Tests of lepton flavor universality and real time event selection on GPUs at LHCb

Dorothea vom Bruch

Aix Marseille Univ, CNRS/IN2P3, CPPM

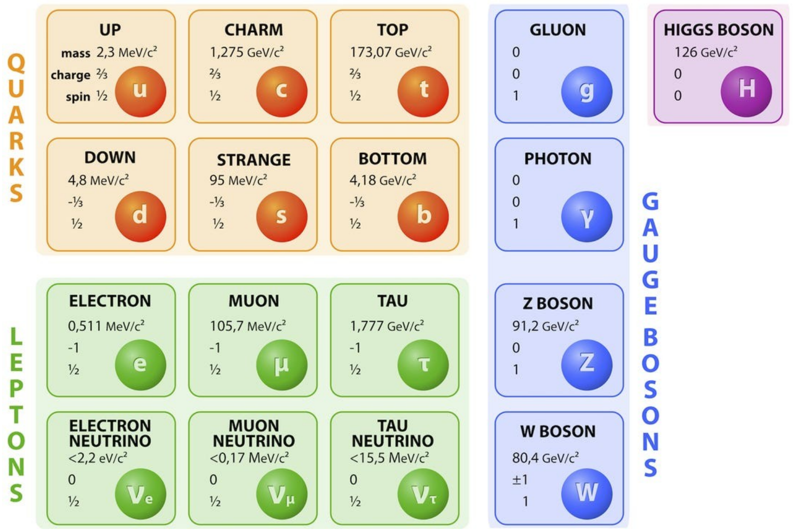
February 1st 2021

CPPM seminar

Search for New Physics

- The Standard Model describes most particle physics phenomena extraordinarily well
- However, some phenomena on the macroscopic scale are not explained:
  - What is dark energy, dark matter?
  - Where does the matter-antimatter asymmetry come from?
- Some characteristics of the SM are also not understood:
  - Why are there three flavor generations?
  - Where does the pattern of masses and mixings of quarks and leptons come from?

→ **Resolve discrepancies with particles or forces at new energy scales**



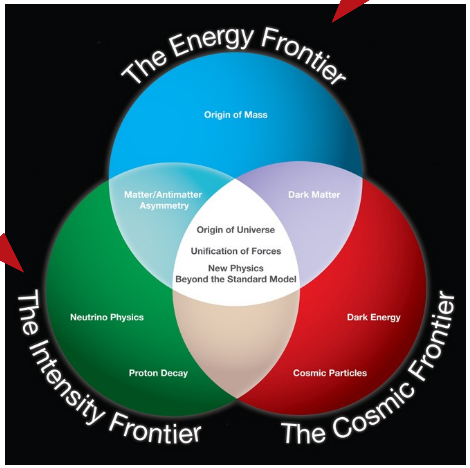
How to search for New Physics?

Indirect searches

- Precision measurements of precisely calculated observables
- For example measuring
  - Lepton flavor universality & violation
  - Angular distributions
  - …

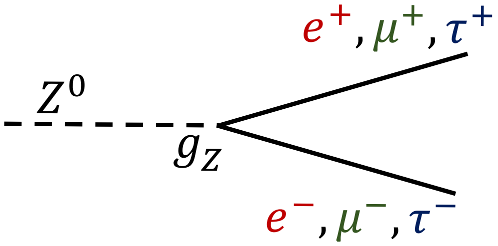
Direct searches

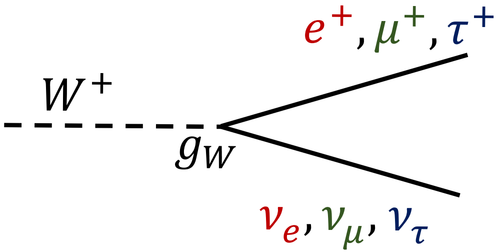
- Directly detect new particles
  - Observe naturally existing particles at dedicated experiments (WIMPs, Axions etc.)
  - Produce them with particle accelerators



Lepton flavor universality

The SM predicts equal couplings between the electroweak bosons and leptons
→ Lepton flavor universality





In ratios of branching fractions form factor uncertainties and dependence from CKM matrix elements partially cancel

→ theoretically clean to probe coupling

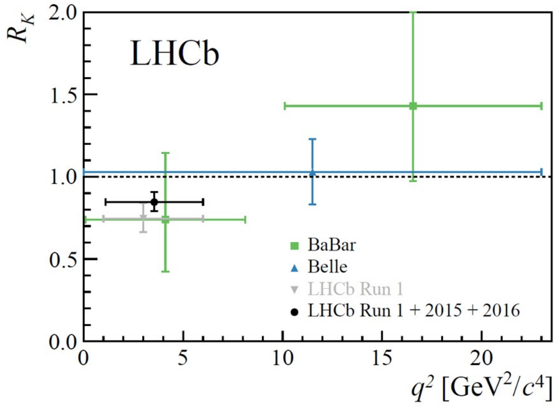
$$\frac{d\Gamma}{dq^2}(H_b \to H_c \tau \nu) \propto G_F^2 |V_{cb}|^2 f(q^2)^2$$

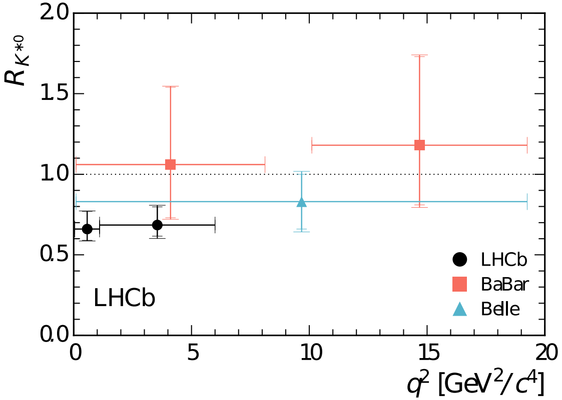
$$R(H_c) = \frac{B(H_b \to H_c \tau \nu)}{B(H_b \to H_c \mu \nu)}$$

Lepton flavor universality in B decays

b → sl$^+$l$^-$

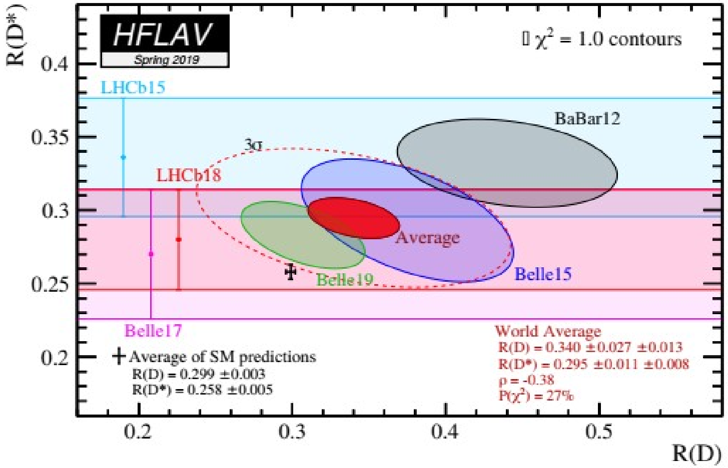
$\mathcal{R}(K^{(*)}) = \mathcal{B}(B\to K^{(*)}\mu^+\mu^-) / \mathcal{B}(B\to K^{(*)}e^+e^-)$





b → clν

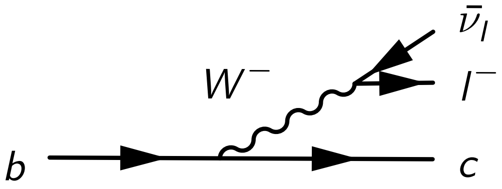
$\mathcal{R}(D^{(*)}) = \mathcal{B}(B\to D^{(*)}\tau\nu_\tau) / \mathcal{B}(B\to D^{(*)}\mu\nu_\mu)$



R(D) and R(D*) compatible with the SM at the 3.1σ level (increased to 3.8σ with latest *theory prediction)*

R(K) and R(K*) are compatible with the SM at 2.5 σ and 2.1-2.5 σ respectively

Semileptonic b → clν decays



Pros

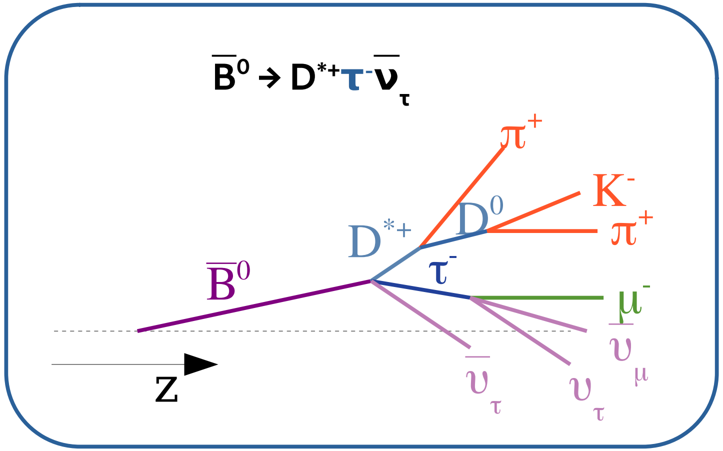
- High branching fraction → abundant
- Only one hadronic current → Theoretically clean

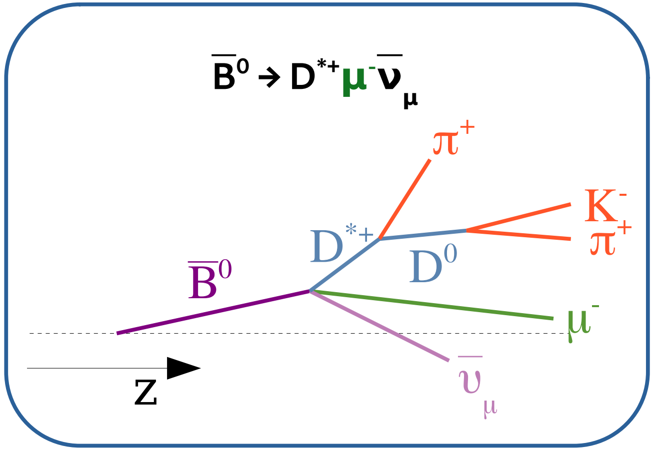
Cons

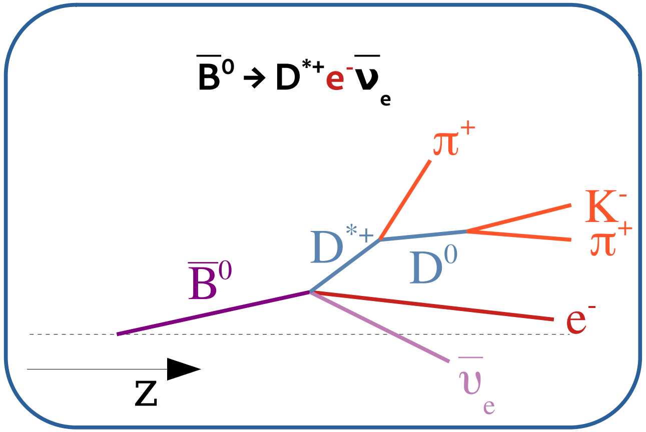
- Partially reconstructed signal (neutrinos!)
  → Experimentally difficult
- Many backgrounds
- Large simulated samples required

LFU tests with $b \to cl\nu$ transitions

- Many different decays to study: $\overline{B}^0 \to D^{(*)} l \nu$, $\overline{B}^0 \to D_s l \nu$, $\overline{B}^0 \to J/\psi l \nu$, $\overline{B}^0 \to \Lambda_c l \nu$
- In this talk: Focus on $\overline{B}^0 \to D^{*+} l \nu$

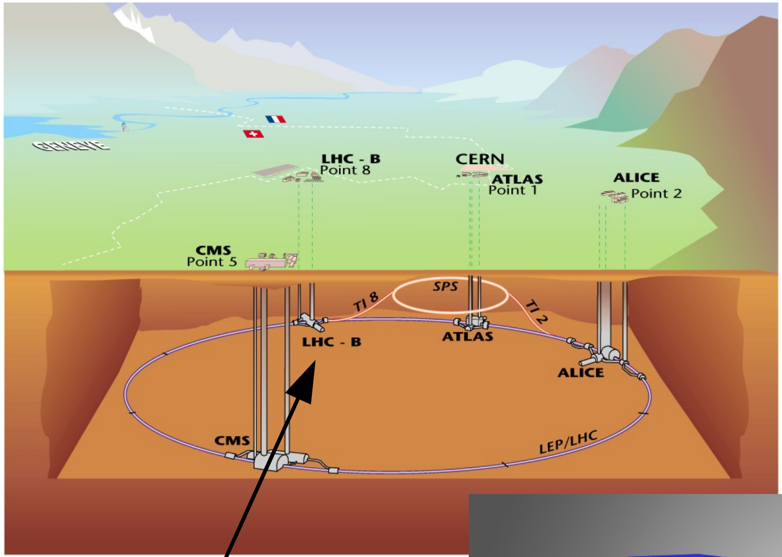




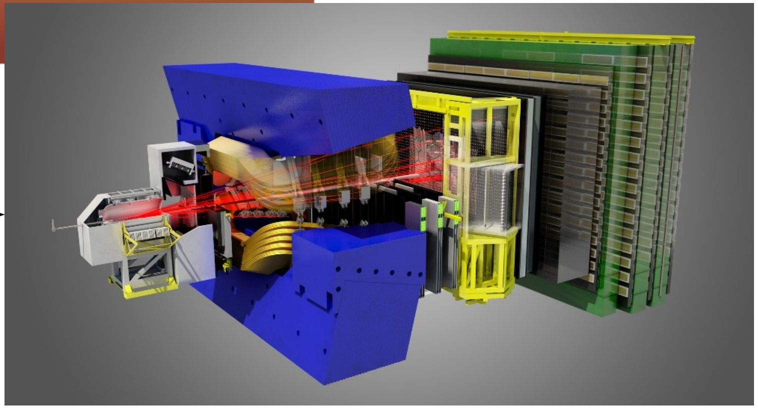


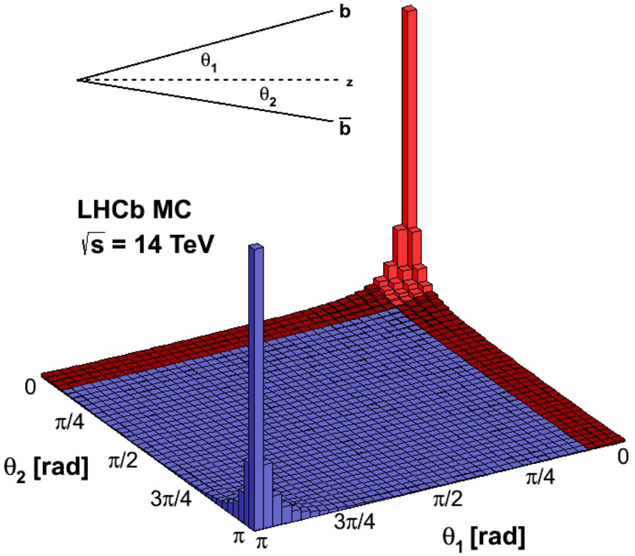
One place to study them: LHCb

LHC @ CERN



General purpose detector in the forward region specialized in beauty and charm hadrons





LHCb detector, 2011 - 2018

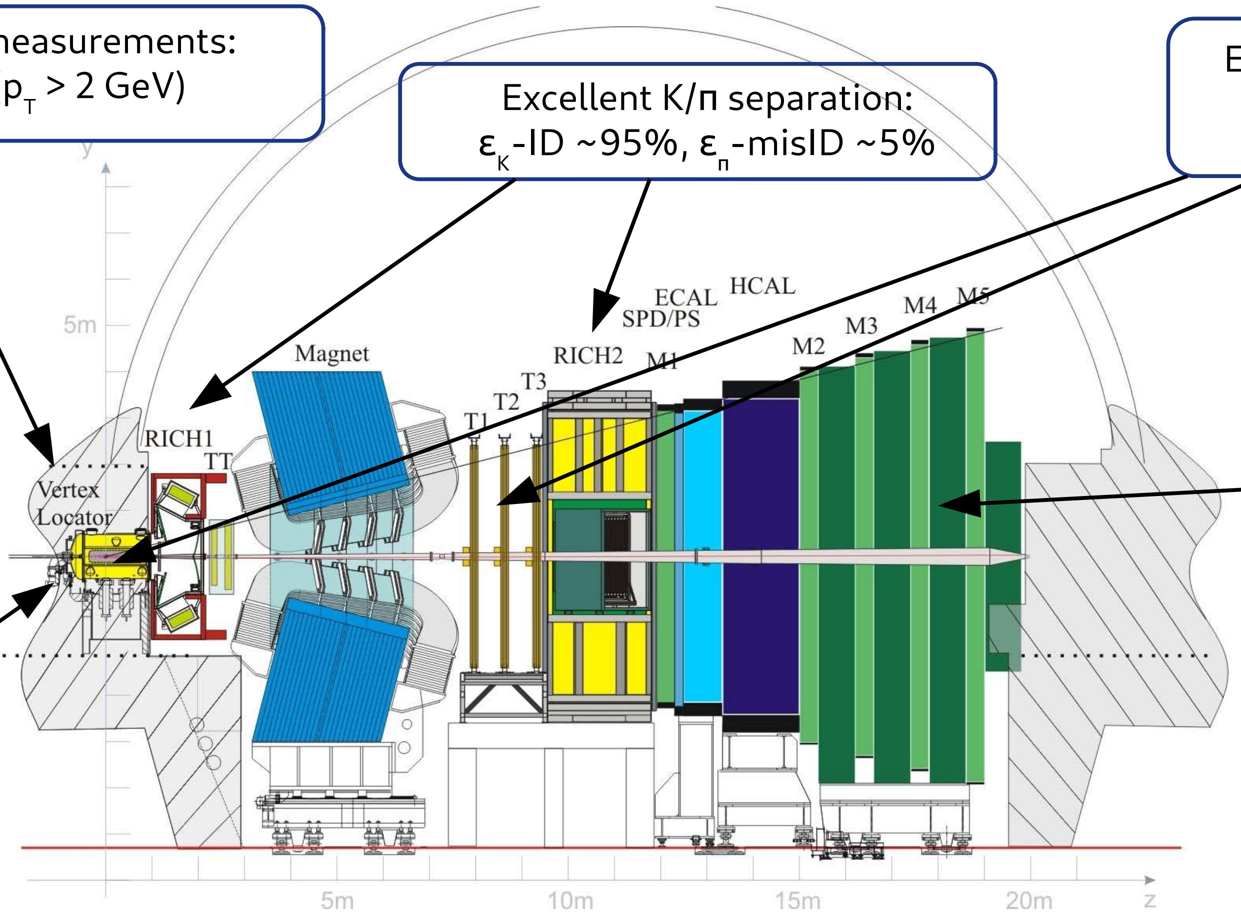
Precise vertex measurements:
σ_{IP} = 20 µm (p_T > 2 GeV)

