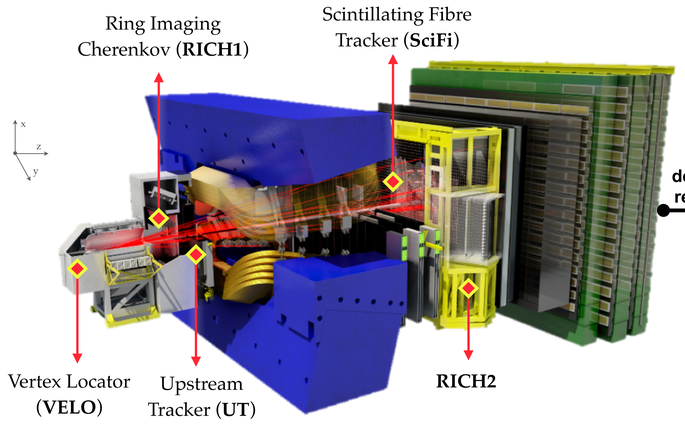
NEW GENERATION OFFLINE SOFTWARE FOR THE LHCb UPGRADE I

- *Sprucing, Analysis Productions and Ntupling* -

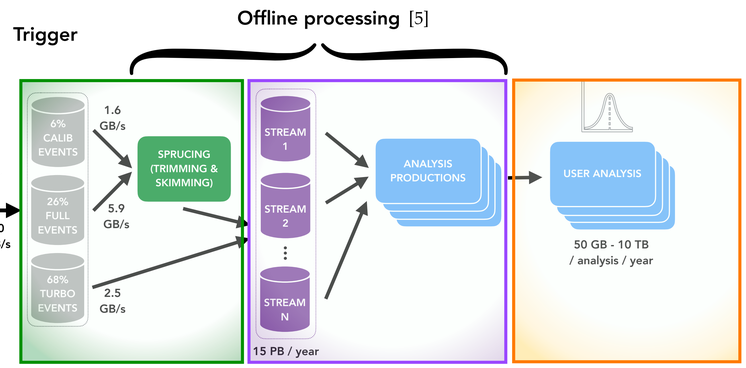
Martina Ferrillo, on behalf of the LHCb Collaboration

University of Zurich[UZH]

LHCb Upgrade I challenges: data processing



: newly installed detectors in LHCb Upgrade I [1 - 4]



Offline LHCb Upgrade data flow, focusing on the three aspects of **Sprucing (1)**, **Analysis Productions (2)** and **User Analysis tools (3)**.

During LS2, more than **90%** of the LHCb active **detector channels** have been **replaced** for Run3 data taking period and beyond.

To cope with the Run3 **crossing rate** of (non-empty) **bunches** of **30 MHz** and accommodate the need for higher efficiency [4], LHCb has deployed a fully software-based trigger.

With an increase by **5x** in **instantaneous luminosity** and a ~2x more efficient trigger [4], LHCb expects ca. **10x signal yield/time** and **3x event size**.

⇒ **Expected increase of ~30x in the data volume !** [6]

To face this unprecedented challenge in data management and coordinate the commensurate software developments, the *offline* **Data Processing & Analysis (DPA)** project has been created in 2020. Within its **scope**:

- The trimming/reduction of the data coming from the trigger (*i.e. Sprucing*) ;
- The centralised production of the NTuples for subsequent data analysis (*i.e. Analysis Productions*) ;
- Software developments for a modern offline analysis framework (*i.e. Offline Analysis tools*) .

(1) Sprucing

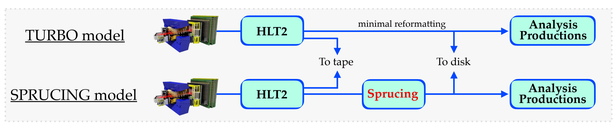
Run 3 persistency models, *i.e.* amount of per-event information to be saved by the trigger:

	Event size	*Persisted objects*	*Saved to disk*
(a) **TURBO**	4-16 kB	***Only*** the signal candidate is saved, discarding the rest of the event.	Yes. No further trimming of data is needed.
(b) **TURBO Selective Persistence**	~16 kB	The signal candidate is saved ***together*** with a custom set of other physics objects.	Yes. No further trimming of data is needed.
(c) **FULL/TURCAL**	48-69 kB	The ***whole*** event information is retained.	No, saved to tape, **not accessible to users**. Move to disk only after further selection (**Sprucing**).

Run 3 baseline / Calibration, other physics ch.

Ca. 70% of the physics of interest is saved to disk in the default Run 3 model (**TURBO SP**) and is immediately available for analysis, while *e.g.* **beauty physics** channels are persisted in the more inclusive **FULL streams**.

