Run-3 offline data processing and analysis at LHCb

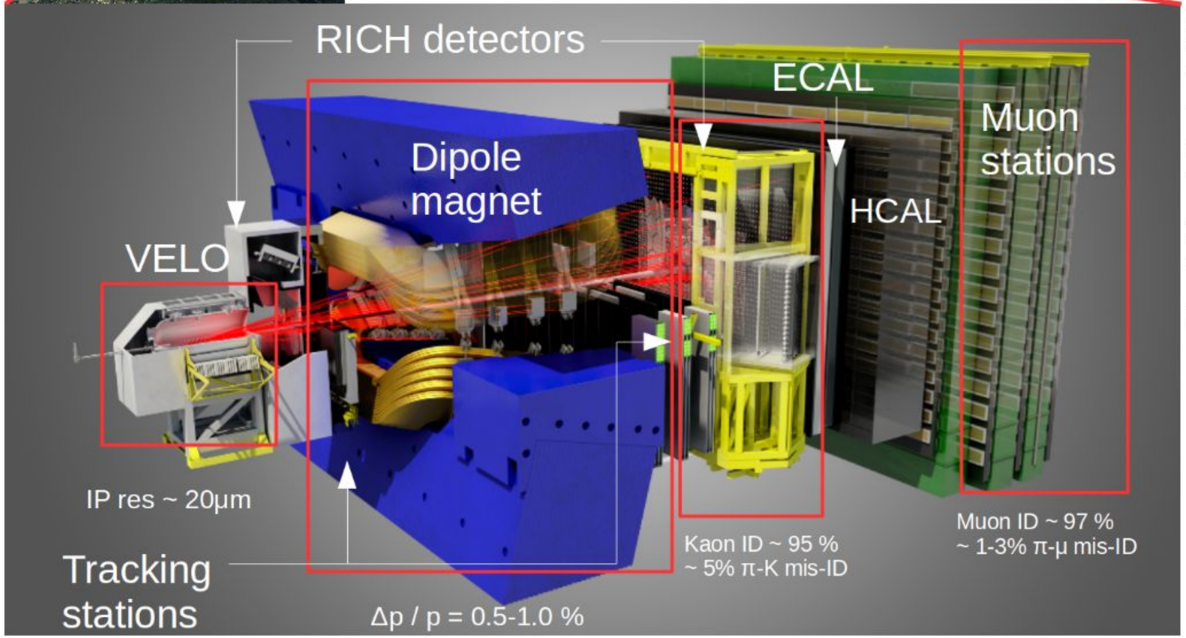
N. Skidmore on behalf of the LHCb collaboration

