Stripping and Sprucing:

Offline data processing in LHCb

Shuqi Sheng (盛书琪), on behalf of the LHCb Collaboration

16/11/2024

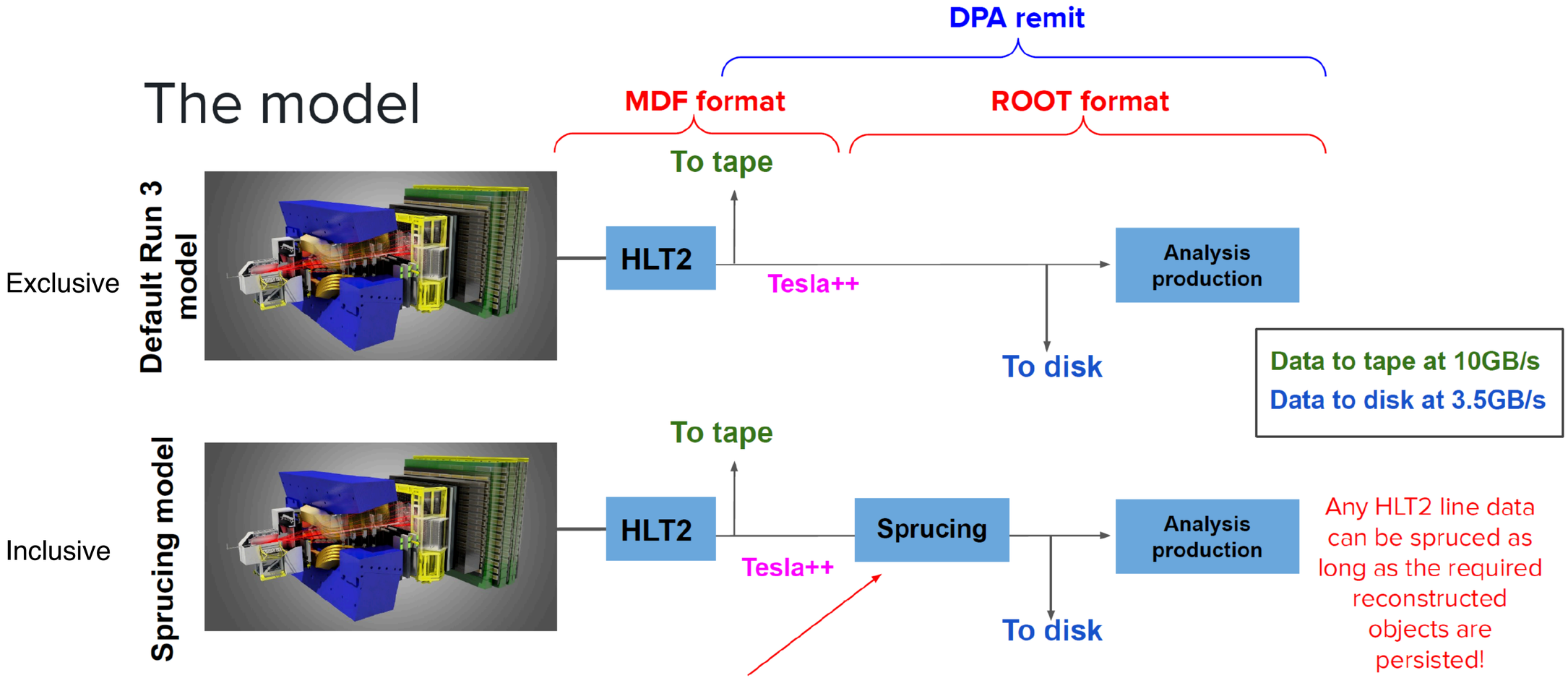
The Offline data challenge

- Data necessarily has a complex journey before it reaches analysts
- Run3 compared to Run2, LHCb revolutionized this process for the benefit of the analyst

Extremely short introduction for DPA

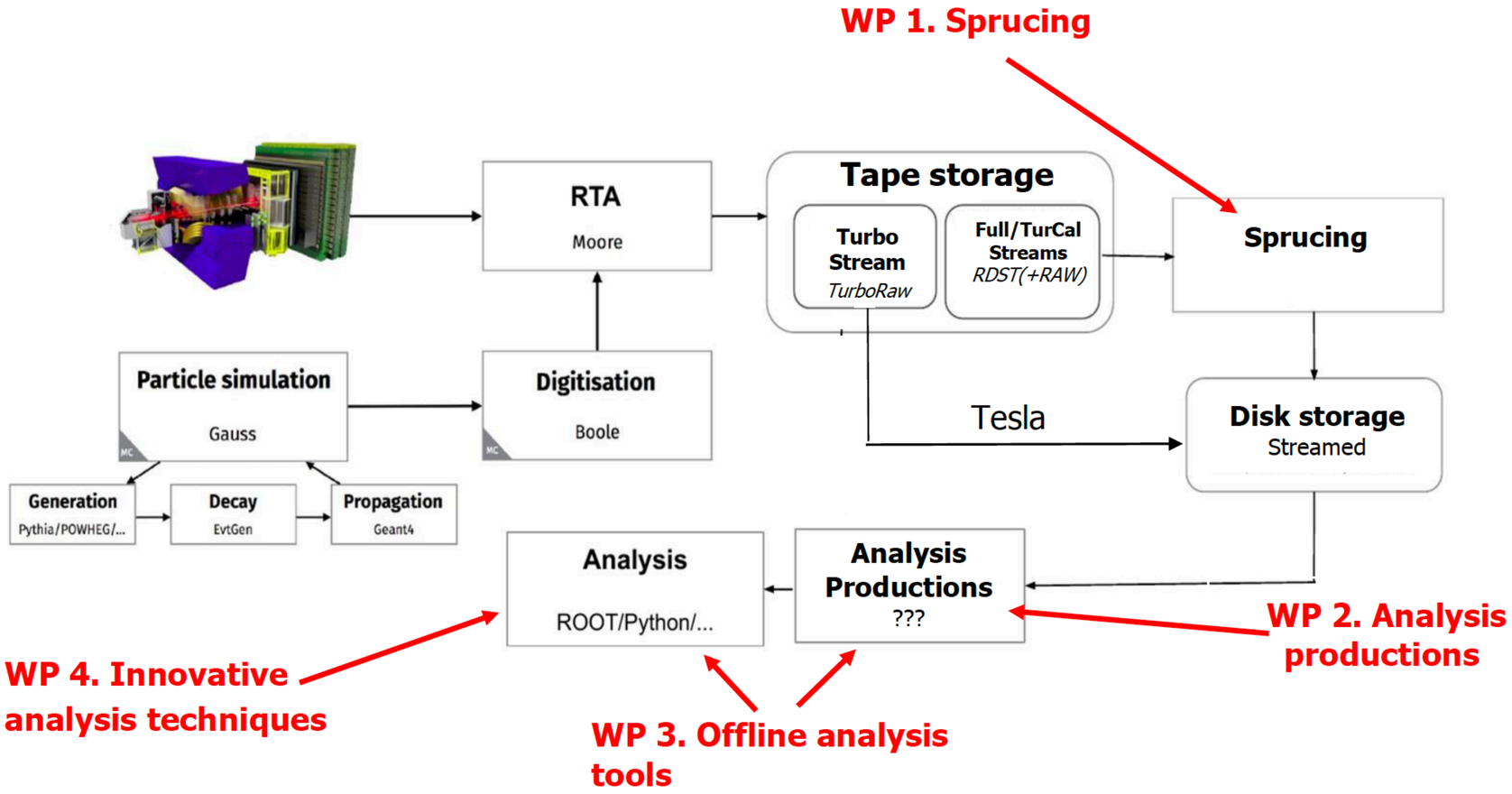
- Introduction
  - **Data processing & Analysis** (DPA) project: for offline data processing and analysis
    - It is built for 2 main ideas in 2020:
      - Centralized skimming and trimming (aka sprucing) of a significant fraction of trigger production
      - Centralized analysis production for physics WGs and users
  - Computing and software
    - RTA (online) + DPA (offline) joint
    - Computing and simulation

DPA in the run3 data/analysis workflow



Most of lines are streamed in Sprucing instead of running exclusive Sprucing.

