Doing More with less: The evolution of the LHCb experiment trigger and selected physics results

Conor Fitzpatrick

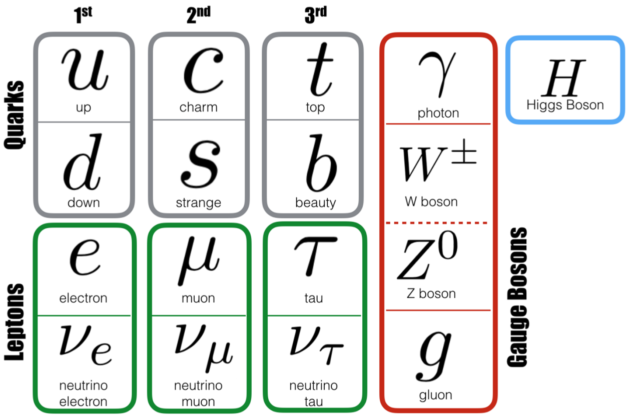
Science Coffee seminar
Lund University

BEAUTY2CHARM
European Research Council
Established by the European Commission

UK Research and Innovation

The Standard Model is amazing

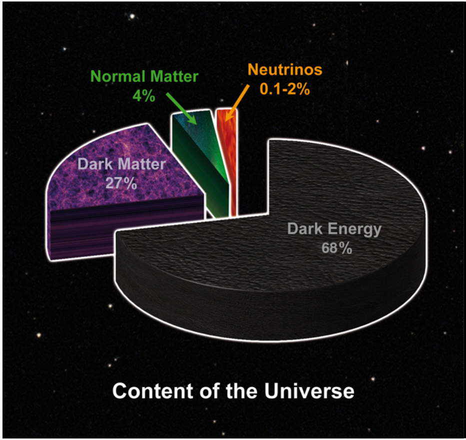
- The standard model (SM) of particle physics is a triumph. Accurate, predictive and testable:

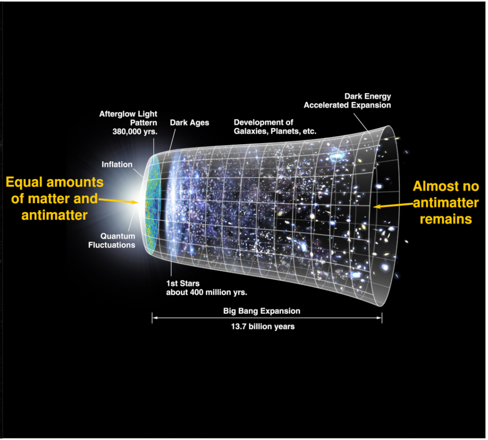


- Beautifully validated by the discovery of the Higgs boson in 2012, predicted in 1964

The Standard Model is incomplete

- The SM does not describe several features required to explain the present universe.
  - For example:

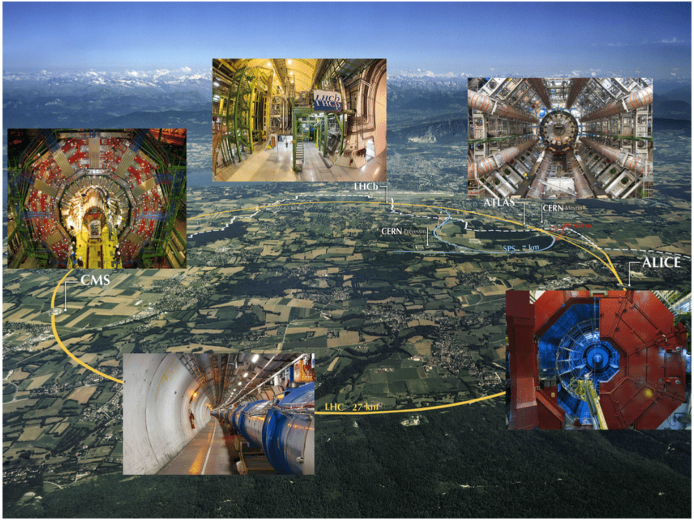




- What is Dark Matter, energy? Where did all the antimatter go?

There must be physics beyond the Standard Model

- Explanations for these features needs new physics (NP).



- The LHC is a laboratory: Several unique ways to search for new physics

C. Fitzpatrick

March 30, 2021

The LHC

- The LHC collides protons at both high energy, and high intensity

C. Fitzpatrick

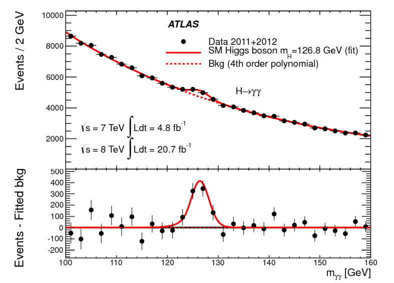
March 30, 2021

The LHC

- The LHC collides protons at both high energy, and high intensity

Direct detection

- If the collision energy is high enough, create and observe new particles



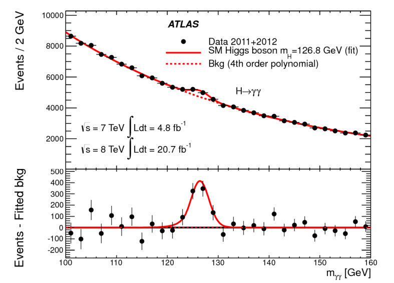
- LHC nearing its design energy
- No direct NP observations so far

The LHC

- The LHC collides protons at both high energy, and high intensity

Direct detection

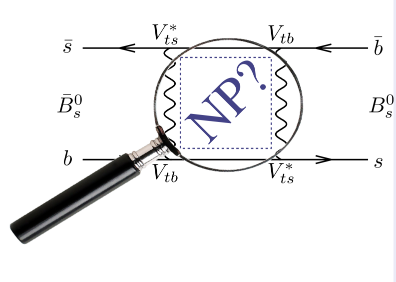
- If the collision energy is high enough, create and observe new particles



- LHC nearing its design energy
- No direct NP observations so far

Indirect detection

- If the intensity is high enough, look for subtle deviations in known processes



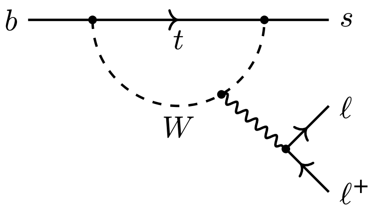
- LHC produces the world's largest precision datasets

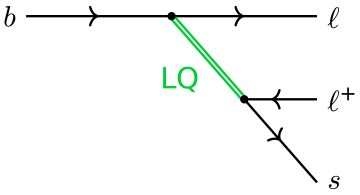
C. Fitzpatrick

March 30, 2021

Where to look for deviations?

