Allen: A High Level Trigger on GPUs for LHCb

Dorothea vom Bruch

LPNHE, CNRS

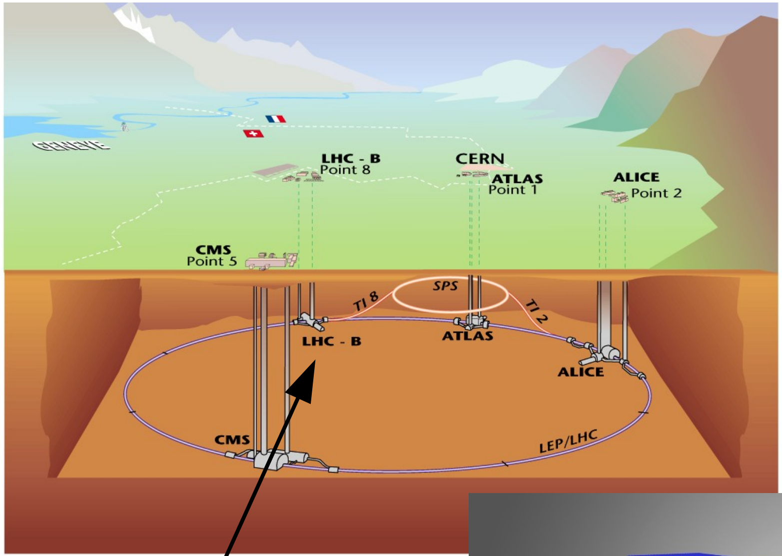
Sorbonne University, Paris Diderot University

January 27th 2020

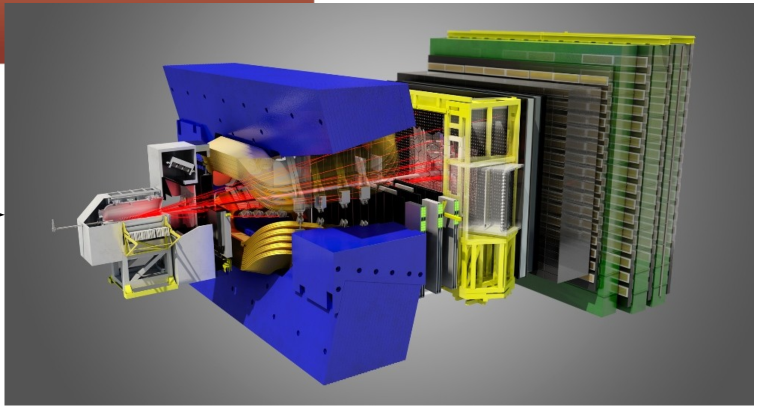
LPNHE seminar

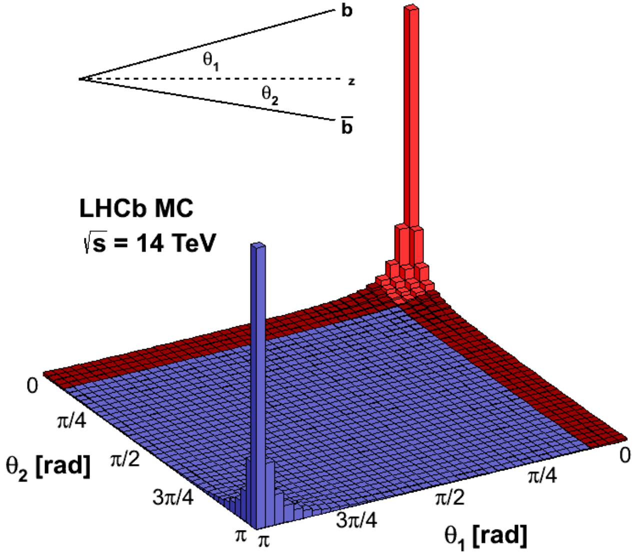
LHCb

LHC @ CERN



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

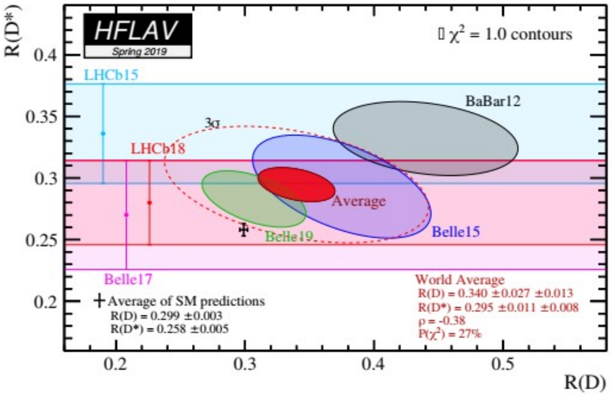




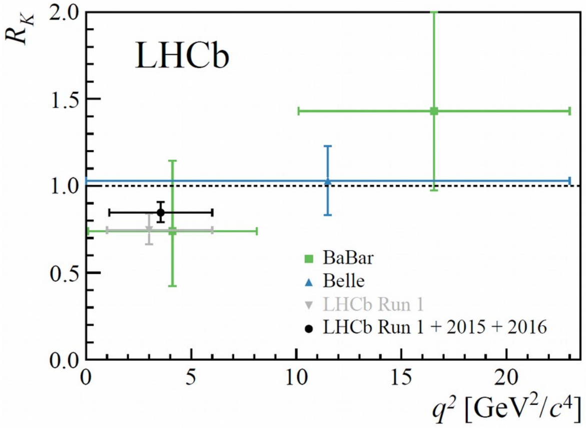
Highlights from Runs 1 & 2

Lepton flavor universality

b → clν

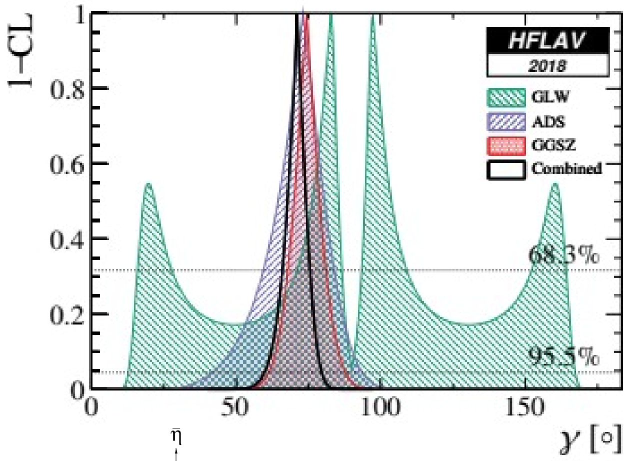


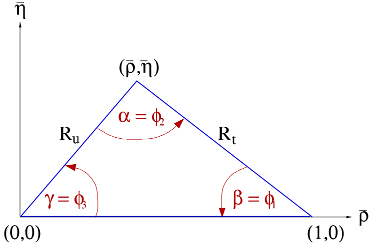
b → sl⁺l⁻



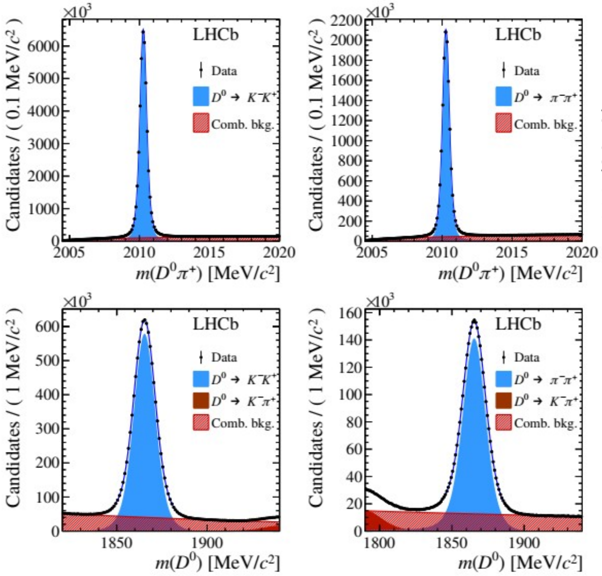
T. Humair, Moriond 2019

Constraining CKM angles



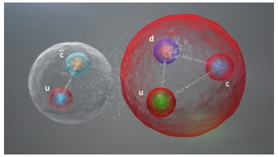


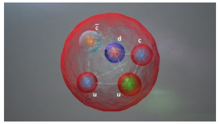
CP violation in charm decays



Phys. Rev. Lett. 122, 211803 (2019)

Pentaquarks

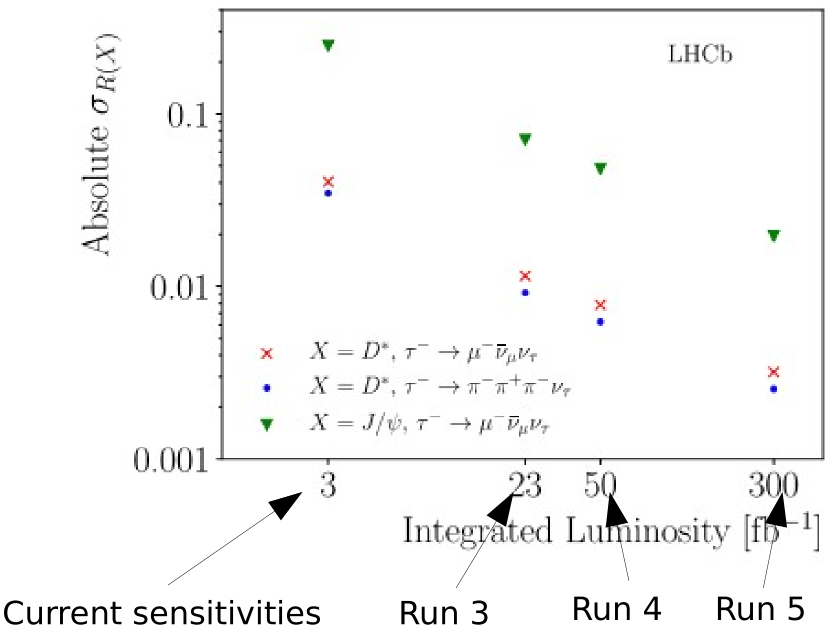




Prospects for Run 3 (2021)

- Crucial to reduce uncertainties
  → manifestation of new physics?
- R(D*) is theoretically clean
  → reduction of statistical uncertainty necessary
- Runs 3 and beyond will shed light on the flavour anomalies currently observed

Projected absolute uncertainties on R(D*)



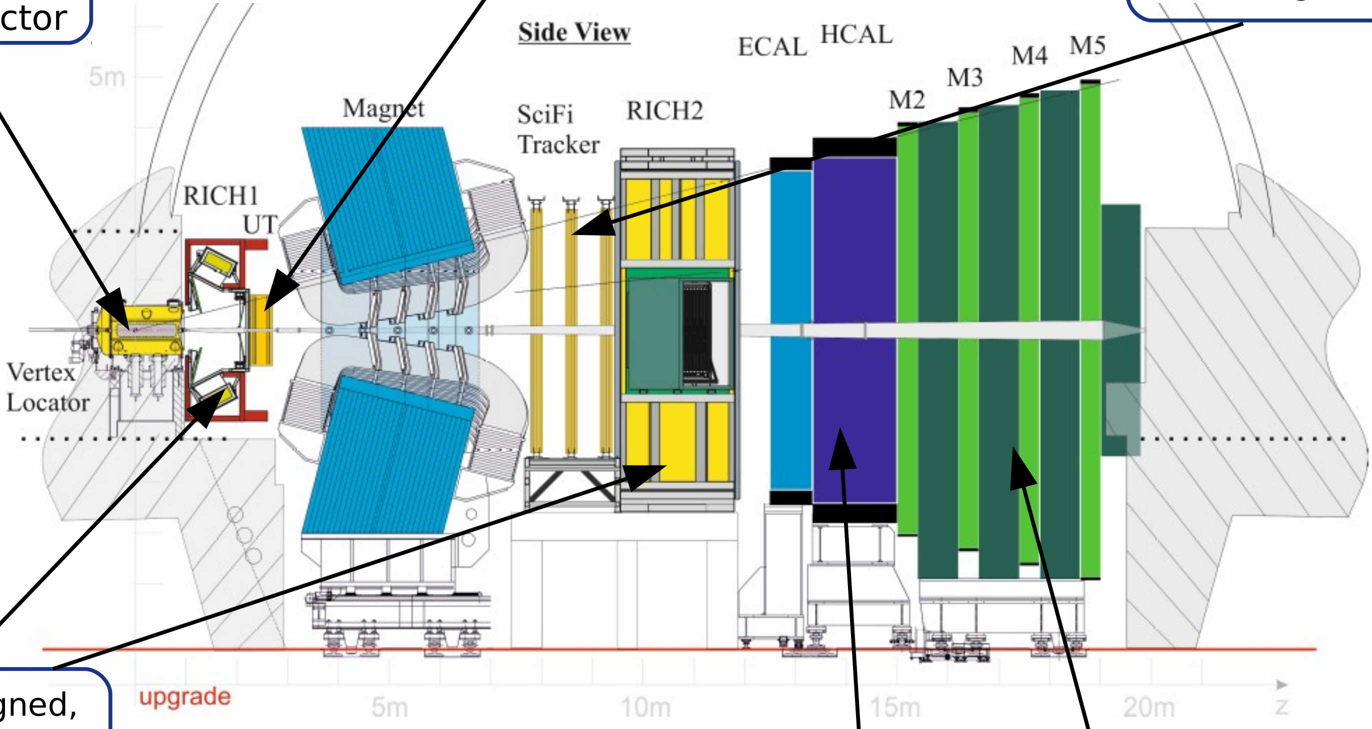
LHCb-PUB-2018-009, CERN-LHCC-2018-027

LHCb upgrade for Run 3 (2021)

