BROOKHAVEN
NATIONAL LABORATORY

The ATLAS Analysis Model Study Group for Run-3

Johannes Elmsheuser and several more ATLAS members

24 July 2019, BNL NPPS meeting

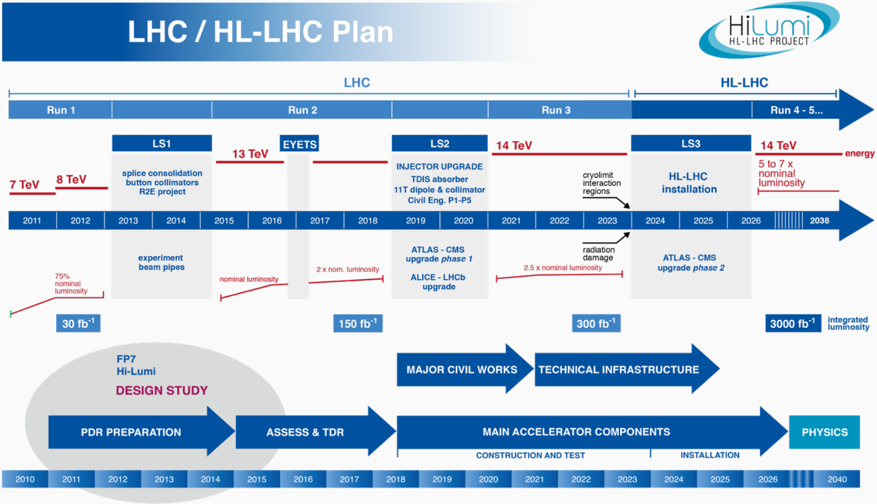
Outline

Introduction

Analysis model study group for Run3 (AMSG-R3)

AMSG-R3 recommendations

Introduction: LHC timelines



LHC / HL-LHC Plan

HiLumi HL-LHC PROJECT

LHC

HL-LHC

Run 1

Run 2

Run 3

Run 4 - 5...

LS1

splice consolidation
button collimators
R2E project

experiment
beam pipes

EYETS

LS2

INJECTOR UPGRADE
TDIS absorber
11T dipole & collimator
Civil Eng. P1-P5

ATLAS - CMS
upgrade *phase 1*

ALICE - LHCb
upgrade

LS3

HL-LHC
installation

ATLAS - CMS
upgrade *phase 2*

7 TeV

8 TeV

13 TeV

14 TeV

14 TeV

energy

5 to 7 x
nominal
luminosity

cryolimit
interaction
regions

radiation
damage

75%
nominal
luminosity

nominal luminosity

2 x nom. luminosity

2.5 x nominal luminosity

2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026 2038

30 fb^{-1}

150 fb^{-1}

300 fb^{-1}

3000 fb^{-1}

integrated
luminosity

FP7
Hi-Lumi
DESIGN STUDY

PDR PREPARATION

ASSESS & TDR

MAJOR CIVIL WORKS

TECHNICAL INFRASTRUCTURE

MAIN ACCELERATOR COMPONENTS

CONSTRUCTION AND TEST

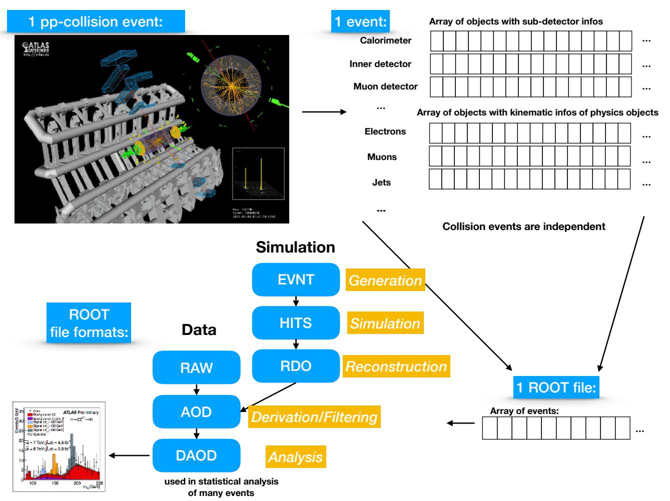
INSTALLATION

PHYSICS

2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026 2040

- Note that $\sqrt{s}$ in Run3 is still uncertain and depends on magnet training in 2021

Introduction: Simplified data analysis workflow for ATLAS



1 pp-collision event:

1 event:

Array of objects with sub-detector infos

Calorimeter ...