- The Standard Model consists of:
  - 3 gauge couplings
  - 2 Higgs parameters
  - 6 quark masses
  - 3 quark mixing angles and 1 phase
  - 3 charged lepton masses (+ 3 neutrino masses)
  - 3 lepton mixing angles + 1 phase
- Flavor parameters are a large part of the SM
- Flavour changing processes are particularly interesting as new interactions can affect what we measure
- SM predictions, particularly for heavy (B,B_s^0) mesons are precise. Deviations are a smoking gun



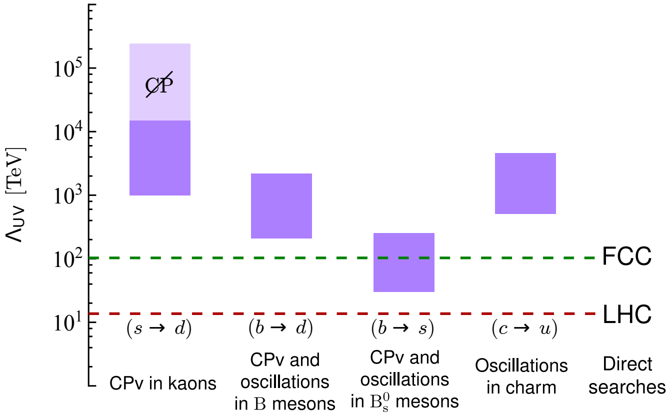


C. Fitzpatrick

March 30, 2021

What can they tell us?

- Quark flavour changing processes are sensitive to energies higher than we can probe directly:

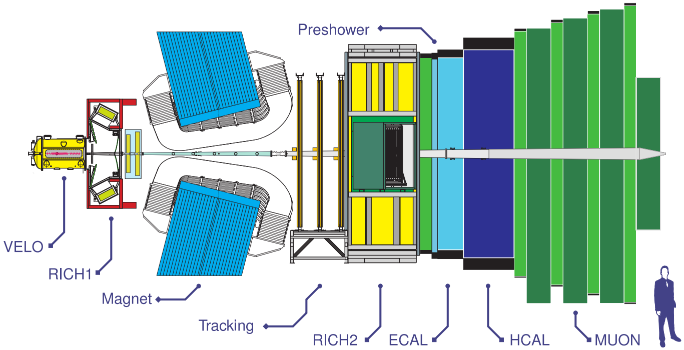


- Sets the energy for future collider designs[1]

[1] M. Neubert, EPS-HEP, 2011

LHCb: The precision flavour experiment

- LHCb was built to exploit the high rates of beauty and charm at the LHC to make measurements of this kind:



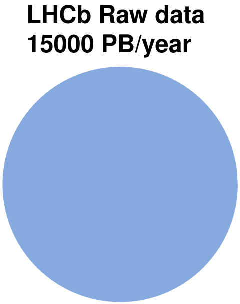
- Precise particle identification (RICH + MUON)
- Excellent decay time resolution: $\sim$45 fs (VELO)
- High purity + efficiency with flexible trigger

What is a trigger?

- High intensity means high data rates:
  - The LHC collides bunches of protons at 30 MHz
  - At the experiments, each collision is about 100kB (LHCb) - 1MB (ATLAS/CMS)
  - LHC operates for about 5×10^6 seconds/year.

What is a trigger?

- High intensity means high data rates:
  - The LHC collides bunches of protons at 30 MHz
  - At the experiments, each collision is about 100kB (LHCb) - 1MB (ATLAS/CMS)
  - LHC operates for about 5×10^6 seconds/year.
- The LHC experiments generate 15-150 exabytes of raw data each

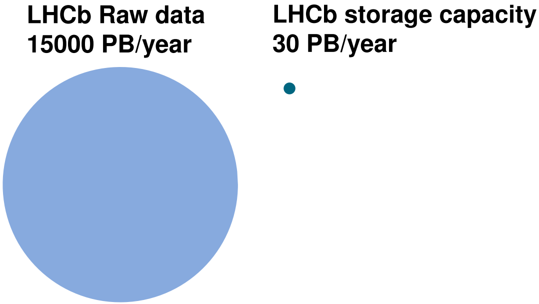


C. Fitzpatrick

March 30, 2021

What is a trigger?

- High intensity means high data rates:
  - The LHC collides bunches of protons at 30 MHz
  - At the experiments, each collision is about 100kB (LHCb) - 1MB (ATLAS/CMS)
  - LHC operates for about 5×10^6 seconds/year.
- The LHC experiments generate 15-150 exabytes of raw data each

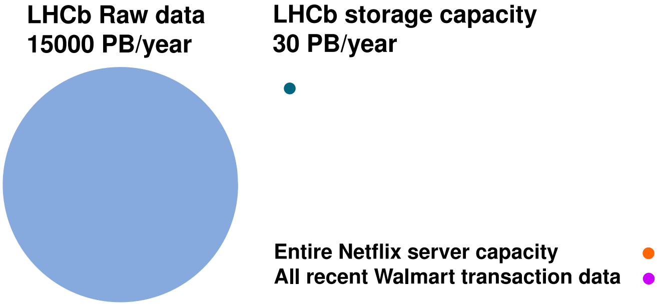


- Storage is limited to tens of PB / year

C. Fitzpatrick
March 30, 2021

What is a trigger?

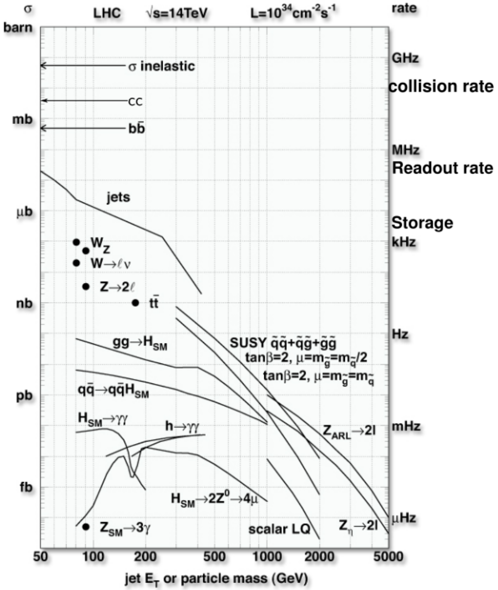
- High intensity means high data rates:
  - The LHC collides bunches of protons at 30 MHz
  - At the experiments, each collision is about 100kB (LHCb) - 1MB (ATLAS/CMS)
  - LHC operates for about 5×10^6 seconds/year.
- The LHC experiments generate 15-150 exabytes of raw data each



- Storage is limited to tens of PB / year
- LHC experiments have **similar storage requirements to fortune 500 companies**