New Vertex locator (Velo) tracking detector

New UT tracking detector

New scintillating fibre (SciFi) tracking detector

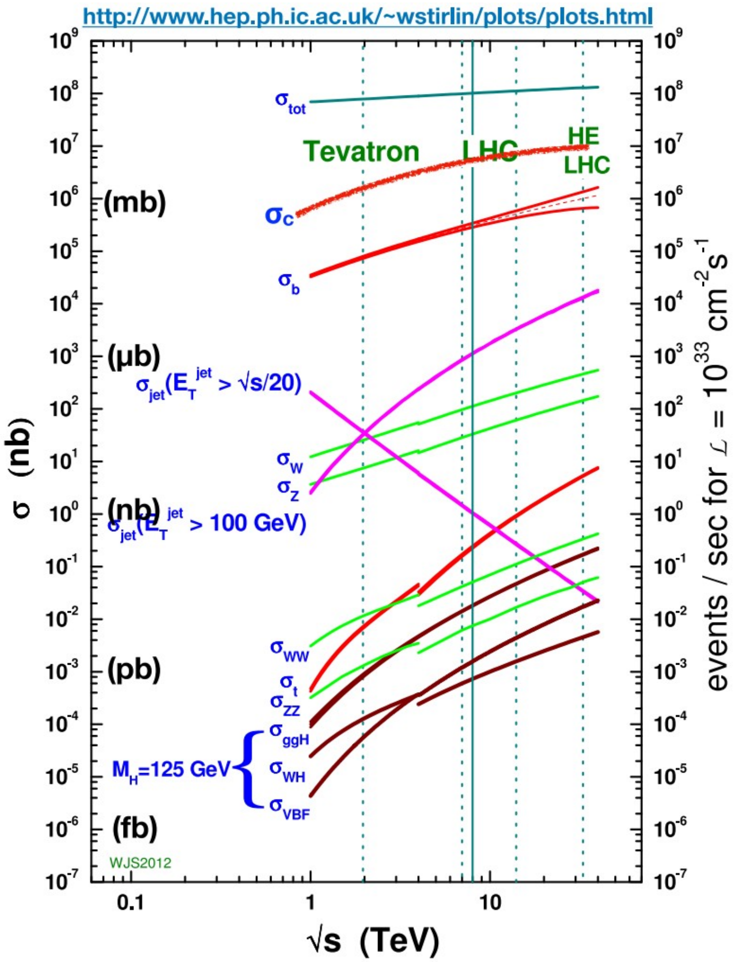


All readout systems renewed

RICH 1 redesigned, new photo detectors for RICH1 & RICH2

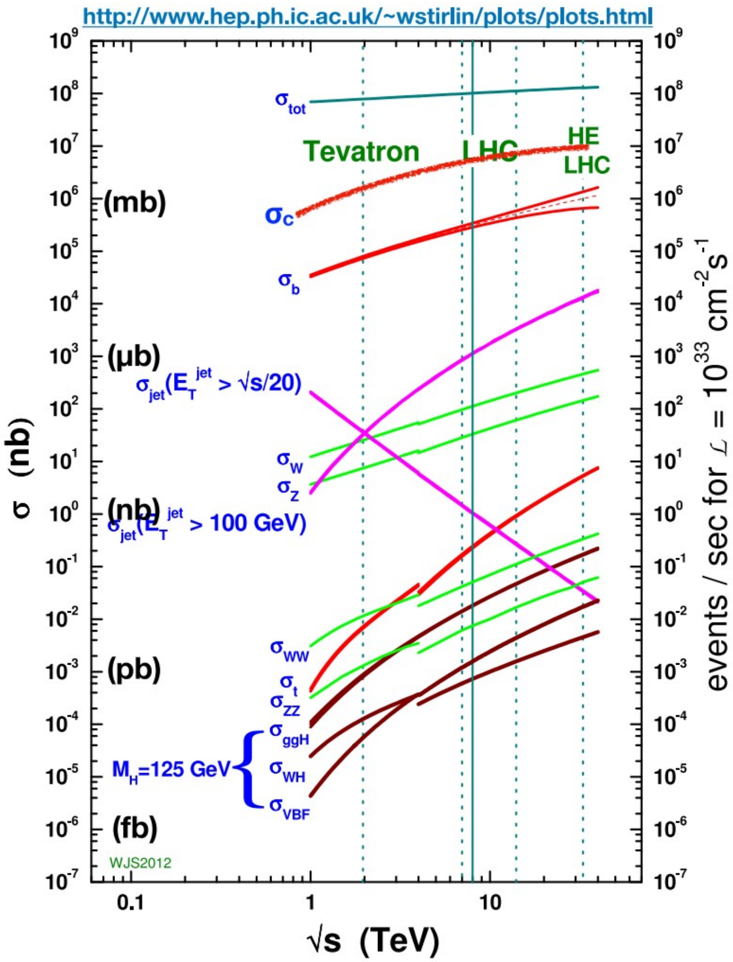
Calorimeters & muons: redundant components removed, new electronics, more shielding

Reaching the MHz signal era



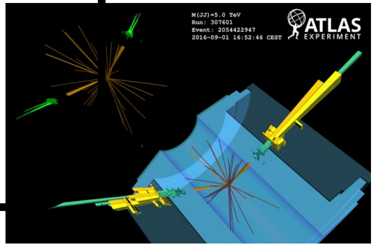
Run 3: Luminosity of $2x10^{33}$ $cm^{-2}s^{-1}$, $\sqrt{s}$ = 14 TeV

Reaching the MHz signal era

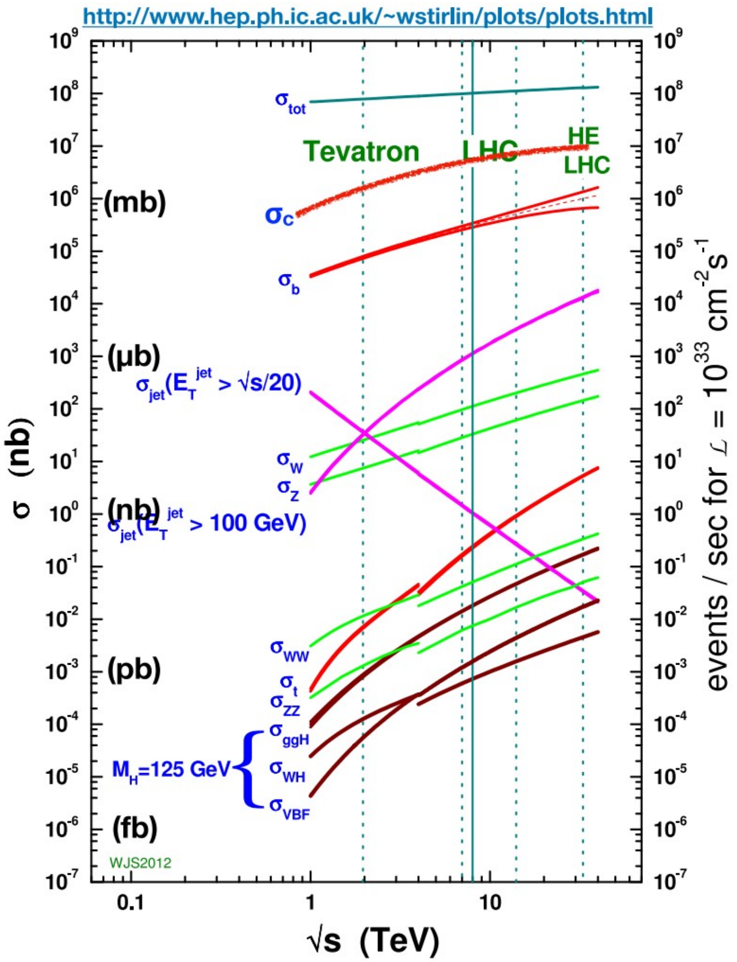


Run 3: Luminosity of 2×10^{33} $cm^{-2}s^{-1}$, $\sqrt{s}$ = 14 TeV

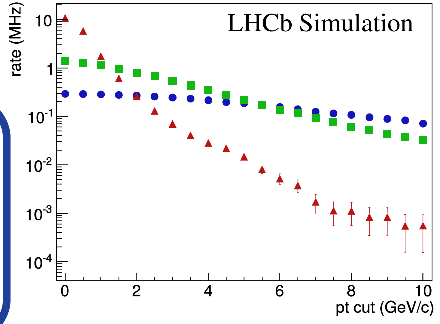
- General purpose LHC experiments: jets, electro-weak physics, Higgs physics
- Local characteristic signatures, e.g. high transverse energy
- Can trigger efficiently at ~100 kHz
  → hardware-level trigger possible



Reaching the MHz signal era

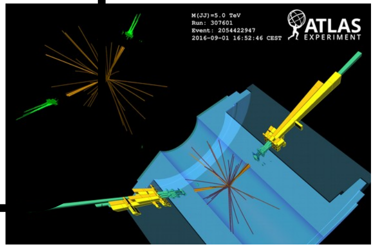


Run 3: Luminosity of 2×10^{33} cm^{-2}s^{-1}, $\sqrt{s}$ = 14 TeV



- Too many interesting events
- No “simple” local criteria for selection
  → hardware-level trigger not an option

- General purpose LHC experiments: jets, electro-weak physics, Higgs physics
- Local characteristic signatures, e.g. high transverse energy
- Can trigger efficiently at ~100 kHz
  → hardware-level trigger possible

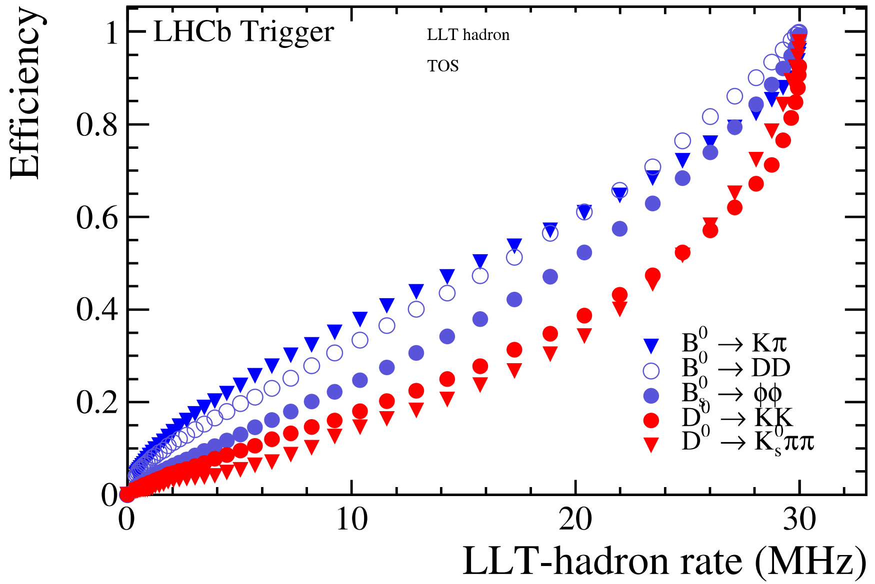


Change in trigger paradigm

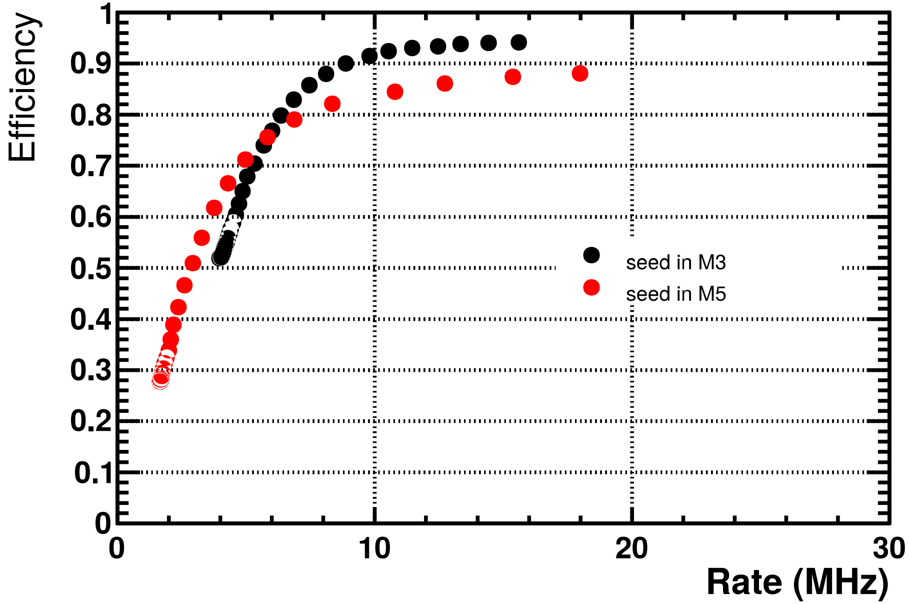
Access as much information about the collision as early as possible

Why no low level trigger?

Low level trigger on E_T from the calorimeter

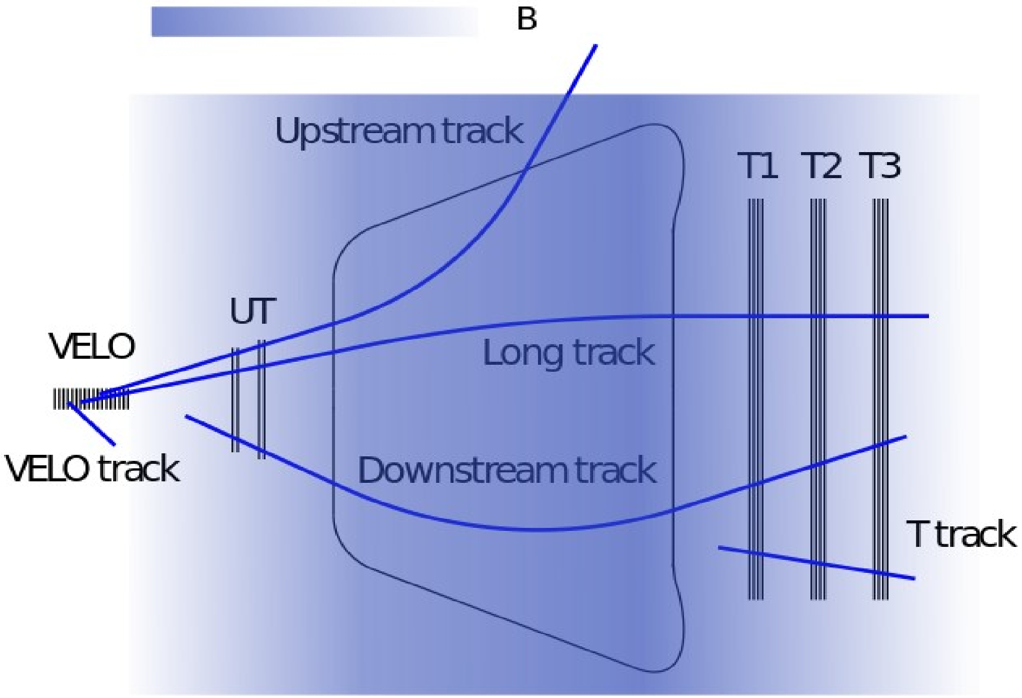


Low level trigger on muon p_T, $B \to K^*\mu\mu$



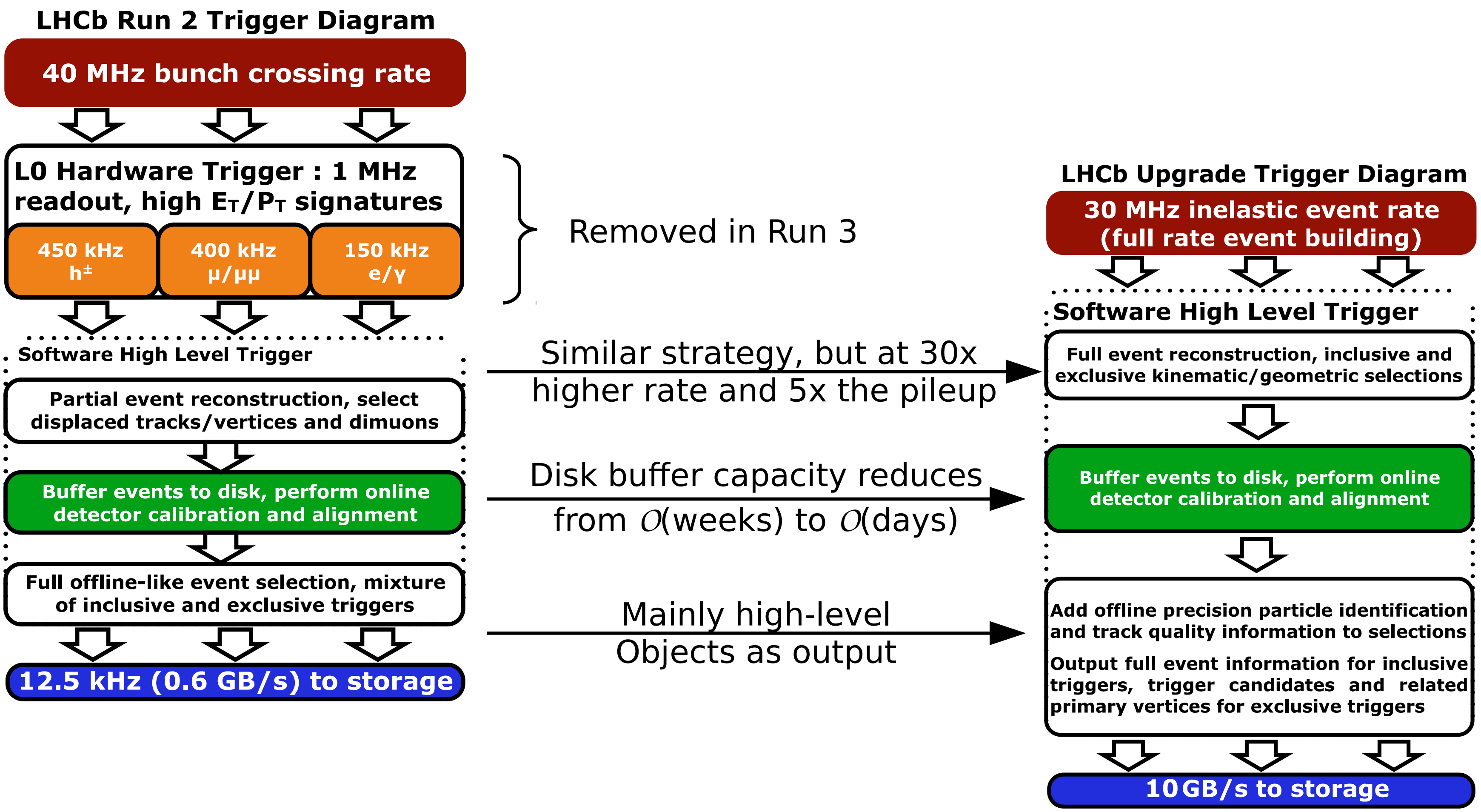
Need track reconstruction at first trigger stage