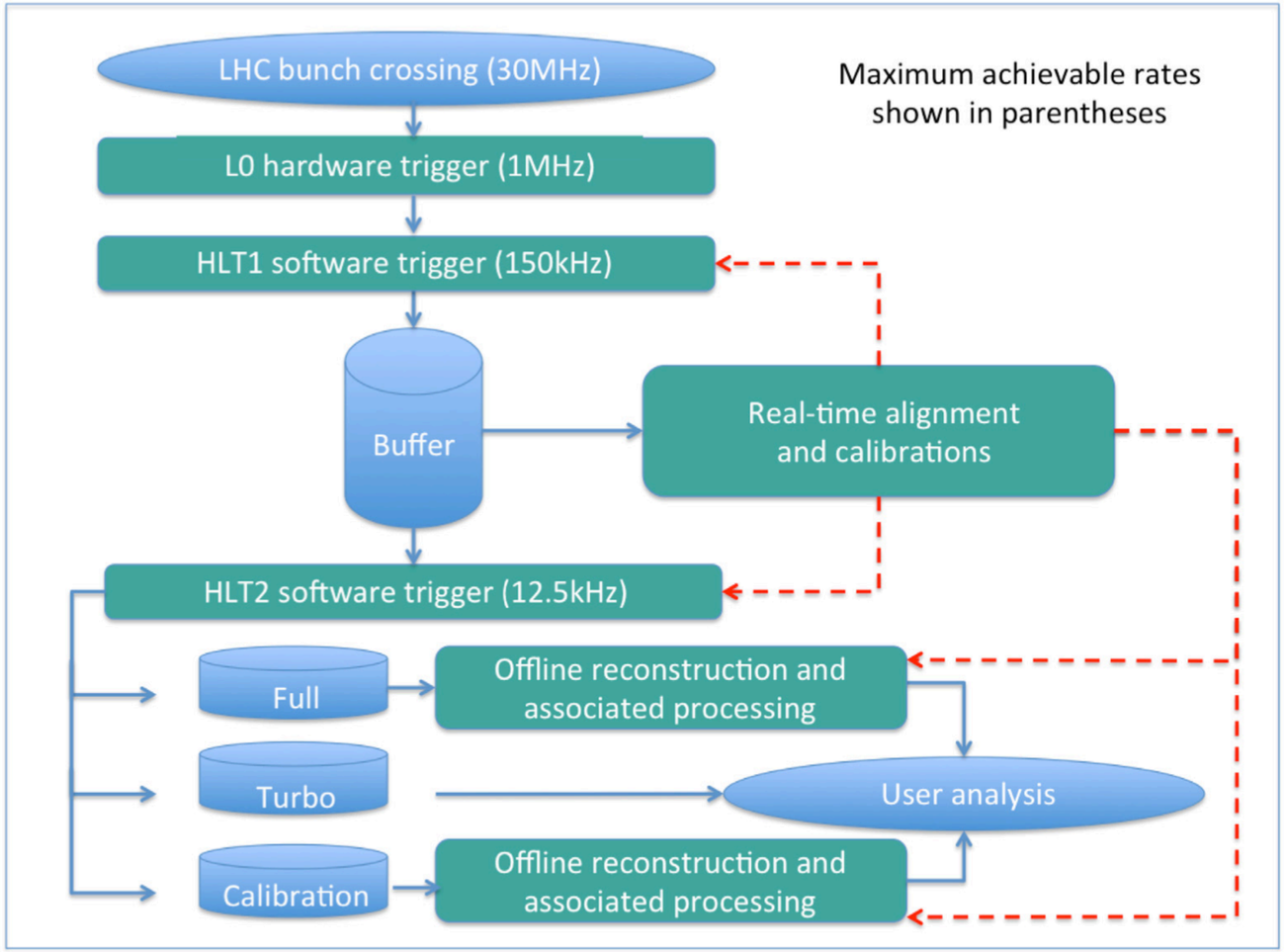
Process and analysis in Run3



WP1 - Sprucing
WP2 - Analysis Production
WP3 - Offline Analysis Tools
WP4 - Innovative Analysis Techniques
WP5 - Legacy Software & Data
WP6 - Analysis Preservation & Open Data
WP7 - Training and documentation

Focus of this talk

Trigger+offline processing in run2



LHCb detector and trigger performance in Run II

Fig. 1. Schematic diagram showing the overall data processing model in Run-II, where the blue solid line represents data flow, and the red dashed line the propagation of calibrations.

Offline data processing in run2

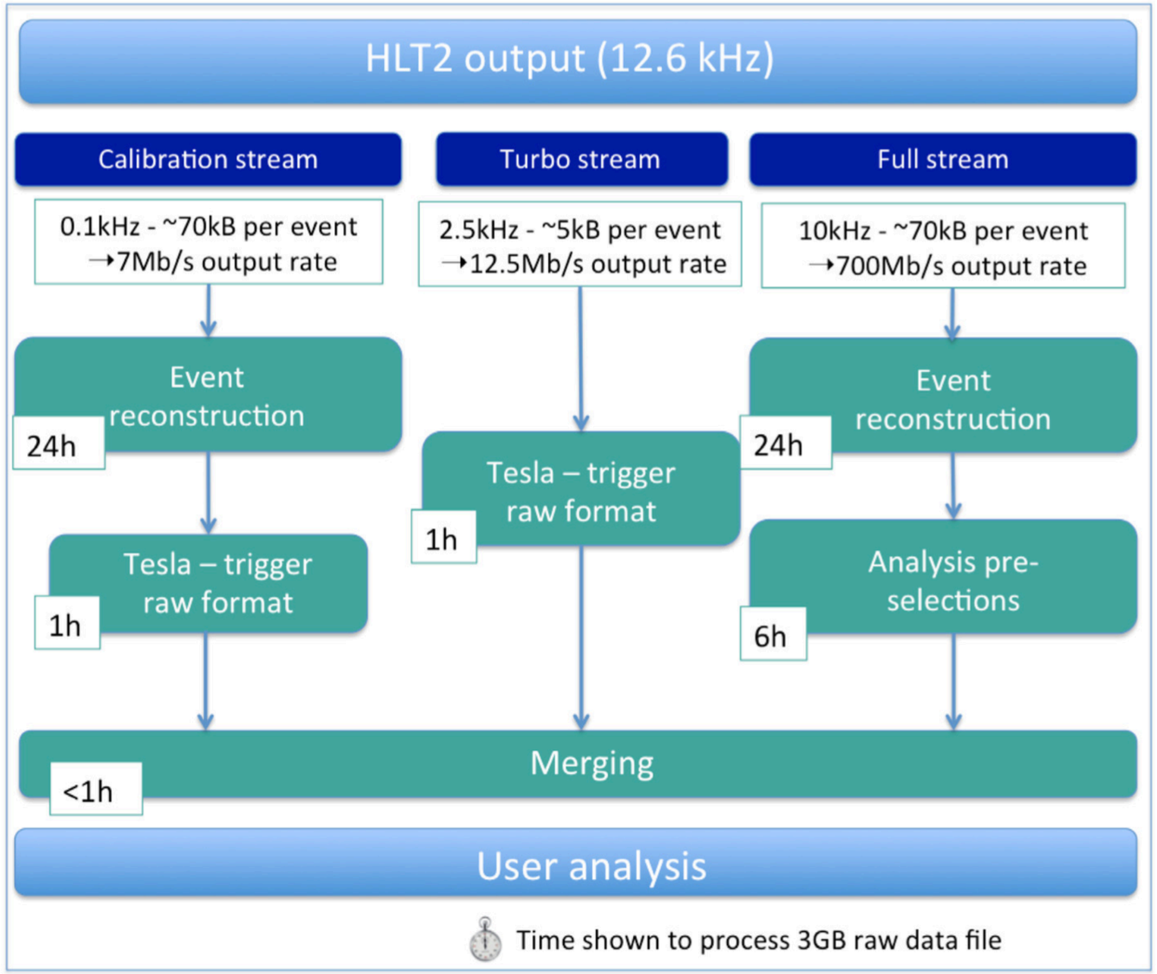


Fig. 2. Turbo data processing versus the traditional approach, as described in Section 4. The time taken for each step in hours is provided for a 3 GB raw data file. In addition, a calibration stream separates events for further processing to calculate data-driven efficiencies for both the Full and Turbo streams.