The LHCb trigger

- A trigger is needed to reduce storage and readout costs
- A *good* trigger does so by keeping more signal than background

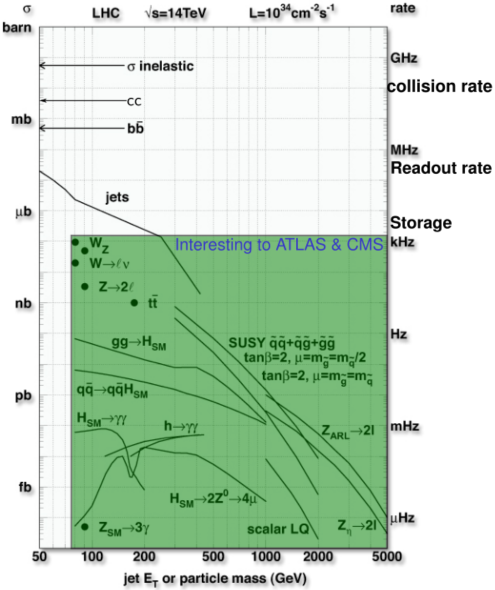


C. Fitzpatrick

March 30, 2021

The LHCb trigger

- A trigger is needed to reduce storage and readout costs
- A *good* trigger does so by keeping more signal than background
- ATLAS/CMS are interested in signatures in the kHz region
  - Readout at 100 kHz is efficient with reasonably straightforward E_T requirements

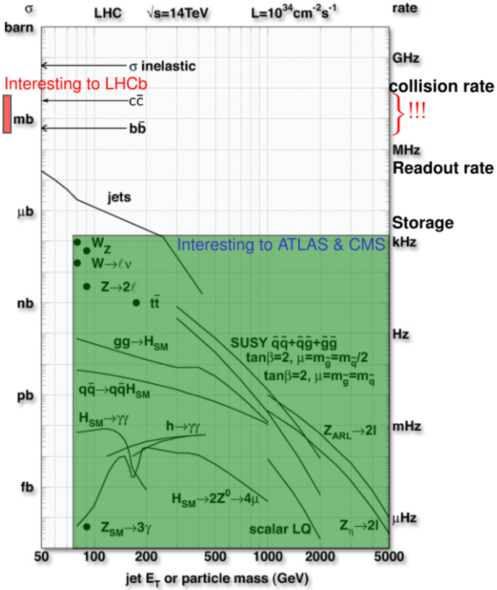


C. Fitzpatrick

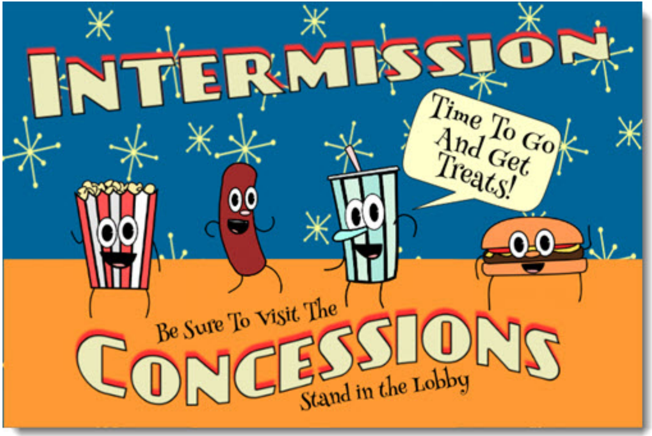
March 30, 2021

The LHCb trigger

- A trigger is needed to reduce storage and readout costs
- A *good* trigger does so by keeping more signal than background
- ATLAS/CMS are interested in signatures in the kHz region
  - Readout at 100 kHz is efficient with reasonably straightforward E_T requirements
- LHCb faces a unique challenge addressed in Runs 1&2 with:
  - Lower luminosity running
  - 1 MHz readout rate



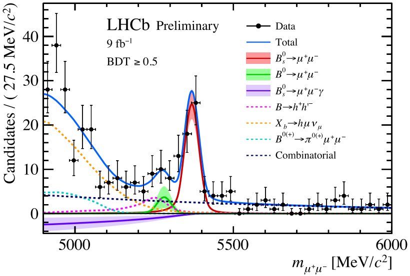
Introduction ends

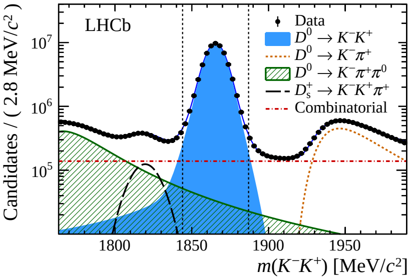


- I hope this has served as an accessible introduction to why flavour physics is of interest, and the trigger challenge LHCb faces
- Going forward I will describe the trigger in more detail, provide performance characteristics and discuss some recent physics results.

The LHCb Run 2 trigger in two plots

- The LHCb trigger in Run 2 had to cover extremes of data taking:

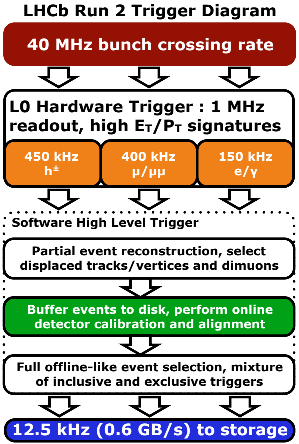




- High efficiency to collect rare decays like $B_s^0 \to \mu\mu$ [2]
- High purity for enormous charm signals like $D^0 \to KK$ [3]
- Requires a high degree of flexibility at high data rates

[2] NEW: LHCb-PAPER-2021-007 in preparation
[3] Phys. Rev. Lett. 122, 211803 (2019)

The Run 2 LHCb Trigger

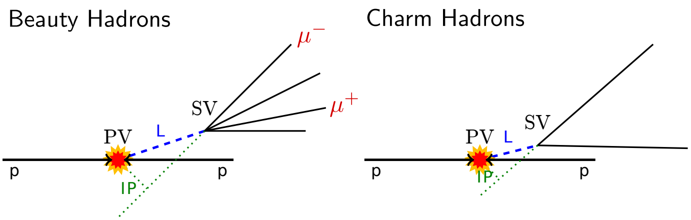


- The LHCb Run 2 trigger (2015-2019)
- Three trigger levels, with a hardware L0 stage:
  - Level-0 trigger buys time to readout the detector with Calo, Muon p_{T} thresholds: $40 \rightarrow 1$ MHz
  - Events built at 1 MHz, sent to HLT farm ($\sim$27000 physical cores)
  - HLT1 has $40\times$ more time, fast tracking followed by inclusive selections 1 MHz $\rightarrow$ 100 kHz
  - HLT2 has $400\times$ more time than L0: Full event reconstruction, inclusive + exclusive selections using whole detector
- Flexibility comes from software-centric HLT design[4]