Tracks in the LHCb detector

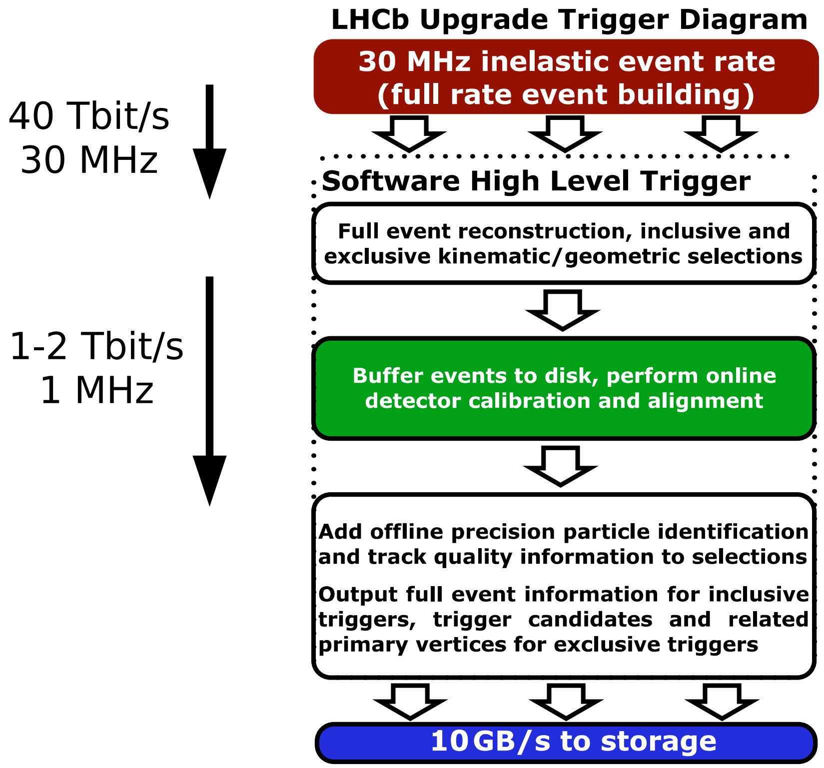


Need information from many subdetectors → read out full detector

Run 2 versus Run 3 trigger



Trigger in Run 3 (2021)

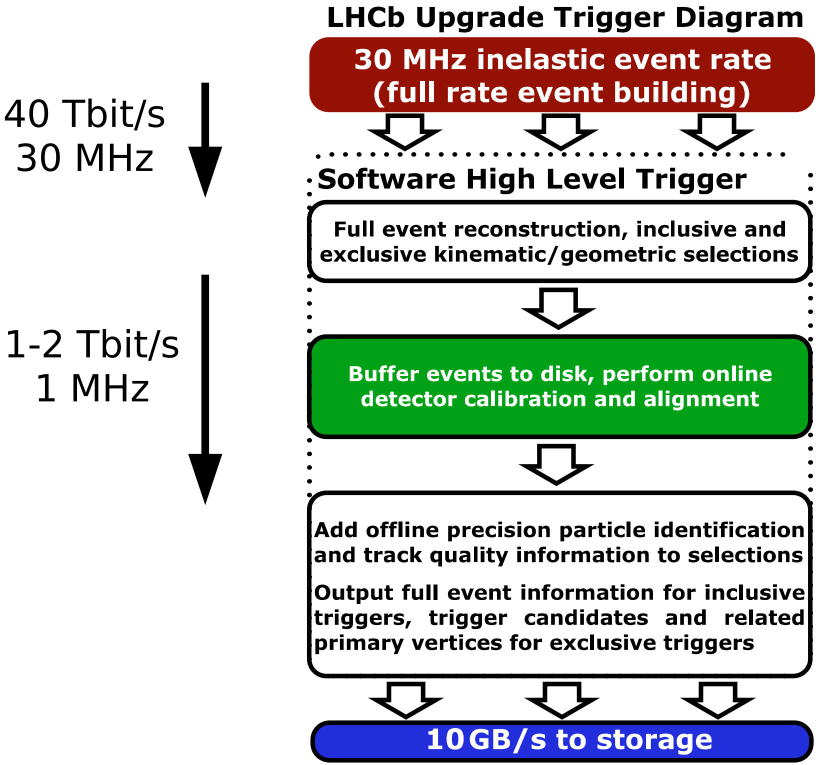


High Level Trigger 1 (HLT1)

- Full charged particle track reconstruction
- Few inclusive single or two-track selections
- Reduce event rate by roughly factor 30

- **High Level Trigger 2 (HLT2)**
- Aligned and calibrated detector
- Offline-quality track reconstruction
- Particle identification
- Full track fitting

Trigger in Run 3 (2021)

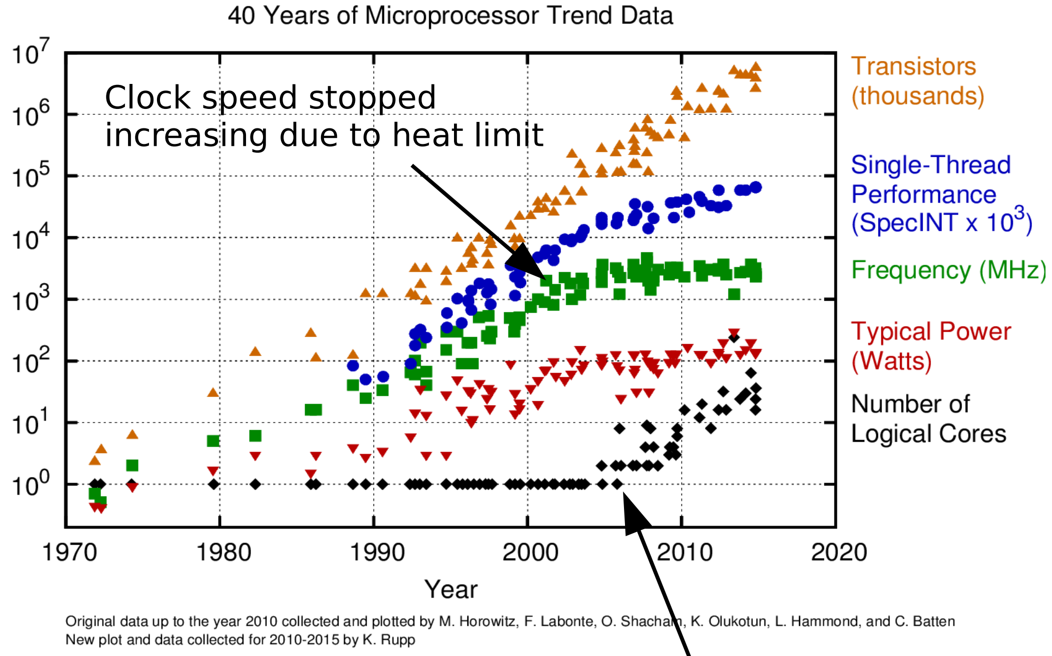


High Level Trigger 1 (HLT1)

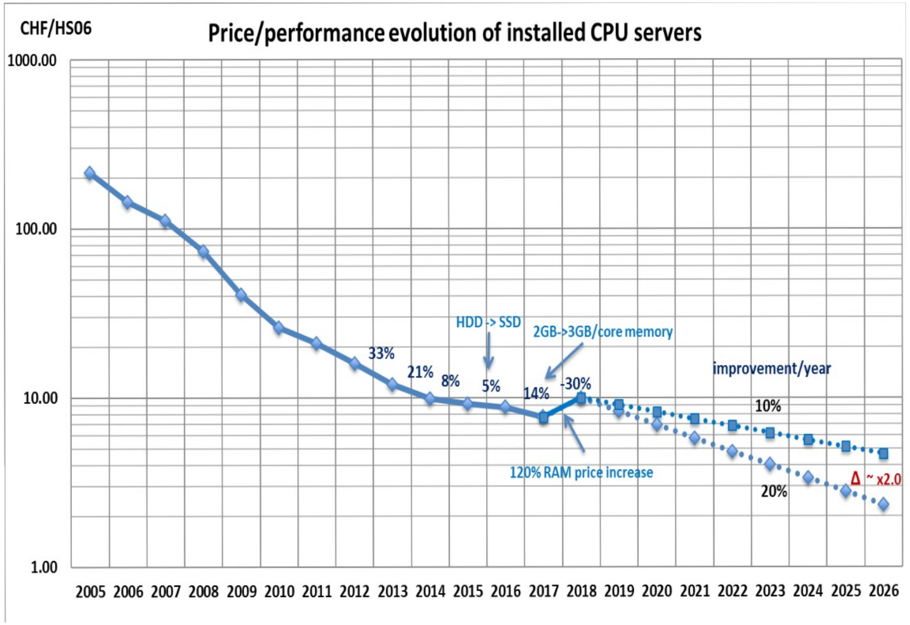
- Full charged particle track reconstruction
- Few inclusive single or two-track selections
- Reduce event rate by roughly factor 30

Track reconstruction @ 30 MHz is a huge computing challenge!

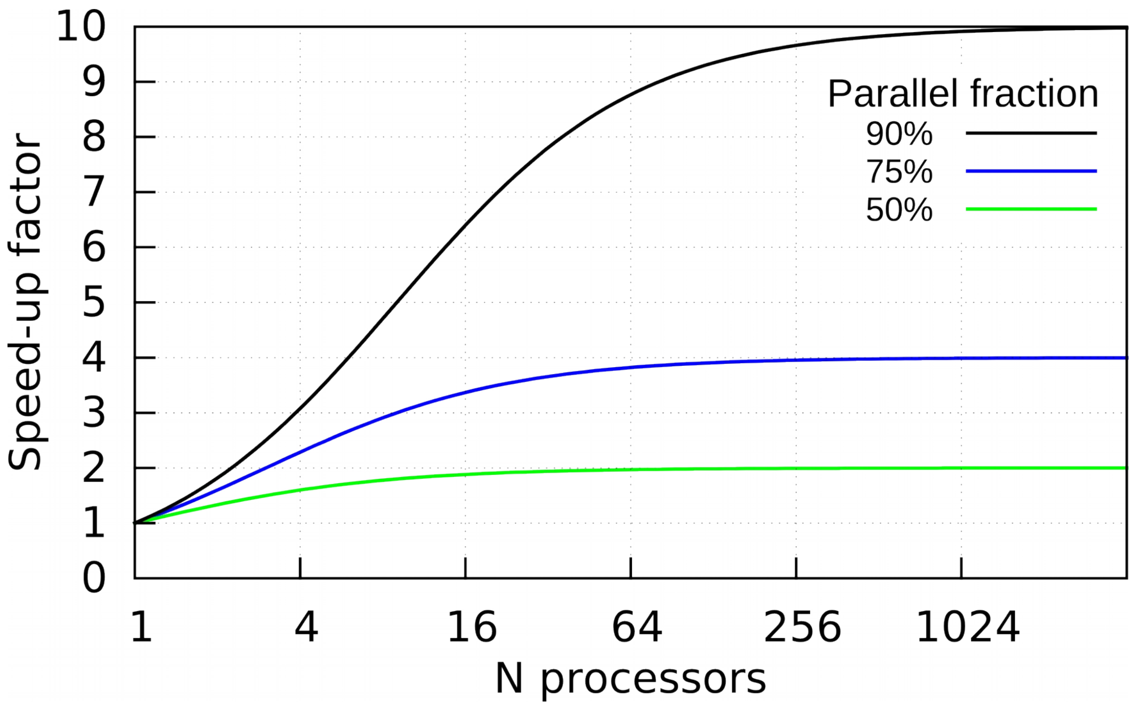
Today's computing landscape



Multiple core processors emerge



Amdahl’s law



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

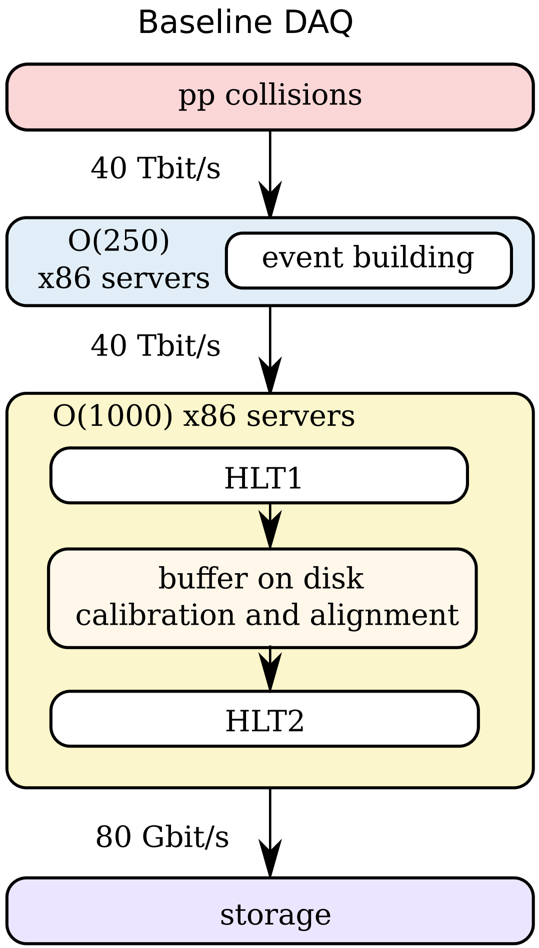
S: sequential part of program

P: parallel part of program

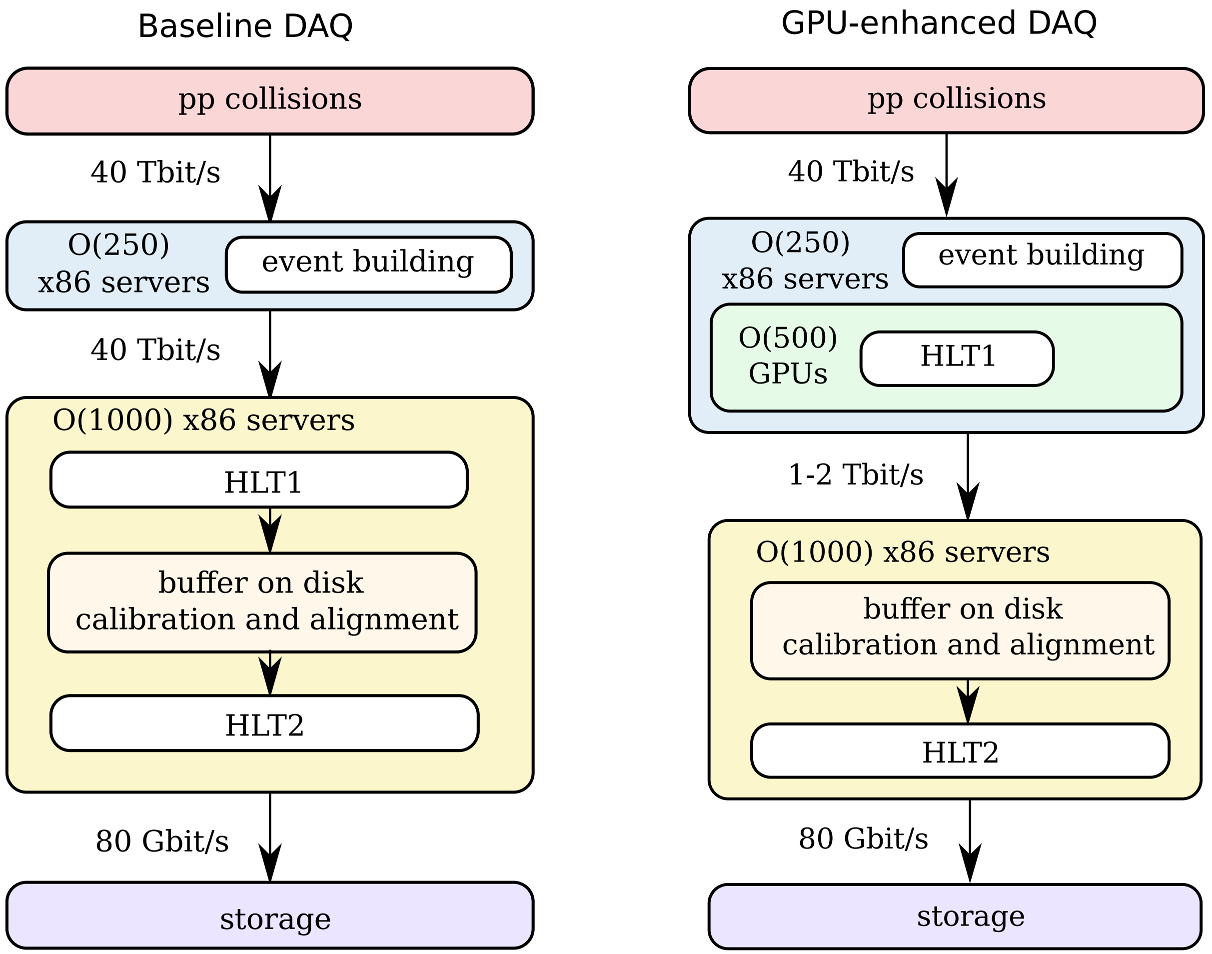
N: number of processors

Can we use the FLOPS available on a GPU to run HLT1 @ 30 MHz?

