Status of Allen, a complete first level trigger (& framework) for GPUs

Allen

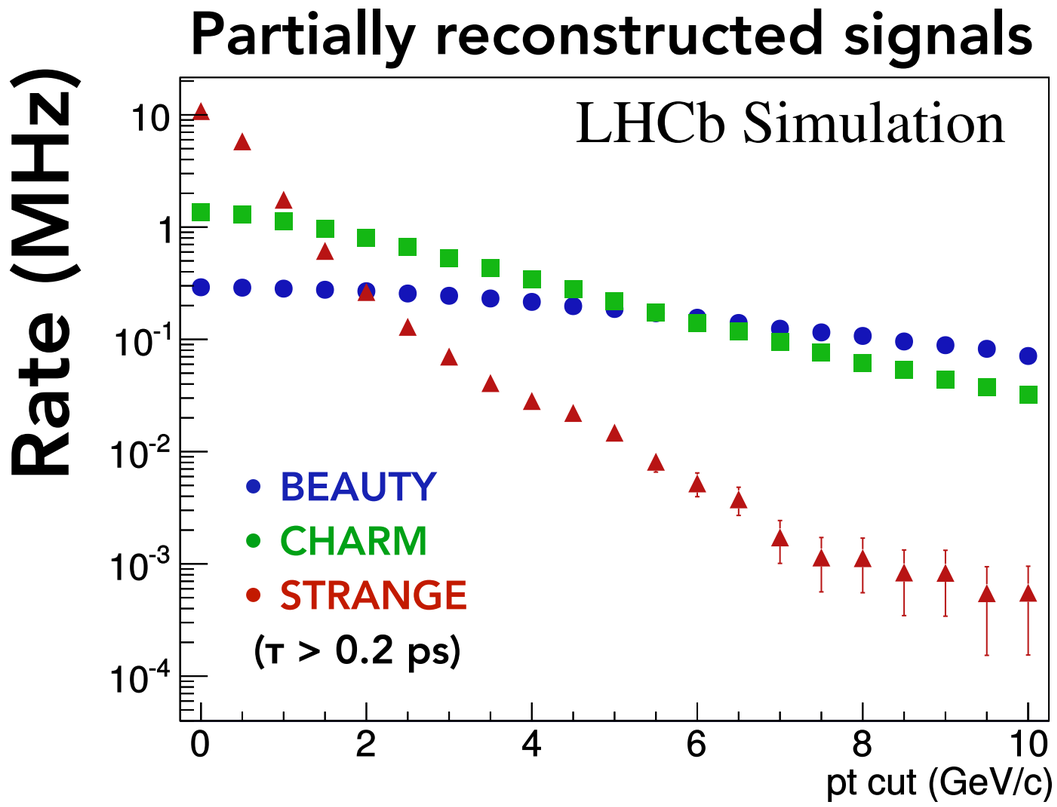
Project ID: 38633

V. V. Gligorov, CNRS/LPNHE

CERN OpenLab Technical Workshop, 22.01.2020

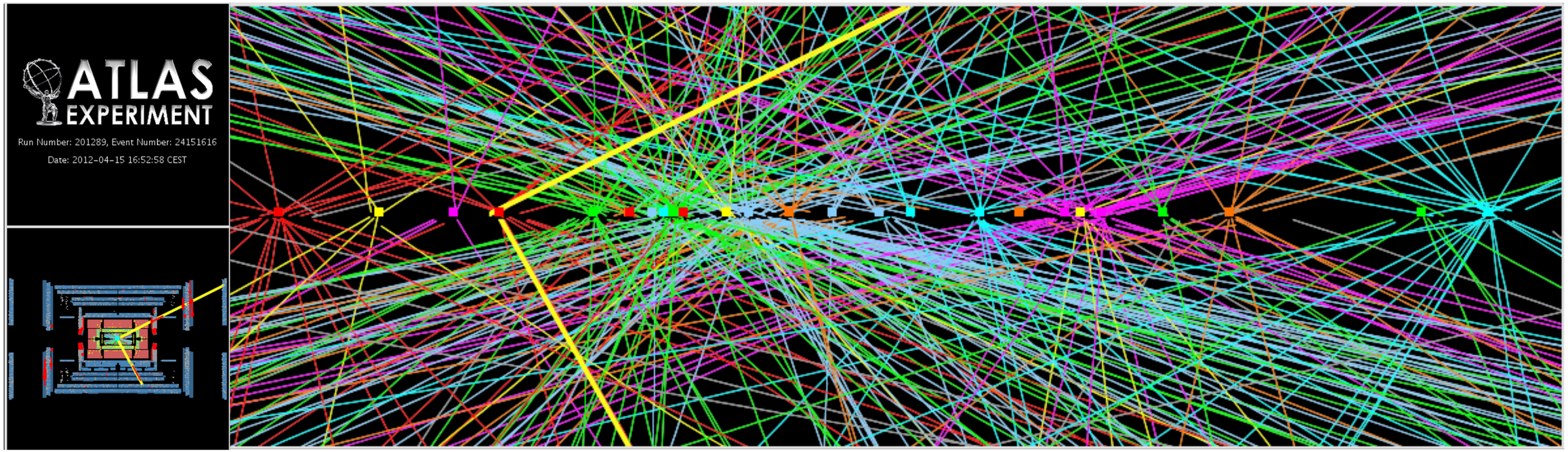
European Research Council
Established by the European Commission

The challenge of the LHCb upgrade in one slide



We will have MHz of signals in our acceptance!
We can only afford to fully store 50 kHz of events