Inner detector ...

Muon detector ...

...

Array of objects with kinematic infos of physics objects

Electrons ...

Muons ...

Jets ...

...

Collision events are independent

Simulation

EVNT – *Generation*

HITS – *Simulation*

RDO – *Reconstruction*

ROOT file formats:

Data

RAW

AOD – *Derivation/Filtering*

DAOD – *Analysis*

used in statistical analysis of many events

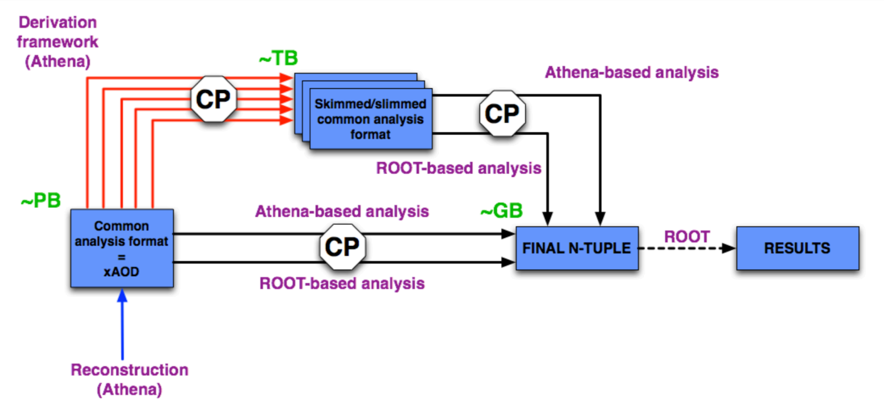
1 ROOT file:

Array of events: ...

In essence: several steps of data processing and then data reduction

First parts on Grid/Cloud/HPC - last step usually on local resources

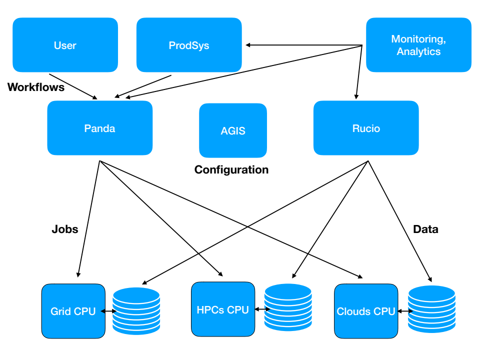
ATLAS Run2 analysis workflows



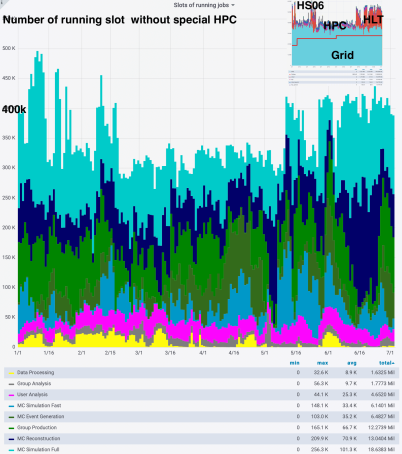
ATLAS DISTRIBUTED COMPUTING OVERVIEW

The ATLAS distributed computing system is centered around:

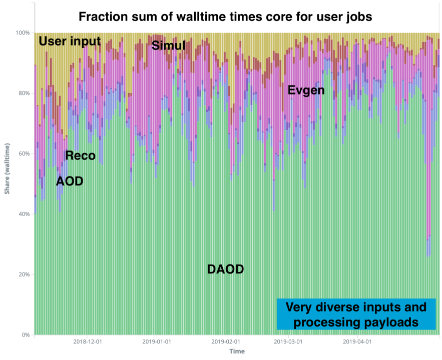
- **Workflow management system**: PanDA
- **Data management system**: Rucio
- Many **additional components**: AGIS, ProdSys, Analytics, ...
- **Resources**: WLCG grid sites, Tier0, HPCs, Boinc, Cloud
- **Shifters**: Grid, Expert and Analysis (ADCoS, CRC, DAST)



