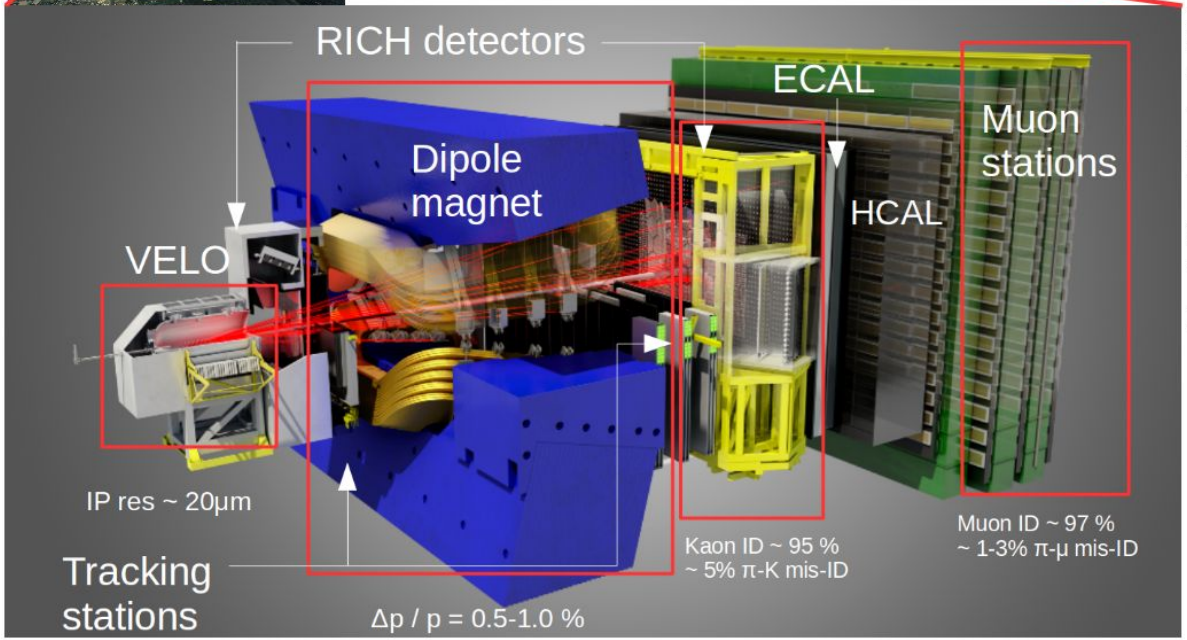
The evolution of the LHCb offline computing towards the Run 3 Upgrade

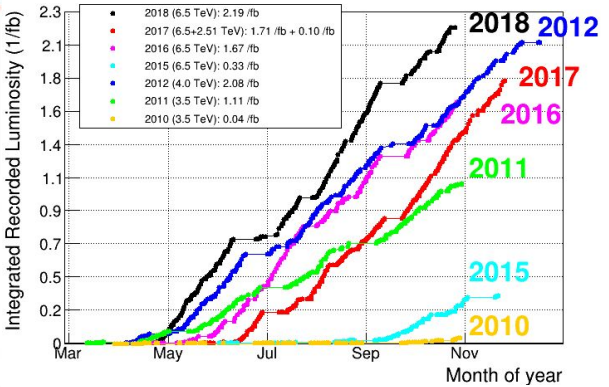
N. Skidmore on behalf of the LHCb collaboration