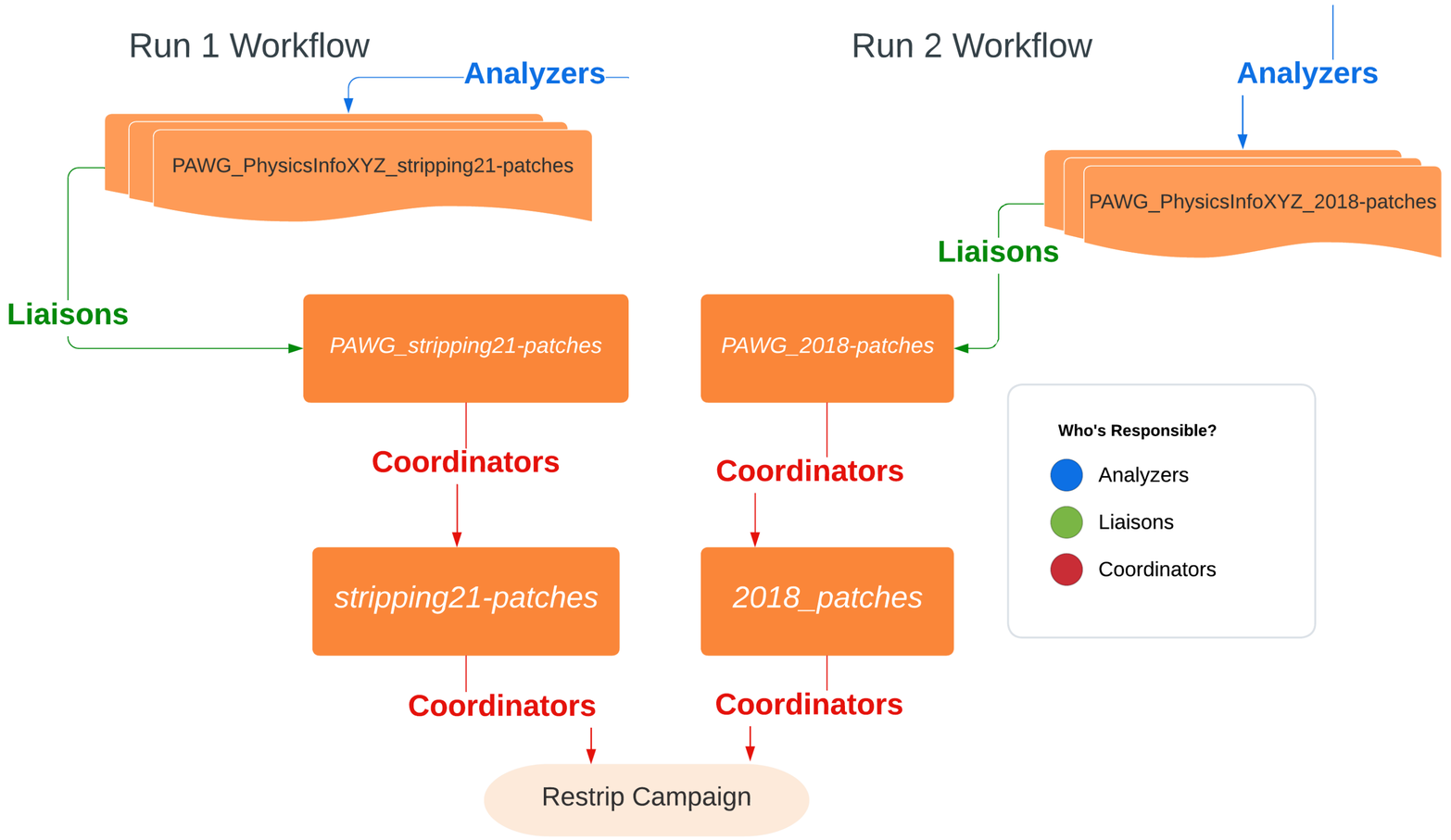
WP5: Legacy software & Data 🪂 Here I am

Long-term support of legacy runs 1 & 2 code and “user-level” (aka stripped) data is paramount

WP5: Legacy software & Data
Maintenance of, and support for, legacy run1&2 software and data samples.

- Legacy datasets will continue to need re-stripping campaign even in the very long term
- Maintenance of the stripping campaign & related software stack, i.e. DaVinci and legacy branch
- Porting out of the stripping campaign into maintained DaVinci branches
- Porting these user-facing processing code compatible with the latest python environment

A re-stripping campaign



- Gitlab workflow for development
- Analyst responsibilities:
  - Develop individual branch and prepare to be merged
  - Ensure the CI tests stay green
- Liaison responsibilities:
  - Test all the WG lines
  - Monitor the rate information and CI tests
  - Help in need for the line development
  - Prepare configuration for all WG lines
  - Finally approval to the merge

What is a Stripping Line

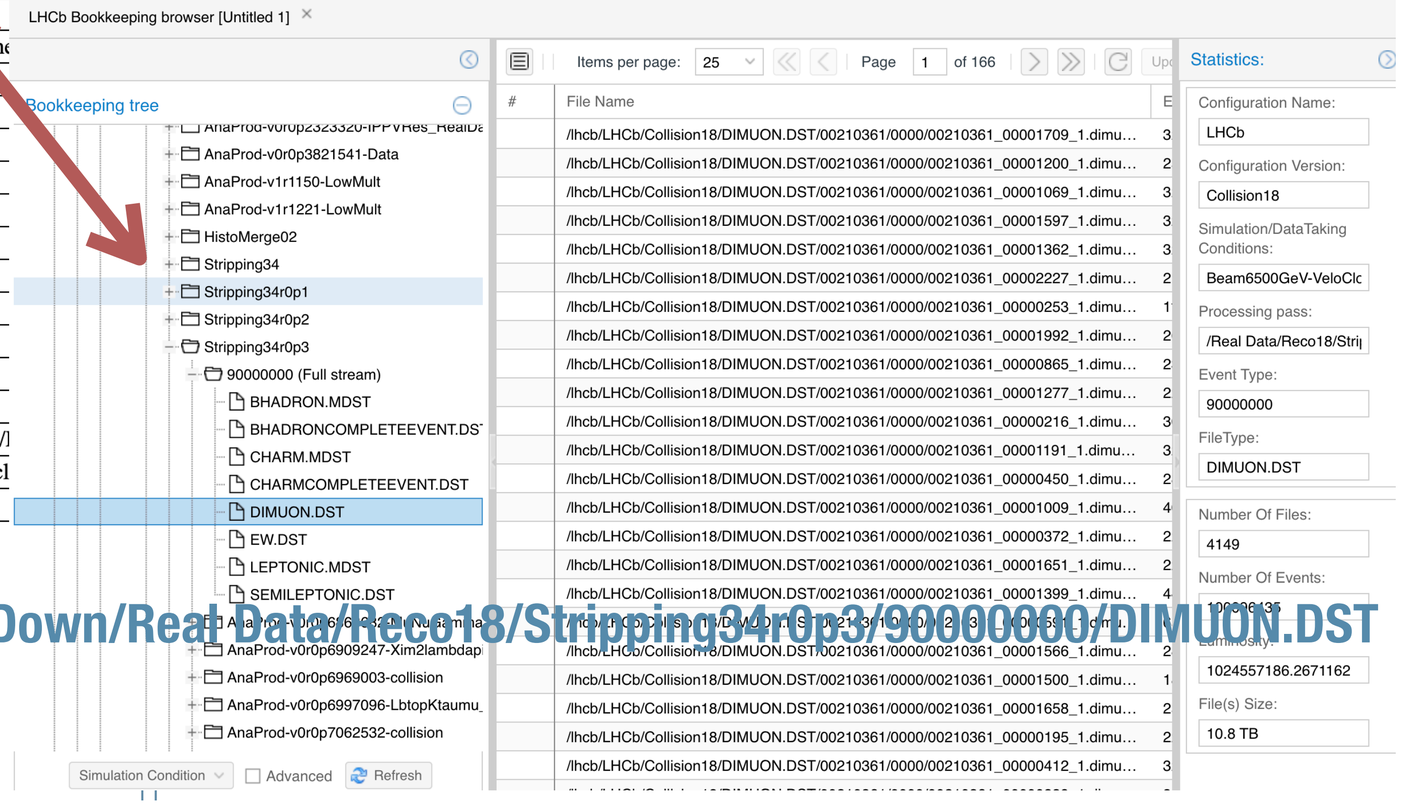
- The Stripping stage has access to all the reconstructed information in the event: analysts can take this and build their candidates
- The output of the Stripping is categorized into streams separated based on physics
  - BHADRON / BHADRONCOMPLETEEVENT / CHARM / CHARMCOMPLETEEVEVT/ DIMUON / EW / LEPTONIC / SEMILEPTONIC
  - A stream is a collection of lines
- Lines can write out either to
  - DST streams: The full reconstructed event is saved (with raw information that can be optionally persisted). Analysts have a lot of flexibility after the Stripping is run.
  - mDST streams: Slimmer format, discards raw event and only keep information concerning the candidates we build in the Stripping

Where is the stripping line?