[4] JINST 14 (2019) P04013

Signatures

- Typical beauty and charm decay topologies:



- $B^{\pm}$ mass $\sim$5.28 GeV, daughter p_T $\mathcal{O}$(1 GeV)
- $\tau \sim$1.6 ps, Flight distance $\sim$1 cm
- Important signature: Detached muons from $B \to J/\psi X$, $J/\psi \to \mu\mu$

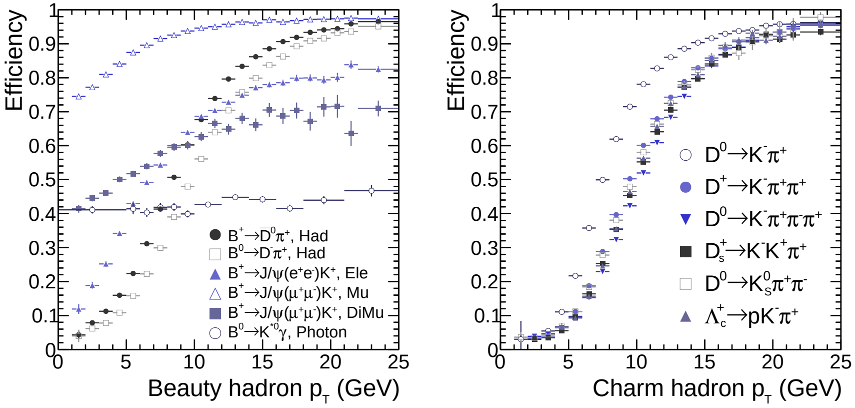
- D^0 mass $\sim$1.86 GeV, appreciable daughter p_T
- $\tau \sim$0.4 ps, Flight distance $\sim$4 mm
- Also produced as 'secondary' charm from B decays.

Underlying Trigger strategy:

- Readout based on simple L0 critera, Fast reconstruction at HLT1: Primary Vertices, High p_T tracks, optional Muon ID, Exclusive and inclusive selections at HLT2 with full reconstruction

Level 0

- L0 uses simple, localised signatures: Transverse energy/momentum thresholds in the muon and calorimeter systems

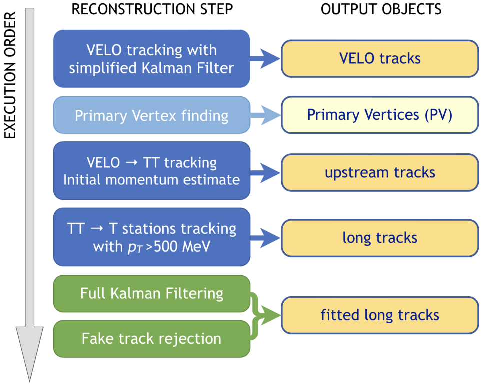


- Genetic algorithm-based bandwidth division balances signal efficiency across entire physics programme within 1 MHz output.
- Typically 40-60% efficient for hadronic beauty 10-30% charm, 90% efficient for muon signatures

C. Fitzpatrick

HLT1

- After readout, events were sent to a 27,000 core CPU farm where the full event is available for processing
- HLT1 performs a fast reconstruction to obtain primary vertices and all tracks above $p_{\mathrm{T}} > 500$ MeV
- These are available for 1- and 2- track MVA selections
- Full muon ID applied to fitted long tracks $p_{\mathrm{T}} > 500$MeV, and an additional fast reconstruction recovers muons with $p_{\mathrm{T}} > 80$MeV.

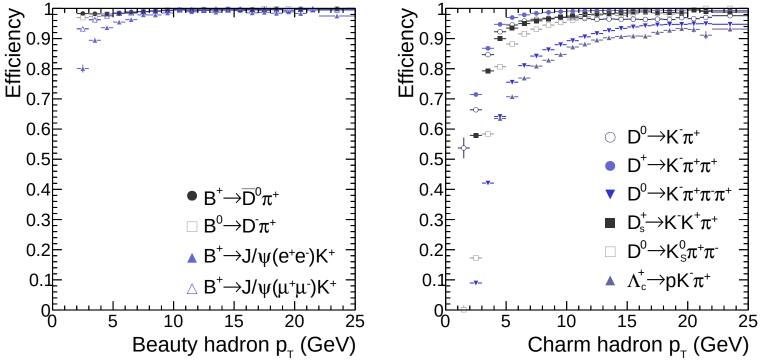


MANCHESTER 1824
The University of Manchester
C. Fitzpatrick

March 30, 2021

HLT1 selections

- Majority of physics at HLT1 selected using 1- and 2- track multivariate algorithms. Rate reduction from 1 MHz $\rightarrow$ 100 kHz:



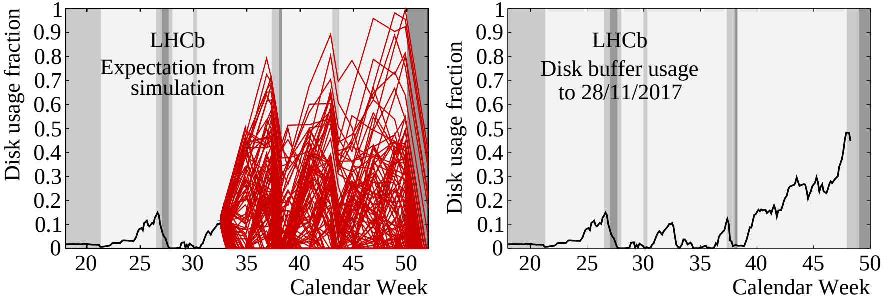
- Extremely efficient ($> 95\%$) for beauty, $70 + \%$ efficient for charm

MANCHESTER 1824
The University of Manchester
C. Fitzpatrick

March 30, 2021

Disk Buffer

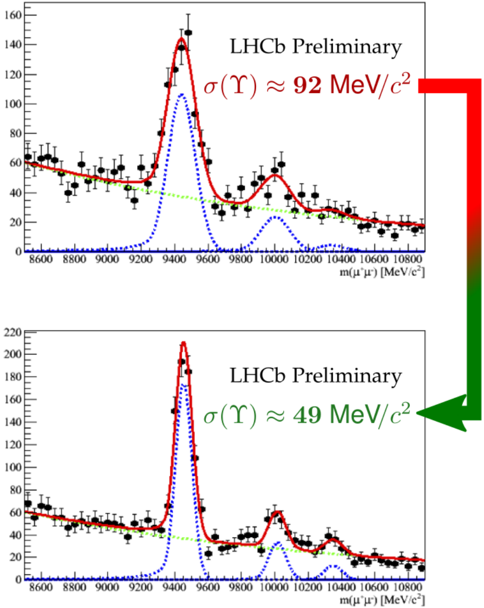
- HLT Farm: off-the shelf servers, with considerable (11PB) disk capacity
- HLT1 gets written to these disks, allowing HLT2 to run asynchronously. Provides nearly a 2 week contingency.