ICHEP 2020 July 2020

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

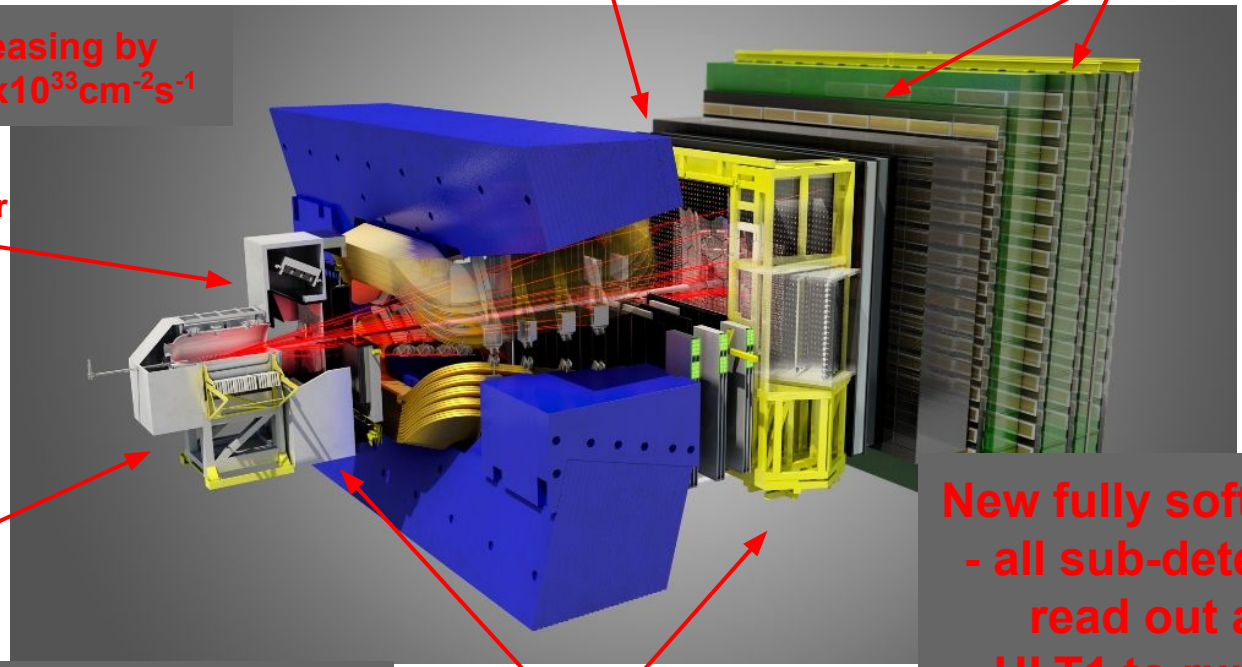
LHCb Upgrade I

New scintillating fibre tracker (SciFi)

New electronics (CALO, MUON)

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

New silicon upstream tracker (UT)



New pixel vertex detector (VELO)

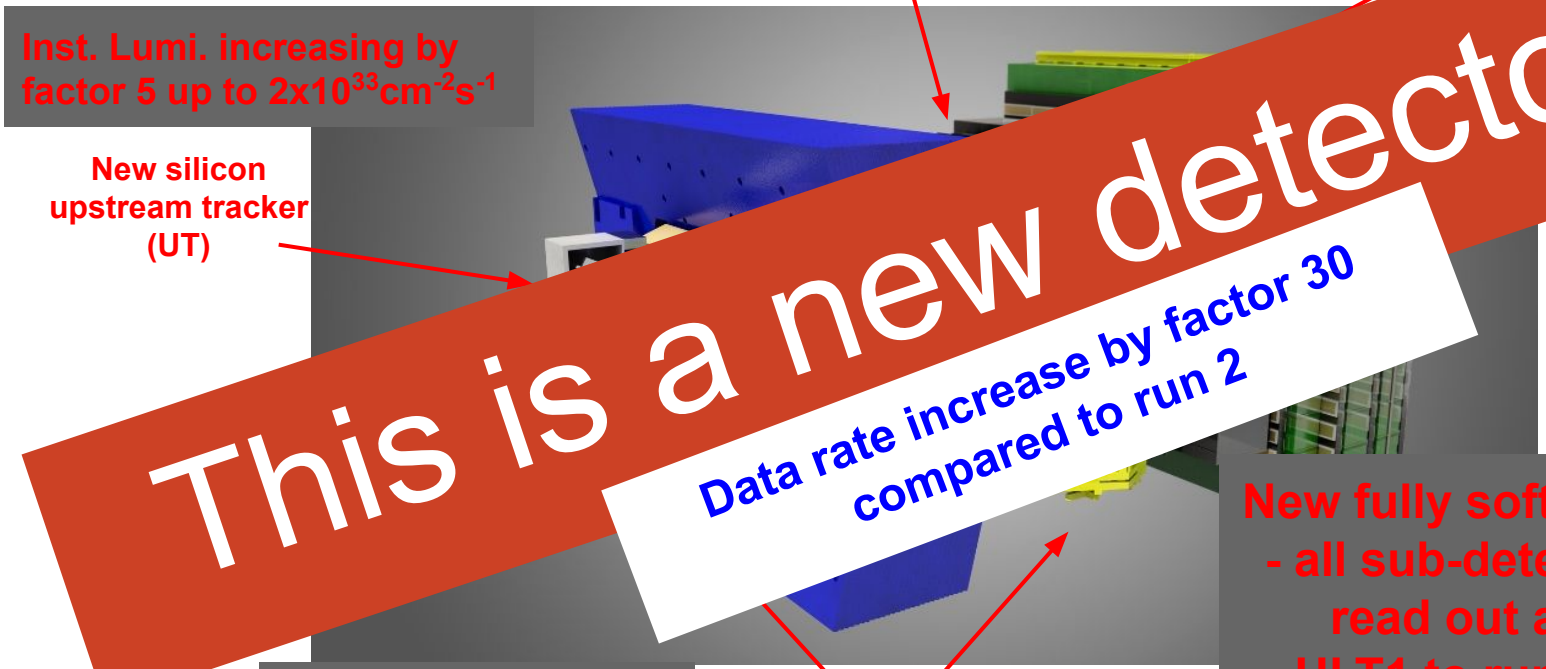
Target Int. Lumi. $50 fb^{-1}$

New optics and photodetectors (RICH)