Excellent K/π separation:
ε_K-ID ~95%, ε_π-misID ~5%

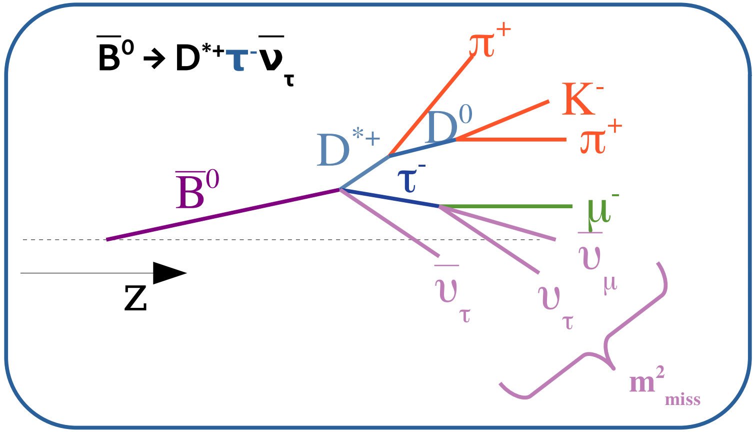
Excellent momentum resolution:
Δp/p ~0.5-1%

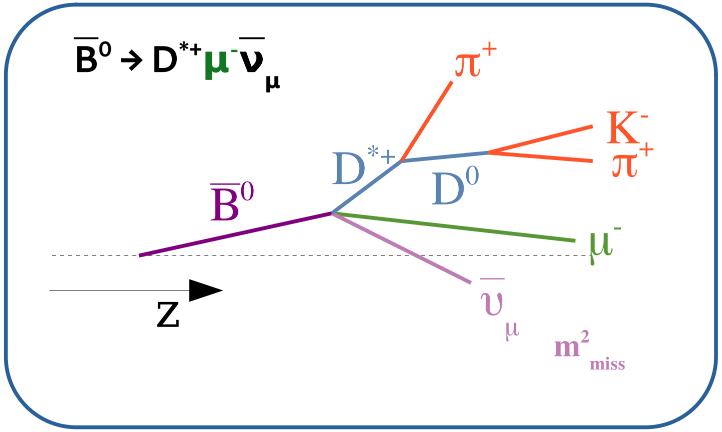
Excellent muon Identification:
ε_μ-ID ~97%

Excellent decay time resolution:
σ_τ ~ 45 fs
for b hadrons



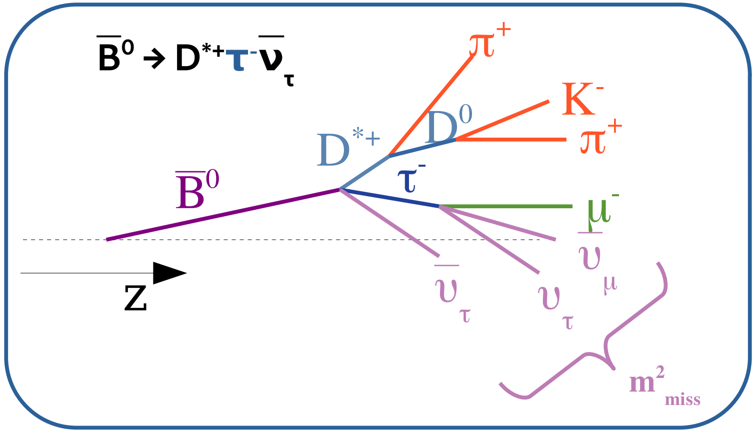
LHCb Run 1 $R(D^*)_{\tau\mu}$ with $\tau \to \mu\nu\nu$

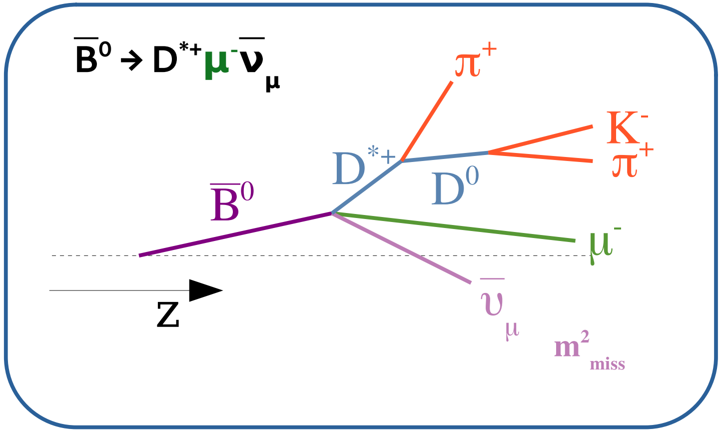


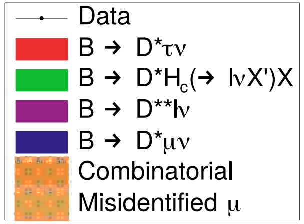


- Same final state
- Tau vertex not well constrained

LHCb Run 1 $R(D^*)_{\tau\mu}$ with $\tau \to \mu\nu\nu$



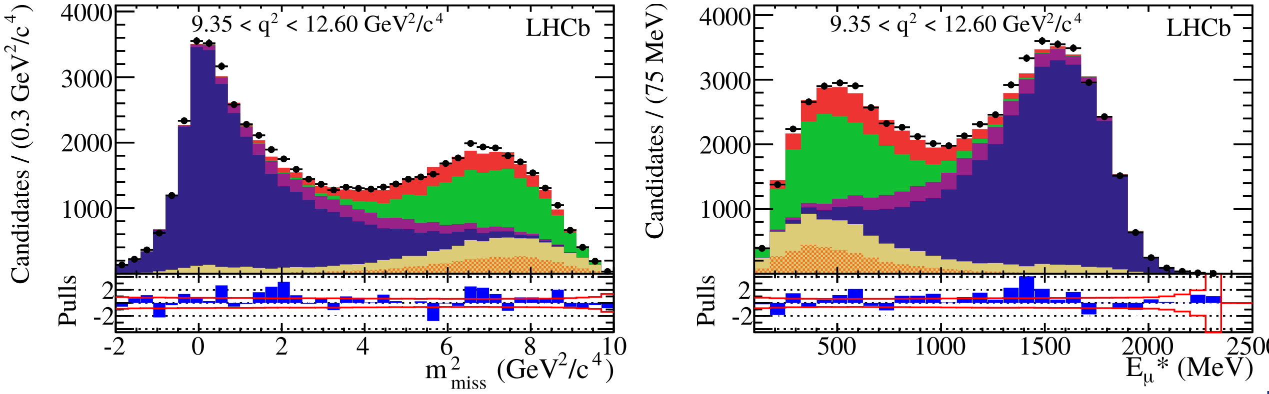




- Same final state
- Tau vertex not well constrained

Separate $\overline{B}^0 \to D^{*+}\tau^-\overline{\nu}_\tau$ **from** $\overline{B}^0 \to D^{*+}\mu^-\overline{\nu}_\mu$ (different colors!!)

$q^2 = (p_B^{\ 2} - p_{D^*}^{\ 2})$



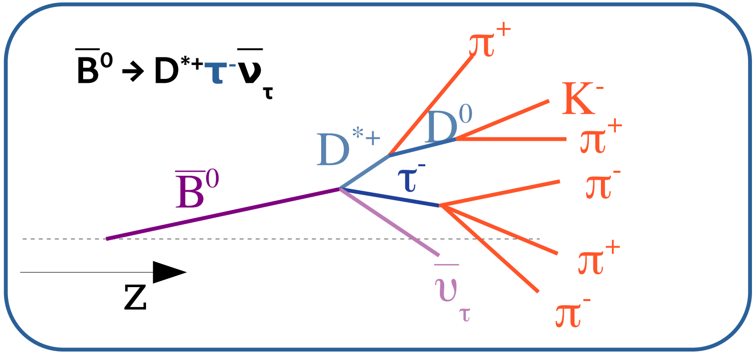
$\mathcal{R}(D^*) = \mathcal{B}(B \to D^{(*)}\tau\nu_\tau) / \mathcal{B}(B \to D^{(*)}\mu\nu_\mu)$

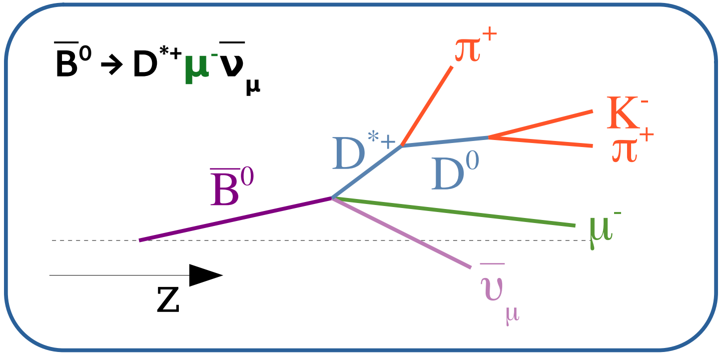
$= 0.336 \pm 0.027\ (\text{stat}) \pm 0.030\ (\text{syst})$

Analysis of Run 2 data ongoing

PRL 115 111803 (2015)

LHCb Run 1 $R(D^*)_{\tau\mu}$ with $\tau^+ \to \pi^+\pi^-\pi^+(\pi^0)\nu$



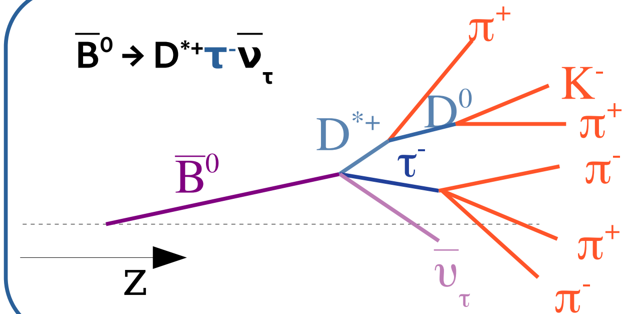


- Tau vertex well constrained
- Different final states
- Measure instead:
  $K(D^*) = \overline{B}^0 \to D^{*+}\tau^-\overline{\nu}_\tau \,/\, \overline{B}^0 \to D^{*+}3\pi^\pm$
- → Need external inputs:
  - $\mathcal{B}(\overline{B}^0 \to D^{*+}3\pi^\pm)$
  - $\mathcal{B}(\overline{B}^0 \to D^{*+}\mu^-\overline{\nu}_\mu)$

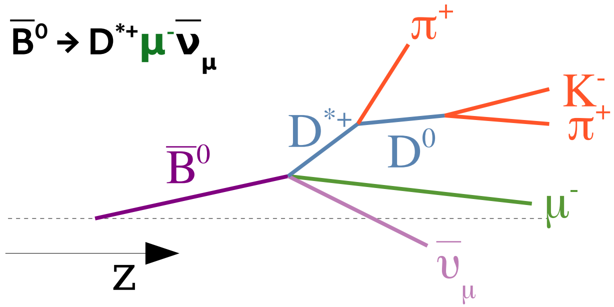
LHCb Run 1 $R(D^*)_{\tau\mu}$ with $\tau^+ \to \pi^+\pi^-\pi^+(\pi^0)\nu$

PRL 120, 171802 (2018)

$\overline{B}^0 \to D^{*+}\tau^-\overline{\nu}_\tau$



$\overline{B}^0 \to D^{*+}\mu^-\overline{\nu}_\mu$

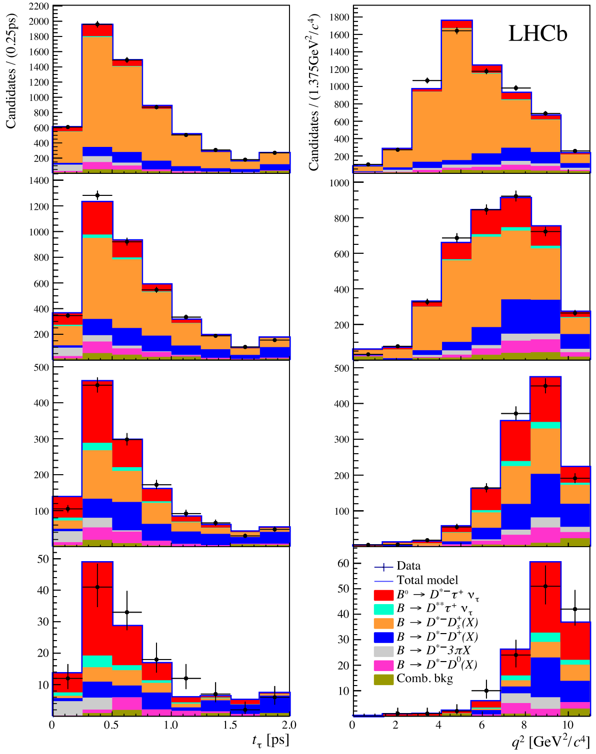


- Tau vertex well constrained
- Different final states
- Measure instead:
  $K(D^*) = \overline{B}^0 \to D^{*+}\tau^-\overline{\nu}_\tau \,/\, \overline{B}^0 \to D^{*+}3\pi^\pm$
- → Need external inputs:
  - $\mathcal{B}(\overline{B}^0 \to D^{*+}3\pi^\pm)$
  - $\mathcal{B}(\overline{B}^0 \to D^{*+}\mu^-\overline{\nu}_\mu)$

$\mathcal{R}(D^*) = 0.291 \pm 0.019$ (stat)

$\pm\, 0.026$ (syst) $\pm\, 0.054$ (ext)

Analysis of Run 2 data ongoing



In bins of BDT output trained to reduce D_s background

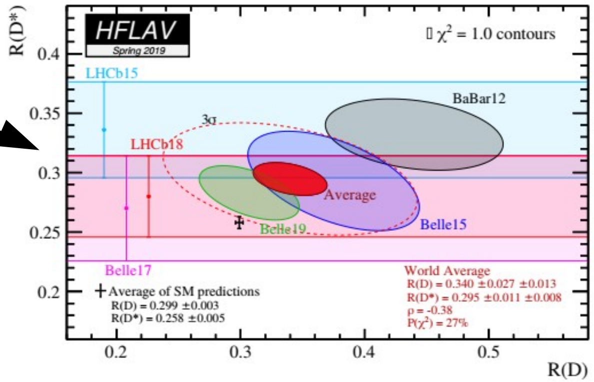
CPPM LHCb group involved

LHCb $R(D^*)_{e\mu}$

- Belle & BaBar have measured $R(D^*)$ for **τ** versus **e and μ,** LHCb only **τ** versus **μ**
- Goal: perform **e-μ-τ** universality test at LHCb
- First step: $R(D^*)_{e\mu} = B(B\to D^* e\nu_\tau) / B(B\to D^*\mu\nu_\mu)$
- First semileptonics analysis with electrons in the final state at LHCb
- First test of e-μ universality with $b \to cl\nu$ at LHCb
- Previously measured by Belle:

$R(D^*)_{e\mu} = 1.01 \pm 0.01\ (\text{stat}) \pm 0.03\ (\text{syst})$

Phys. Rev. D 100, 052007 (2019)



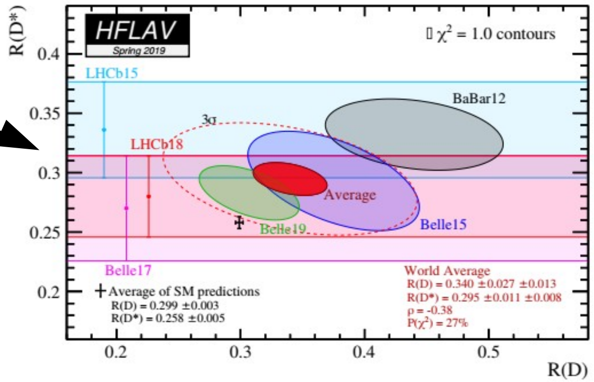
LHCb $R(D^*)_{e\mu}$

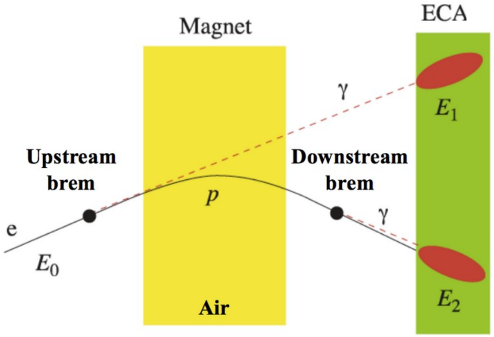
- Belle & BaBar have measured $R(D^*)$ for **τ** versus **e and μ,** LHCb only **τ** versus **μ**
- Goal: perform **e-μ-τ** universality test at LHCb
- First step: $R(D^*)_{e\mu} = B(B \to D^* e \nu_\tau) / B(B \to D^* \mu \nu_\mu)$
- First semileptonics analysis with electrons in the final state at LHCb
- First test of e-μ universality with $b \to cl\nu$ at LHCb
- Previously measured by Belle:

