Allen: A High Level Trigger on GPUs for LHCb

Physics and throughput performance

Dorothea vom Bruch

on behalf of the LHCb collaboration

LPNHE, CNRS

Sorbonne University, Paris Diderot University

November 6th 2019

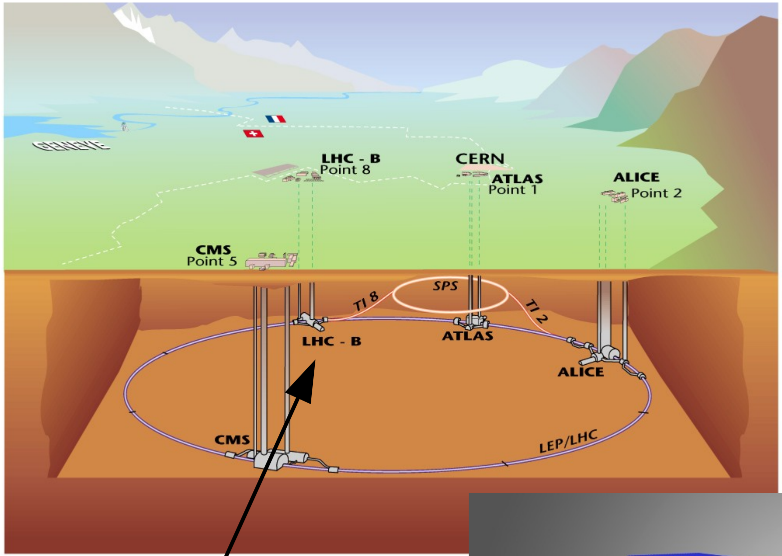
CHEP 2019, Adelaide

European Research Council

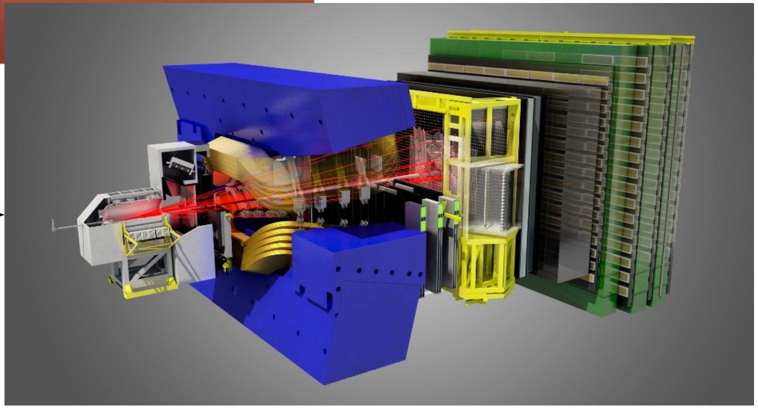
Established by the European Commission

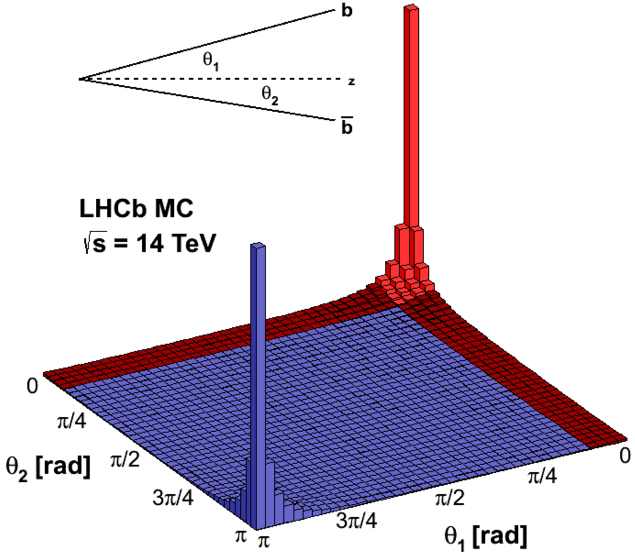
LHCb

LHC @ CERN

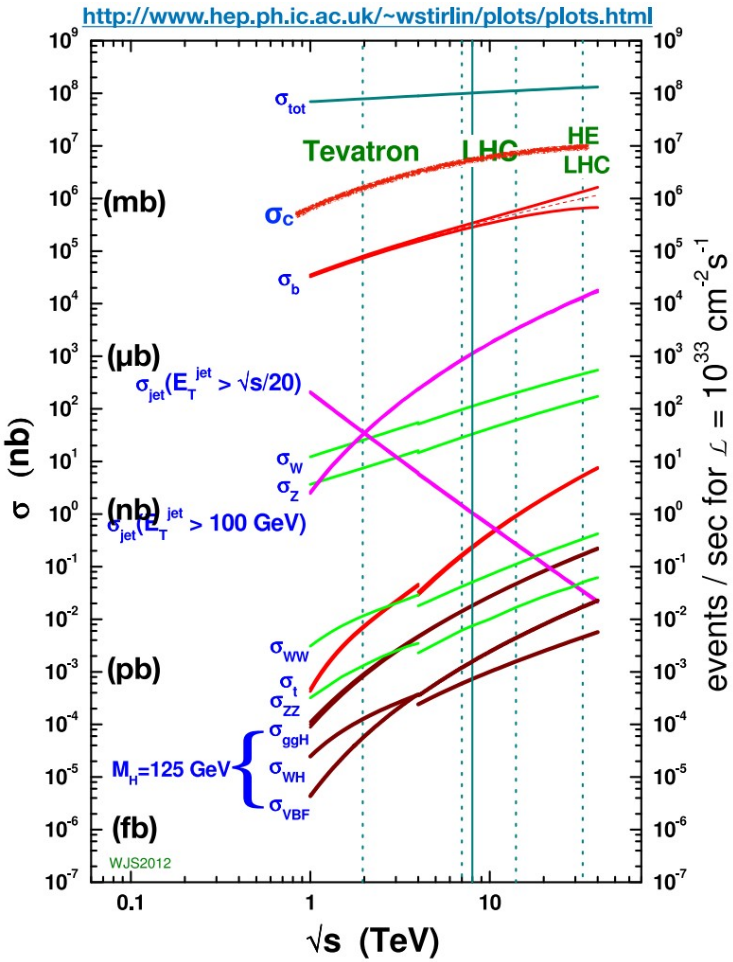


General purpose detector in the forward region specialized in beauty and charm hadrons



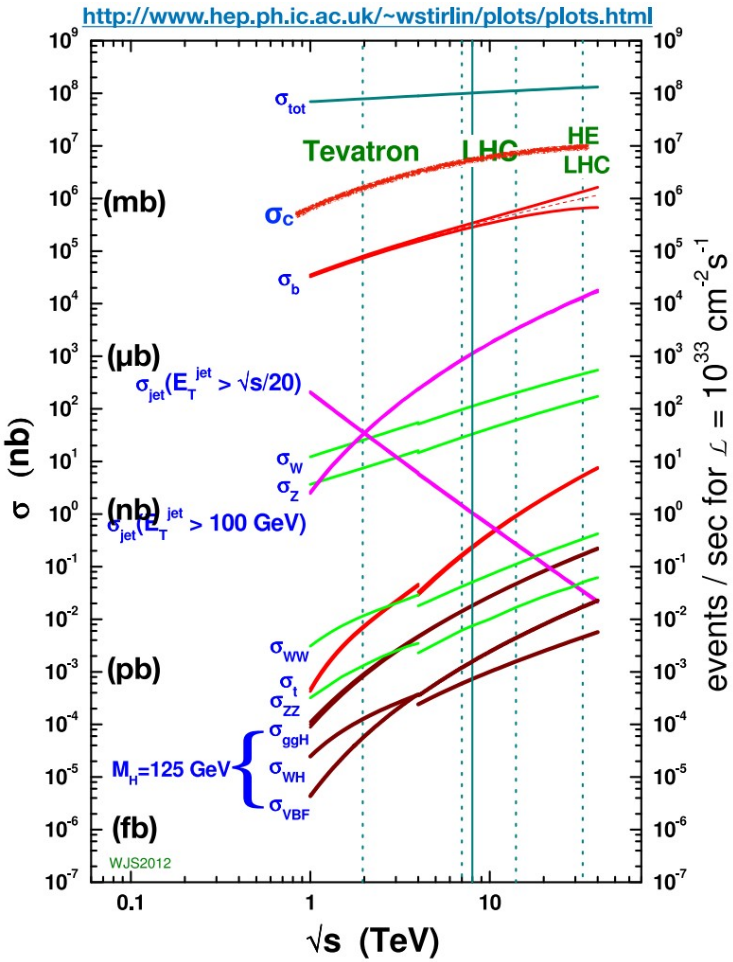


Reaching the MHz signal era



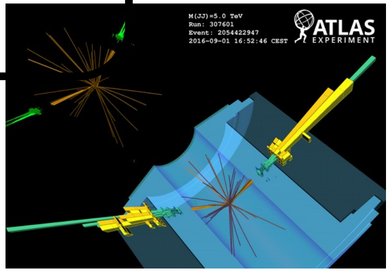
Run 3: Luminosity of $2x10^{33}$ $cm^{-2}s^{-1}$, $\sqrt{s}$ = 14 TeV

Reaching the MHz signal era

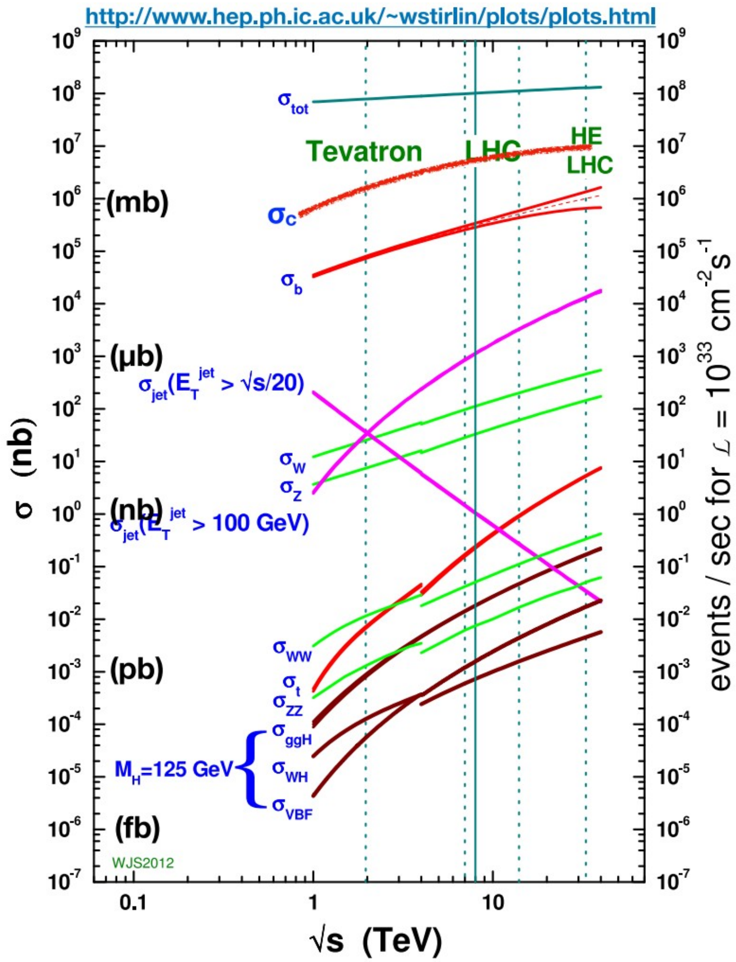


Run 3: Luminosity of 2×10^{33} $cm^{-2}s^{-1}$, $\sqrt{s}$ = 14 TeV

- General purpose LHC experiments
- Local characteristic signatures
- Can trigger efficiently at ~100 kHz
- Hardware-level trigger possible



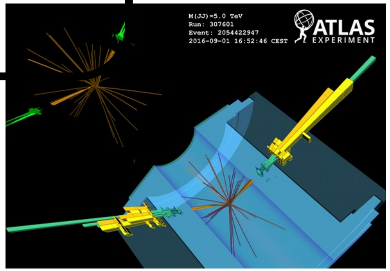
Reaching the MHz signal era



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

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

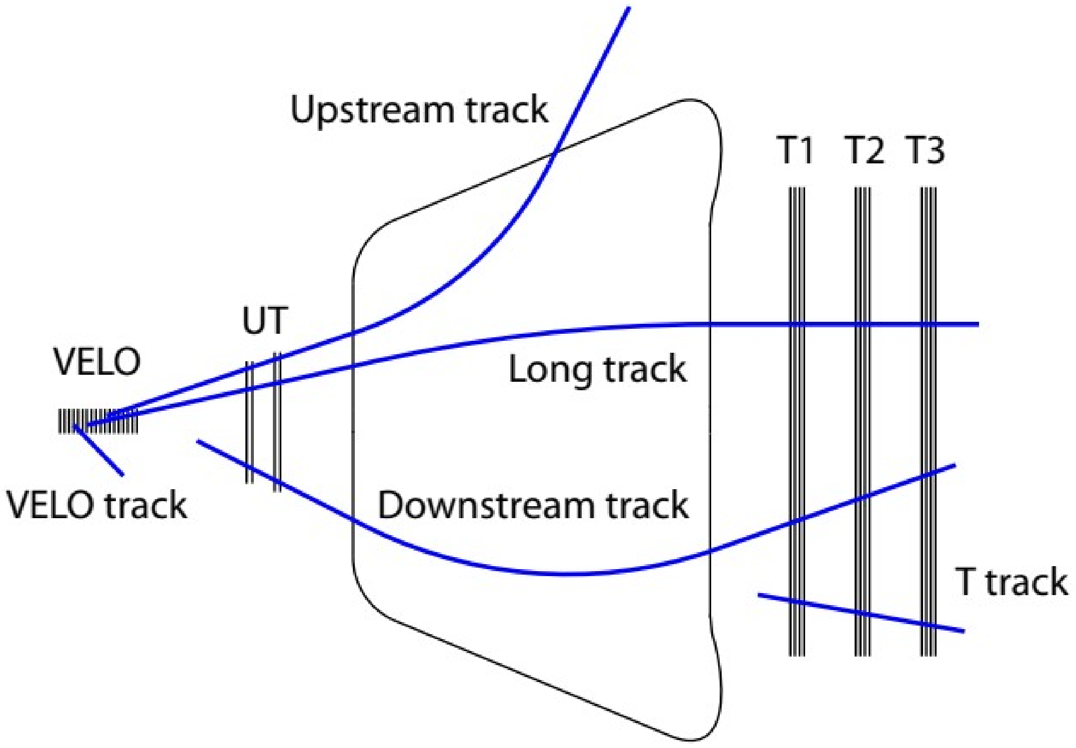
- General purpose LHC experiments
- Local characteristic signatures
- Can trigger efficiently at ~100 kHz
- Hardware-level trigger possible