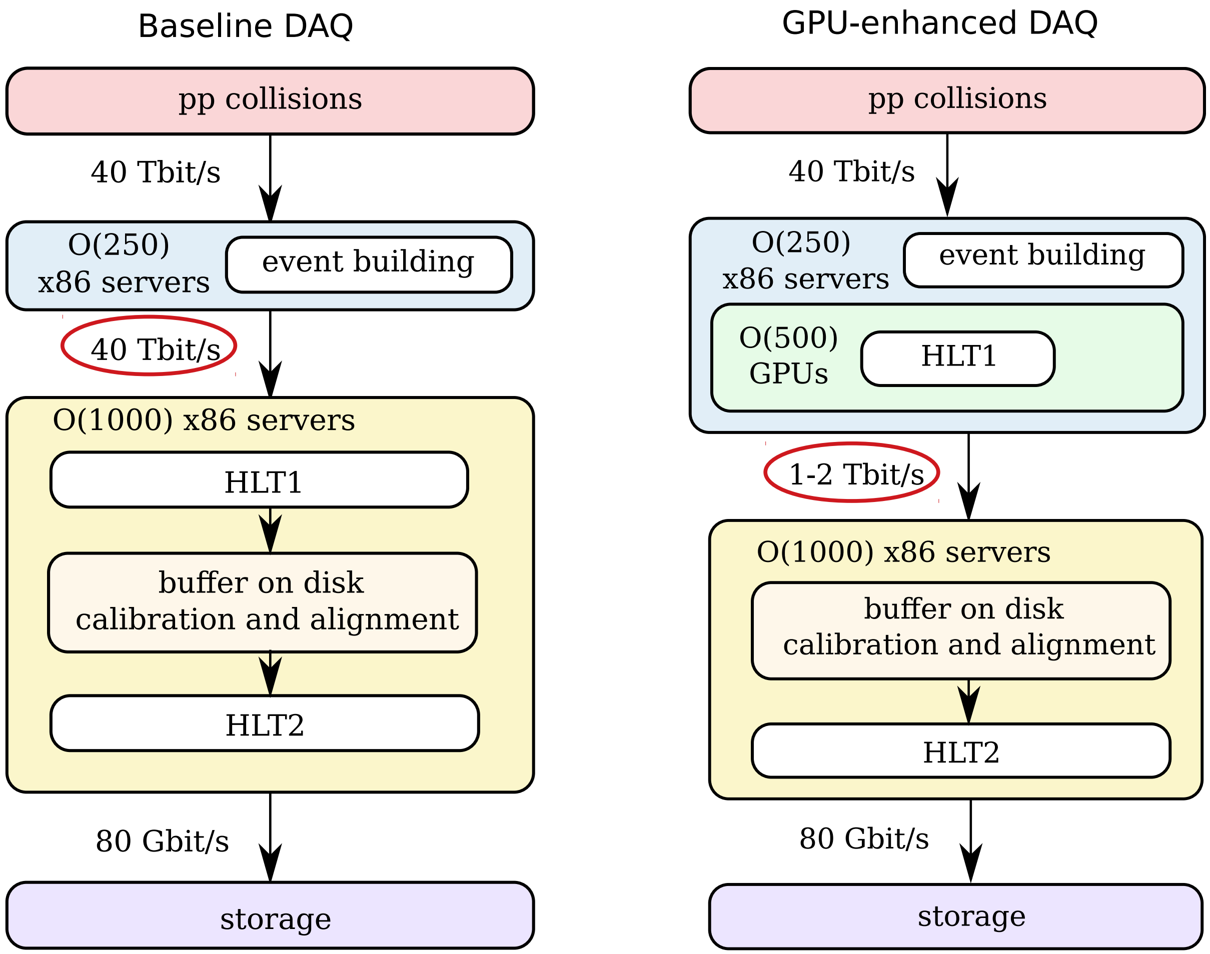
Where to place the GPUs?



Where to place the GPUs?



Where to place the GPUs?



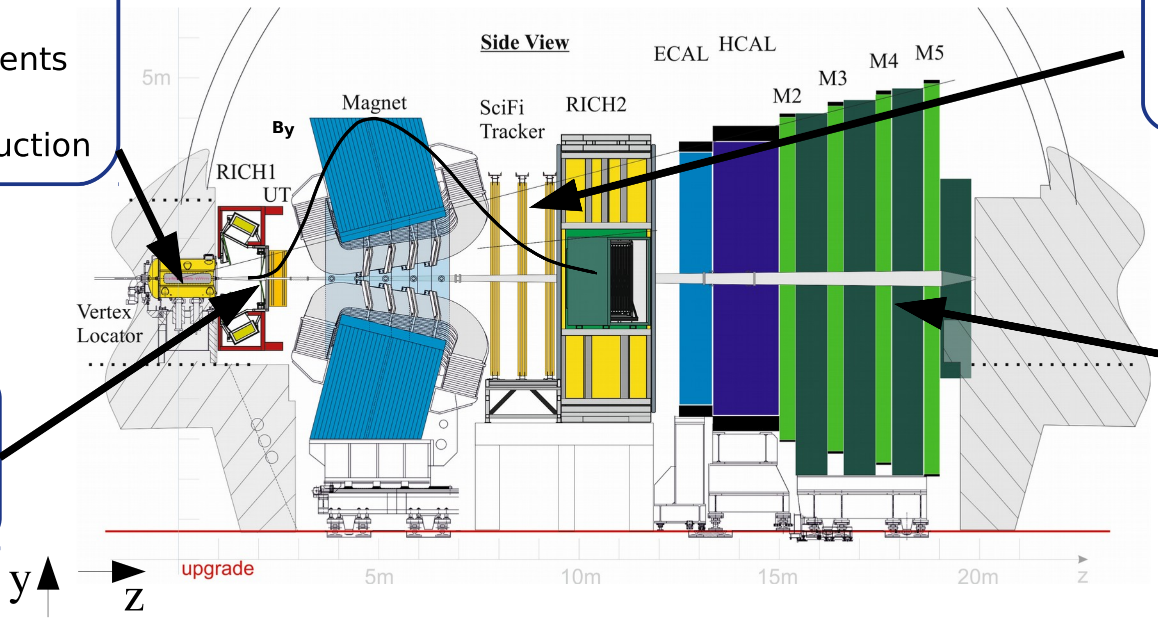
GPUs naturally integrate into LHCb's DAQ

If HLT1 can run on 500 GPUs
→ Save money on network
→ Buy GPUs instead

LHCb HLT1 elements

Velo
- Decode raw data
- Clustering of measurements
- Track reconstruction
- Primary vertex reconstruction

SciFi
- Decode raw data
- Track reconstruction



Muons
- Decode raw data
- Match hits to tracks

UT
- Decode raw data
- Track reconstruction

Track fit: Kalman filter

Find secondary vertices

Selections
- 1-track selection
- 2-track selection
- Based on p, p_t, displacement, vertex criteria and muon identification

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!

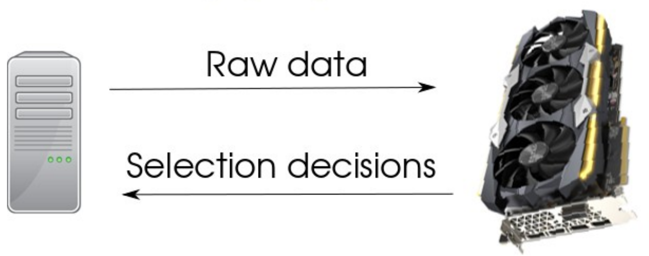
The Allen project

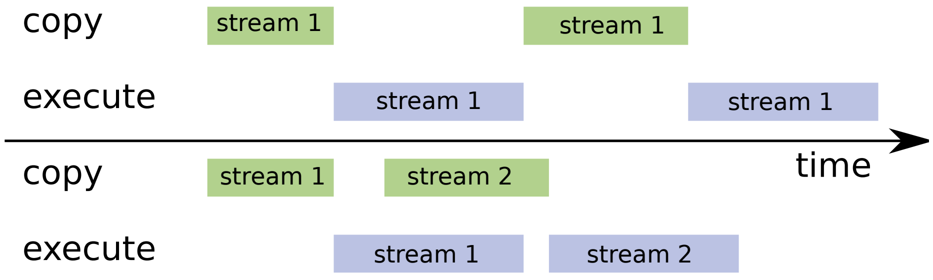
- Fully standalone software project: https://gitlab.cern.ch/lhcb/Allen
- Only requirements: a C++17 compliant compiler & CUDA v10.1
- Built-in physics validation
- Configurable sequence, custom memory manager
- Cross-architecture compatibility

- Project started in February 2018
- Roughly 14 part-time developers, mostly students
- 2 almost full-time
- After 15 months of development time:
  project reviewed as viable solution for Run 3 (starting in 2021)

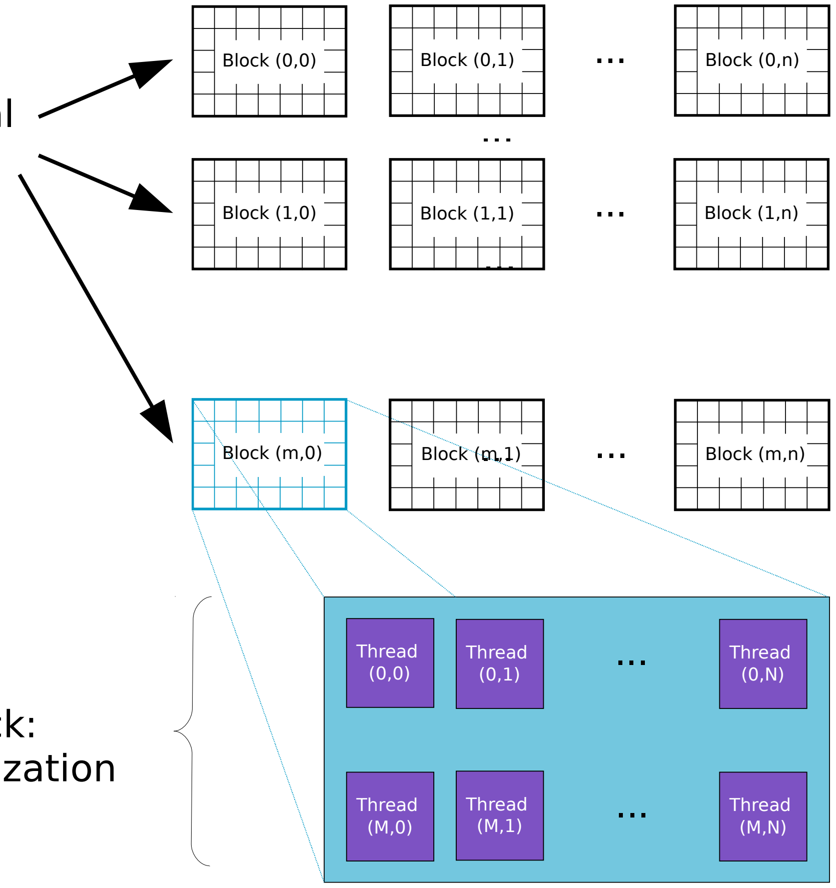
- Named after Frances E. Allen

HLT1 on GPUs





Individual events



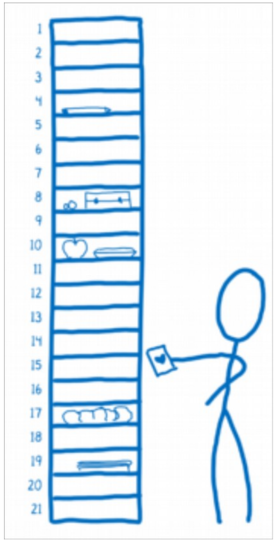
Within one block:
intra-event parallelization

Software framework

Static scheduler
For sequence of algorithms

```
SEQUENCE_T(
  velo_estimate_input_size_t,
  prefix_sum_velo_clusters_t,
  velo_masked_clustering_t,
  velo_calculate_phi_and_sort_t,
  velo_fill_candidates_t,
  velo_search_by_triplet_t,
  velo_weak_tracks_adder_t)
```

Memory manager for
GPU memory



Bachelor, Master and PhD students contribute in only a few months time

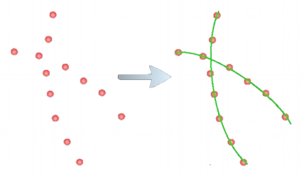
Recurrent tasks of HLT1

Raw data decoding

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

Track reconstruction

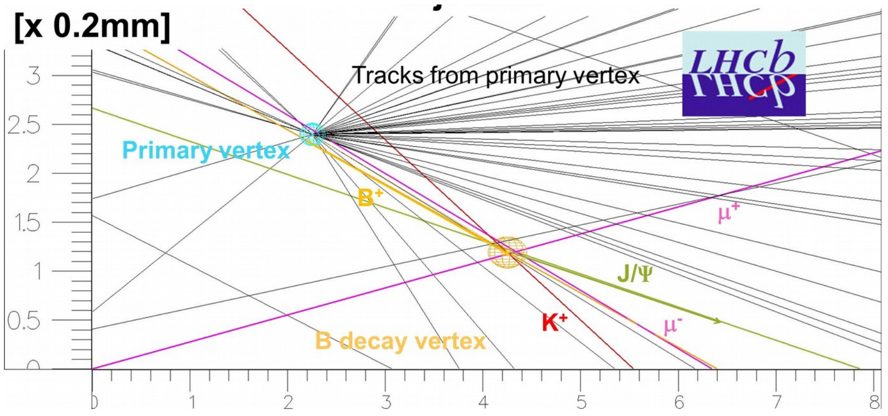
- Consists of two steps:
  - Pattern recognition: Which hits belong to which track? → Huge combinatorics
  - Track fitting: Done for every track
- Parallelizable over events, combinations of hits, and tracks



f(x) = … +/- …

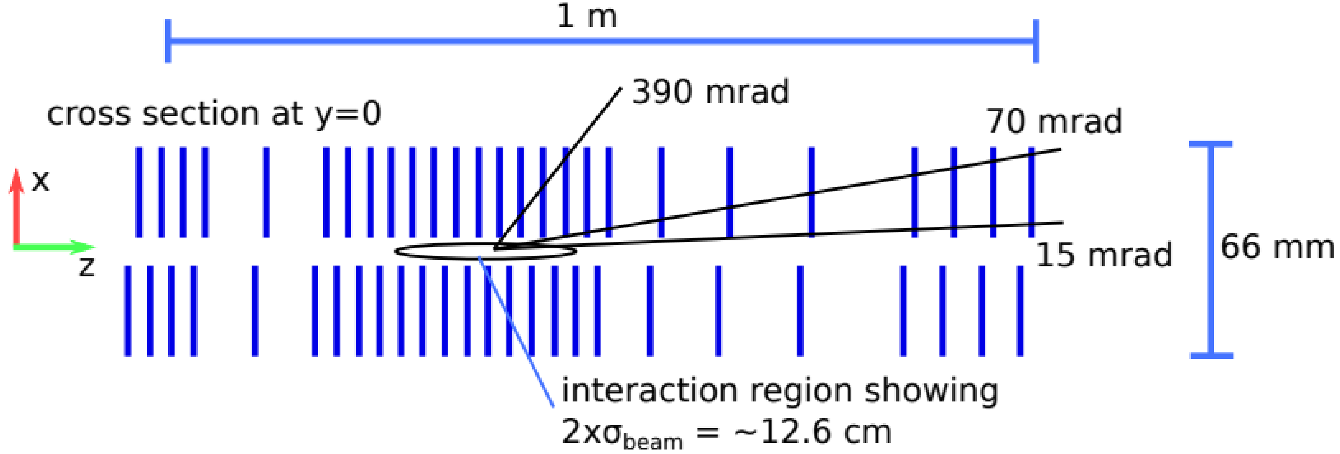
Vertex finding

- Where did proton-proton collisions take place?
- Where did particles decay within the detector volume?
- Parallelizable over events, combinations of tracks