From selection to compression : real-time analysis

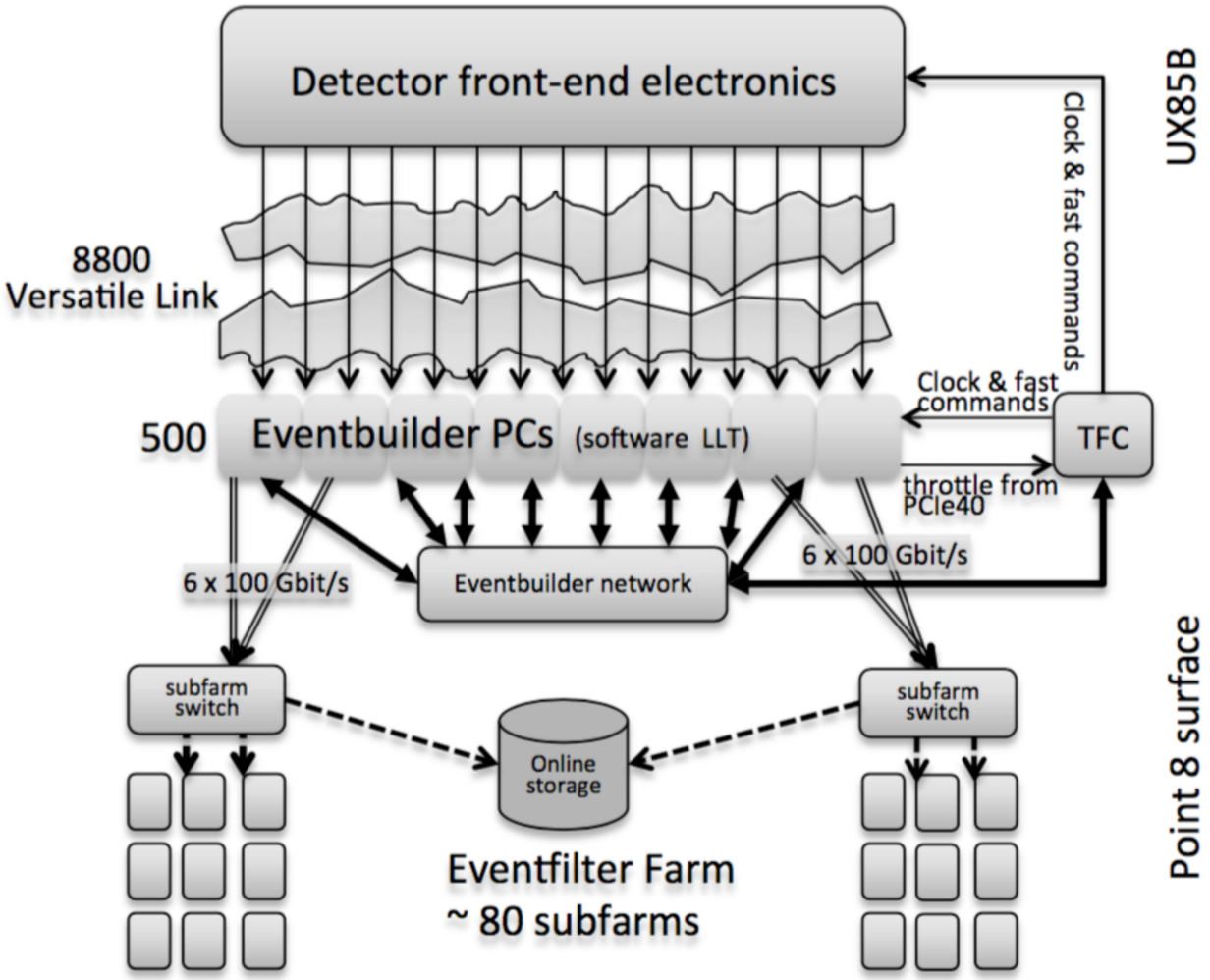


Create an order of magnitude more room for signal by compressing and removing pileup in real-time!

Or in a picture…

Typical triggers select signal needles in Standard Model haystacks
Real-time analysis sorts and compresses haystacks of needles

From this follows the LHCb DAQ design for the upgrade



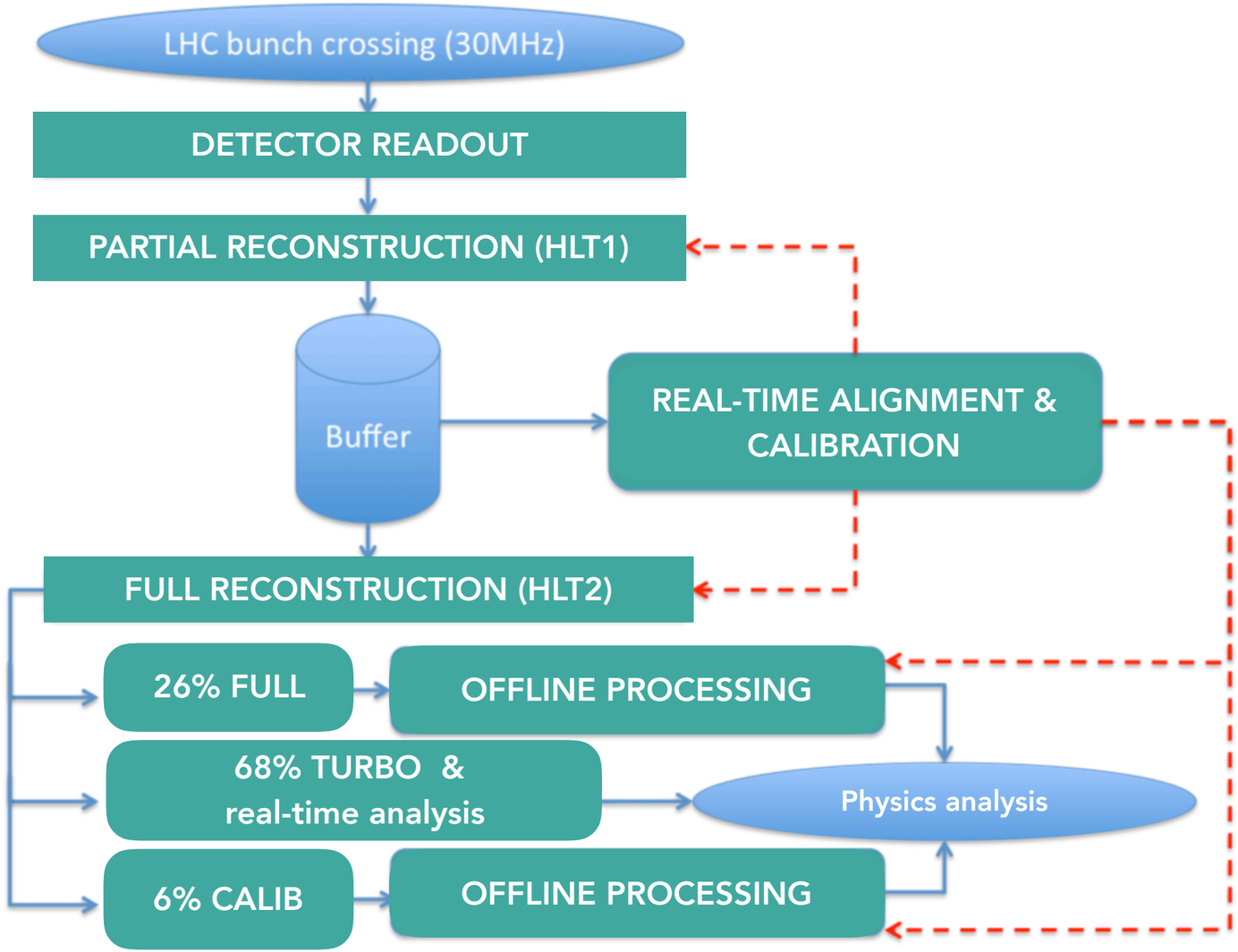
40 Tbit/s full event building & processing in a data centre