Change in trigger paradigm

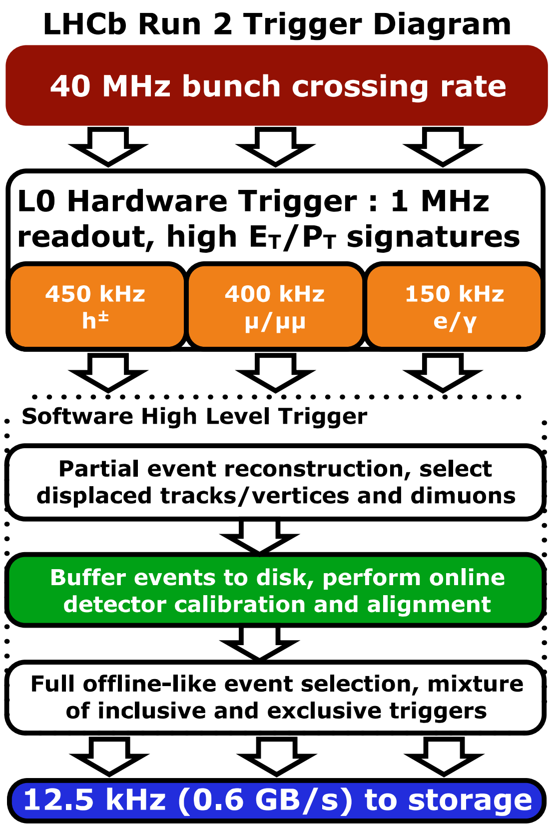
Access as much information about the collision as early as possible

Tracks in the LHCb detector

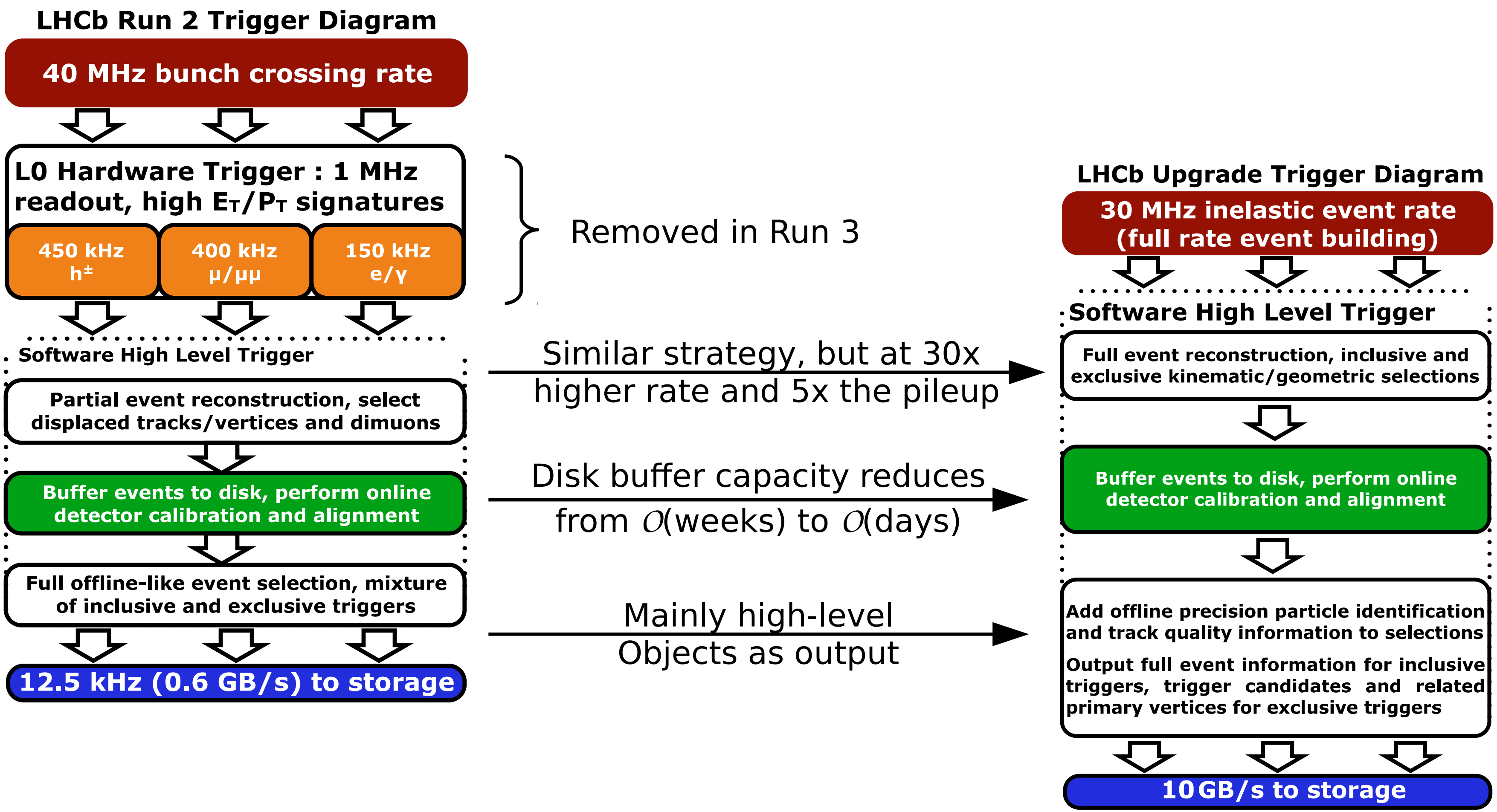


Need information from many subdetectors → read out full detector

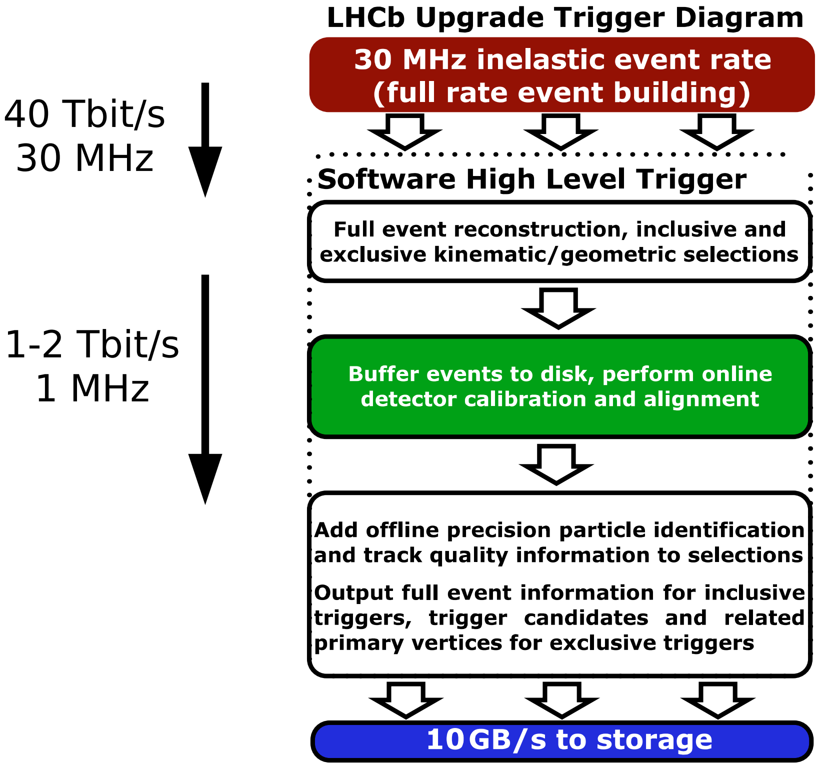
Trigger upgrade for Run 3 (2021)



Trigger upgrade for Run 3 (2021)



Trigger in Run 3 (2021)

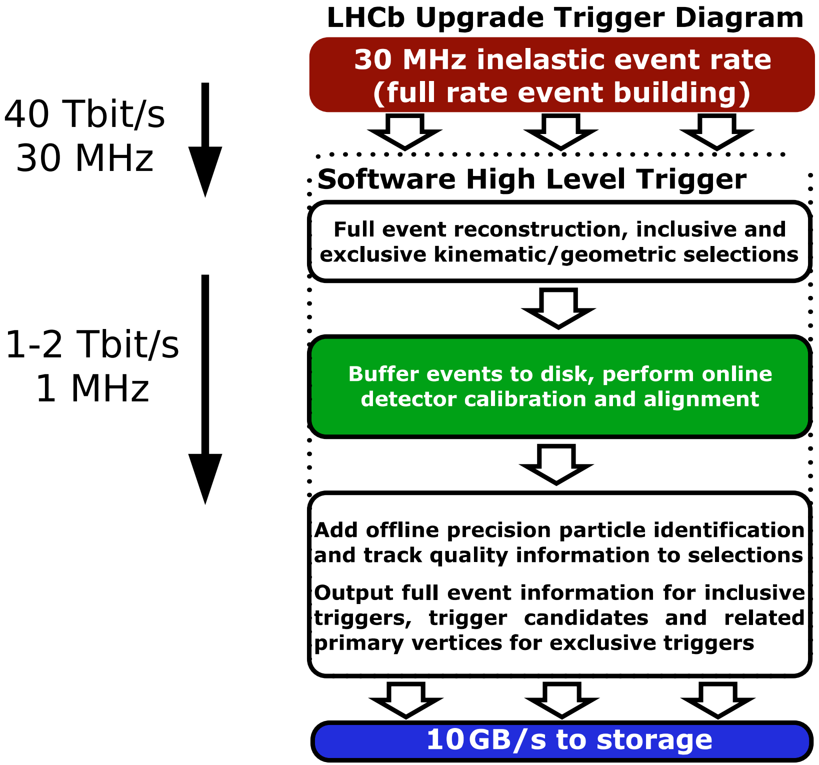


High Level Trigger 1 (HLT1)

- 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)

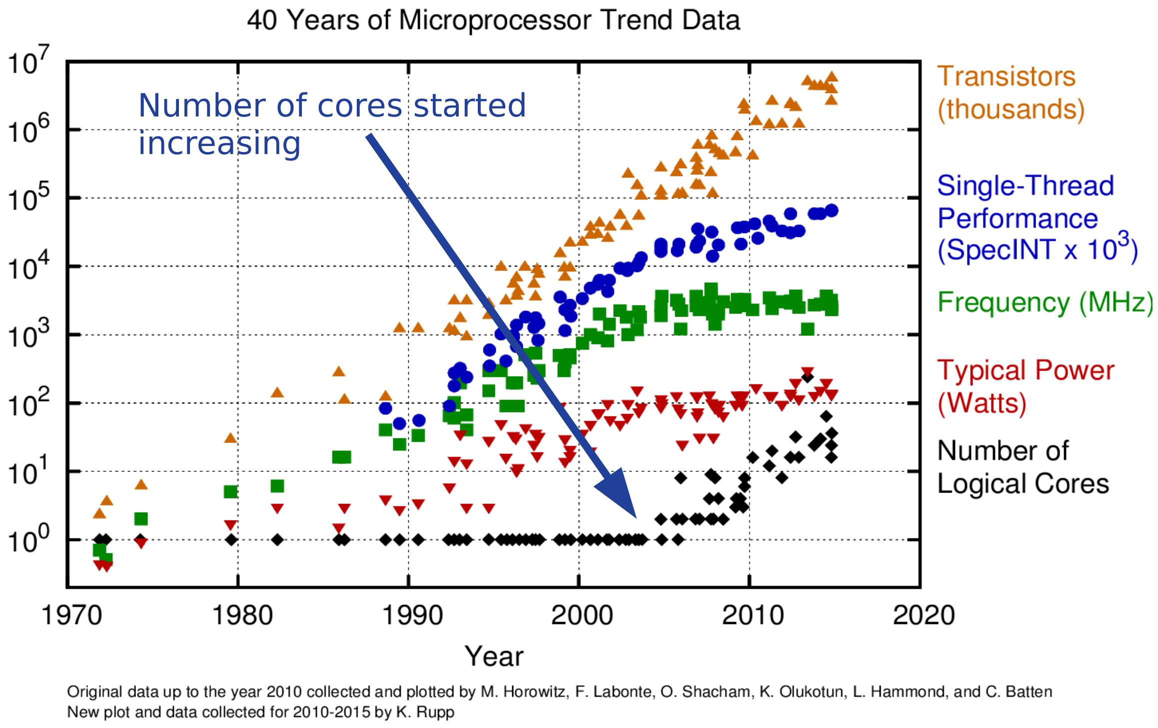


High Level Trigger 1 (HLT1)

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

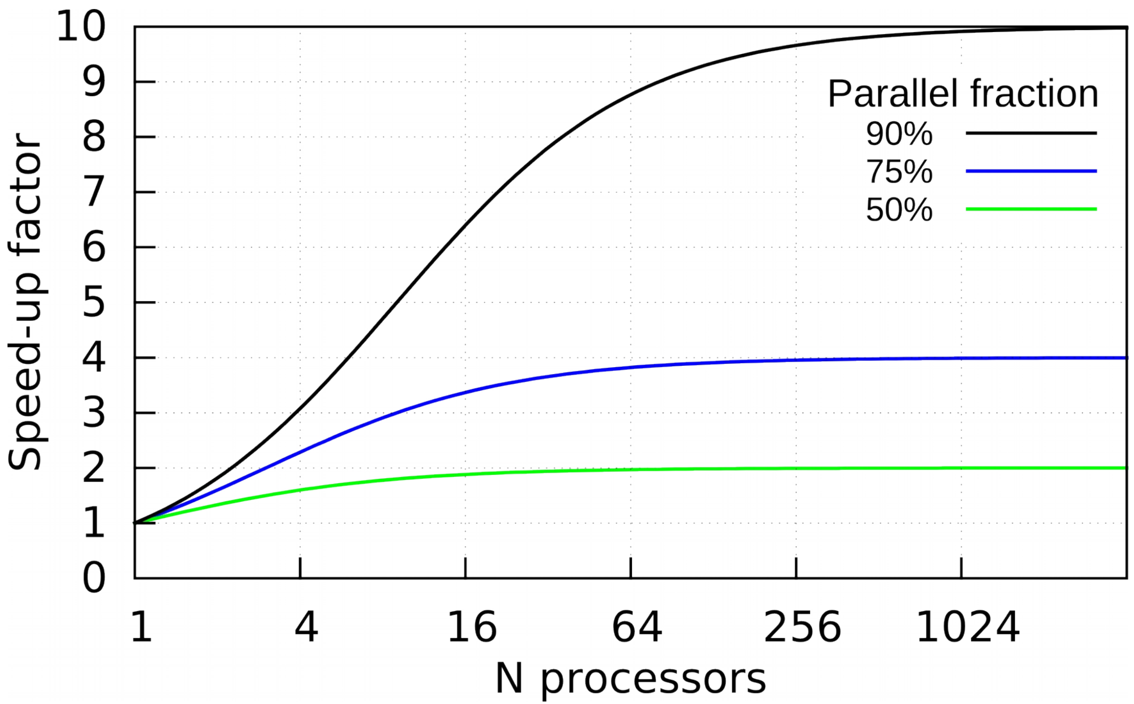
Track reconstruction @ 30 MHz is a huge computing challenge!