FULL streams are saved to **tape**. Before being **moved to disk** storage they undergo the **Sprucing** stage, a set of offline selections (**sprucing lines**) to reduce the event size, running concurrently with **data-taking** and during **Winter shutdowns**. The Sprucing framework is shared with the high-level Trigger stage (**HLT2**), thus making the trigger and sprucing lines interchangeable [7].



(2) Analysis Productions

After the Sprucing/Turbo(SP) stage, the data is saved to disk in **Streams** (files) grouped according to the physics of interest (common sprucing/trigger selections).

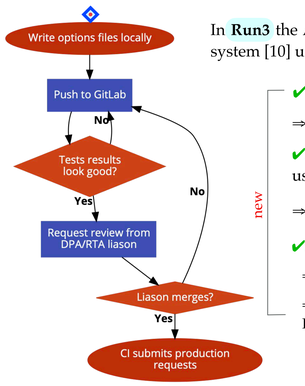
Those files can be processed to further **reduce** the **data volume** by a factor up to $O(10^3)$.

⇒ **Analysis Productions**

Run1-2 scenario - job submission via Ganga system - presents unavoidable **scaling issues**:

old
- ✘ : **Many thousands** of jobs for each analysis, with subsequent long production times;
- ✘ : **Inefficient** use of distributed computing resources in case of buggy jobs. Manual resubmission for each job to be handled by analysts.

In accordance with **CERN Open Data** policy [8], part of the LHCb dataset is made available to the general public. To ensure NTuples can be built in the future without any prior knowledge of the LHCb software, the **NTuple Wizard** web application is being developed [9].



Flowchart for Run3 Analysis Productions

In **Run3** the Analysis Productions will run through the **LHCbDIRAC** grid system [10] used for processing of Data and Simulation:

new
- ✔: **Heavily automated** and **centrally managed productions** ;
- ⇒ Automatic handling of files grouping and job failures.
- ✔: **Optimisation** of computing resources usage by **validating** user jobs configuration **prior** to submission;
- ⇒ Continuous Integration tests + approval from working group liaisons.
- ✔: **Analysis preservation** infrastructure.
- ⇒ Save info about file origin in the Bookkeeping system.
- ⇒ Automatically retrieve the output location of each Analysis Production.

◆ Analysis Productions **job options**: intuitive configuration via YAML files (**shared infrastructure** design with Sprucing and Offline Analysis Tools)

(3) User analysis tools

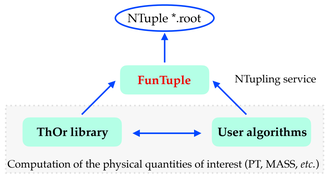
In **Run1-2** the production of **custom** user **NTuples** (*NTupling*) was performed through the DaVinci application.

old
- ✔: Easy to implement building blocks (*a.k.a.* TupleTools) to save the needed event information.
- ✘ : **Redundancy** in the saved variables (often 500+) ⇒ Up to **10 TB** of data for a single analysis!

To meet the demanding requirements posed by the huge increase in the data volume in **Run 3** and beyond, a **new** optimised **DaVinci framework** has been developed:

new
- ✔: Sharing the same general and selection-specific framework with the Trigger (*e.g.* Moore THOr) [11] ;
  - ⇒ 1:1 **consistency** between the **online** and **offline** selection frameworks .
- ✔: Modern, thread-safe and accommodating parallelisation ;
- ✔: Unit-testing routines for debugging and CI within the CERN GitLab platform ;

The **core functionality** of the new DaVinci resides in the **NTupling** tool, *i.e.* **FunTuple** :




```
'''
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-'
#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
})
```

particles

variables

Key features

Separate the **NTupling stage** from the computation of the **physical quantities**, which has to be addressed by dedicated functions or user algorithms.

Intuitive and self-explanatory structure. Total **flexibility** in the choice of the **variables** to save, leading by design to **light-weight** custom **NTuples**.

Transparent application **configuration** in dedicated YAML or JSON files (same as for Analysis Productions).

```
Upgrade_Bd2KstarMuMu_ldst:
    filenames:
    - '/path_to_inputfile_1/file1.dst'
    - '/path_to_inputfile_2/file2.dst'
    qualifiers:
        data_type: Upgrade
        input_type: DST
        simulation: true
```

References and Resources

[1] CERN-LHCC-2013-021
[2] CERN-LHCC-2013-022
[3] CERN-LHCC-2014-001
[4] CERN-LHCC-2014-016
[5] LHCB-FIGURE-2020-016
[6] CERN-LHCC-2018-014
[7] N. Skidmore, Run-3 offline data processing and analysis at LHCb, EPS-HEP 2021
[8] CERN-OPEN-2020-013
[9] LHCb Collaboration, NTupling Wizard
[10] LHCbDIRAC documentation
[11] LHCb Collaboration, ThOr Functors

LEPTON PHOTON 2021 MAN

30th International Symposium on Lepton Photon Interactions at High Energies

10-14 Jan 2022, *University of Manchester (UK)*