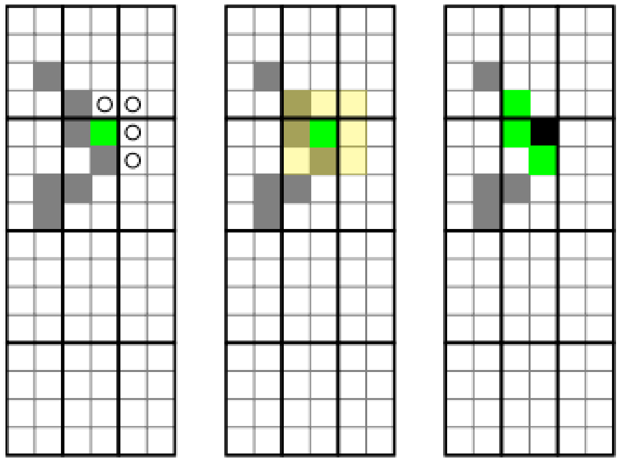


Velo detector: clustering

26 planes of silicon pixel detectors

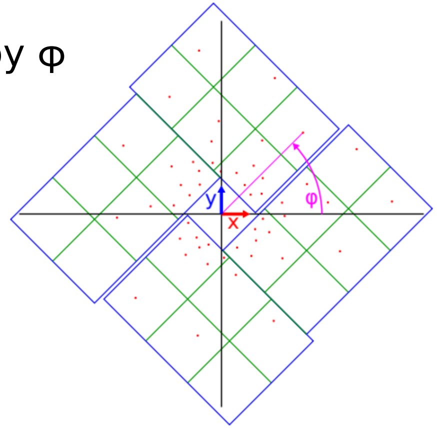


Clustering with bit masks

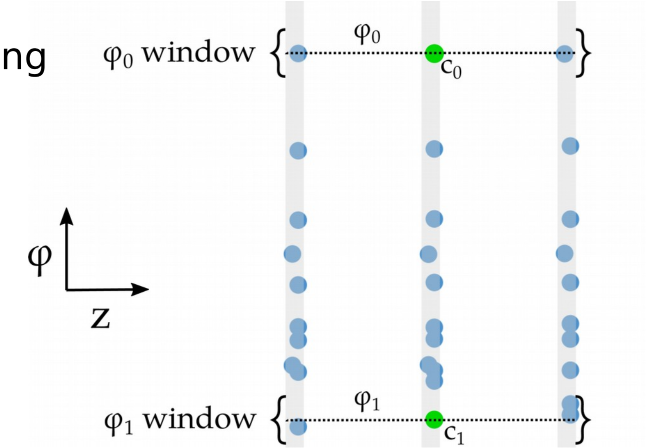


Velo detector: track reconstruction

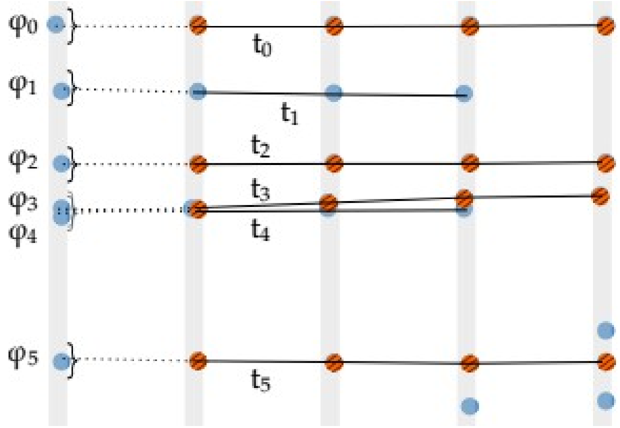
1) Sort hits by φ



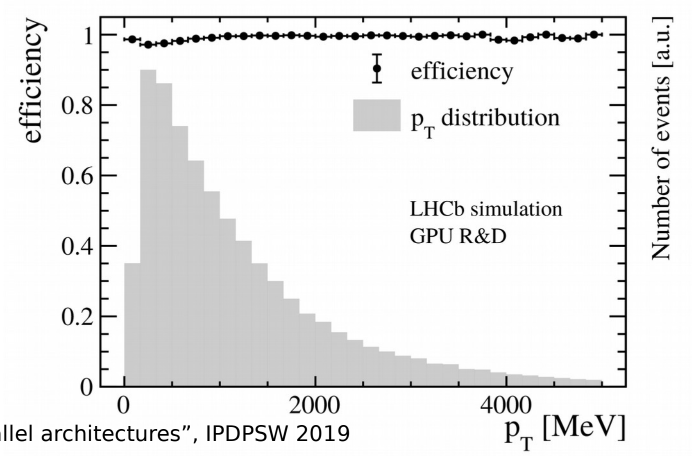
2) Triplet seeding



3) Triplet forwarding

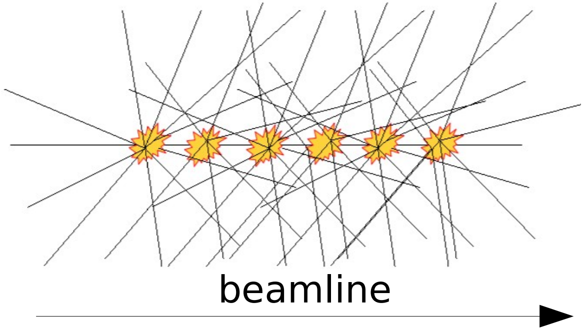


Track reconstruction efficiency for tracks originating from B decays



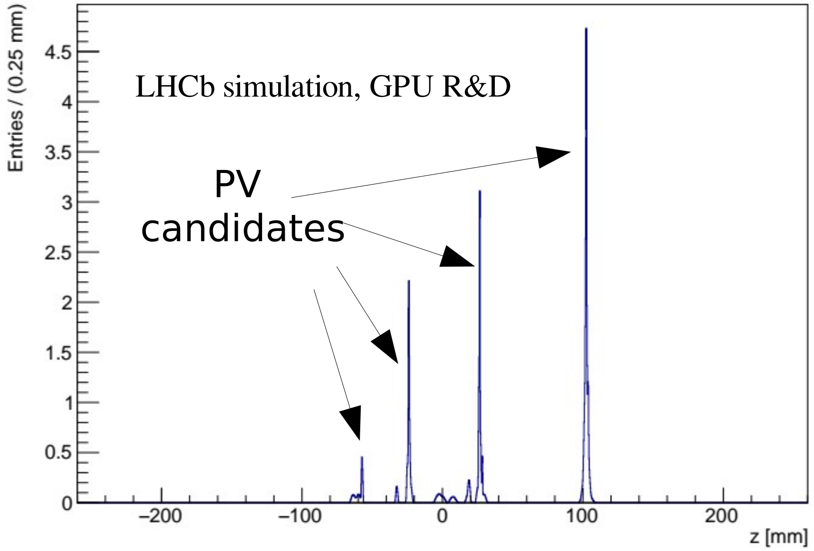
D. Campora, N. Neufeld, A. Riscos Núñez: "A fast local algorithm for track reconstruction on parallel architectures", IPDPSW 2019

Velo detector: primary vertex reconstruction

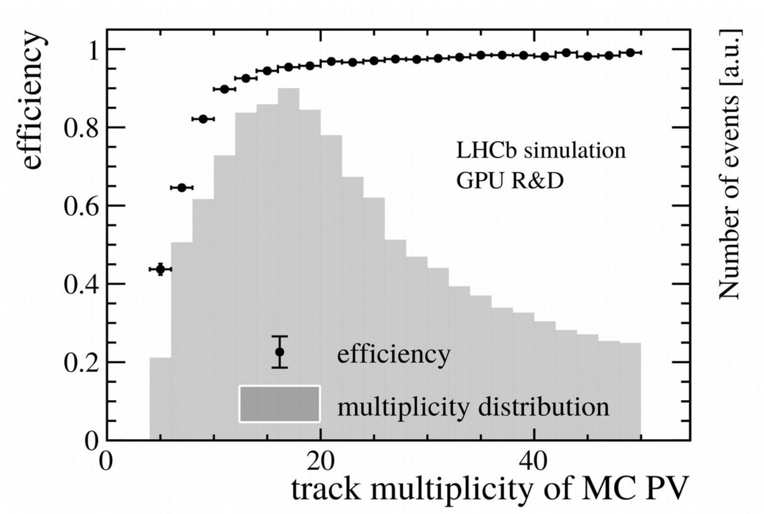


Florian Reiss (LPNHE)

Point of closest approach of tracks to beamline

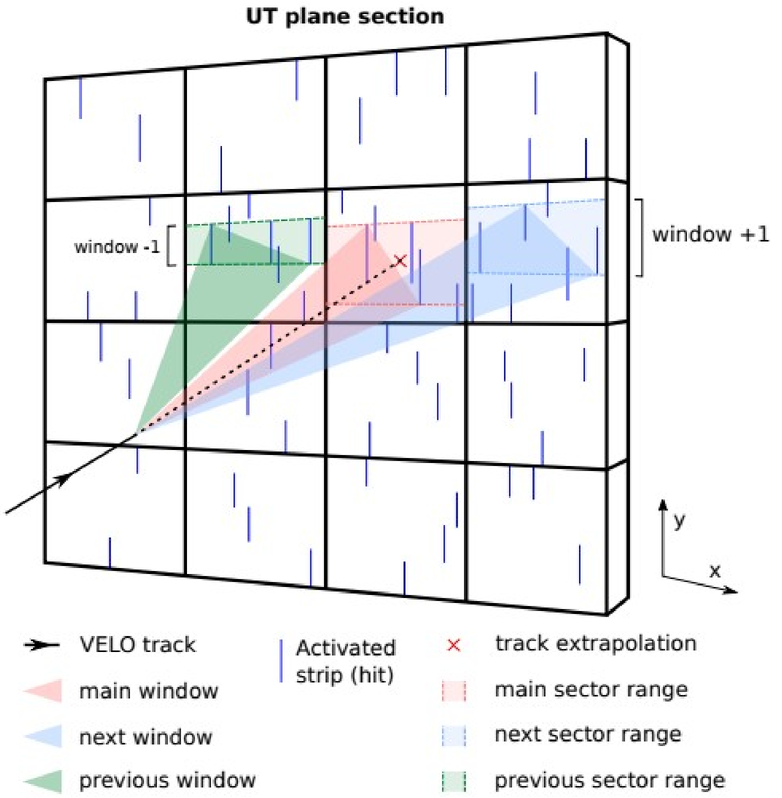


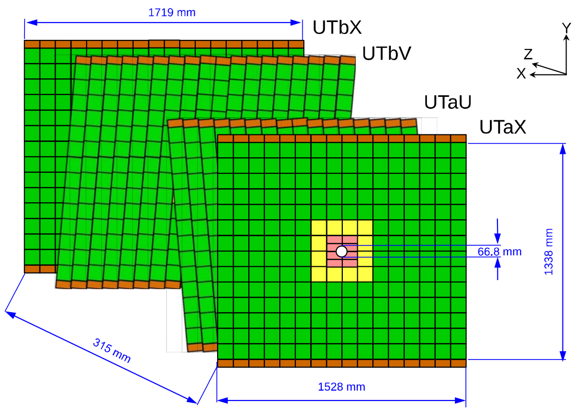
PV reconstruction efficiency



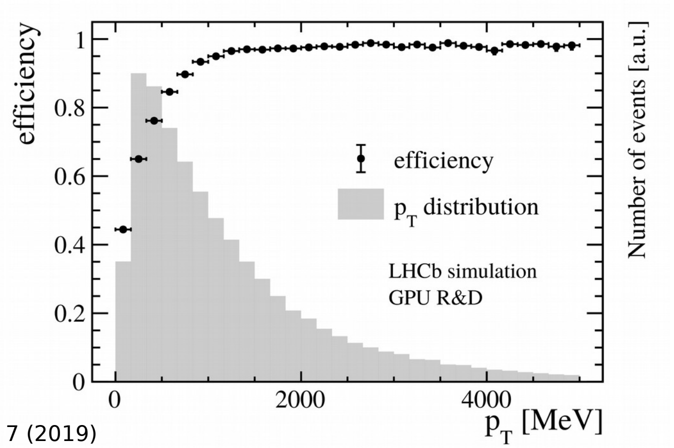
UT detector: track reconstruction

4 planes of silicon strip detectors





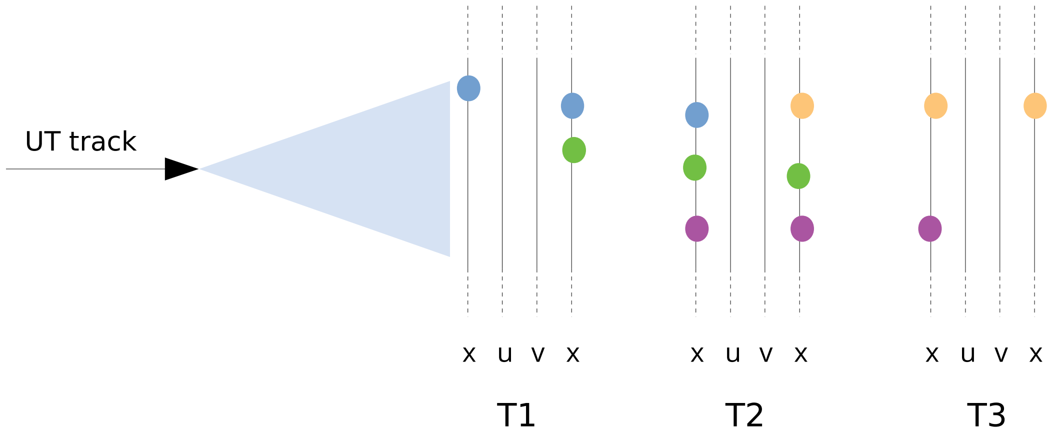
Track reconstruction efficiency for tracks originating from B decays