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

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

LHCb Upgrade I



New scintillating fibre tracker (SciFi)

New electronics (CALO, MUON)

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

New silicon upstream tracker (UT)

This is a new detector

Data rate increase by factor 30 compared to run 2

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

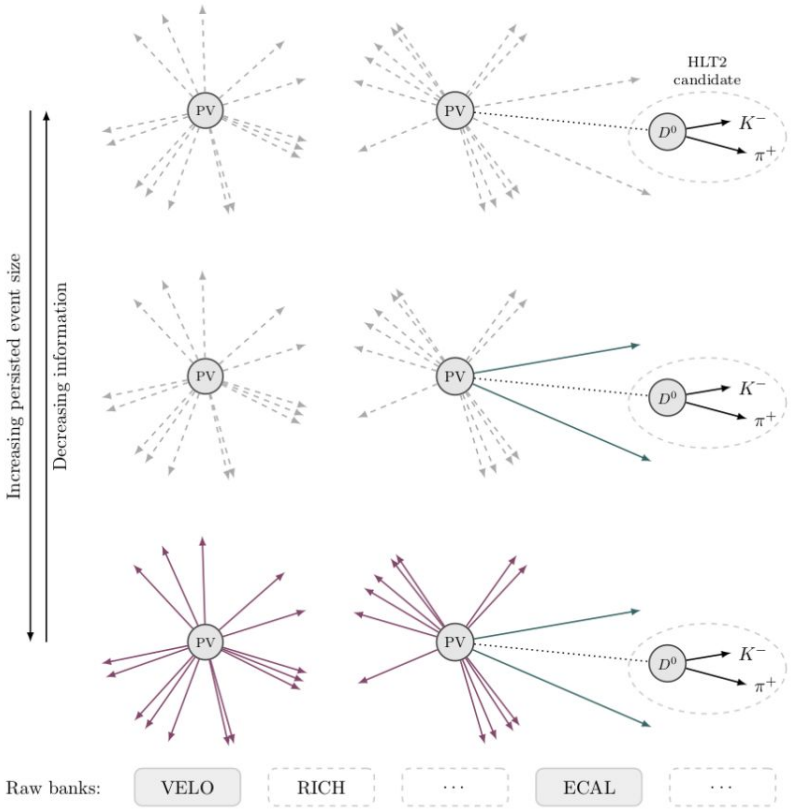
Target Int. Lumi. $50 fb^{-1}$

New optics and photodetectors (RICH)

Data flow evolution - Persistency models (Run 2)

Run 2

- 7 kB
- 16 kB
- 48 kB
- 69 kB



Turbo - save only **reconstructed** objects involved in the trigger decay candidate

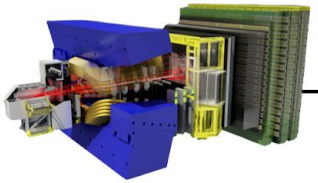
Turbo Selective Persistence (SP) - save **additional** reconstructed objects from event such as other tracks from PV

Full - **whole raw event** is saved. Run 1 model. Re-reconstruction performed **offline**

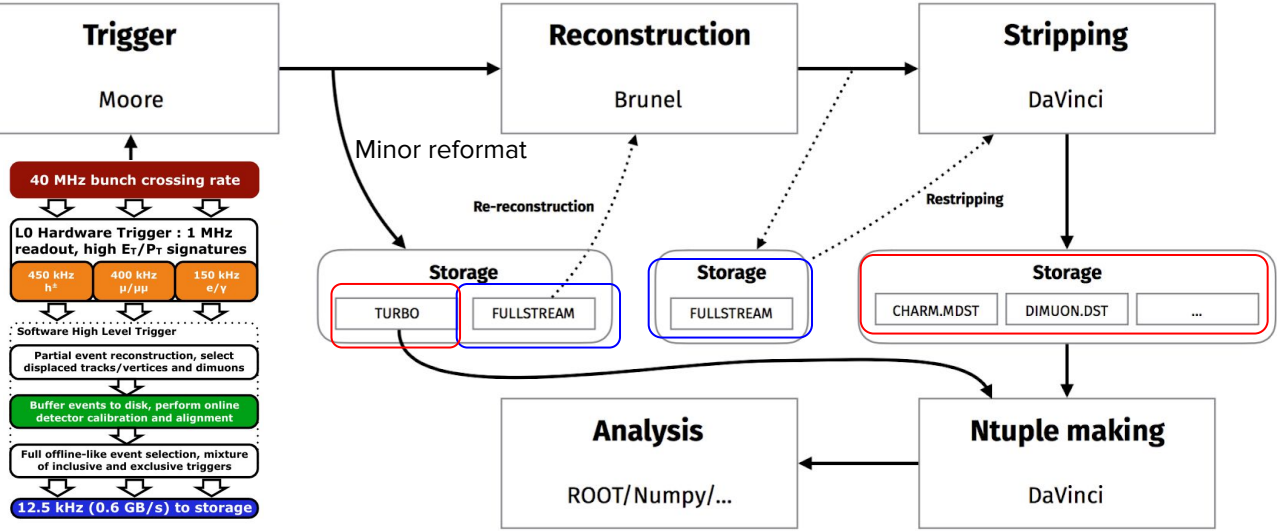
Run 2 only - no offline reconstruction/slimming required