LHCb upgrade dataflow

40 Tb/s
30 MHz non-empty pp

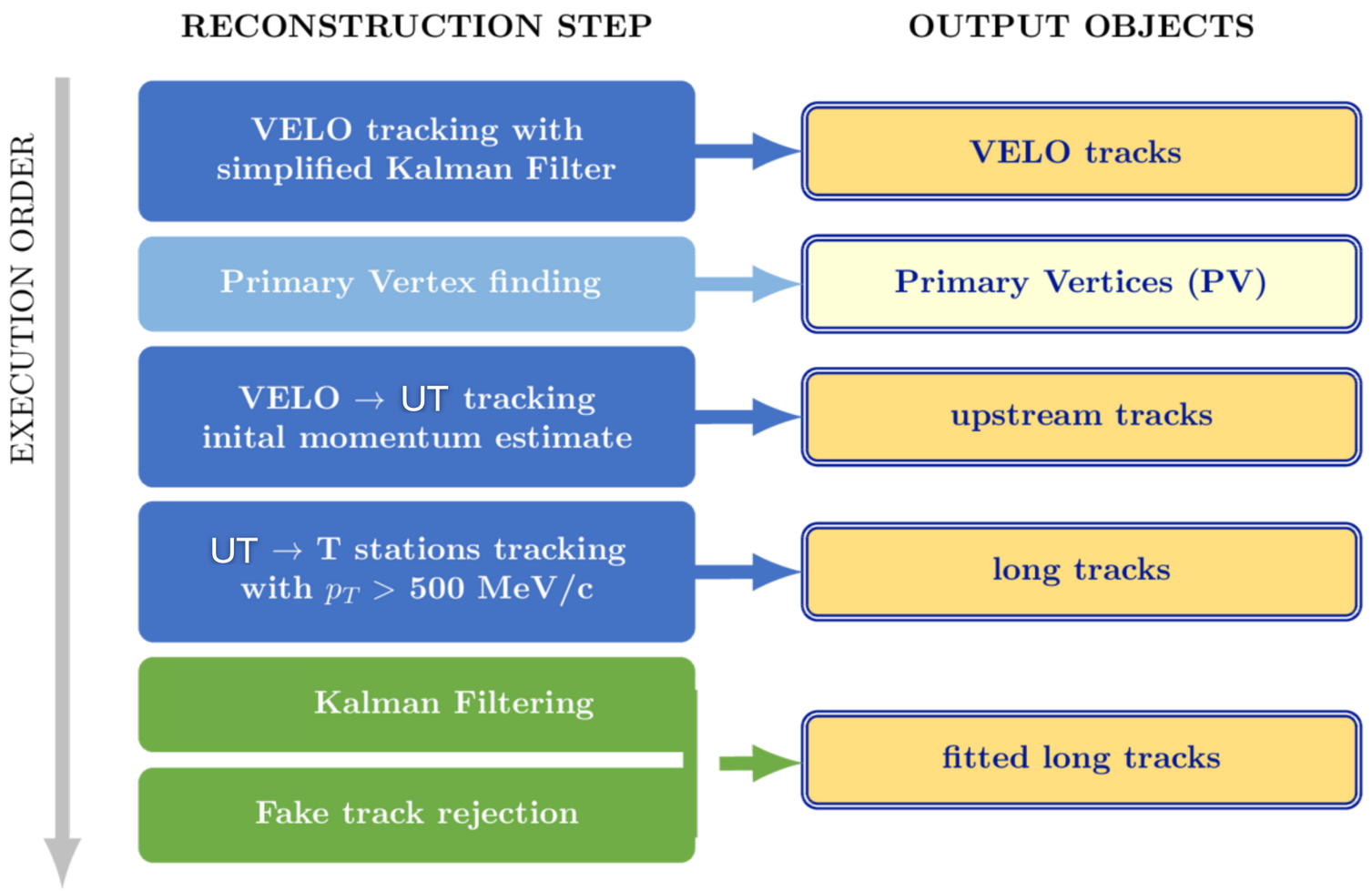
1-2 Tb/s
0.5 - 1.5 MHz

80 Gb/s



HLT1 is a huge high-throughput challenge — budget of a few M$

What is the physics content of HLT1 which runs @30 MHz?



"Traditional" inclusive selections selecting bunch crossings.
Based on charged particles, so require 30 MHz tracking at $2 \cdot 10^{33}$!

Pause and compare this to ATLAS/CMS HL-LHC processing

CMS detector	LHC Run-2	HL-LHC Phase-2	
Peak ⟨PU⟩	60	140	200
L1 accept rate (maximum)	100 kHz	500 kHz	750 kHz
Event Size	2.0 MB [a]	5.7 MB [b]	7.4 MB
Event Network throughput	1.6 Tb/s	23 Tb/s	44 Tb/s
Event Network buffer (60 seconds)	12 TB	171 TB	333 TB
HLT accept rate	1 kHz	5 kHz	7.5 kHz
HLT computing power [c]	0.5 MHS06	4.5 MHS06	9.2 MHS06
Storage throughput	2.5 GB/s	31 GB/s	61 GB/s
Storage capacity needed (1 day)	0.2 PB	2.7 PB	5.3 PB

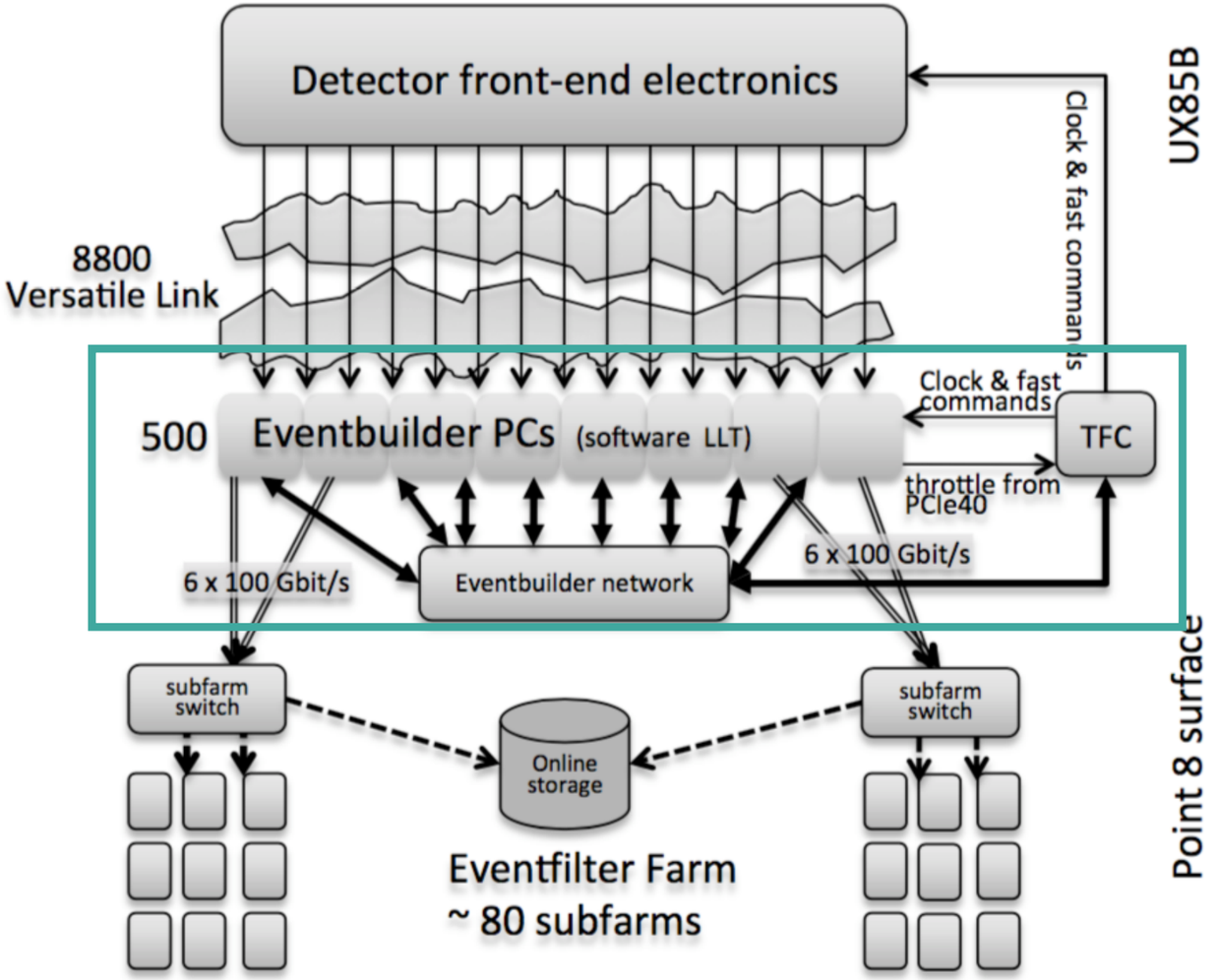
The LHCb upgrade has to handle the same data volume in real-time as ATLAS/CMS HL-LHC upgrades! But earlier and for less money...

