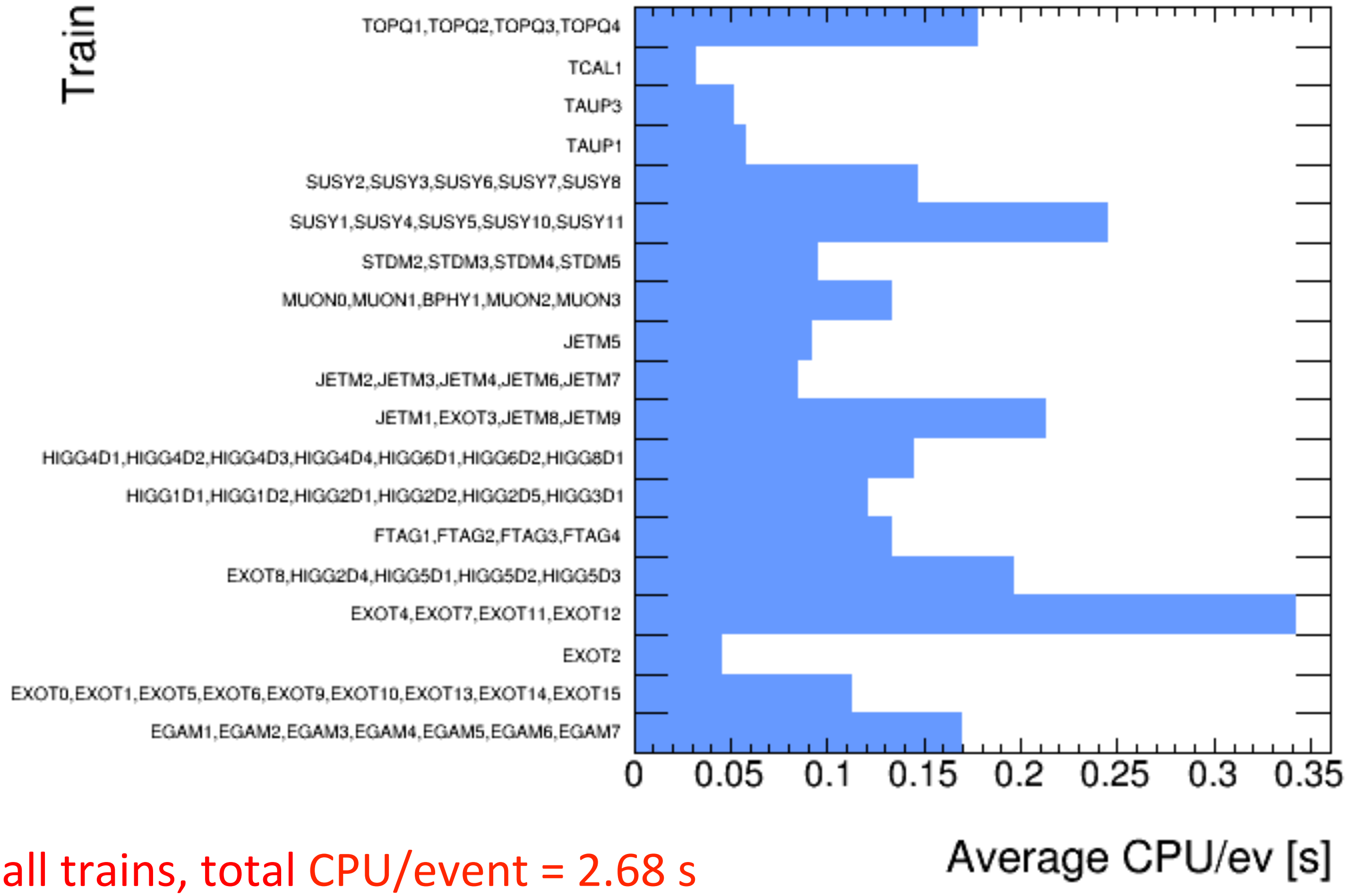
- The following parameters recently added as default:
  - cpuTimeUnit=HS06sPerEvent;cpuTime=0; ramUnit=MBPerCore;cloud=WORLD

- DAOD merging is slow compared to AOD->DAOD making, the merging rate is 15-20 Hz currently, it isn't so bad as the merging is done in serial. Software experts have investigated on possibilities for improvement during the Software TIM meeting at LBNL in November however w/o much success.

- When a site went offline before the merging jobs finished then these jobs at the site expired after 24h, this really slowed down the derivation production. It was investigated at the ADC TIM meeting in Sitges and the brokerage was changed to send the merging jobs anywhere after 24h. Site reliability is important for these merge jobs.