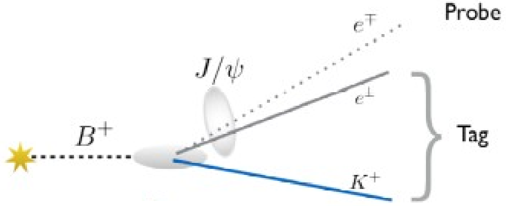
  $R(D^*)_{e\mu} = 1.01 \pm 0.01\ (\text{stat}) \pm 0.03\ (\text{syst})$

  Phys. Rev. D 100, 052007 (2019)

- Similar approach to $R(D^*)_{\tau\mu}$ with $\tau \to \mu\nu\nu$
- Challenge: electron reconstruction and yield at LHCb
- Signal yields from 3d template fit
- Efficiencies from simulation
- Electron reconstruction efficiency from tag-and-probe method on data







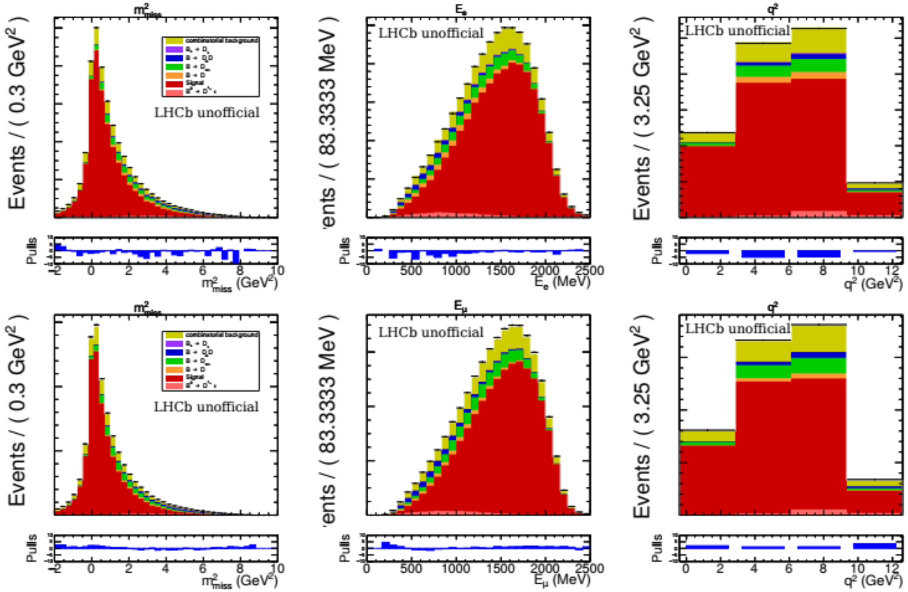
JINST 14 (2019) P11023

LHCb $R(D^*)_{e\mu}$

$$\mathcal{R}(D^*)_{e\mu} = \mathcal{B}(B \to D^* e \nu_\tau) / \mathcal{B}(B \to D^* \mu \nu_\mu)$$

$\overline{B}^0 \to D^{*+} e^- \overline{\nu}_\tau$

$\overline{B}^0 \to D^{*+} \mu^- \overline{\nu}_\mu$

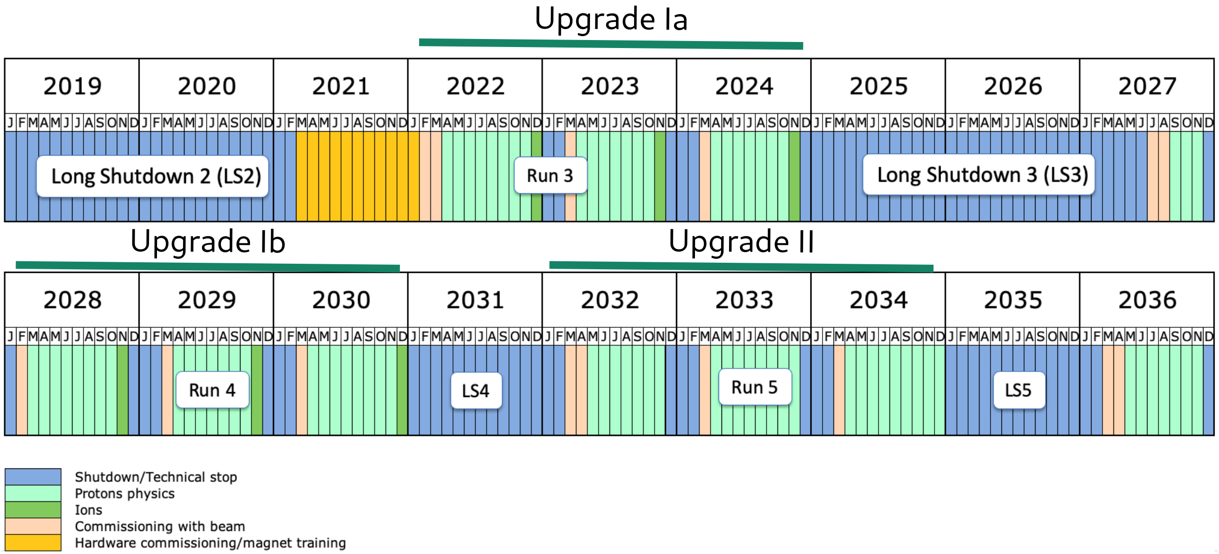


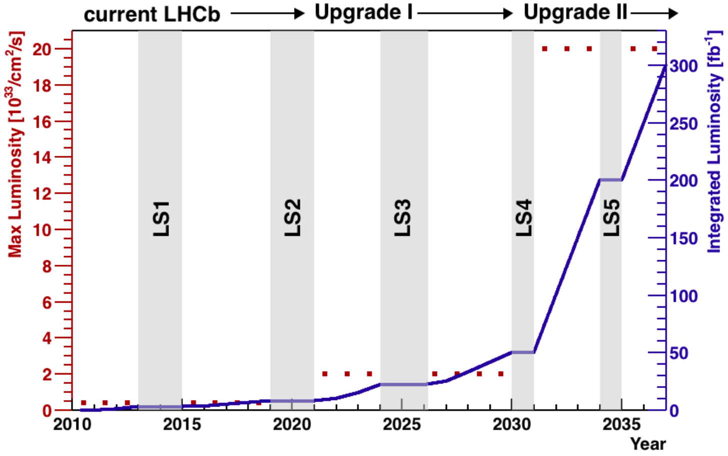
- Two distinct data samples
- Phase space chosen for similar electron and muon reconstruction efficiencies
- Looks like the uncertainties are under control
- Analysis under finalization

Proof of principle for future analyses with $b \to c e \nu$ and $b \to c \tau (\to e \nu\nu)\nu$ at LHCb
Can measure $\mathcal{R}(D^*)_{e\tau} = \mathcal{B}(B^0 \to D \tau (\to e \nu\nu)\nu) / \mathcal{B}(B \to D^* e \nu)$ in the future

CPPM LHCb group involved

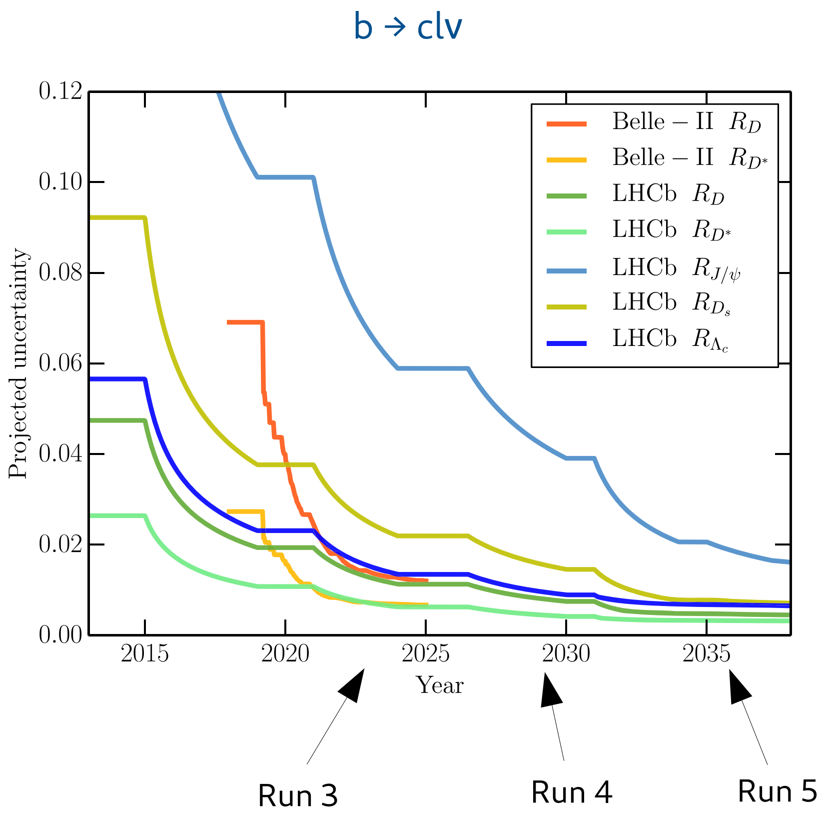
Much more data in the future!

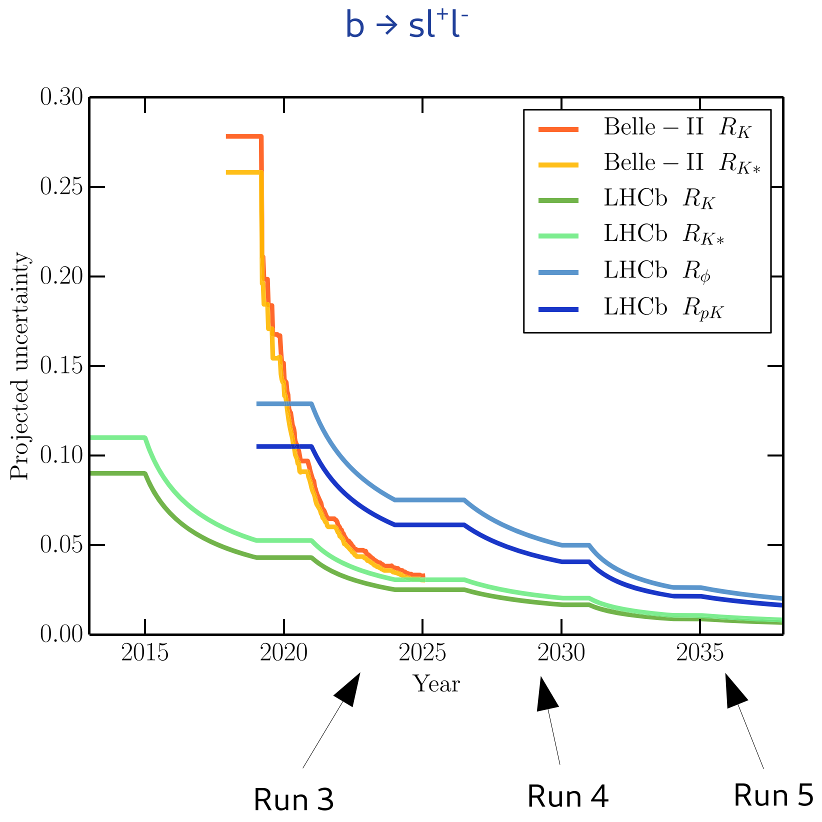




- Higher and higher precision needed in the search for new physics
- More data especially useful for theoretically clean & statistically limited observables, such as $b \to cl\nu$ transitions

Prospects for R(X)