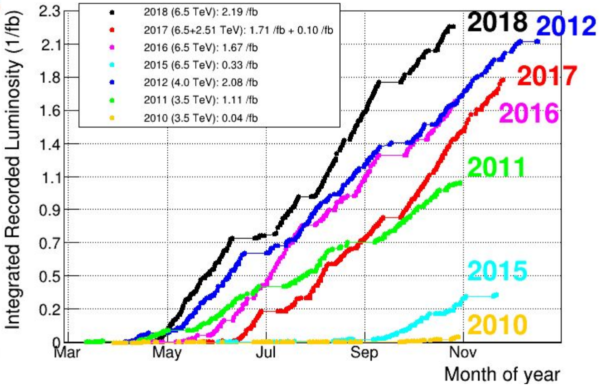
EPS-HEP2021 conference

July 2021

LHCb in run 1 + 2

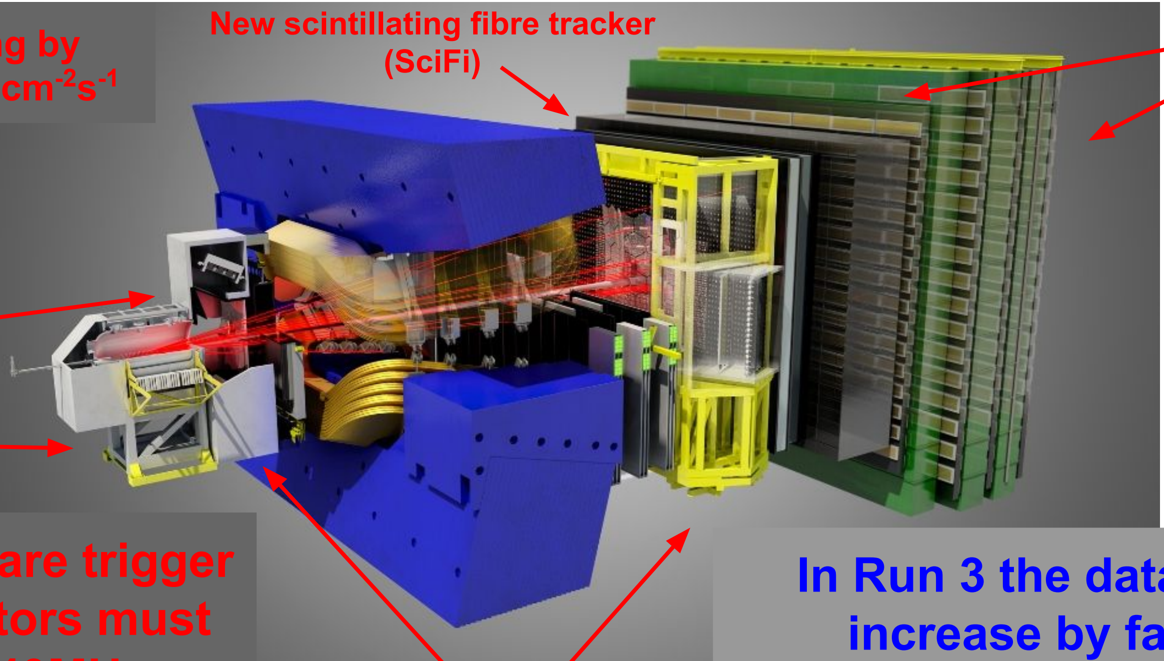
LHCb is a forward arm spectrometer specialising in decays of beauty and charm hadrons





9 fb^{-1} collected in run 1 + 2
$>10^{12}$ bb pairs in acceptance

LHCb Upgrade I - Hardware **and** software Upgrade



Inst. Lumi. increasing by factor 5 up to $2 \times 10^{33} cm^{-2} s^{-1}$

Target Int. Lumi. for Run 3 $50 fb^{-1}$

LHCB-TDR-12
LHCB-TDR-13
LHCB-TDR-14
LHCB-TDR-15
LHCB-TDR-16
LHCB-TDR-17
LHCB-TDR-18

New fully software trigger - all sub-detectors must read out at 40MHz HLT1 to run on GPUs

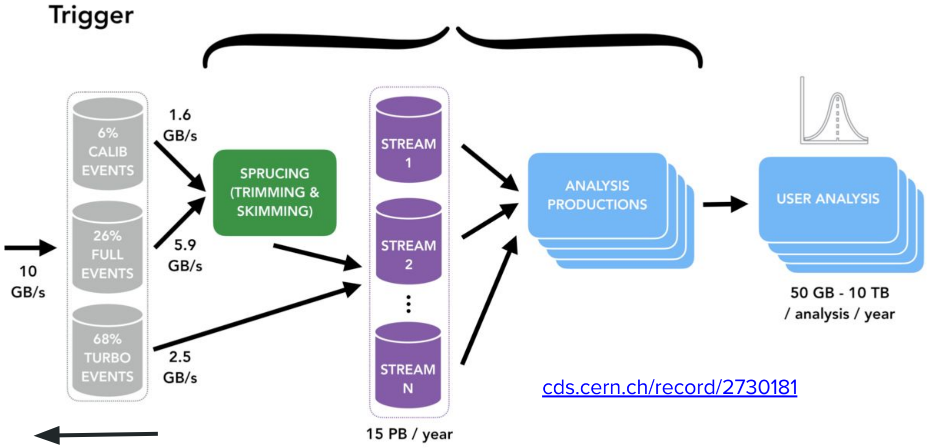
In Run 3 the data rate will increase by factor 30 compared to run 2...