Allen developed as one solution to this problem

LHCb-ANA-20XX-YYY
May 31, 2019

Proposal for an HLT1 implementation on GPUs for the LHCb experiment

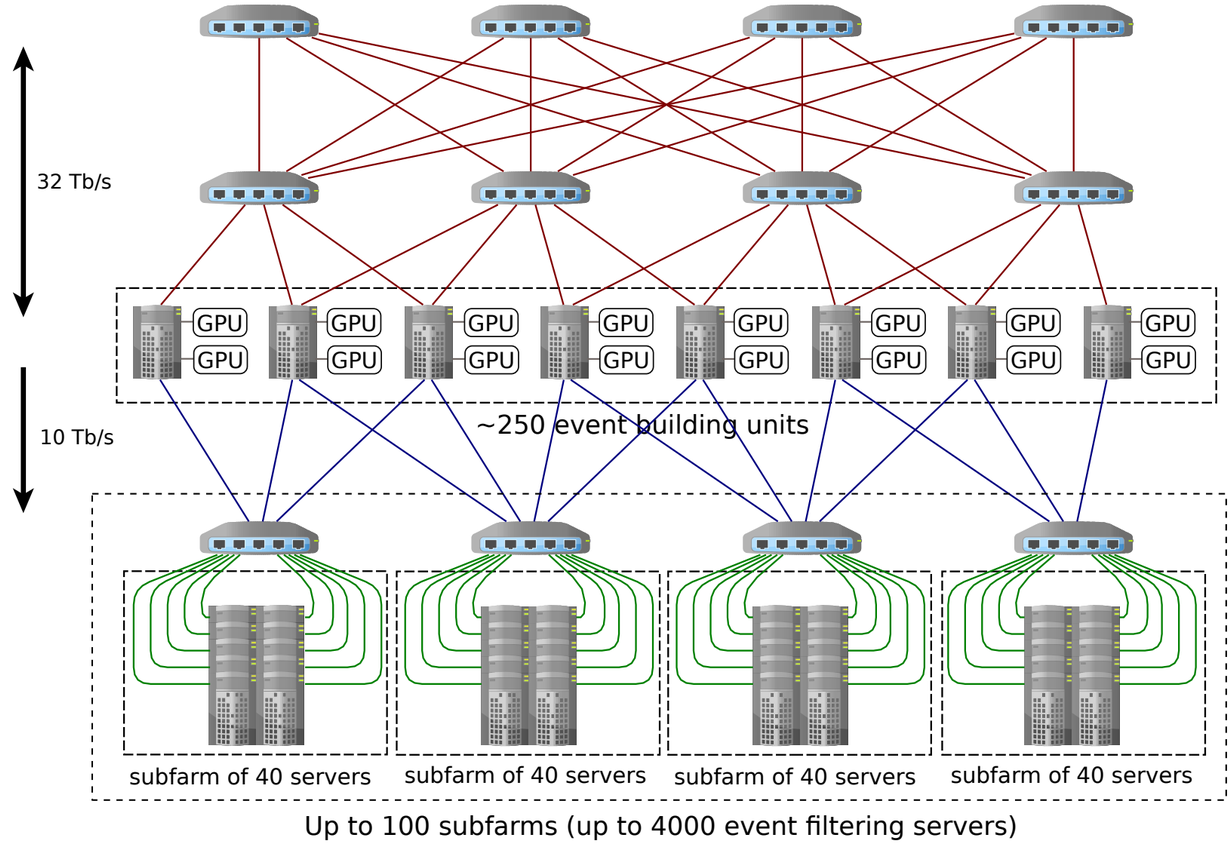
R. Aaij[1], J. Albrecht[2], M. Belous[a,3], T. Boettcher[4], A. Brea Rodríguez[5], D. vom Bruch[6], D. H. Cámpora Pérez[b,7], A. Casais Vidal[5], P. Fernandez Declara[c,7], L. Funke[2], V. V. Gligorov[6], B. Jashal[9], N. Kazeev[a,3], D. Martínez Santos[5], F. Pisani[d,e,7], D. Pliushchenko[f,3], S. Popov[a,3], M. Rangel[10], F. Reiss[6], C. Sánchez Mayordomo[9], R. Schwemmer[7], M. Sokoloff[11], A. Ustyuzhanin[a,3], X. Vilasís-Cardona[8], M. Williams[4]



Exploit flexibility of LHCb's DAQ to implement HLT1 on O(500) GPUs in the event-building servers. No comparisons to the fully viable and complete CPU baseline here — cost-benefit analysis is ongoing to decide if LHCb will use Allen in Run 3.