CPU Usage for Data: Average CPU/event per train

Target: ~2s/event for data for all derivations, with 20 trains ~100 ms/event per train

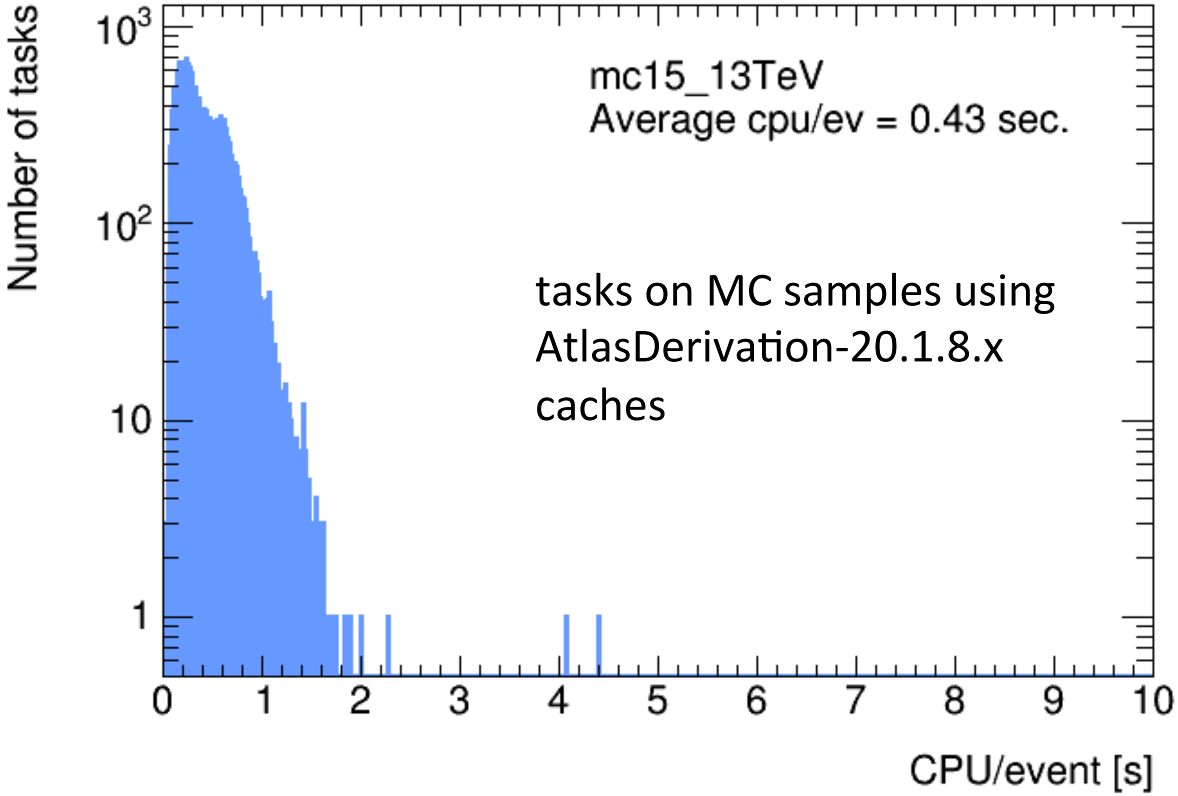


open-ended derivation production of the 25ns runs

For all trains, total CPU/event = 2.68 s

CPU Usage for MC: Average CPU/event

- Train model is not respected for MC derivations as mentioned earlier; not all derivations run on the same MC samples, submitted by the group contacts in a random way, not possible to assemble long trains.
- Usage is pretty good; average CPU/event is 0.43 s. and average walltime/event is 0.57 s.

