Triggers

Luke Grazette, Ivan Cambon Bouzas

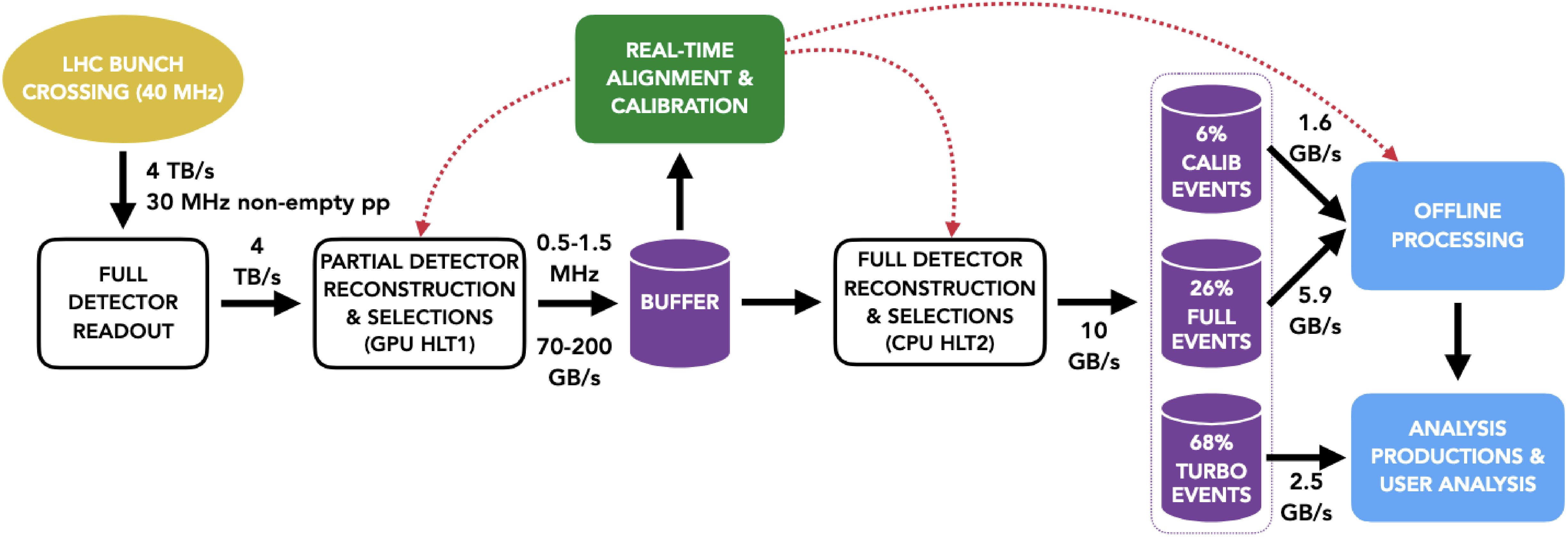
Contents

- Lecture-style (Now)
  - Re-introduction to the LHCb DataFlow
  - Hlt1
  - Hlt2 and the persistency Model
  - Sprucing

- Hands-on Session (Later)
  - Running Hlt2 and interpreting the output
  - Configuring Hlt2 algorithms and writing lines
  - HltEffChecker and other useful tools

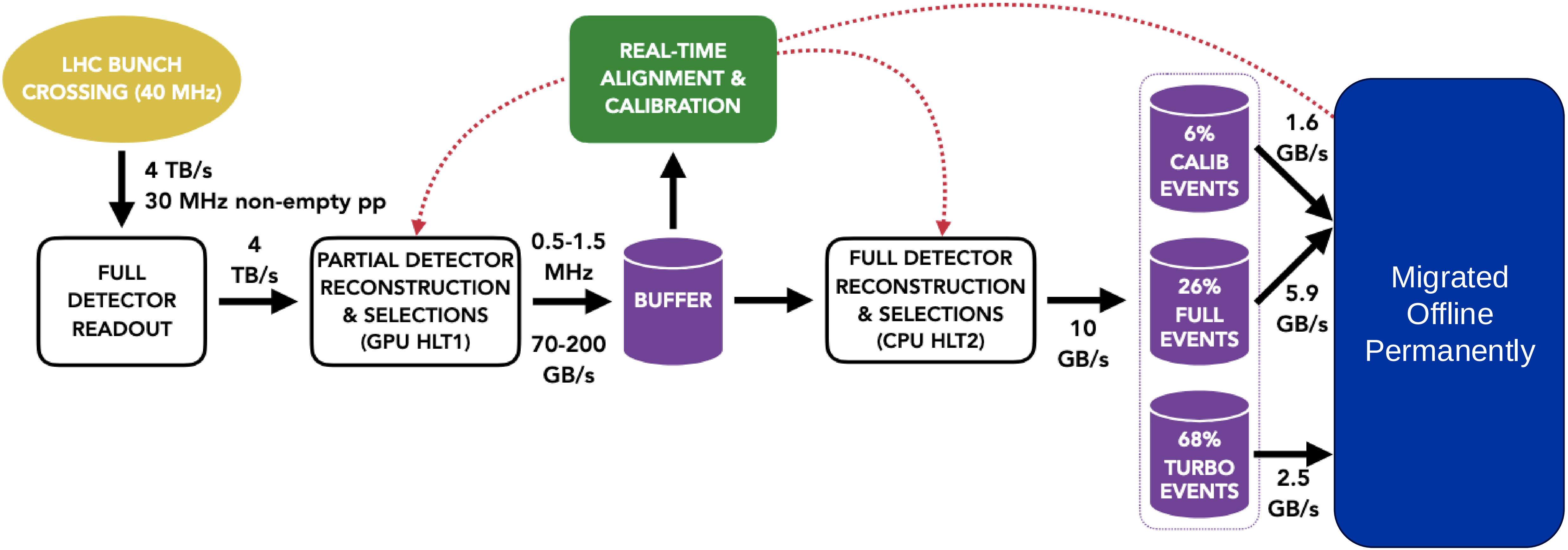
Dataflow

The LHCb Upgrade *Online* DataFlow



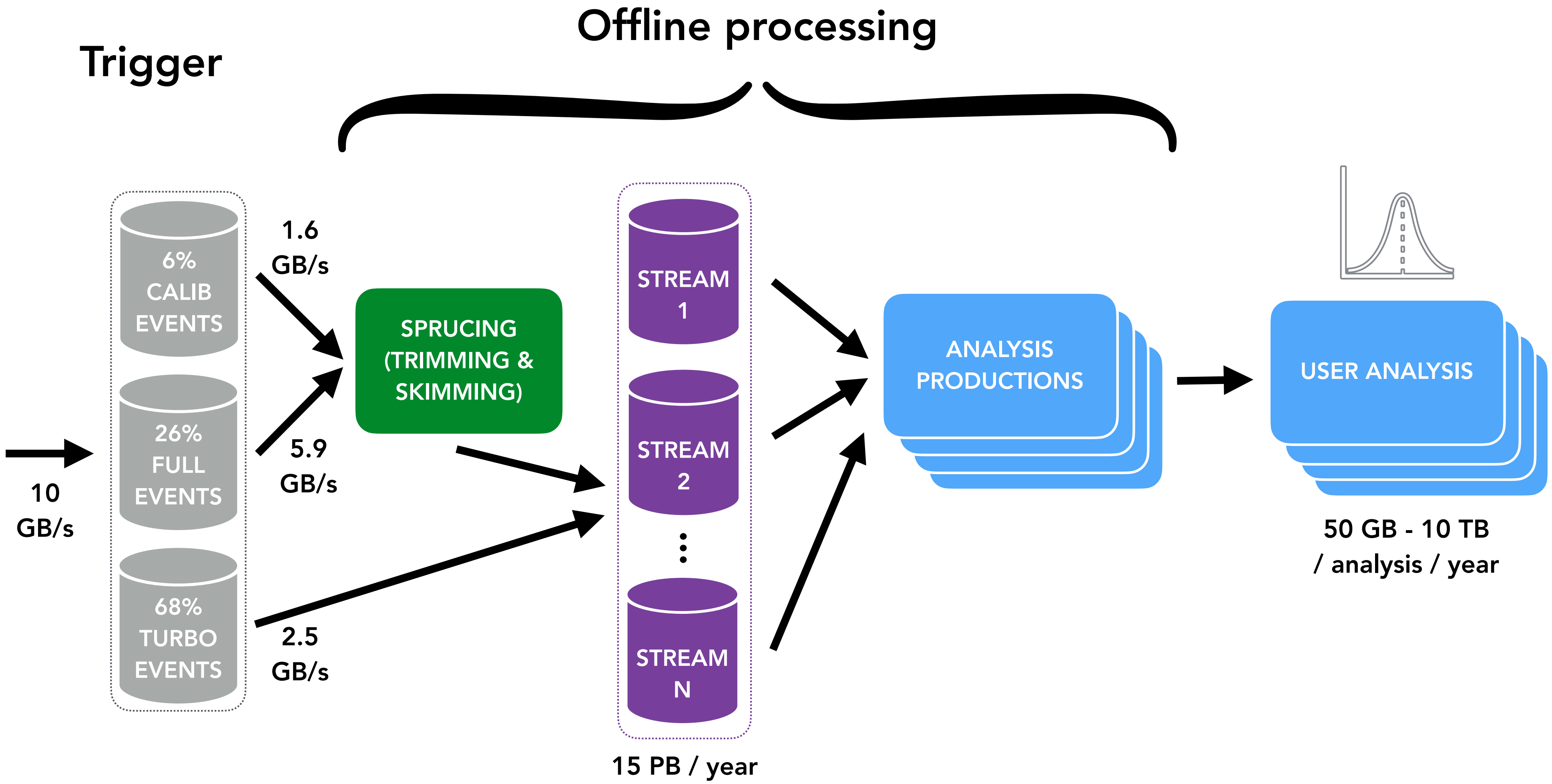
The LHCb Upgrade Dataflow.
All numbers are taken from the LHCb Upgrade Trigger and Online TDR and the LHCb Upgrade Computing Model TDR

The LHCb Upgrade *Online* DataFlow



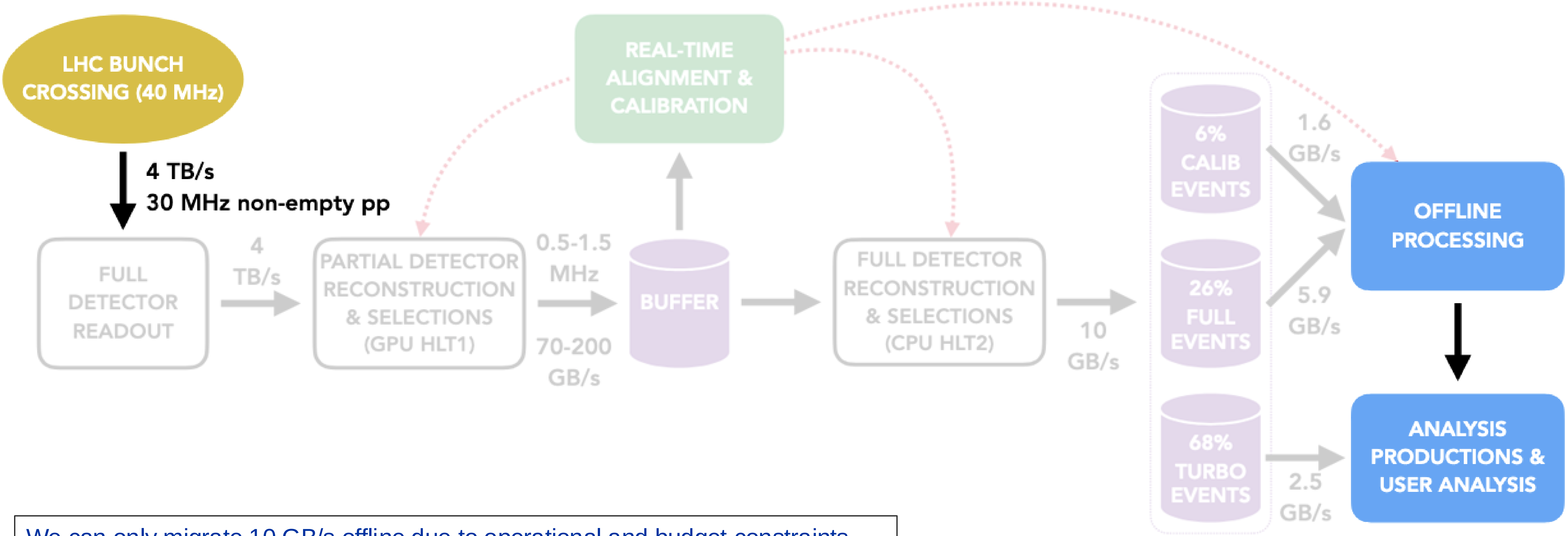
The LHCb Upgrade Dataflow.
All numbers are taken from the LHCb Upgrade Trigger and Online TDR and the LHCb Upgrade Computing Model TDR

The LHCb Upgrade *Offline* DataFlow



The LHCb Upgrade Dataflow.
All numbers are taken from the LHCb Upgrade Trigger and Online TDR and the LHCb Upgrade Computing Model TDR

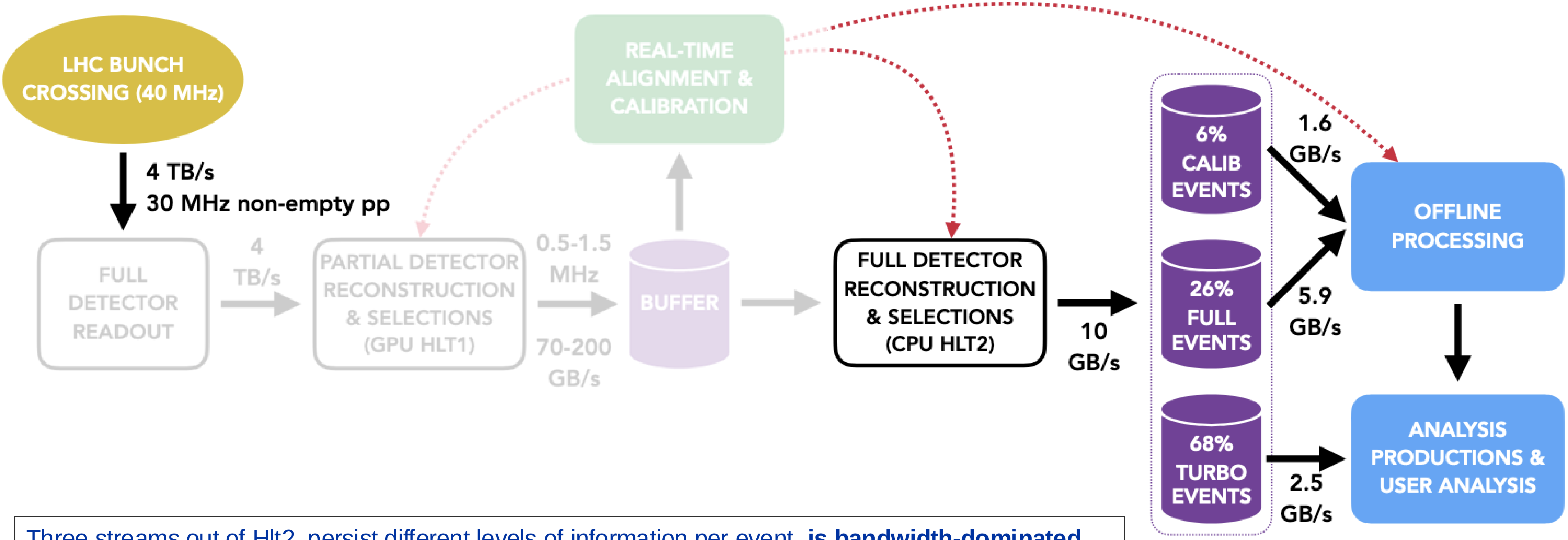
Why do we need a trigger?



We can only migrate 10 GB/s offline due to operational and budget constraints.

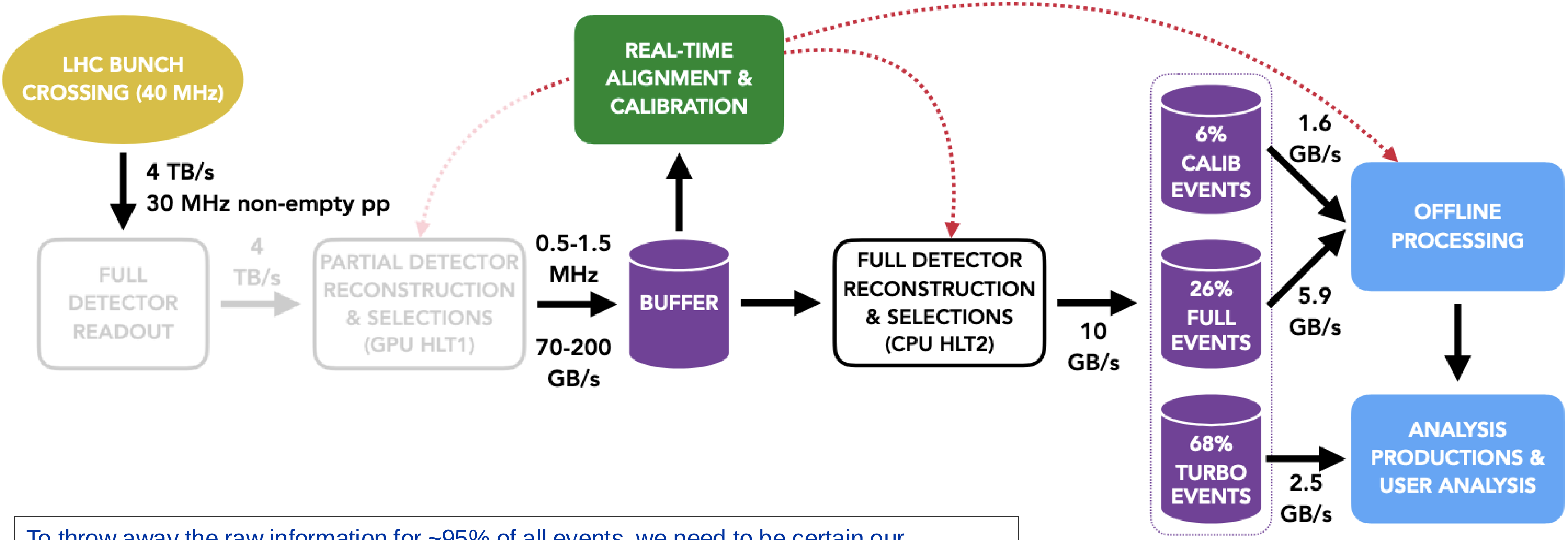
- Find a factor 400 data reduction somewhere.
- Throw away events, and/or by reduce the size of events.