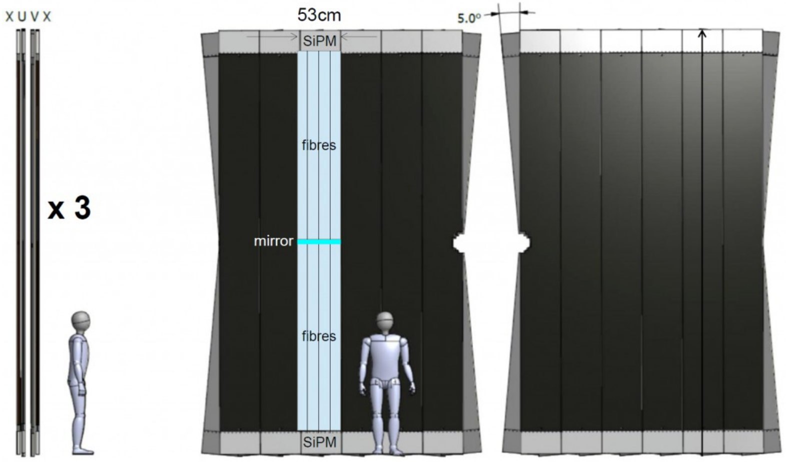


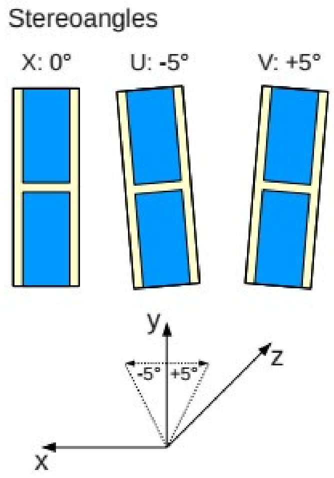
P. Fernandez Declara, D. Campora Perez, J. Garcia-Blas, D. vom Bruch, J. Daniel Garca, N. Neufeld , IEEE Access 7 (2019)

SciFi detector

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

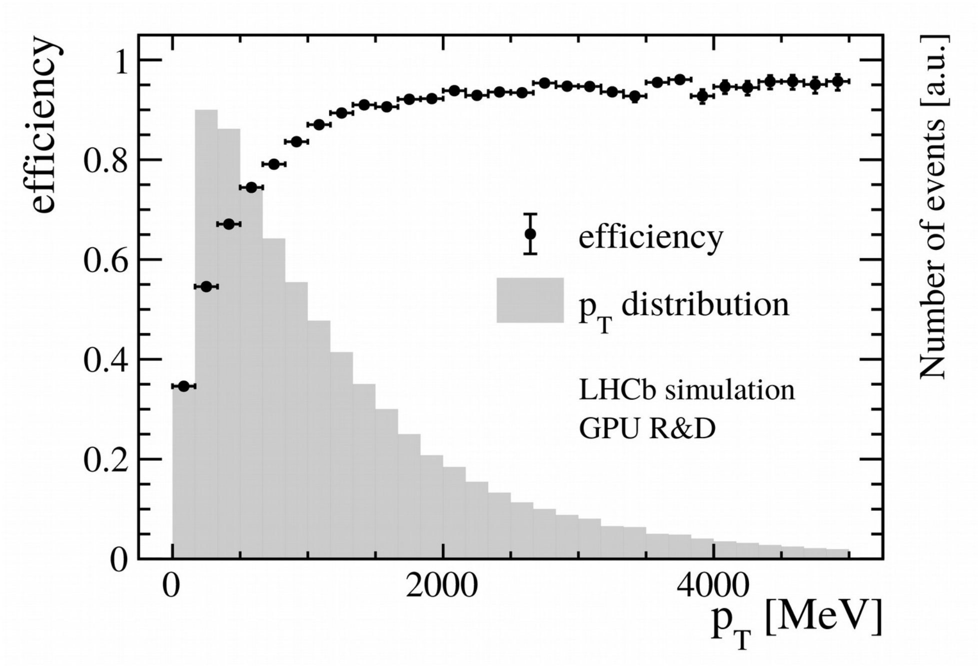




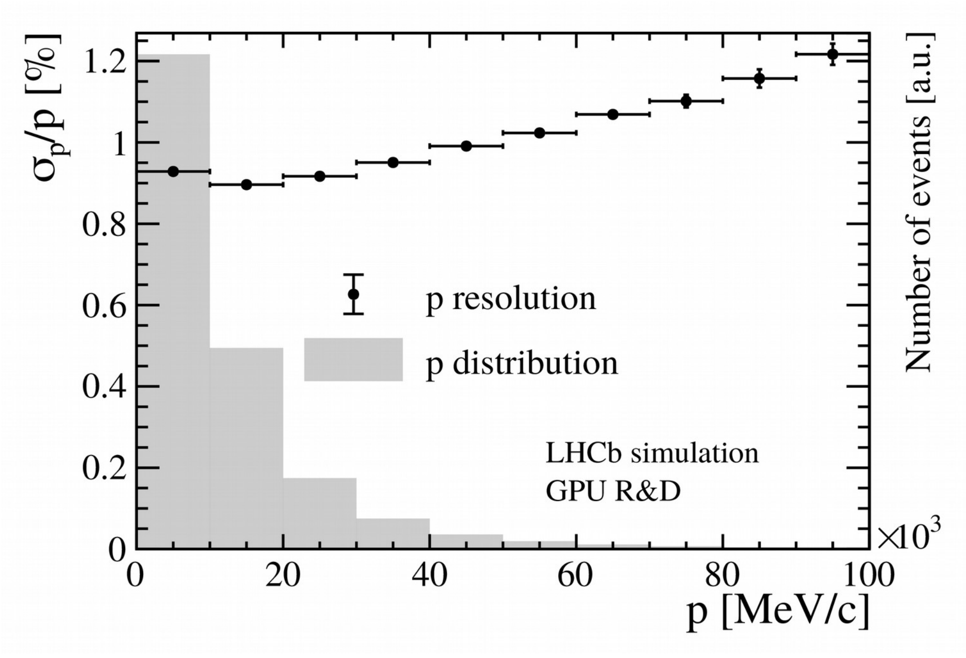


SciFi detector: track reconstruction

Track reconstruction efficiency for tracks originating from B decays

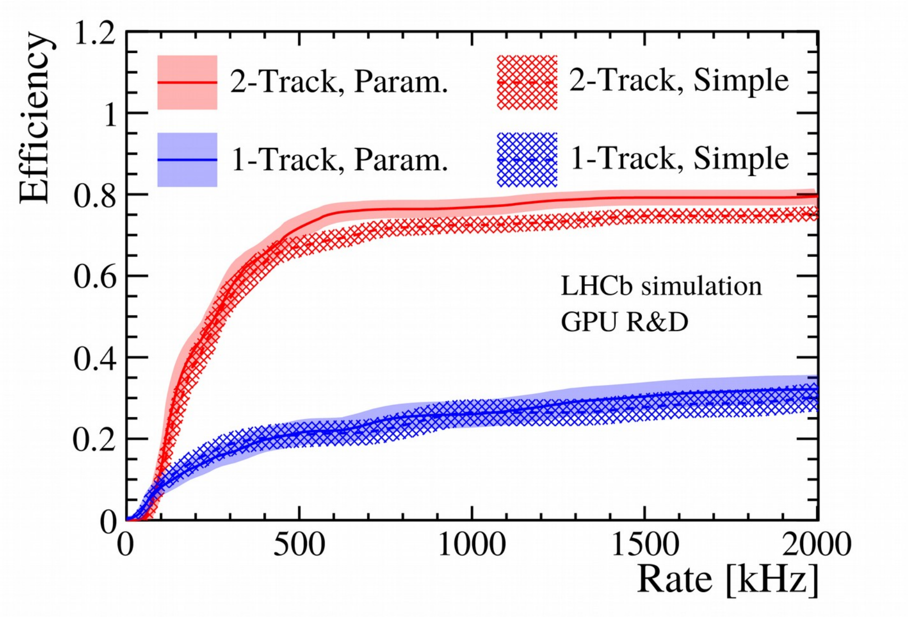


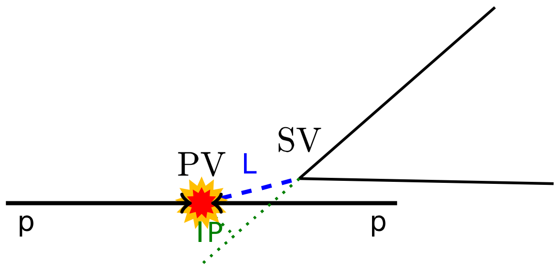
Momentum resolution



Kalman filter

Improved track description → better impact parameter resolution



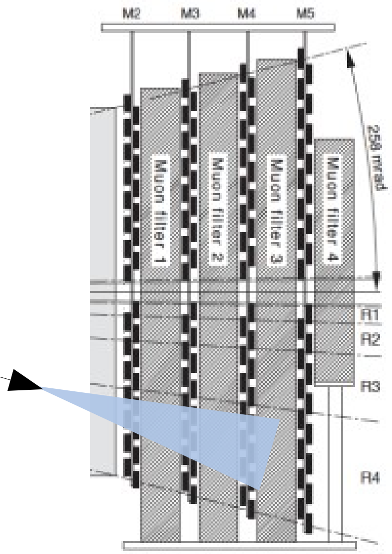


- Simple: Simplified Kalman filter with constant momentum assumption
- Param.: Parameterized Kalman filter with momentum estimate from SciFi track reconstruction

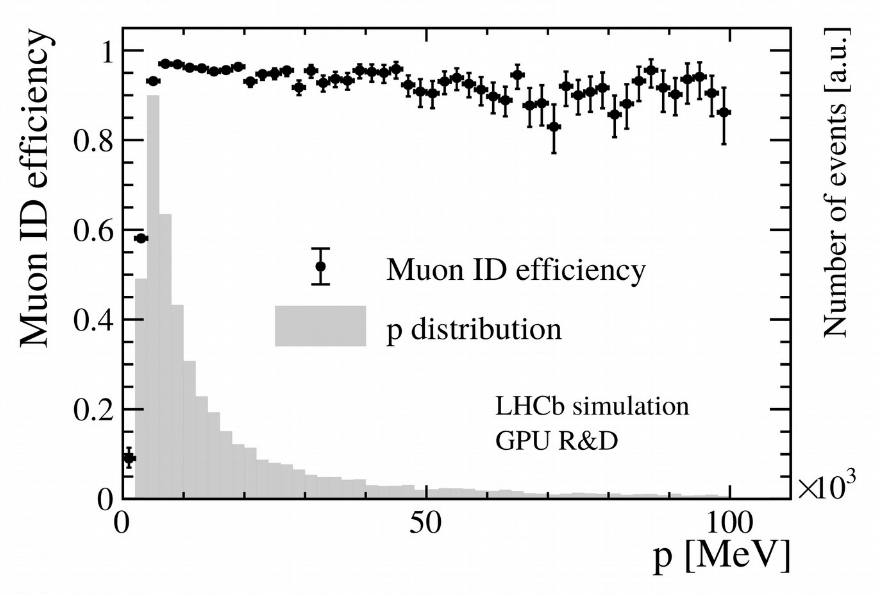
Muon identification

Four multi-wire proportional chambers
Interleaved with iron walls

SciFi track

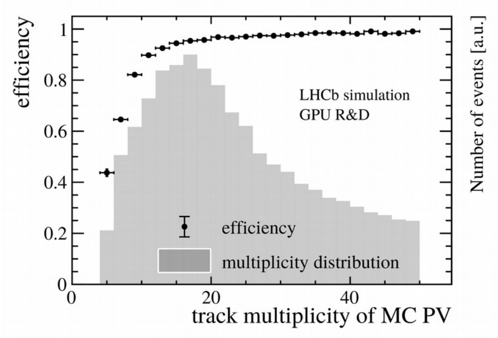


Muon identification efficiency

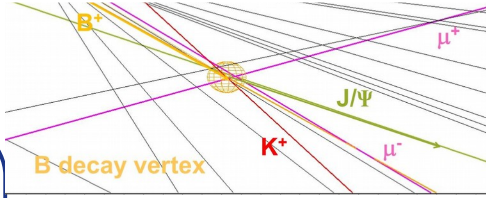


Ingredients for selections

Primary vertices



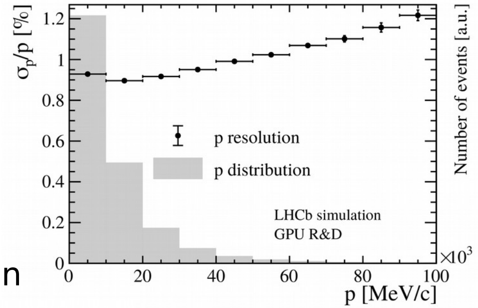
Secondary vertices



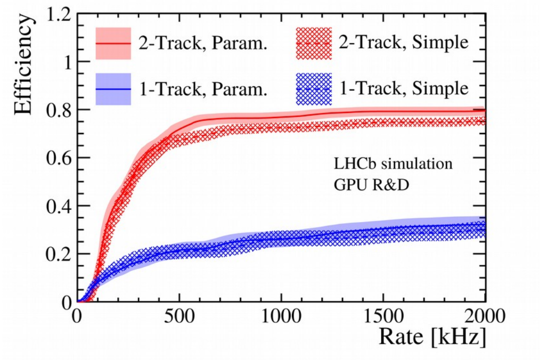
Selections

- 1-track selection
- 2-track selection
- Based on p, p_t, displacement, vertex criteria and muon identification

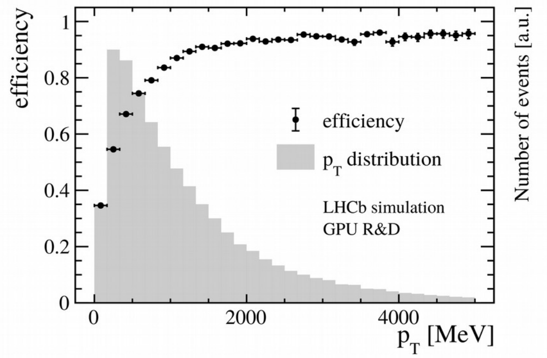
Momentum



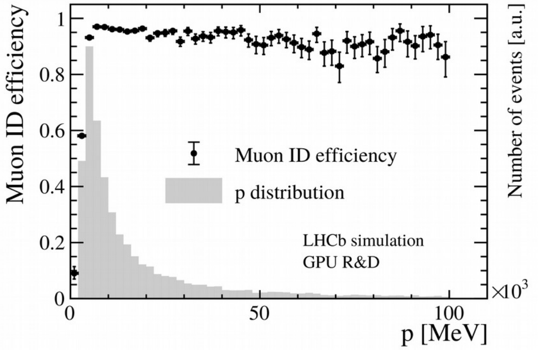
Impact parameter



Tracks



Muon identification



Event selection

Trigger	Rate [kHz]
1-Track	215 ± 18
2-Track	659 ± 31
High-p_T muon	5 ± 3
Displaced dimuon	74 ± 10
High-mass dimuon	134 ± 14
Total	999 ± 38

Event rate reduced from 30 MHz to 1 MHz

Selection efficiencies

Allen

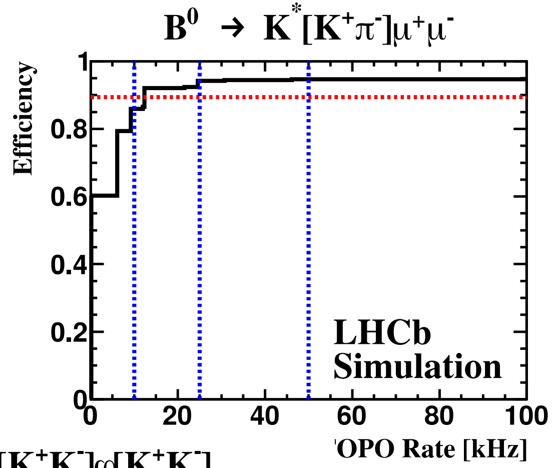
Selection efficiencies, values given in %