Dimuon

https://lhcbdoc.web.cern.ch/lhcbdoc/stripping/config/stripping34r0p3/index.html

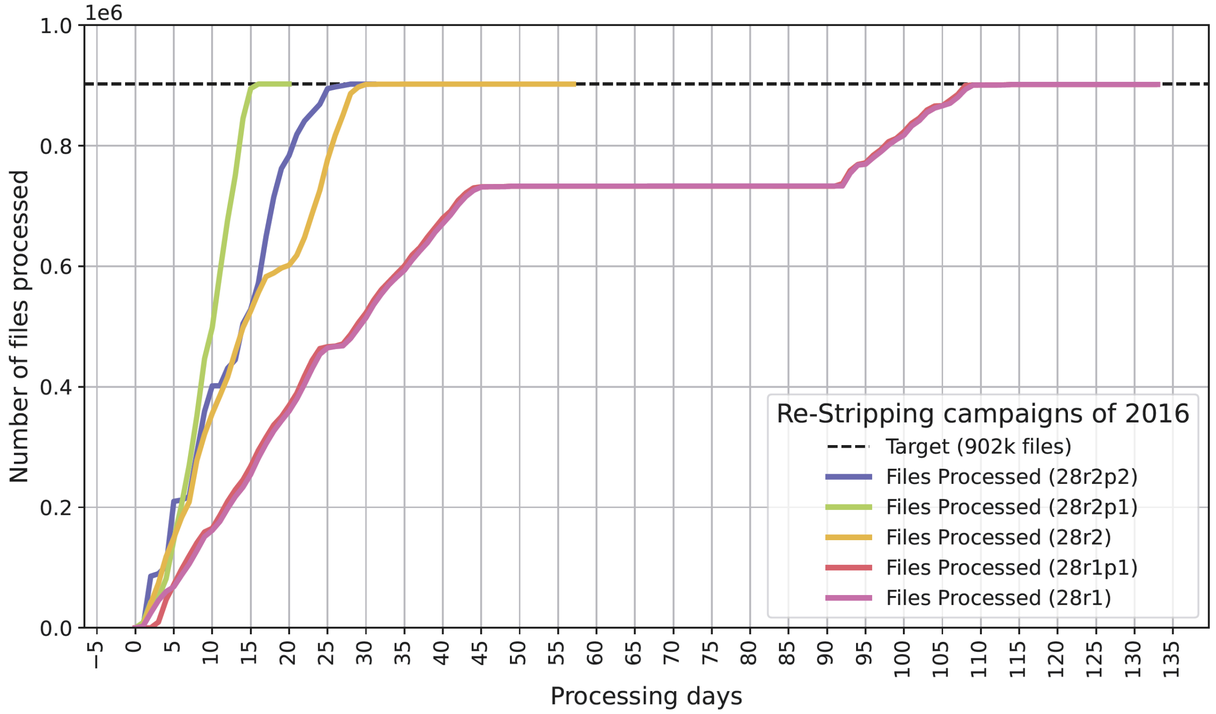
StrippingB2JpsiX_eta2pipipi0Line	Phys/B2JpsiX_eta2pipipi0Line/Particles	1.0
StrippingB2JpsiX_etaprime2pipipi0Line	Phys/B2JpsiX_etaprime2pipipi0Line/Particles	1.0
StrippingB2JpsiX_omega2pipipi0Line	Phys/B2JpsiX_omega2pipipi0Line/Particles	1.0
StrippingB2JpsiX_phi2pipipi0Line	Phys/B2JpsiX_phi2pipipi0Line/Particles	1.0
StrippingB2KX2KKDDDarkBosonLine	Phys/B2KX2KKDDDarkBosonLine/Particles	1.0
StrippingB2KX2KKDDSSDarkBosonLine	Phys/B2KX2KKDDSSDarkBosonLine/Particles	0.100
StrippingB2KX2MuMuDDDarkBosonLine	Phys/B2KX2MuMuDDDarkBosonLine/Particles	1.0
StrippingB2KX2MuMuDDSSDarkBosonLine	Phys/B2KX2MuMuDDSSDarkBosonLine/Particles	0.100
StrippingB2KpiX2KKDDDarkBosonLine	Phys/B2KpiX2KKDDDarkBosonLine/Particles	1.0
StrippingB2KpiX2KKDDSSDarkBosonLine	Phys/B2KpiX2KKDDSSDarkBosonLine/Particles	0.100
StrippingB2KpiX2MuMuDDDarkBosonLine	Phys/B2KpiX2MuMuDDDarkBosonLine/Particles	1.0
StrippingB2KpiX2MuMuDDSSDarkBosonLine	Phys/B2KpiX2MuMuDDSSDarkBosonLine/Particles	
StrippingB2MuMuMuMuB2DetachedDimuonAndJpsiLine	Phys/B2MuMuMuMuB2DetachedDimuonAndJpsiLine	
StrippingBc2Ds2MuMuLine	Phys/Bc2Ds2MuMuLine/Particles	
StrippingBc2Ds2MuMu_SSLine	Phys/Bc2Ds2MuMu_SSLine/Particles	
StrippingBu2LLK_mmLine_fulldst	Phys/Bu2LLK_mmLine_fulldst/Particles	
StrippingMultiLepton_3mDetLine	Phys/MultiLepton_3mDetLine/Particles	
StrippingMultiLepton_3mDetSSLine	Phys/MultiLepton_3mDetSSLine/Particles	
StrippingMultiLepton_3mPromptLine	Phys/MultiLepton_3mPromptLine/Particles	
StrippingMultiLepton_3mPromptSSLine	Phys/MultiLepton_3mPromptSSLine/Particles	
StrippingMultiLepton_B22mu2eXTightLine	Phys/MultiLepton_B22mu2eXTightLine/Particles	
StrippingMultiLepton_B24mLine	Phys/MultiLepton_B24mLine/Particles	
StrippingMultiLepton_B24muXTightLine	Phys/MultiLepton_B24muXTightLine/Particles	
StrippingMultiLepton_B26mLine	Phys/MultiLepton_B26mLine/Particles	
StrippingMultiLepton_Incl2mu2muLongLivedDownLine	Phys/MultiLepton_Incl2mu2muLongLivedDownLine/I	
StrippingMultiLepton_Incl2mu2muLongLivedLine	Phys/MultiLepton_Incl2mu2muLongLivedLine/Particl	
StrippingMultiLepton_InclDet2mu2muLine	Phys/MultiLepton_InclDet2mu2muLine/Particles	

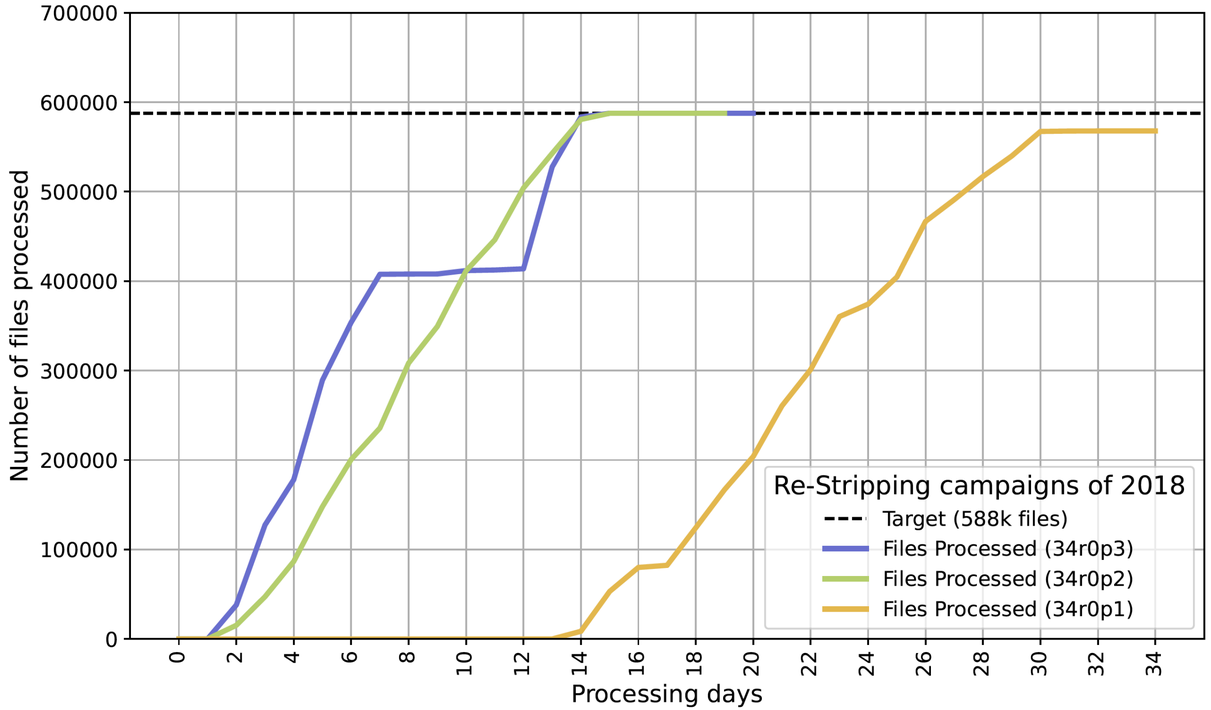


/LHCb/Collision18/Beam6500GeV-VeloClosed-MagDown/Real Data/Reco18/Stripping34r0p3/90000000/DIMUON.DST

Handshakes with Computer Team

- Live feedback of samples processing allows to catch any serious oversights in development
- All campaigns took approximately 2 weeks

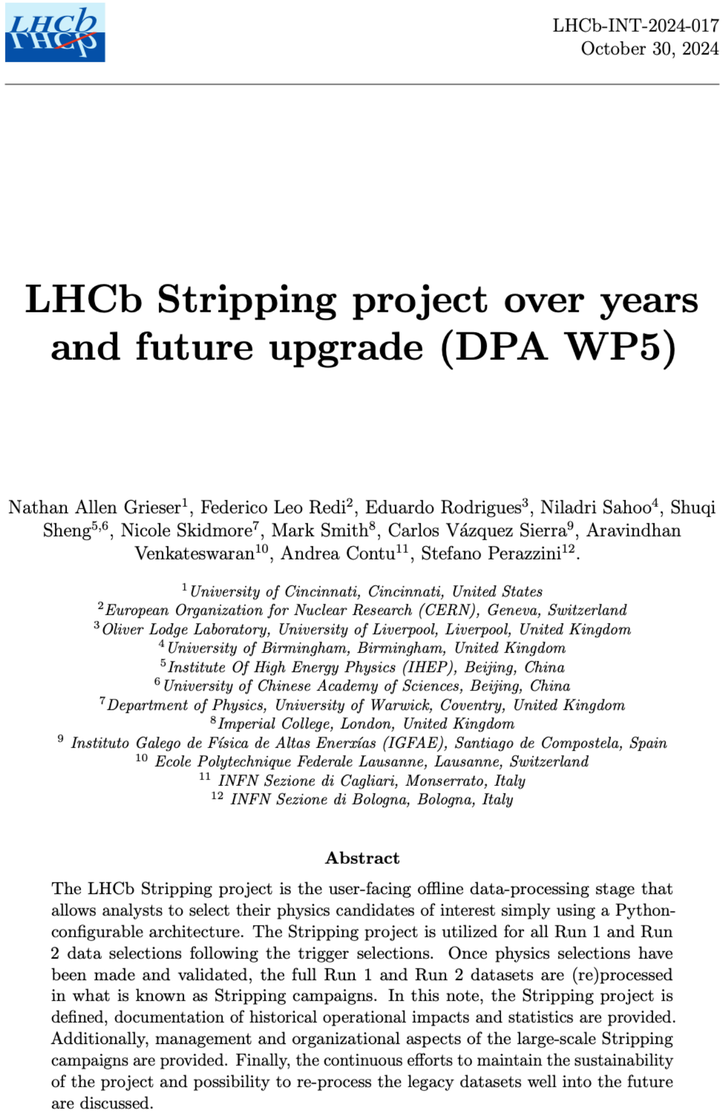




How is Stripping going?