CPU resource usage (2019) and analysis input (2019)



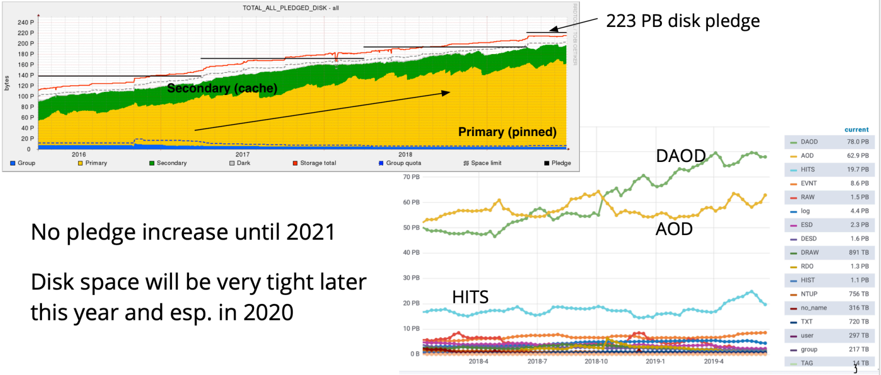
	min	max	avg	total▲
Data Processing	0	32.6 K	8.9 K	1.6325 Mil
Group Analysis	0	56.3 K	9.7 K	1.7773 Mil
User Analysis	0	44.1 K	25.3 K	4.6520 Mil
MC Simulation Fast	0	148.1 K	33.4 K	6.1401 Mil
MC Event Generation	0	103.0 K	35.2 K	6.4827 Mil
Group Production	0	165.1 K	66.7 K	12.2739 Mil
MC Reconstruction	0	209.9 K	70.9 K	13.0404 Mil
MC Simulation Full	0	256.3 K	101.3 K	18.6383 Mil



Very diverse inputs and processing payloads

- 10-20% of analysis share on the Grid/Cloud - not HPC - mainly single core serial processing payloads
- Very diverse inputs and processing payloads in analysis
- In addition lots of final analysis happens on local batch farm or computers on individual ntuples

ATLAS DISK SPACE EVOLUTION

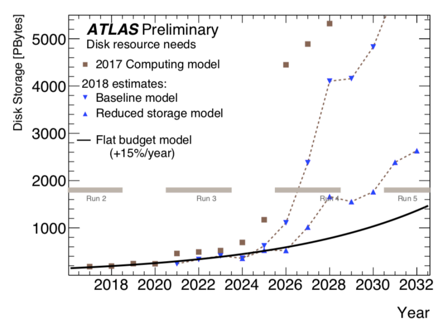


No pledge increase until 2021

Disk space will be very tight later this year and esp. in 2020

- Mainly Analysis formats on DISK (AOD/DAOD)
- Only 1-2 replicas possible because of large sample sizes
- Many event duplication from AOD to DAOD
- In addition TAPE ≈ 253 PB used and pledge of 315 PB

ATLAS disk space projections



Run3: Initial assumption resources will be: 1.5 × (resources in 2018) Consistent with ”flat budget”

Outline

Introduction

Analysis model study group for Run3 (AMSG-R3)

AMSG-R3 recommendations

AMSG-R3 GROUP MANDATE AND DOCUMENTATION

- Analysis Model Study Group for Run3 (AMSG-R3) was setup last autumn consisting of ≈ 10 persons in consultation with many domain experts
- Concluded last month with a document and set of recommendations
- Mandate in essence:

  Collect options to save at least 30% disk space overall (for the same data/MC sample), harmonise analysis and give directions for further savings for the HL-LHC.
- Presentation at CHEP19 about AMSG-R3 recommendation and current status