Architecture of a GPU trigger @ 30 MHz

Arxiv 1912.09161

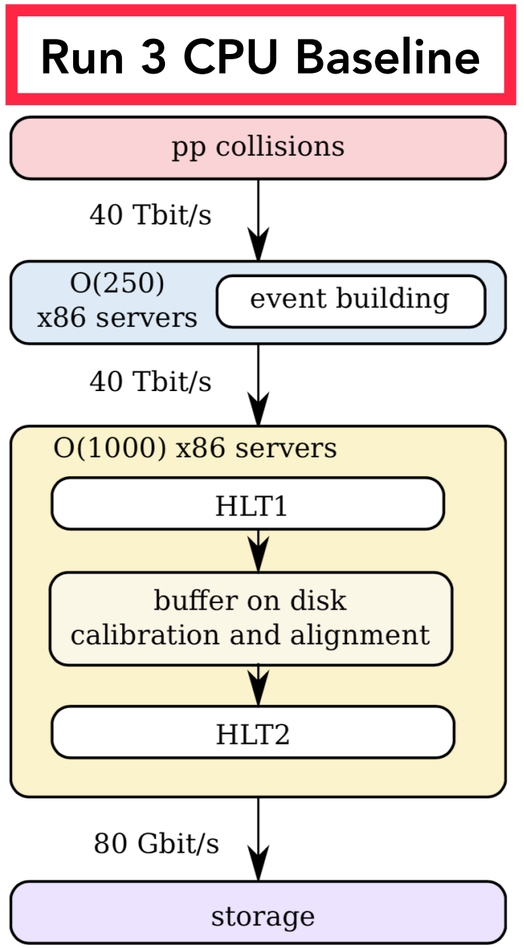


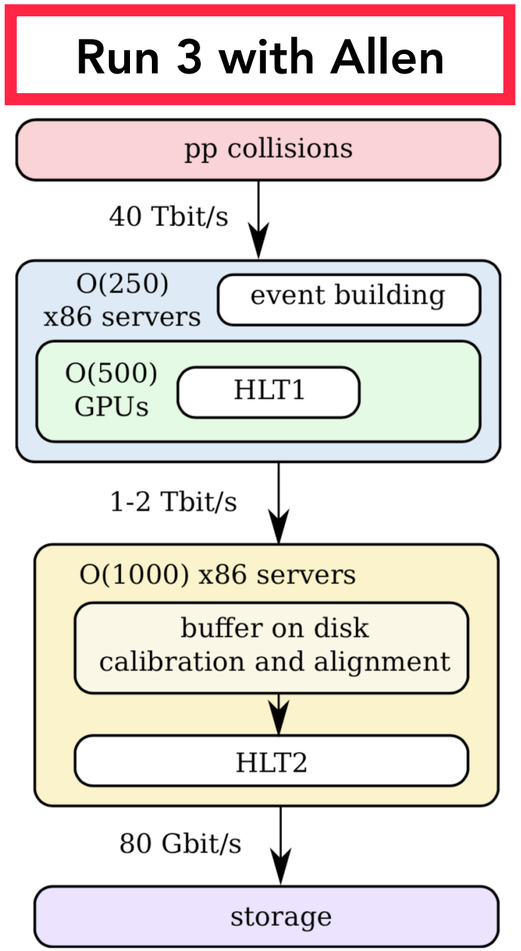
100G IB

10GbE

25GbE

PCIe

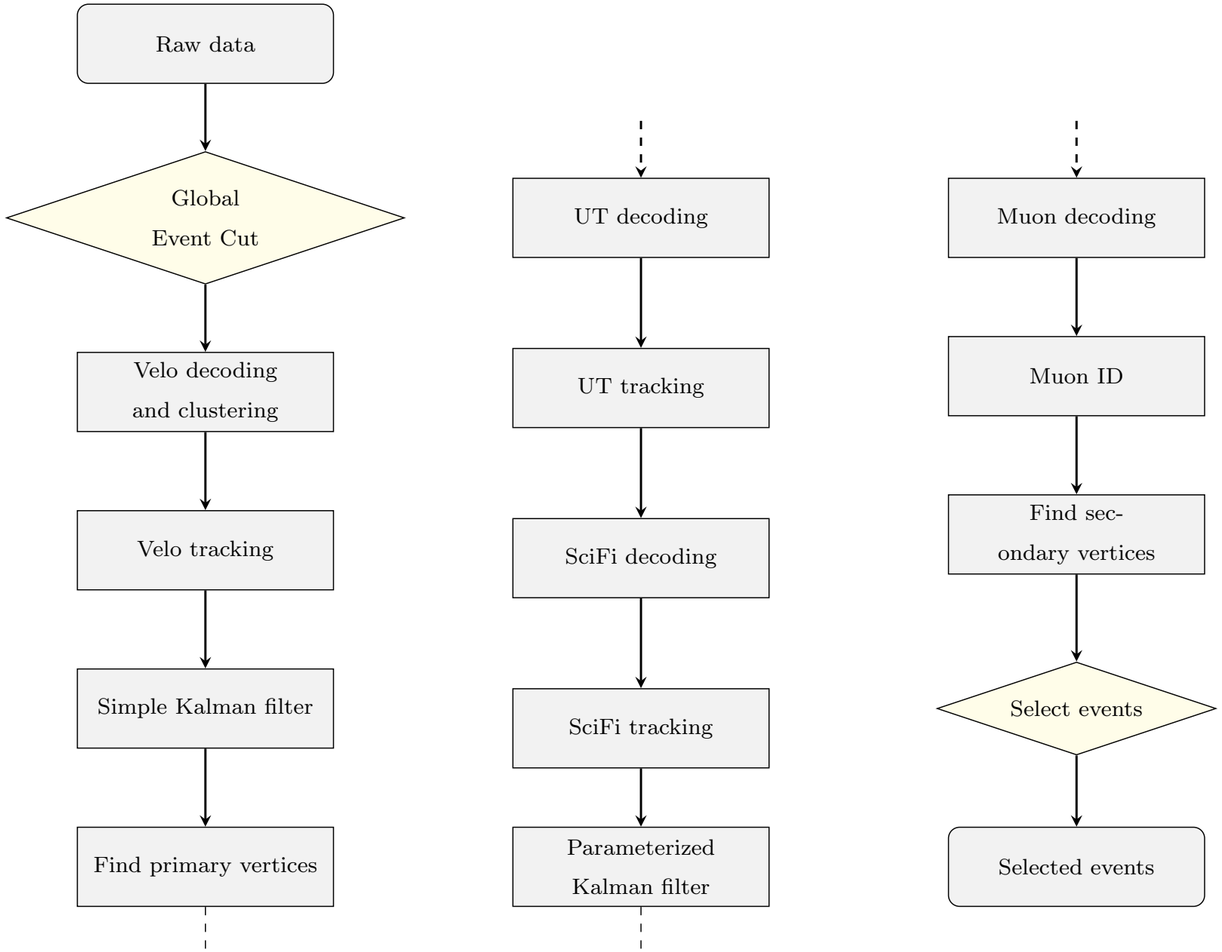




Use empty slots in event building servers — opportunistic but efficient.

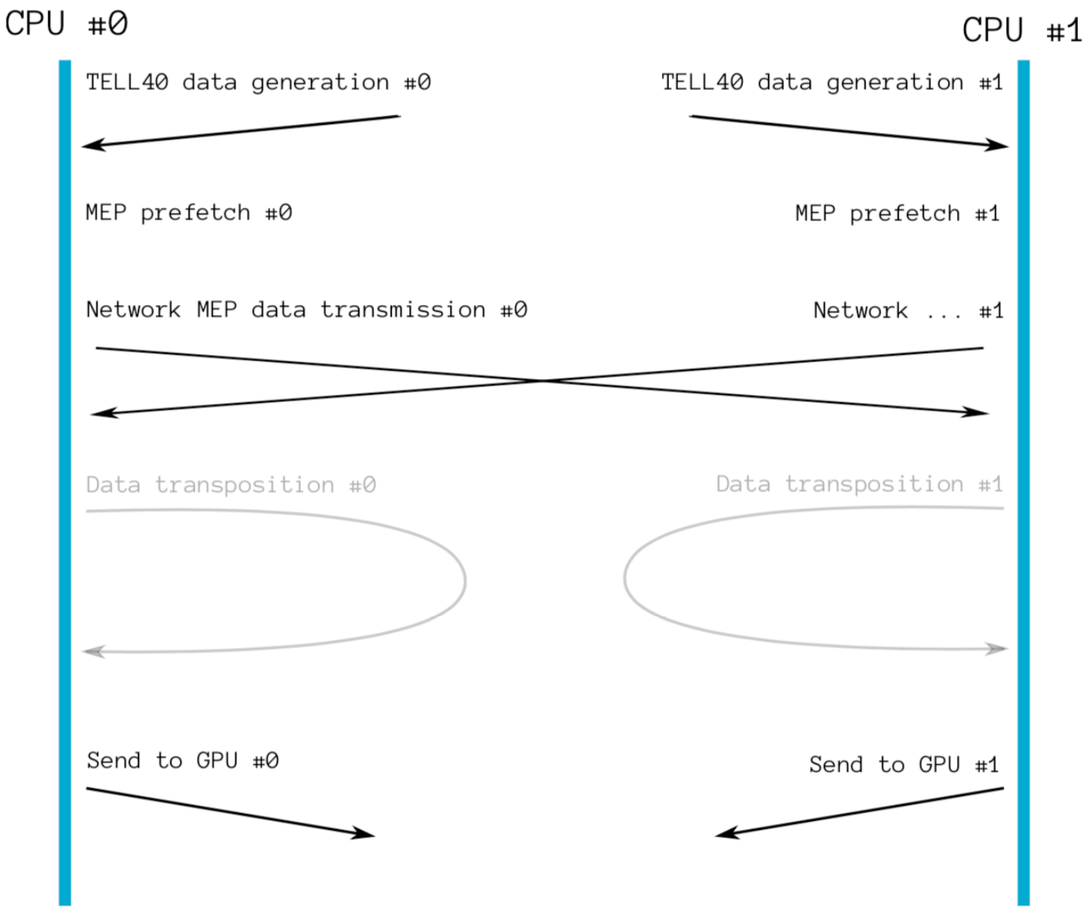
Each GPU consumes ~4 GB/s and outputs ~0.1–0.2 GB/s, well within PCIe3 limit.