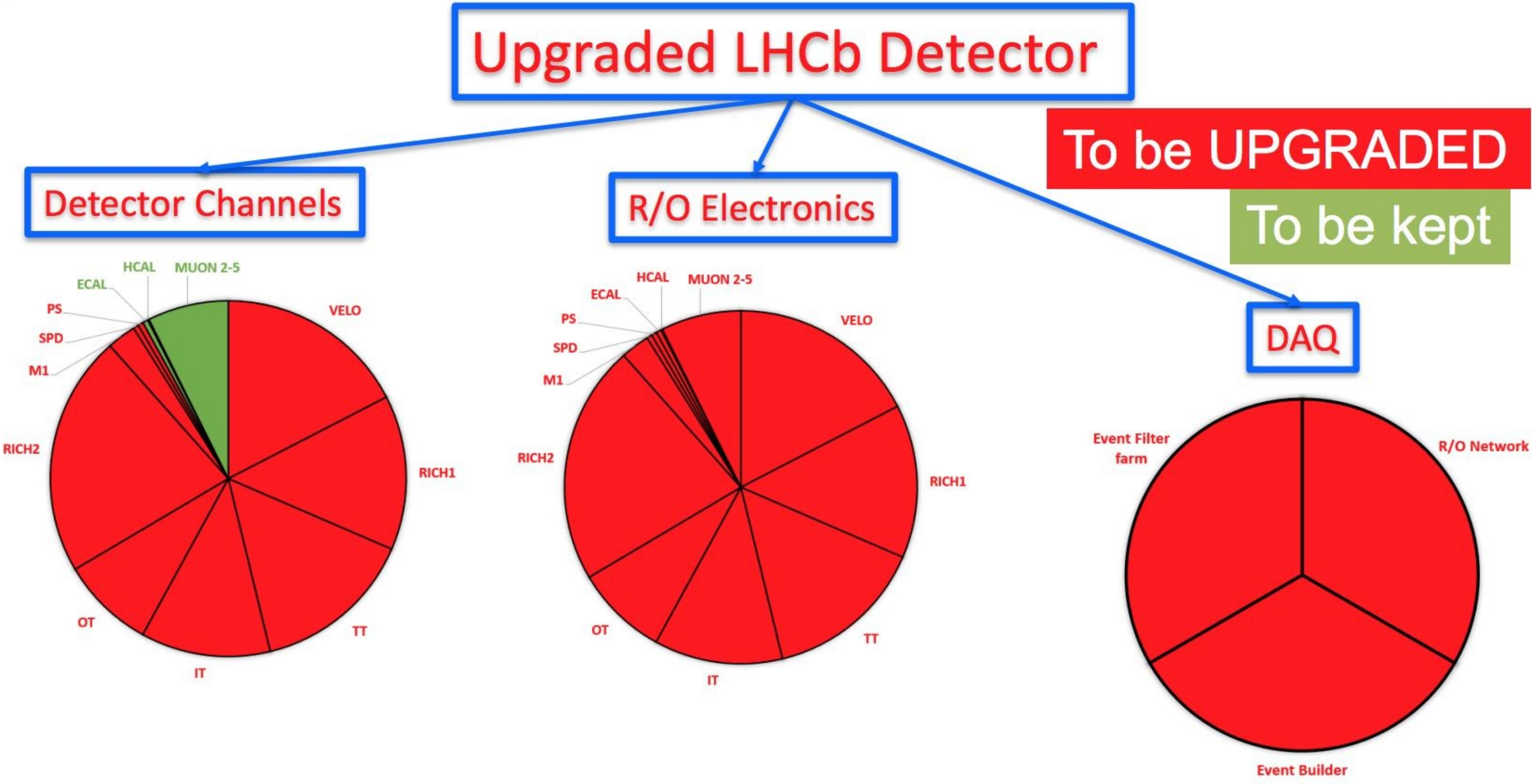




Run 3 and beyond will shed light on the flavor anomalies observed today

Upgrade 1 of LHCb for Run 3

LHCb Upgrade I

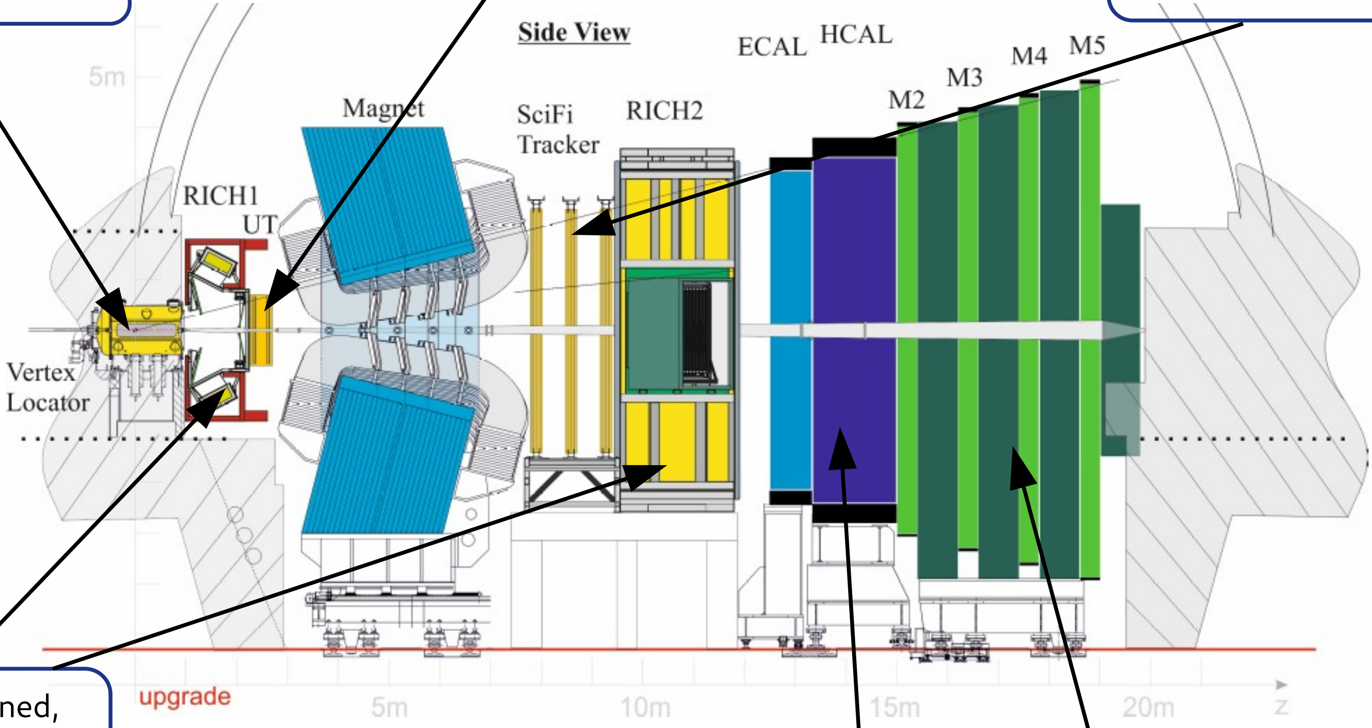


LHCb Upgrade I

New Vertex locator (Velo) tracking detector

New UT tracking detector

New scintillating fibre (SciFi) tracking detector

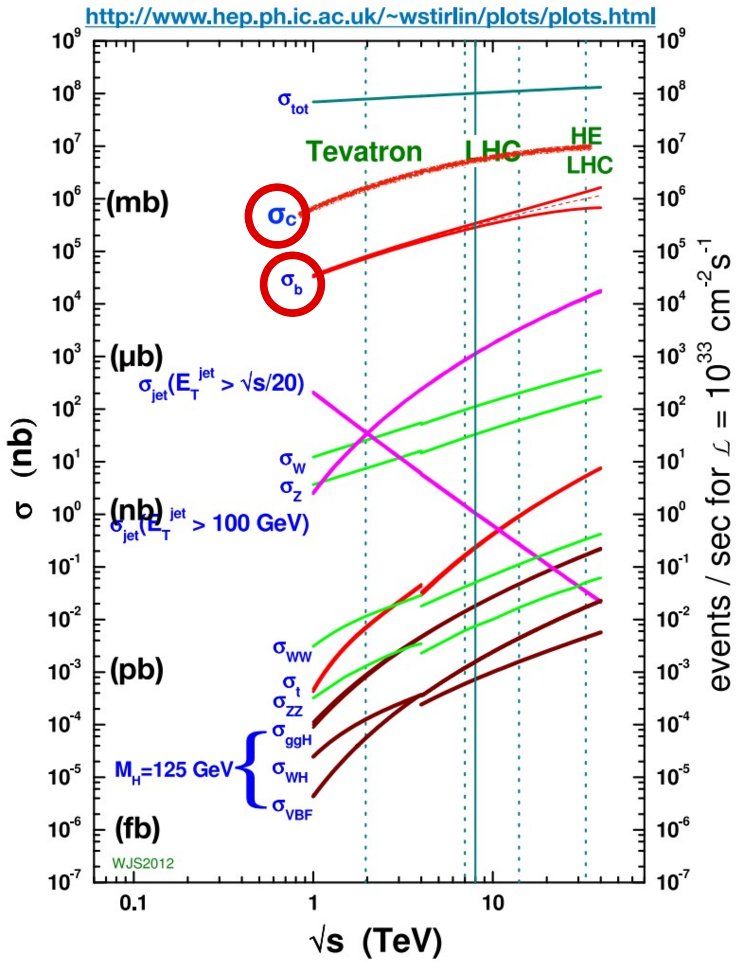


All readout systems renewed

RICH 1 redesigned, new photo detectors for RICH1 & RICH2

Calorimeters & muons: redundant components removed, new electronics, more shielding

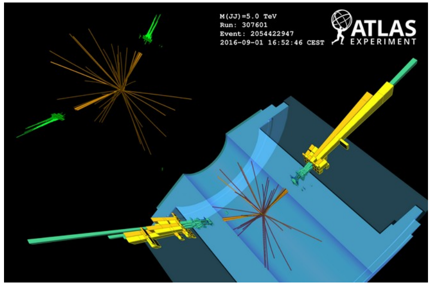
The MHz signal era



Run 3: Luminosity of $2x10^{33}$ $cm^{-2}s^{-1}$, $\sqrt{s}$ = 14 TeV

General purpose LHC experiments:

- Mainly direct searches
- Local characteristic signatures
- Signal rates up to ~100 kHz



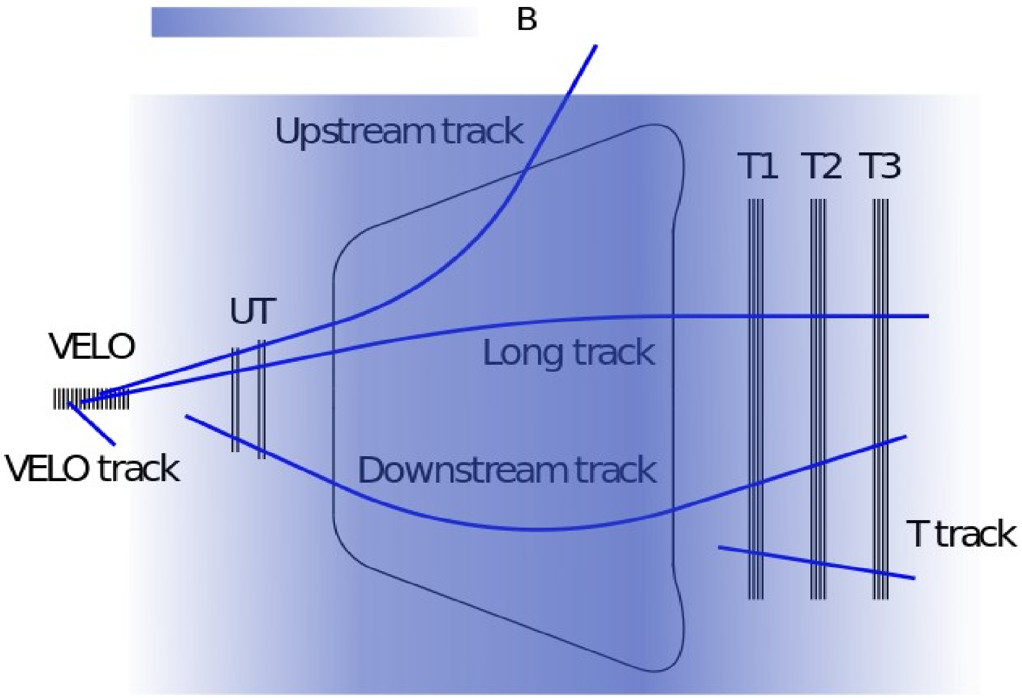
LHCb:

- Intensity frontier
- No “simple” local criteria for selection
- Signal rates up to ~MHz

Change in trigger paradigm

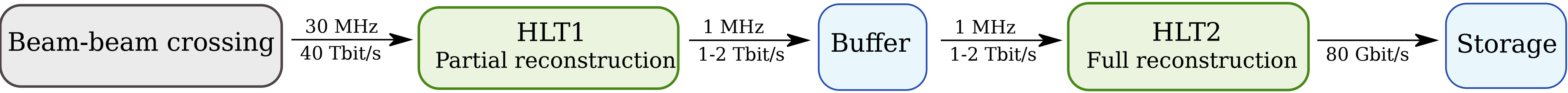
Access as much information about the collision as early as possible

Tracks in the LHCb detector



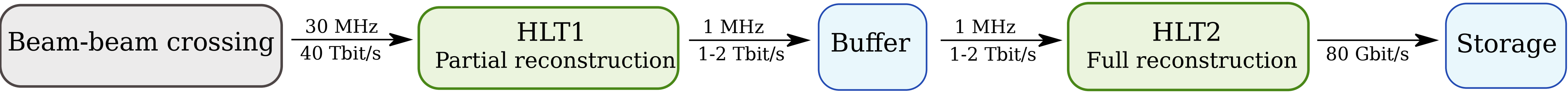
Need information from many subdetectors → read out full detector

Data selection only in software



- **High Level Trigger 1 (HLT1):**
  - Full charged particle track reconstruction
  - Few inclusive single and two-track selections
- **High Level Trigger 2 (HLT2):**
  - Real-time aligned and calibrated detector
  - Offline-quality track reconstruction
  - Particle identification
  - Full track fit

Data selection only in software



- **High Level Trigger 1 (HLT1):**
  - Full charged particle track reconstruction
  - Few inclusive single and two-track selections
- **High Level Trigger 2 (HLT2):**
  - Real-time aligned and calibrated detector
  - Offline-quality track reconstruction
  - Particle identification
  - Full track fit

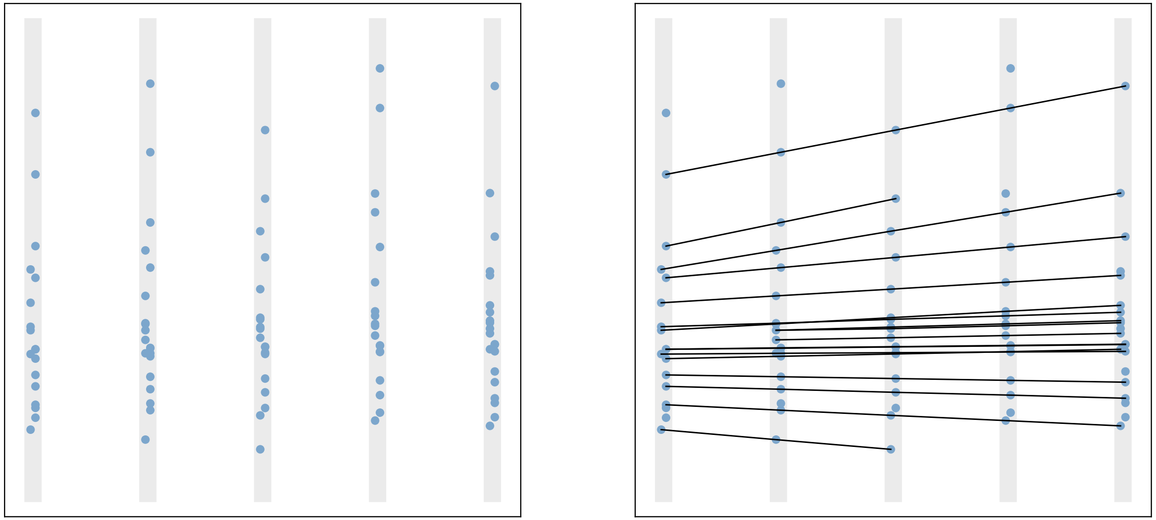
Comparison to Run II trigger

- 5 x higher pileup
- 30 x higher rate into HLT1
- Up to 10 x efficiency improvement for some physics channels

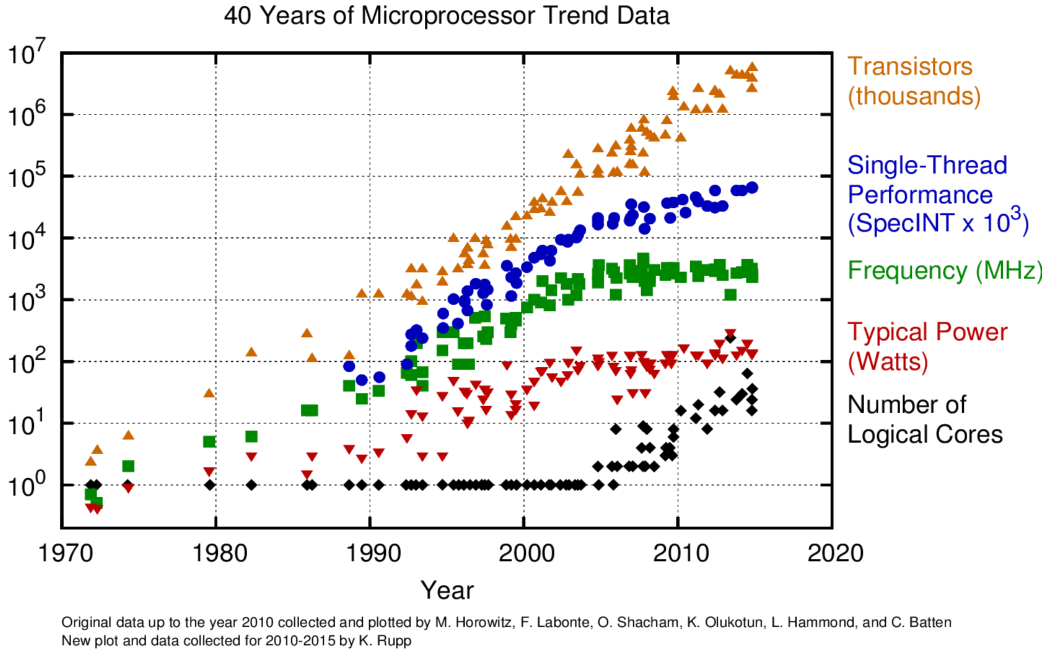
Huge computing challenge

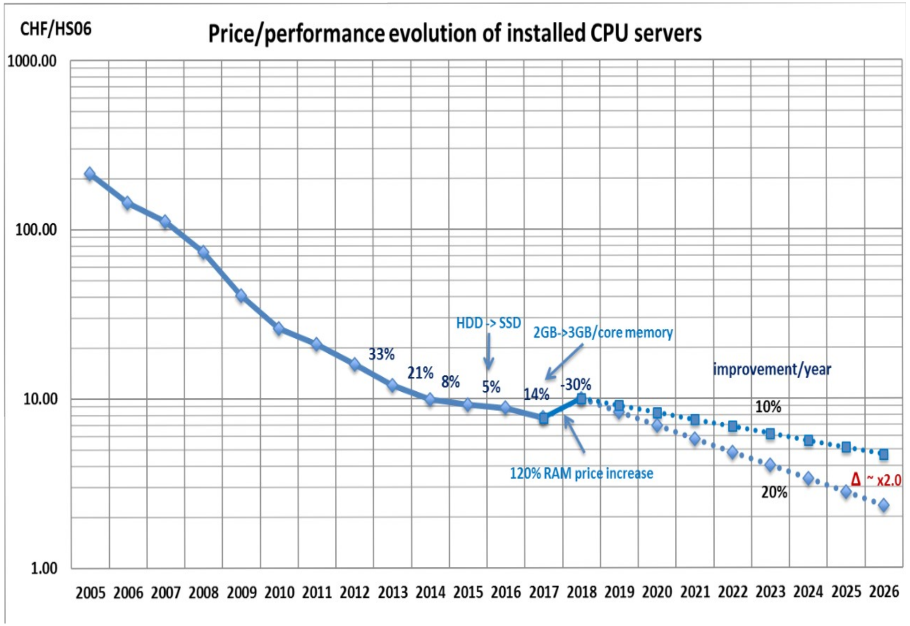
Track reconstruction @ 30 MHz

- Connect the dots to go from measurements to particle trajectories
- Many possible connections → huge combinatorics
- Do this for three sub-detectors, 30 million times per second



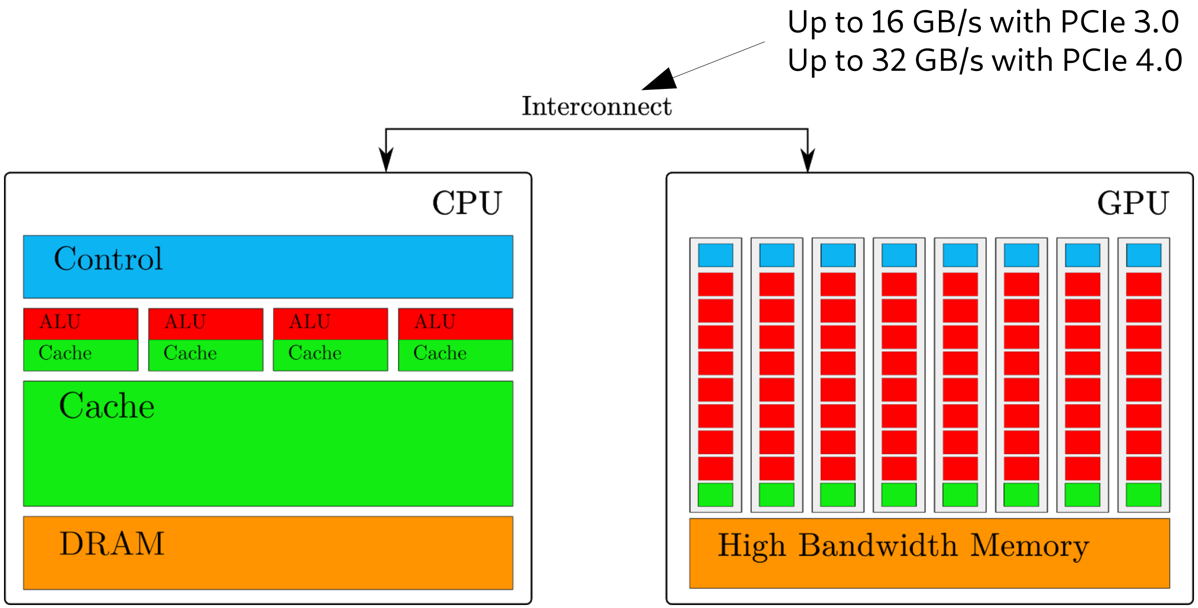
Today's computing landscape





Can we use the FLOPS available on the highly parallel architecture of Graphics Processing Units (GPUs) to run HLT1 @ 30 MHz?