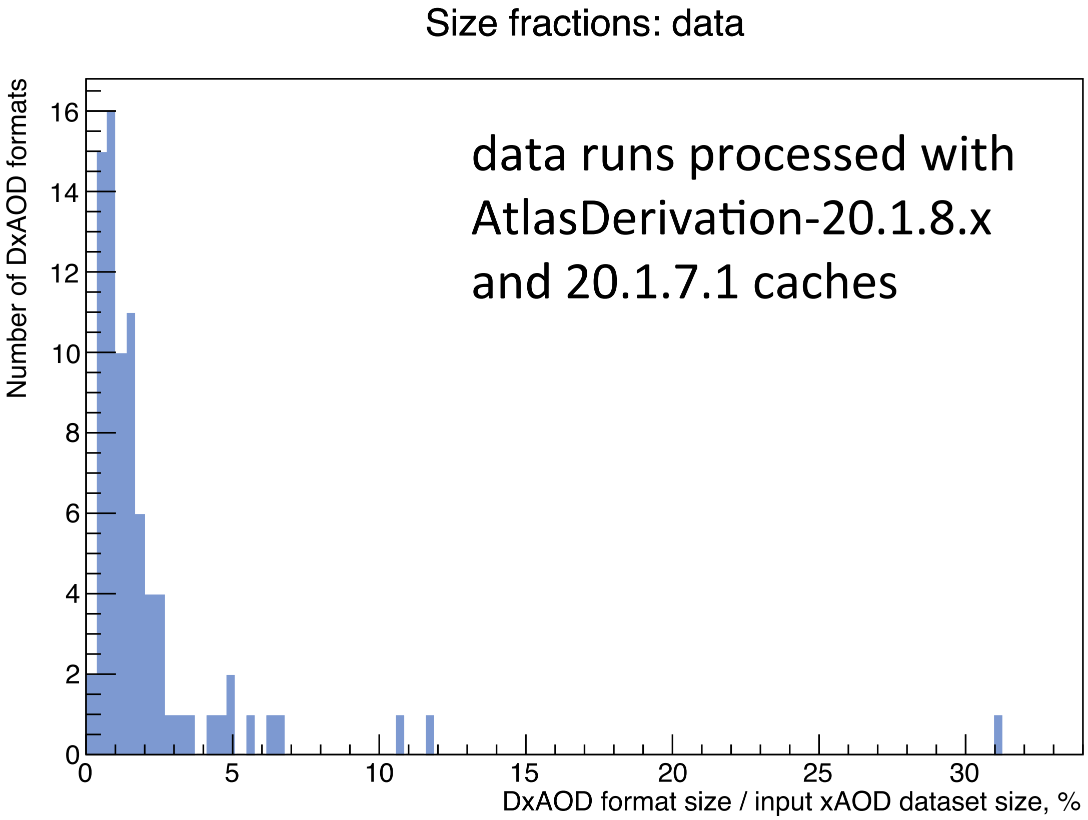


tasks on MC samples using AtlasDerivation-20.1.8.x caches

Disk Usage: Size fractions

Design budget: total DAOD size ≤ total AOD size

Target size fraction for a given derivation: ~1 %



data runs processed with AtlasDerivation-20.1.8.x and 20.1.7.1 caches

data:

Total input AOD:	0.38 TB
Total output DAOD:	0.65 TB

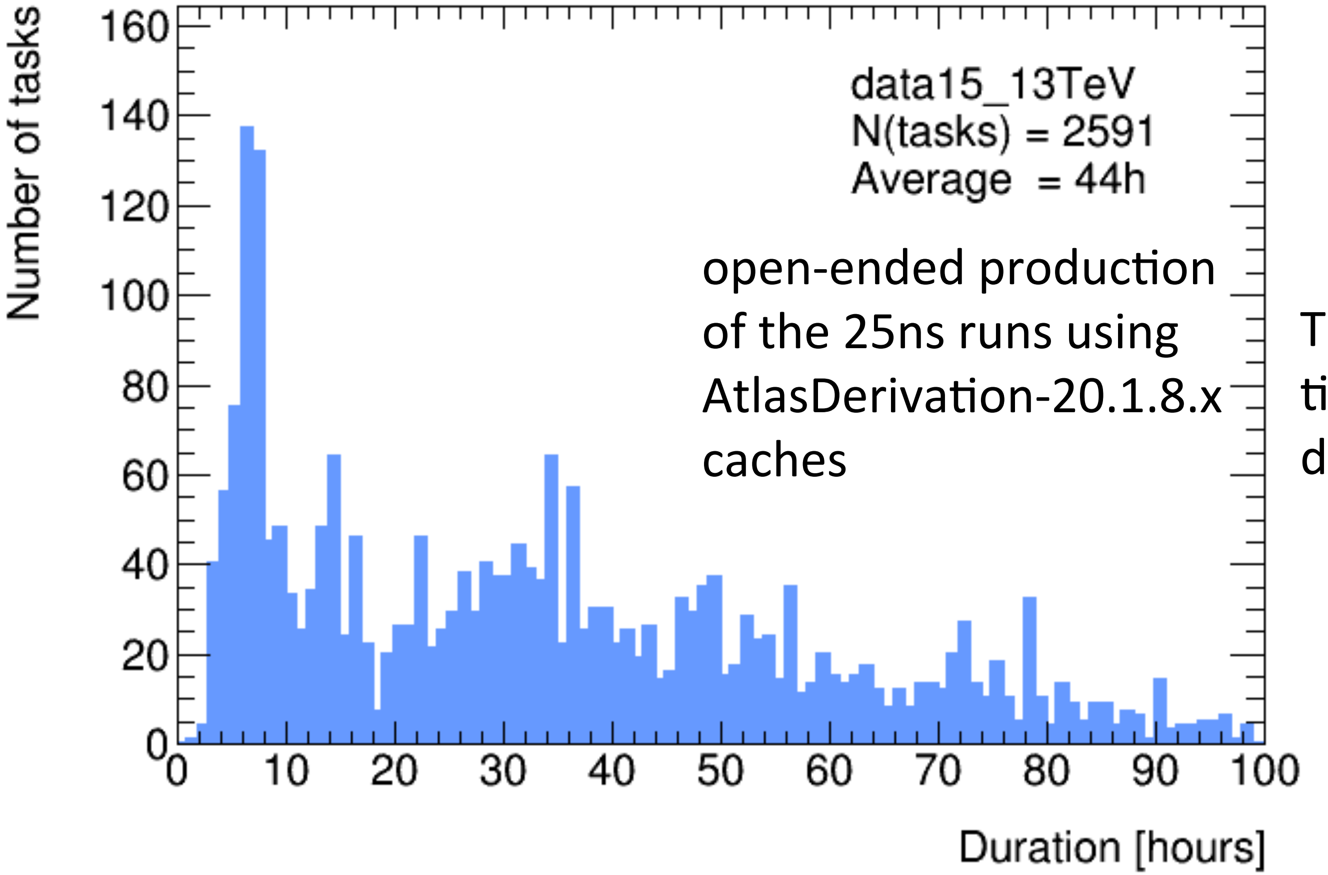
MC:

Total input AOD:	1.83 PB
Total output DAOD:	1.46 PB

- One copy of the output above.