Data flow evolution Run 2

- Most physics uses **full** persistency model
  - Offline re-reconstruction
  - Skimming/slimming required - stripping
- In run 2 Turbo (+SP) model adopted for some physics



Tape storage - not accessible to users
Disk storage - available to users



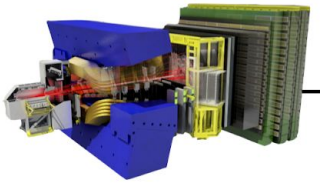
Online

Offline

Data flow evolution Upgrade

All reconstruction, alignment and calibration performed **online**

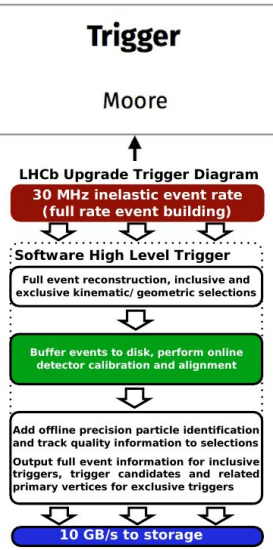
- TURBO (+SP) stream ~immediately available on disk
- FULL stream saves **full reconstructed event*** - RAW data can be removed
- TURCAL stream for calibration saves FULL+RAW info

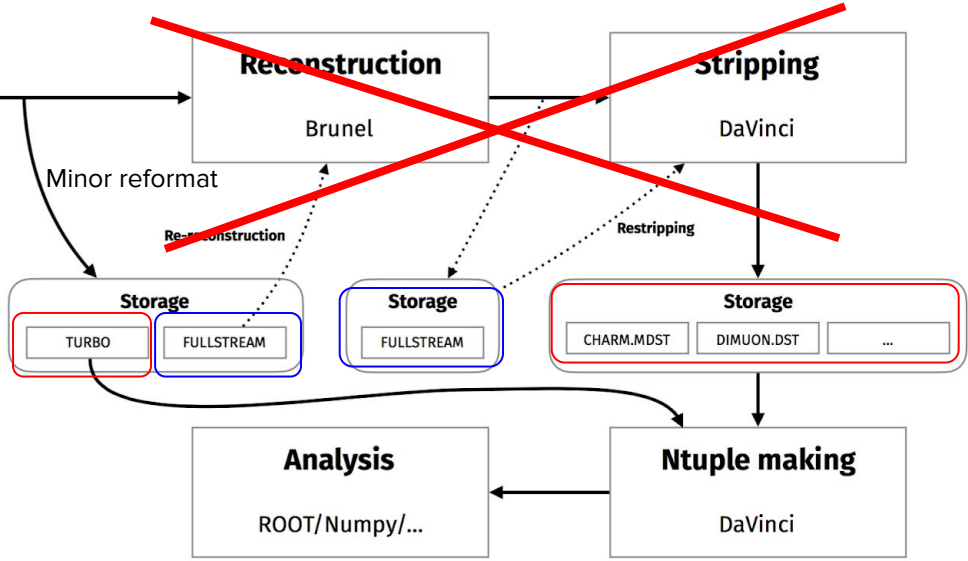


Tape storage - not accessible to users

Disk storage - available to users

For LHCb upgrade trigger see talk #520 and #521





*Slightly simplified view, full stream is extreme instance of Turbo SP

Online

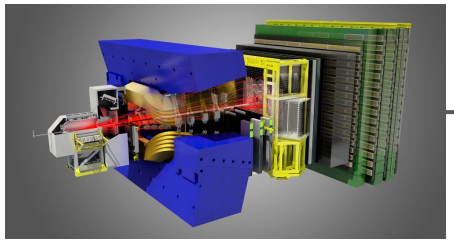
Offline

Data flow evolution Upgrade

Cannot save all HLT output straight to **disk**!