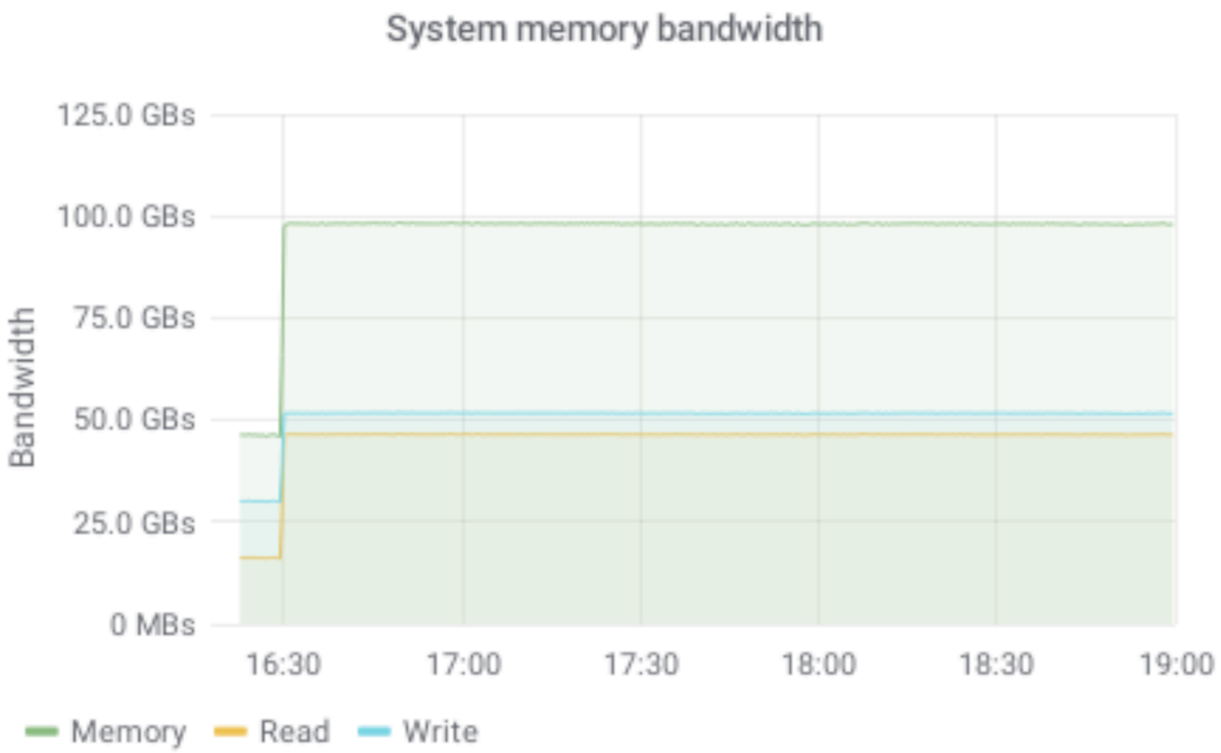
Allen is not just one or two algorithms



Allen is a complete data processing solution! A wide range of configurable reconstruction and selection algorithms, monitoring and writing of provenance information. Integration with Gaudi is provided for the LHCb production environment but Allen has no inherent reliance on the LHCb codebase. I/O is handled with minimal reliance on server CPU.

Integration test (LHCb figure 2019-016)