- Review of the group derivations started in the new year to reduce the sizes before 2016 data taking starts, groups to evaluate the impact of size reductions of 20% and 40%.

Task Durations for Data Derivations



open-ended production of the 25ns runs using AtlasDerivation-20.1.8.x caches

Target turnaround time is 48h for derivations.

- 48h calibration loop after end of run
- 36-48h for Tier-0 reconstruction
- 3-4h for Tier-0 AOD merging

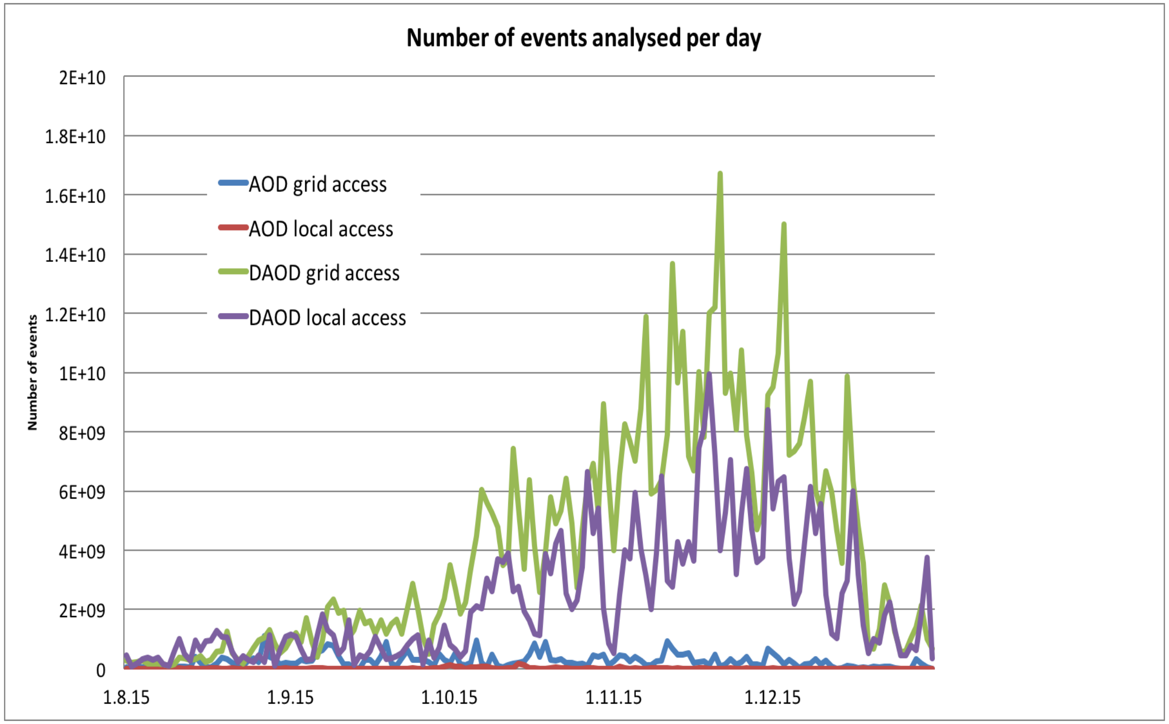
40% of the tasks complete after 24h
70% of the tasks complete after 48h
~6 days turnaround after end of run

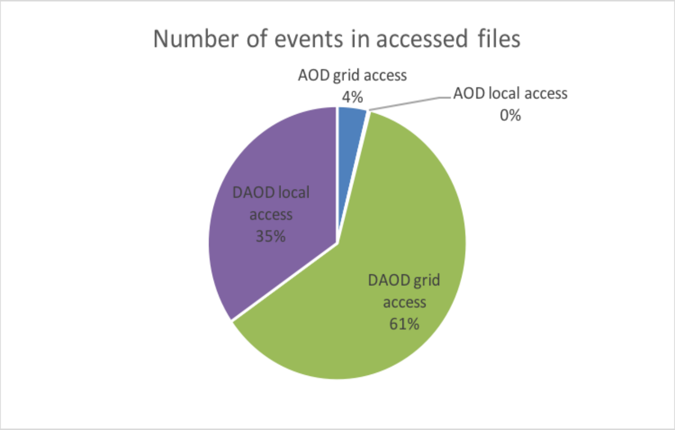
Note: Little faster for MC, 60% and 75% respectively (backup slide).

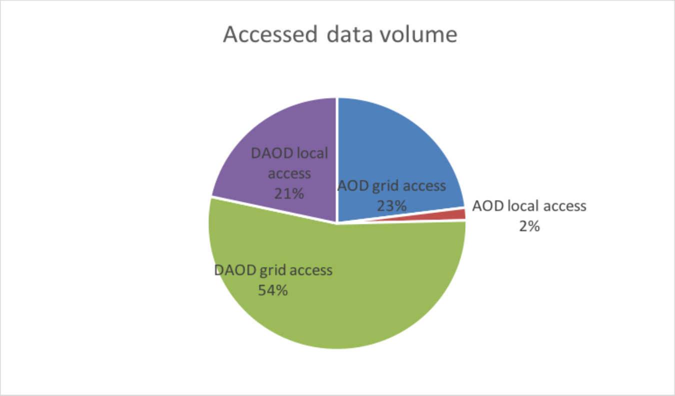
Replicas, Lifetime and Data Deletion Policy