- Utilise cheap **tape storage** for bulk of bandwidth (full stream)
- Rely on central offline slimming/skimming
- Safer option for some physics/allows data mining

Default model Turbo



To tape

HLT2

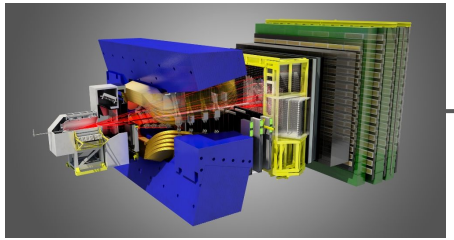
Minor reformat

Data analysis

To disk

Default model - >70% of physics

Sprucing model FULL and TURCAL



To tape

HLT2

Sprucing

Data analysis

To disk

Use cases - topological, inclusive triggers, datamining

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

Data volume strategies

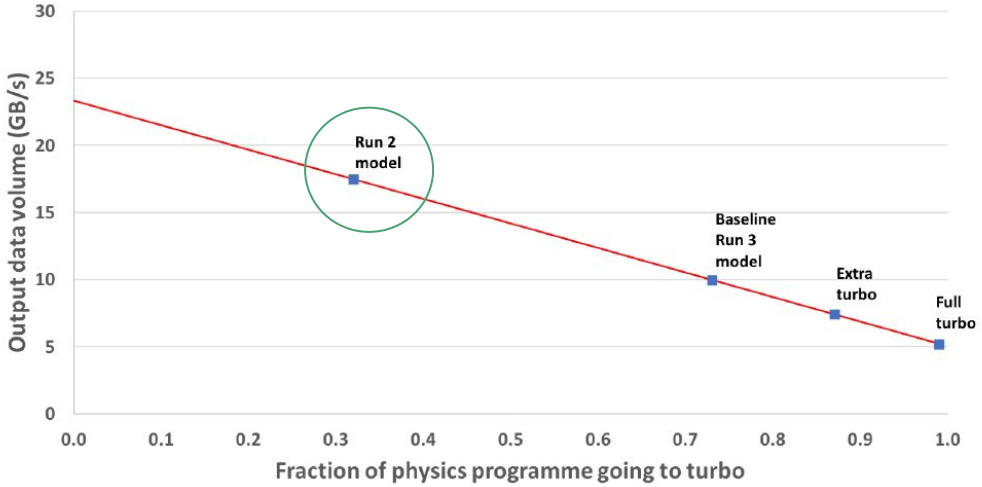
In Run 2
- 29% of physics rate to turbo stream

In upgrade
- Luminosity increase factor 5
- HLT efficiency increase - no L0 hardware trigger
- Raw event size increase due to pile up

With no changes to Run 2 model HLT2 output is 17.4GB/s*!

Run 2

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%



*live LHC seconds

Data volume strategies

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

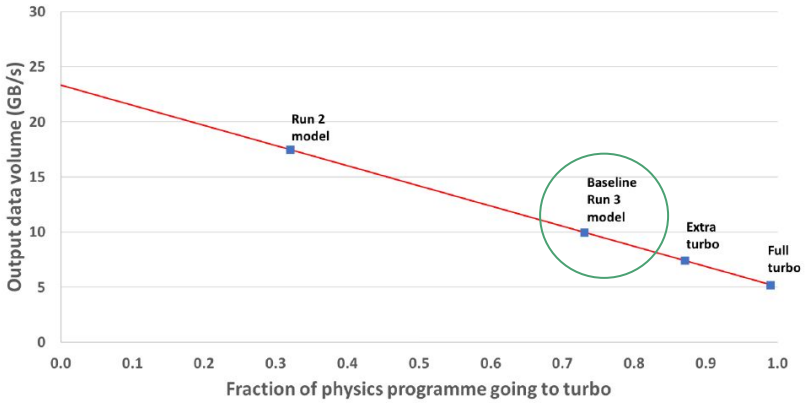
Baseline

Rely on large fraction of physics channels moving to turbo model

- Huge **migration** of physics selections to HLT framework - effort ongoing
- Now have to be **optimised for speed** as selections run online