Hlt2



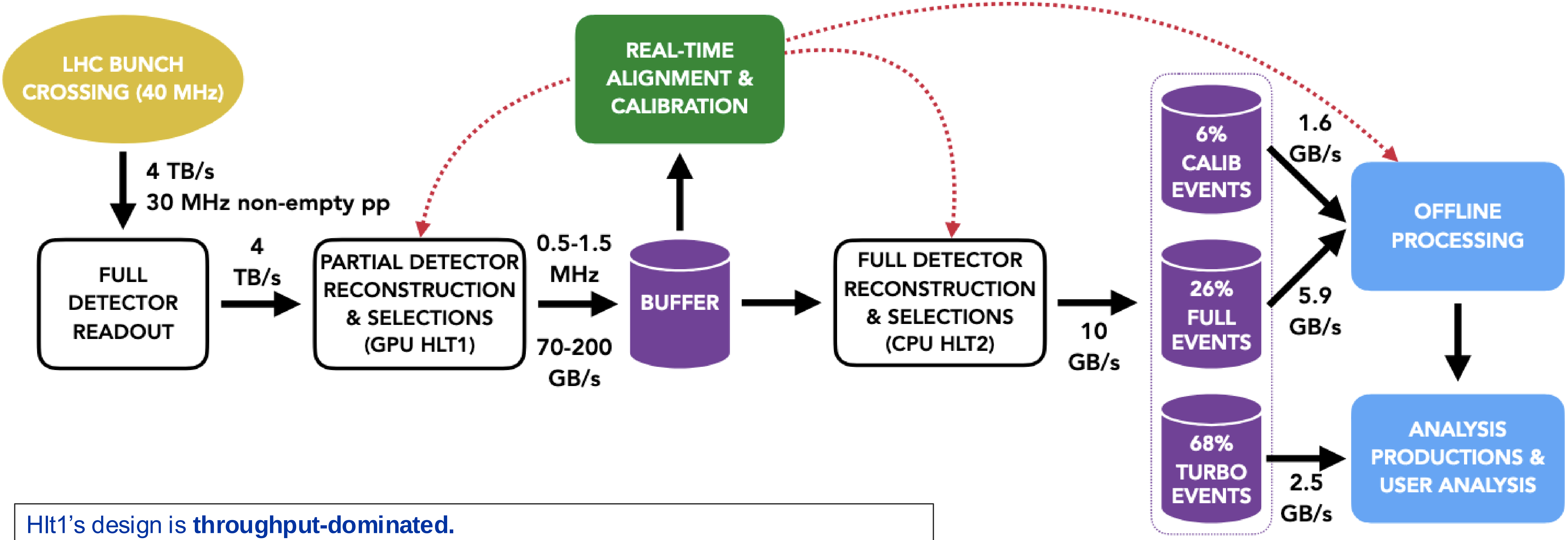
Three streams out of Hlt2, persist different levels of information per event, **is bandwidth-dominated**.
- **Calibration**: all raw banks (i.e. can re-reconstruct tracks offline)
- **Full**: all reconstructed tracks
- **Turbo**: all signal candidate tracks

Online Alignment + Calibration



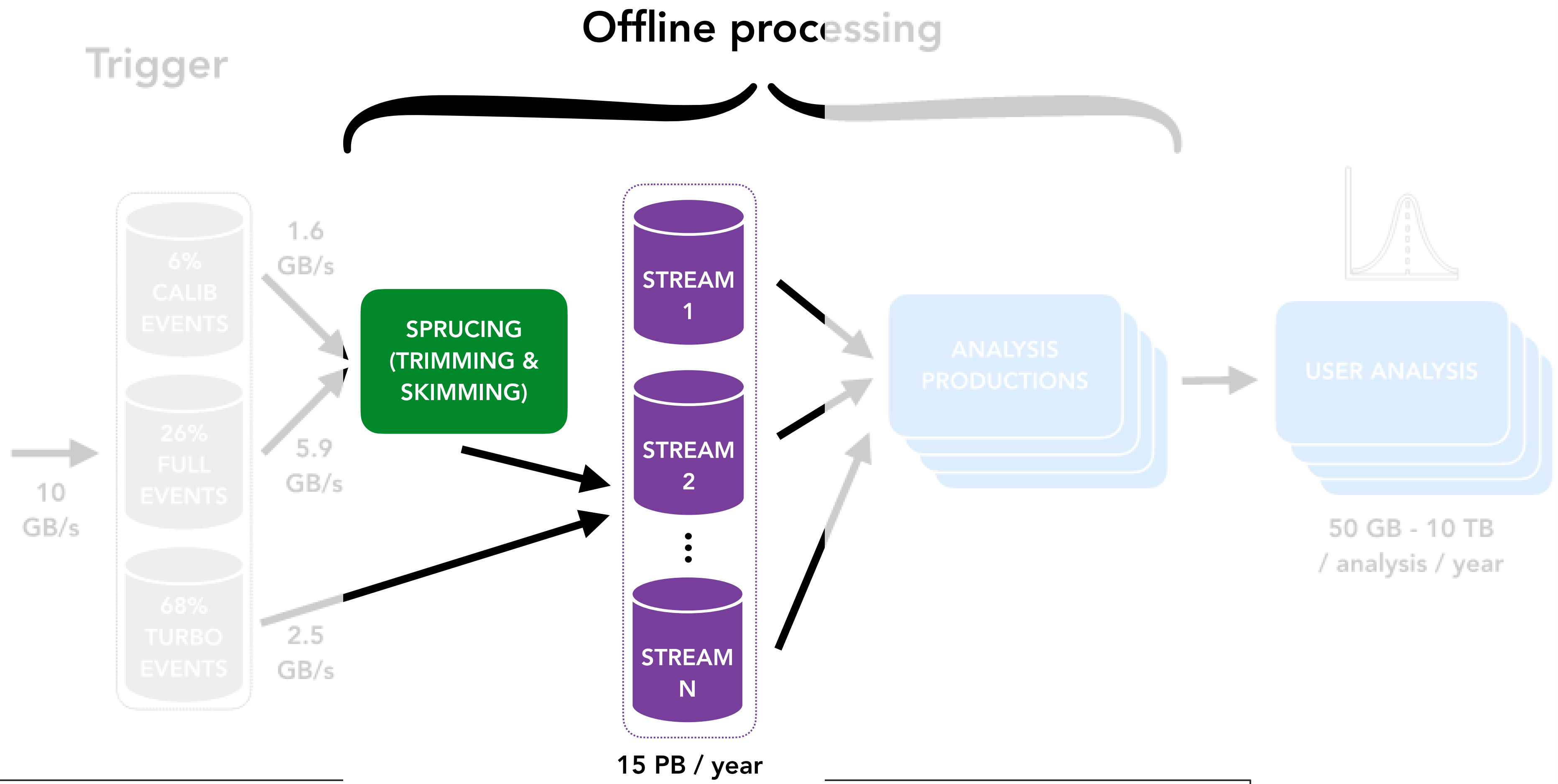
To throw away the raw information for ~95% of all events, we need to be certain our reconstruction is performing at an ***'offline-quality'.***
Only possible by the real-time alignment and calibration performed on the data before HLT2.

Hlt1



Hlt1's design is **throughput-dominated.**
This was to facilitate the removal of the L0, a hardware trigger used in Run2.
Reconstructing tracking and primary-vertex information via GPUs for parallelisation:
- 30:1 reduction in events, removing large hadronic backgrounds
- decays of beauty and charm hadrons can be more efficiently selected

Sprucing



Non-destructive further processing of data migrated offline (HLT2 output).
- Removal of persisted information per-event + further selections to reduce events
- Output stored 'to tape', i.e. constantly accessible by analysts. **bandwidth-dominated**
- Run concurrently to data-taking, and also in re-processing campaigns (for example: End-of-Year ReSprucing 2024)

Hlt1

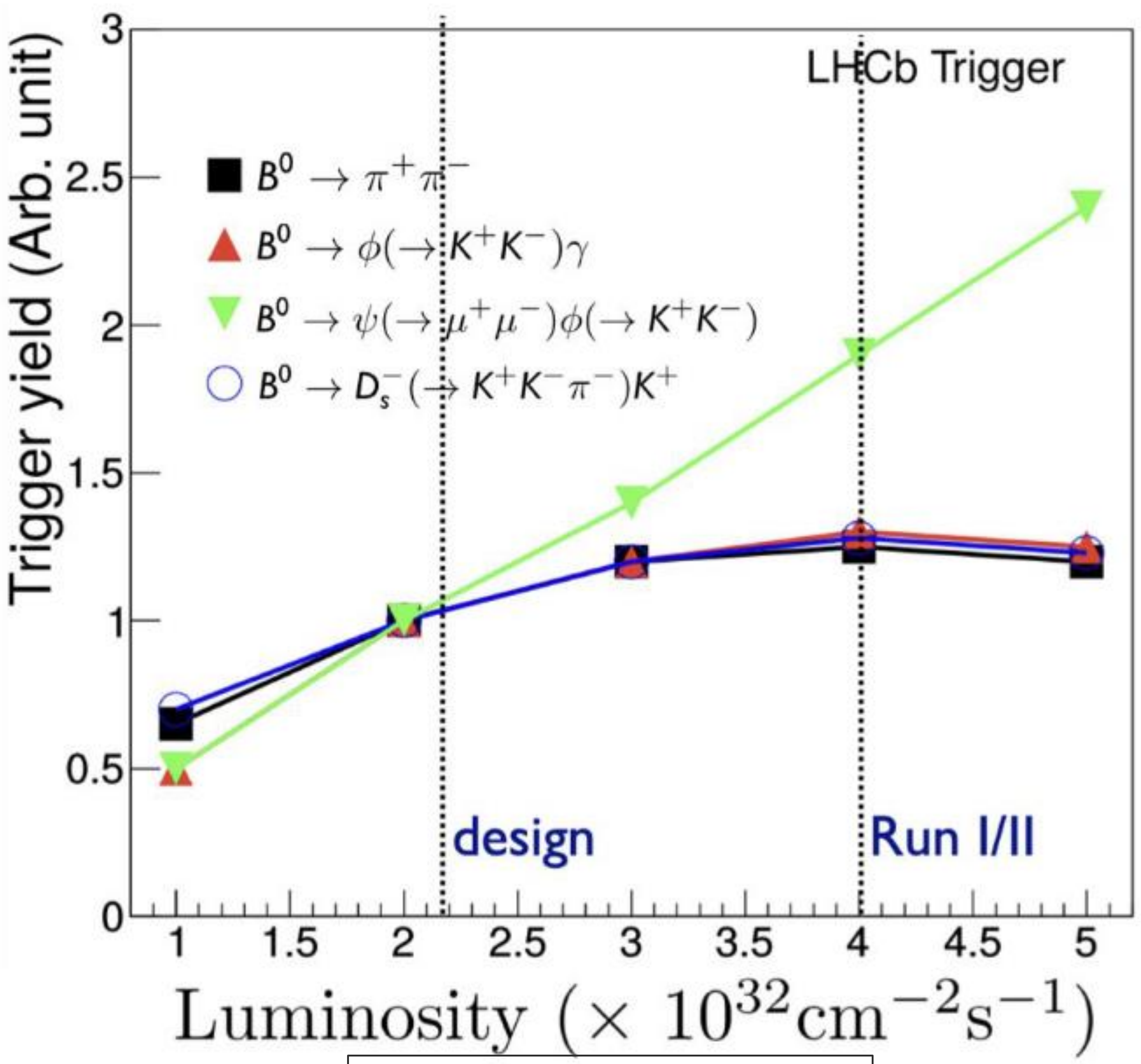
Issues of the L0 Trigger in Run1/2

The L0 trigger was a *hardware* trigger used during Run1 and 2, which ran before HLT1. (*”really fast electronics”)*

It selected high p_T, E_T signatures and reduced the rate from 30 to 1 MHz.

Unfortunately caused significant inefficiencies for heavy flavour modes ☹, and hadronic signatures saturated.

The luminosity increased even further for Run3 to $2 \ * 10^{33}$, i.e., a factor 5 further increase but with a much reduced physics gain…



J. Phys. Conf. Ser. **878** 012012

Fully-Software Triggers

Removing the hardware trigger
- real-time analysis (RTA) for the full selection of data.
- Novel fully-software triggers!
- More holistic, flexible and efficient

Consequences:
- Full detector readout at 40 MHz
- HLT1 reconstruction to run at 30 MHz

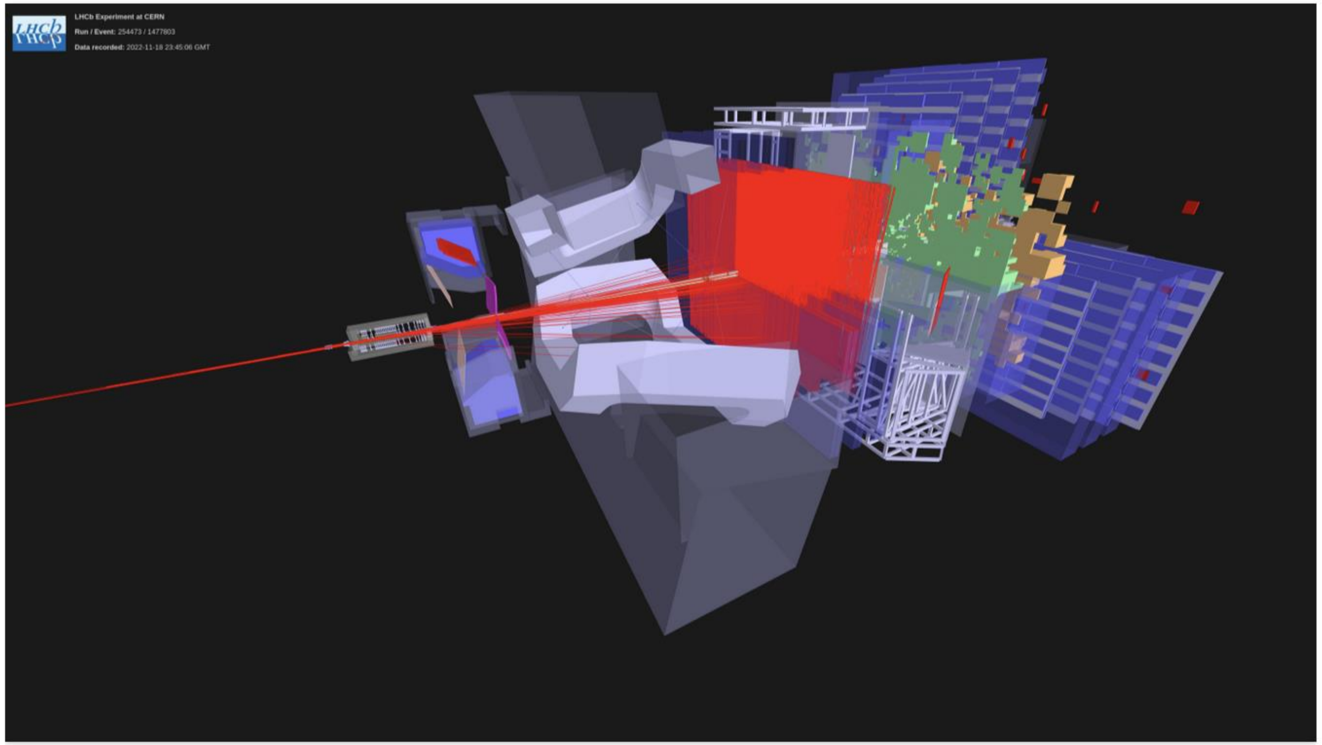
Required a huge effort to upgrade DAQ:
- **~1 million electronic channels**
- **~500 custom aggregating cards**

~ ~150 Computer servers

LHCb begins using unique approach to process collision data in real-time

Using a new system called real-time analysis, the LHCb collaboration has made filtering and analysing experiment data simpler and faster