See talk by Marianna

Data Processing and Analysis (DPA) in Run 3

- Increased data rate in Run 3 poses significant Offline data processing and analysis challenges
- Coordination of these activities by DPA project
    - Software project on same level as detector projects

DPA remit



cds.cern.ch/record/2730181

DPA project
Offline processing/selections/analysis

- WP1 - Sprucing
- WP2 - Analysis Productions
- WP3 - Offline Analysis Tools
- WP4 - Innovative Analysis Techniques
- WP5 - Legacy Software & Data
- WP6 - Analysis Preservation & Open Data

Six work packages

See talk by Peilian

DPA work packages

WP1: Sprucing

Offline, central data skimming and slimming

Sharing of HLT2 framework

Ensemble of “Sprucing selections” from physics WGs run concurrently with data and during EoYS

WP2: Analysis Productions

Centralised nTuple production using DIRAC production system

Maximal automation

Inbuilt testing/validation and analysis preservation

WP3: Offline analysis tools

Offline analysis application sharing HLT2/Sprucing tools

Modern/flexible design to be used with AnaProds

Thread safe application

WP4: Innovative analysis

R&D for innovative analysis techniques to be adopted in the future by LHCb

Quantum computing

WP5: Legacy data & software

Continued re-stripping (slim/skim) of legacy Run 1+2 data

Maintenance of legacy software stacks for Run 1+2 data

WP6: Analysis preservation and Open data

Release of Run 1 LHCb data to CERN Open Data portal

Guidelines and tools for analysis preservation

LHCb use of CERN’s CAP and REANA

WP1 : Sprucing

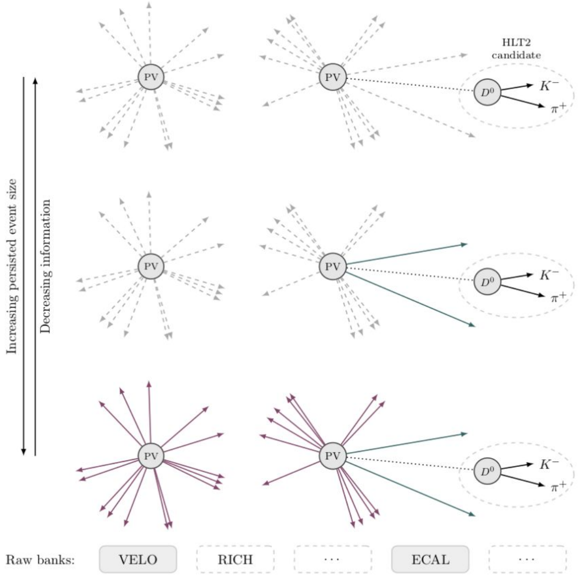
In Run 3 event persistency is customisable depending on the Physics involved

(2018 event size)

TURBO
- 7 kB

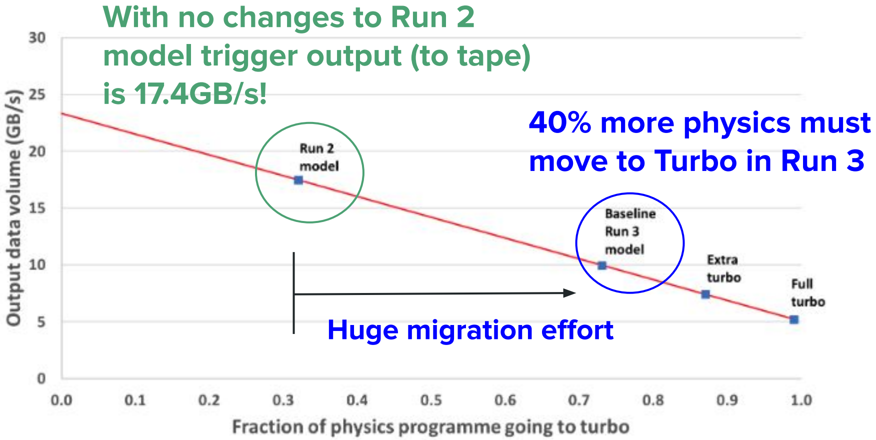
TURBO SP
- 16 kB

FULL/TURCAL
- 48 kB
- 69 kB



In Run 2 - 32% of physics events went to TURBO

stream	event size (kB)	event rate (kHz)	rate fraction	throughput (GB/s)	bandwidth fraction
FULL	70	7.0	65%	0.49	75%
Turbo	35	3.1	29%	0.11	17%
TurCal	85	0.6	6%	0.05	8%
total	61	10.8	100%	0.65	100%