- DAOD replication policy:
  - One initial output at Tier1 datadisk, one copy at Tier2 datadisk, one secondary copy at US Tier2 datadisk (if the primary is not there already).
- Lifetime policy:
  - Reduced to 6 months for data+MC from 24 (data)/12 (MC) months in October based on 2015 production experience, extension is 6 months. If no access after 6 months they are marked as secondary (subject to deletion when space is needed on pledged resources).
- Data deletion policy:
  - Obsoletion of old derivations (n-2 version and older or buggy) are done once a month. The list of datasets to be deleted is announced to the group derivation contacts via the e-groups and they have a week to notify if they like to keep the data, after that the corresponding tasks are marked as obsoleted and the data is deleted from all resources (Rucio-aware, including the local groupdisk).
  - Recent complaints from groups on deleting the data from local groupdisks, whether they can be kept there. Discussion at CREM, decision is to continue with the same policy and try to tighten the communication with group contacts so that they are not late for objections. Further consultation with the Physics Coordination is ongoing. Thoughts from Tier3 representatives?

DAOD vs AOD Access by Users







T. Beermann, , D. Benjamin, M. Lassnig, I. Vukotic

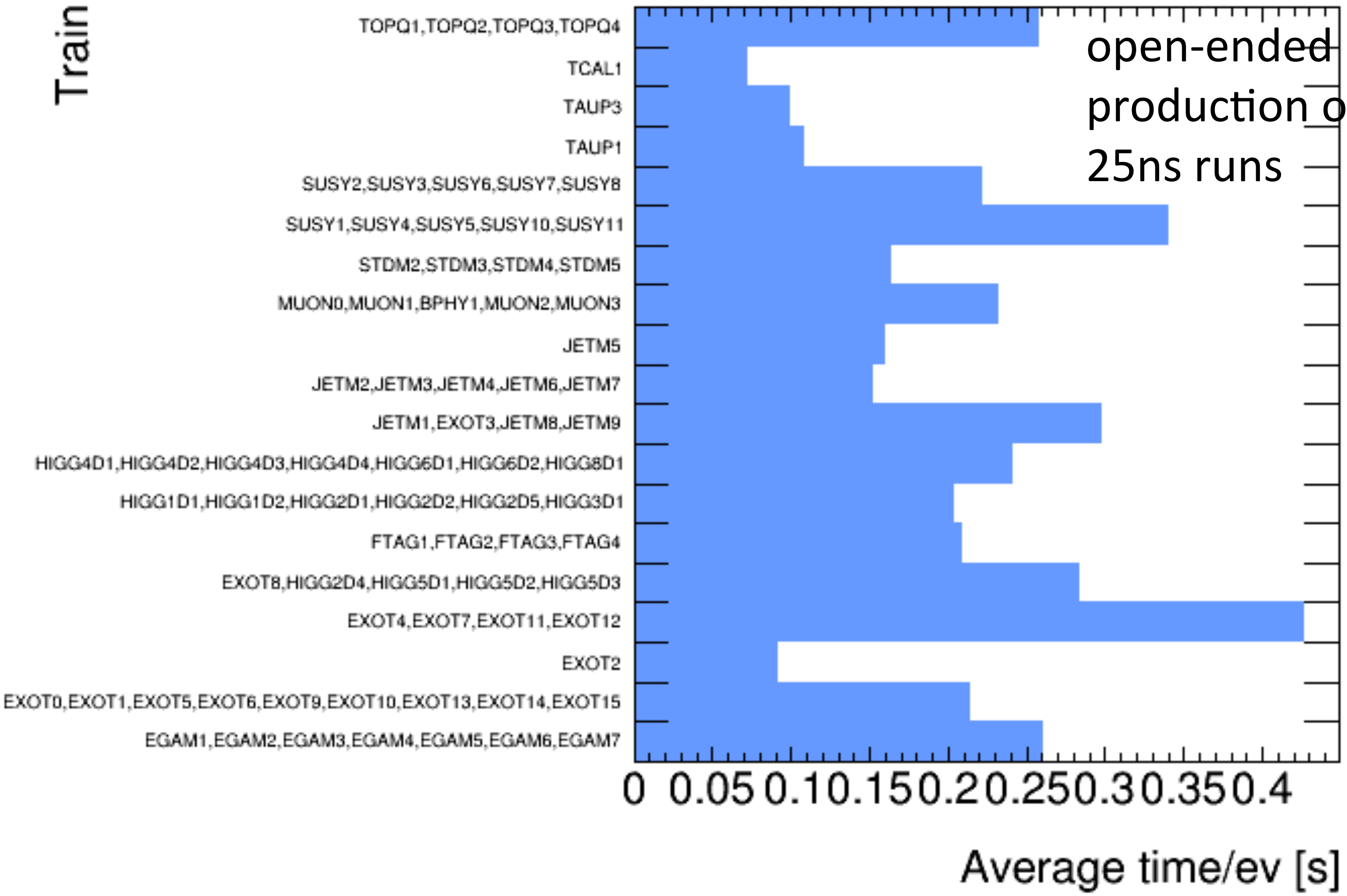
- Local access of AODs and DAODs are accounted since August 2015.
- Success of the new analysis model, 96% of the analysed events are read from DAODs, corresponding to 75% of the total data volume.