GPU architecture design



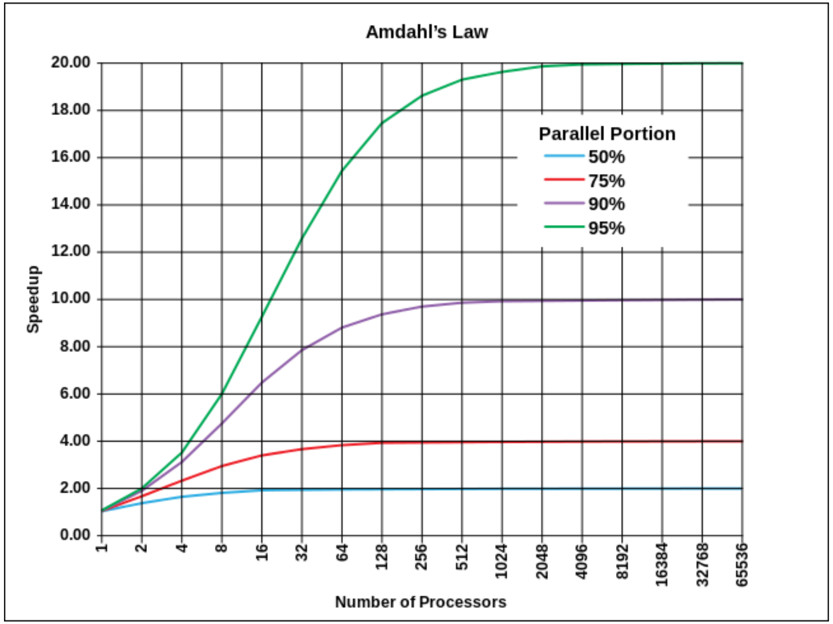
- Low core count / powerful ALU
- Complex control unit
- Large chaches

→ **Latency optimized**

- High core count
- No complex control unit
- Small chaches

→ **Throughput optimized**

When to go parallel? → Amdahl's law



Speedup in latency = 1 / (S + P/N)

- S: sequential part of program
- P: parallel part of program
- N: number of processors

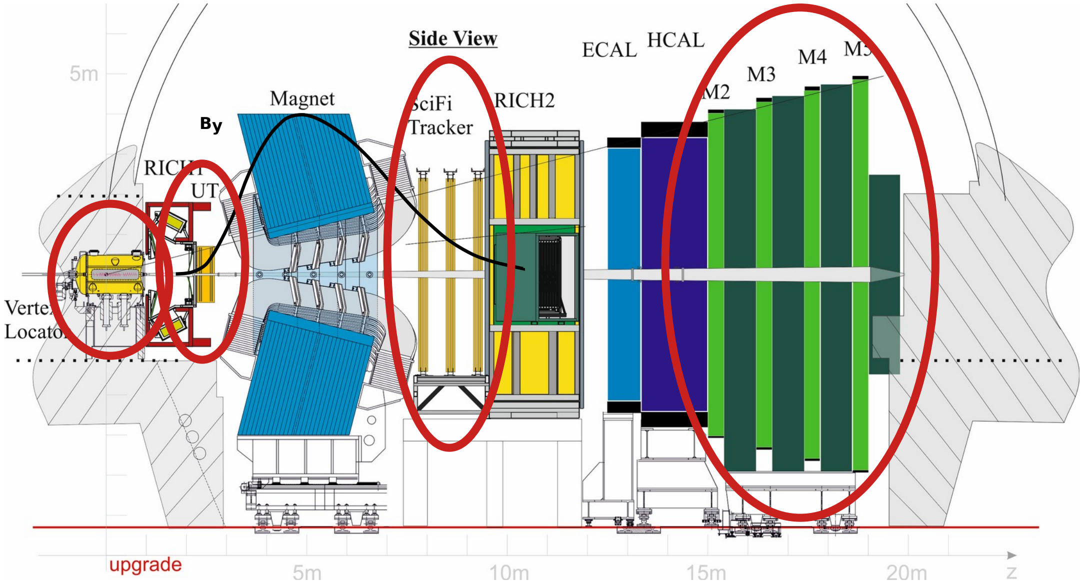
Parallel

Sequential

Consider how much of the problem can actually be parallelized!

LHCb HLT1 elements

- Decode binary payload of four sub-detectors
- Reconstruct charged particle trajectories
- Identify muons
- Reconstruct primary and secondary decay vertices
- Select pp-bunch collisions based on
  - Single-track properties
  - Secondary vertex properties



Manageable amount of algorithms with highly parallelizable tasks

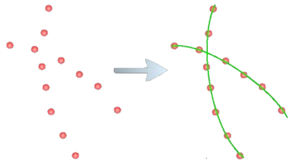
Common parallelization techniques

Raw data decoding

- Transform binary payload from subdetector raw banks into collections of hits (x,y,z) in LHCb coordinate system
- Parallelize over all subdetectors and readout units

Track reconstruction

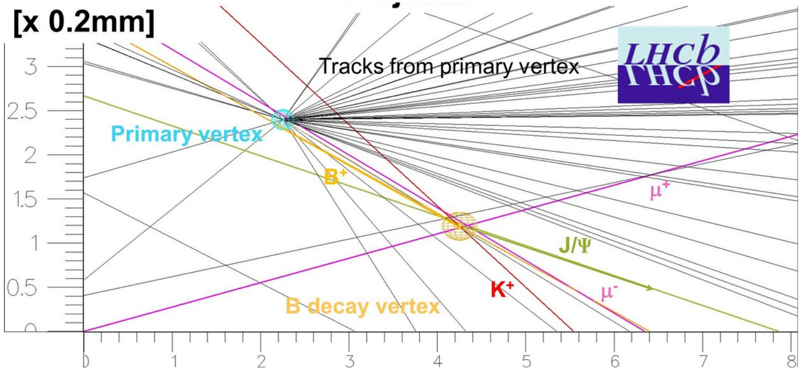
- Consists of two steps:
    - Pattern recognition: Which hits belong to which track?
    - Track fitting: Done for every track
- Parallelize over combinations of hits and tracks



f(x) = … +/- …

Vertex finding

- Reconstruct primary and secondary vertices
- Parallelize across combinations of tracks and vertex seeds



How does HLT1 map to GPUs?

Characteristics of LHCb HLT1	**Characteristics of GPUs**
Intrinsically parallel problem: - Run events in parallel - Reconstruct tracks in parallel	Good for - Data-intensive parallelizable applications - High throughput applications

How does HLT1 map to GPUs?

Characteristics of LHCb HLT1	**Characteristics of GPUs**
Intrinsically parallel problem: - Run events in parallel - Reconstruct tracks in parallel	Good for - Data-intensive parallelizable applications - High throughput applications
Huge compute load	Many TFLOPS

How does HLT1 map to GPUs?

Characteristics of LHCb HLT1	**Characteristics of GPUs**
Intrinsically parallel problem: - Run events in parallel - Reconstruct tracks in parallel	Good for - Data-intensive parallelizable applications - High throughput applications
Huge compute load	Many TFLOPS
Full data stream from all detectors is read out → no stringent latency requirements	Higher latency than CPUs, not as predictable as FPGAs

How does HLT1 map to GPUs?

Characteristics of LHCb HLT1	**Characteristics of GPUs**
Intrinsically parallel problem: - Run events in parallel - Reconstruct tracks in parallel	Good for - Data-intensive parallelizable applications - High throughput applications
Huge compute load	Many TFLOPS
Full data stream from all detectors is read out → no stringent latency requirements	Higher latency than CPUs, not as predictable as FPGAs
Small raw event data (~100 kB)	Connection via PCIe → limited I/O bandwidth

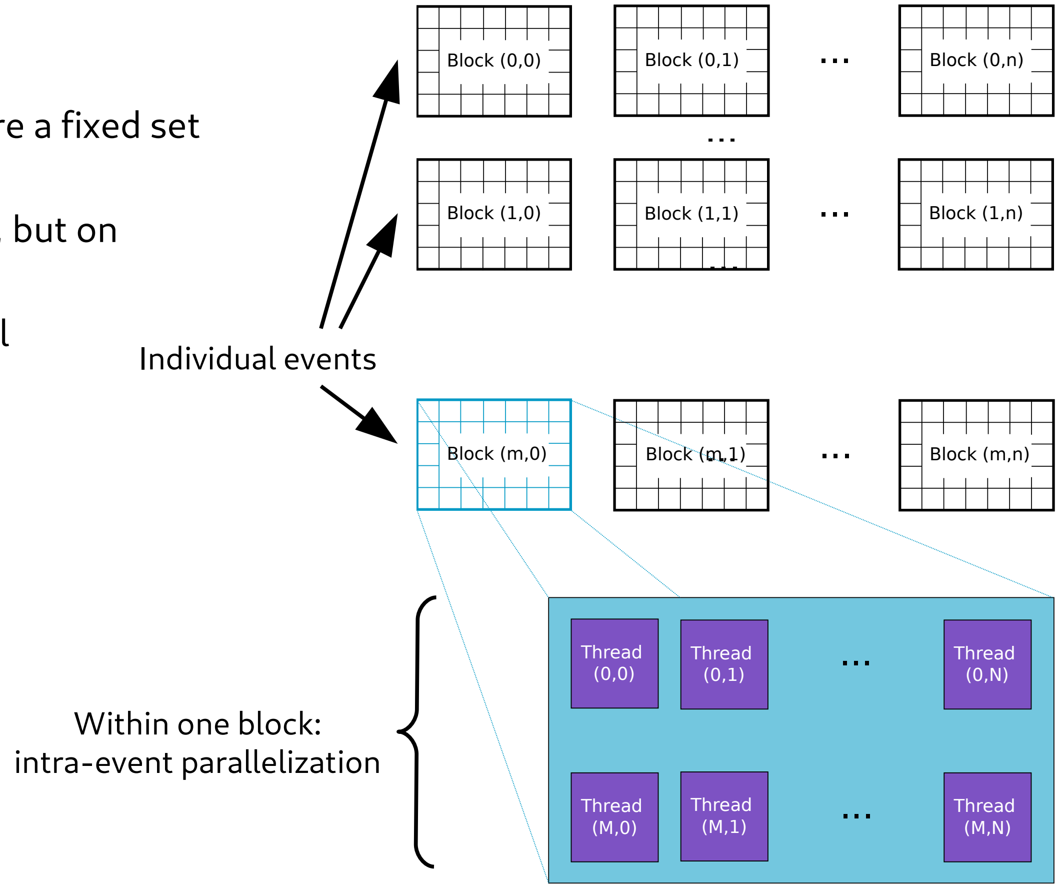
How does HLT1 map to GPUs?

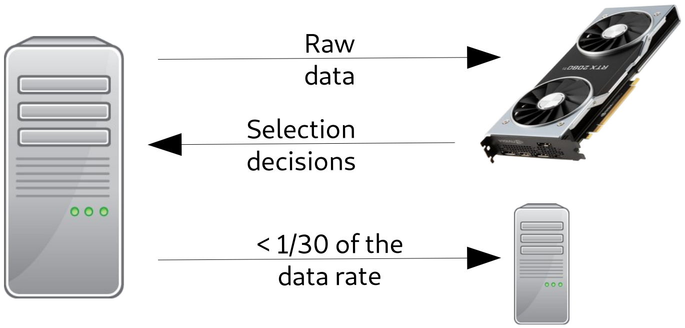
Characteristics of LHCb HLT1	**Characteristics of GPUs**
Intrinsically parallel problem: - Run events in parallel - Reconstruct tracks in parallel	Good for - Data-intensive parallelizable applications - High throughput applications
Huge compute load	Many TFLOPS
Full data stream from all detectors is read out → no stringent latency requirements	Higher latency than CPUs, not as predictable as FPGAs
Small raw event data (~100 kB)	Connection via PCIe → limited I/O bandwidth
Small event raw data (~100 kB)	Thousands of events fit into O(10) GB of memory

Perfect fit!

HLT1 on GPUs

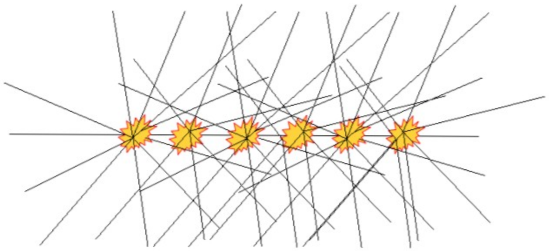
- GPU code is executed on many “threads”
- These threads are organized in a “grid”, where a fixed set of threads is grouped into one “block”
- Each thread processes the same instructions, but on different data
- Thousands of events are processed in parallel
- In addition: intra-event parallelization

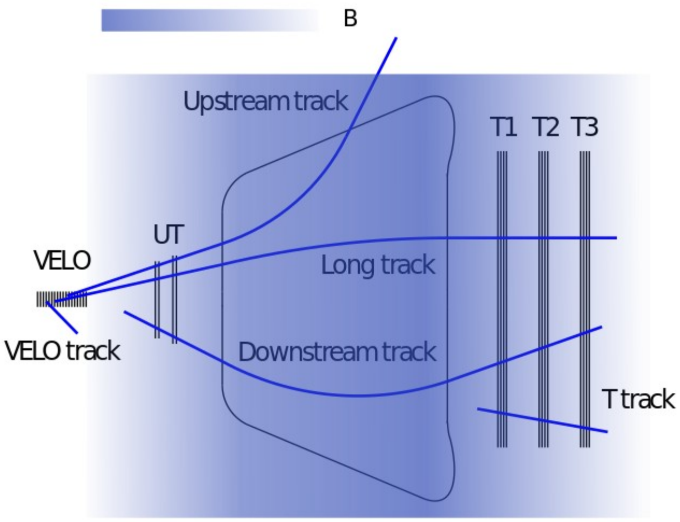




LHCb: Characteristics for pattern recognition

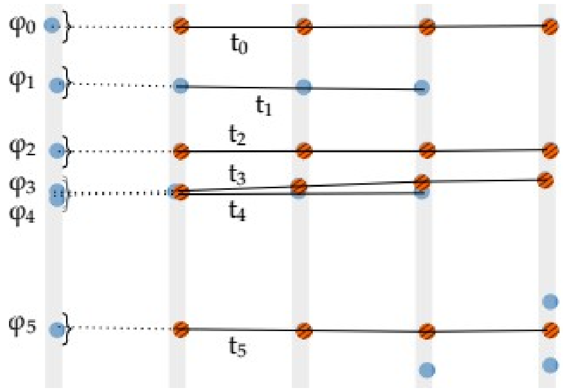
- Average pile up of 6
- Few hundred - few thousand hits in subdetectors
- Tens to hundreds of tracks in subdetectors
- Velo tracks are input for:
  - Primary vertex finding
  - Track forwarding to other detectors
- Mainly straight line tracks
- Large bend between UT and SciFi detectors
- Most tracks have p_T < 2 GeV/c



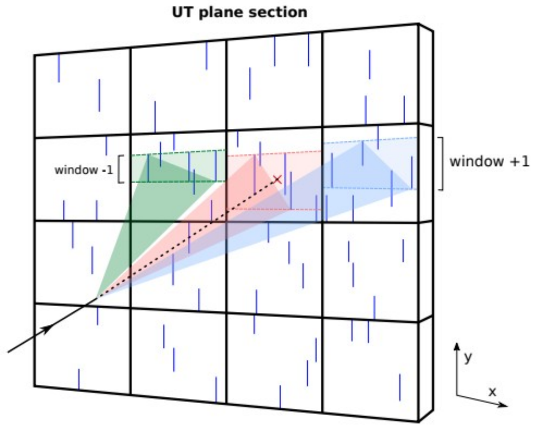


Parallelization of reconstruction tasks

Search for combinations
of hits in parallel

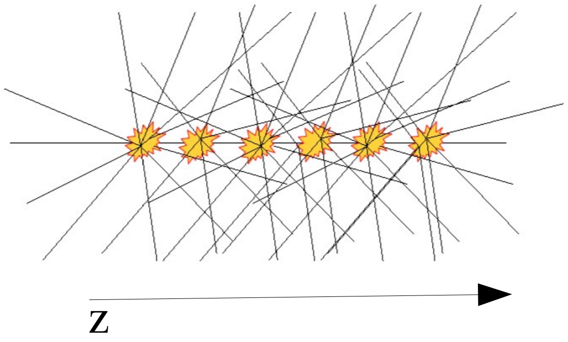


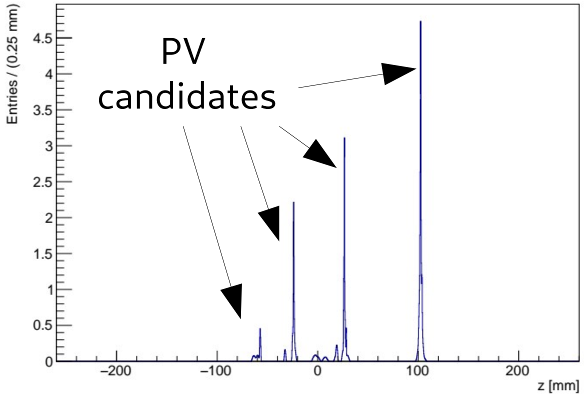
Store objects (for example hits)
In best suited memory layout