WP1 : Sprucing

Output data volume	Turbo physics fraction	Physics channels left in FULL
10 GB/s	73%	EW, high PT, (semi)leptonic and some hadronic B-physics, leptonic charm decays and general LFV searches
7.5 GB/s	87%	EW, high PT, some leptonic B-physics, some LFV searches and leptonic searches
5 GB/s	99%	None

Baseline

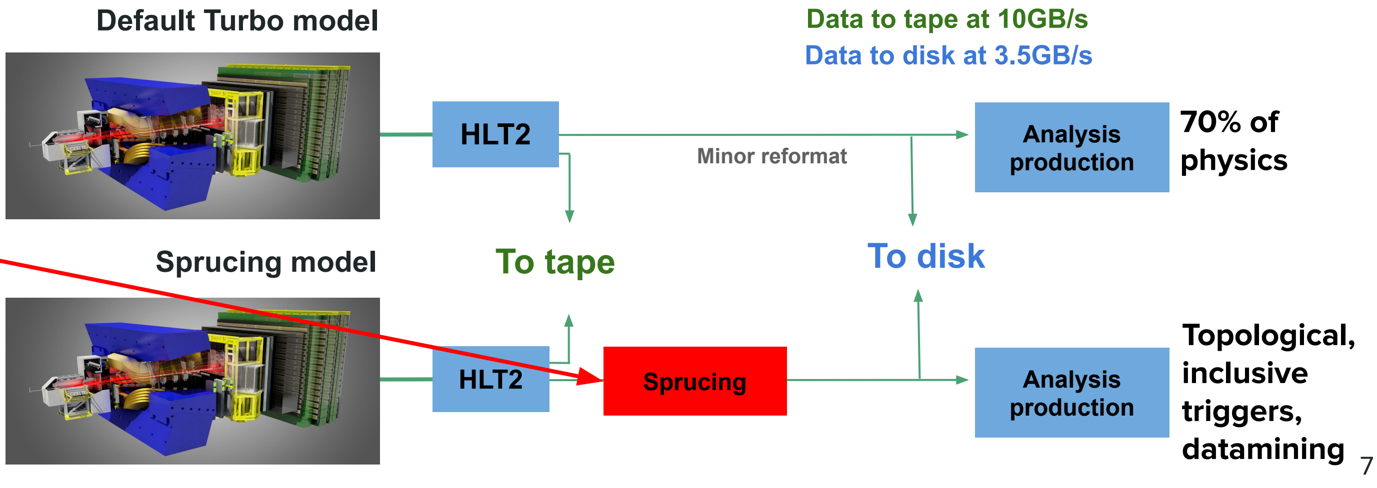
For physics left in FULL stream - Sprucing:

- Utilise cheap **tape storage** for bulk of bandwidth (FULL stream)
- Perform central offline slimming/skimming

Sprucing

A further offline stage of data reduction/selection between tape and disk storage when HLT2 line throughput is too large to go straight to disk.

Utilise same selection framework as HLT2



WP2 : Analysis productions (AnaProd)