Those that cannot move to turbo will follow offline sprucing model

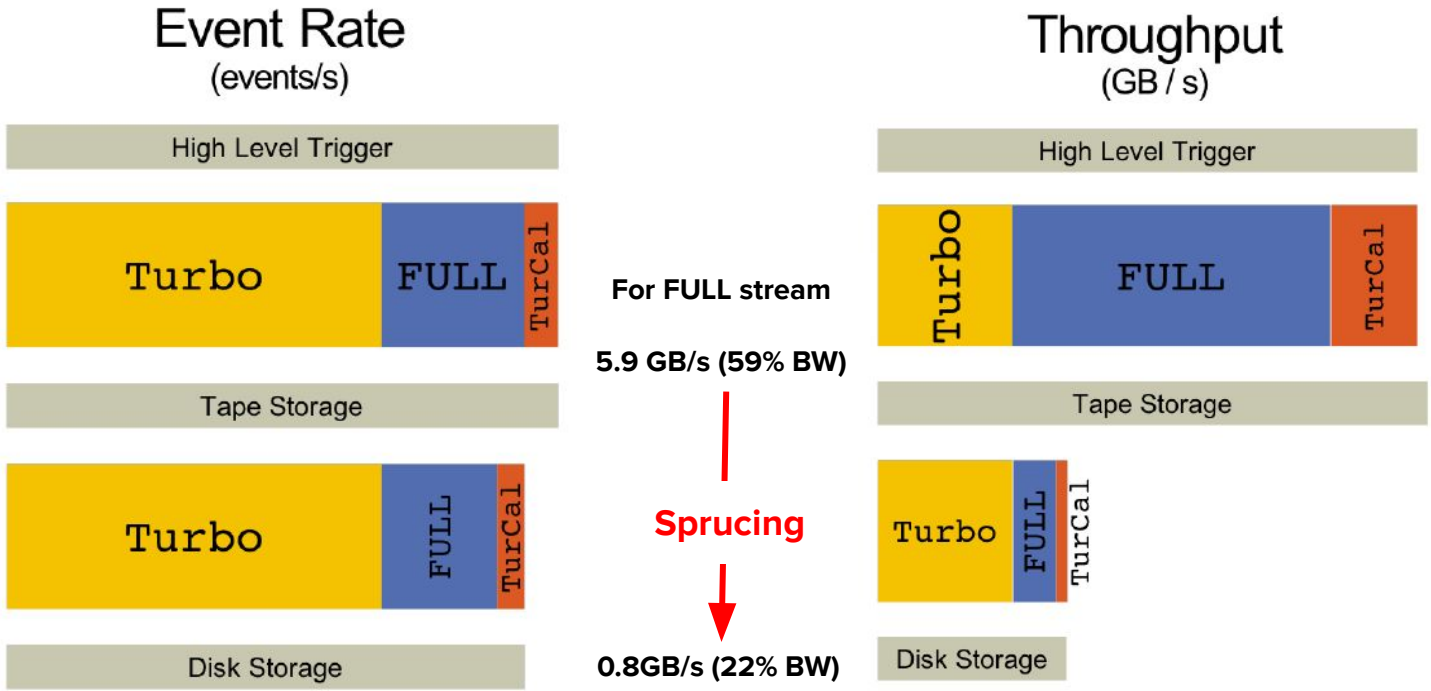
- Note sprucing selections use **same codebase** as HLT2



Data volume strategies

For selections at LHCb see talk #519

Baseline scenario



Total data to tape at 10GB/s

For FULL stream

5.9 GB/s (59% BW)

Sprucing

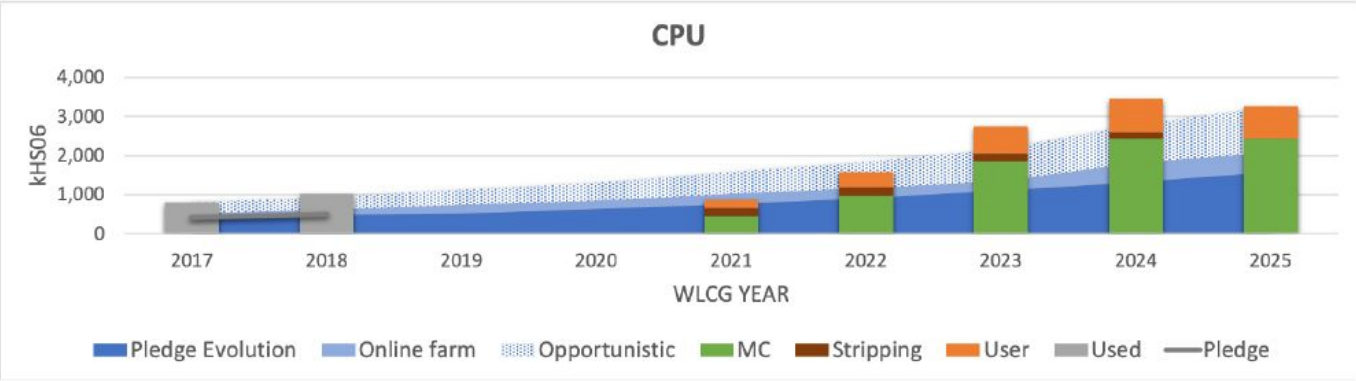
0.8GB/s (22% BW)

Total data to disk at 3.5GB/s

Simulation

For Fast simulations at LHCb see talk #516

- Real data will dominate disk storage but simulation will dominate CPU needs - **90% of total offline CPU** resources
- Decrease in time required to simulate events crucial to fully exploit the larger datasets
  - Measurements hinting at SM tensions have systematic uncertainties dominated by **limited MC statistics**
- Fast simulation options are crucial to exploit the run 3 dataset