- An internal note has been prepared titled “LHCb Stripping project over years and future upgrade”
- It is planed to be public for collaboration during the next LHCb week in December
- A paper will be submitted at the beginning of next year

https://cds.cern.ch/record/2915185



LHCb-INT-2024-017
October 30, 2024

LHCb Stripping project over years and future upgrade (DPA WP5)

Nathan Allen Grieser[1], Federico Leo Redi[2], Eduardo Rodrigues[3], Niladri Sahoo[4], Shuqi Sheng[5,6], Nicole Skidmore[7], Mark Smith[8], Carlos Vázquez Sierra[9], Aravindhan Venkateswaran[10], Andrea Contu[11], Stefano Perazzini[12].

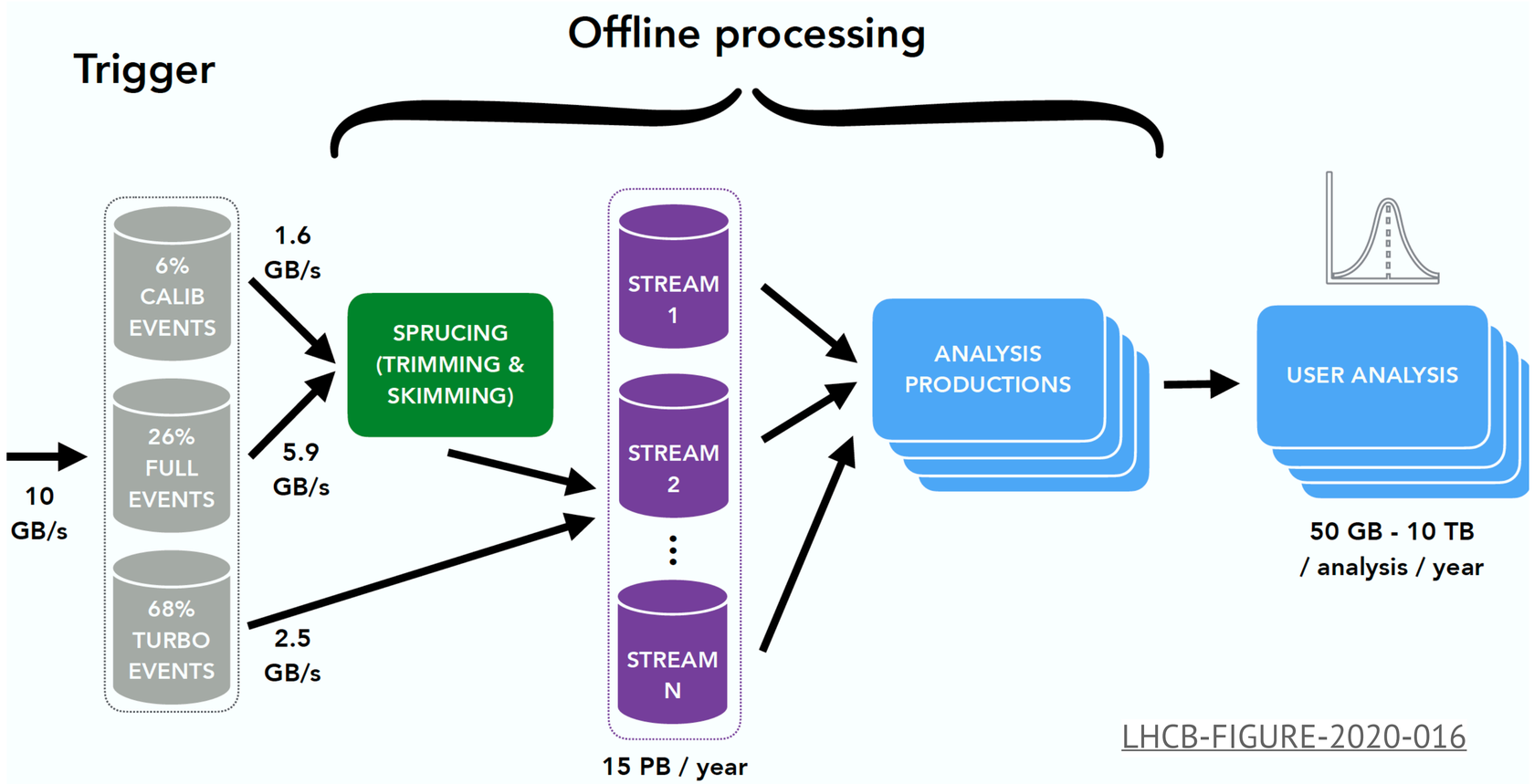
[1] *University of Cincinnati, Cincinnati, United States*
[2] *European Organization for Nuclear Research (CERN), Geneva, Switzerland*
[3] *Oliver Lodge Laboratory, University of Liverpool, Liverpool, United Kingdom*
[4] *University of Birmingham, Birmingham, United Kingdom*
[5] *Institute Of High Energy Physics (IHEP), Beijing, China*
[6] *University of Chinese Academy of Sciences, Beijing, China*
[7] *Department of Physics, University of Warwick, Coventry, United Kingdom*
[8] *Imperial College, London, United Kingdom*
[9] *Instituto Galego de Física de Altas Enerxías (IGFAE), Santiago de Compostela, Spain*
[10] *Ecole Polytechnique Federale Lausanne, Lausanne, Switzerland*
[11] *INFN Sezione di Cagliari, Monserrato, Italy*
[12] *INFN Sezione di Bologna, Bologna, Italy*

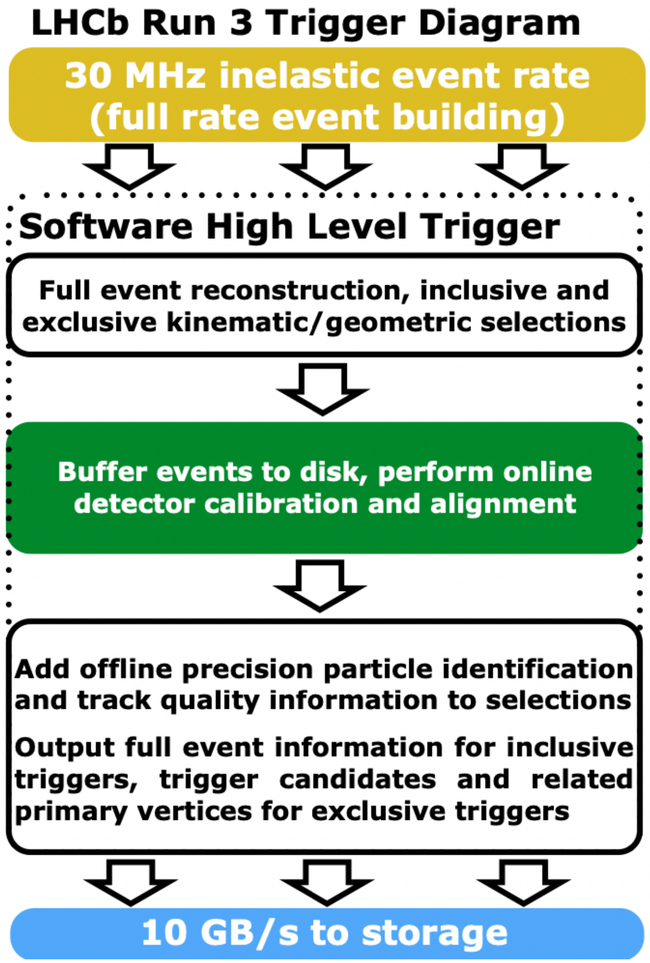
Abstract

The LHCb Stripping project is the user-facing offline data-processing stage that allows analysts to select their physics candidates of interest simply using a Python-configurable architecture. The Stripping project is utilized for all Run 1 and Run 2 data selections following the trigger selections. Once physics selections have been made and validated, the full Run 1 and Run 2 datasets are (re)processed in what is known as Stripping campaigns. In this note, the Stripping project is defined, documentation of historical operational impacts and statistics are provided. Additionally, management and organizational aspects of the large-scale Stripping campaigns are provided. Finally, the continuous efforts to maintain the sustainability of the project and possibility to re-process the legacy datasets well into the future are discussed.