Outlook

- No change in the workflow for 2016 data taking is foreseen.
- Improvements over the past year helped to smooth operations; open-ended production on data, handing out MC derivation submissions to group contacts, major AtlasDerivation caches aligned with machine restarts after technical stops.
- Target turnaround time is 48h for derivation production. Need some effort to give an automatic push to tasks in order to finish on schedule. Site reliability is important.
- Two ongoing efforts:
    - DAOD size reduction; groups are reviewing the content of their derivations and report to derivation coordination team.
    - Submitting MC and derivation production at the same time to avoid time delays in producing the derivations (though not significant with the current system), a long standing request from the groups. Goal is to achieve this with MC15c production with the common background samples across the groups.
- Questions, comments from you for further improvements?

Backup

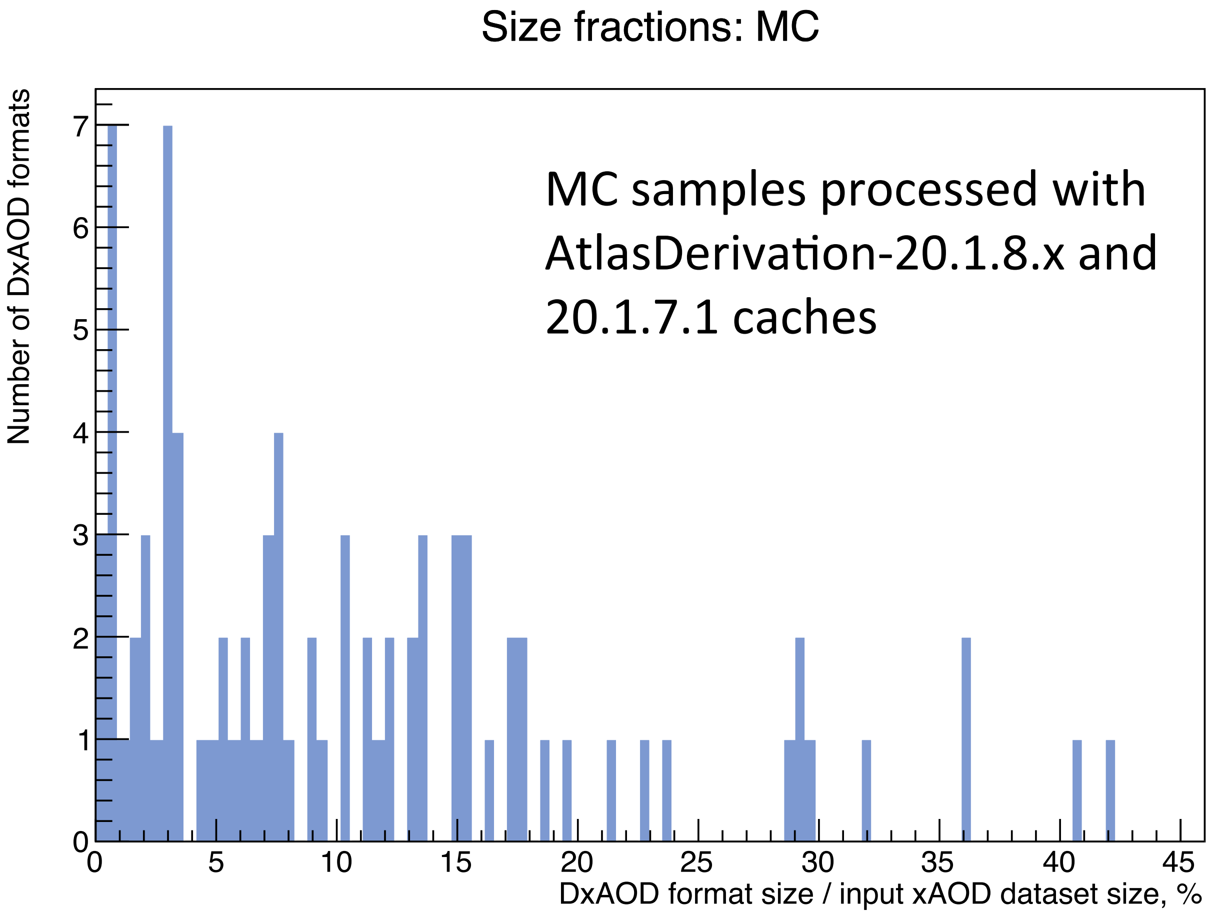
Average walltime/event per train for data