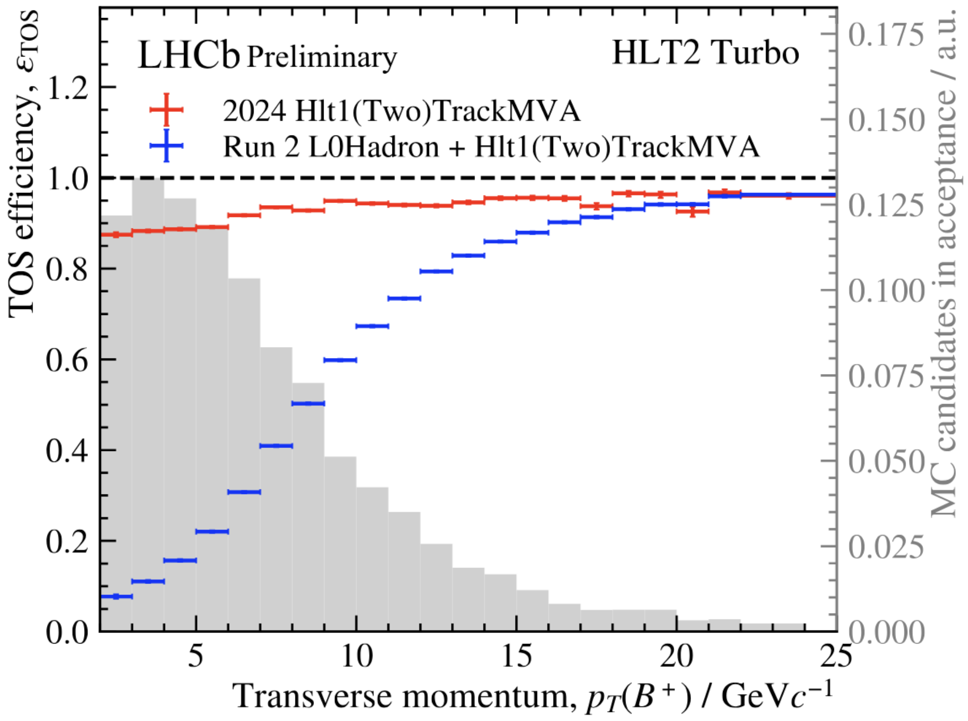
1 MARCH, 2023 | By LHCb collaboration



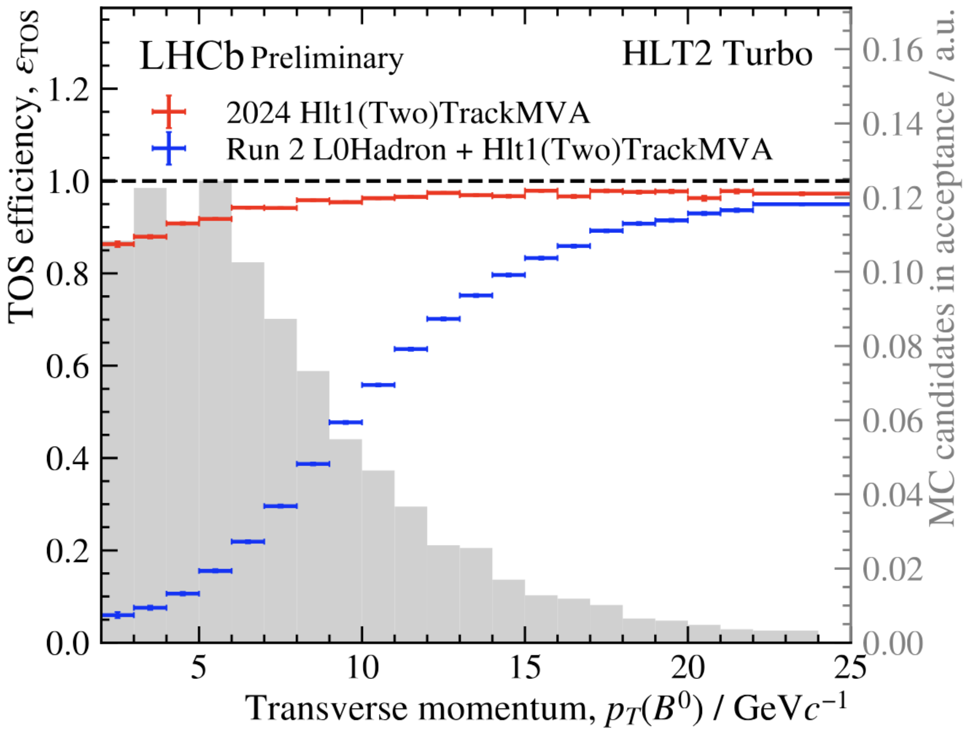
(Image: CERN)

home.web.cern.ch/

Was it worth it?



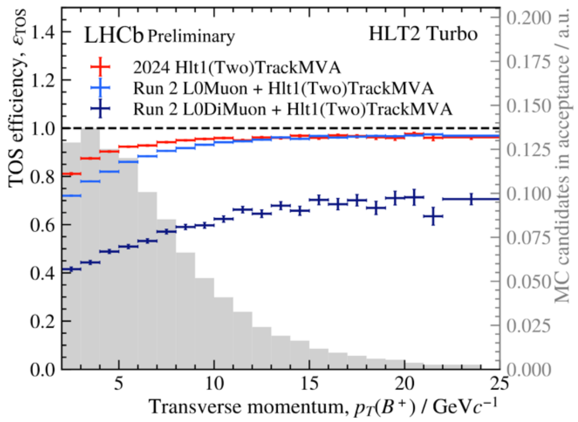
(a) TOS efficiencies in $B^+ \to \overline{D}^0 (K^+\pi^-)\pi^+$.



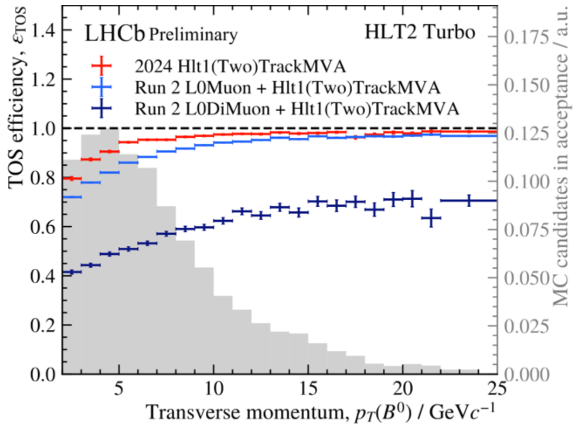
(b) TOS efficiencies in $B^0 \to D^- (K^+\pi^-\pi^-)\pi^+$.

LHCB-FIGURE-2024-030

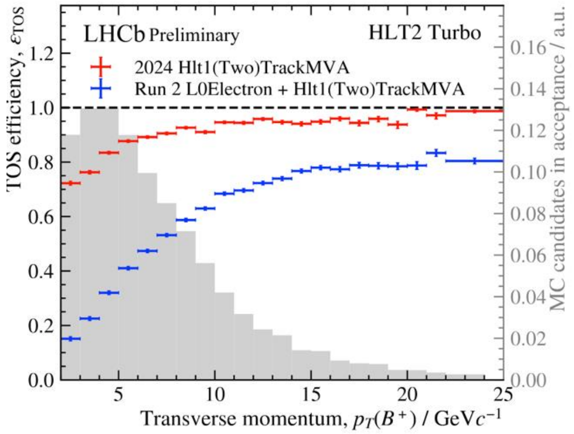
Was it worth it?



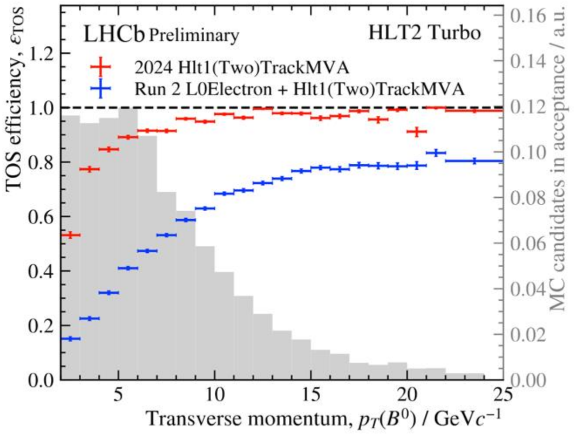
(a) TOS efficiencies in $B^+ \to J/\psi\,(\mu^+\mu^-)\,K^+$.



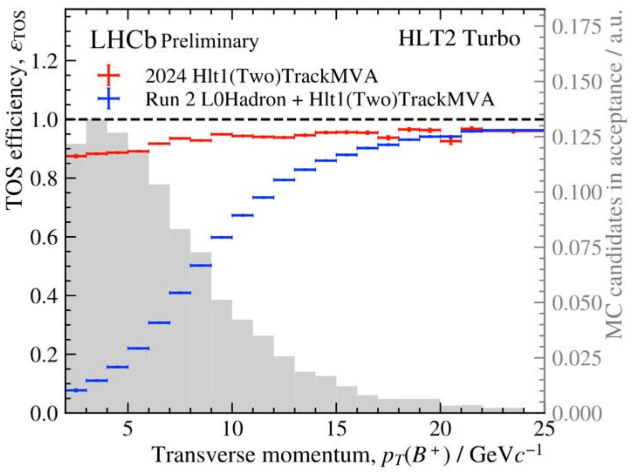
(b) TOS efficiencies in $B^0 \to J/\psi\,(\mu^+\mu^-)\,K^{*0}$.



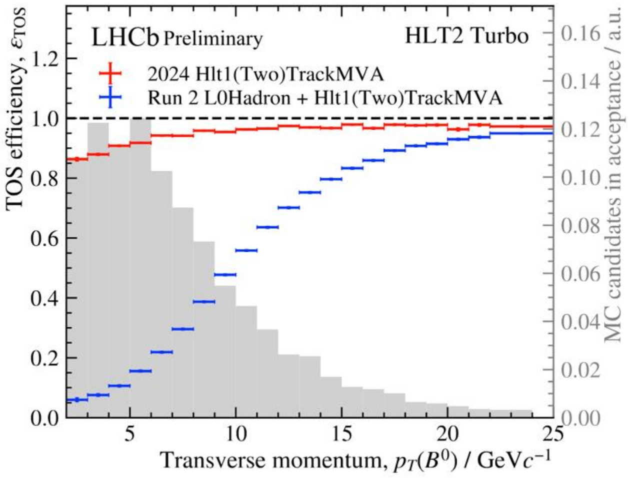
(a) TOS efficiencies in $B^+ \to J/\psi\,(e^+e^-)\,K^+$.



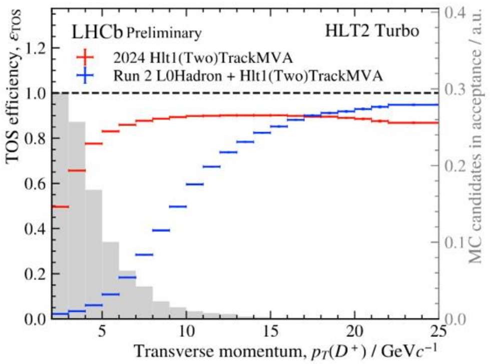
(b) TOS efficiencies in $B^0 \to J/\psi\,(e^+e^-)\,K^{*0}$.



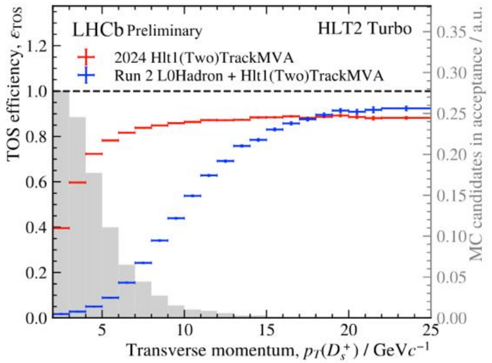
(a) TOS efficiencies in $B^+ \to \overline{D}^0\,(K^+\pi^-)\,\pi^+$.



(b) TOS efficiencies in $B^0 \to D^-\,(K^+\pi^-\pi^-)\,\pi^+$.

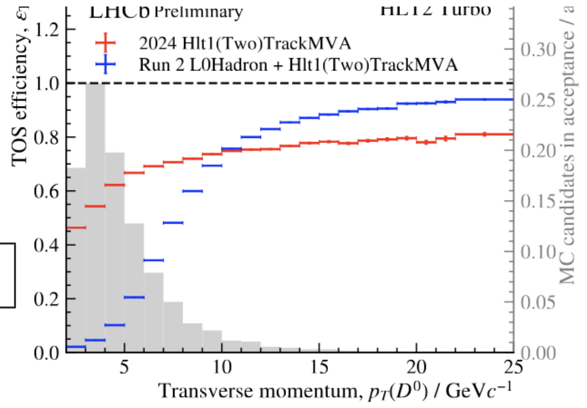


(a) TOS efficiencies in $D^+ \to K^-\pi^+\pi^+$.



(b) TOS efficiencies in $D_s^+ \to K^+K^-\pi^+$.

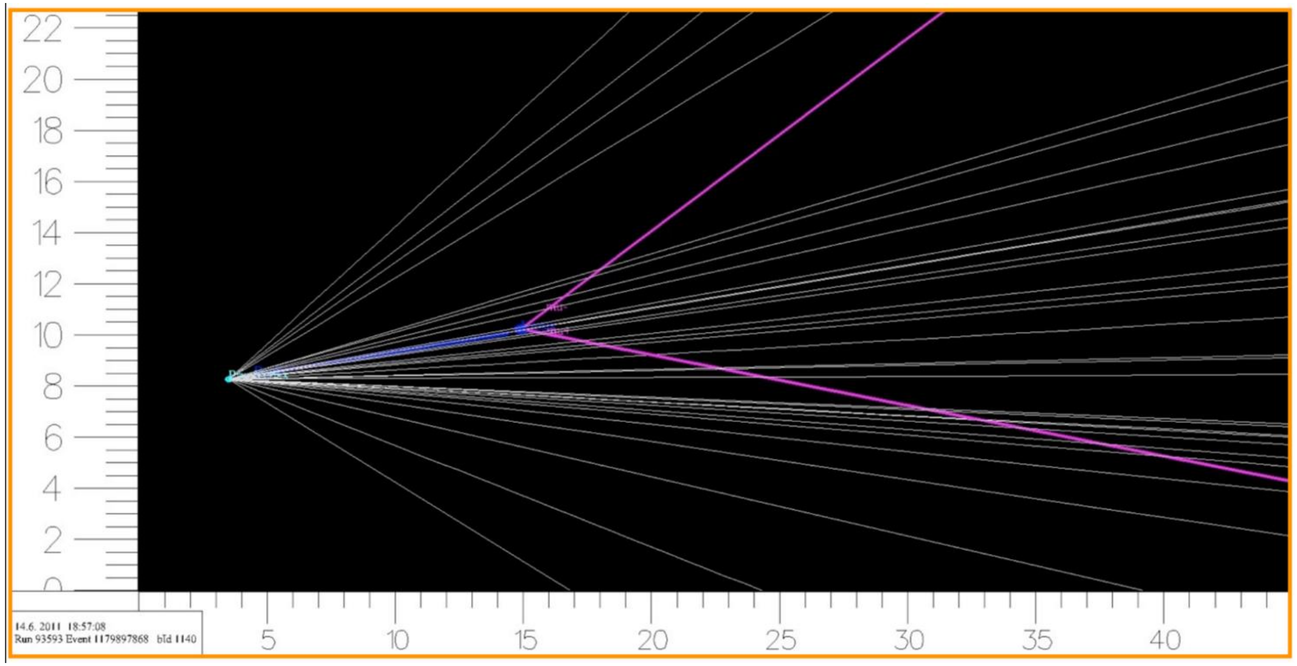
LHCB-FIGURE-2024-030



(c) TOS efficiencies in $D^0 \to K^-\pi^+$.

Partial Reconstruction

What is done:
 - Track reconstruction: trajectories of charged particles inside LHCb tracking detectors
 - PV Reconstruction: extrapolating reconstructed tracks back to the collision point.
 - muon and electron ID: “simplest“ of the particle IDs, possible to do within timing constraint
 - “Simple” trigger algorithms: Up to two-body topological combinations for trigger.



Partial Reconstruction

What is done:
- Track reconstruction: trajectories of charged particles inside LHCb tracking detectors
- PV Reconstruction: extrapolating reconstructed tracks back to the collision point.
- muon and electron ID: “simplest“ of the particle IDs, possible to do within timing constraint
- “Simple” trigger algorithms: Up to two-body topological combinations for trigger.

What must be skipped:
- ‘Expensive’ tracking: Complex descriptions of material budget and magnet field interactions.
- hadron ID: RICH reconstruction
- Arbitrarily complex trigger algorithms: N-body topologies

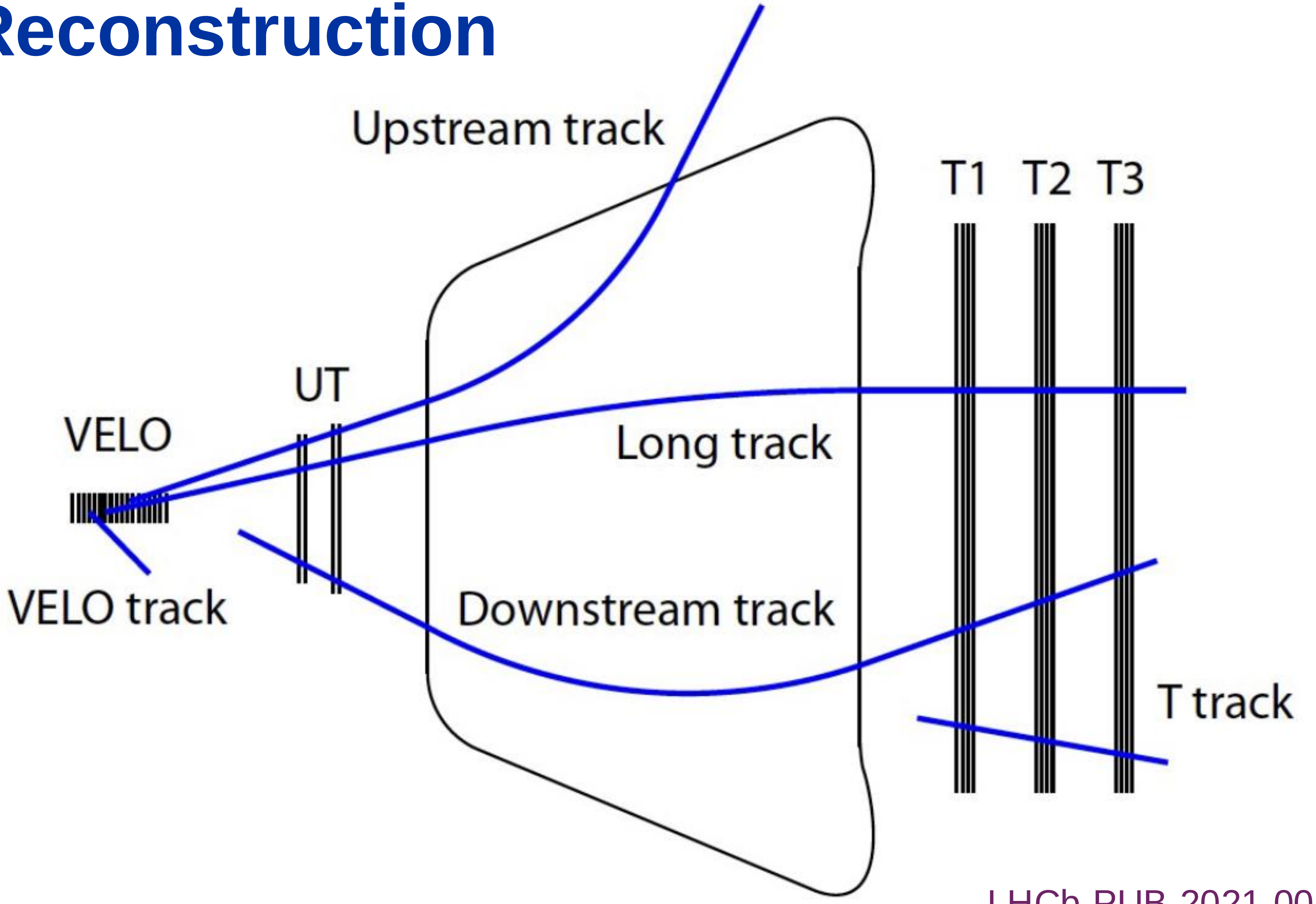
Partial Reconstruction

What is done:
- Track reconstruction: trajectories of charged particles inside LHCb tracking detectors
- PV Reconstruction: extrapolating reconstructed tracks back to the collision point.
- muon and electron ID: “simplest“ of the particle IDs, possible to do within timing constraint
- “Simple” trigger algorithms: Up to two-body topological combinations for trigger.