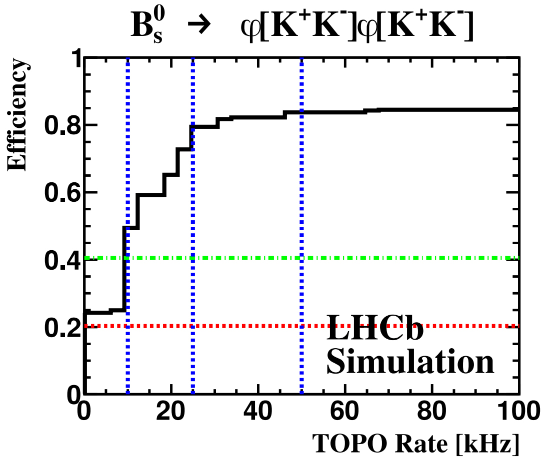
Signal	GEC	TIS -OR- TOS	TOS	GEC × TOS
$B^0 \to K^{*0}\mu^+\mu^-$	89 ± 2	91 ± 2	89 ± 2	79 ± 3
$B^0 \to K^{*0}e^+e^-$	84 ± 3	69 ± 4	62 ± 4	52 ± 4
$B^0_s \to \phi\phi$	83 ± 3	76 ± 3	69 ± 3	57 ± 3
$D^+_s \to K^+K^-\pi^+$	82 ± 4	59 ± 5	43 ± 5	35 ± 4
$Z \to \mu^+\mu^-$	78 ± 1	99 ± 0	99 ± 0	77 ± 1

TIS: Trigger independent from signal
TOS: Trigger on signal

Consistent physics performance with TDR, which assumed running on x86 architecture

Technical Design Report



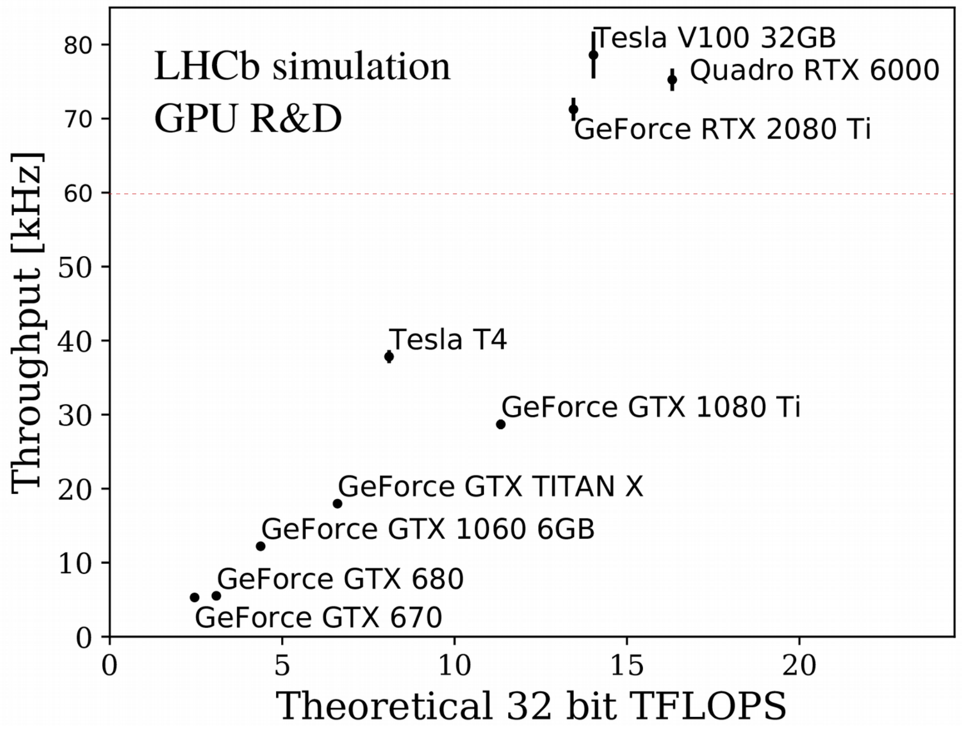


Full HLT1 running on GPUs

Physics performance matches HLT1 requirements

What about the throughput performance?

Throughput on various GPUs



Full HLT1 sequence

The system can run on 500 GPUs
→ network cost saving → no additional cost from using GPUs

Allen publication

- First publication submitted: arXiv:1912.09161

Cornell University

the Si

arXiv.org > physics > arXiv:1912.09161

Search...

Help | Advance

Physics > Instrumentation and Detectors

Allen: A high level trigger on GPUs for LHCb

R. Aaij, J. Albrecht, M. Belous, P. Billoir, T. Boettcher, A. Brea Rodríguez, D. vom Bruch, D. H. Cámpora Pérez, A. Casais Vidal, D. C. Craik, P. Fernandez Declara, L. Funke, V. V. Gligorov, B. Jashal, N. Kazeev, D. Martínez Santos, F. Pisani, D. Pliushchenko, S. Popov, R. Quagliani, M. Rangel, F. Reiss, C. Sánchez Mayordomo, R. Schwemmer, M. Sokoloff, H. Stevens, A. Ustyuzhanin, X. Vilasís Cardona, M. Williams

(Submitted on 19 Dec 2019)

We describe a fully GPU-based implementation of the first level trigger for the upgrade of the LHCb detector, due to start data taking in 2021. We demonstrate that our implementation, named Allen, can process the 40 Tbit/s data rate of the upgraded LHCb detector and perform a wide variety of pattern recognition tasks. These include finding the trajectories of charged particles, finding proton-proton collision points, identifying particles as hadrons or muons, and finding the displaced decay vertices of long-lived particles. We further demonstrate that Allen can be implemented in around 500 scientific or consumer GPU cards, that it is not I/O bound, and can be operated at the full LHC collision rate of 30 MHz. Allen is the first complete high-throughput GPU trigger proposed for a HEP experiment.

Comments: 12 pages, 12 figures, submitted to Computing and Software for Big Science
Subjects: **Instrumentation and Detectors (physics.ins-det)**; High Energy Physics - Experiment (hep-ex)
Cite as: arXiv:1912.09161 [physics.ins-det]
(or arXiv:1912.09161v1 [physics.ins-det] for this version)

Bibliographic data

[Enable Bibex(What is Bibex?)]

Submission history

From: Dorothea Vom Bruch [view email]
[v1] Thu, 19 Dec 2019 12:37:20 UTC (1,143 KB)

Integration test with event building server

Impact on event building when running Allen?

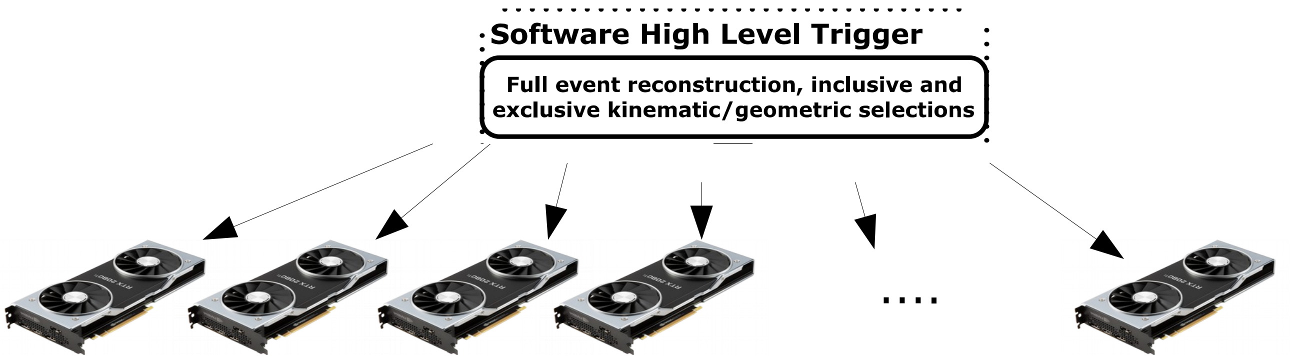
Monitoring temperatures, memory bandwidths, processing rate, ...

The Allen team

Summary

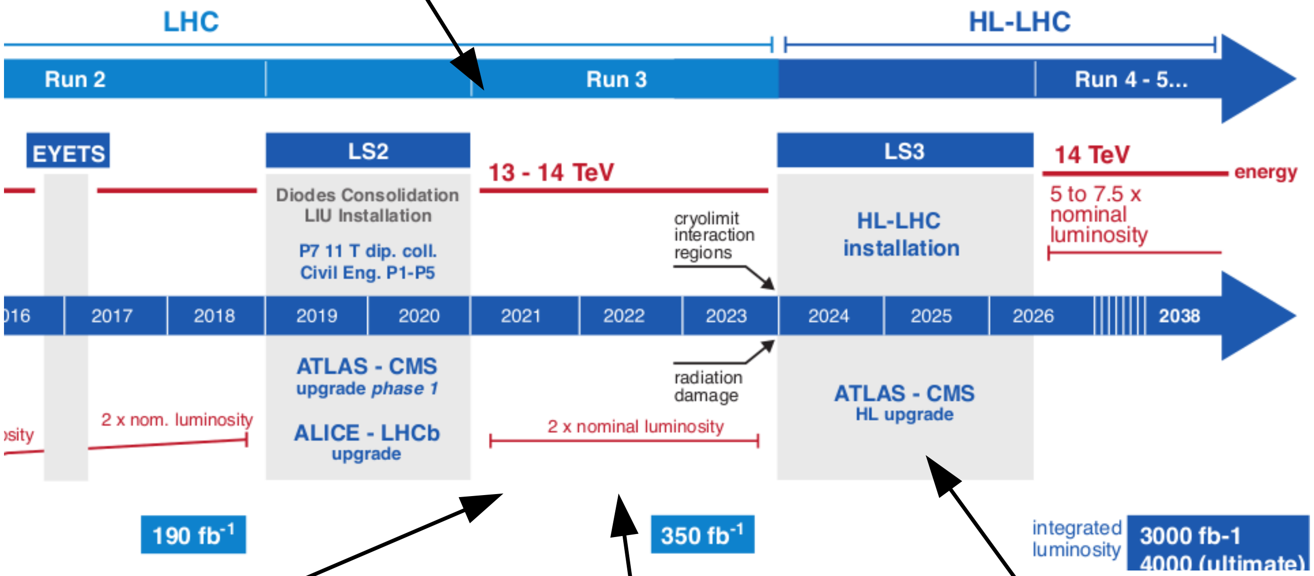
- Allen is the first complete high throughput trigger implementation on GPUs
- Baseline HLT1 can run on GPUs
- Efficient selections enable full exploitation of statistics in Run 3

  → crucial to explore new physics scenarios
- Scaling of GPU performance → maximize physics discovery potential of LHCb



LHC Schedule

Suggested HLT1 on GPUs for LHCb @ 40 Tbit/s



CMS: demonstrator of GPUs in high level trigger

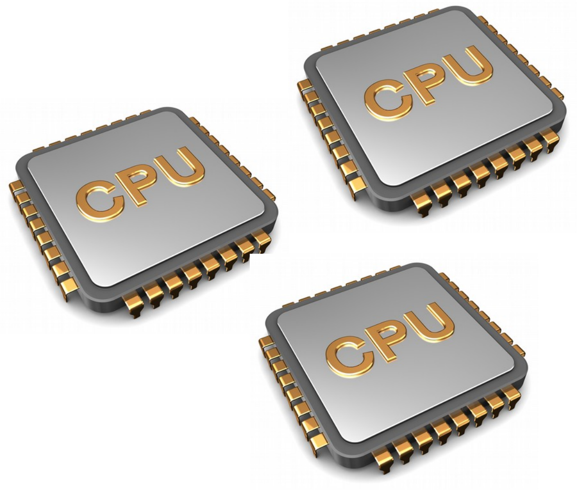
ALICE: GPUs used for data compression @ 5 Tbit/s

How will data rates be handled after LS3?

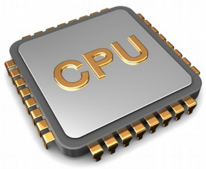
Outlook

- Features of Allen are not tied to HLT1 for LHCb
- High-throughput applications can profit from a similar setup
- Ideas for parallelization of algorithms can be useful for other applications

- Higher luminosity not only challenges real-time event selection
- Simulation production also requires major computing resources
- Need common effort to make best use of heterogeneous computing within high energy physics


CPU
CPU
CPU


FUTURE
PAST


CPU

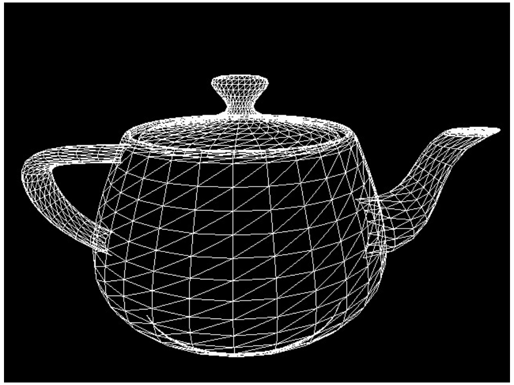
???

Backup

Graphics requirements

Graphics pipeline

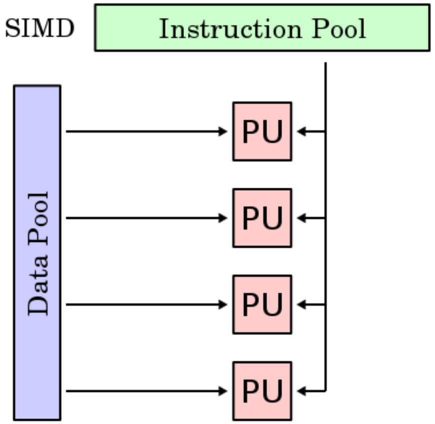
- Huge amount of arithmetic on independent data:
  - Transforming positions
  - Generating pixel colors
  - Applying material properties and light situation to every pixel



Hardware needs

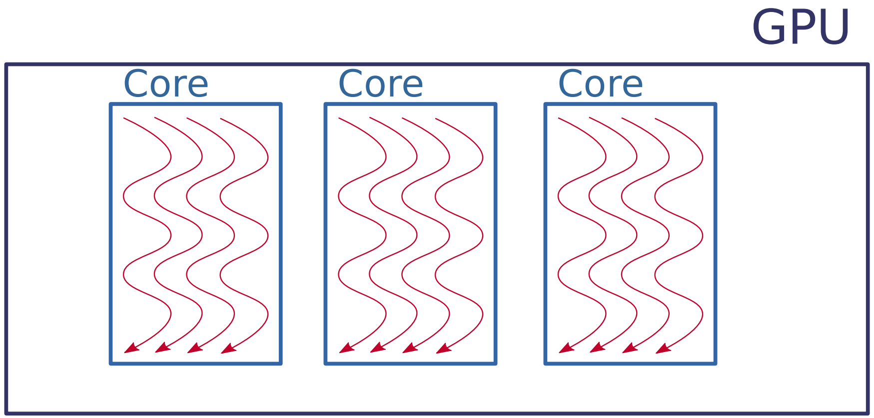
- Access memory simultaneously and contiguously
- Bandwidth more important than latency
- Floating point and fixed-function logic

→ Single instruction applied to multiple data: SIMT

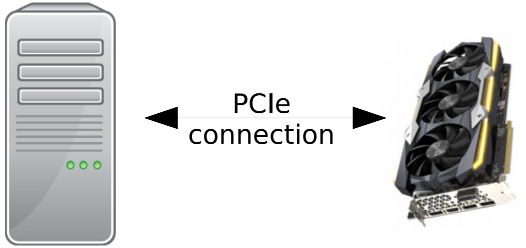


GPU in a nutshell

- Core: multiple SIMT threads grouped together
- GPU: many cores grouped together