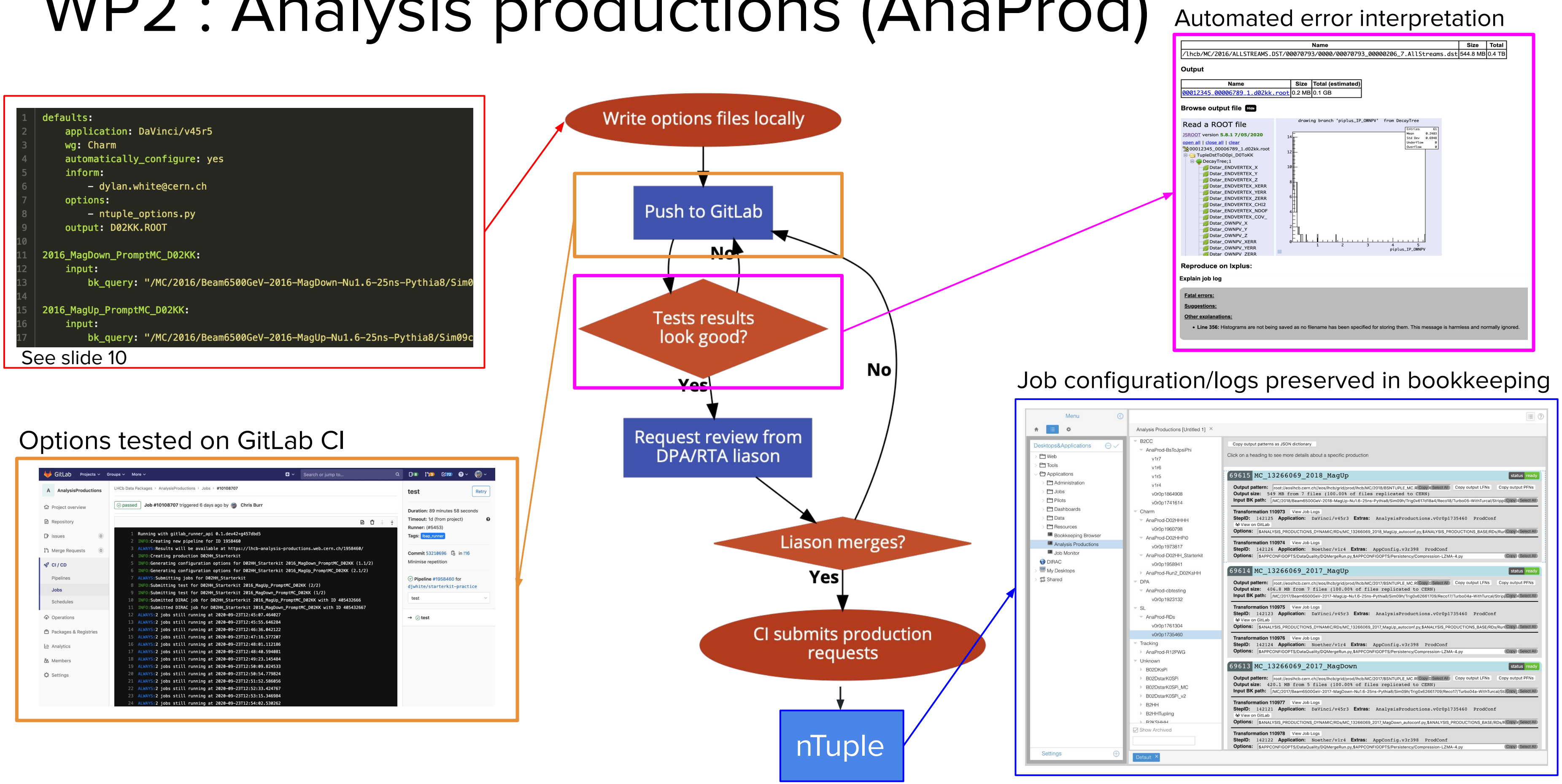
In Run 1 + 2 analysts created nTuples individually from data on disk using Ganga… does not scale well for Run 3

- 1000s of faulty jobs can be submitted instantly (10% of user jobs fail)
- Time consuming - O(weeks) for Run 1 + 2 tuples - failed jobs re-submitted manually by user
- No analysis preservation infrastructure

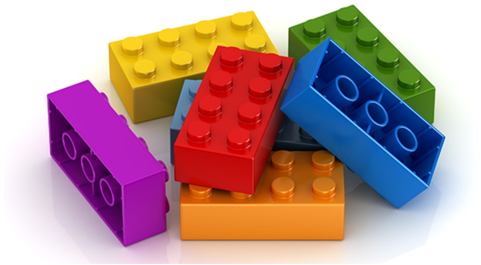
Analysis productions submit nTupling jobs centrally using DIRAC transformation System (AnaProds already in use for legacy data)

- Does not require analyst to babysit grid jobs
- Options tested automatically upon push to GitLab (CI). Final approval must be given from PAWG liaisons
- Job details/configuration/logs automatically preserved in LHCb bookkeeping/EOS
- Automated error interpretation/advice
- Results displayed on webpage

WP2 : Analysis productions (AnaProd)