Split problem into
independent tasks

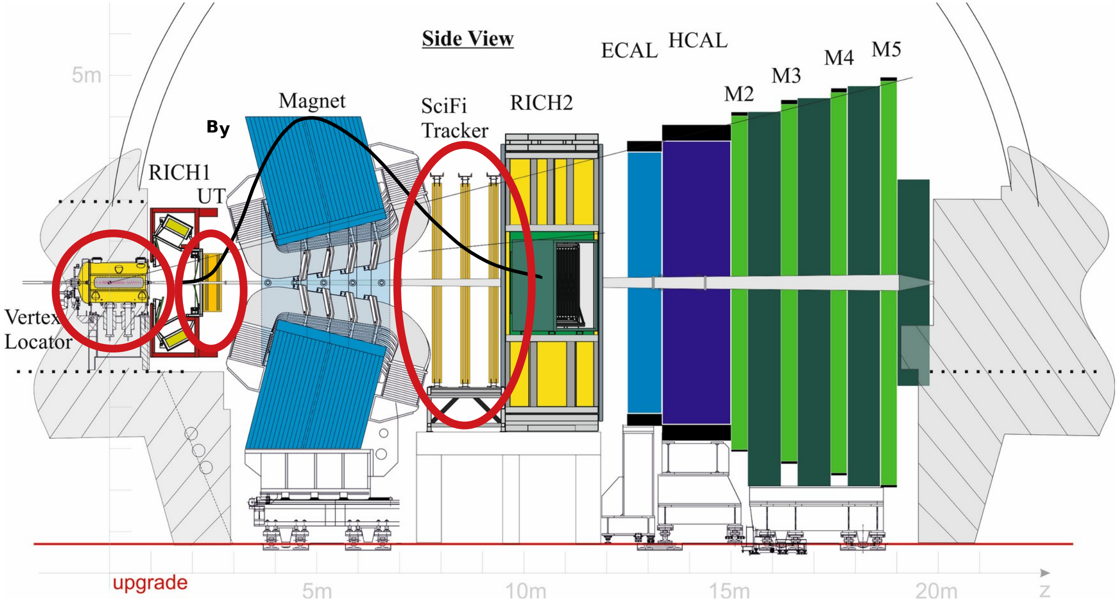
Example: primary vertex
(PV) reconstruction



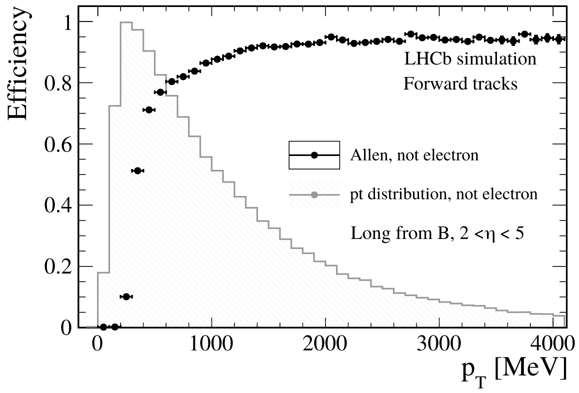


Track reconstruction performance

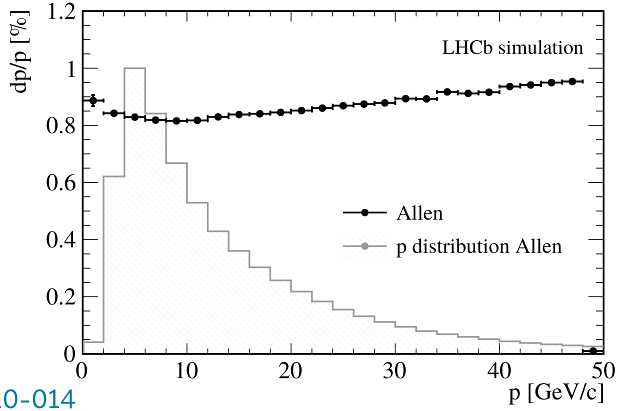
Tracks reconstructed in the Velo, UT & SciFi detectors



Track reconstruction efficiency for tracks originating from B decays



Momentum resolution



LHCb-FIGURE-2020-014

HLT1: Trigger selections

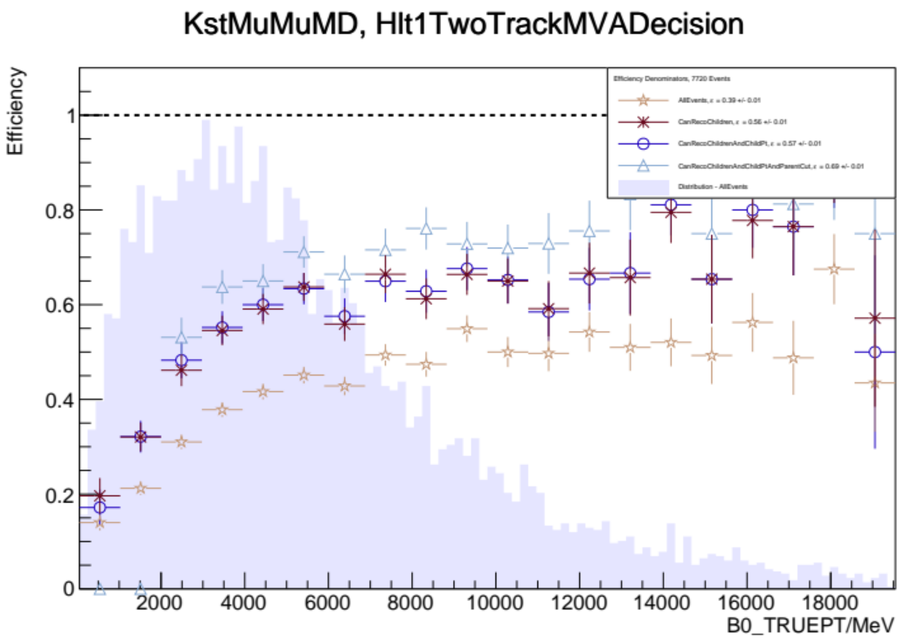
	Trigger	Rate [kHz]
Monitoring & calibration lines	ErrorEvent	0 ± 0
	PassThrough	30000 ± 0
	NoBeams	5 ± 3
	BeamOne	18 ± 5
	BeamTwo	8 ± 3
	BothBeams	4 ± 2
	ODINNoBias	0 ± 0
	ODINLumi	1 ± 1
	GECPassthrough	27822 ± 52
	VeloMicroBias	26 ± 6
Physics selections	TrackMVA	409 ± 23
	TrackMuonMVA	23 ± 6
	SingleHighPtMuon	7 ± 3
	TwoTrackMVA	503 ± 26
	DiMuonHighMass	131 ± 13
	DiMuonLowMass	177 ± 15
	DiMuonSoft	8 ± 3
	D2KPi	93 ± 11
	D2PiPi	34 ± 7
	D2KK	76 ± 10
	Total w/o pass through lines	1157 ± 39

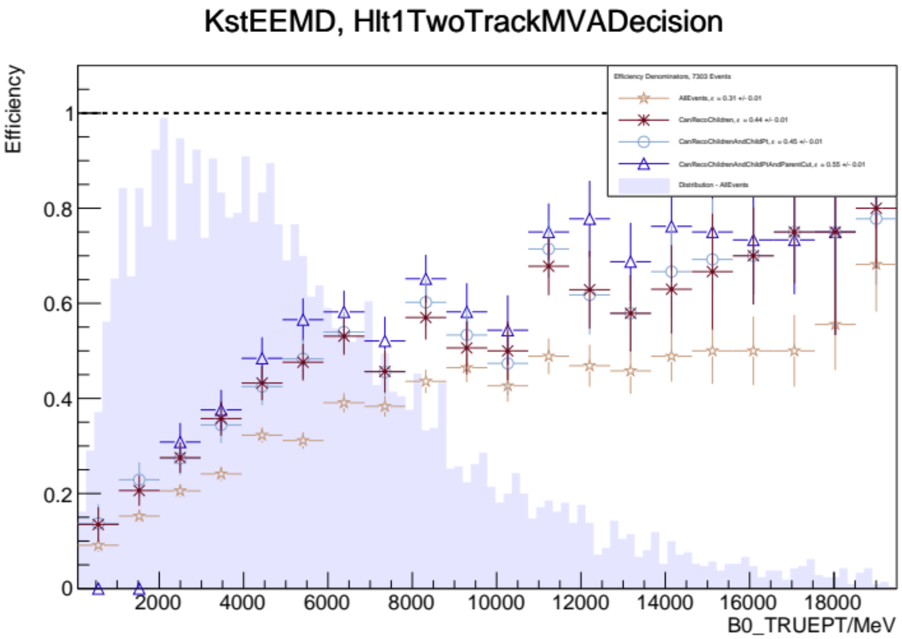
Event rate reduced by factor 30

Alignment

LHCb-FIGURE-2020-014

HLT1: Selection efficiencies



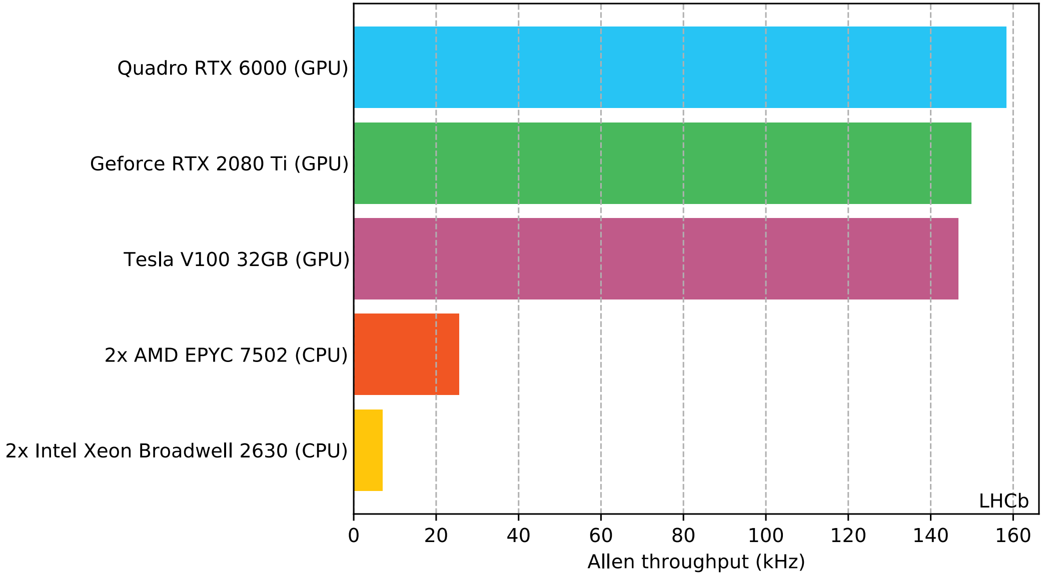


CERN-LHCC-2020-006

Selection efficiencies for electron and muon final states similar

In Run 2: Electron selection efficiency roughly factor two worse than muons due to hardware level trigger

Computing performance



LHCb-FIGURE-2020-014

- Require about 215 GPU cards to process full HLT1 @ 30 MHz
- Have slots for 500 cards
- Significant throughput increase on latest Nvidia GPUs (RTX 3080, RTX 3090)

The Allen project

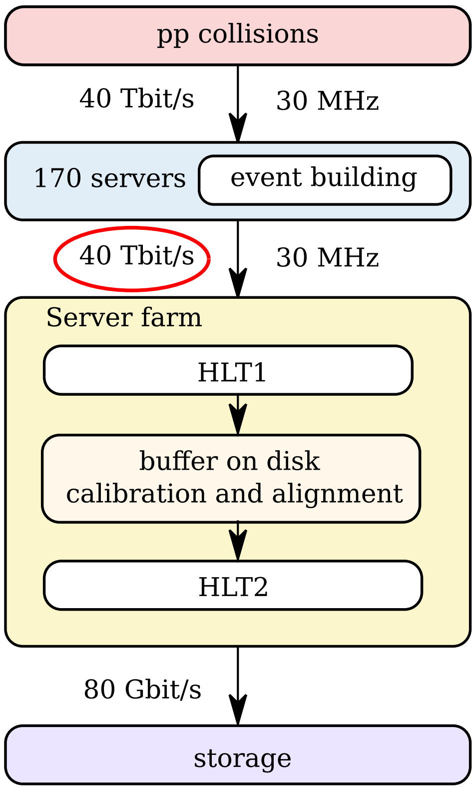
- Fully standalone software project: https://gitlab.cern.ch/lhcb/Allen
- Framework developed for processing HLT1 on GPUs
- Runs on CPUs, Nvidia GPUs, AMD GPUs
- GPU code written in CUDA
- Cross-architecture compatibility via macros (ROCm for AMD, c++ for CPUs)
- Configuration via python
- Memory manager for GPU memory

- Named after Frances E. Allen

History: HLT1 architecture choice

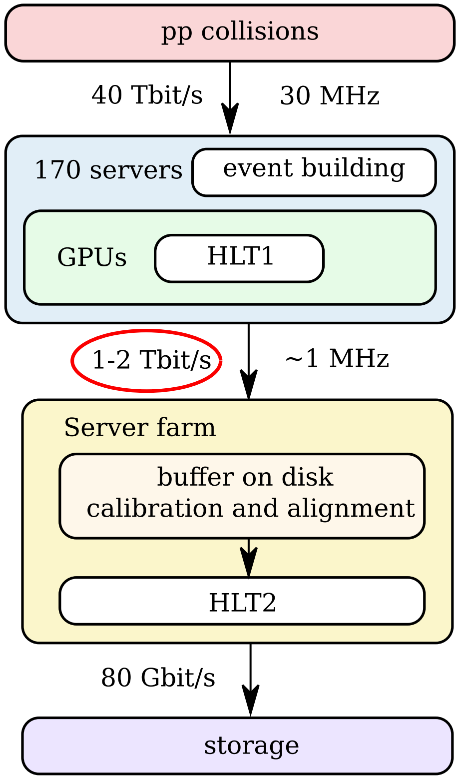
Proposal in TDR (2014)

CERN-LHCC-2014-016



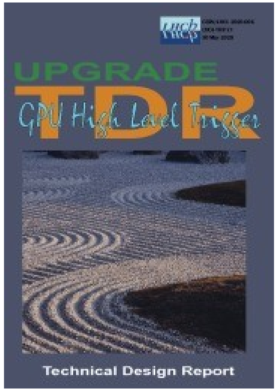
Updated strategy (as of 5/2020)

CERN-LHCC-2020-006

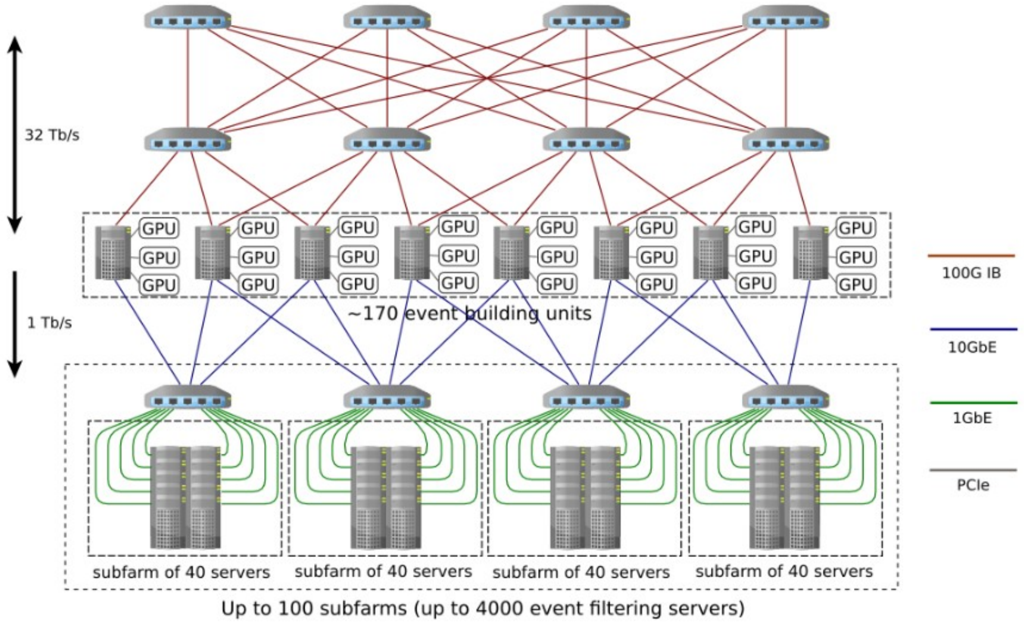


- Developed two solutions simultaneously
- Both the multi-threaded CPU & the GPU HLT1 fulfilled the requirements from the 2014 TDR
- LHCb was in the luxury situation to choose among them
- Compared physics performance & price-performance

  → decided for GPU solution



Towards commissioning



- Communication with event builder network
- Final data formats of sub-detector raw data
- Monitoring: histograms, counters
- As sub-detectors are commissioned, run algorithms on first data
  - Cosmic tracks
  - Calorimeter clusters (sources)

CPPM LHCb group involved

Integration test with event building server

Impact on event building when running HLT1 on GPUs inside the event building servers?