Architecture for high level trigger?



Graphics Processing Units (GPUs) have thousands of cores

Amdahl’s law



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

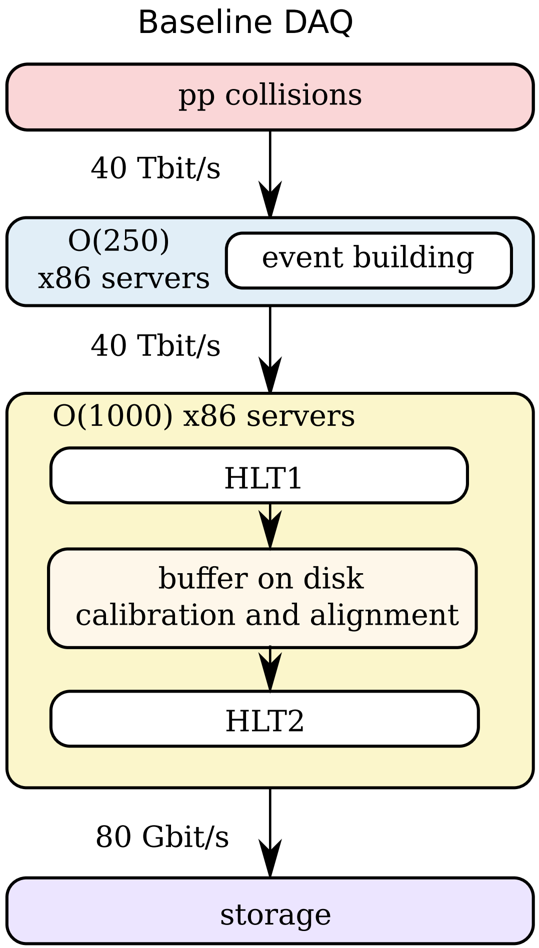
S: sequential part of program

P: parallel part of program

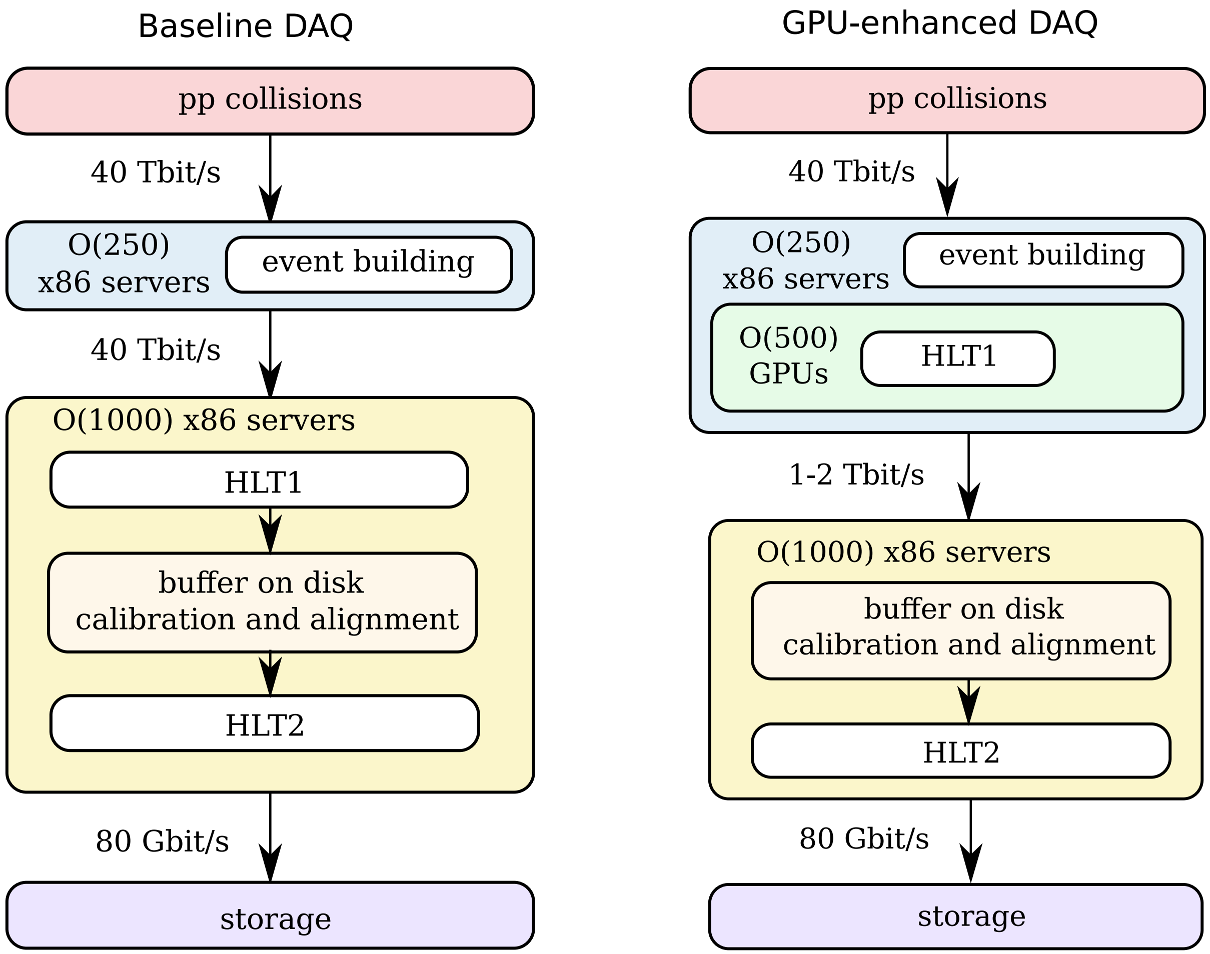
N: number of processors

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

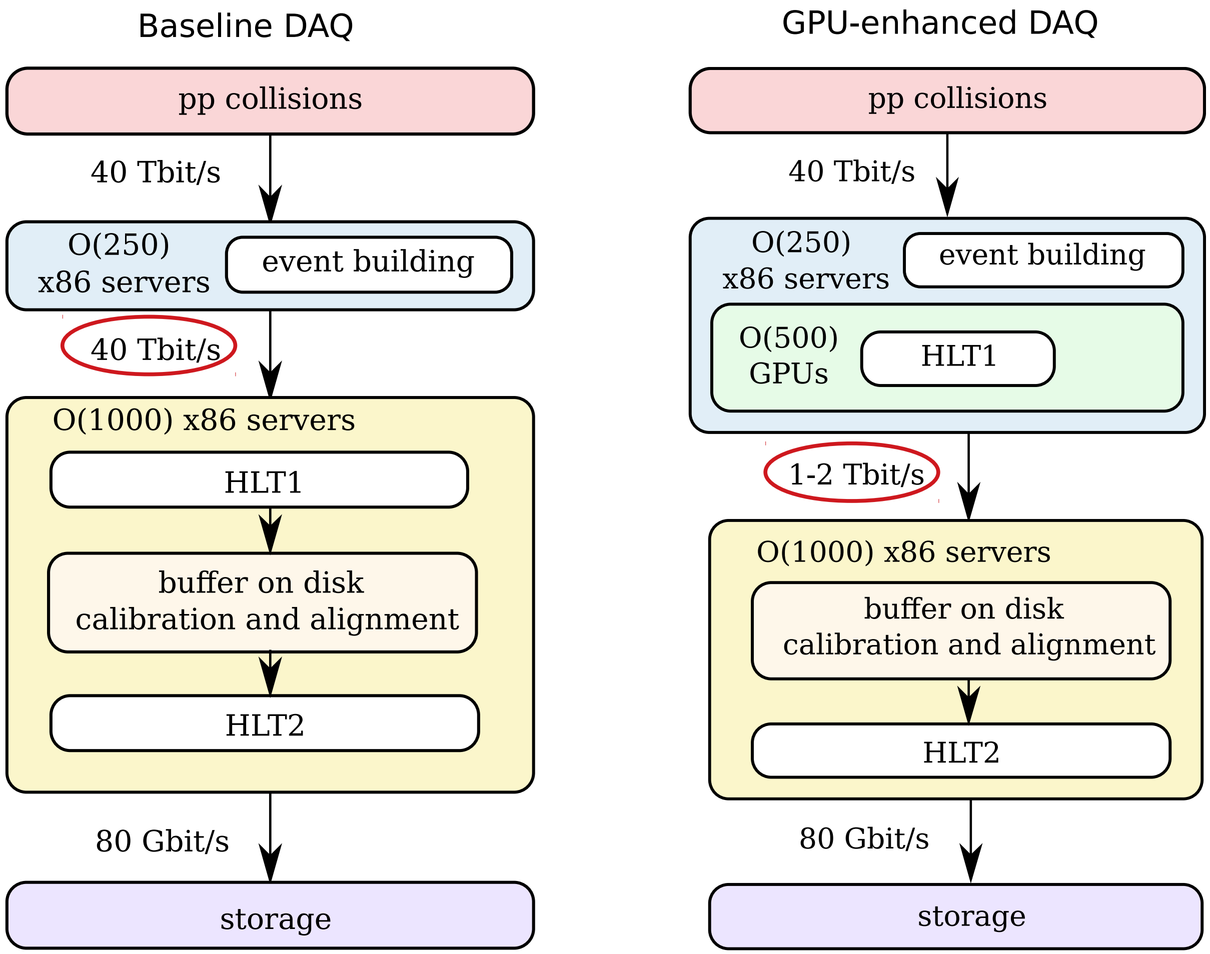
Where to place the GPUs?



Where to place the GPUs?



Where to place the GPUs?

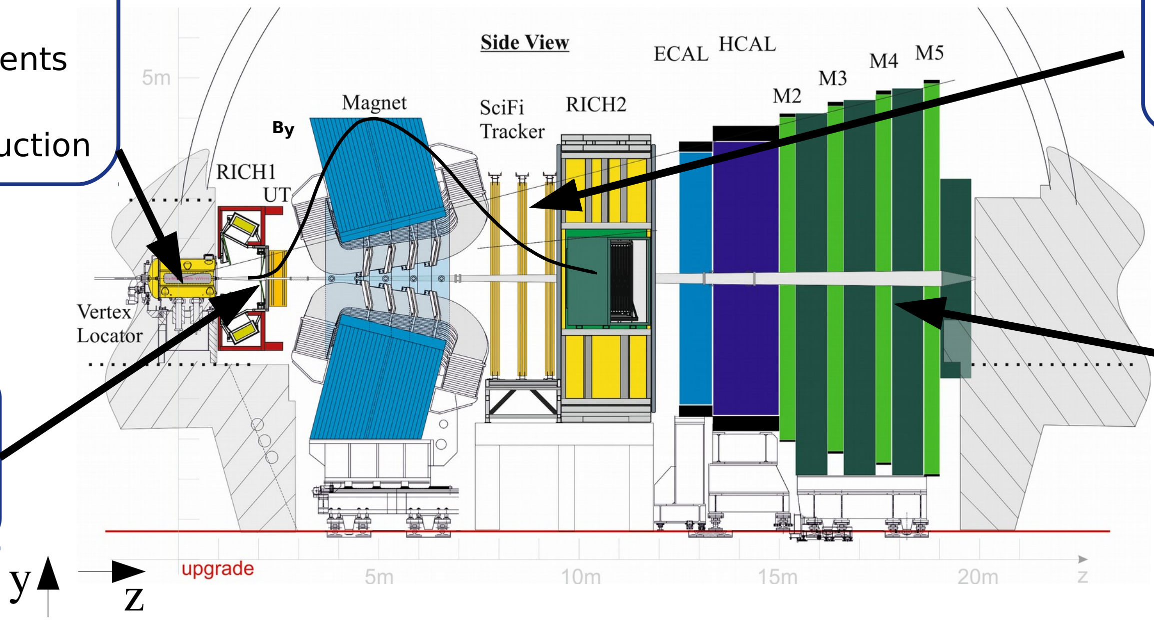


GPUs naturally integrate into LHCb’s DAQ

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	GPUs have 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	GPUs have 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	GPUs have 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 R&D project