WP3 : Offline analysis tools



In Run 1+2 nTuples used “TupleTools” from DaVinci (user analysis) application

- “TupleTools” create and save variable branches for typical use cases eg. TupleToolTrackInfo
- Very easy to implement but adds lots of redundant branches - can easily save 500+ variables
- 500GB - 10TB of data for a single Run 1+2 analysis - nTuples tend to be only used for one analysis

For Run 3 complete redesign of DaVinci framework - **FUNTUPLE**

- Same thread-safe functors as in HLT2/Sprucing used to create light-weight nTuples
  - Consistency between Online and Offline selections/tools/algorithms
- Analyst has full control over exactly which variables for which particles are persisted in nTuple
- AnaProds will run analysts’ DaVinci options

Define particle

```
'''
Tuple observables related to Jpsi -> mu+ mu-
'''
#FunTuple: make branches to tuple
branches = {}
branches['Jpsi'] = 'J/psi(1S) -> mu+ mu-'
branches['MuPlus'] = 'J/psi(1S) -> ^mu+ mu-'
```

Define variables

```
#FunTuple: make collection of functors for Jpsi
variables_jpsi = FunctorCollection({
    'LOKI_P':'P',
    'LOKI_MAXPT':'TRACK_MAX_PT',
    'LOKI_NTRCKS_ABV_THRSHLD':'NINTREE(ISBASIC & (PT > 15*MeV))',
    'LOKI_Muonp_PT':'CHILD(PT, 1)',
    'LOKI_Muonm_PT':'CHILD(PT, 2)',
    'THOR_PT':F.PT
})
```

WP4 : R&D Innovative analysis techniques

Think tank for innovative analysis techniques and exploitation of new analysis facilities with heterogeneous computing resources (GPU/CPU/FPGA)

Worldwide LHC Computing Grid (WLCG) consists of ~ 1M CPU cores over 170 sites

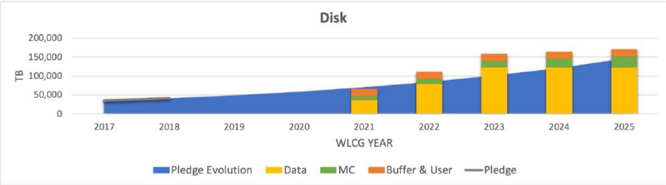
- Most sites have no GPUs yet - push towards High Performance Computing (HPC) centers providing large GPU resources
- Potential to utilise LHCb's HLT1 GPU farm during detector downtime

Need development such that significant LHCb payloads can run on GPUs

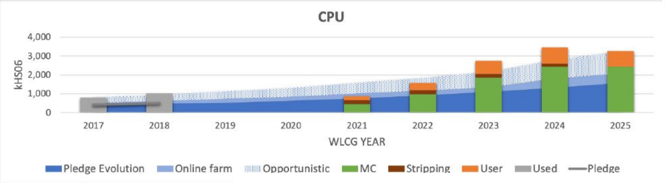
- Use advanced algorithms such as Generative Adversarial Networks (GANs) to train models describing LHCb sub-detector response - GPUs speed up GAN training - *Ultra-fast simulation*
- Users using GPUs for analysis, e.g. TensorFlow for model fitting (Zfit) - particularly for complex amplitude analysis models with large statistics

First investigations into use of Quantum Machine Learning for jet tagging (see backup)

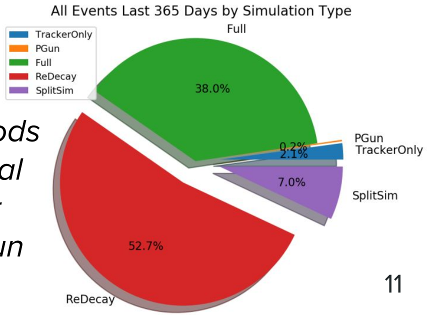
In Run 3 LHCb will produce ~ 15PB of data on disk per year



Real data will dominate disk storage but simulation will dominate CPU needs - 90% of total offline CPU resources