Data transfer to a GPU



PCIe generation	16 lanes	Year
3.0	15.75 GB/s	2010
4.0	31.5 GB/s	2017

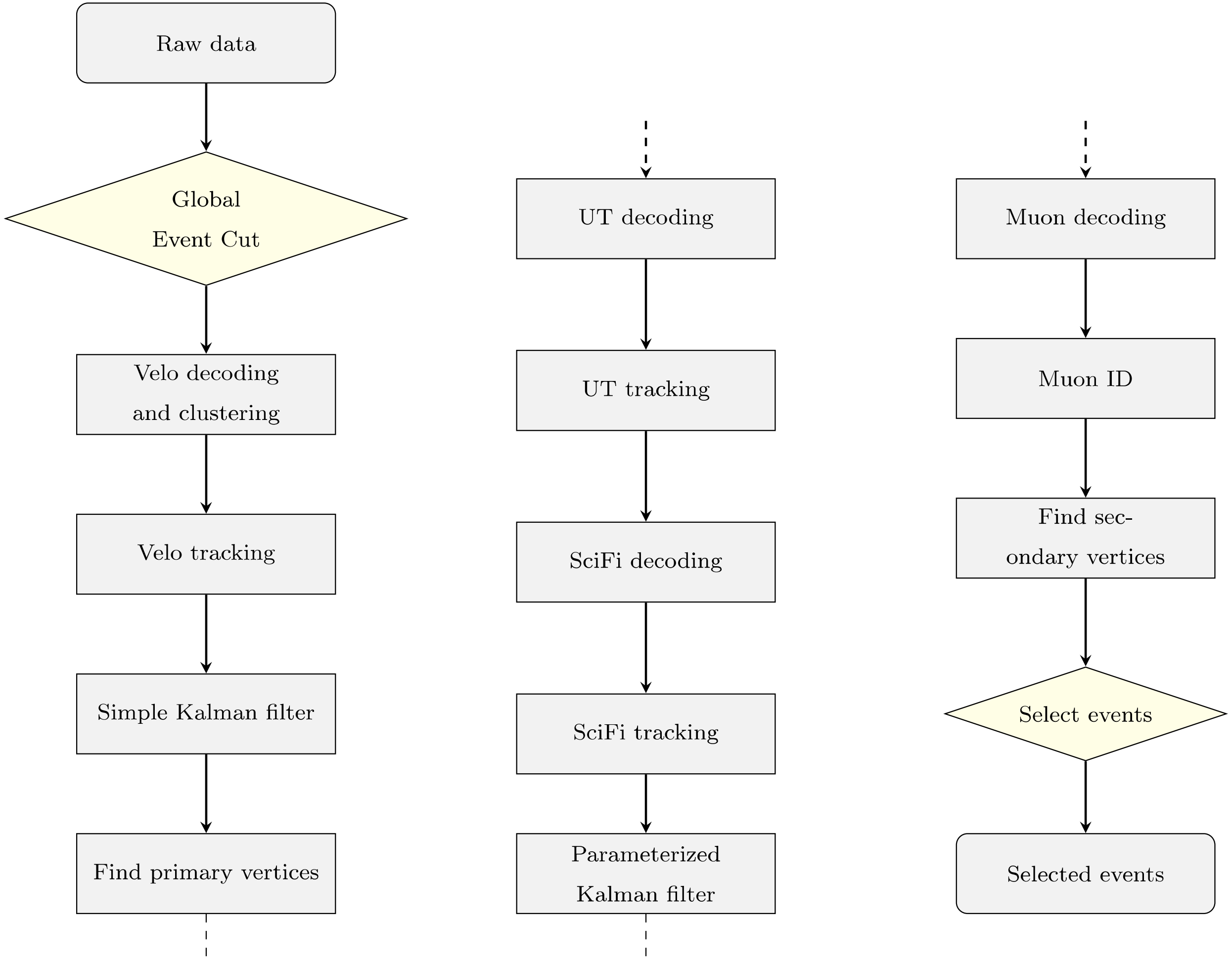
Selections

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

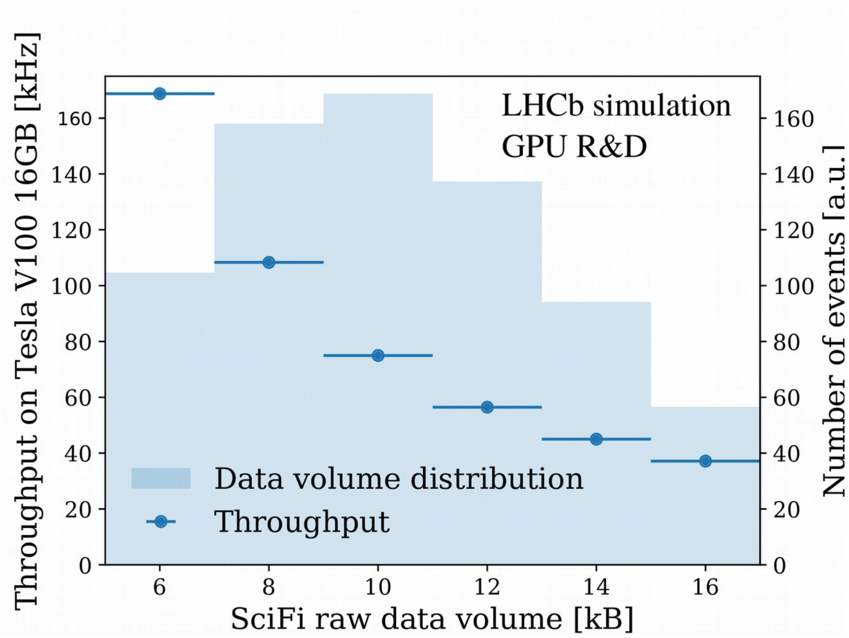
Criteria applied to signal decays in efficiency calculations

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)

HLT1 algorithms in Allen



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

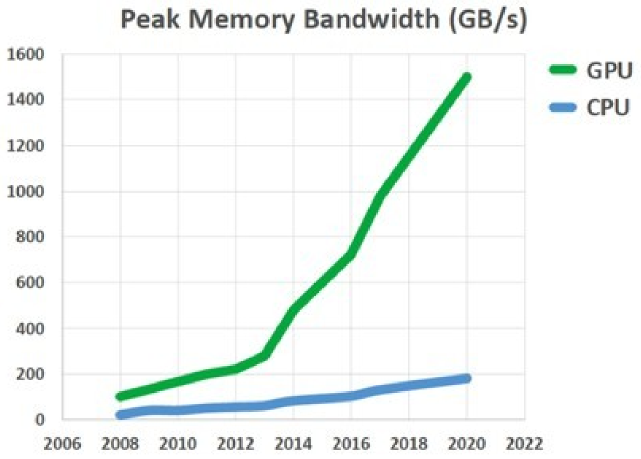
CUDA streams

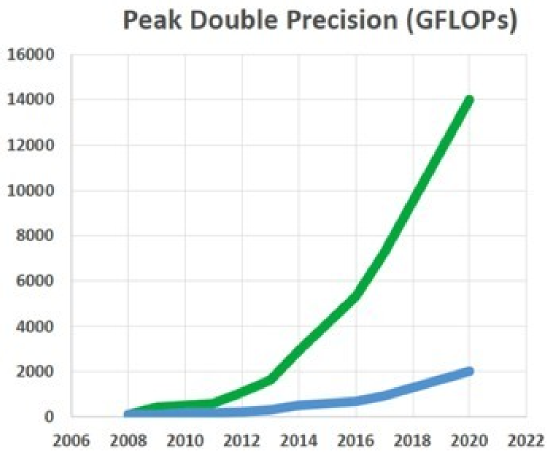
Allen settings	Threads (-t)	Memory (-m)	Number of events (-n)	Repetitions (-r)
High	12	700	1000	100
Low	2	700	1000	100

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

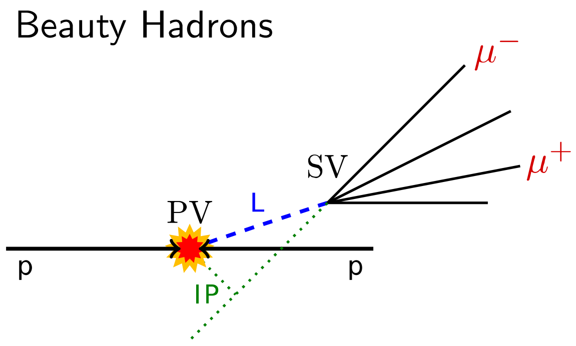
Computing costs and prospects

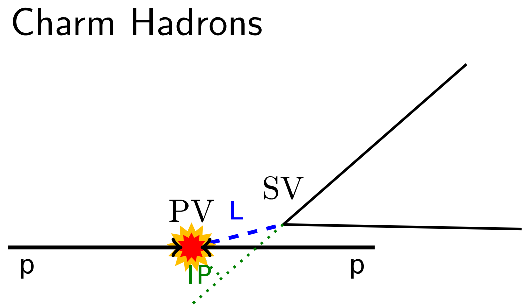
Architecture	GPU	CPU
Performance	> 60 kHz	35 kHz on 20 cores
Amount	500	1500 (12C / 24T each)
Type	RTX 2080 Ti GPUs	Intel Xeon Silver 4116 2.1 G + RAM
Price	0.5M $	2M $
Compactness (servers)	250 (< 100 with PCIe4)	750 (dual-socket)





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

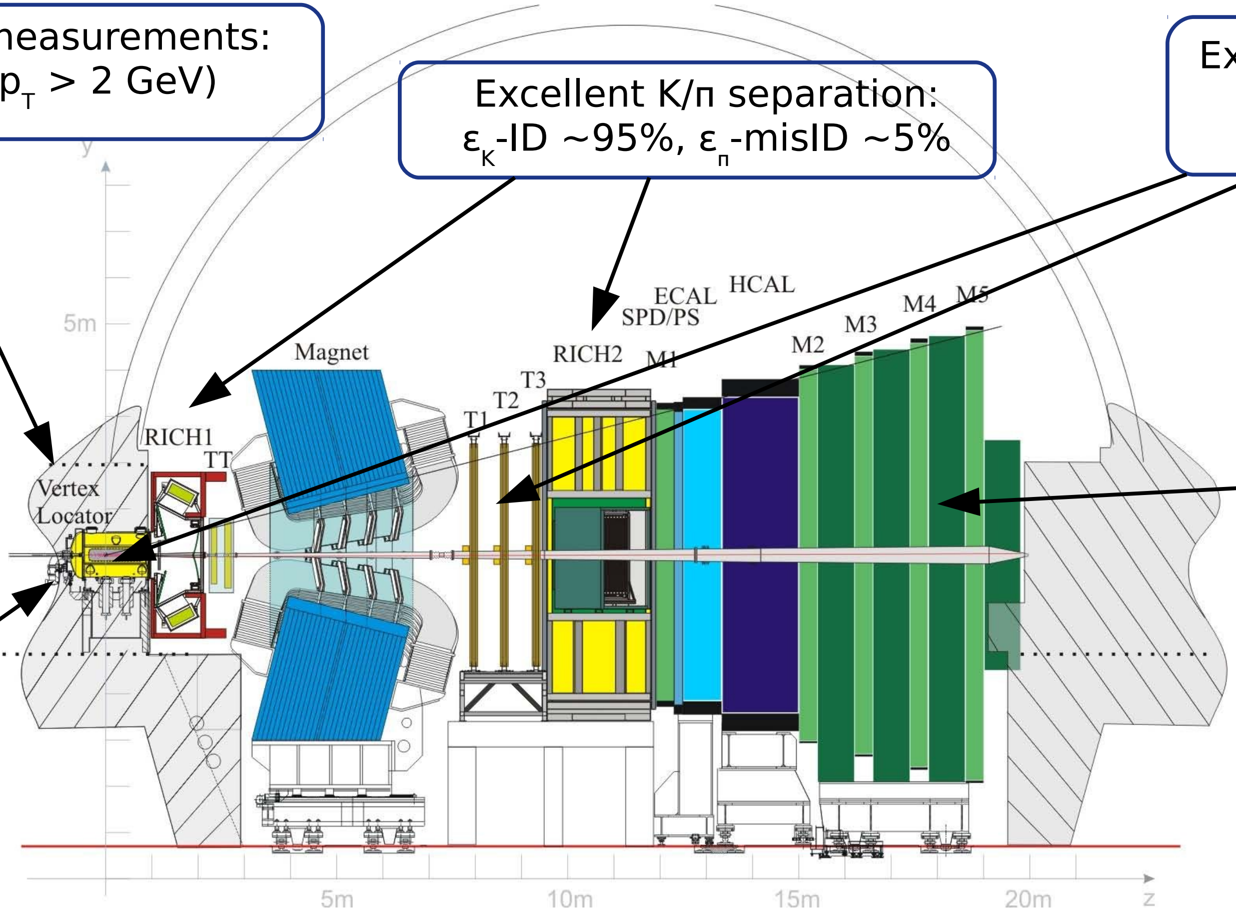
PV: Primary vertex
SV: Secondary vertex
IP: Impact parameter: distance between point of closest approach of a track and a PV

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%

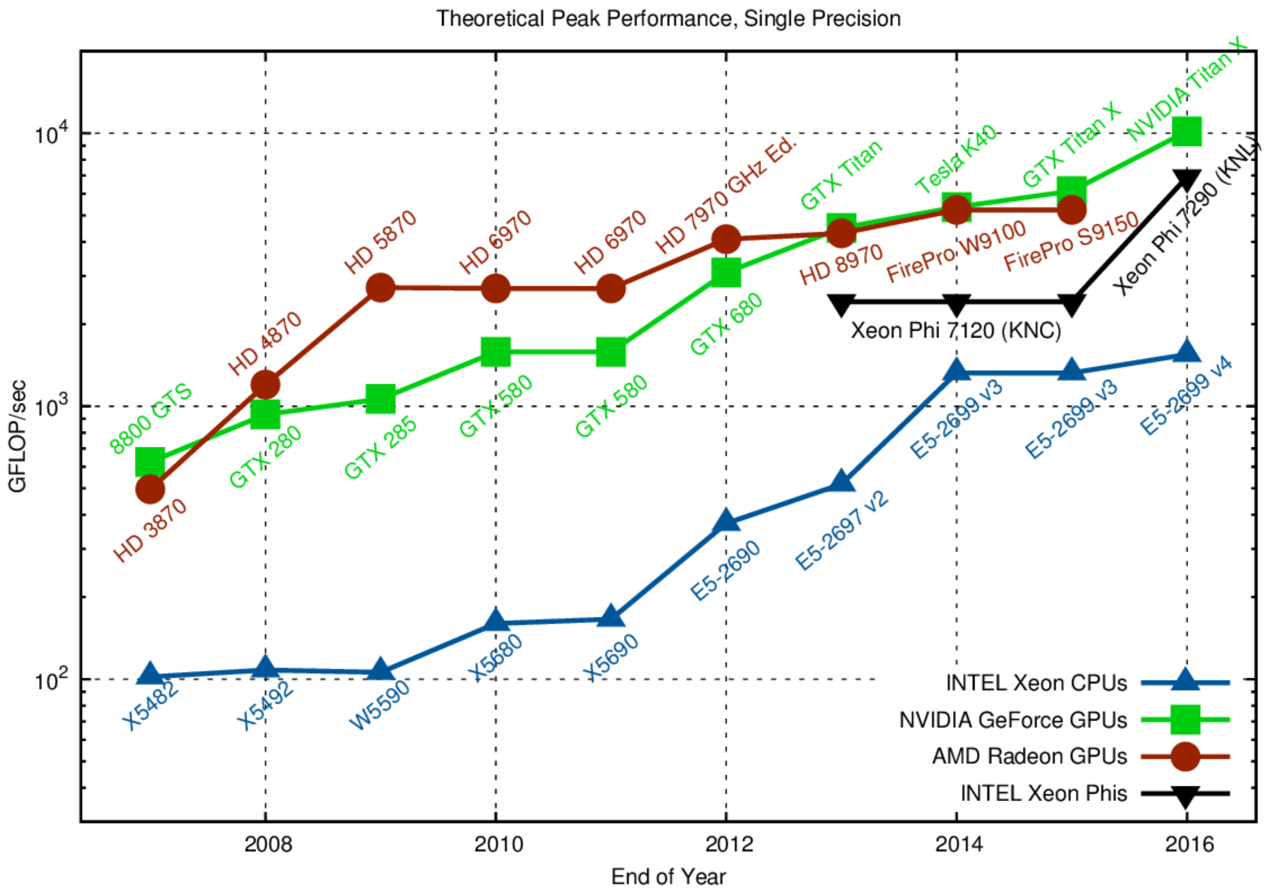
Excellent momentum resolution:
Δp/p ~0.5-1%



Excellent muon Identification:
ε_μ-ID ~97%

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

GPU performance over time



Throughput of x86 HLT1