Monitoring temperatures, memory bandwidths, processing rate, ...
Tested in production server candidate in October 2019
→ To be repeated this year with latest hardware

Summary

- LHCb plays major role in studying the flavor anomalies in B decays
- Combination of various $b \to cl\nu$ measurements crucial to uncover possible New Physics
- A measurement with all three lepton species $R(D^*)_{\tau e \mu}$ will be possible at LHCb in Run 3

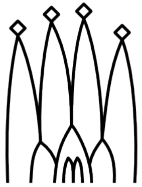
- Upgrade I for Run 3 basically turns LHCb into a new experiment
- Need software-only real-time selection @ 40 Tbit/s to exploit full physics potential
- Developed first complete high-throughput GPU trigger for an HEP experiment to tackle computing challenge
- Enough computing headroom to add more complex algorithms → even higher physics gain
- Heterogeneous trigger prepares LHCb for future upgrades (400 Tbit/s in Run 5)

Backup

Allen software framework

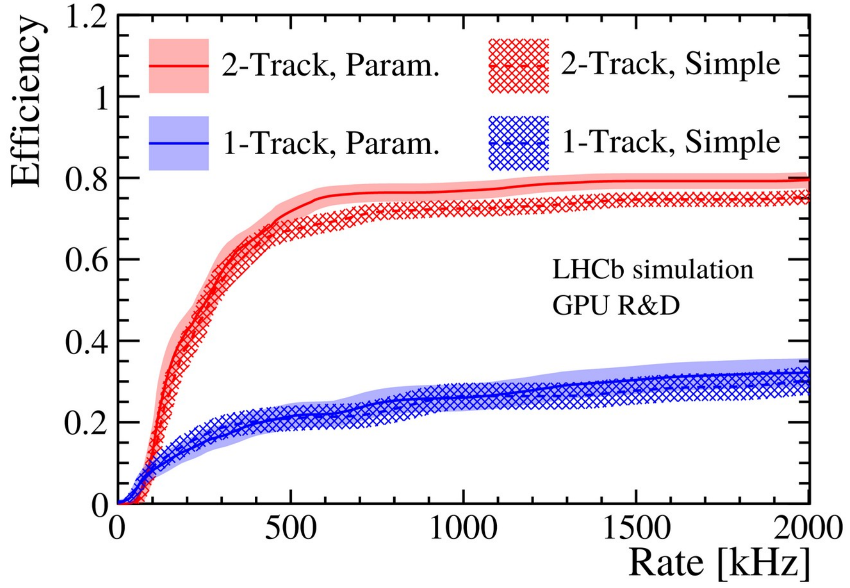
- Algorithm sequences defined in python and generated at compile time
  - Algorithms to run with inputs / outputs, properties (minimum momentum cut-off etc.)
- Memory manager:
  - Large chunk of GPU memory allocated at start-up
  - Pieces of memory assigned to algorithms by memory manager
  - Memory size has to be known at compile time
- Cross-architecture compatibility via macros & few coding guide lines
- Support three modes:
  - Standalone project
  - Compiling with Gaudi for data acquisition
  - Compiling with Gaudi for simulation workflow and offline studies

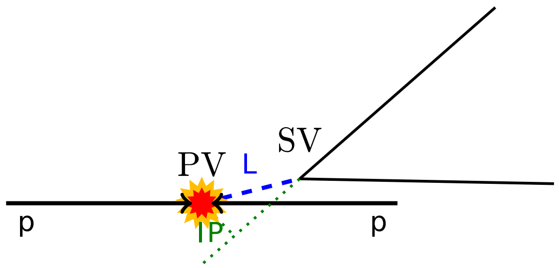




Kalman filter

Improved track description → better impact parameter resolution

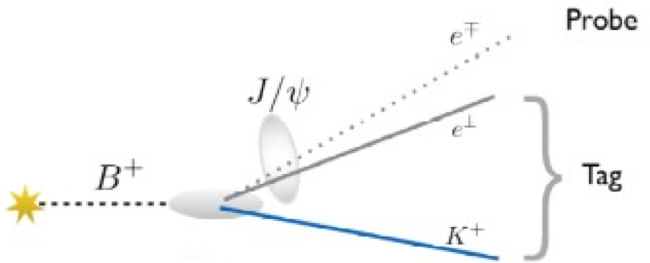




- Simple: Simplified Kalman filter with constant momentum assumption
- Param.: Parameterized Kalman filter with momentum estimate from SciFi track reconstruction

Systematic uncertainties in the future

- Main systematics:
  - Limited size of simulated samples
  - Modeling of fit components: Dedicated control samples & simulated samples
  - Data-driven method to obtain electron reconstruction efficiency from dedicated control samples
- Large statistics simulation is a software challenge, not part of this talk
- Larger data sets will improve precision of components from dedicated control samples



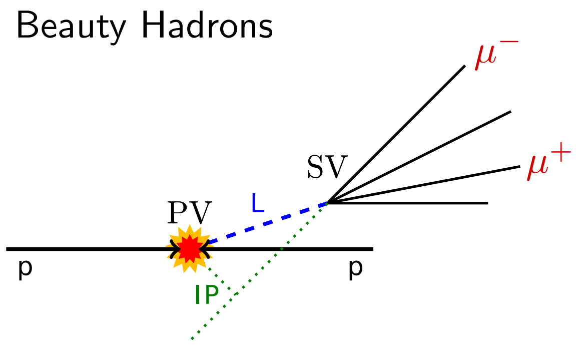
One electron is only reconstructed in the vertex detector

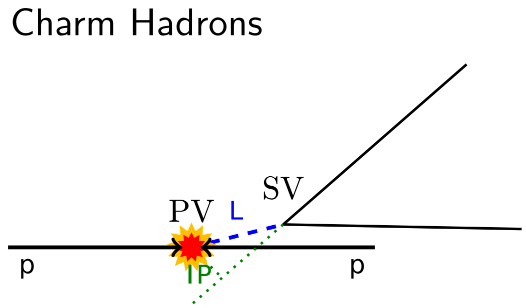
One electron is fully reconstructed and paired with a kaon

Background is suppressed through B and J/ψ mass constraints
This also provides the probe electron's momentum

After upgrade I (Run 3), trigger efficiencies for electrons will significantly improve (see second part of the talk)

Beauty and charm decays





- $B^{\pm/0}$ mass ~5.3 GeV

  → Daughter p_T $\mathcal{O}$(1 GeV)
- τ ~1.6 ps → flight distance ~1cm
- Detached muons from B→J/ΨX, J/Ψ → $\mu^+\mu^-$
- Displaced tracks with high p_T

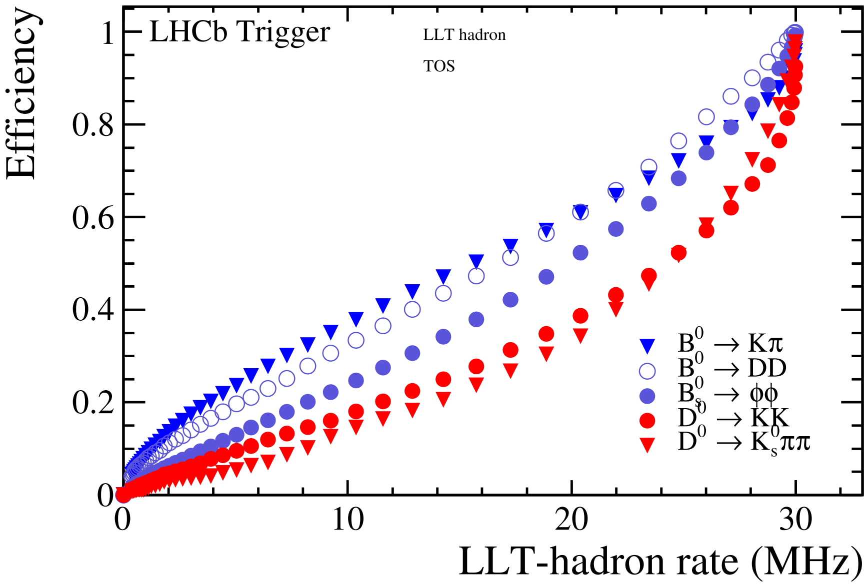
- $D^{\pm/0}$ mass ~1.9 GeV

  → Daughter p_T $\mathcal{O}$(700 MeV)
- τ ~0.4 ps → flight distance ~4mm
- Also produced from B decays

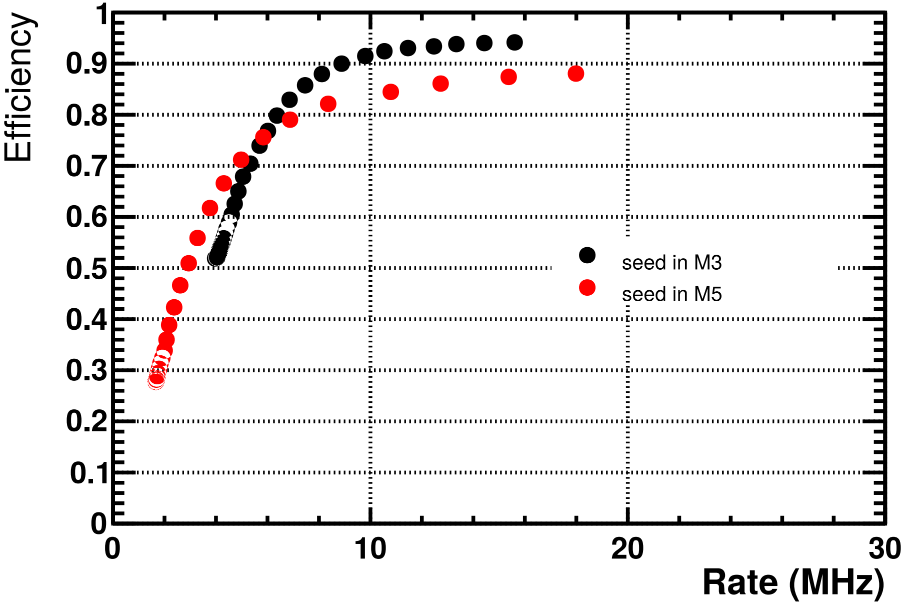
PV: Primary vertex
SV: Secondary vertex
IP: Impact parameter: distance between point of closest approach of a track and a PV

Why no low level trigger?

Low level trigger on E_T from the calorimeter



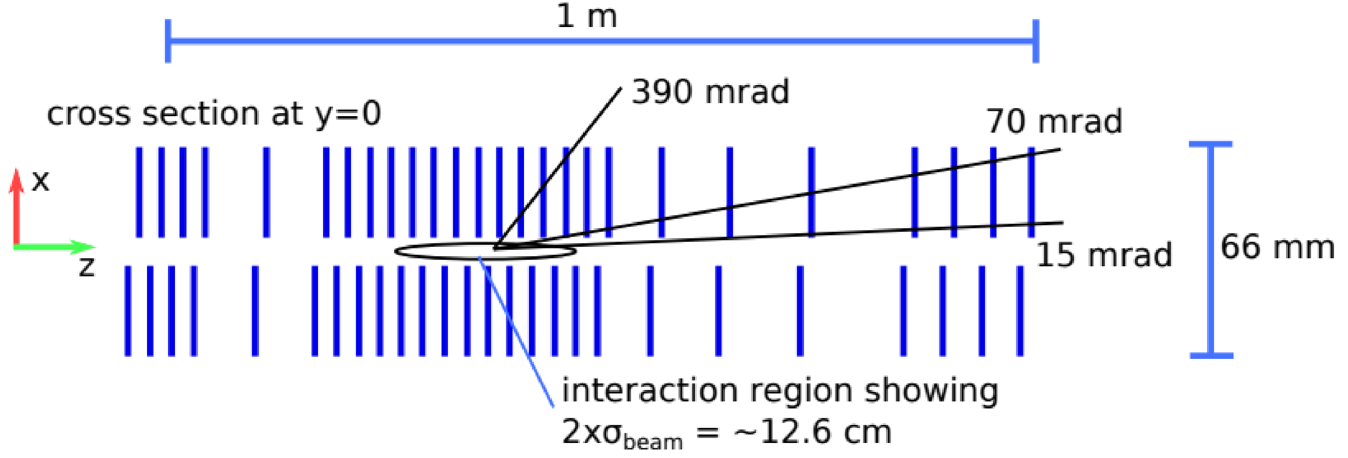
Low level trigger on muon p_T, $B \to K^*\mu\mu$



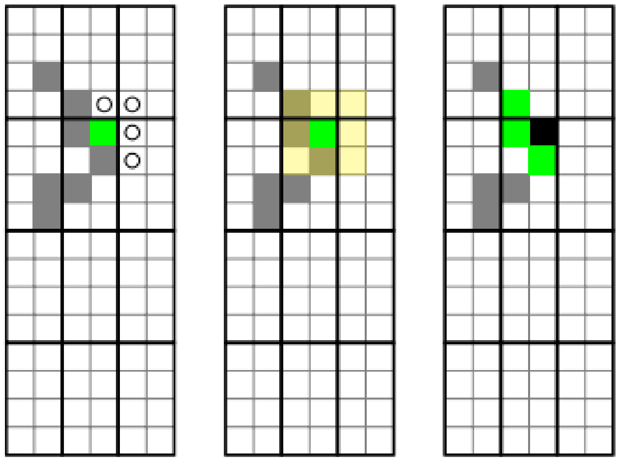
Need track reconstruction at first trigger stage

Velo detector: clustering

26 planes of silicon pixel detectors

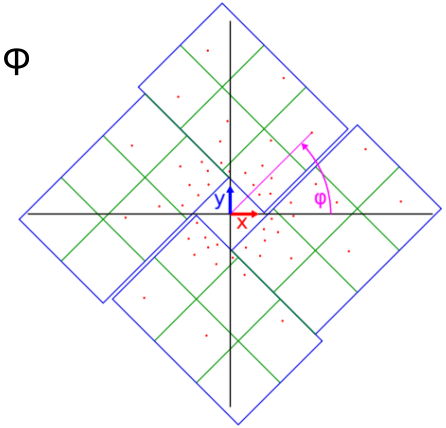


Clustering with bit masks

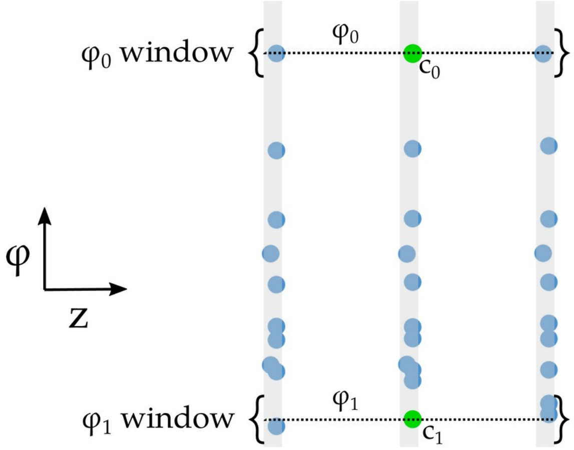


Velo detector: track reconstruction

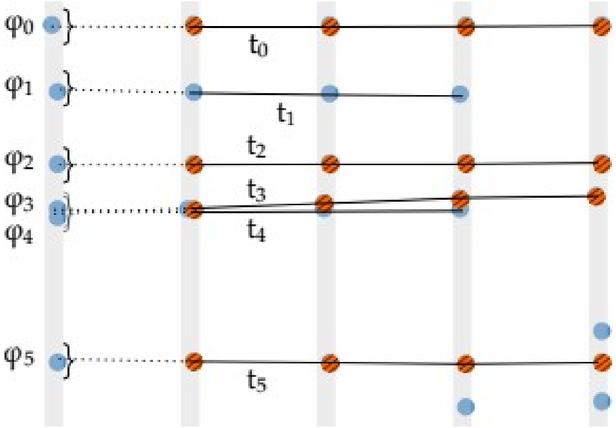
1) Sort hits by φ



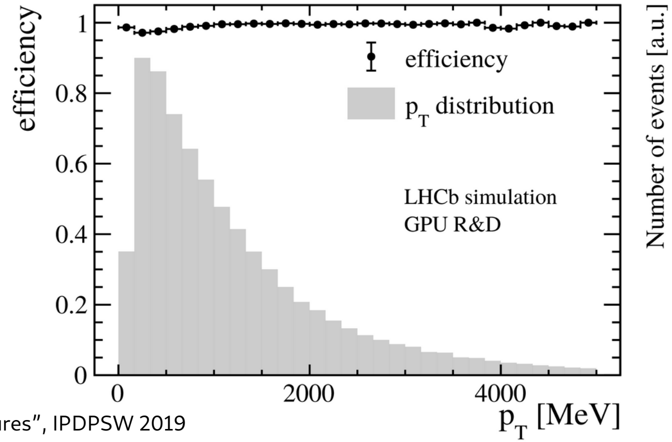
2) Triplet seeding



3) Triplet forwarding

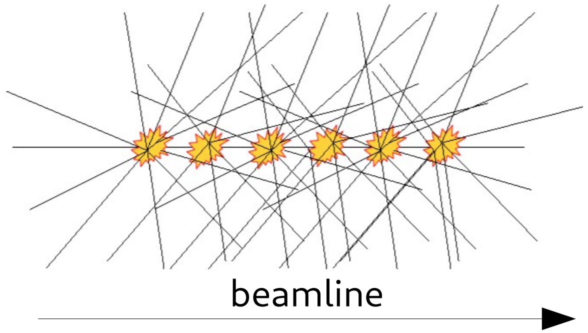


Track reconstruction efficiency for tracks originating from B decays

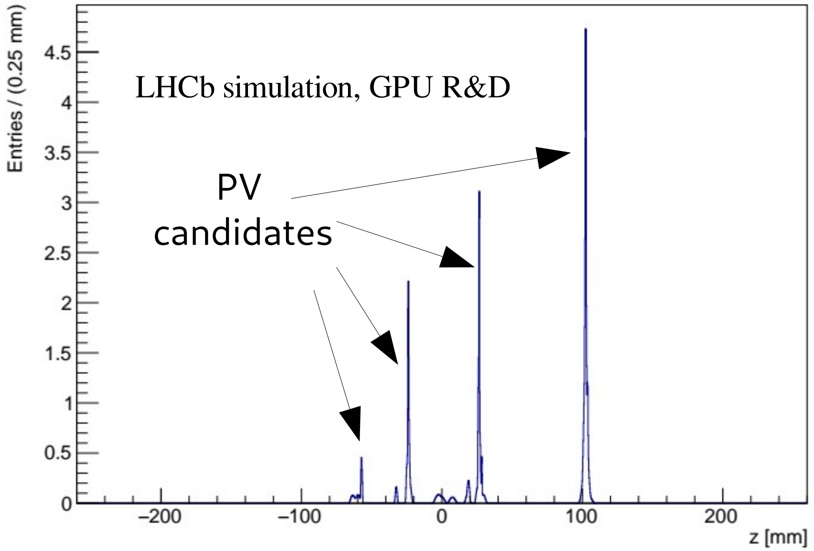


D. Campora, N. Neufeld, A. Riscos Núñez: "A fast local algorithm for track reconstruction on parallel architectures", IPDPSW 2019