Contents

The Run3 offline LHCb dataflow



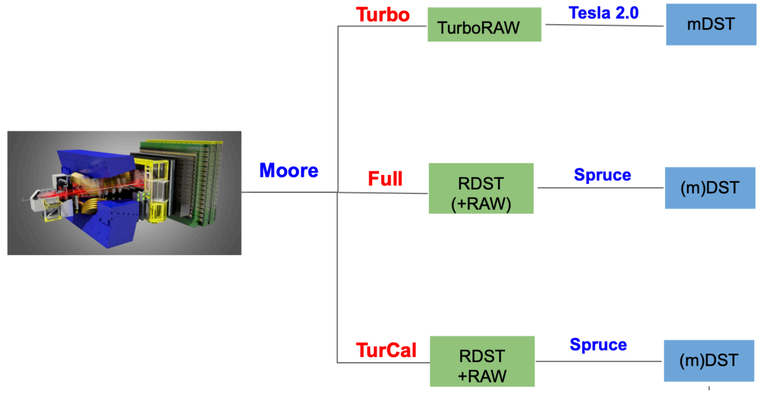


WP1: Sprucing

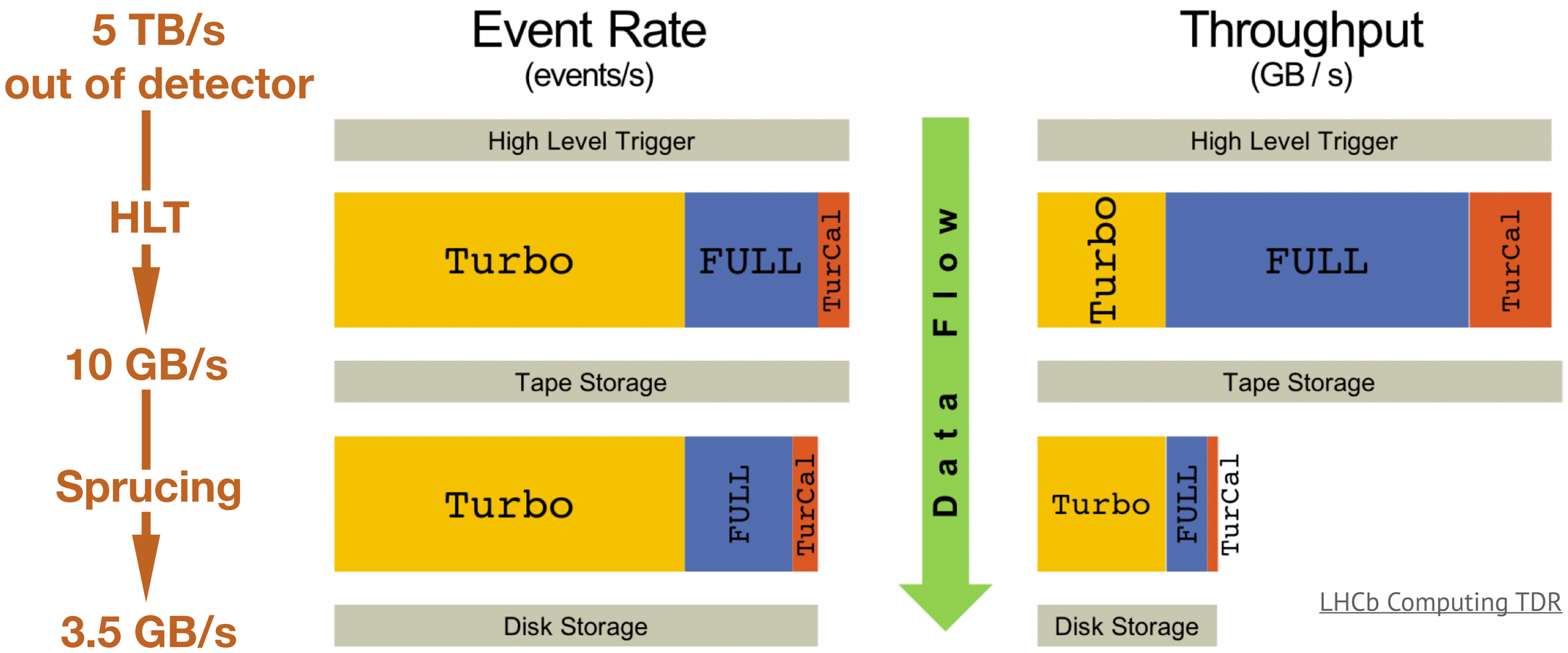
WP1: Sprucing
Centralized offline data selection/streaming for data that cannot go (initially) to TURBO stream.

Consistent selection, formatting and streaming of data using common applications, allowing efficient assess to all data given disk sources

- Sprucing line using same framework as HLT2 lines - Moore
- Sprucing line write out same format as HLT2 TURBO lines
- Sprucing data will be streamed - optional configuration
- Sprucing needs to be able to run all HLT2 output including TURBO



Sprucing: manage tape/disk persistency

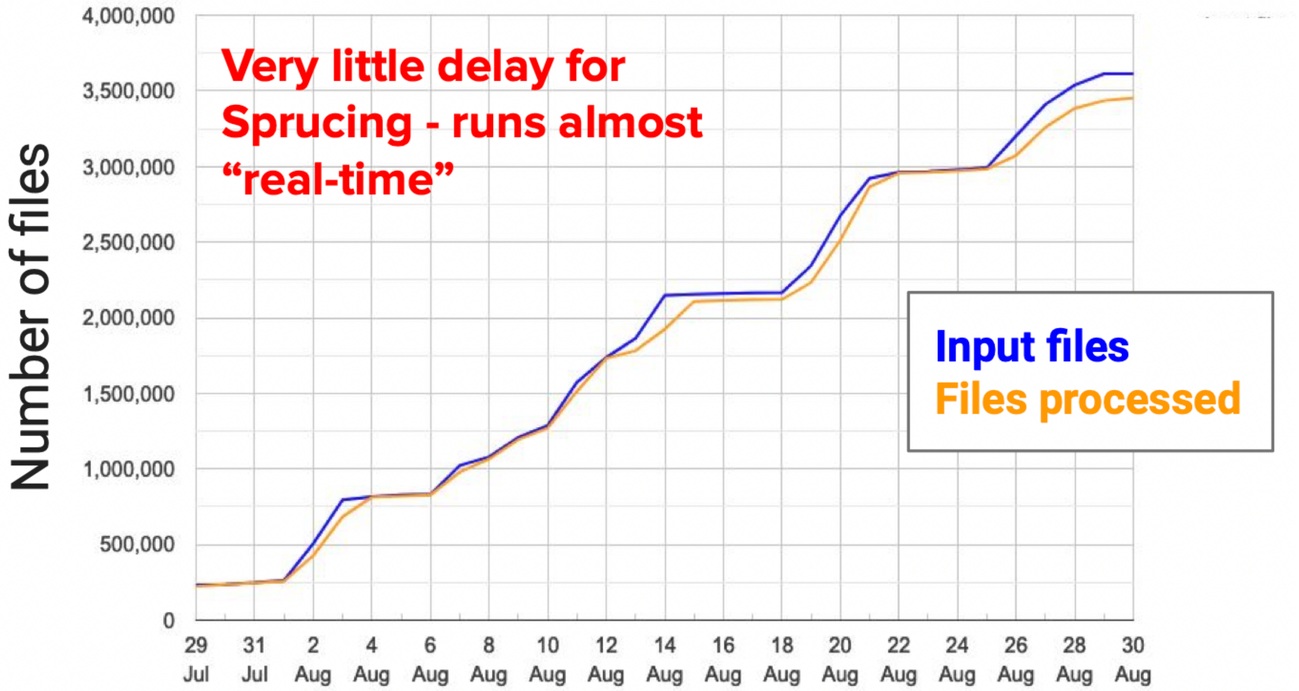


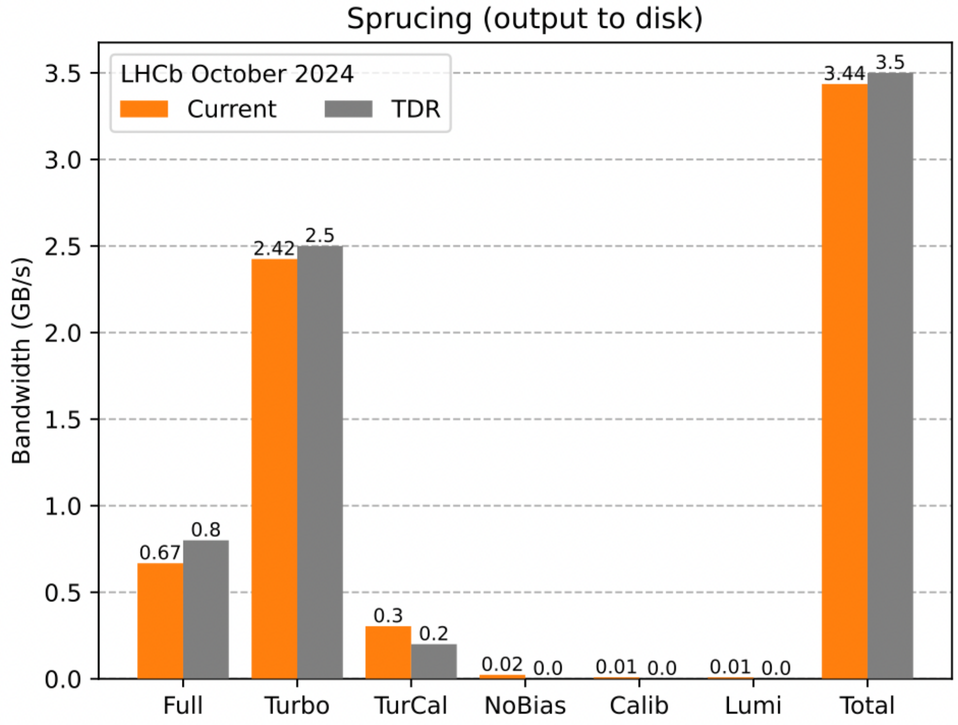
LHCb Computing TDR

Can keep inclusively selected full events on tape for future exploitation in yearly re-sprucing campaigns

Sprucing: manage tape/disk persistency

- LHCb spruced 35 PB data in 2024 over 3 physics streams and 2 technical streams
- Sprucing bandwidth of all streams to disk on HLT1-filtered real data from Run 30751





WP2: Analysis production - declarative ntupling

- A central submission system
- Exploit DIRAC transformation system
- Full job testing on Gitlab CI

Application + Job options + Data to run on

```
defaults:
  application: DaVinci/v45r4
  wg: WG
  automatically_configure: yes
  turbo: no
  inform:
    - someone@cern.ch
  options:
    - make_ntuple.py
  output: DVNtuple.root

{%- set datasets = [
  (11, 3500, '14', '21r1'),
  (12, 4000, '14', '21'),
  (15, 6500, '15a', '24r2'),
  (16, 6500, '16', '28r2'),
  (17, 6500, '17', '29r2'),
  (18, 6500, '18', '34'),
]%}

{%- for year, energy, reco, strip in datasets %}
  {%- for polarity in ['MagDown', 'MagUp'] %}

My_20{{year}}_{{polarity}}_job:
  input:
    bk_query: /LHCb/Collision{{year}}/Beam{{energy}}GeV-VeloClosed-{{polarity}}/Real Data/Reco

  {%- endfor %}
{%- endfor %}
```

Analysis production - declarative ntupling

Comprehensive job testing through Gitlab pipelines

Looks good!
Ran for 213720 events in dirac:951865446
Production was submitted as 141617.