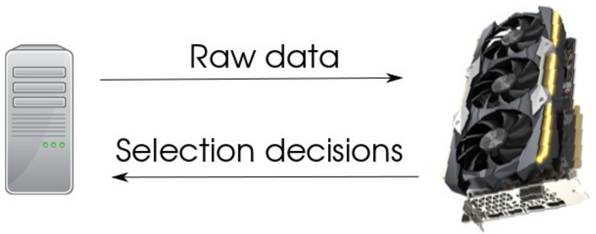
- Fully standalone software project: https://gitlab.cern.ch/lhcb/Allen
- Only requirements:
  C++17 compliant compiler, CUDA v10, boost, ZeroMQ
- Built-in physics validation
- Configurable sequence, custom memory manager
- Cross-architecture compatibility

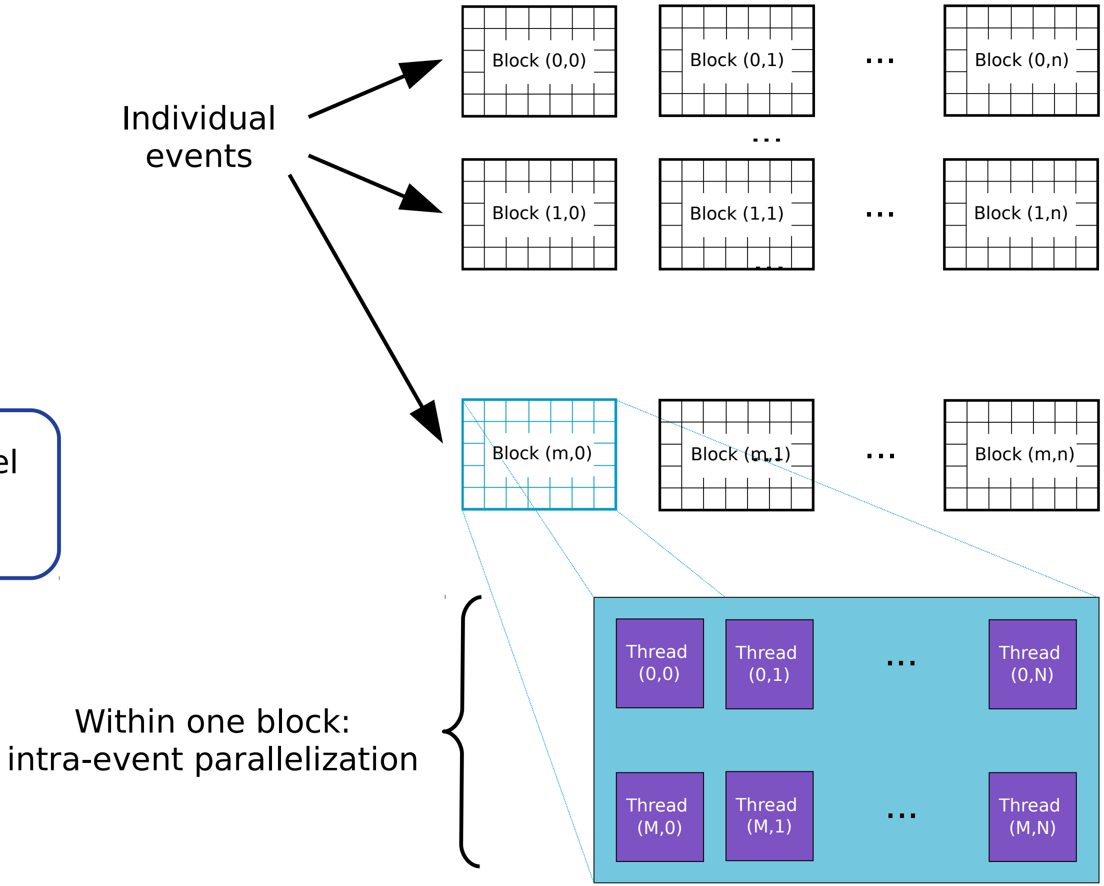
- Project started in February 2018
- After 15 months of development time:
  project reviewed as viable solution for Run 3 (starting in 2021)

- Talk on software challenges by D. Cámpora: Monday, Track 5

- Named after Frances E. Allen

HLT1 on GPUs

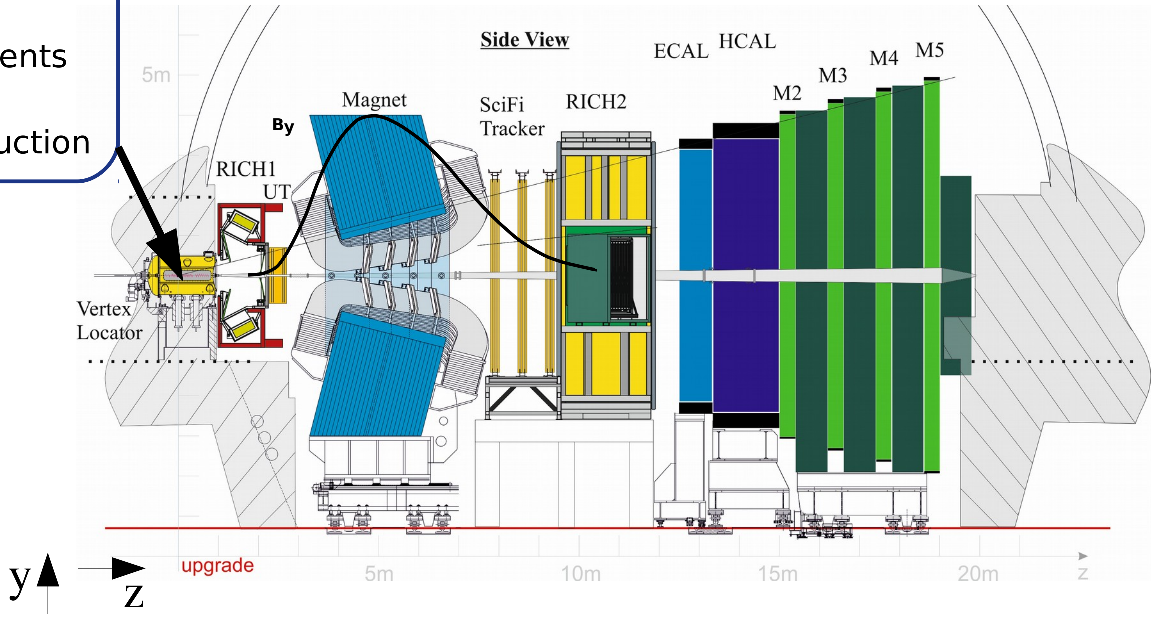




- Process thousands of events in parallel
- Single precision only

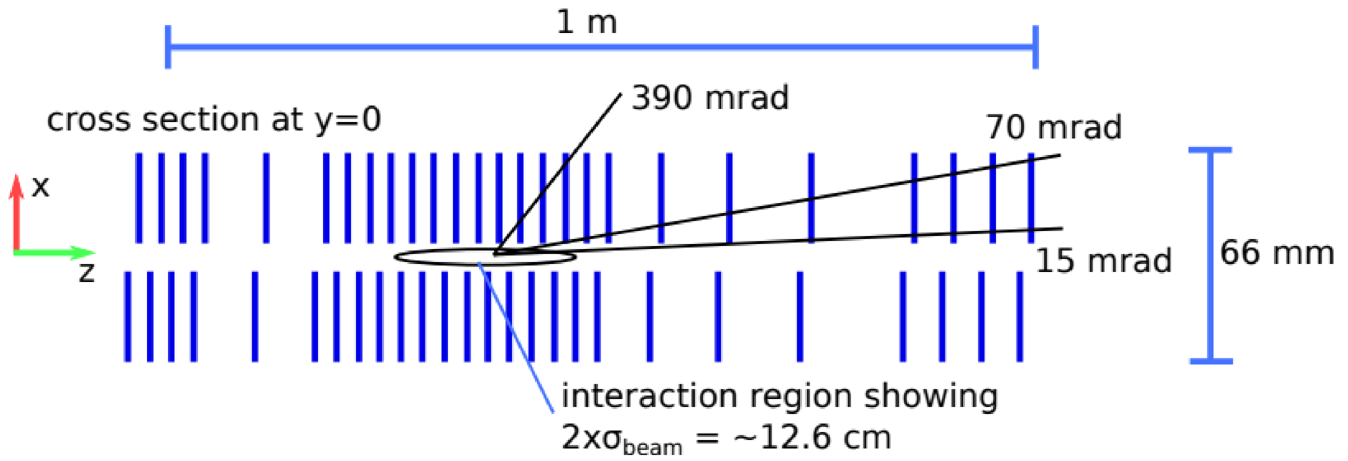
Velo detector

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

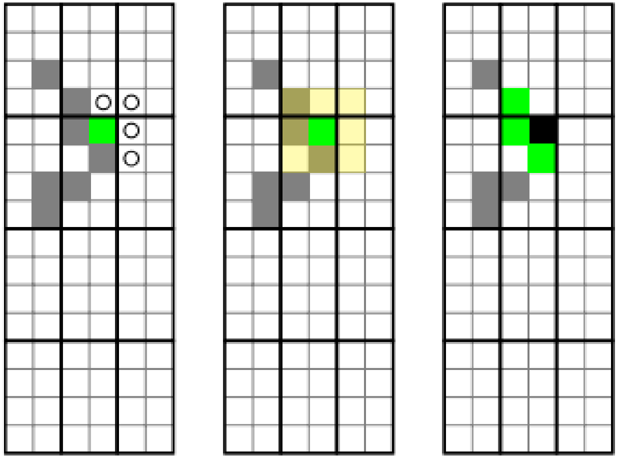


Velo detector: clustering

26 planes of silicon pixel detectors

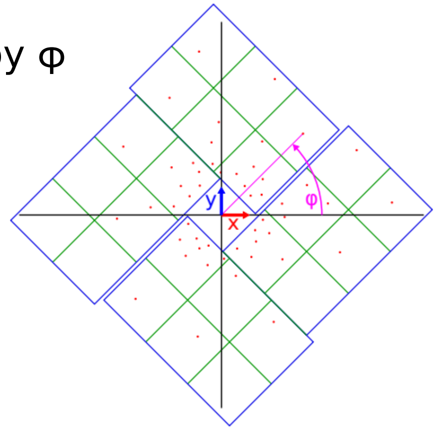


Clustering with bit masks

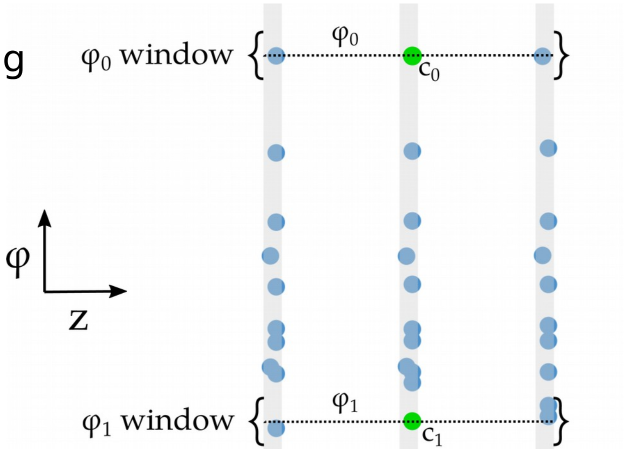


Velo detector: track reconstruction

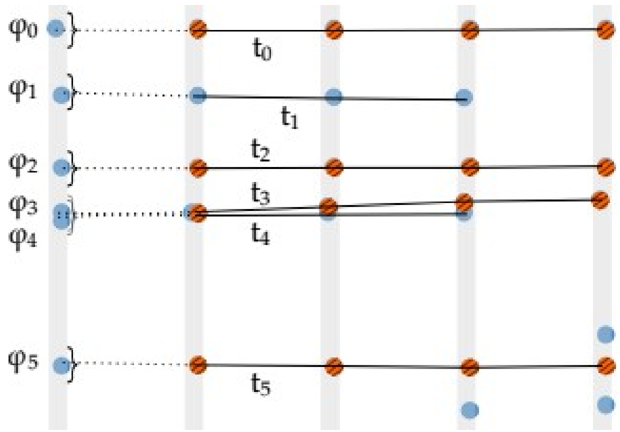
1) Sort hits by φ

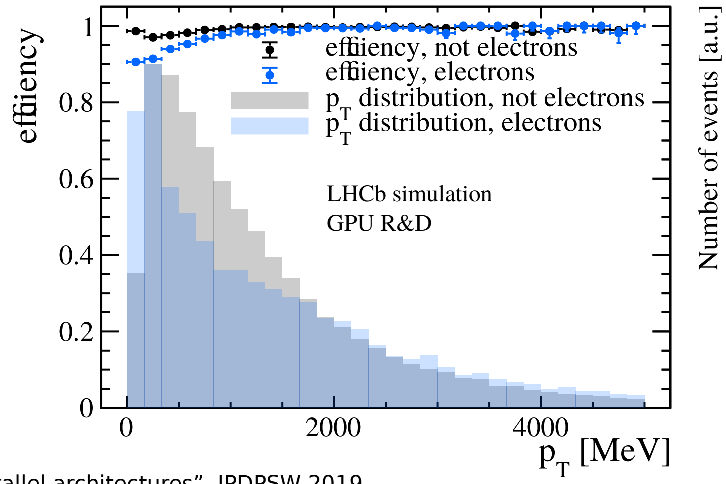


2) Triplet seeding



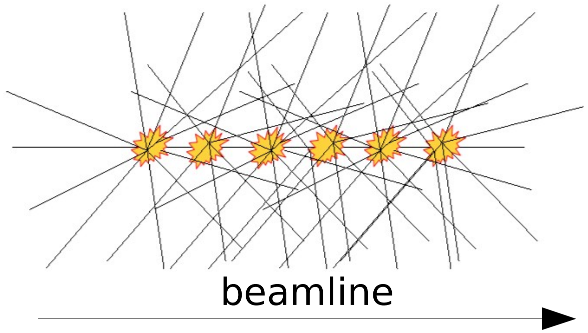
3) Triplet forwarding



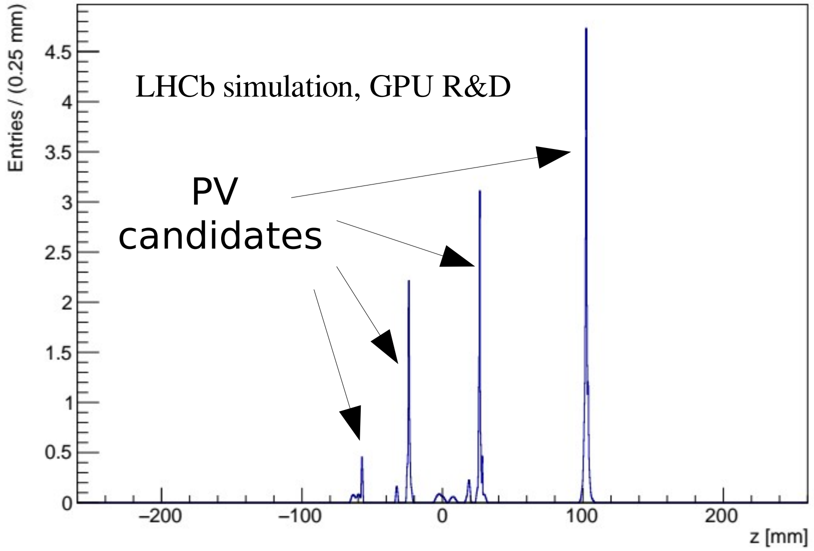


D. Campora, N. Neufeld, A. Riscos Núñez: “A fast local algorithm for track reconstruction on parallel architectures”, IPDPSW 2019