- Effectively doubles trigger CPU capacity, Farm is used twice for HLT, excess used for simulation
- Buffer simulated during data taking, allowing HLT1 output to be tuned
- Asynchronous HLT has another big advantage though. . .

Real-time Alignment + Calibration

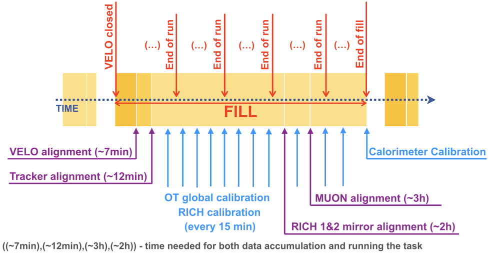
- With Run 2 signal rates, efficient & pure output required full reconstruction at HLT2
  - Online selections → offline selections
  - Reduces systematic uncertainties and workload for analysts
- Alignment and calibration of full detector in the trigger needed
- While HLT1 is written to disk, alignment & calibration tasks run

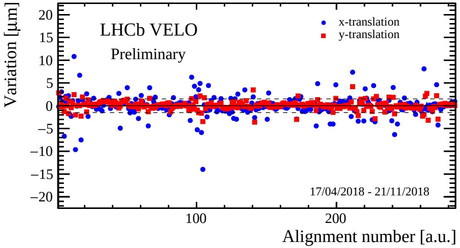


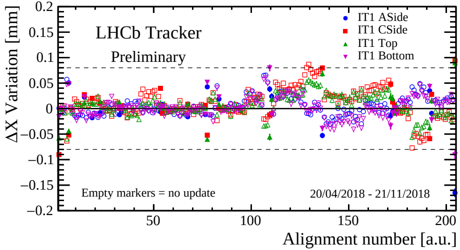
C. Fitzpatrick

March 30, 2021

A fully aligned detector



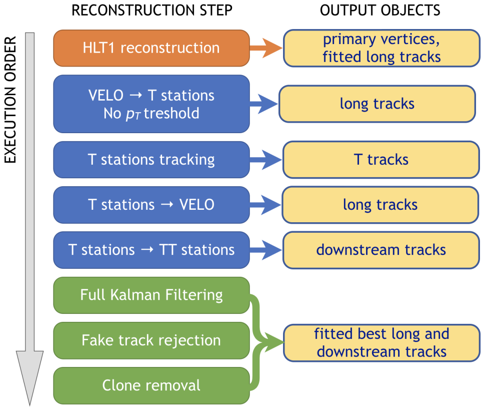




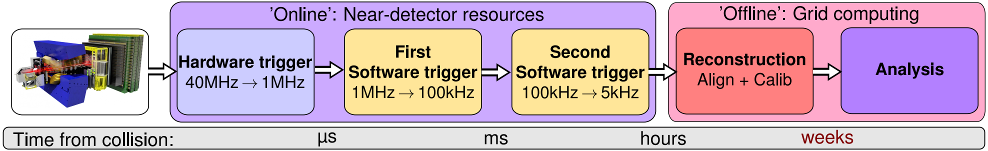
- All detectors were aligned & calibrated in-situ using the full HLT1 output rate
- Updates applied automatically if needed prior to HLT2 starting

HLT 2: Full event reconstruction

- At HLT2 the full reconstruction is performed down to 0 p_{T}
- Long and downstream tracks are avaiable for physics
- Full Particle ID is available (RICH, MUON, CALO)
- All quantities are now 'offline quality' after alignment & calibration
- Several hundred inclusive & exclusive selections, resulting in 600MB/s sent offline for analysis

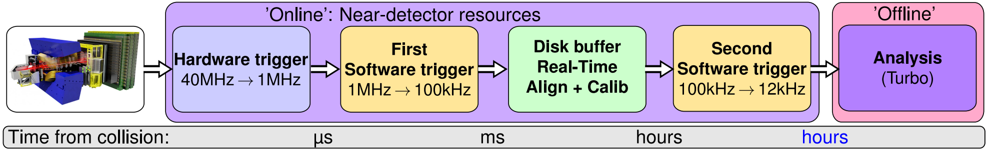


HLT2: Reduced event formats



- Trigger *rates* aren't important, output *bandwidth* is
- Offline reprocessing previously needed to recover best quality

HLT2: Reduced event formats



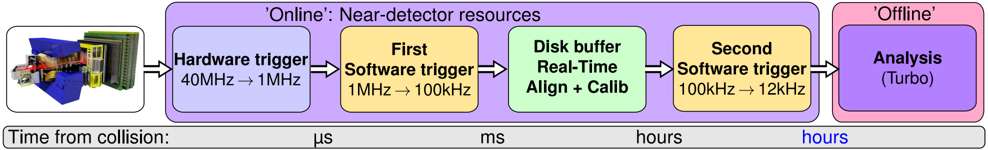
- Trigger *rates* aren't important, output *bandwidth* is
- Offline reprocessing previously needed to recover best quality
- After alignment: online == offline, why reprocess? Do analysis on trigger objects at HLT2, write only the relevant objects offline
- Significant reduction in event size → higher rates for the same bandwidth

More with less
Introduction
Run 2 Trigger
L0
HLT1
Buffer
Alignment & Calibration
HLT2
Selected physics results
Turbo enabled
Oscillations
Anomalies
Run 3 (Upgrade 1)
Triggerless readout
GPU HLT1
Conclusions

C. Fitzpatrick

March 30, 2021

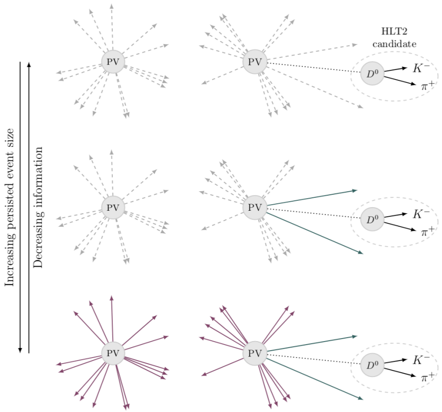
HLT2: Reduced event formats



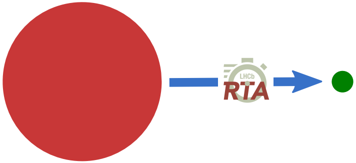
- Trigger *rates* aren't important, output *bandwidth* is
- Offline reprocessing previously needed to recover best quality
- After alignment: online == offline, why reprocess? Do analysis on trigger objects at HLT2, write only the relevant objects offline
- Significant reduction in event size → higher rates for the same bandwidth
- Added bonus: offline CPU freed up for simulation.