*from TDR - LHC schedule since changed

Simulation

For Fast simulations at LHCb see talk #516

Successful adoption of fast simulation in Run 1 and 2

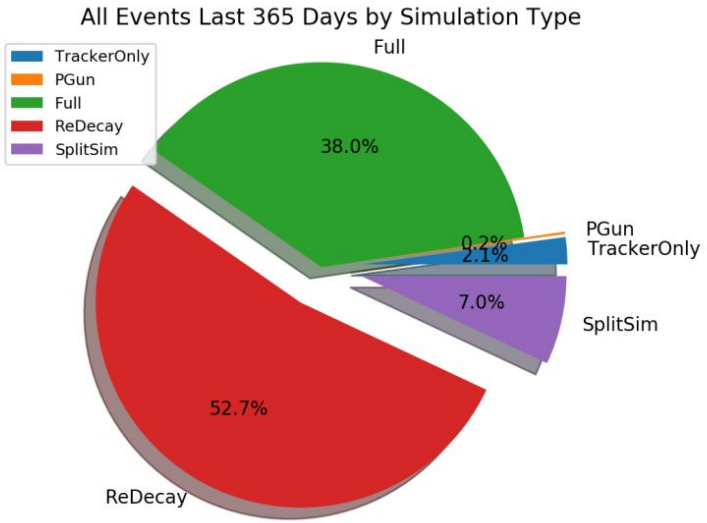
Full - full detector simulation

PGun - single signal particle spawned with kinematics configured to follow distribution (no full Pythia event) Factor 50 speed increase

ReDecay - re-use the underlying event but generate and simulate new signal decays every time Eur. Phys. J. C 78 (2018) 1009 Factor 10-20 speed increase

TrackerOnly simulation **-** Factor 10 speed increase

SplitSim - only simulate full event if required condition is passed eg. if a photon converts to $e^+ e^-$ Speed up depends on condition



Analyst data tuples

In Run 1 + 2 analysts create **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

In run 3 submit jobs centrally using **DIRAC transformation System** (Analysis Productions)

- MC data is already produced this way
- Does not require analyst to babysit jobs
- Jobs can be **tested automatically** with GitLab CI
- Job details/configuration/logs **automatically preserved** in LHCb bookkeeping/EOS
- Automated error interpretation/advice
- Results displayed on webpage

Analysis production job for RDs

`MC_13266069_2012_MagUp`

Status	Commit	Requested	Processed	Runtime	Kept
Success	e9ba8301	-1 events	565 events	0:02:24	True

Input

Name	Size	Total
/lhcb/MC/2012/RXCHAD.STRIP.DST/00108364/0000/00108364_00000007_1.rxchad.strip.dst	131.8 MB	0.0 TB

Output

Name	Size	Total (estimated)
00012345_00006789_1.bsntuple_mc.root	3.2 MB	0.3 GB

Browse output file Show

Reproduce locally Show

Job log Show

DIRAC log Show

Offline analysis tools

Tuples produced using **TupleTools** - creation and saving of 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
- **Redesign** of tools such that this redundancy is minimised

LHCb collaboration uses a wide range of tools C++ /Python/ ROOT/ uproot/ numpy/ pandas/...)

Custom user environments (for use on distributed computing) limited by CVMFS distributions

- Experimenting with providing analysts the ability to install **Conda environments** on CVMFS
- **Singularity containers** (CERNVM) are used for running legacy applications on grid - looking to expand

NumPy

pandas

Heterogenous resources

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

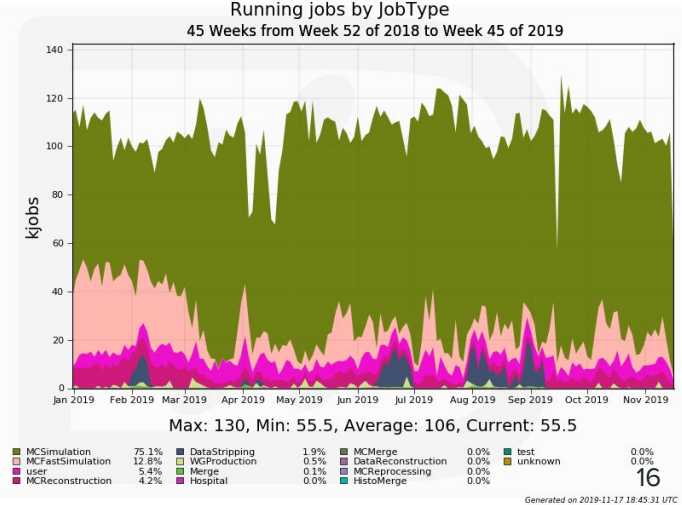
Most sites have no GPUs yet - push towards High Performance Computing (HPC) centers providing **large GPU resources**

Potential to **utilise HLT1 GPU farm** like current HLT CPU farm during detector downtime

Need development such that **significant LHCb payloads can run on GPUs**

- User analysis utilising eg. TensorFlow for ML and fitting but small share of LHCb's CPU
- Full detector simulation main payload but Geant4 has no GPU compatibility yet (work ongoing outside LHCb)

GPU batch cluster at CERN to develop/run GPU workflows



Data Processing and Analysis (DPA) project

Run 3 offline data volume necessitates a **more coordinated approach** to offline data processing addressing the aforementioned points

- Software projects carry same status as sub-detector projects

WP1: Sprucing

Centralised offline data selection/streaming for data that cannot go (initially) to TURBO stream.