What must be skipped:
- ‘Expensive’ tracking: Complex descriptions of material budget and magnet field interactions.
- hadron ID: RICH reconstruction
- Arbitrarily complex trigger algorithms: N-body topologies

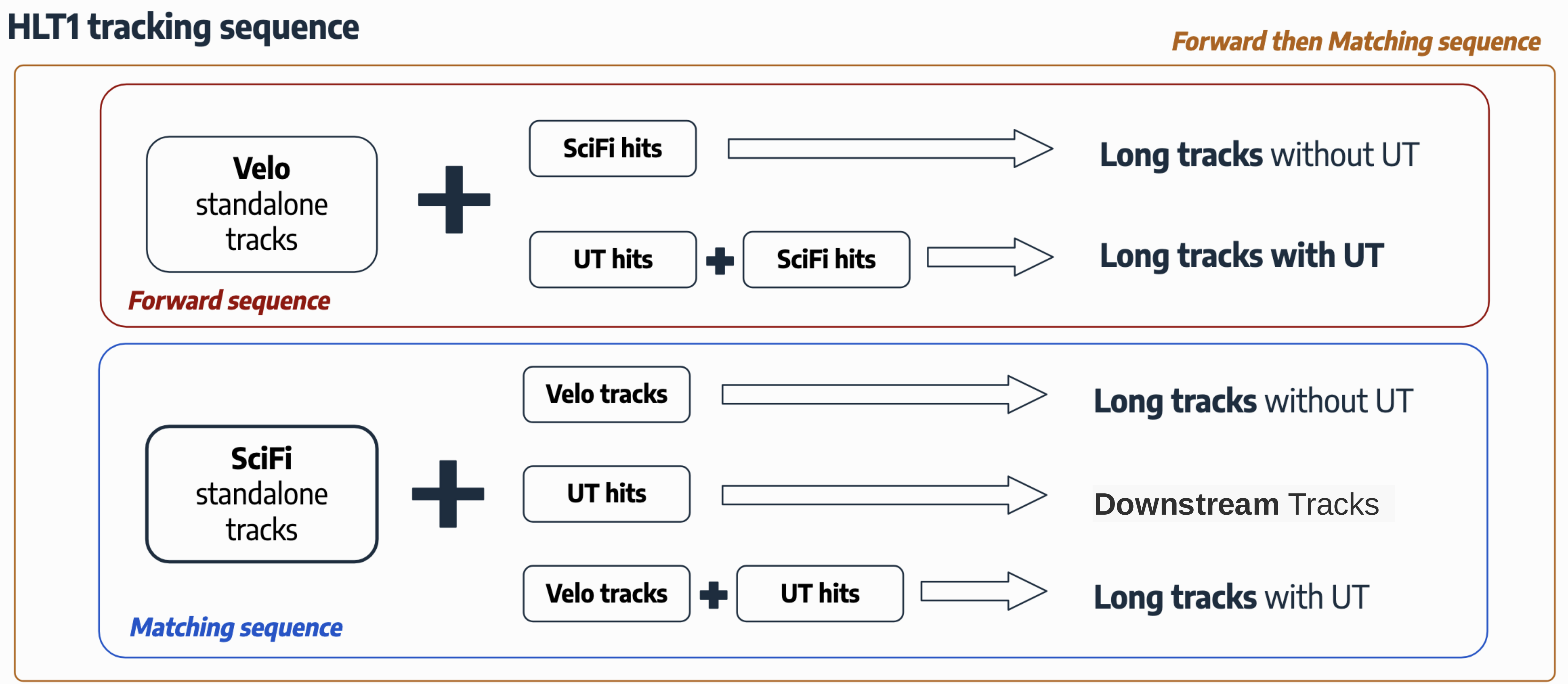
The TDR aimed for 1 MHz output rate, but this year was possible to operate at 20% higher than that!
- Can loosen some important triggers to get more physics potential
- more ‘wiggle-room’ for ‘adventurous’ trigger activities, like *Downstream* tracking.

Track Reconstruction



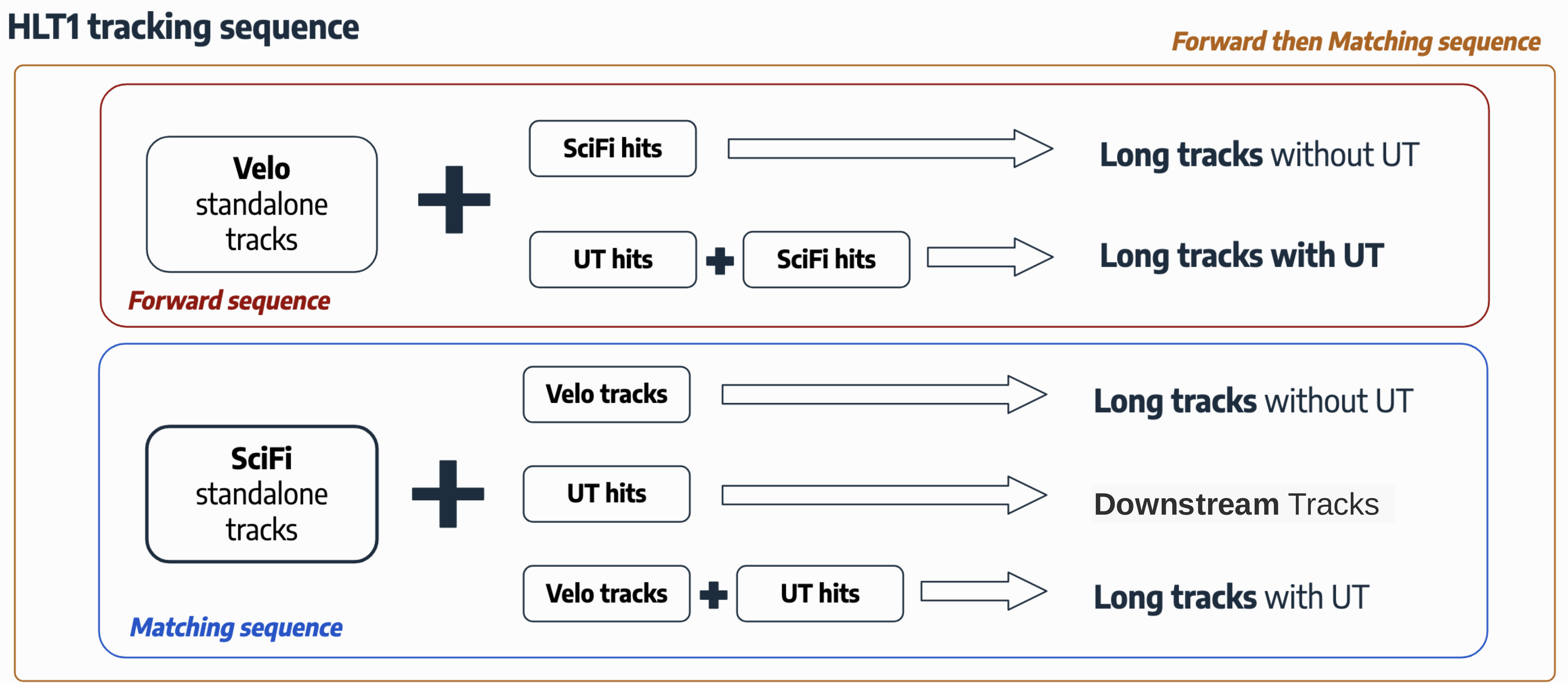
LHCb-PUB-2021-005

Track Reconstruction



J. Zhou's talk @CHEP24. The different Hlt1 tracking sequence. For most of this year *forward_then_matching* was used

Track Reconstruction



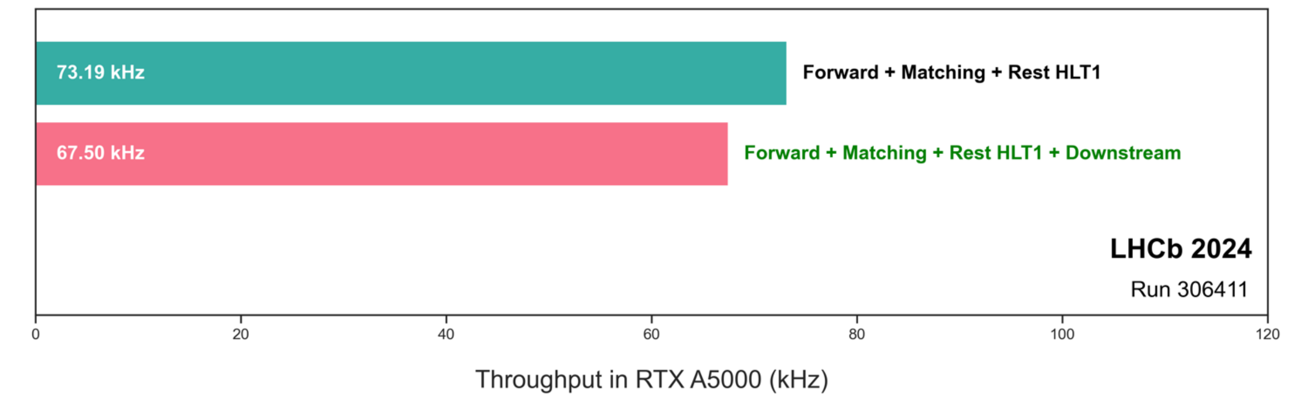
J. Zhou's talk @CHEP24. The different Hlt1 tracking sequence. For most of this year *forward_then_matching* was used

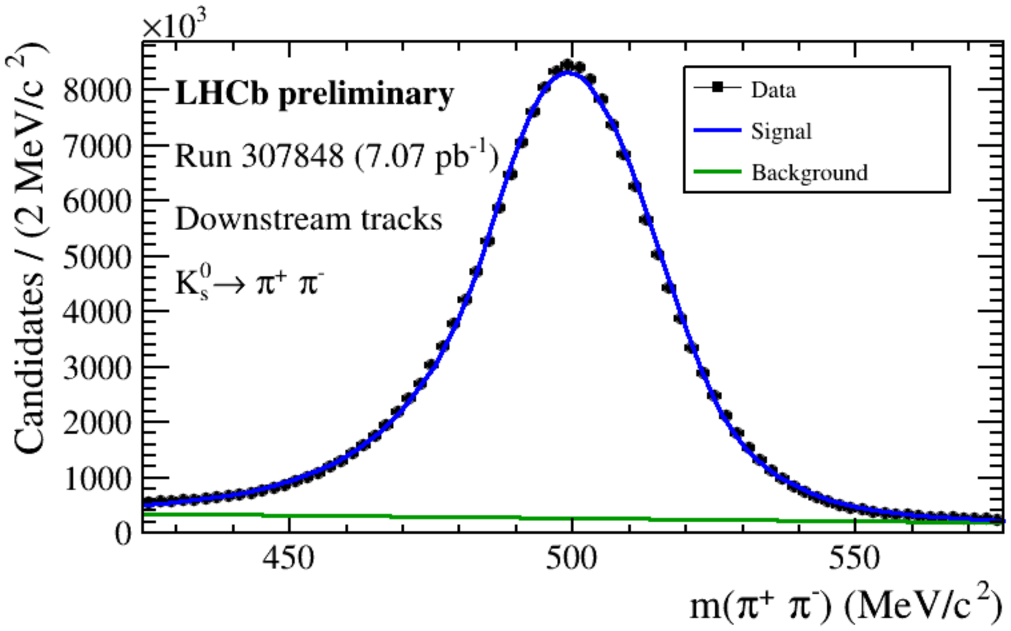
Downstream Tracking

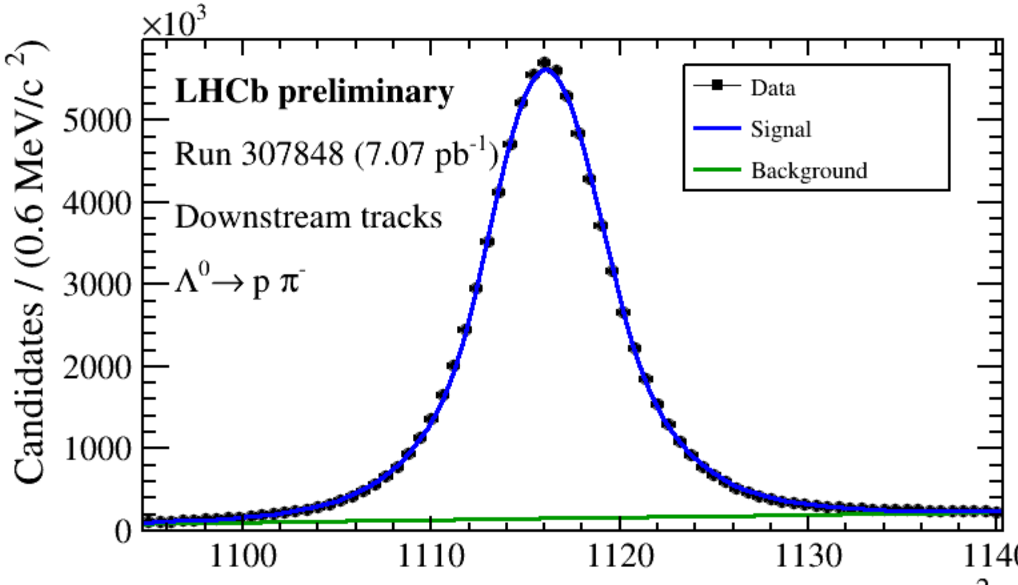
The main purposes of Downstream tracking is to improve the reconstruction efficiencies of decays occurring outside the VELO detector.

Since October, this has been included in data-taking.

For ~10% throughput decrease (still above threshold), can now detect decays that were previously ‘invisible’ to HLT1!
Improved trigger efficiency and exotic/BSM reach.