open-ended derivation production of the 25ns runs

For all trains total walltime/event = 4.17 s

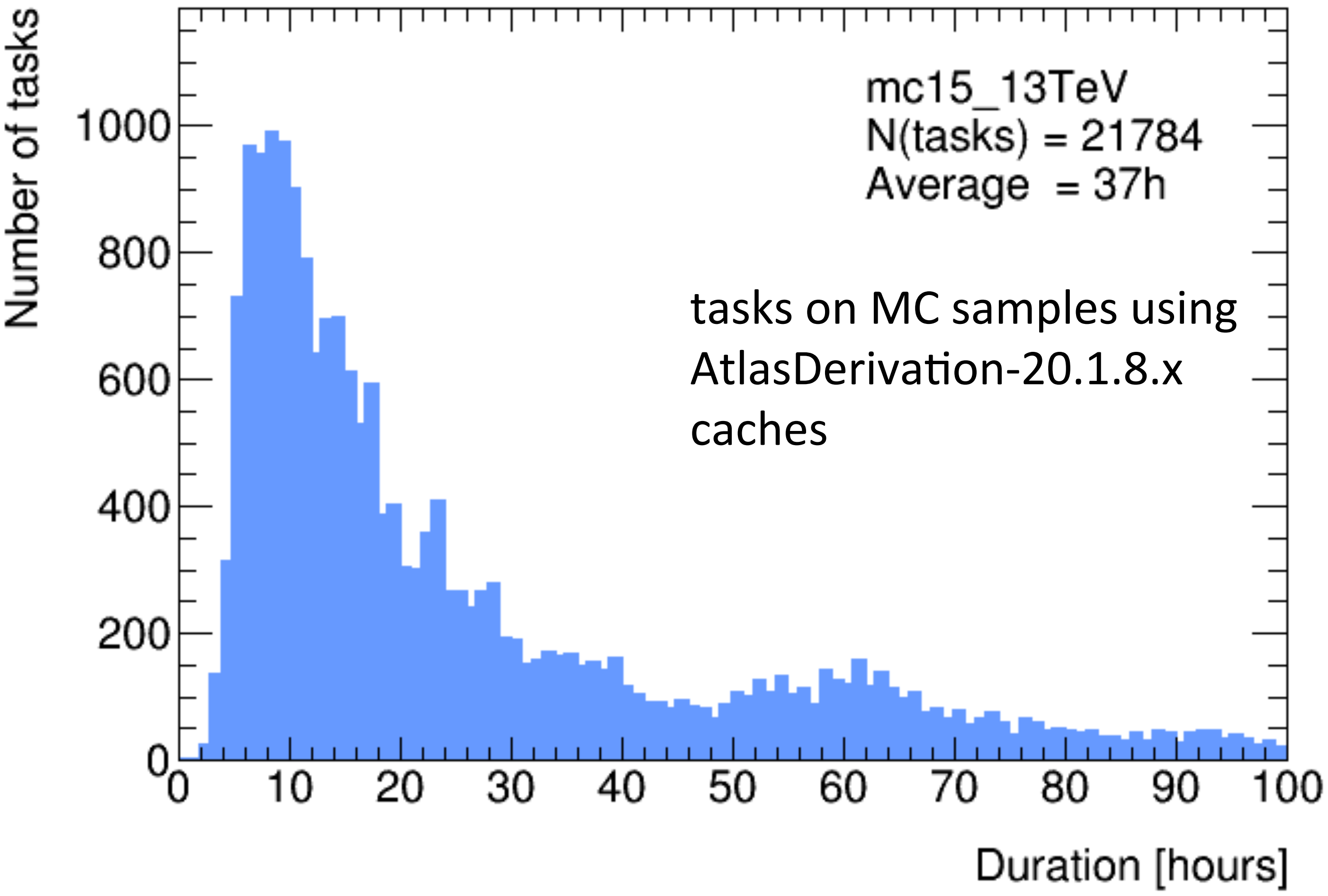
Size Fractions - MC



MC samples processed with AtlasDerivation-20.1.8.x and 20.1.7.1 caches

Total input AOD:	1.83 PB
Total output DAOD:	1.46 PB

Task durations for MC derivations

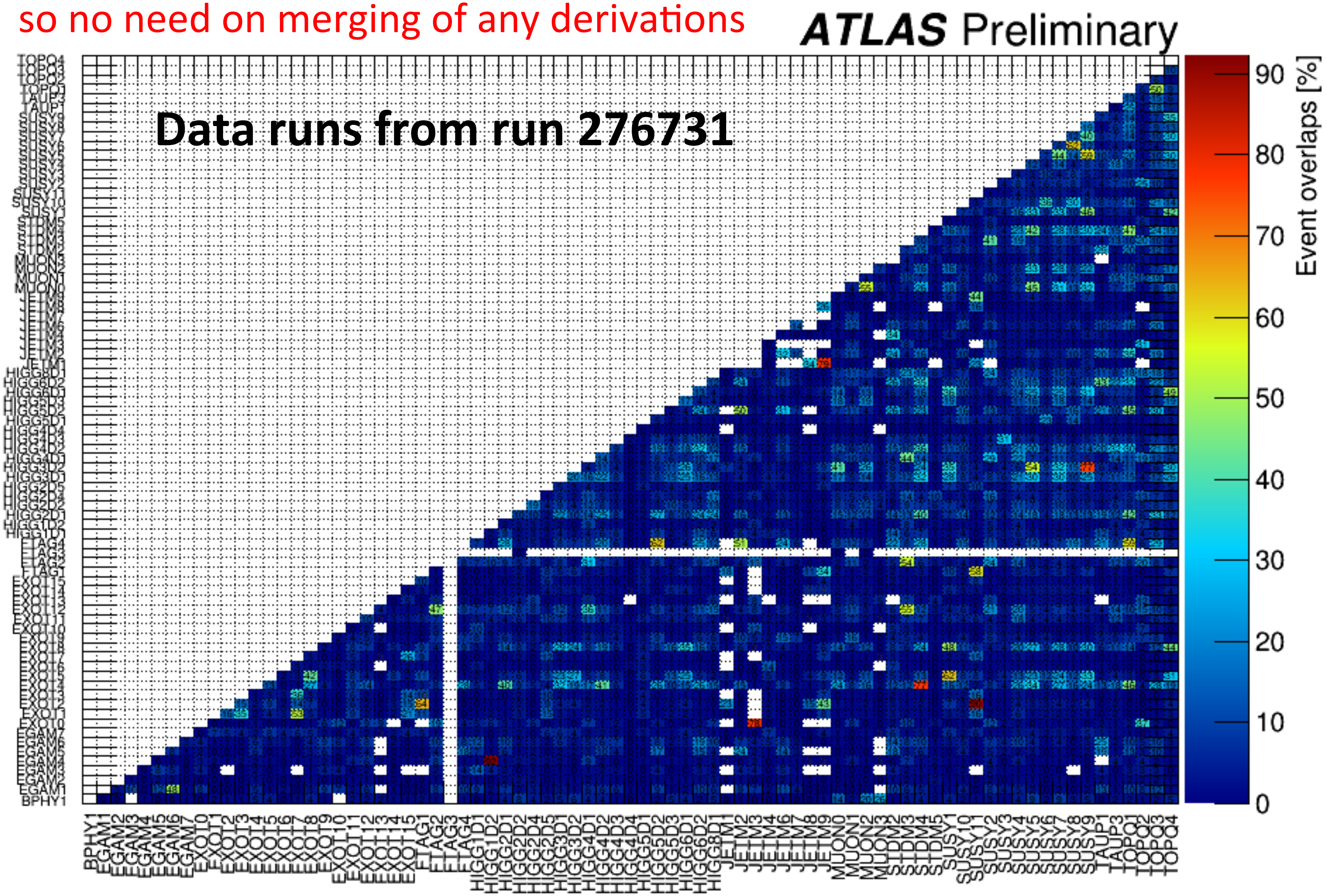


tasks on MC samples using AtlasDerivation-20.1.8.x caches

60% of the tasks complete after 24h
75% of the tasks complete after 48h

Event Overlaps in Derivations

No two formats share more than 70% of their events AND 70% of their variables so no need on merging of any derivations


ATLAS Preliminary
Data runs from run 276731
Event overlaps [%]
TOPQ4
TOPQ3
TOPQ2
TOPQ1
TAUP3
TAUP1
SUSY9
SUSY8
SUSY7
SUSY6
SUSY5
SUSY4
SUSY3
SUSY2
SUSY11
SUSY10
SUSY1
STDM5
STDM4
STDM3
STDM2
MUON3
MUON2
MUON1
MUON0
JETM9
JETM8