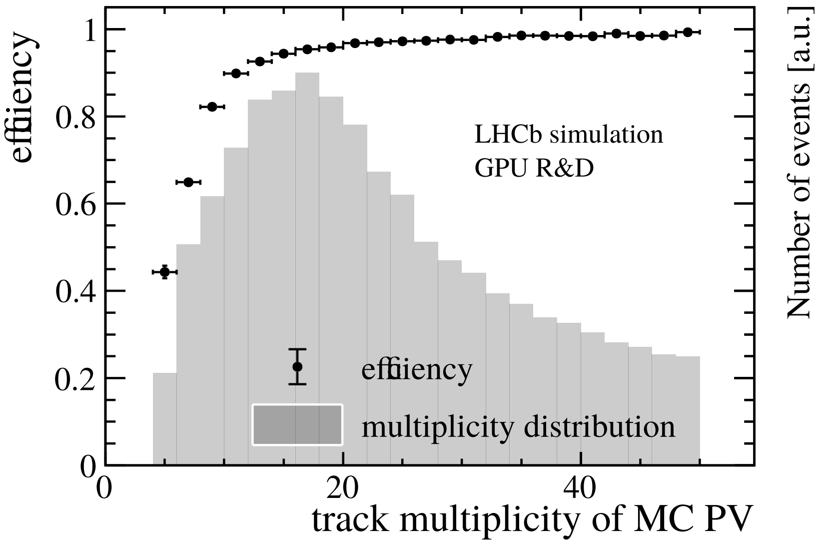
Velo detector: primary vertex reconstruction



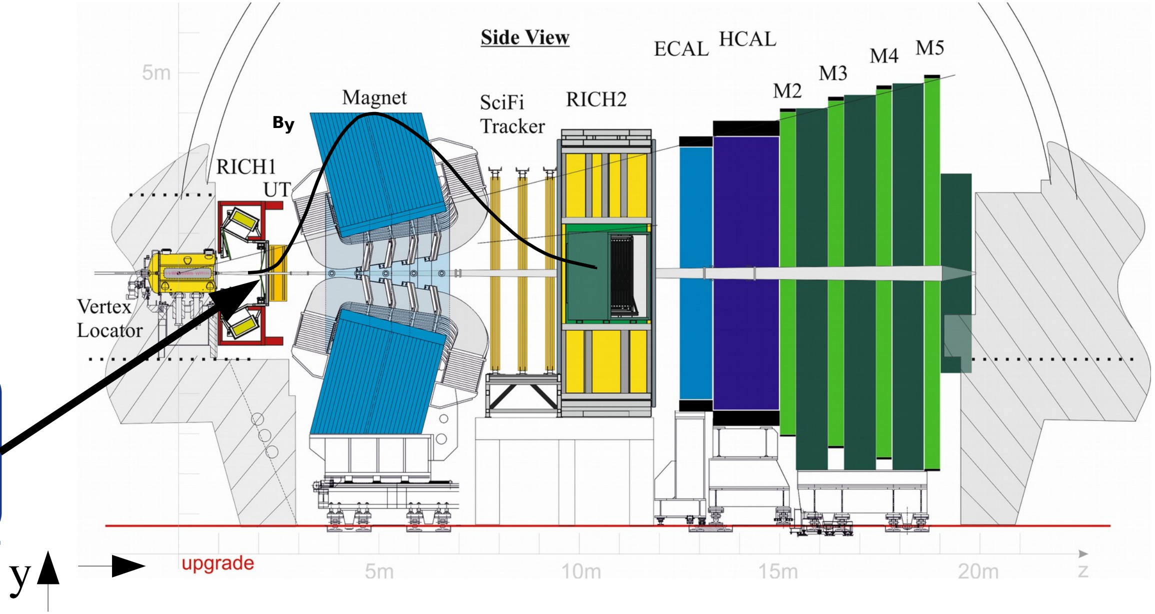
Point of closest approach of tracks to beamline



PV reconstruction efficiency



UT detector

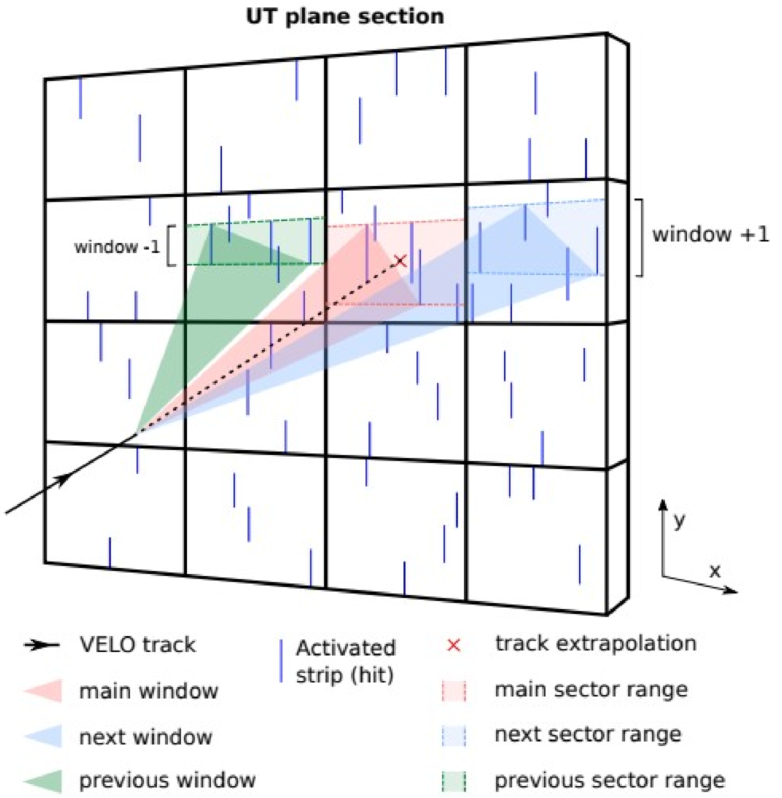


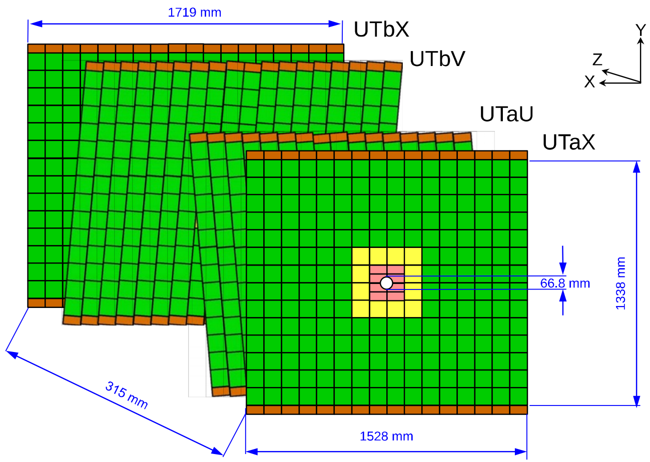
UT

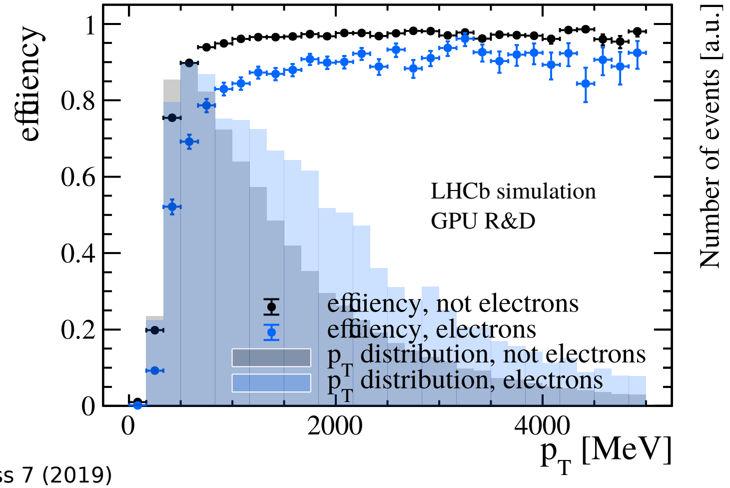
- Decode raw data
- Track reconstruction

UT detector: track reconstruction

4 planes of silicon strip detectors

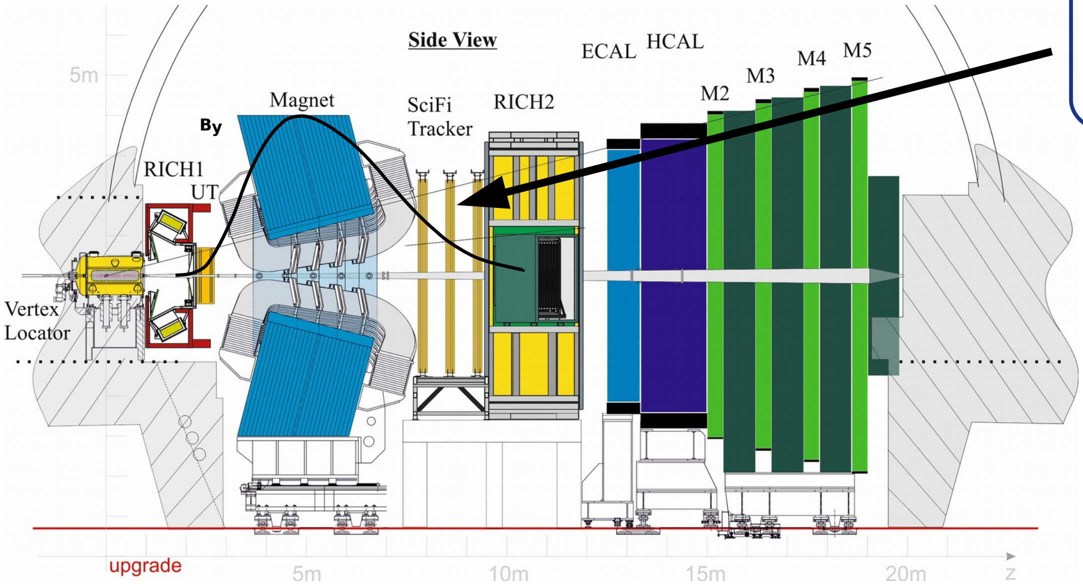






P. Fernandez Declara, D. Campora Perez, J. Garcia-Blas, D. vom Bruch, J. Daniel Garca, N. Neufeld , IEEE Access 7 (2019)

SciFi detector

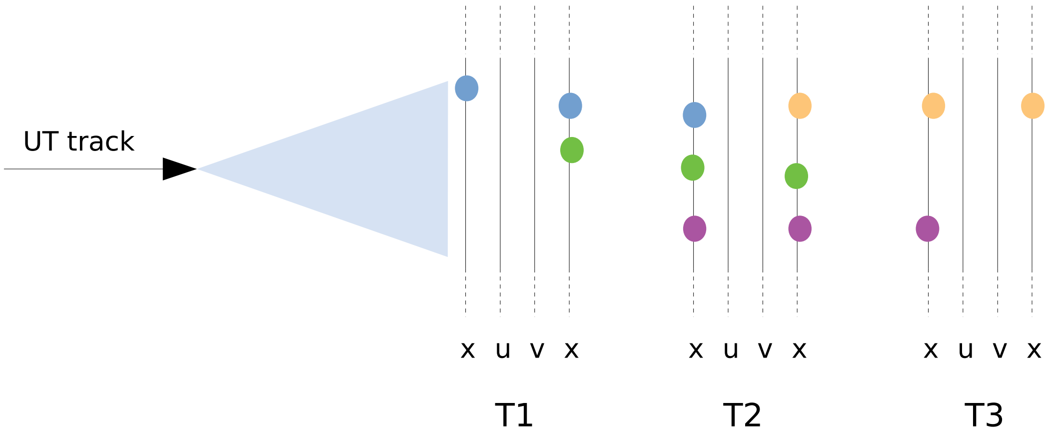


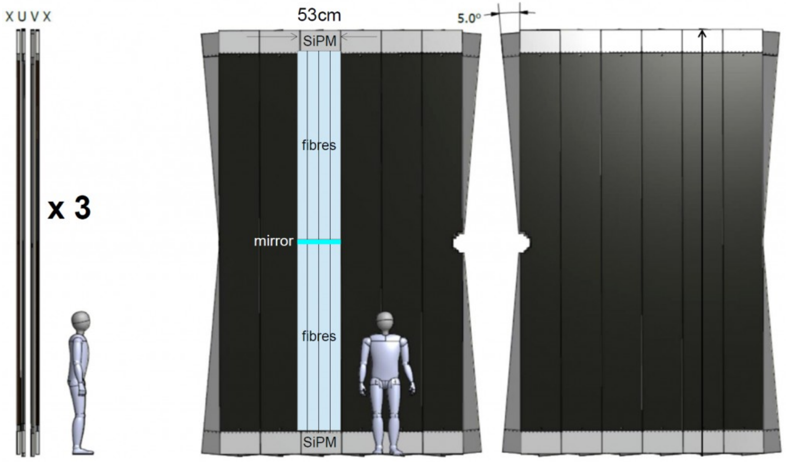
SciFi

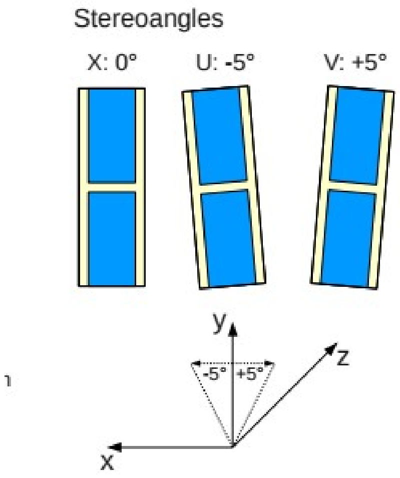
- Decode raw data
- Track reconstruction