Turbo

- Turbo is LHCb's Real-Time Analysis paradigm for reduced event format data[5]



- High degree of flexibility: Save only as much of the event as is needed for analysis
  - Keep all reconstructed objects, drop the raw event: < 100kB
  - Keep only objects used to trigger: 7kB
  - 'Selective Persistence' objects used to trigger + user-defined selection: $7 \rightarrow 100$kB



LHCb Raw data
100 kB per collision

LHCb RTA-enabled data
7kB per collision

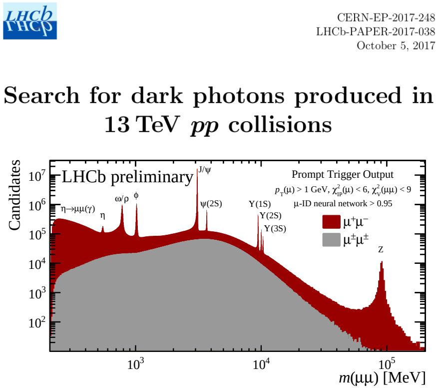
[5] arXiv:1604.05596, JINST 14 (2019) P04006

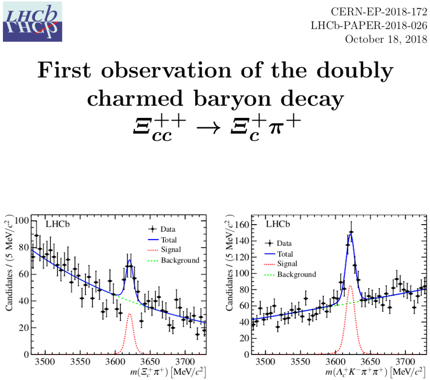
C. Fitzpatrick

March 30, 2021

Turbo in Run 2

- 528 trigger lines at HLT2. 50% were Turbo
- 25% of the trigger rate was Turbo but it counted for only 10% of the bandwidth
- Many analyses would not have been possible without Turbo[6]

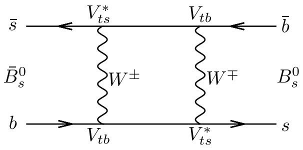




[6] Phys. Rev. Lett. 120, 061801 (2018), Phys. Rev. Lett. 121, 162002 (2018)

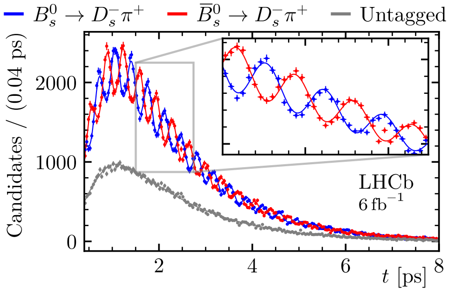
Beautiful oscillations

- One interesting feature of the flavour changing process is that neutral mesons can oscillate via the weak current:



- Time-evolution: Oscillates with frequency Δm_s
- Can be measured using flavour specific decays:
  - $B_s^0 \to \overline{B}_s^0 \to D_s^+\pi^-$
  - $\overline{B}_s^0 \to B_s^0 \to D_s^-\pi^+$
- Uses knowledge of the initial $B_s^0/\overline{B}_s^0$ flavour and excellent decay time resolution

- Latest LHCb measurement is extremely precise:



- $\Delta m_s = 17.7683 \pm 0.0051 \pm 0.0032\text{ps}^{-1}$ PRELIMINARY[7]
- Beautiful demonstration of quantum mechanics

[7] LHCb-PAPER-2021-005 in preparation

C. Fitzpatrick

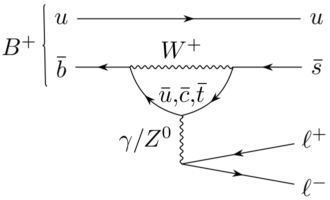
March 30, 2021

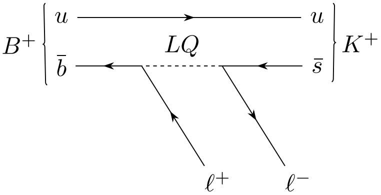
The anomalies: R_K and friends

- In the SM couplings of gauge bosons to leptons are independent of lepton flavour
- Accounting for phase space and helicity suppression effects we would expect:

$$R_K = \frac{\mathcal{B}(\mathrm{B}^+ \to \mathrm{K}^+ \mu^- \mu^+)}{\mathcal{B}(\mathrm{B}^+ \to \mathrm{K}^+ \mathrm{e}^- \mathrm{e}^+)} \simeq 1(\mathrm{SM})$$

- QCD uncertainties predicted at the level of 10^{-4} and QED corrections below 1%[8]
- Deviations from $R_K = 1$ could imply new physics processes:

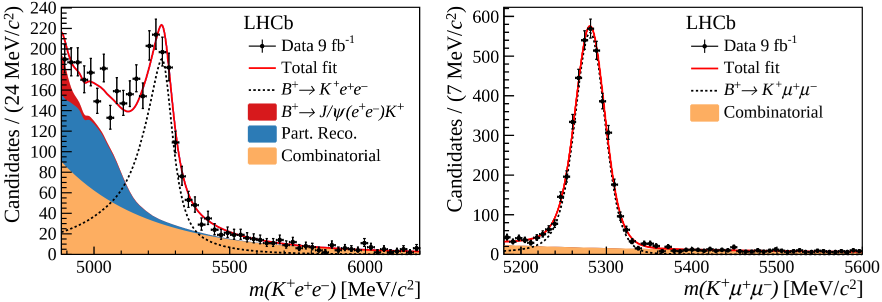




[8] JHEP12 (2007) 040, EPJC 76 (2016) 8, 440

Measuring R_K

- Biggest challenge: Muons and electrons are detected and triggered differently.
  - Electrons lose energy to Bremsstrahlung, must be recovered by looking for compatible clusters in the calorimeter
  - Trigger thresholds for electrons are higher than muons due to noisier environment. Electron channel uses several trigger categories.