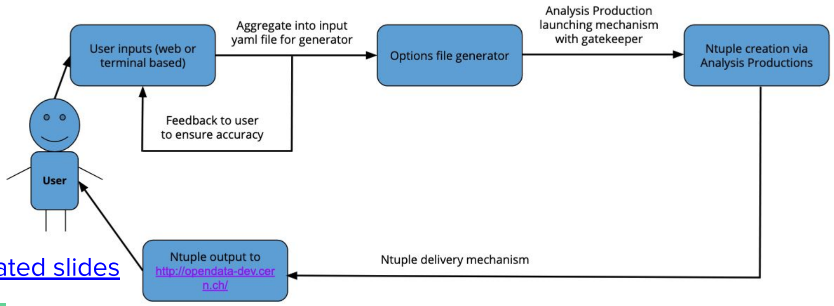
Impressive adoption of fast simulation methods at LHCb. Essential until full detector simulation can run on GPUs



WP6: Analysis preservation and open data

Run 1 LHCb data to be released on [CERN Open Data portal]

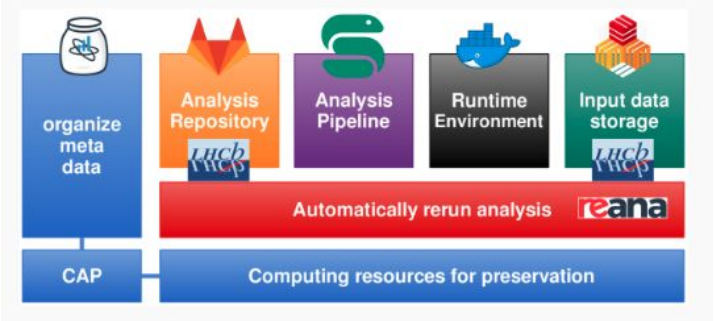
- Development of Open Data nTuple wizard
  - Auto-generates options from intuitive user input - no knowledge of LHCb software required
  - Launches AnaProd (See slide 8-9)
  - Returns nTuple to user
- Much smaller storage and bandwidth requirements on Open Data Portal
- Will be used by collaboration members to analyse Run 1+2 data



[Dedicated slides]

Analysis preservation has been in place at LHCb since 2017 - building on this

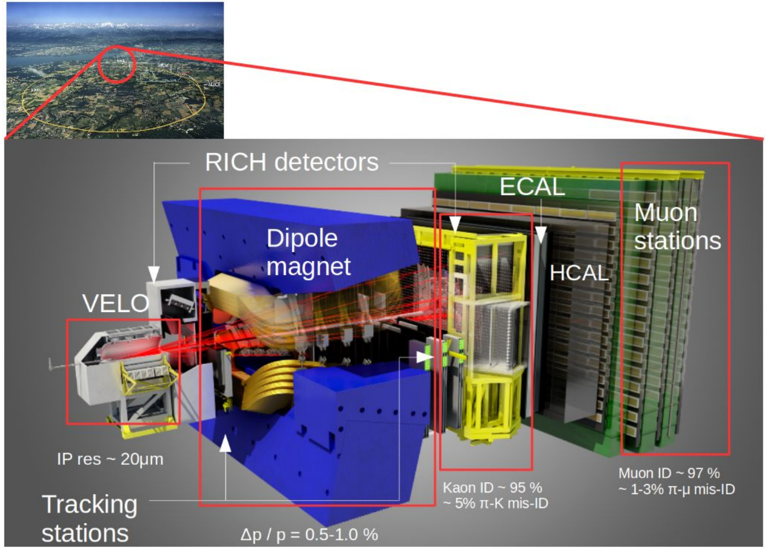
...analysis code should be preserved in a long-term archive such as a physics analysis gitlab group, the input ntuples should also be preserved in a long-term archive (EOS) and sufficient documentation to enable a (technically competent) LHCb member to run the code in a standard environment such as lxplus should be included with the code.