SciFi detector

- 12 layers of scintillating fibres
- Efficiency of fibres ~98-99%
- Describe trajectories in magnetic field with parameterizations
  → no need to load large field map into GPU memory

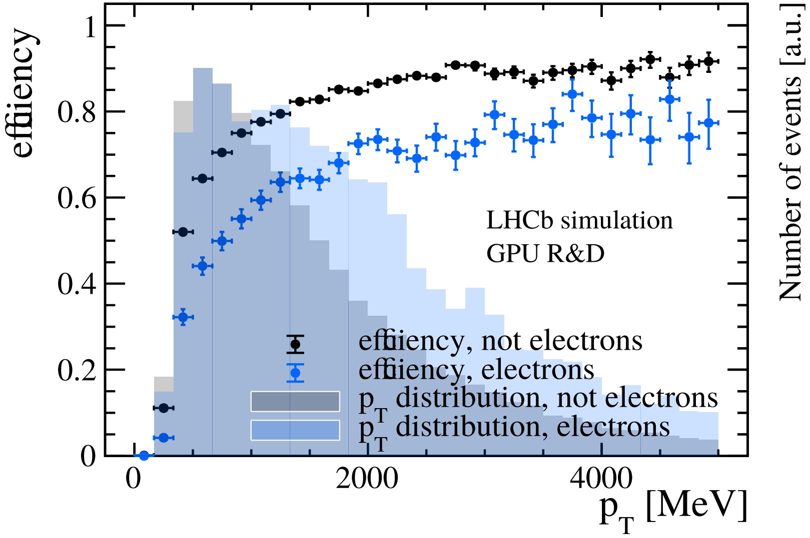




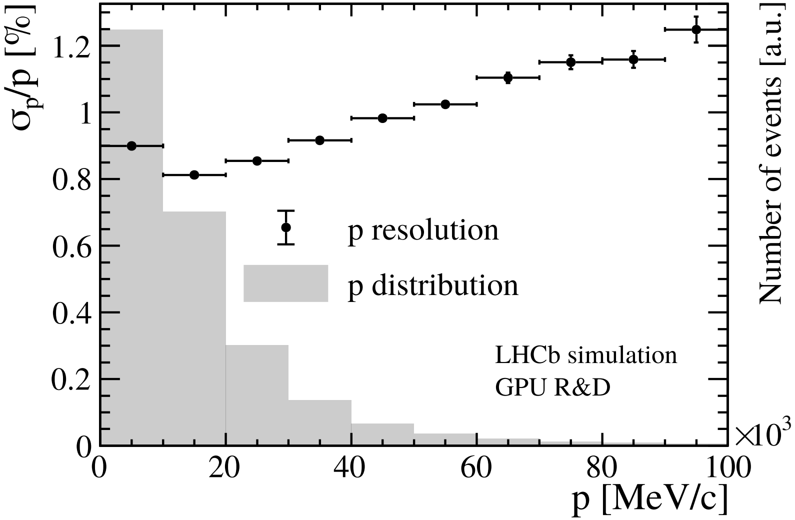


SciFi detector: track reconstruction

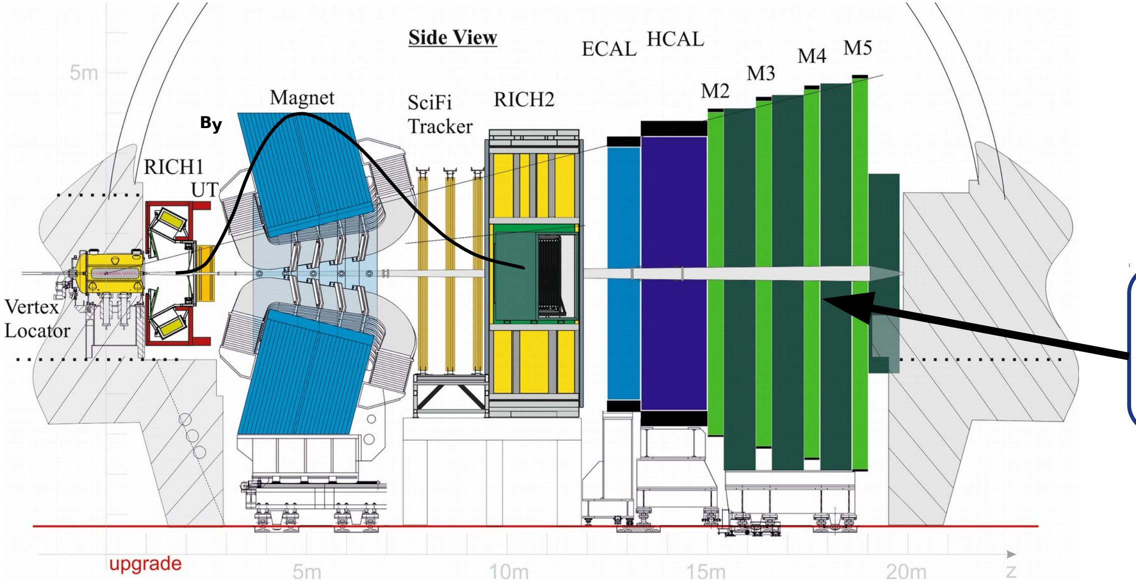
Track reconstruction efficiency for tracks originating from B decays



Momentum resolution



Muon chambers



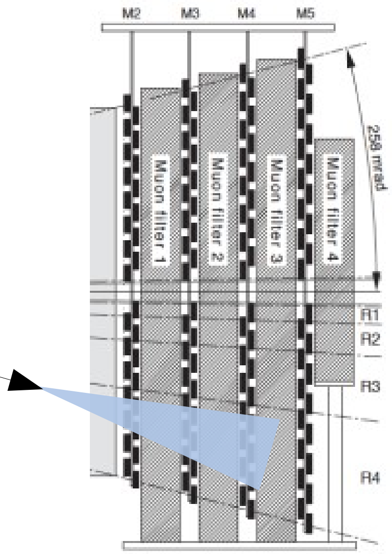
Muons

- Decode raw data
- Match hits to tracks

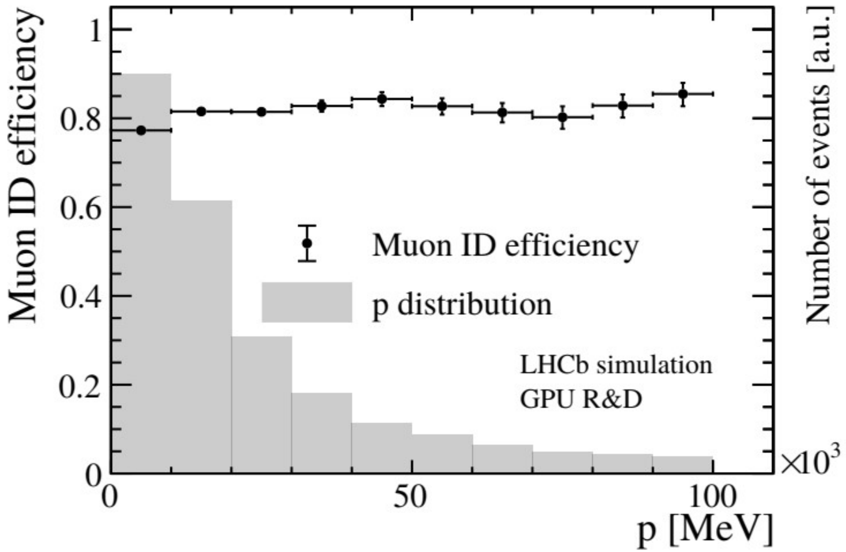
Muon identification

Four multi-wire proportional chambers
Interleaved with iron walls

SciFi track

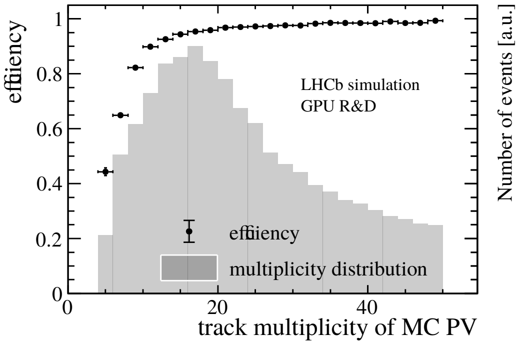


Muon identification efficiency

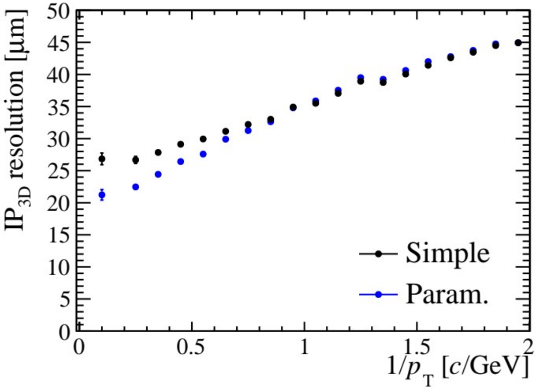


Ingredients for selections

Primary vertices



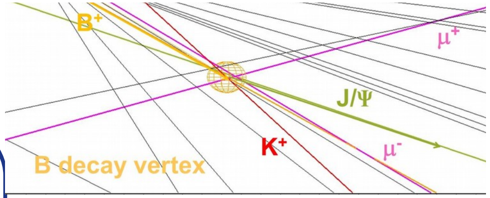
Impact parameter



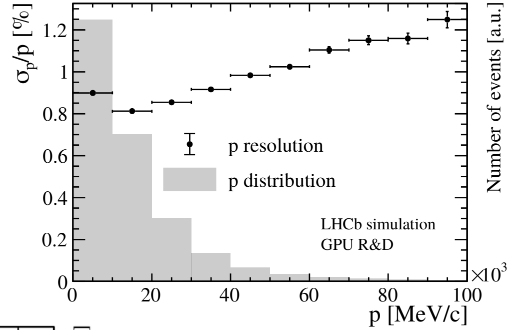
Selections

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

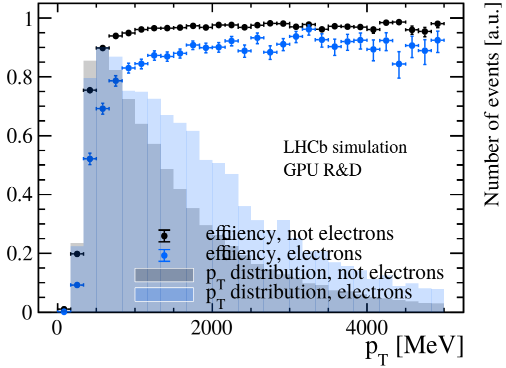
Secondary vertices



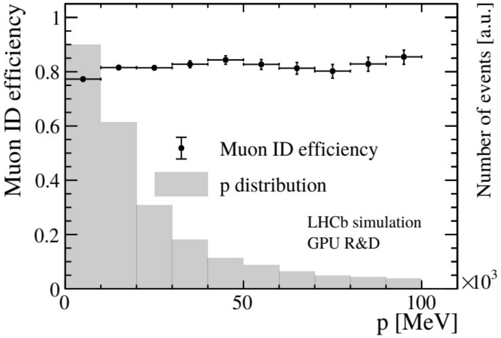
Momentum



Tracks



Muon identification



Event selection

Trigger	Rate [kHz]
1-Track	249 ±18
2-Track	663 ±30
High-p_T muon	1 ±1
Displaced dimuon	50 ±8
High-mass dimuon	101 ±12
Total	971 ±36

Event rate reduced from 30 MHz to 1 MHz

Selection efficiencies, values given in %

Signal	GEC	TIS -OR- TOS	TOS	GEC $\times$ TOS
$B^0 \to K^{*0}\mu^+\mu^-$	89 ±2	85 ±2	78 ±3	69 ±3
$B^0 \to K^{*0}e^+e^-$	84 ±3	69 ±4	62 ±4	53 ±3
$B_s^0 \to \phi\phi$	83 ±3	70 ±3	65 ±4	54 ±3
$D_s^+ \to K^+K^-\pi^+$	82 ±4	62 ±5	38 ±5	32 ±4
$Z \to \mu^+\mu^-$	78 ±1	97 ±1	97 ±1	75 ±1

GEC: Global event cut
TIS: Trigger independent from signal
TOS: Trigger on signal

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

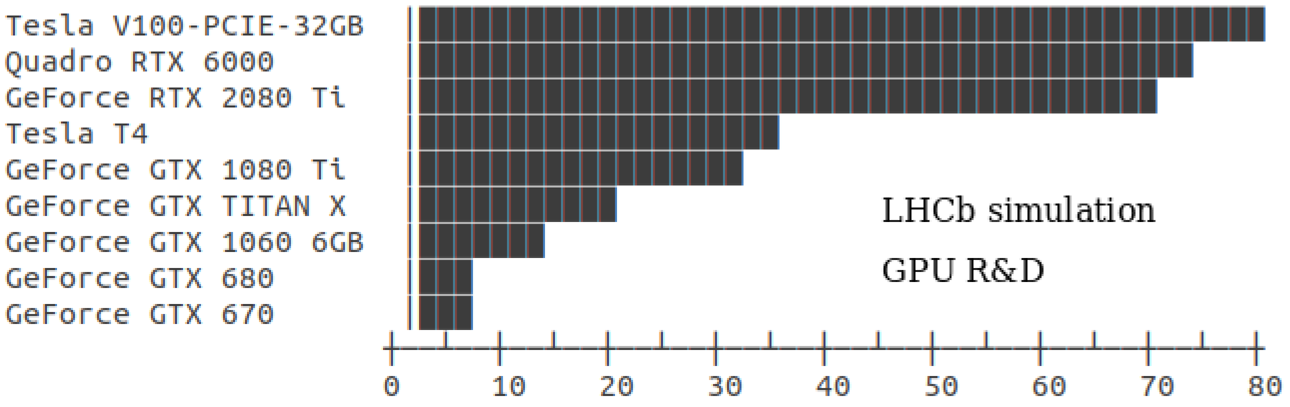
Full HLT1 running on GPUs

Physics performance matches HLT1 requirements

What about the throughput performance?