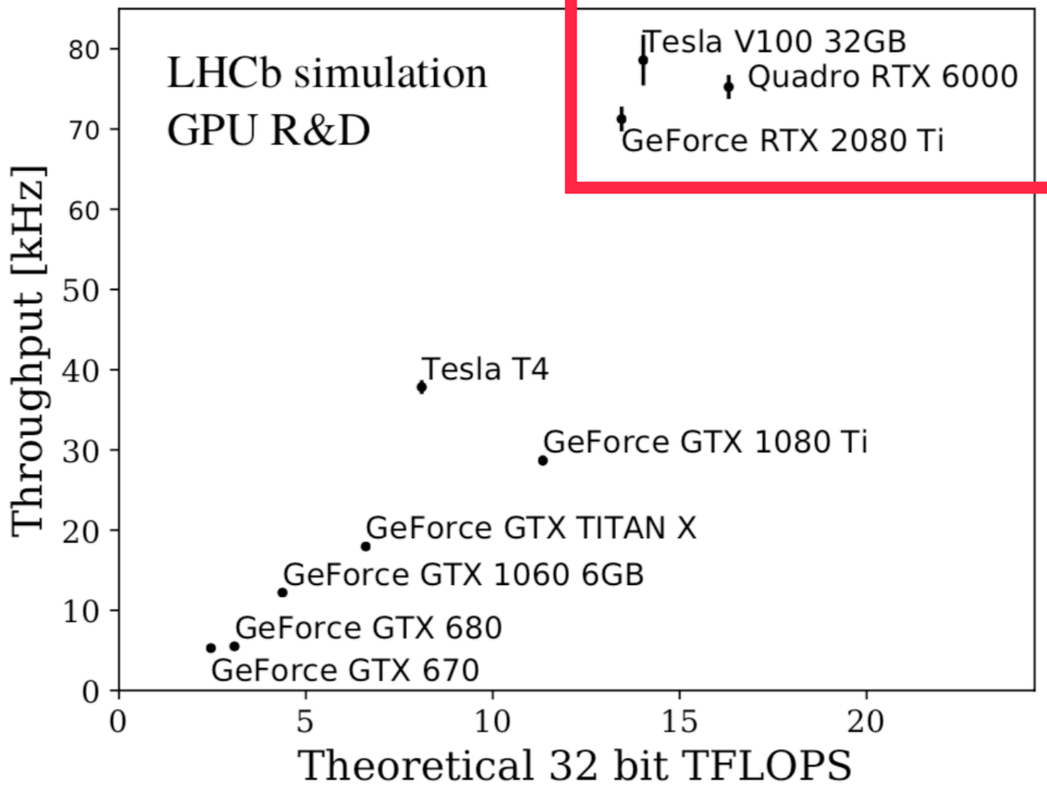


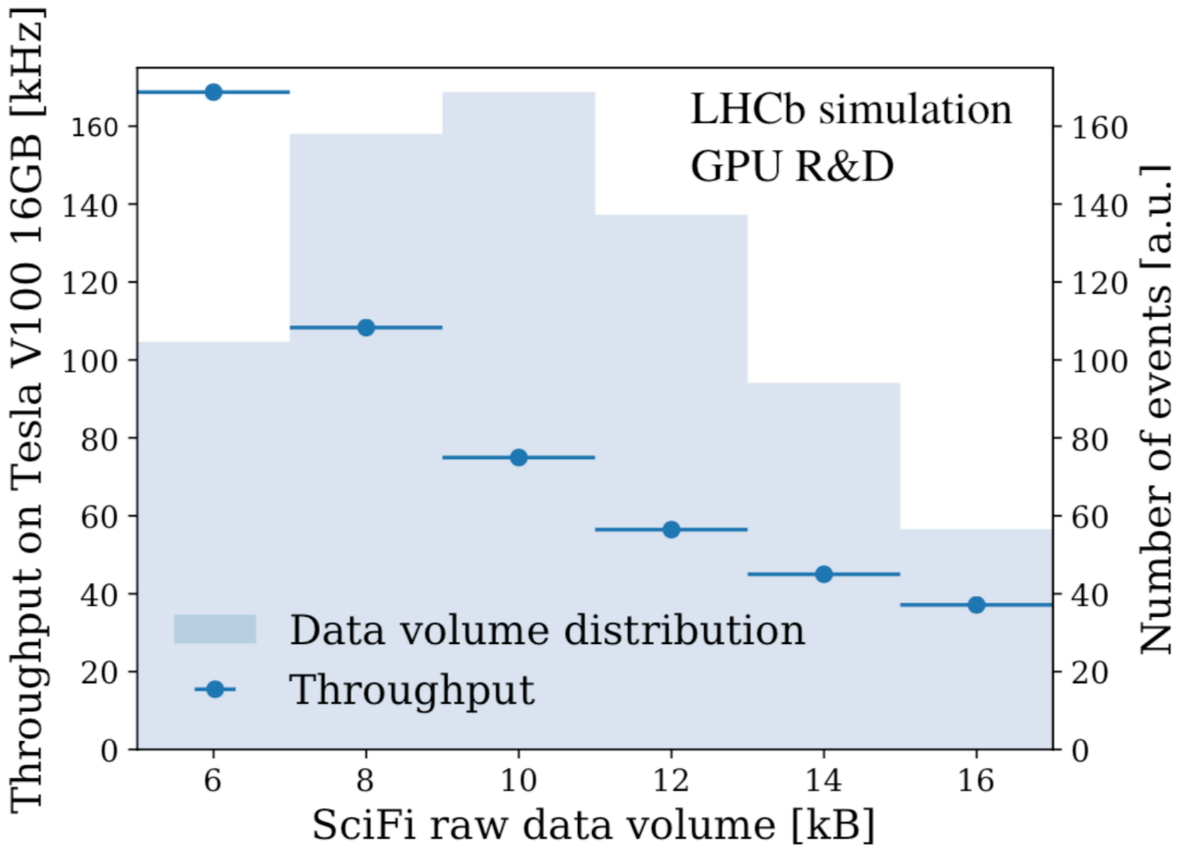


Test carried out with two GPUs and two readout boards in an Intel server: stable performance observed, I/O does not bottleneck the processing. A second harsher test in an AMD EPYC server with three GPUs and three readout boards across two NUMA domains is ongoing.

GPU throughput scaling

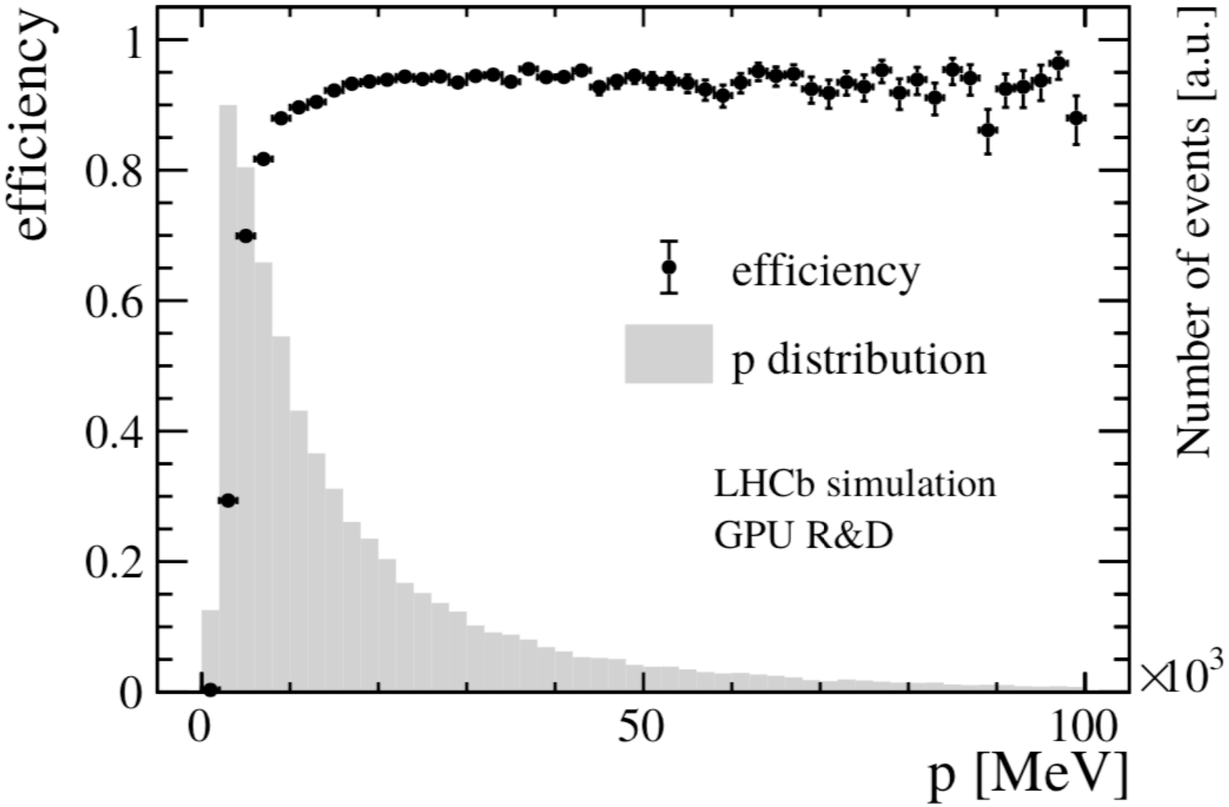
Compatible with 30 MHz on O(500) GPUs!