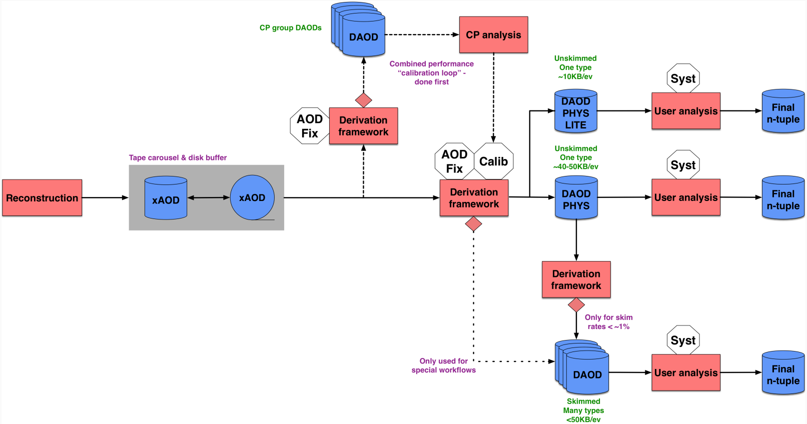
Outline

Introduction

Analysis model study group for Run3 (AMSG-R3)

AMSG-R3 recommendations

NEW PRODUCTION WORKFLOWS AND FORMATS



DAOD_PHYS:
50 kB/event, combined single DAOD format (for MC, but also DATA)

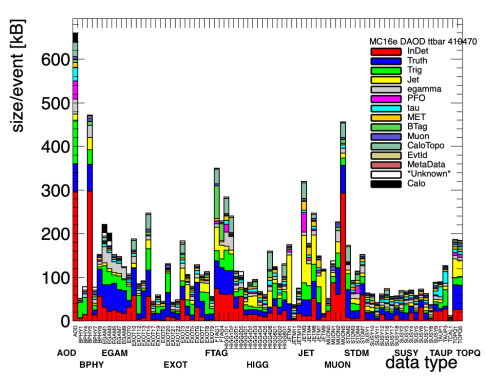
DAOD_PHYSLITE:
10 kB/event, very condensed and calibrated objects, very important for HL-LHC

today's DAODs:
Significantly reduce number of today's DAODs

AODs:
Larger fraction only available on TAPE

AOD/DAOD CONTENT REDUCTION

MC16e ttbar 410470, 79 DAODs, 1 AOD, AMI tag
e6337_e5984_s3126_r10724_r10726_p3654



Tracks/InDet

- tracks selection criteria for $<\mu>\approx 60$
- track covariance matrix: drop elements in the DAOD, use lossy compression
- split into 2 categories: tracks associated to primary vertex and not - store less detail for PU tracks

Truth

- remove any duplication in MC truth records
- enforce TRUTH3 in physics DAODs

Trigger

- AODRun3_Large (wish 50 kB) and AODRun3_Small (wish 5 kB, for MC)
- Introduce dedicated DAOD_Trigger

Lossy Compression:
use lossy float compression of variables where physics allows this

Summary of the AMSG-R3 recommendations

Formats	Introduce **DAOD_PHYS** with ∼50 kB/event
	Introduce **DAOD_PHYSLITE** with ∼10 kB/event and **calibrated objects**
	Reduce number DAODs formats, use these for CP, systematic and R&D studies
Production	Stop open-ended production for data DAODs
	Use a **tape carousel model for AOD** inputs in parts of the DAOD production
	Consider caps on sizes of individual DAOD type datasets
	Bring Rucio redirector with global name space into production
	Smart DAOD replica placement on the grid sites
	Increase usage of docker/singularity containers for analysis and group ntuple production
	Central skimming of DAOD_PHYS into physics DAODs will still be offered
AOD/DAOD content	Significantly **reduced track, trigger, truth** information, use **calibrated objects**
	Apply **lossy compression** for most variables in AOD/DAODs where feasible and applicable
	Avoid any information duplication in the AOD/DAODs containers

Simple disk space model with Run2 numbers

- Simple model of Run2 AOD+DAODs: 131.9 PB
- One possible model using Run2 numbers:
  - 4 DAOD_PHYS+DAOD_PHYSLITE (MC+DATA) replicas
  - 0.5 AOD replica (aka TAPE buffer)
  - 50% of today's MC+DATA DAOD

	MC				Data			
	AOD	DAOD	DAOD PHYS	DAOD PHYS LITE	AOD	DAOD	DAOD PHYS	DAOD PHYS LITE
events	$3 \cdot 10^{10}$	$1 \cdot 10^{11}$	$3 \cdot 10^{10}$	$3 \cdot 10^{10}$	$2 \cdot 10^{10}$	$1 \cdot 10^{11}$	$2 \cdot 10^{10}$	$2 \cdot 10^{10}$
size/event [kB]	600	100	70	10	400	50	40	10
disk space [PB]	18.0	10.0	2.1	0.3	8.0	5.0	0.8	0.2
other versions	1.5	2	2	2	1.5	2	2	2
repl. fac.	0.5	1	4	4	0.5	2	4	4
Sum [PB]	13.5	20.0	16.8	2.4	6.0	20.0	6.4	1.6