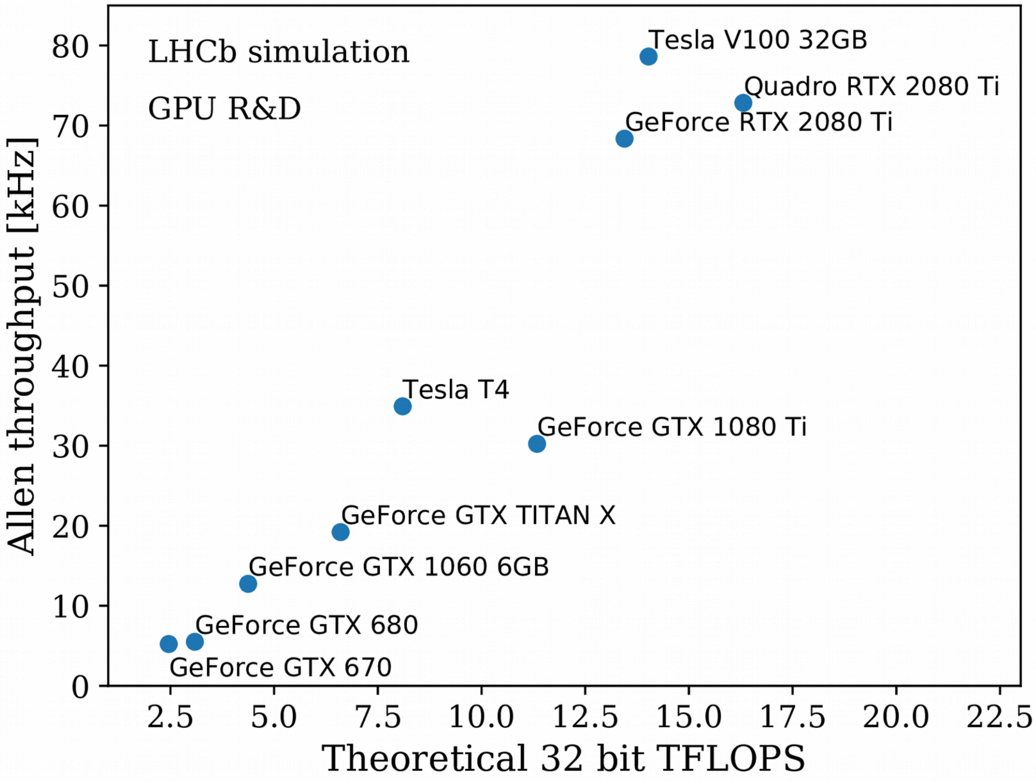
Throughput on various GPUs

Throughput of the full HLT1 sequence



HLT1 can run on 500 GPUs
→ Buy GPUs instead of expensive network

Allen scalability with GPU model



The Allen team

LPNHE Paris

Yandex

Universidad de Sevilla

USC Universidade de Santiago de Compostela

Universidad Carlos III de Madrid

Massachusetts Institute of Technology

University of Cincinnati

Nikhef

technische universität dortmund

CERN

UFRJ

DKE

Higher School of Economics National Research University

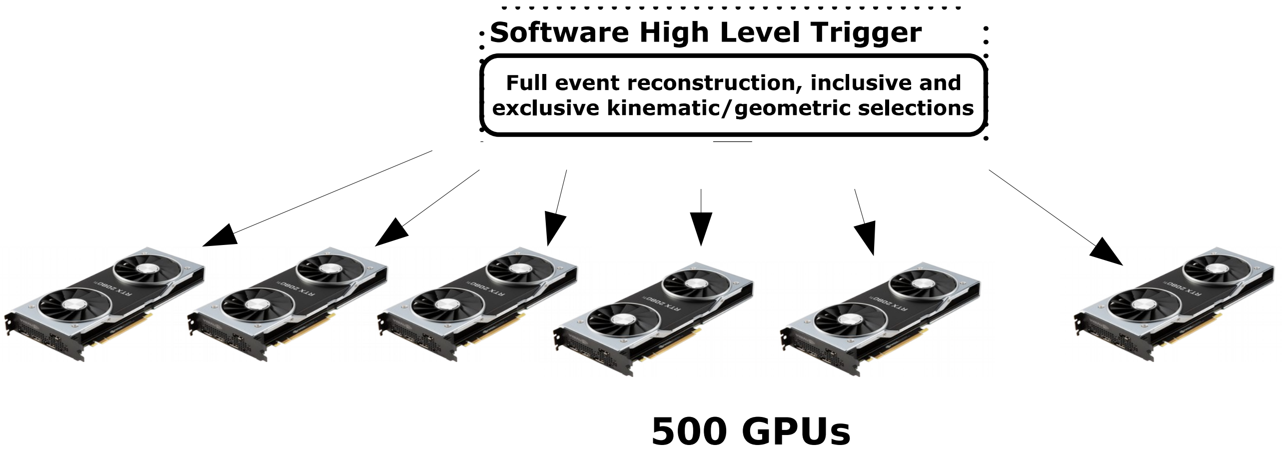
laSalle Universitat Ramon Llull

IFIC Institut de Física Corpuscular

Excelencia Severo Ochoa

Summary

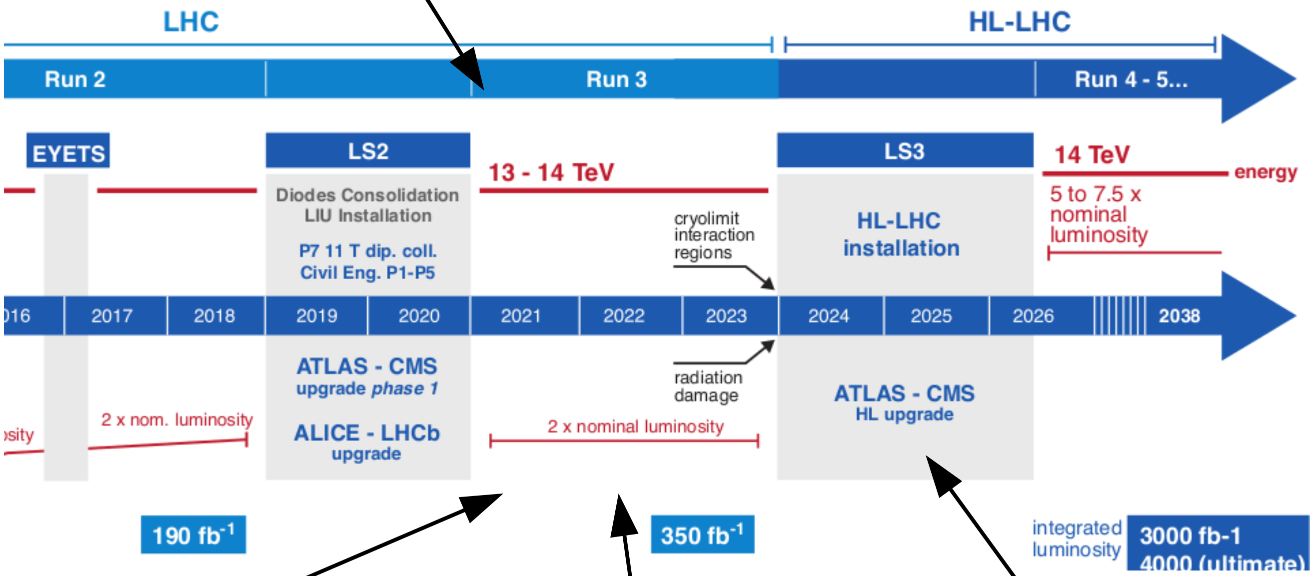
- Allen is the first complete high throughput trigger implementation on GPUs
- Developed a compact, modular and scalable framework
- Baseline HLT1 can run on GPUs
- Scaling of GPU performance → maximize physics discovery potential of LHCb
- Integration tests ongoing (see talk by D. Cámpora, Monday Track 5)
- HLT1 on GPUs is being considered as alternative to the baseline x86 architecture



Backup

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?

Graphics requirements

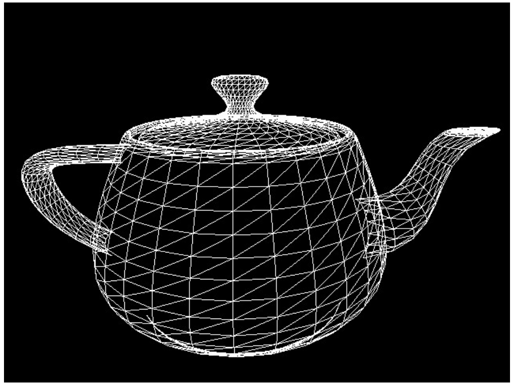
Graphics pipeline

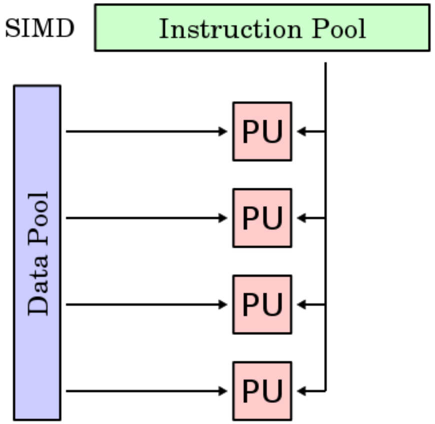
- Huge amount of arithmetic on independent data:
  - Transforming positions
  - Generating pixel colors
  - Applying material properties and light situation to every pixel

Hardware needs

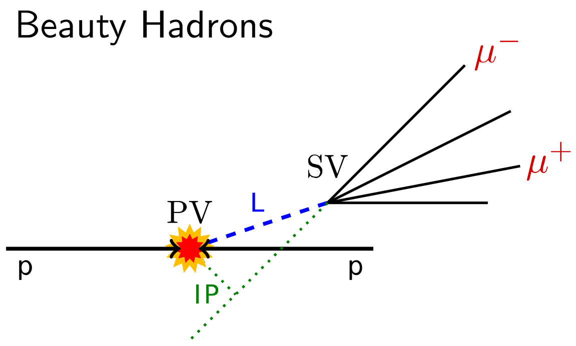
- Access memory simultaneously and contiguously
- Bandwidth more important than latency
- Floating point and fixed-function logic

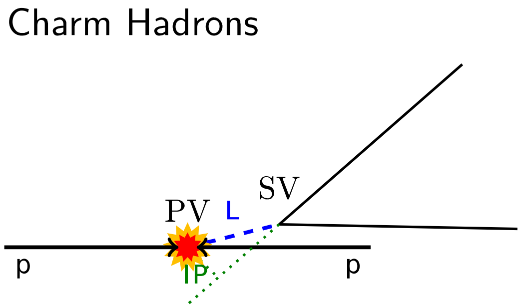
→ Single instruction applied to multiple data: SIMT





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

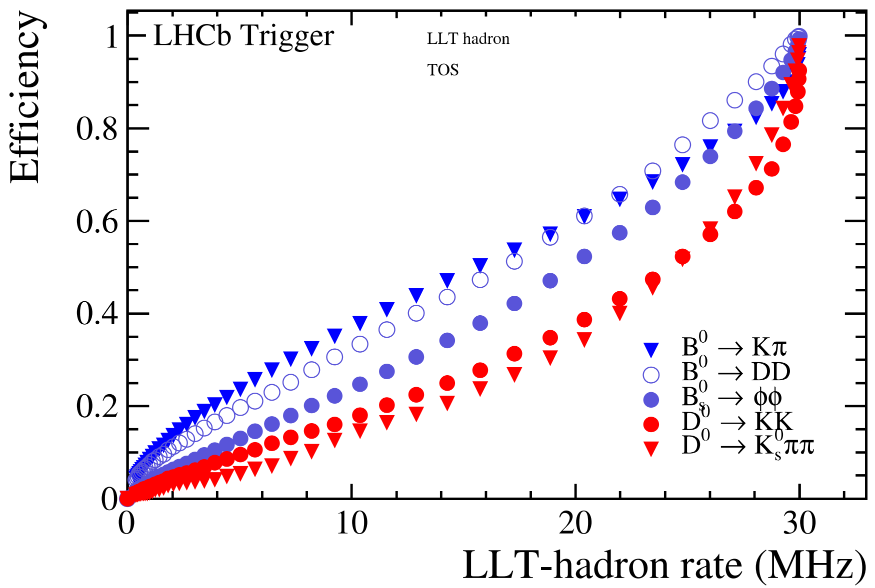
- $D^{\pm/0}$ mass ~1.9 GeV

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

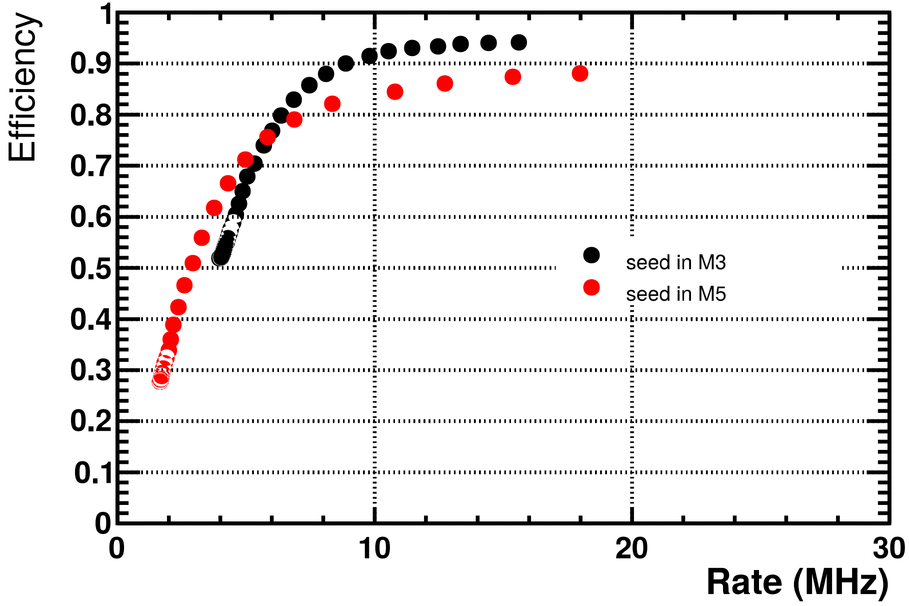
PV: Primary vertex
SV: Secondary vertex
IP: Impact parameter: distance between point of closest approach of a track and a PV

Why no low level trigger?