Coordinator: Nicole Skidmore (Bonn)

WP2: Analysis productions

An upgrade to the WG productions of Run 2, done centrally.

Coordinator: Chris Burr (CERN)

WP3: Offline analysis tools

LHCb software and tools for offline analysis and analysis preservation.
Forum for discussion about, and usage of, modern analysis packages outside the LHCb hat.

Coordinator: Patrick Koppenburg (Nikhef)

WP4: Innovative analysis techniques

Think Tank for new ideas – exploitation of new analysis facilities (clusters, cloud), possibly with heterogeneous computing resources, etc.
Co-ordinate exploratory work, aiming for successful proofs of concept to become mainstream in LHCb.

Coordinator: Donatella Lucchesi (Padova)

WP5: Legacy software and data

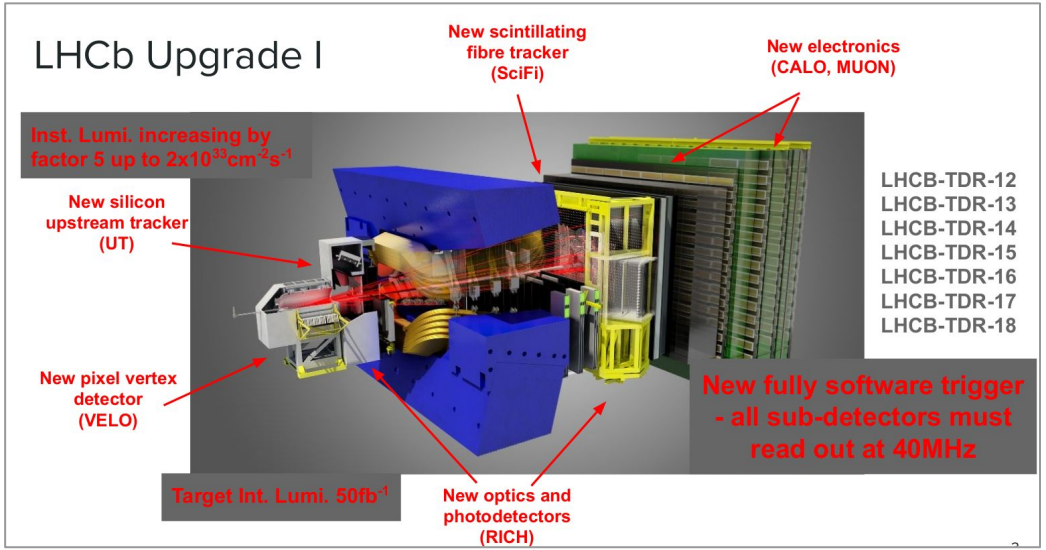
Maintenance of, and support for, legacy runs 1 & 2 software and data samples.

Coordinator: Alison Tully (EPFL)

Project leader: Eduardo Rodrigues (Liverpool)

Conclusions

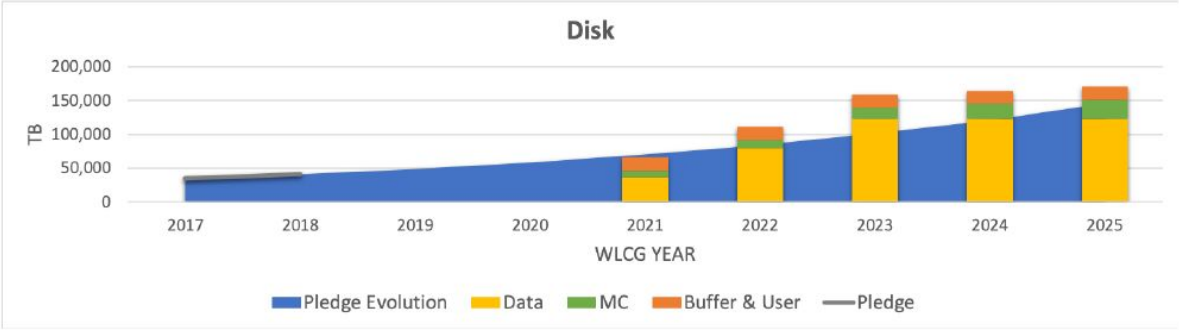
- LHCb will have to process data offline an order of magnitude larger than in Run 2
- LHCb is progressing well to meet the offline demands that run 3 will bring
  - Revised treatment of event persistency models and central offline data processing
  - Developments in fast simulation methods
  - Initiatives towards use of accelerators
  - Creation of offline Data Processing and Analysis project to coordinate efforts in collaboration with computing team



Backup

Resources at LHCb

In Run 3 LHCb will produce ~ 15PB of user accessible data per year



Real data dominates disk storage

But simulation dominates CPU - mitigation strategies using **fast simulation**

*from TDR - LHC schedule since changed