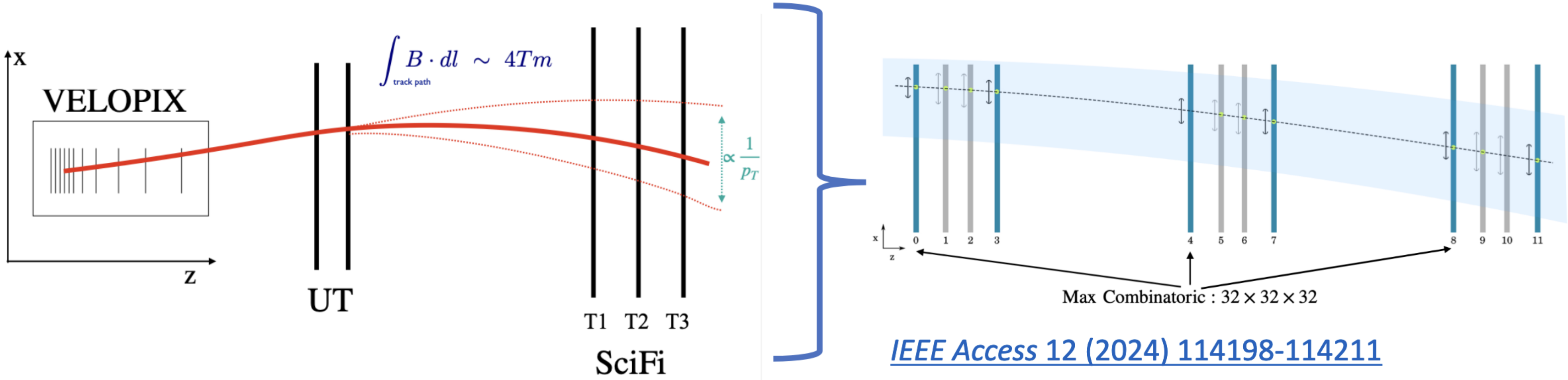




LHCB-FIGURE-2024-035

Reconstruction in GPU

- How to fully exploit parallelization power of GPUs?
- Parallelization levels when reconstructing tracks traversing the whole LHCb detector:
  1. Over events, independent p-p collisions
  2. Over input tracks, extrapolate straight tracks in VELO+UT into the magnetic field reaching the SciFi
  3. Over hits in SciFi, meaning possible extrapolations segments



IEEE Access 12 (2024) 114198-114211

21.10.24 Alessandro Scarabotto - LHCb Run3 trigger 22

A. Scarabotto's talk @CHEP24. Performance of the LHCb heterogeneous software trigger

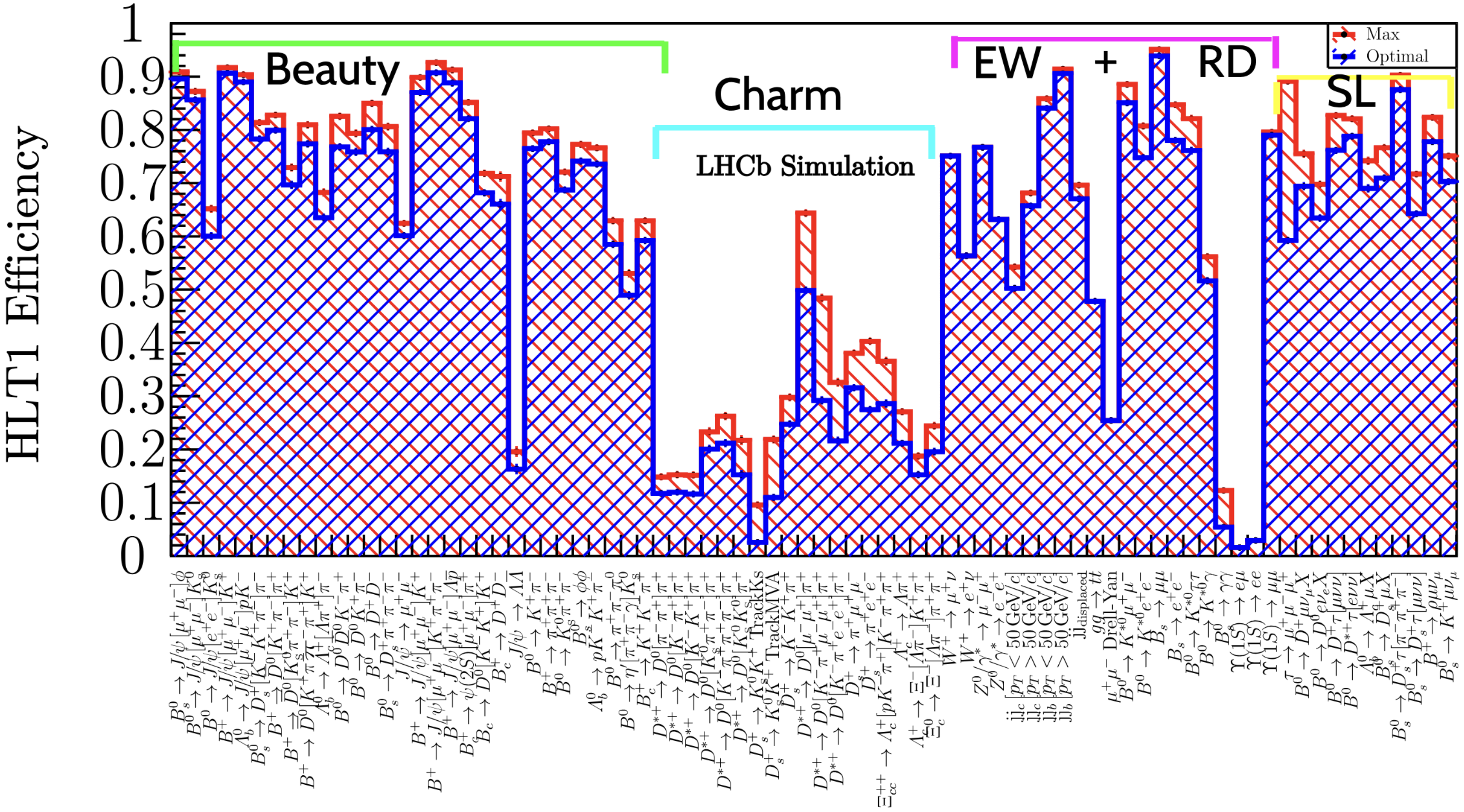
How to share Rate (Bandwidth) fairly?

~50 lines, broad physics.

An automated procedure to determine trigger parameters that maximises the physics output within the Rate constraints.

Allows fast turnaround for re-optimisations as data-taking conditions change.

What does it mean to share fairly in this context?
The *Physics Planning Group* can provide weights prioritising certain lines according to the experiment's interests.

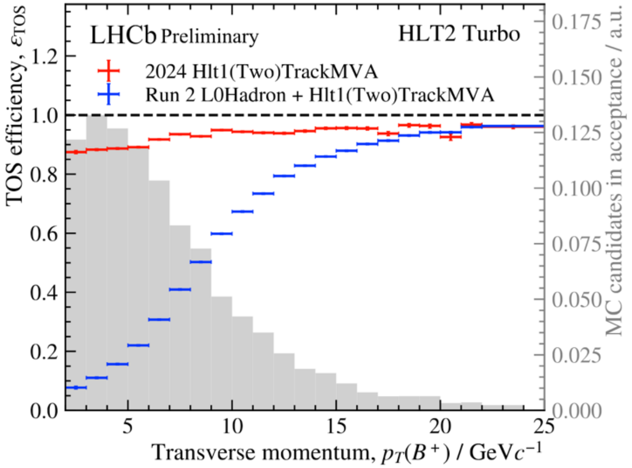


J. Horswill talk @CHEP24. MC Reconstructible Efficiencies for the lines considered in the Hlt1 bandwidth automation.

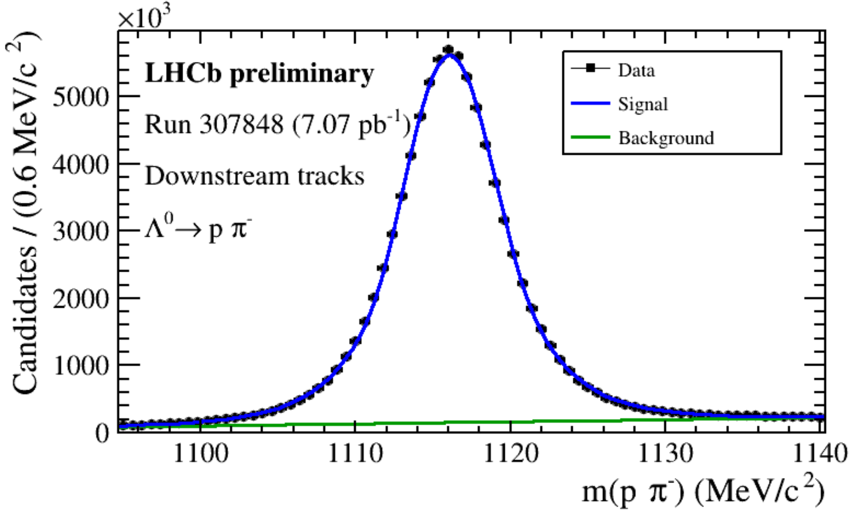
Hlt1 summary

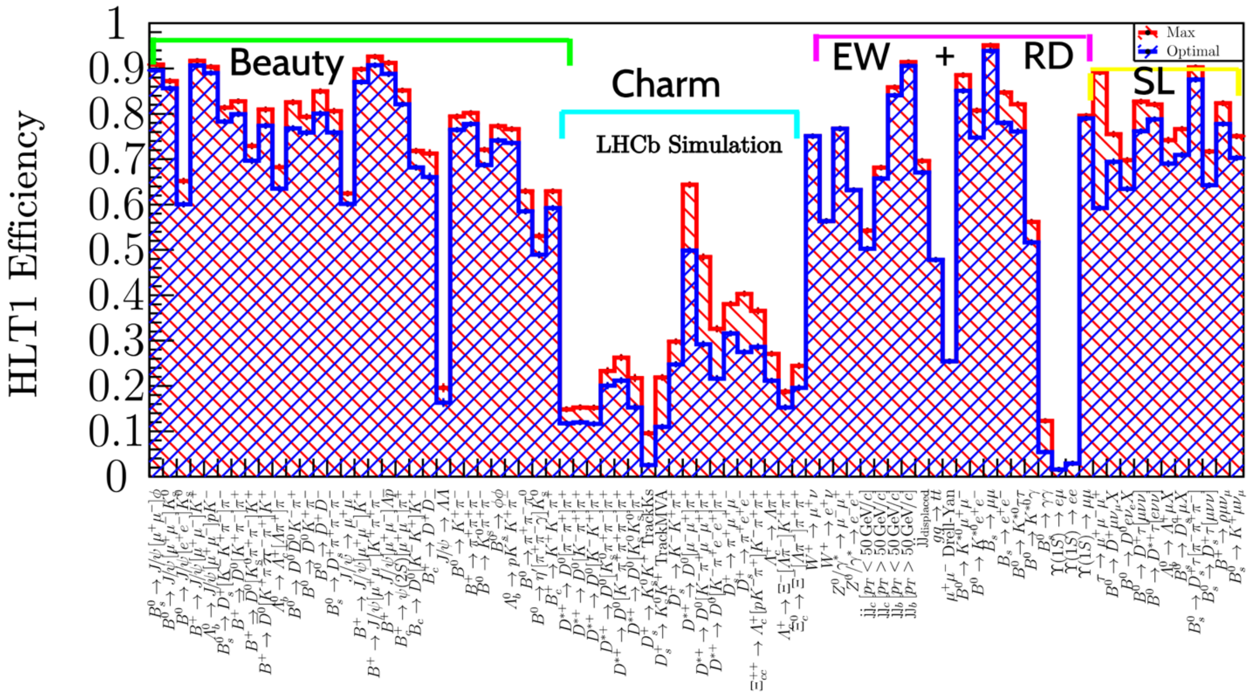
Meets the throughput demands since removing the L0 hardware trigger via:
- GPU parallelisation
- Partial reconstruction

Even with this high demand, we were able to run at a higher Rate and Throughput than originally designed.
- Higher efficiencies than Run1/2
- Automated fair division of rate
- Interesting new physics possibilities.



(a) TOS efficiencies in $B^+ \to \bar{D}^0 (K^+\pi^-)\pi^+$.





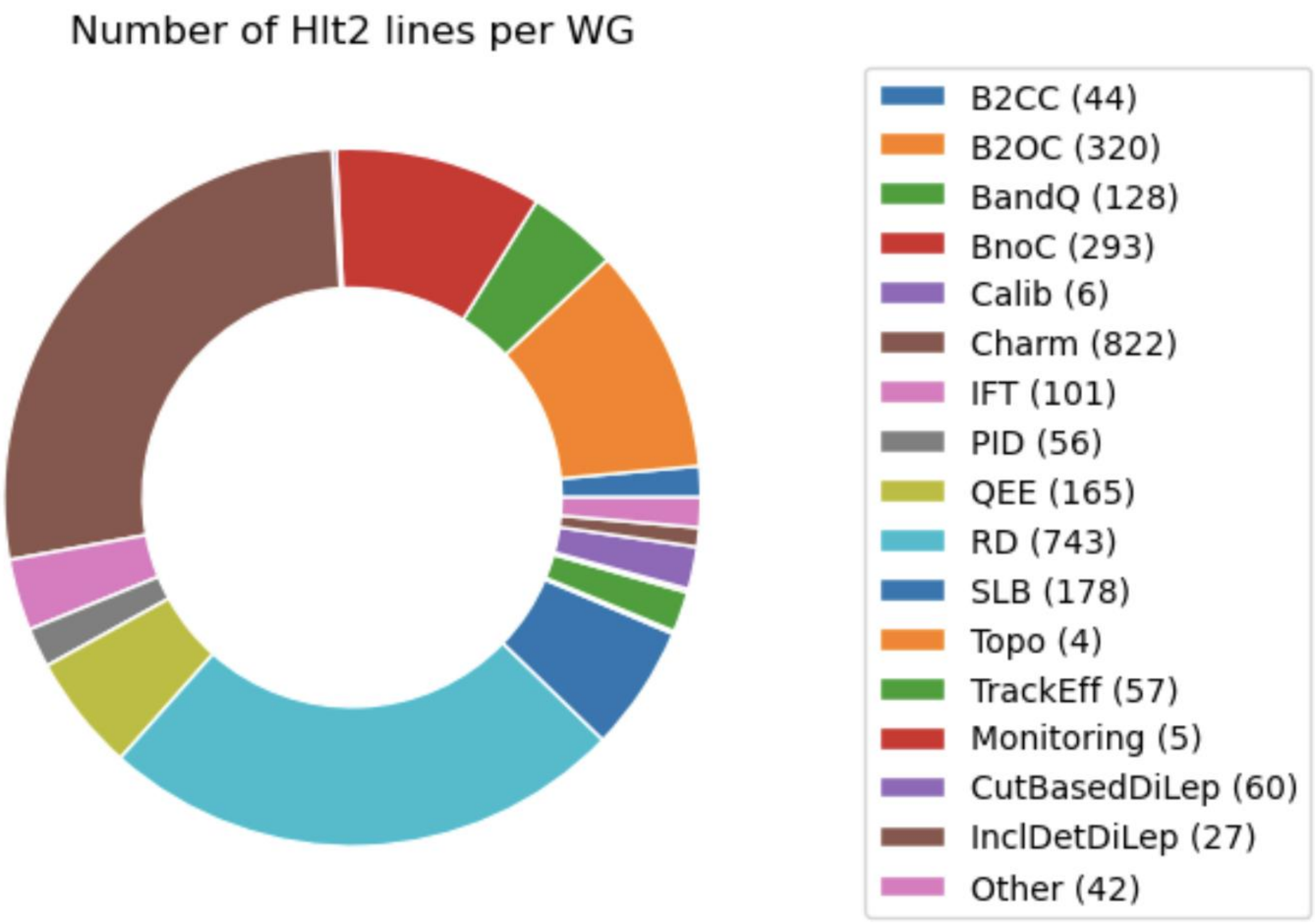
Hlt2 and the Persistency Model

Breadth

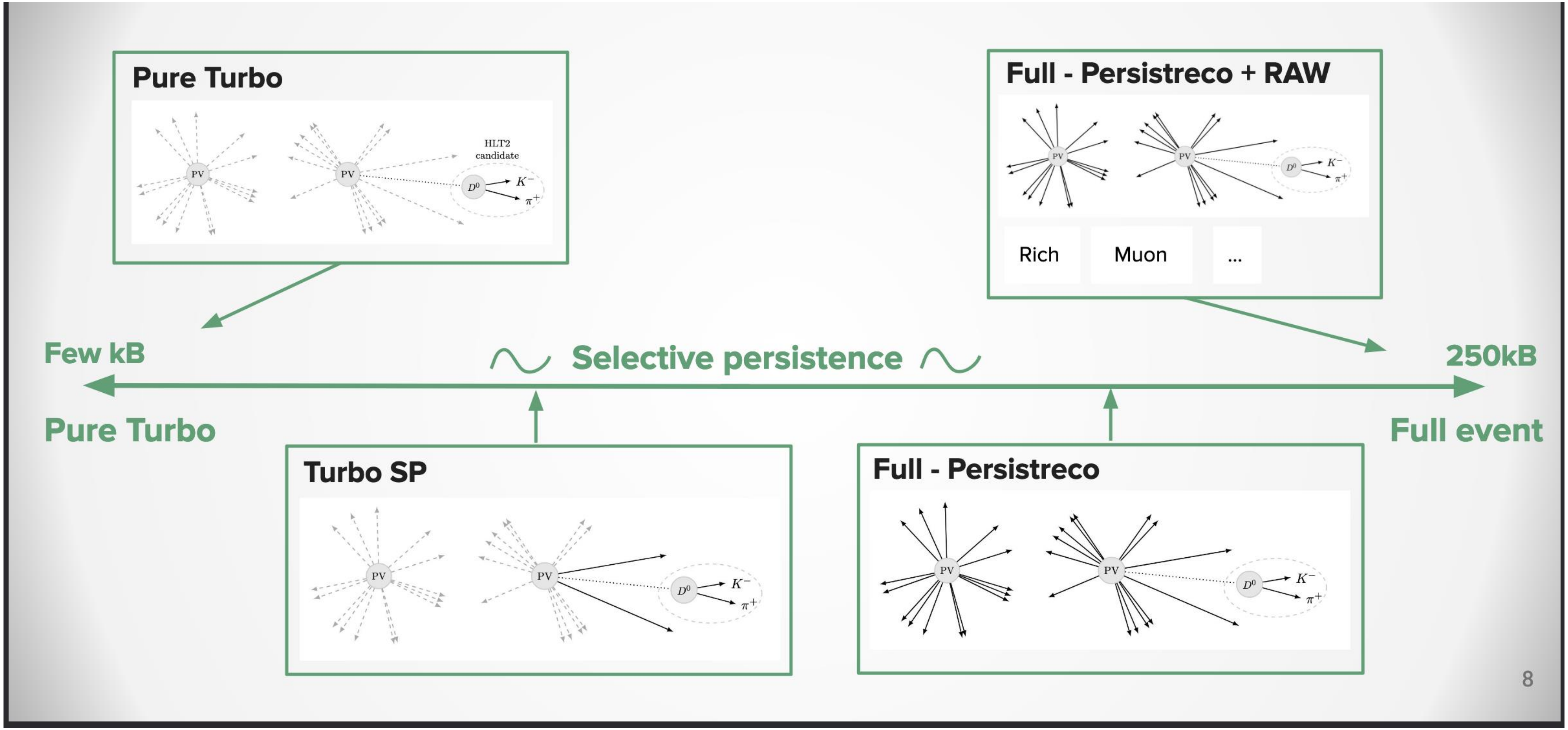
3056 lines across the collaboration, ranging across all of the physics programme.
- ~10 working groups
- 100s of authors,
- 10,000s of algorithms.

While HLT1 can automate its bandwidth division, due to the sheer dimensionality, Hlt2 (&Sprucing) require division 'by hand'.

The PPG provide limits of bandwidth per physics working group.