Low level trigger on E_T from the calorimeter



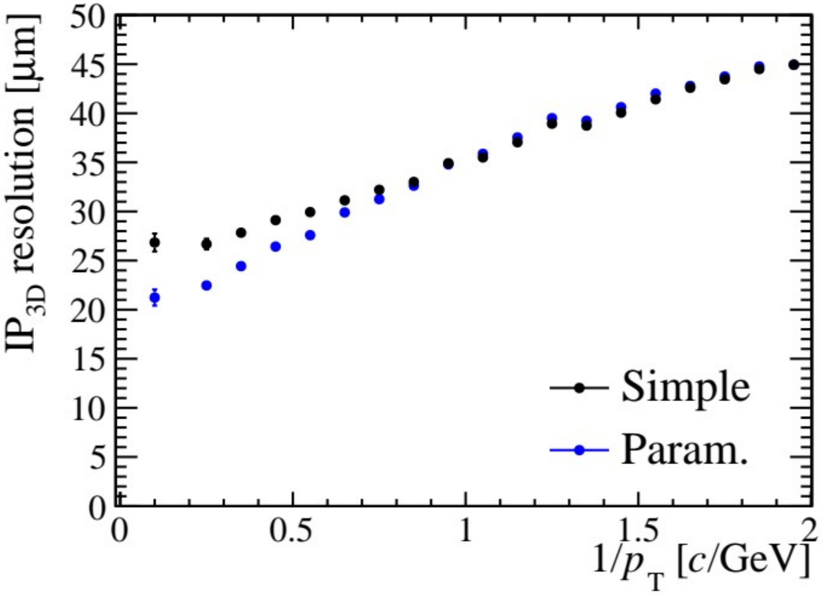
Low level trigger on muon p_T, $B \to K^*\mu\mu$

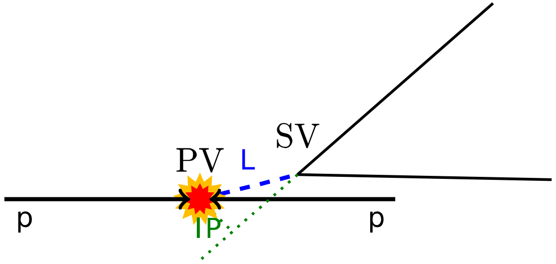


Need track reconstruction at first trigger stage

Kalman filter

Improved track description → better impact parameter resolution

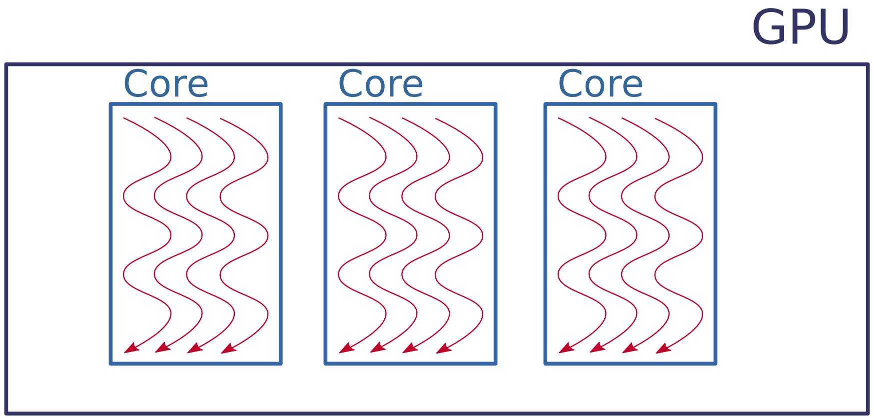




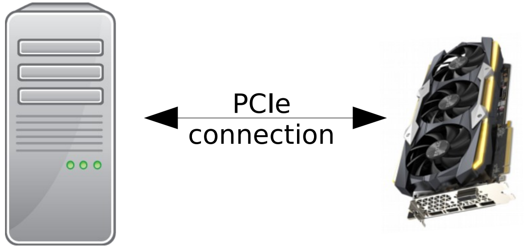
- Simple: Simplified Kalman filter with constant momentum assumption
- Param.: Parameterized Kalman filter with momentum estimate from SciFi track reconstruction

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

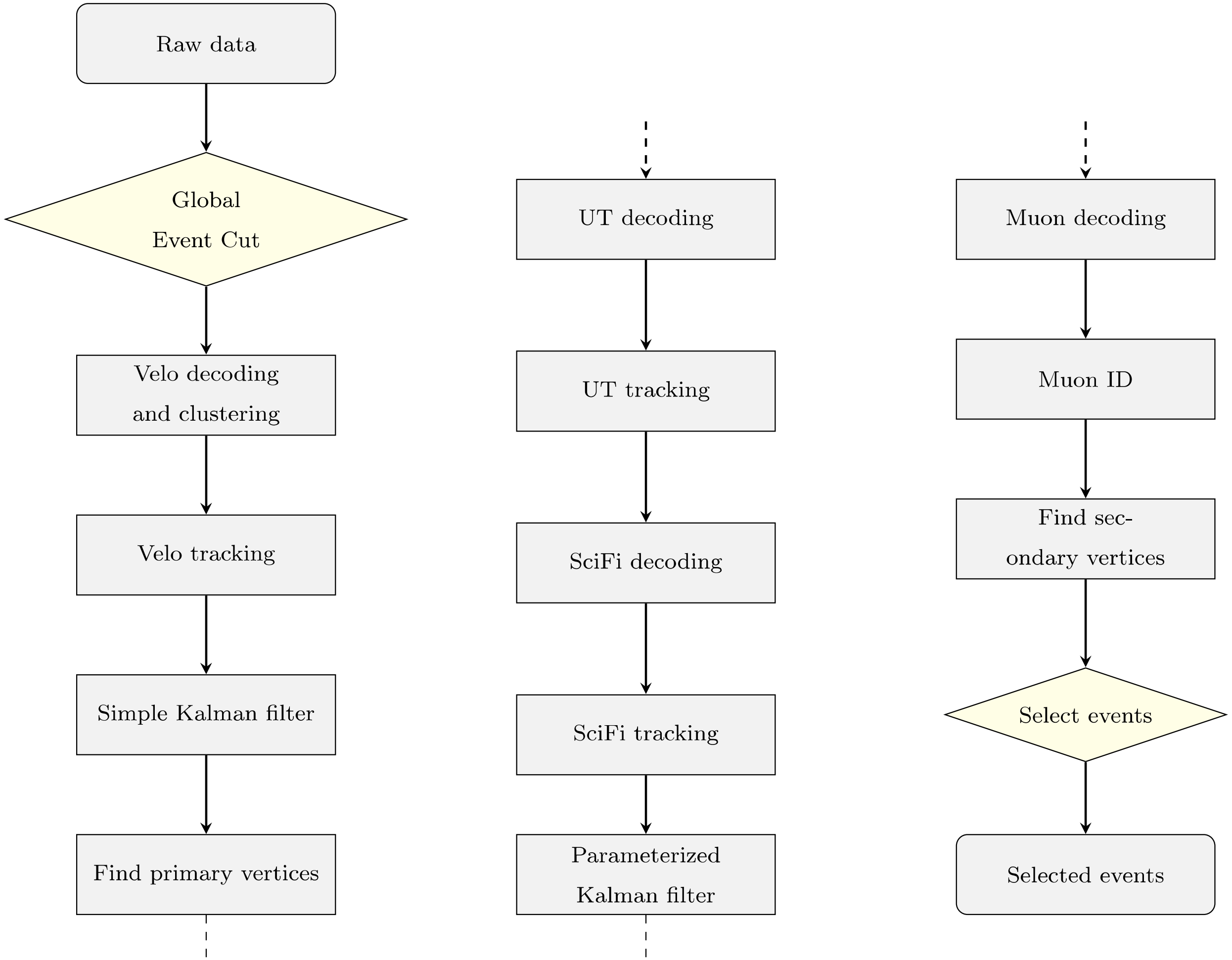
Selections

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

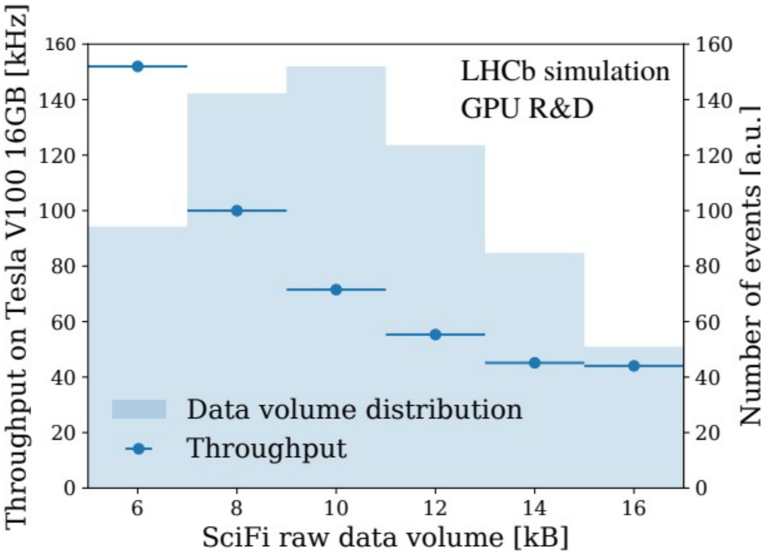
Criteria applied to signal decays in efficiency calculations

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



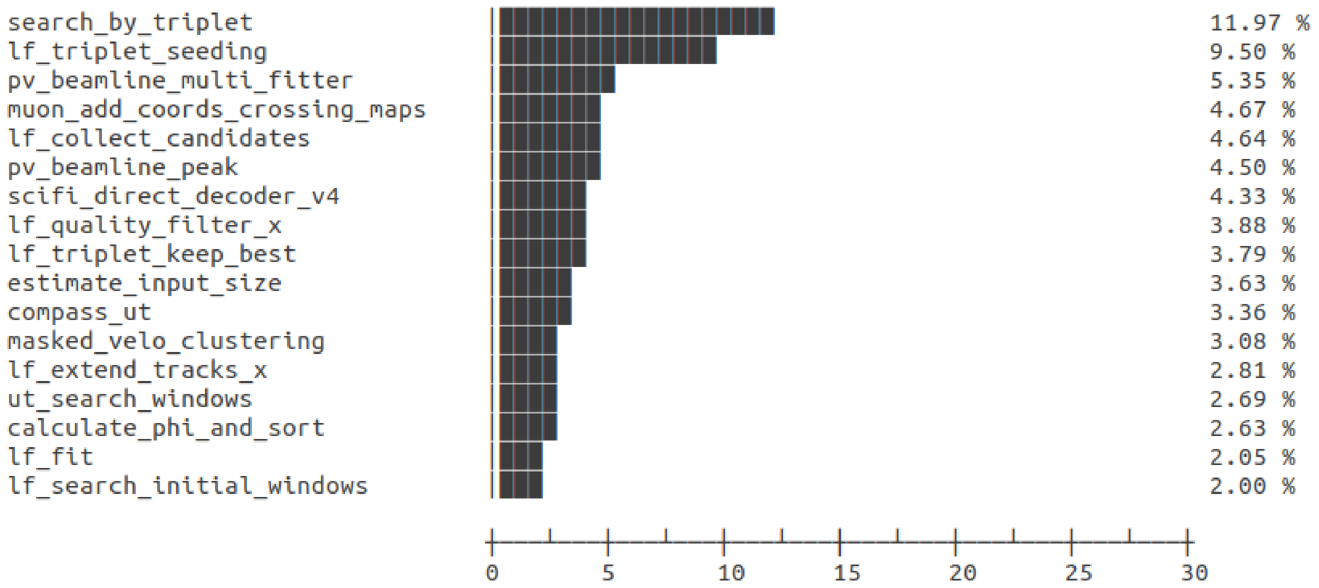
Throughput versus occupancy



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

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

Algorithm breakdown



Showing only algorithms contributing ≥ 2%

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

Throughput of x86 HLT1