JETM7
JETM6
JETM4
JETM3
JETM2
JETM1
HIGG8D1
HIGG6D2
HIGG6D1
HIGG5D3
HIGG5D2
HIGG5D1
HIGG4D4
HIGG4D3
HIGG4D2
HIGG4D1
HIGG3D2
HIGG3D1
HIGG2D5
HIGG2D4
HIGG2D2
HIGG2D1
HIGG1D2
HIGG1D1
FTAG4
FTAG3
FTAG2
FTAG1
EXOT15
EXOT14
EXOT13
EXOT12
EXOT11
EXOT10
EXOT9
EXOT8
EXOT7
EXOT6
EXOT5
EXOT4
EXOT3
EXOT2
EXOT1
EXOT0
EGAM7
EGAM6
EGAM5
EGAM4
EGAM3
EGAM2
EGAM1
BPHY1
BPHY1
EGAM1
EGAM2
EGAM3
EGAM4
EGAM5
EGAM6
EGAM7
EXOT0
EXOT1
EXOT2
EXOT3
EXOT4
EXOT5
EXOT6
EXOT7
EXOT8
EXOT9
EXOT10
EXOT11
EXOT12
EXOT13
EXOT14
EXOT15
FTAG1
FTAG2
FTAG3
FTAG4
HIGG1D1
HIGG1D2
HIGG2D1
HIGG2D2
HIGG2D4
HIGG2D5
HIGG3D1
HIGG3D2
HIGG4D1
HIGG4D2
HIGG4D3
HIGG4D4
HIGG5D1
HIGG5D2
HIGG5D3
HIGG6D1
HIGG6D2
HIGG8D1
JETM1
JETM2
JETM3
JETM4
JETM6
JETM7
JETM8
JETM9
MUON0
MUON1
MUON2
MUON3
STDM2
STDM3
STDM4
STDM5
SUSY1
SUSY10
SUSY11
SUSY2
SUSY3
SUSY4
SUSY5
SUSY6
SUSY7
SUSY8
SUSY9
TAUP1
TAUP3
TOPQ1
TOPQ2
TOPQ3
TOPQ4