WG	Events processed	Run time	Peak memory
B2OC	213,720	00:25:31	2.56 GB

Inputs and outputs

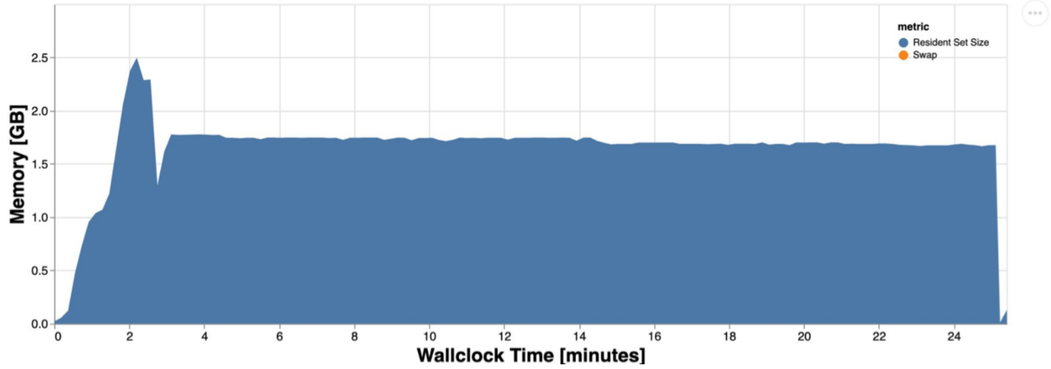
	Path	EOS path	Open in JSRoot	Size (this job)	Size (entire sample)
Input	LFN:/lhcb/LHCb/Collision24/B2OC.DST/00241744/0000/00241744_00007538_1.b2oc.dst			4.43 GB	101.87 TB
	LFN:/lhcb/LHCb/Collision24/B2OC.DST/00241585/0021/00241585_00210894_1.b2oc.dst			4.15 GB	
Intermediate	00012345_00006789_1.data.root	Copy		2.89 MB	
Output	00012345_00006789_2.data.root	Copy		2.76 MB	~ 32.82 GB

Report on estimated output size and memory usage

Steps

	No.	Application	Data packages	DB tags	Options	ProcPass	Visible	Test info
👍	1	DaVinci/v64r10 (x86_64_v2-el9-clang16-opt)	AnalysisProductions/v1r1874	—	ⓘ	AnaProd-v1r1874-Bs2DsPi_2024Data_24c3	✓	ⓘ
👍	2	LHCb/v55r1	AnalysisProductions/v1r1874	—	ⓘ	merged	✗	ⓘ

Resource Consumption



Interactive logs warning/error highlighting

Logs show less output

DaVinci_1.log | prmon_1.txt | prodConf_DaVinci_1.json | LHCb_2.log | prmon_2.txt | prodConf_LHCb_2.json | DIRAC.log

Copy Download

```
1  Overriding DIRACSYSCONFIG to /tmp/tmp1_mfqfq8,/tmp/pilot.cfg
2  Restarting process with ['/cvmfs/lhcb.cern.ch/lhcbdirac/versions/v11.0.48-1727212764/Linux-x86_64/bin/dirac-production-request-run-local', '/tm
3  Executing workflow locally
4  Executing from /tmp/951865446
5  Executing job at temp directory /tmp/951865446/Local_99hwjh07_JobDir
6  File not found Request_0_AnalysisProduction_AnaProd-v1r1874-Bs2DsPi_2024Data_24c3_EventType_94000000_B2OC_1.xml
7  Job has input data requirement, will attempt to resolve data for DIRAC.LocalProdTest.local
8  Replica Lookup Time: 0.48 seconds
9  Metadata Lookup Time: 0.12 seconds
10 Job has a specific policy setting: DIRAC.WorkloadManagementSystem.Client.DownloadInputData
```

apd python packages allows for easy data files retrieval

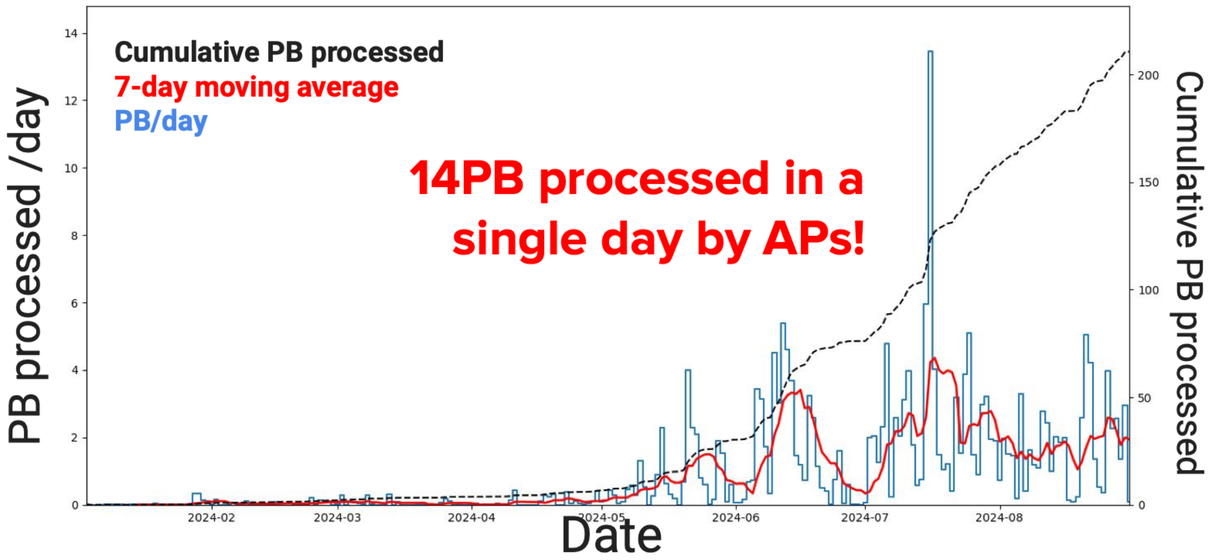
```python
from apd import AnalysisData

datasets = AnalysisData("b2oc", "bs2dspi_run3")
bs2dspi_2024data_magdown_24c2_pfns = datasets(polarity="magdown", eventtype="94000000", datatype="2024")
```

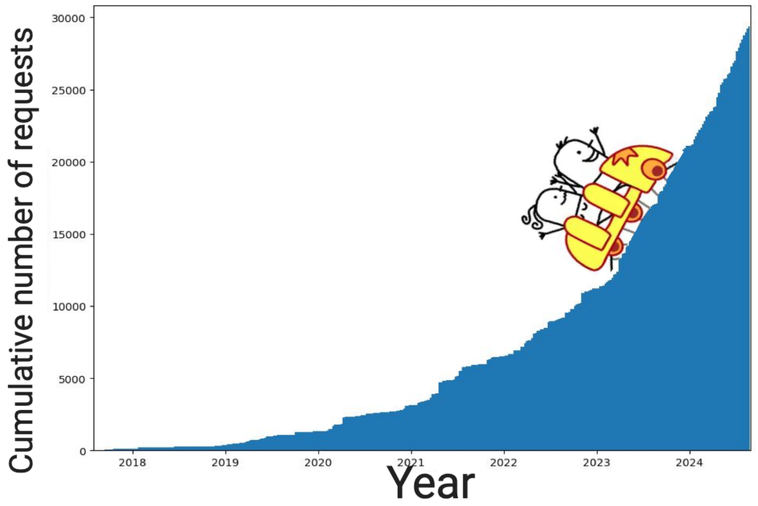
Full data provenance with datasets tagged by analysis

Analysis production - declarative ntupling

- Full adoption of analysis productions at LHCb
  - Over 1200 Run3 APs have been submitted so far
  - 700+ “live” APs picking up data as it was Spruced