Persistency Model



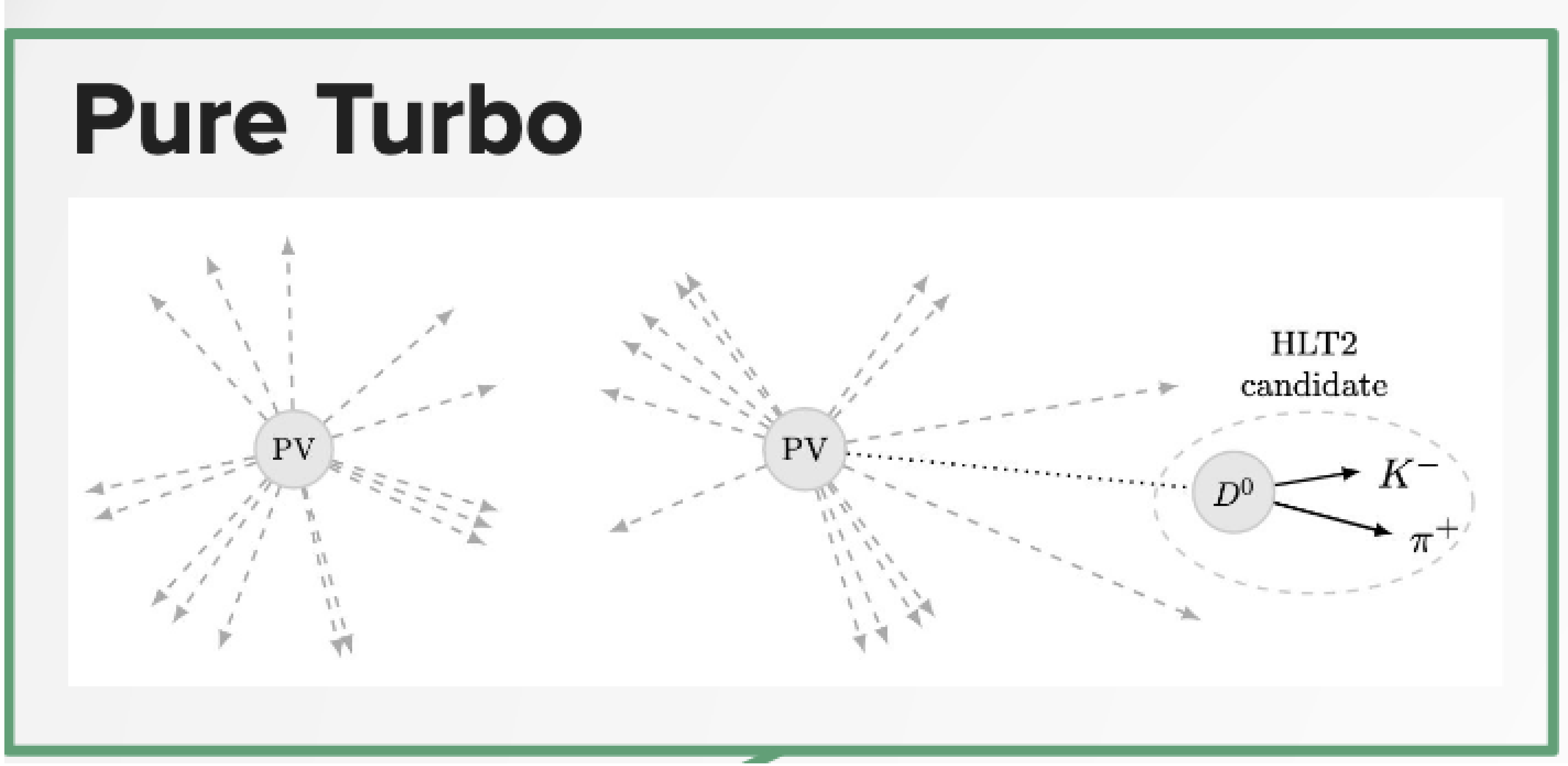
N. Skidmore's talk @CHEP24. By trimming information within an event, we can reduce file size.

Persistency Model

The ‘cheapest’ persistency.

Envisioned to contain ~73% of the physics content for the LHCb.

Approx. a 25:1 reduction in data size

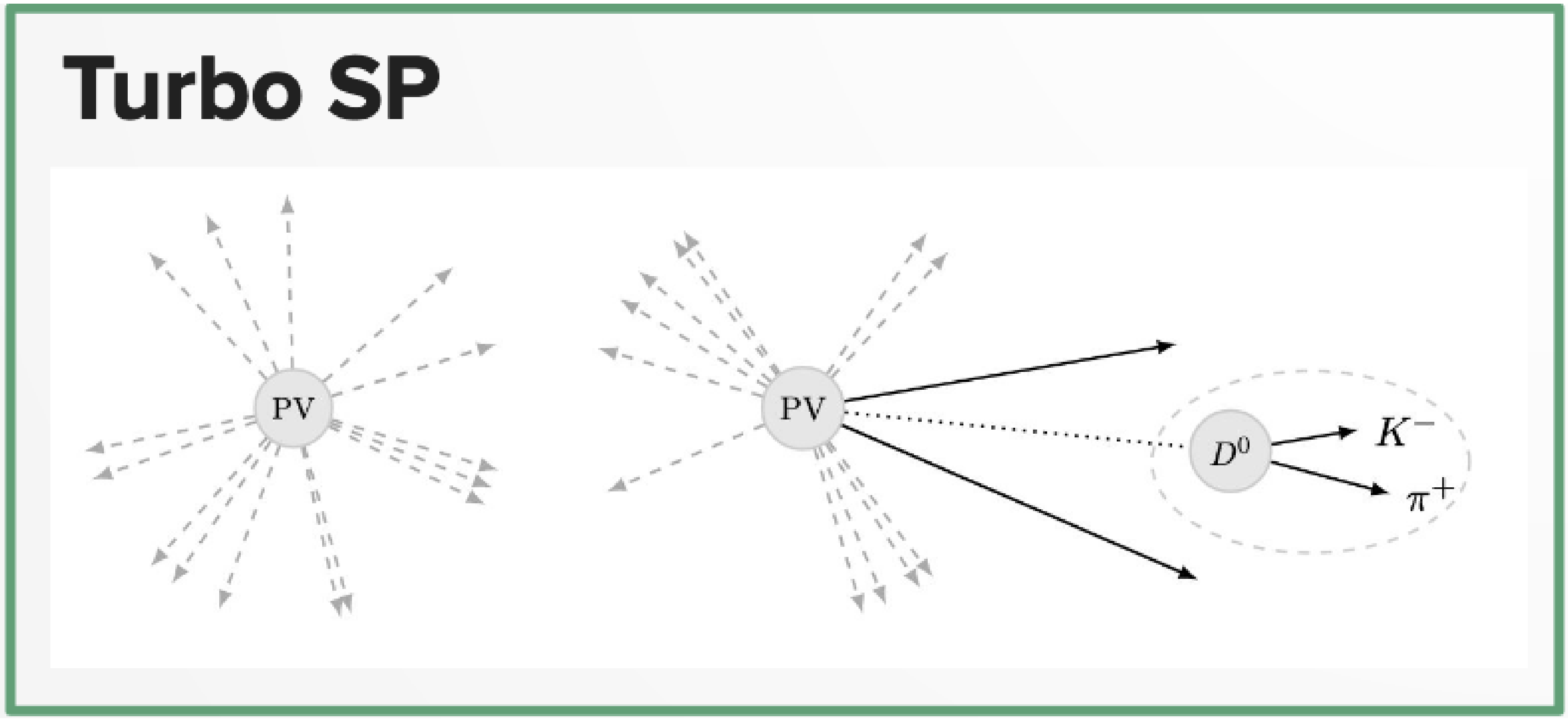


Persistency Model

Turbo + selective persistence

Can keep extra from the event per-candidate.
For instance:
 - Keeping particles in a cone around the signal to calculate ‘Isolation’.

Approx. 3:1 data reduction, dependent on what is kept.
Has to be justified



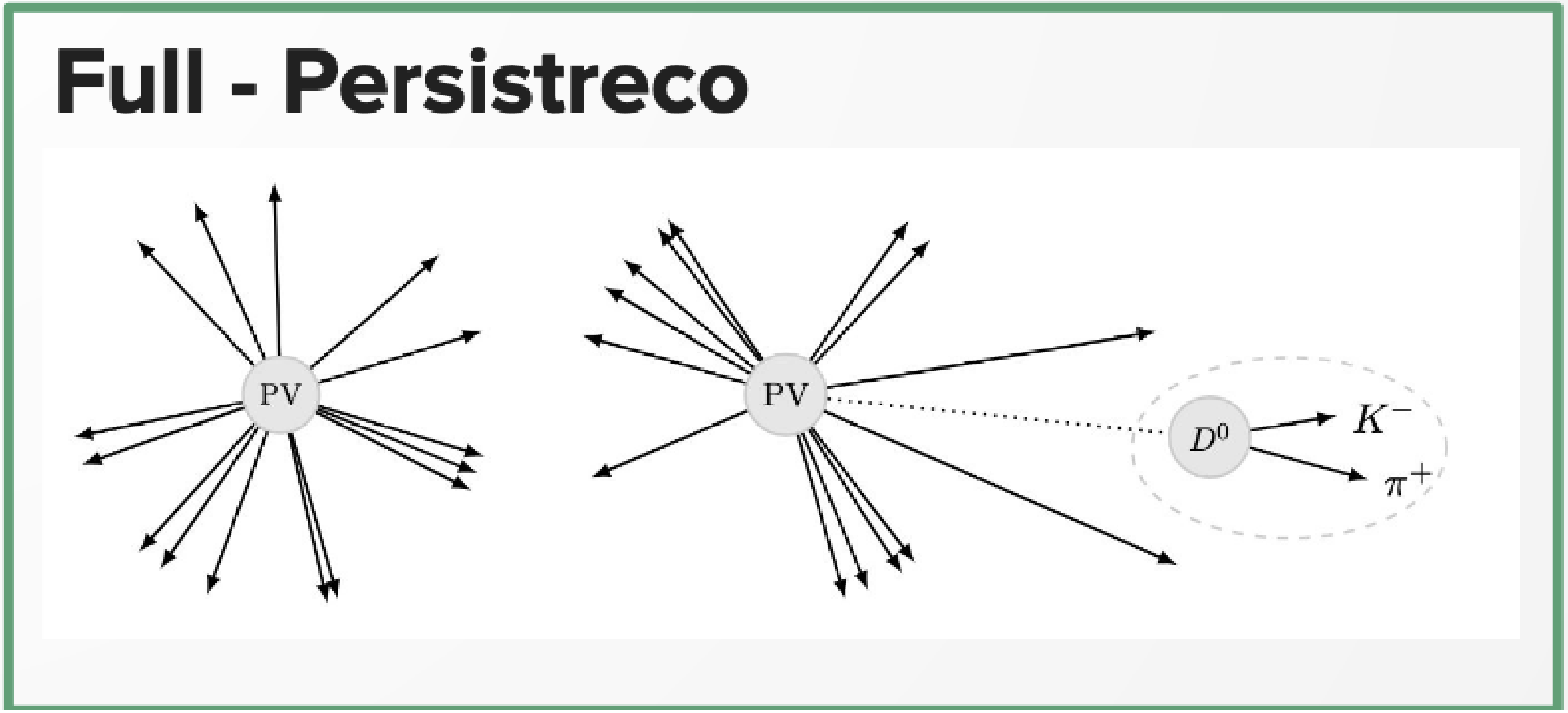
Persistency Model

Full persistence

Keeps all reconstructed tracks from the event.

Approx. 3:2 data reduction. Requires further selections @Sprucing.

Common use case of “Inclusive” decay modes. i.e. lines that capture several physics decays at once.

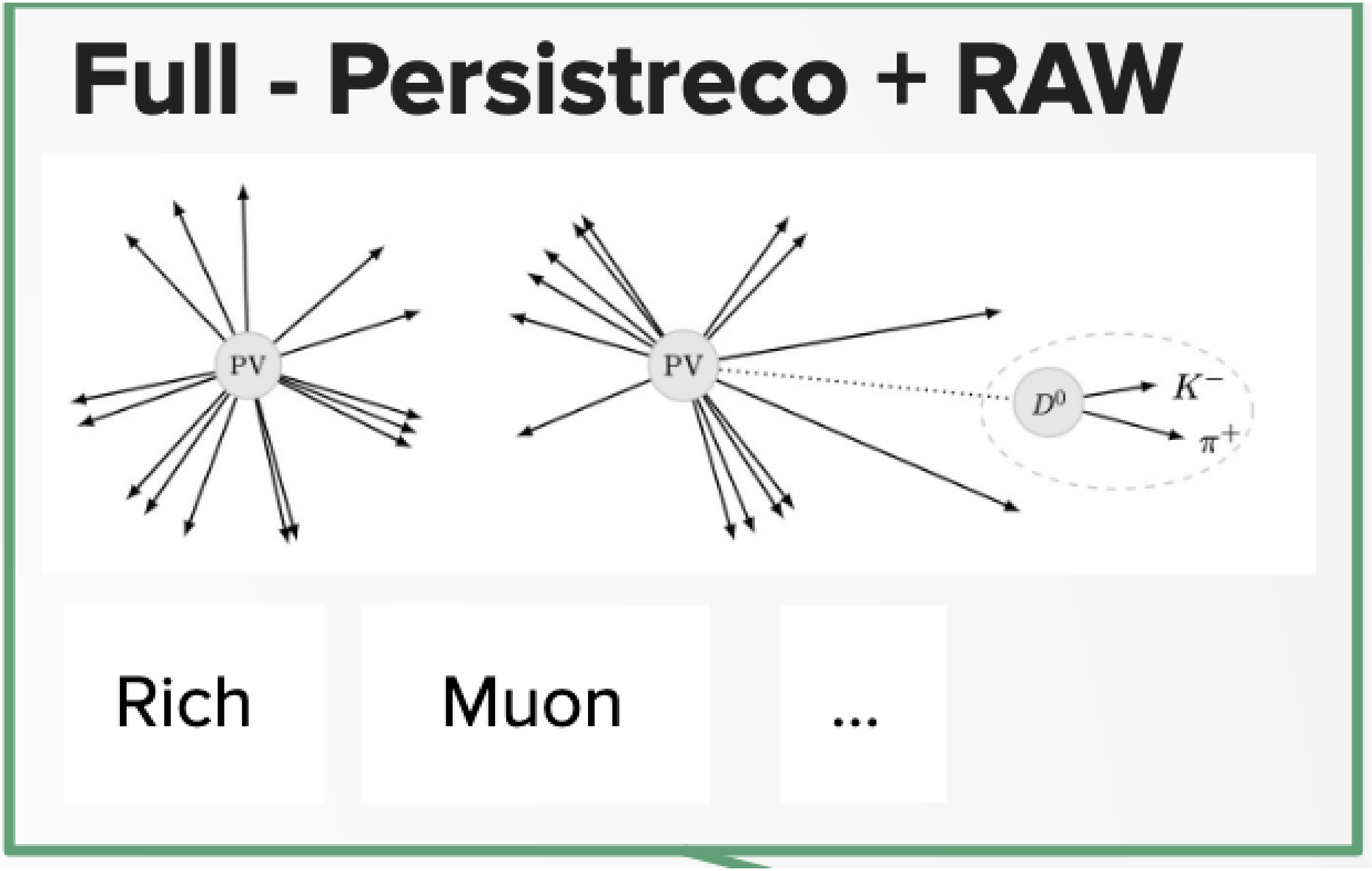


Persistency Model

Keeps all reconstructed tracks from the event and also all raw banks

No data reduction. Requires further selections @Sprucing.