- Sum: 85.1 PB, Potential saving: 45.9 PB

Summary and Conclusions

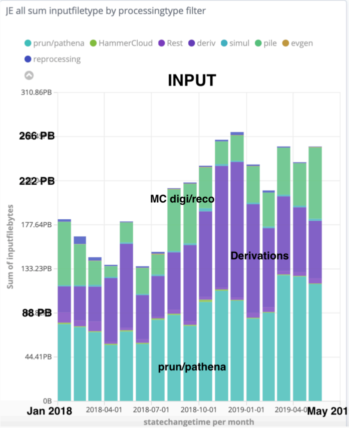
- AMSG-R3 note with recommendations available and finished
- DAOD_PHYS prototype is available and collecting feedback from different physics groups
- DAOD_PHYSLITE very important for HL-LHC, but urgently have to find new developers
- Lossy compression interesting additional way to shrink format sizes - latest ROOT 6.18.00 offers truncation options for TLeafF16/TLeafD32 (see link)
- Additional work has to be carried out by analysis software, trigger and combined performance groups

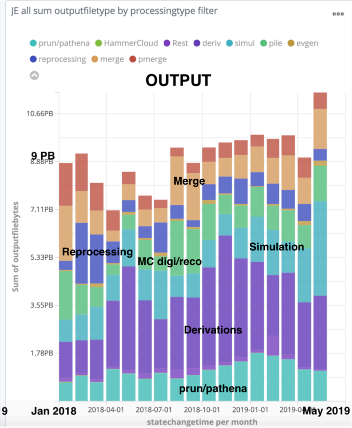
BACKUP

”blind” lossy compression with $t\bar{t}$ MC file

	DAOD_PHYS			DAOD_PHYSLITE			AOD		
	Compr.	Default	Ratio	Compr.	Default	Ratio	Compr.	Default	Ratio
	[kB]	[kB]		[kB]	[kB]		[kB]	[kB]	
MetaData	0.23	0.23	1.00	0.18	0.18	1.00	1.14	1.16	0.99
BTag	0.97	0.98	0.99	0.08	0.08	1.00	7.74	9.20	0.84
Muon	1.43	1.73	0.83	0.47	0.47	1.00	14.17	17.59	0.81
Truth	1.91	2.80	0.68	2.37	2.52	0.94	43.56	61.04	0.71
PFO	2.35	3.01	0.78				33.69	44.61	0.76
EvtId	1.93	3.07	0.63	1.77	1.76	1.00	1.56	2.10	0.74
tau	4.03	6.11	0.66	2.06	3.76	0.55	25.36	37.85	0.67
MET	7.35	7.42	0.99	3.45	3.44	1.00	12.70	13.16	0.96
egamma	5.31	8.22	0.65	0.15	0.15	1.00	30.16	41.61	0.72
Jet	9.62	12.00	0.80	0.76	0.76	1.00	15.78	20.85	0.76
Trig	42.52	47.15	0.90	33.23	33.20	1.00	132.32	165.25	0.80
InDet	35.70	58.20	0.61	0.60	0.60	1.00	193.43	307.24	0.63
CaloTopo				0.45	0.45	1.00	24.89	35.01	0.71
Calo							18.06	18.07	1.00
Analysis jet/e/μ/τ/γ				1.72	2.26	0.76			
Total	113.32	150.92	0.75	47.27	49.63	0.95	554.94	775.25	0.72
Total-Trig	70.80	103.77	0.68	14.04	16.43	0.85	422.63	609.99	0.69

Processing input and output volumes PanDA in past 17 months





- Grid **input** processing volume ≈200-250 PB/month - 30-50% derivation production, 30-50% analysis
- Copied to worker node - files might be accessed multiple times on the worker node (digi-reco)
- Grid **output** volume: ≈ 8-9 PB/month of which 2-5 PB/month derivation production
- Tier0 batch is not included here and adds to the input/output volumes