  - Electron mode still has poorer q^2 and mass resolution due to detector effects

C. Fitzpatrick
March 30, 2021

Controlling detection differences

- The J/ψ and ψ(2S) resonant modes are used extensively to cancel and cross-check differences.
- Double-ratio taken with J/ψ mode cancels out several systematics:

$$R_K = \frac{\mathcal{B}(B^+ \to K^+ \mu^- \mu^+)}{\mathcal{B}(B^+ \to K^+ J/\psi(\mu^- \mu^+))} \bigg/ \frac{\mathcal{B}(B^+ \to K^+ e^- e^+)}{\mathcal{B}(B^+ \to K^+ J/\psi(e^- e^+))}$$

- Analysts measure $r_{J/\psi} = 0.981 \pm 0.020$ and $r_{\psi(2S)} = 0.997 \pm 0.011$
- Confirms expectation and constitutes most precise LFU test in the ψ(2S) mode.
- Dominant sources of systematic uncertainty:
  - Choice of fit model
  - Finite sample size of calibration samples

C. Fitzpatrick

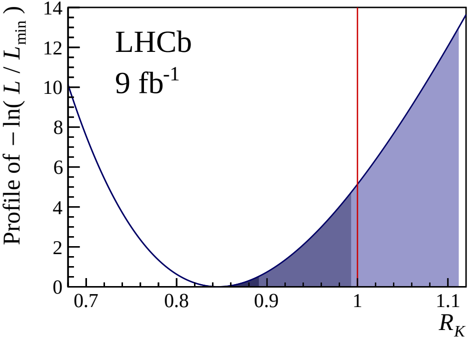
March 30, 2021

R_K Result

- R_K determined on $1.1 < q^2 < 6.0\text{GeV}^2$

$$R_K = 0.846^{+0.042}_{-0.039}\ (\text{stat})^{+0.013}_{-0.012}\ (\text{syst})^9$$

- p-value of SM hypothesis is 0.001.

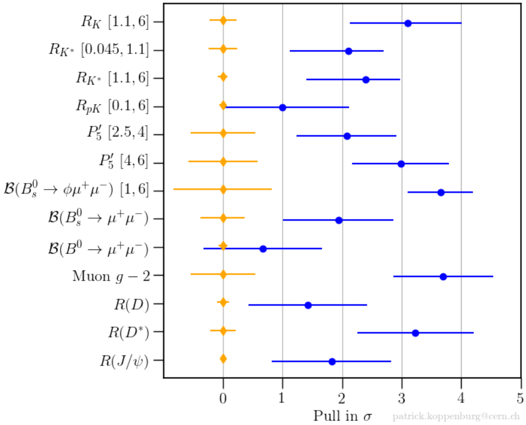


- Accounting for the 1% theory uncertainty, Evidence of LFU violation at 3.1σ

[9] LHCb-PAPER-2021-004 PRELIMINARY

Is 3.1σ cause for cautious excitement?

- Status of anomalies involving muons[10]:

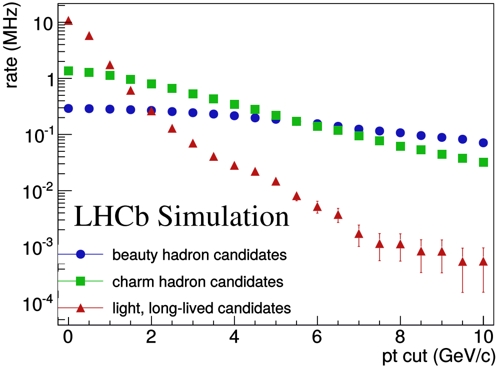


- There appears to be a trend in measurements involving muons
- The coming $g-2$ update may help clarify the situation independently of LHCb.

[10] Figure by P. Koppenburg

The MHz signal era

- LHCb is currently upgrading most of its detectors
- Starting in 2022, The 'new' LHCb will run at five times the collision rate:



- Even after simple trigger criteria, MHz of signals [11]

[11] LHCb-PUB-2014-027

The MHz signal era

- LHCb is currently upgrading most of its detectors
- Starting in 2022, The 'new' LHCb will run at five times the collision rate:

- Even after simple trigger criteria, MHz of signals [11]

[11] LHCb-PUB-2014-027

So what 'stuff' can we throw away?

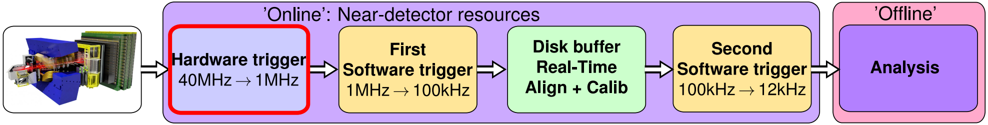
- The problem is no longer one of rejecting (trivial) background
- Fundamentally changes what it means to trigger

- Instead, we need to categorise different 'signals'
  - Requires access to as much of the event as possible, as early as possible

C. Fitzpatrick

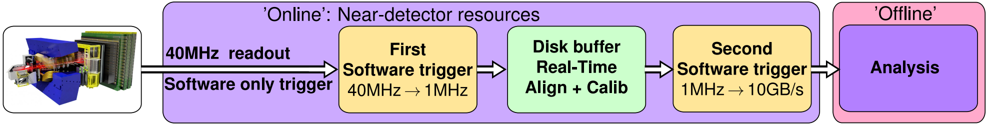
March 30, 2021

Reading out at 30MHz



- The L0 trigger cannot reduce the rate to the 1 MHz readout limit without throwing away signal

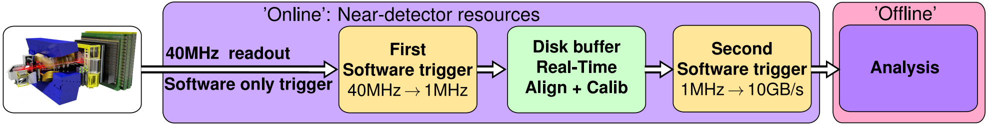
Reading out at 30MHz



- The L0 trigger cannot reduce the rate to the 1 MHz readout limit without throwing away signal

- The software triggers are pure: Can use the full event to make the decision

Reading out at 30MHz

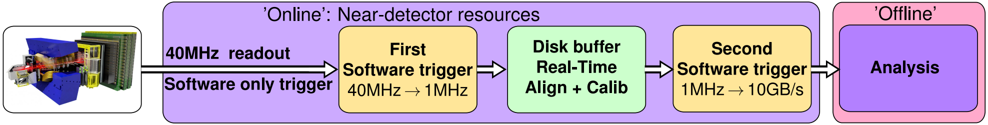


- The L0 trigger cannot reduce the rate to the 1 MHz readout limit without throwing away signal

- The software triggers are pure: Can use the full event to make the decision
- Solution: Readout and reconstruct 30 MHz of collisions in software