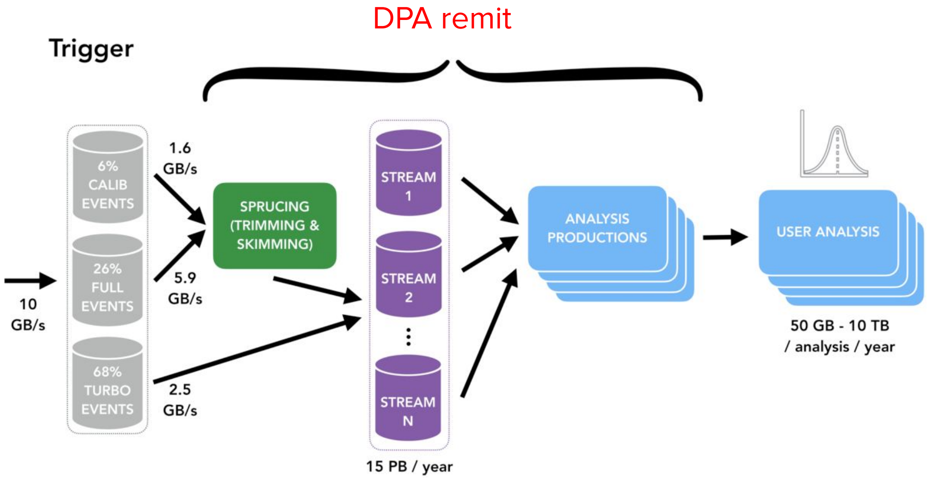


- Developing tag-based access to AnaProd output removing need for copying files/hard coding paths
- Support for snakemake (significant adoption in LHCb) in REANA

Summary

- LHCb will have to process data offline an order of magnitude larger than in Run 2 in 2022
- LHCb is progressing well to meet the Offline demands that run 3 will bring, coordinated by the DPA project





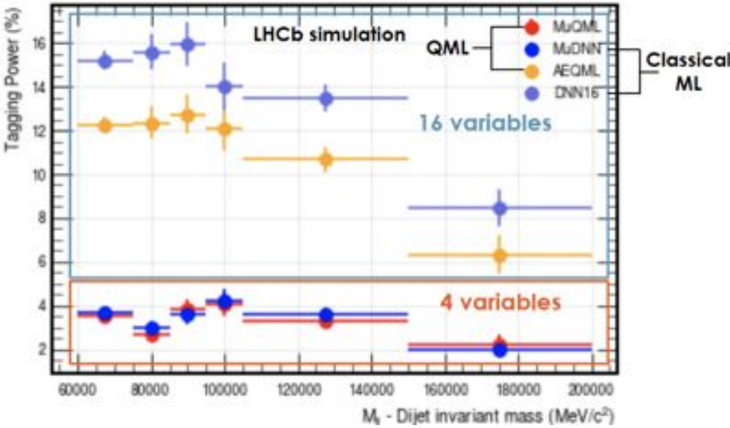
Backup

WP4 : R&D Innovative analysis techniques

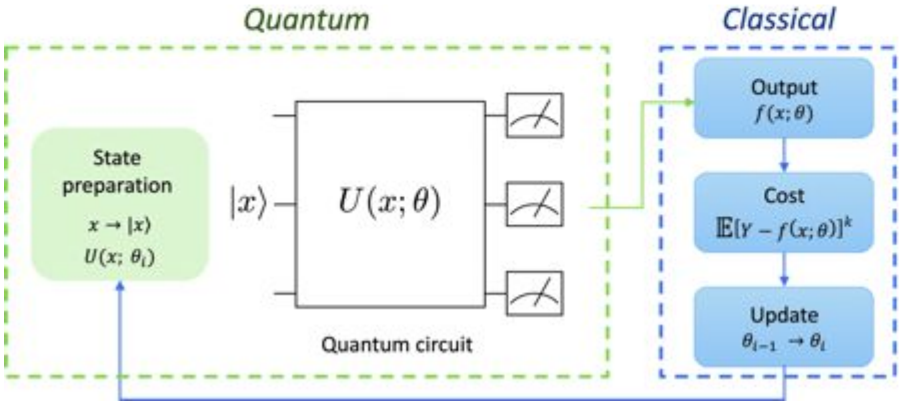
Brand new!

Quantum Machine Learning model for b vs $\bar{b}$ jet tagging

- Quantum Circuit with parameterized gates
- Variables from jet particles encoded in a quantum state
- The state is processed by trainable quantum gates
- Measurements on the final state are mapped to labels (b or $\bar{b}$)
- Parameters (of the gates) are optimized via a Gradient Descent minimization of a cost function (training)



QML vs Classical ML: comparable performances



Quantum models can allow the study of correlations among particles inside the jet meaning insight on jet substructure and better identification!

For more information see PyHEP 2021 talk