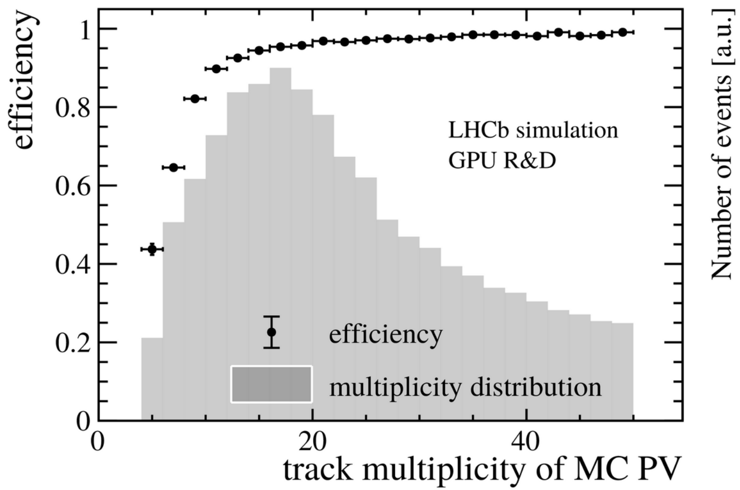
Velo detector: primary vertex reconstruction



Point of closest approach of tracks to beamline

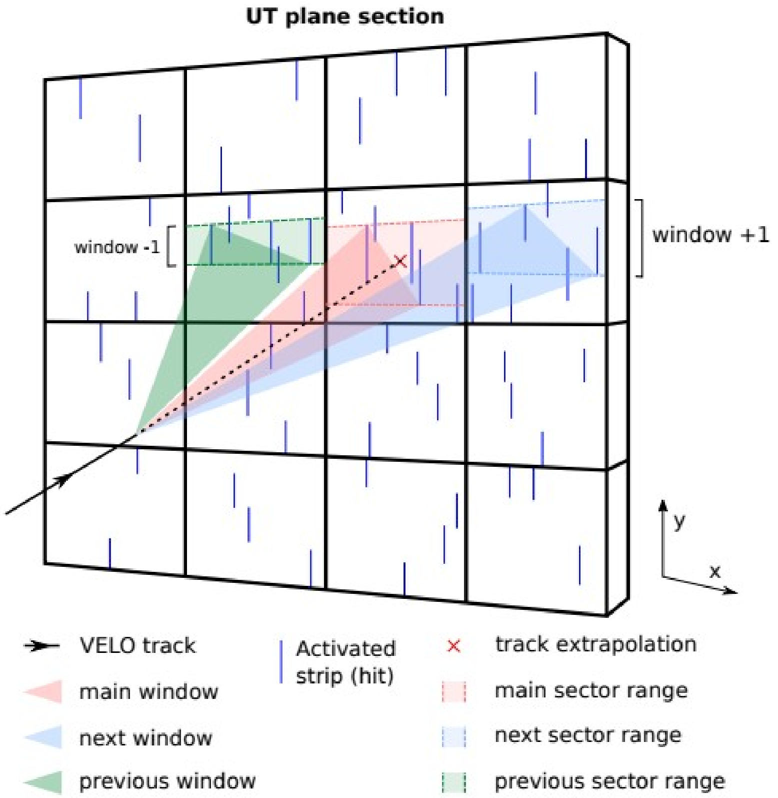


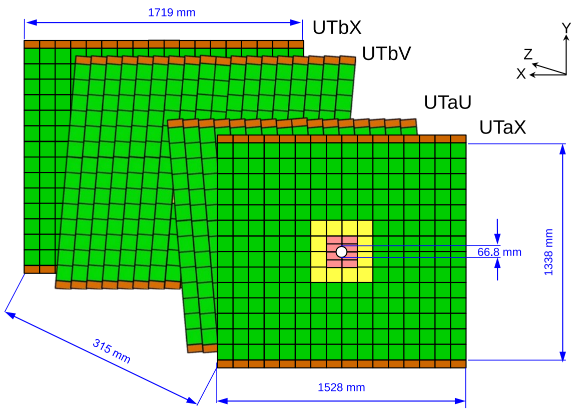
PV reconstruction efficiency



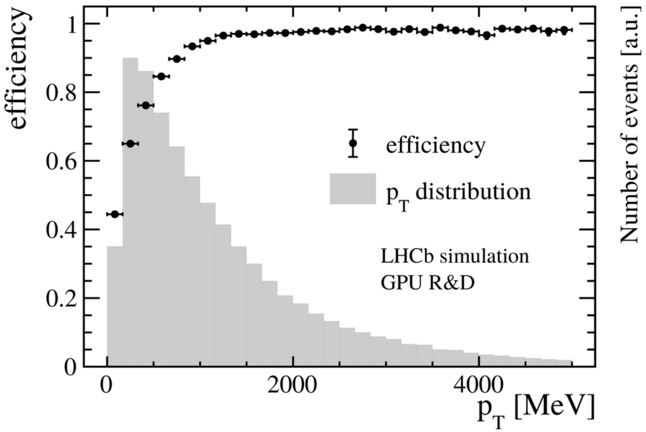
UT detector: track reconstruction

4 planes of silicon strip detectors





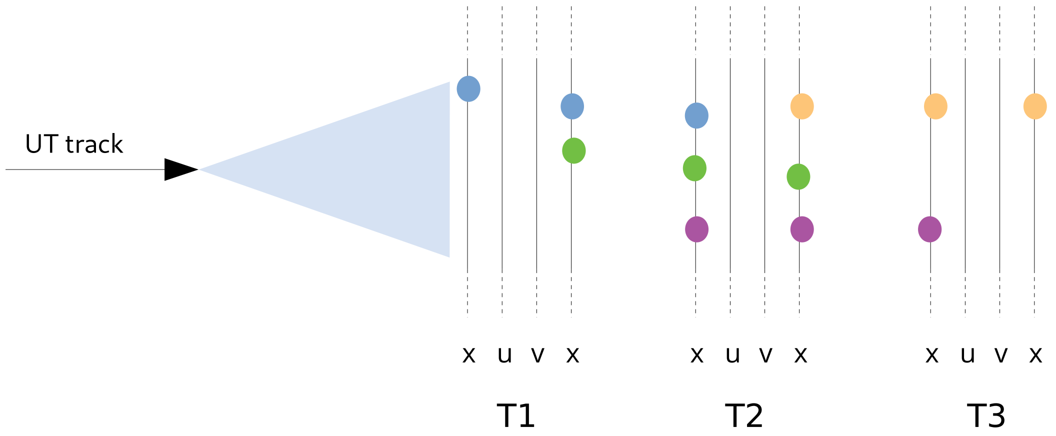
Track reconstruction efficiency for tracks originating from B decays

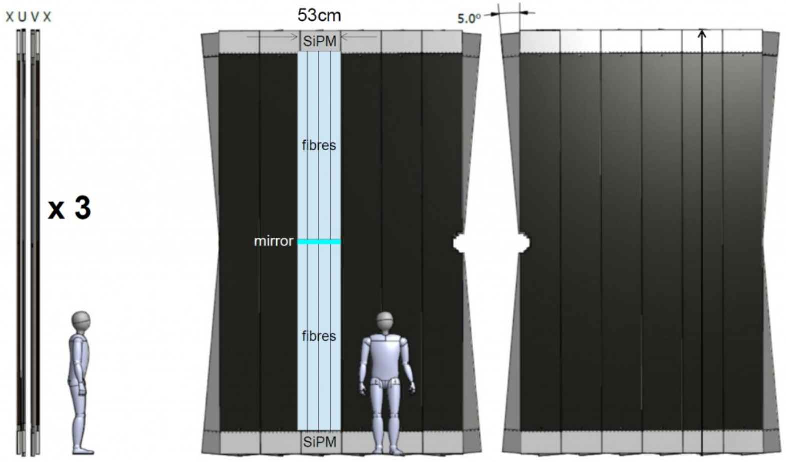


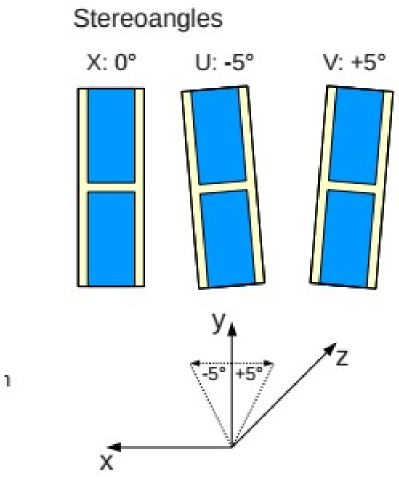
P. Fernandez Declara, D. Campora Perez, J. Garcia-Blas, D. vom Bruch, J. Daniel Garca, N. Neufeld , IEEE Access 7 (2019)

SciFi detector

12 layers of scintillating fibres
Efficiency of fibres ~ 98-99%

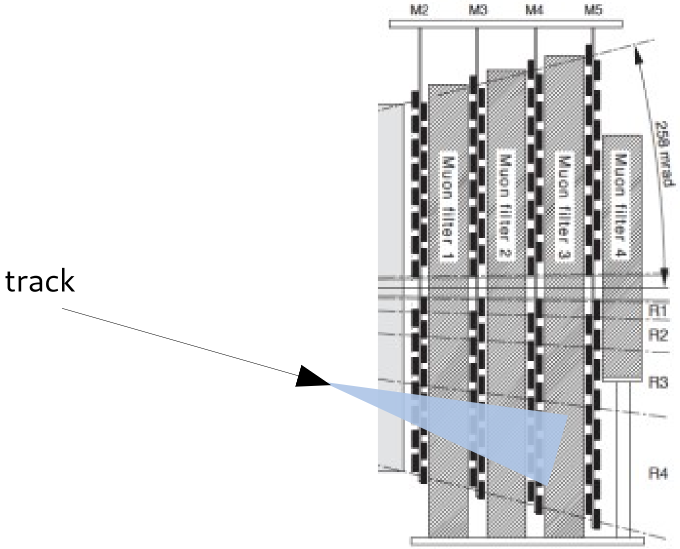




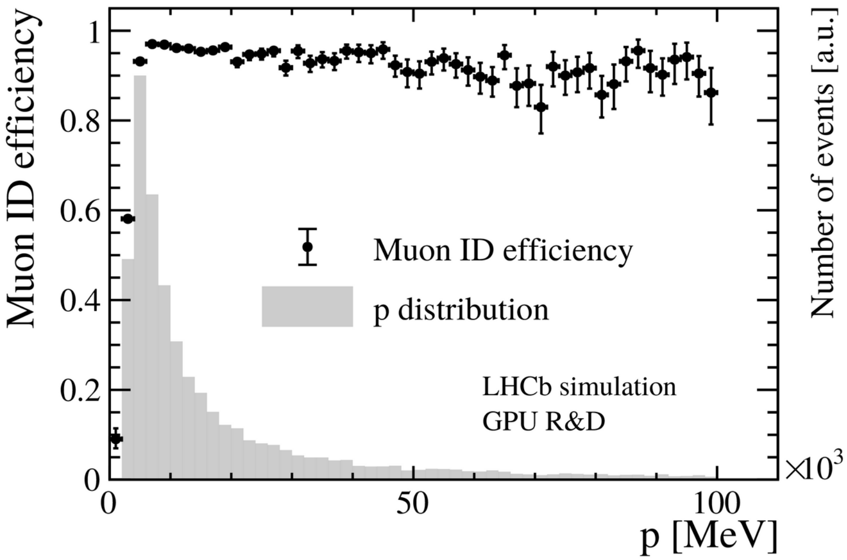


Muon identification

Four multi-wire proportional chambers
Interleaved with iron walls



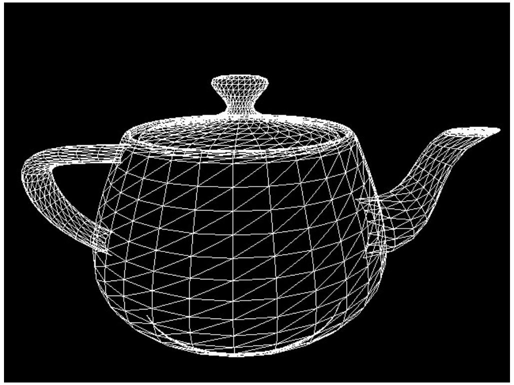
Muon identification efficiency



Graphics requirements

Graphics pipeline

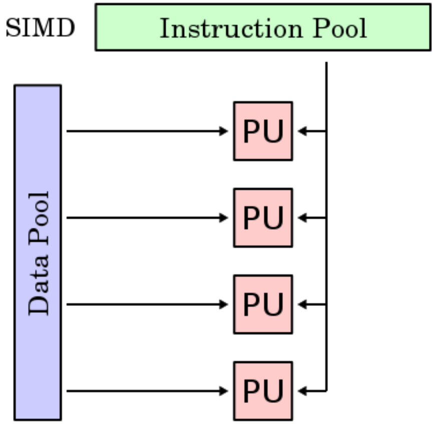
- Huge amount of arithmetic on independent data:
    - Transforming positions
    - Generating pixel colors
    - Applying material properties and light situation to every pixel



Hardware needs

- Access memory simultaneously and contiguously
- Bandwidth more important than latency
- Floating point and fixed-function logic

→ Single instruction applied to multiple data: SIMT



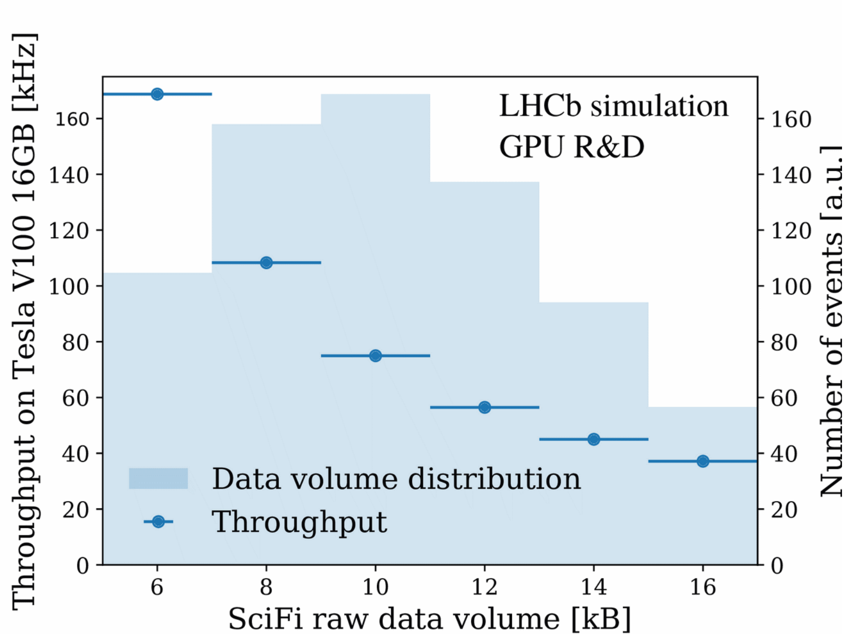
Selections

Selection name	Criteria
1-Track	Single displaced track with high p_T
2-Track	Two-track vertex with significant displacement and p_T
High-p_T muon	Single muon with high p_T
Displaced diumuon	Displaced di-muon vertex
High-mass dimuon	Di-muon vertex with mass near or larger than the J/Ψ

Criteria applied to signal decays in efficiency calculations

b and c hadrons	$p_{\mathrm{T}} > 2$ GeV $\tau > 0.2$ ps
b and c hadron children	$p_{\mathrm{T}} > 200$ MeV $2 < \eta < 5$ reconstructible in the Velo and SciFi detector (long track)
Z children	$p_{\mathrm{T}} > 20$ GeV $2 < \eta < 5$ reconstructible in the Velo and SciFi detector (long track)

Throughput versus occupancy



- Data volume proportional to occupancy
- Low performance decrease at high occupancy

  → will be able to handle real data (likely higher in occupancy than simulation)

GPUs for throughput measurement

CUDA streams

Allen settings	Threads (-t)	Memory (-m)	Number of events (-n)	Repetitions (-r)
High	12	700	1000	100
Low	2	700	1000	100

Card	# cores	Max freq. (GHz)	Cache (MiB, L2)	DRAM (GiB)	DRAM type	CUDA cap.	Allen settings
Geforce GTX 670	1344	1.06	0.5	1.95	GDDR5	3.0	Low
Geforce GTX 680	1536	1.14	0.5	1.95	GDDR5	3.0	Low
Geforce GTX 780 Ti	2880	0.93	1.5	2.95	GDDR5	3.5	Low
Geforce GTX 980	2048	1.29	2	2.01	GDDR5	5.2	Low
Geforce GTX TITAN X	3072	1.08	3	11.92	GDDR5	5.2	High
Geforce GTX 1060 6G	1280	1.81	1.5	5.94	GDDR5	6.1	Low
Geforce GTX 1080 Ti	3584	1.67	2.75	10.92	GDDR5	6.1	High
Geforce RTX 2080 Ti	4352	1.545	6	10.92	GDDR5	7.5	High
Tesla T4	2560	1.59	4	15.72	GDDR6	7.5	High
Tesla V100 32GB	5120	1.37	6	32	HBM2	7.0	High