C. Fitzpatrick
March 30, 2021

Reading out at 30MHz

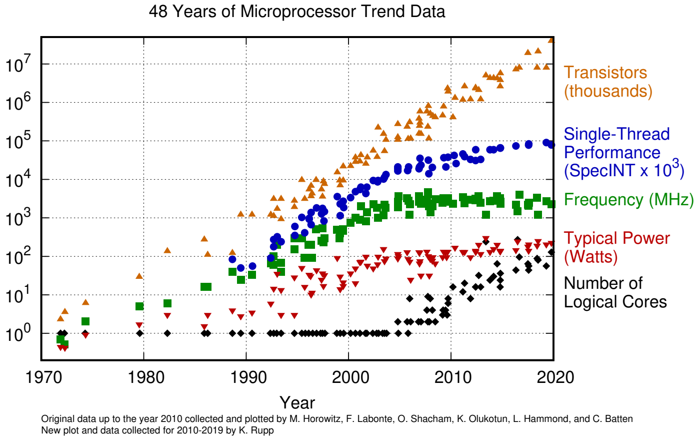


- The L0 trigger cannot reduce the rate to the 1 MHz readout limit without throwing away signal

- The software triggers are pure: Can use the full event to make the decision
- Solution: Readout and reconstruct 30 MHz of collisions in software
  - HLT1 similar to the Run 2 design but now must operate at the 30 MHz visible interaction rate
  - HLT2 input rate increased to 1 MHz and will produce **mostly turbo output** at 10GB/s

Doing more with less

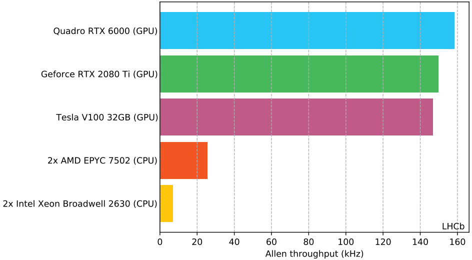
- 30 × the HLT1 input rate without 30 × the cash is a challenge

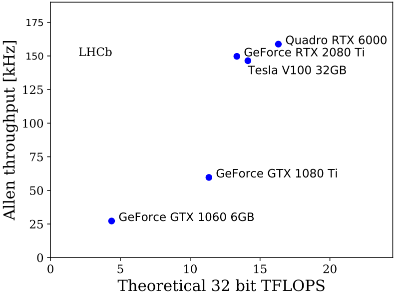


- Processing technologies have transitioned from higher CPU frequencies to incresed parallelism.
- Requires a dramatic change in how we design and run our software
- LHCb is at the leading edge of a wider trend in HEP data processing

Introducing Allen, an HLT1 exclusively on GPUs

- R&D efforts for Run 3 followed two technology options:
  - CPU: LHCb-TDR-016
    - Transition to a fully multithreaded HLT1 & HLT2
    - Exploit vectorisation where possible, restructure data formats
    - Make use of a lightweight scheduler to maximise CPU utilisation
  - GPU: LHCb-TDR-021
    - Implement entire HLT1 reconstruction and triger on GPU
    - Rewrite all HLT1 algorithms in Allen, a new CUDA framework
    - For now, keep HLT2 on CPU

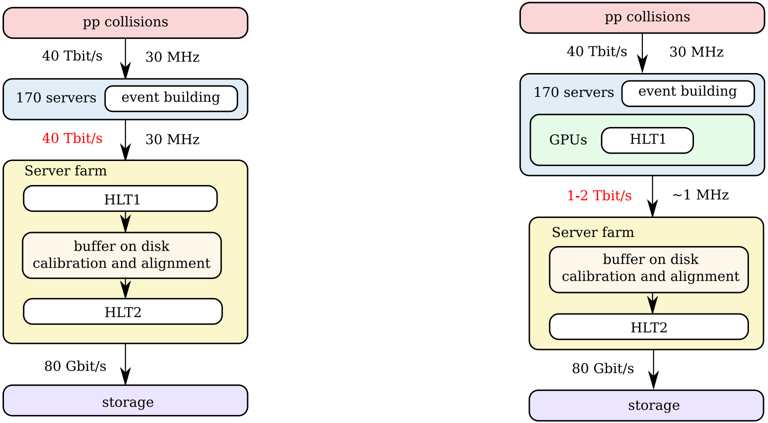




- After delivery of both options, we performed a global cost optimisation to determine the Run 3 baseline

Cost considerations

- Events are built on dedicated nodes in both scenarios
- These are then processed by HLT1 on the filter farm (CPU) or GPU cards (Allen)



- Significant cost saving comes from reduced network infrastructure → GPU HLT1 adopted as baseline
- Performance scaling (previous slide) shows great promise for expansion with future GPU generations

C. Fitzpatrick

Lessons learned from Allen

- The LHCb experience with GPUs/Allen provides some ideas for future R&D:
- While a CPU-based HLT1 is capable of performing a 30MHz reconstruction, the GPU option does so at lower total cost
- General considerations that apply online and offline:
  - Allen shows that we can use GPUs as quasi-standalone trigger processors in a domain where we have the equivalent of a few microseconds per event for the reconstruction and selection.
  - Works well if you keep overheads to a minimum and use the deep memory buffer of the host CPU nodes to smooth out I/O fluctuations.
  - Programmed correctly GPUs can handle complex and even somewhat non-linear data/control flows, as well as complex memory allocation patterns. As with all high-throughput computing the bottlenecks are related to memory management, not TFLOPs for computation.
  - No application is off-limits for GPUs anymore.

Conclusions

- LHCb has an exciting programme of precision measurements and interesting anomalies to explore with Run 3. However:
- LHCb signal rates in the first LHCb Upgrade change the definition of a trigger:
  - 'Rejects background' $\rightarrow$ 'categorises signal'
  - 'Reduces rate' $\rightarrow$ 'Reduces bandwidth'
- To efficiently categorise MHz signals, LHCb will use a triggerless readout into a 30 MHz GPU-based first level trigger
- Offline quality selections mean only subset of the event has to be saved for analysis at the CPU-based HLT2
  - Requires fully aligned & calibrated detector in the trigger
- The RTA paradigm allows LHCb to do **More Physics** with **Less Bandwidth**
  - The upgraded trigger will reduce systematic uncertainties and increase signal efficiencies for future LFU measurements

C. Fitzpatrick

March 30, 2021

The BEAUTY2CHARM team

BEAUTY2CHARM
European Research Council
Established by the European Commission



- ERC-StG-852642 BEAUTY2CHARM aims to:
  - Commission the LHCb upgrade trigger, reconstruction & data processing
  - Exploit the upgrade data to test the SM with $B \to DX$ and $B \to DD$ decays
- PDRAs Nicole, Eva (ex. Lund!) and PhD student Jonathan
- Will be joined by an additional student this year