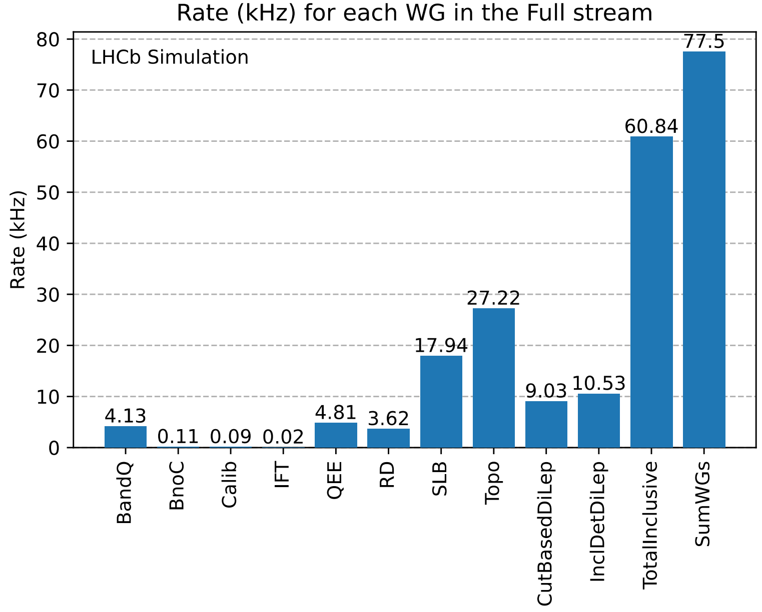
Only used by some calibration lines that require it

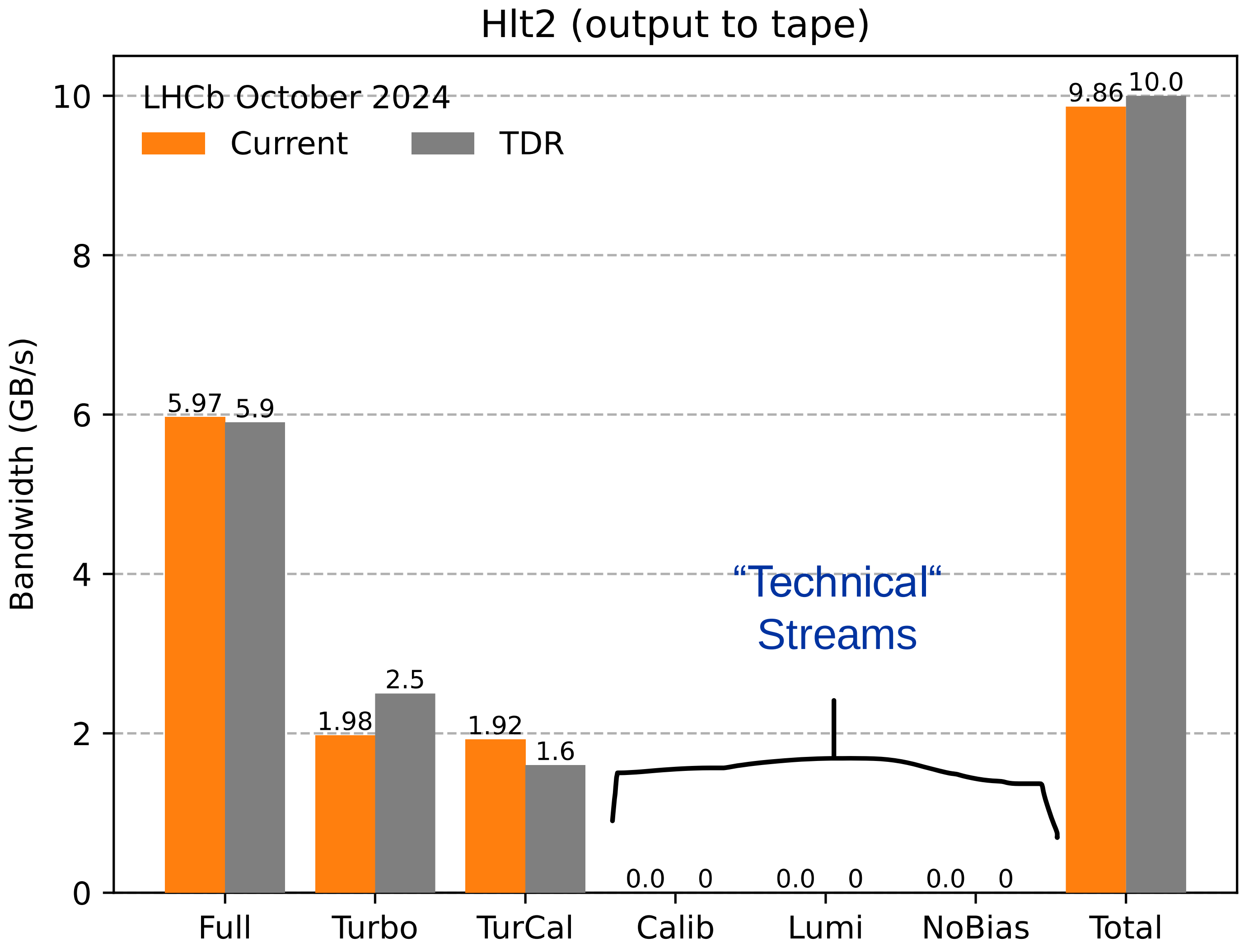


N. Skidmore's talk @CHEP24. By trimming information within an event, we can reduce file size.

Bandwidth

To facilitate the BW constraints, there is reporting of the collaboration's bandwidth on a per-change and per-day level.
Includes '*overviews*' and information as granular as '*table of average event size for all 3000 lines*'.





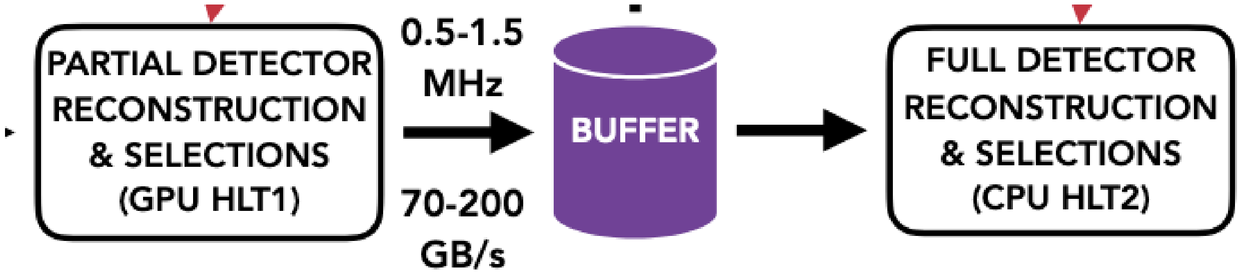
LHCb-FIGURE-2024-034 and R.J. Hunter's talk @CHEP24.

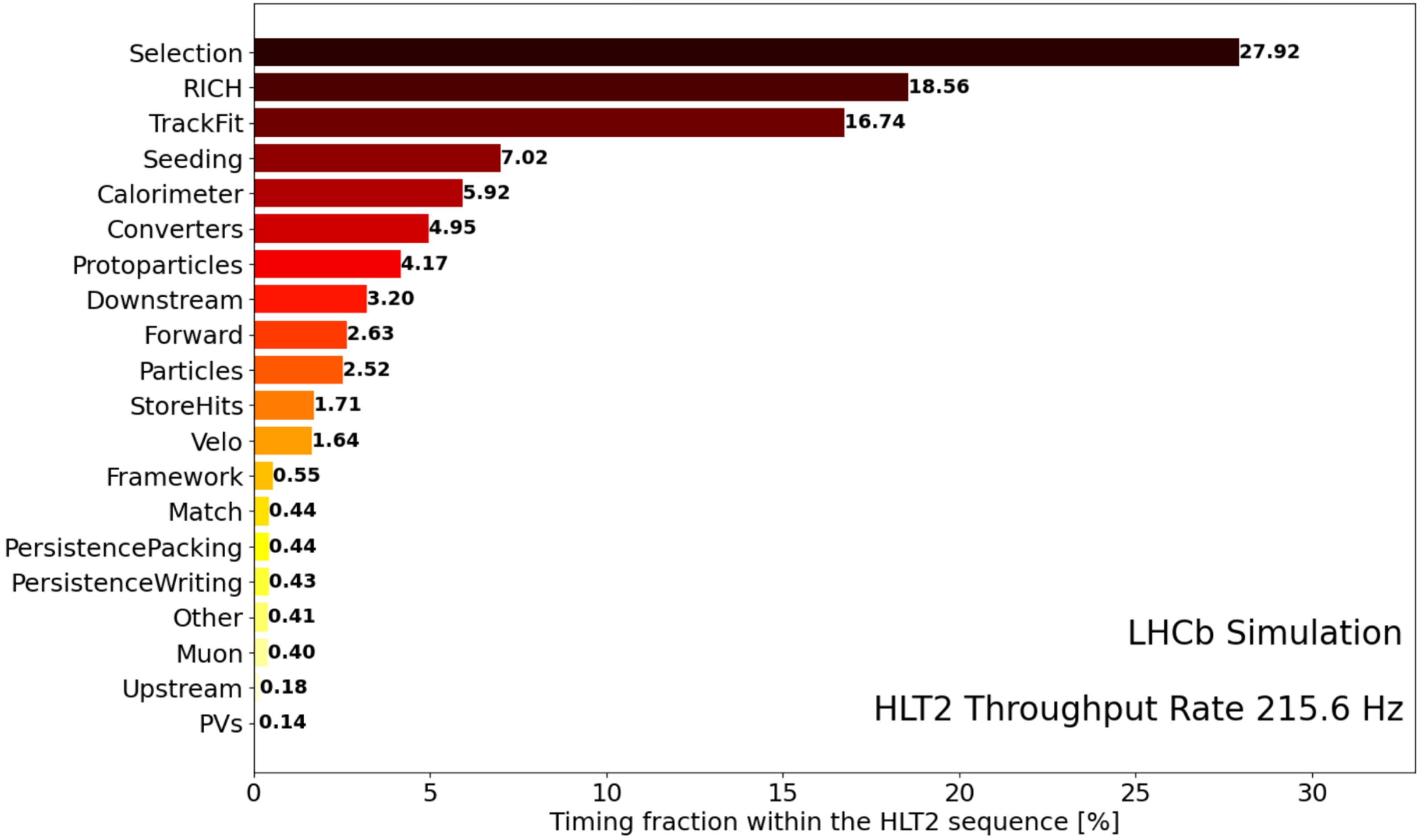
Throughput

HLT2 also has throughput constraints. It needs to be able to process HLT1 output faster than we can fill up the buffer.

Approximately, there's an LHC beam efficiency of <50% over time, so HLT2 must run at least twice as fast as HLT1.

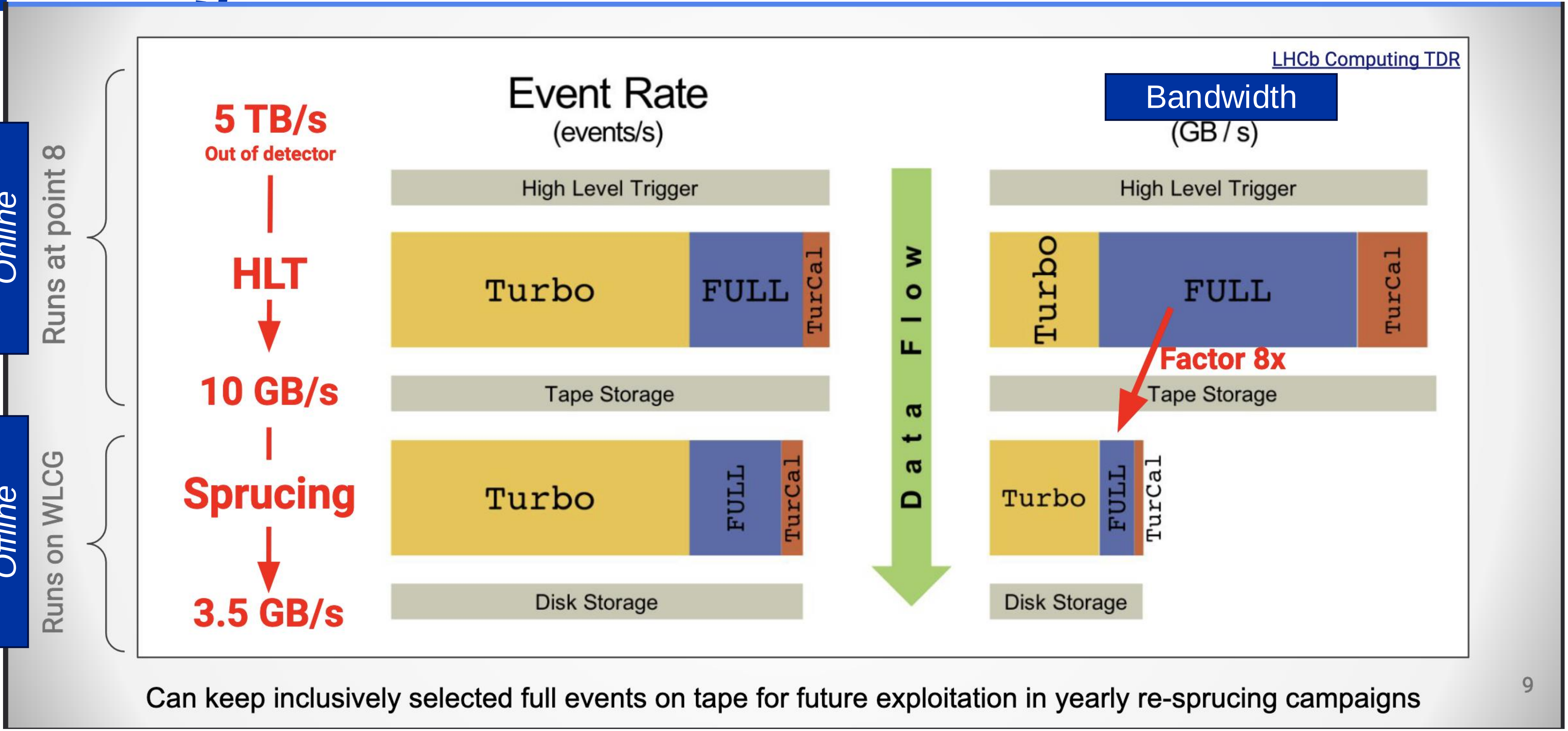
With the current ~4500 CPUs we achieved a HLT2 throughput of 900 kHz, well above the minimum of 500 kHZ.
This is ***so high*** that for 2025 there's work ongoing to try to increase HLT1's output even higher (***1.5/1.6 MHz***) to gain more physics potential.





Sprucing

Sprucing's Role



N. Skidmore's talk @CHEP24. Sprucing further reduces size for inclusive full-events to data size on Grid reasonable

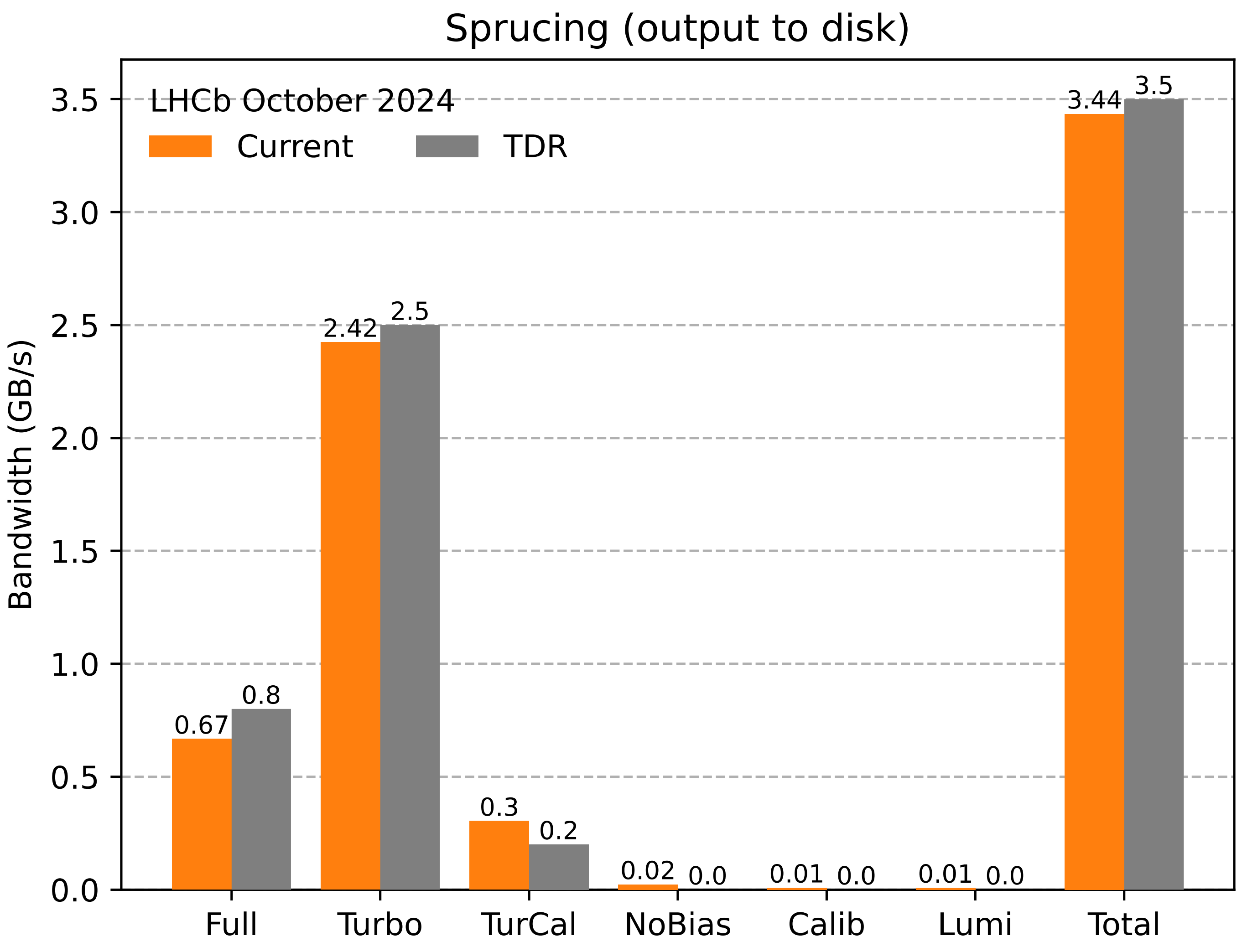
Bandwidth dominated

Like Hlt2, Sprucing is dominated by bandwidth considerations.

10GB/s from HLT2 ->
3.5 GB/s from Sprucing.

For Full, it’s a new selection stage, taking advantage of the persisted reconstructed tracks.
For Turbo and TurCal it’s mostly “PassThrough”. Moves rawbanks around, performs compression and such.

Full-stream can then be re-spruced in later campaigns.