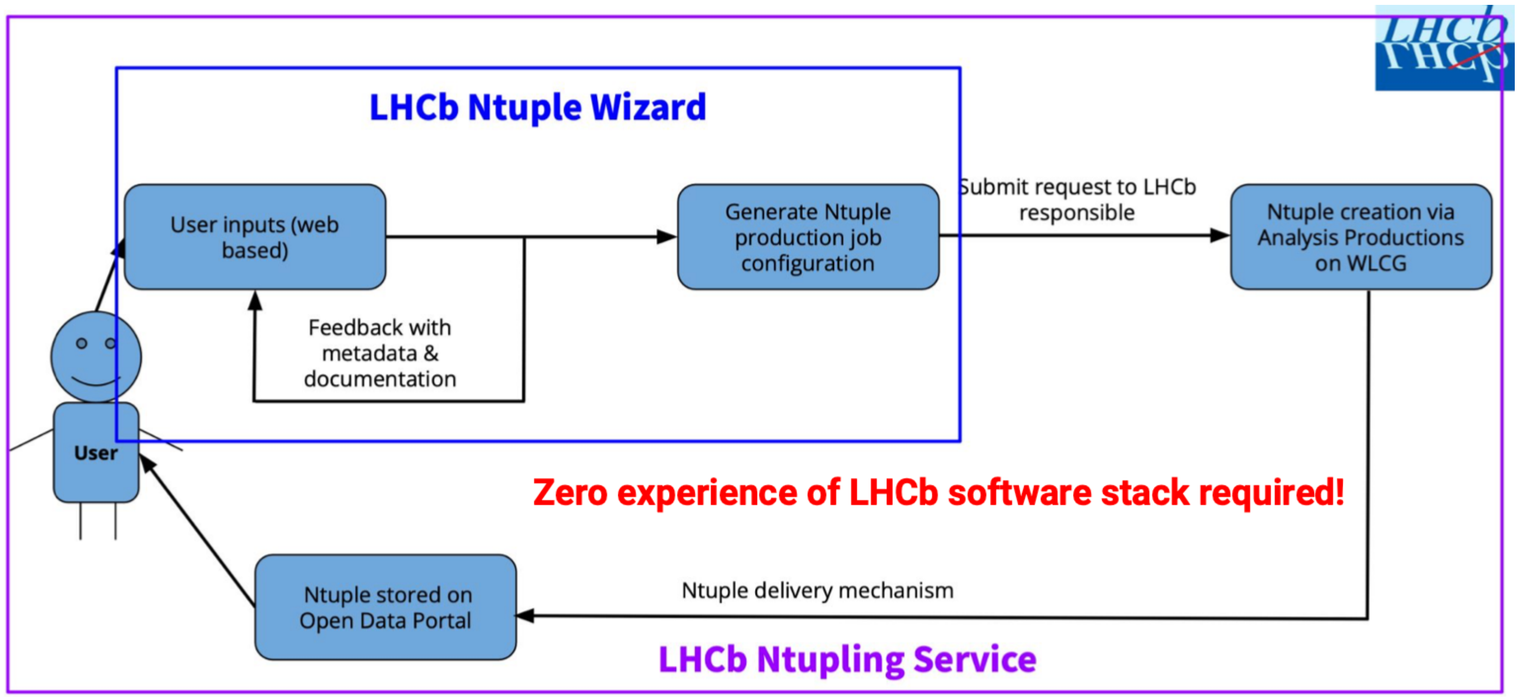
PB processed per day by APs with moving average



Cumulative APs submitted since invention

Open data

- LHCb released its full Run1 dataset ~ 800TB
  - Need scalable solution going forward - **NTuple Wizard!**
  - First public release of LHCb Ntupling Service expected in 2025, efforts ongoing to release all LHCb Run 2 data by 2028



Stay tuned!

Summary

- LHCb continues to have a thriving legacy physics program
- Software and development is maintained to allow for large data reprocessing campaigns
- LHCb Run3 necessitated an overhaul of the offline dataflow
  - Sprucing provides a method for saving high event persistency on tape for future exploitation but also manageable disk requirements
  - Analysis production for ntuple creation are one of the single-most transformative changes at LHCb saving countless person-hours
- With such changes in dataflow, LHCb is creating the Run3 StarterKit for onboarding and continuous reference material

Thank you!

Back Up

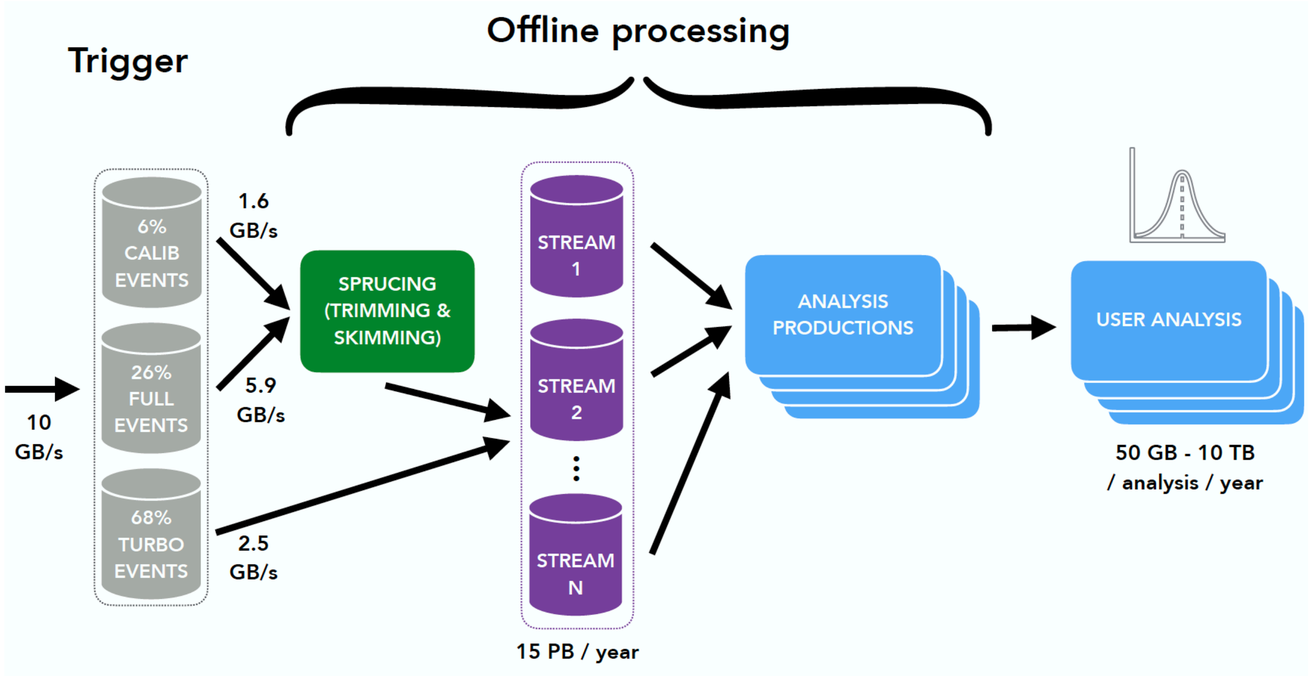
WP2: Analysis production

Proposal for heavily automated, centralized productions using the DIRAC transformation system

WP2: Analysis production
An upgrade to the WG production of Run1,2,3 done centrally.

- Ntuple production with user jobs
- Move the typical analysis workflow
  - Make ntuples → process data with custom scripts →start the actual analysis work to the centralized production system
- Benefit
  - Far greater automation ⇢less time investment for analysts (no baby-sitting grid jobs)
  - Long-term preservation of analysis step

DPA work packages



LHCb upgrade dataflow focusing on the offline aspects.

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

DPA

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

- WP1 - Sprucing
  - The Sprucing runs in two forms
    - **Passthrough** is for HLT2 TURBO stream (changing the file format from MDST to DST and creating File Summary Records (FSRs) for luminosity information)
    - **Exclusive** is for HLT2 Full Stream
- WP2

Development - What goes into a MR?

- Put the developed lines in the “correct” position
- Liaison and coordinator take the responsibilities to review

adding D0->K3pi mode for PromptCharm and moving it to BandQ

Edit | Code

Merged **Ivan Polyakov** requested to merge `BandQ_D02K3pi-ForDiCharm-v…` into `BandQ_2018-patches` 10 months ago

Overview 4 | Commits 1 | Pipelines 4 | Changes 2

All threads resolved! | Add a to do

replacing !1755 (closed).

main motivation is the `Tcc->D0 D0 pi+` channel where adding D0->K3pi mode may increase statistic by x2.

Implemented by substituting `D02Kpi` with `D0`, which is a merge of `D02Kpi` and `D02K3pi`. Thus, all lines like of DiCharm, TriCharm, ... and so on will benefit.

As in default_config the WG specified was 'BandQ' and lived under B&Q line dictionaries asked by @ngrieser to move corresponding file to BandQ folder. Both `TestMyStrippingLine.py` and `TestMyWGfromSelections.py` execute fine and resulting 0000.Charm.mdst contains events from the bewly added mode. See rates below:

```
|                                          *Decision name*|*Rate,%*|*Accepted*| *Mult*|*ms/evt*|
|_StrippingGlobal_                                        |  3.8700|       387|       |   9.265|
|_StrippingSequenceStreamCharm_                           |  3.8700|       387|       |   3.515|
|!StrippingD02KpiForPromptCharm                           |  0.4500|        45|  1.022|   0.120|
|!StrippingD02K3piForPromptCharm                          |  0.1100|        11|  1.182|   0.076|
|!StrippingDstarForPromptCharm                            |  0.6100|        61|  1.016|   0.255|
```

2 Assignees Edit
Ivan Polyakov
Vanya Belyaev

Reviewer Edit
Shuqi Sheng

Labels Edit
B&Q

Milestone Edit
2023 Re-Strip Campaign -- Development and Testing (expired)