Summaries

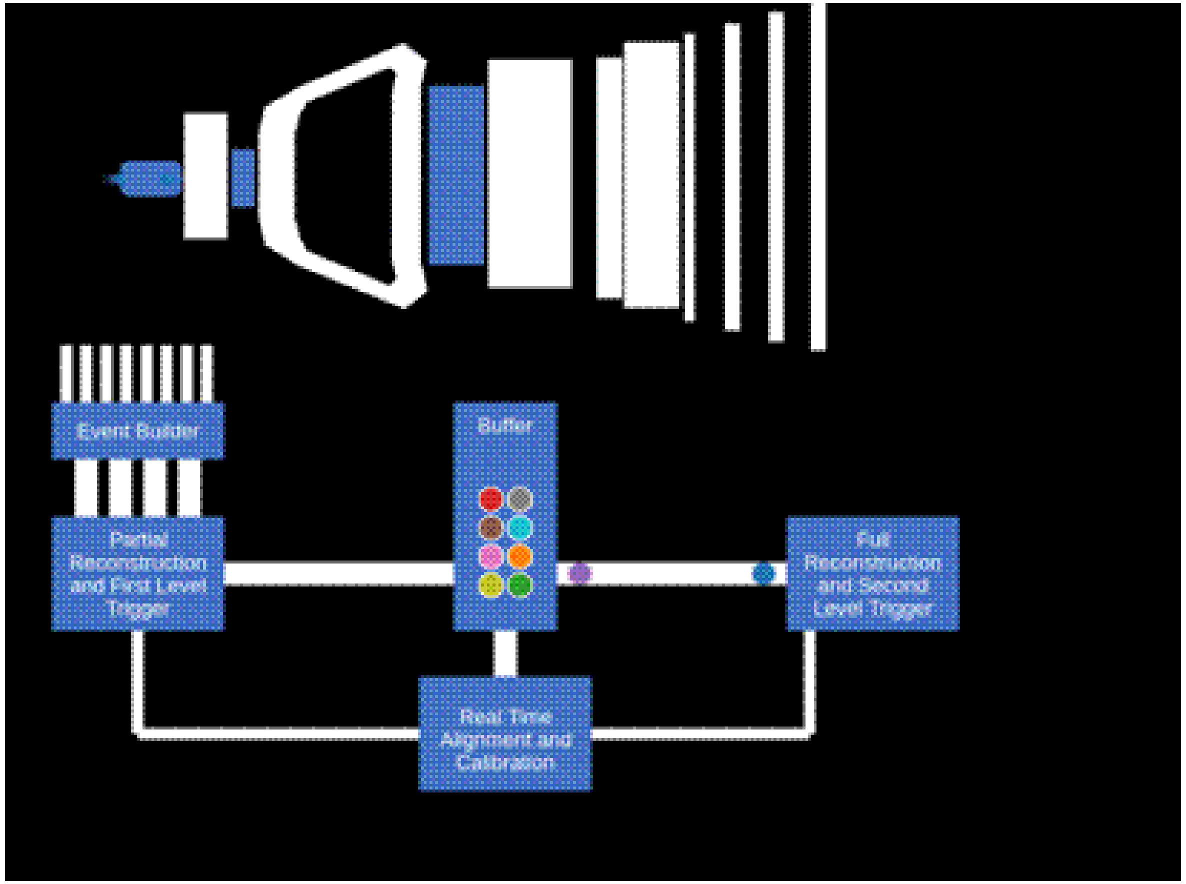
DataFlow

- Hlt1 and Hlt2 allow a 400x reduction in data saved per second (bandwidth), while keeping within operational constraints and high physics efficiency.

- This data is then migrated offline, stored permanently. **(10 GB/s)**

- Offline selections + pruning (Sprucing) is then performed and **15PB/ year** of data is made accessible to analysts via the WLCG

- Re-Sprucing is carried out to further refine offline selections on the permanent data.

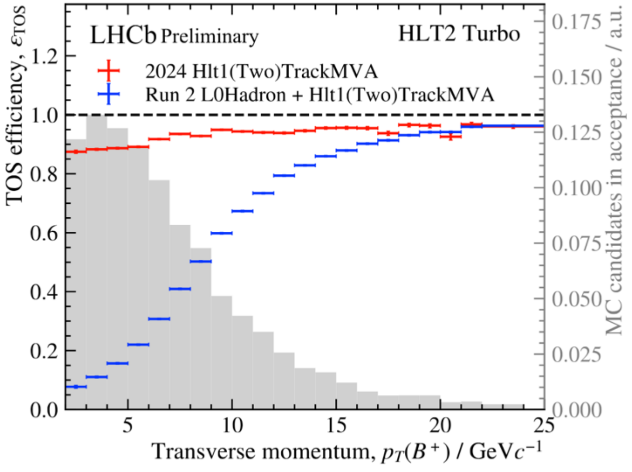


lhcb-outreach.web.cern
LHCb’s unique approach to real-time data processing

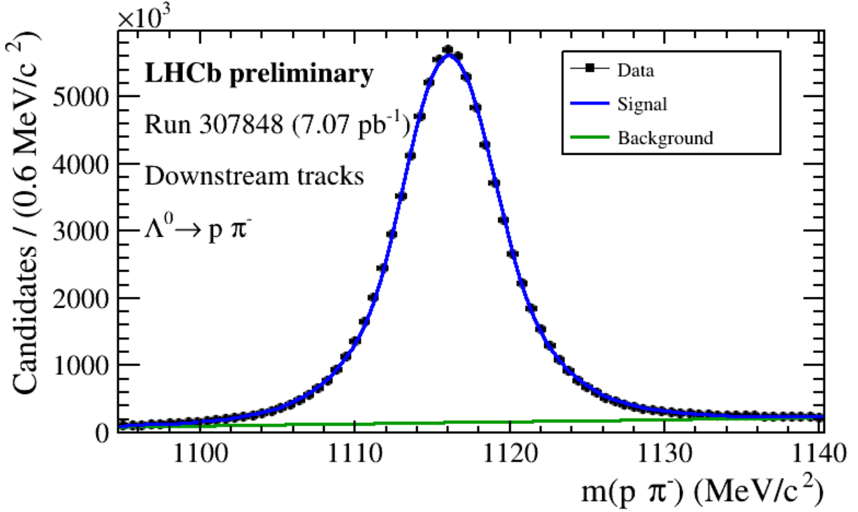
HLT1

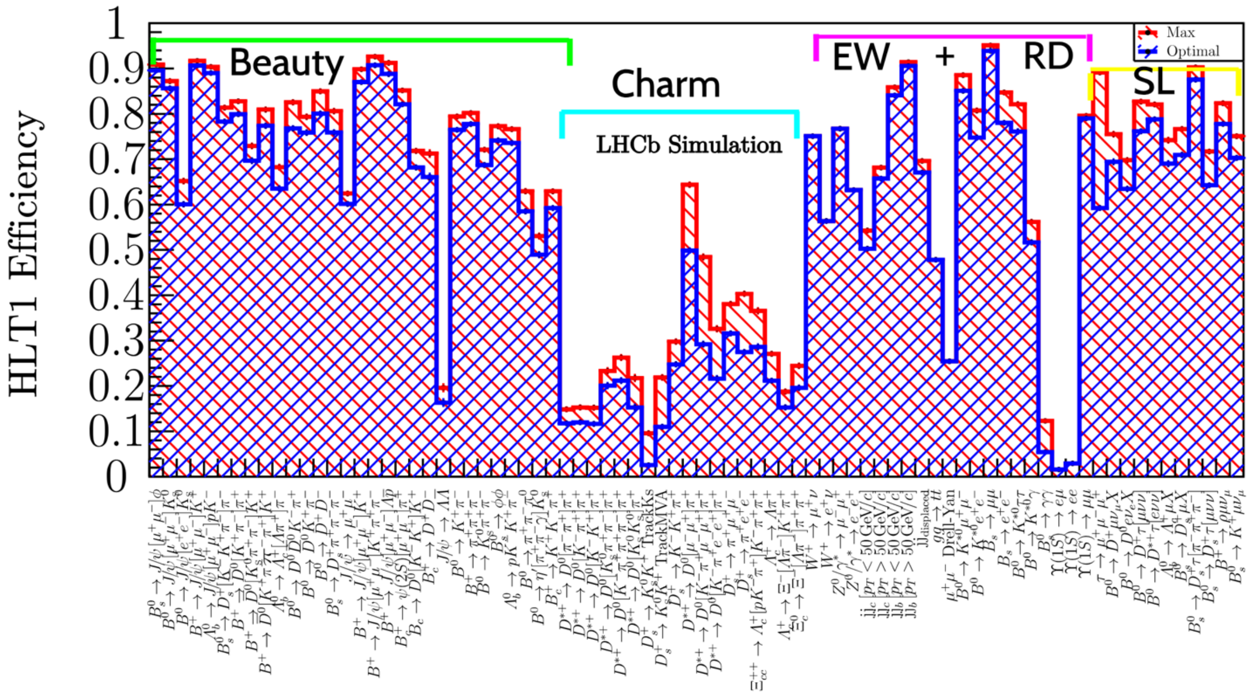
Meets the throughput demands since removing the L0 hardware trigger via:
- GPU parallelisation
- Partial reconstruction

Even with this high demand, we were able to run at a higher Rate and Throughput than originally designed.
- Higher efficiencies than Run1/2
- Automated fair division of rate
- Interesting new physics possibilities.



(a) TOS efficiencies in $B^+ \to \overline{D}^0 (K^+\pi^-)\pi^+$.



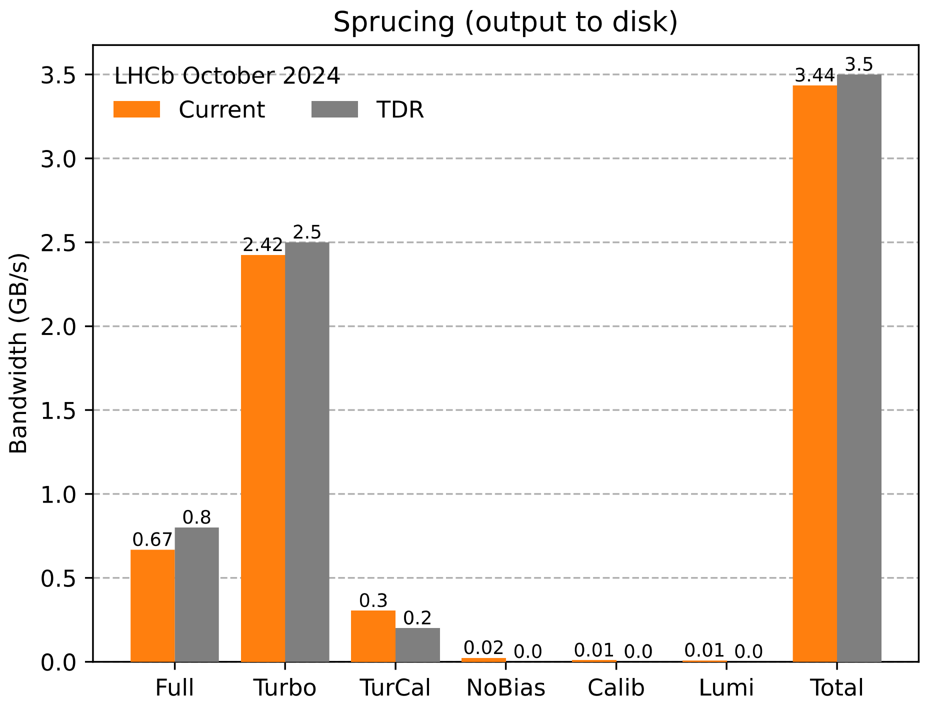


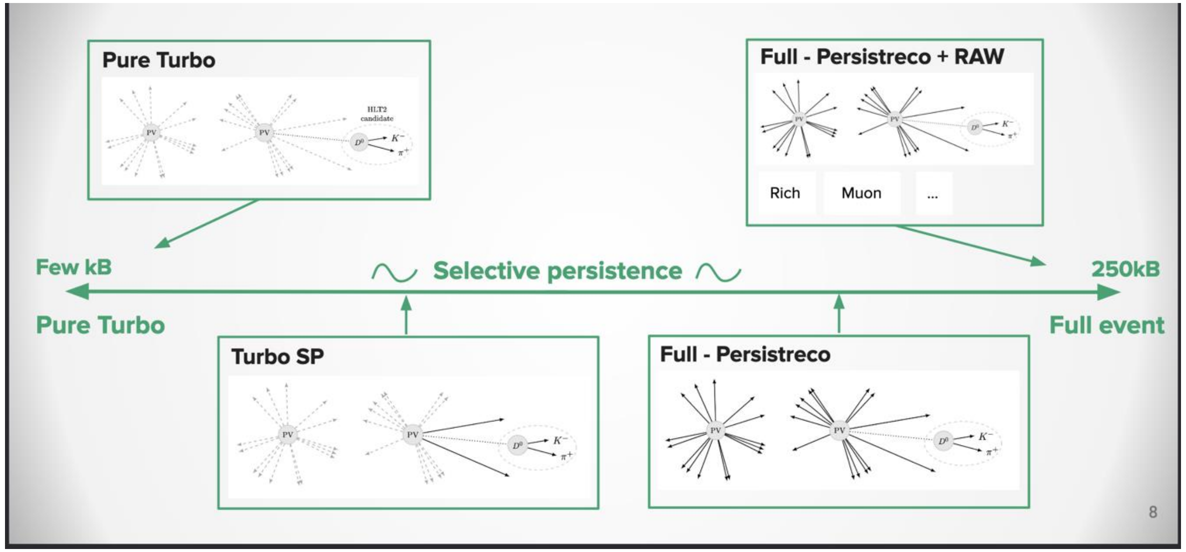
HLT2 and Sprucing

Taking advantage of a persistency model allows massive reductions in average event size for the majority of events.
Thus improving physics reach for over 4000 selection lines,, spread across the physics working groups and persistency streams.

Flexibility to support inclusive and exclusive lines and aiming towards.

Run3 2024 measurements highly prioritized currently, several currently aiming at winter conferences and many more after that.





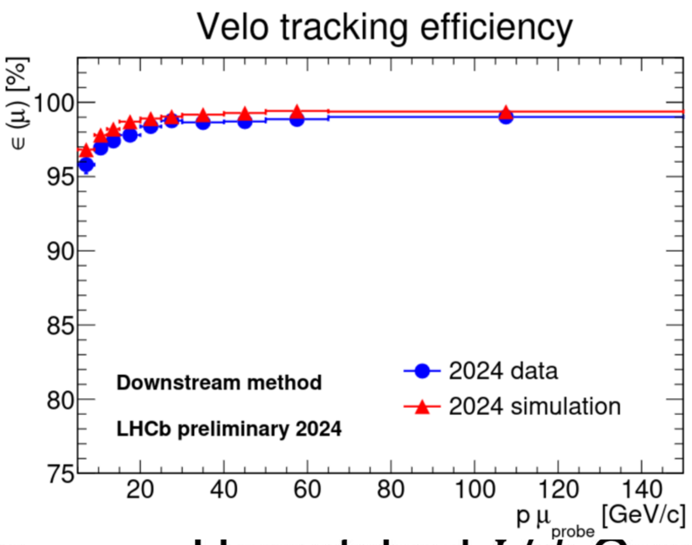
Backups

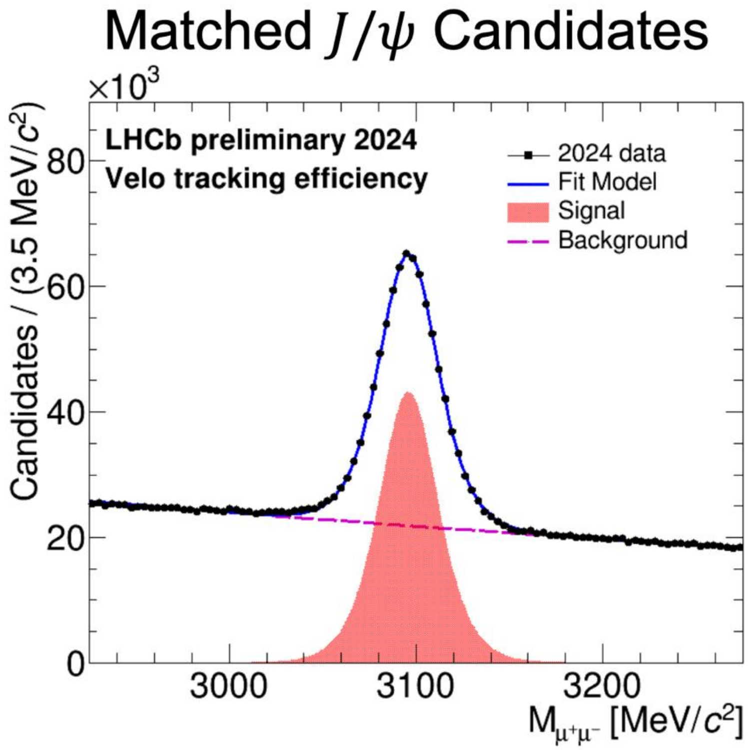
Buffer, Online Alignment and Calibration, skipping :(

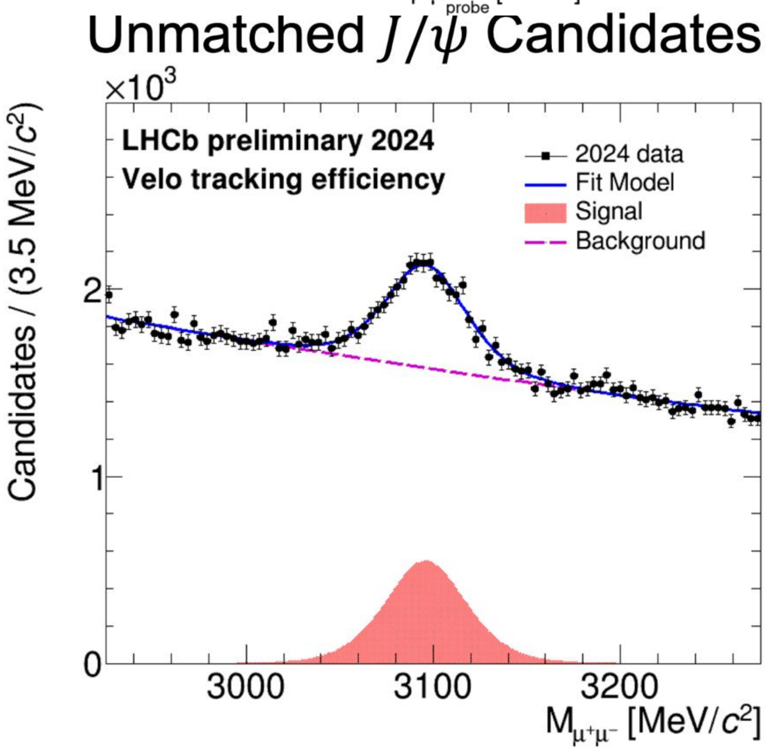
This is truly a very important part of the LHCb, without it the persistency model falls apart and we need to find even higher reductions in rate, harming our efficiencies.

I would recommend 113th LHCb Week AnC summary as an overview of the recent process.

There is also a talk upcoming at the 114th LHCb Week, w/**c next week.**







R. Caspary @CHEP24. Showing data-driven evaluation of tracking efficiencies at the LHCb

Architecture

https://lhcbdoc.web.cern.ch/lhcbdoc/moore/master/design/architecture.html