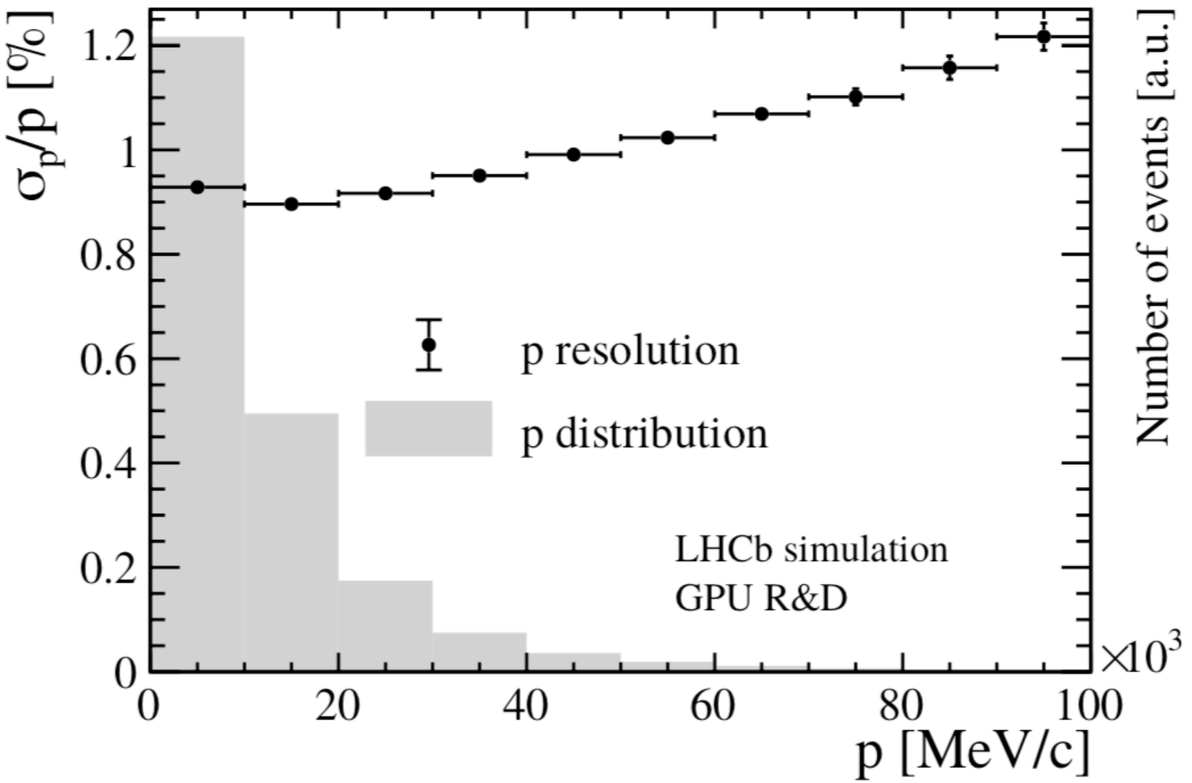




Linear scaling of throughput vs. occupancy, and throughput vs. the theoretical TFLOPS of each card. Efficient use of hardware!

Reconstruction performance





Charged particle reconstruction at 30 MHz down to zero transverse momentum with order 1% momentum resolution.

Physics performance

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
$B_s^0 \to \phi\phi$	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

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

Can reduce the data rate by required factor 30 while keeping >50% selection efficiency for most signals of interest.

Personal observations on working in a hybrid world

1. The computing landscape increasingly consists of hybrid architectures. We are developing the skills to thrive in this world!
2. If the basic principles of high throughput computing are respected, a well designed software architecture will perform on x86, GPU, or FPGA systems. Functional design and uniform API helps to achieve this.
3. High-throughput software is far from what universities teach physics students no matter the architecture. Learning CUDA, HLS or C++17 is the same for them. Recognise the importance of new skills in the field.
4. Real-time processing is a home for API designers, physicists and selection authors, throughput experts, algorithm designers… it's a very diverse community and personal architecture preferences are real. It is more work to keep a diverse community coherent, but it's worth it.

Huge thanks to NVIDIA and OpenLab for all the support and encouragement over the last 24 months!

Conclusions and final thoughts

LHCb 2032

>1000 Eb/year

ATLAS+CMS 2027

260 Eb/year

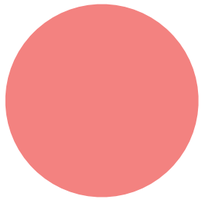
Square Kilometre Array (2030s)

~30000 Eb/year

Sequence genome of all humans on Earth

8000 Eb

Global internet dataflow 2021



2800 Eb/year

Master all tools to tackle the data deluge of next decades

Backup

LHCb analysis methodology and role of calibration samples

Trigger Efficiency

Tag-and-probe calibration method exists & widely used

Tracking efficiency

Tag-and-probe

Existing	Developing
μ	e,π,K,p

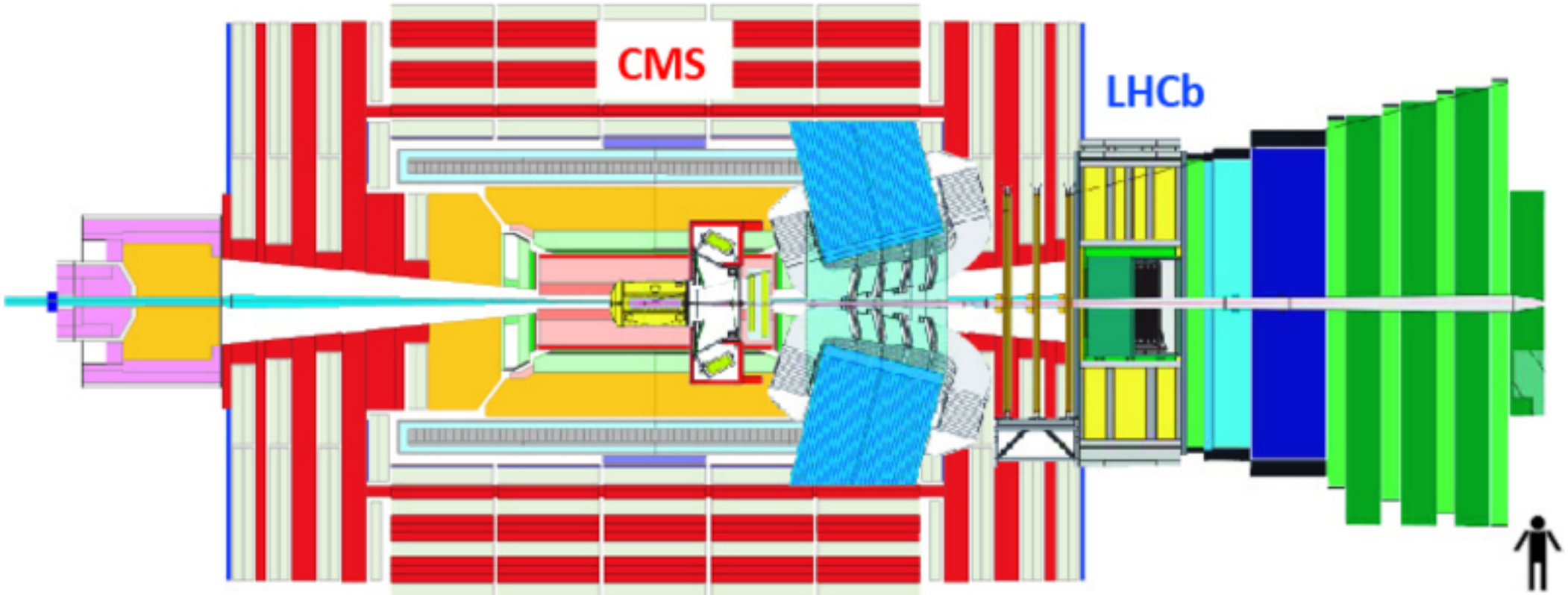
Particle identification

Tag-and-probe

Tag-and-probe calibrations exist for all charged particle species and for π^0/γ, with new sources added over time to improve coverage

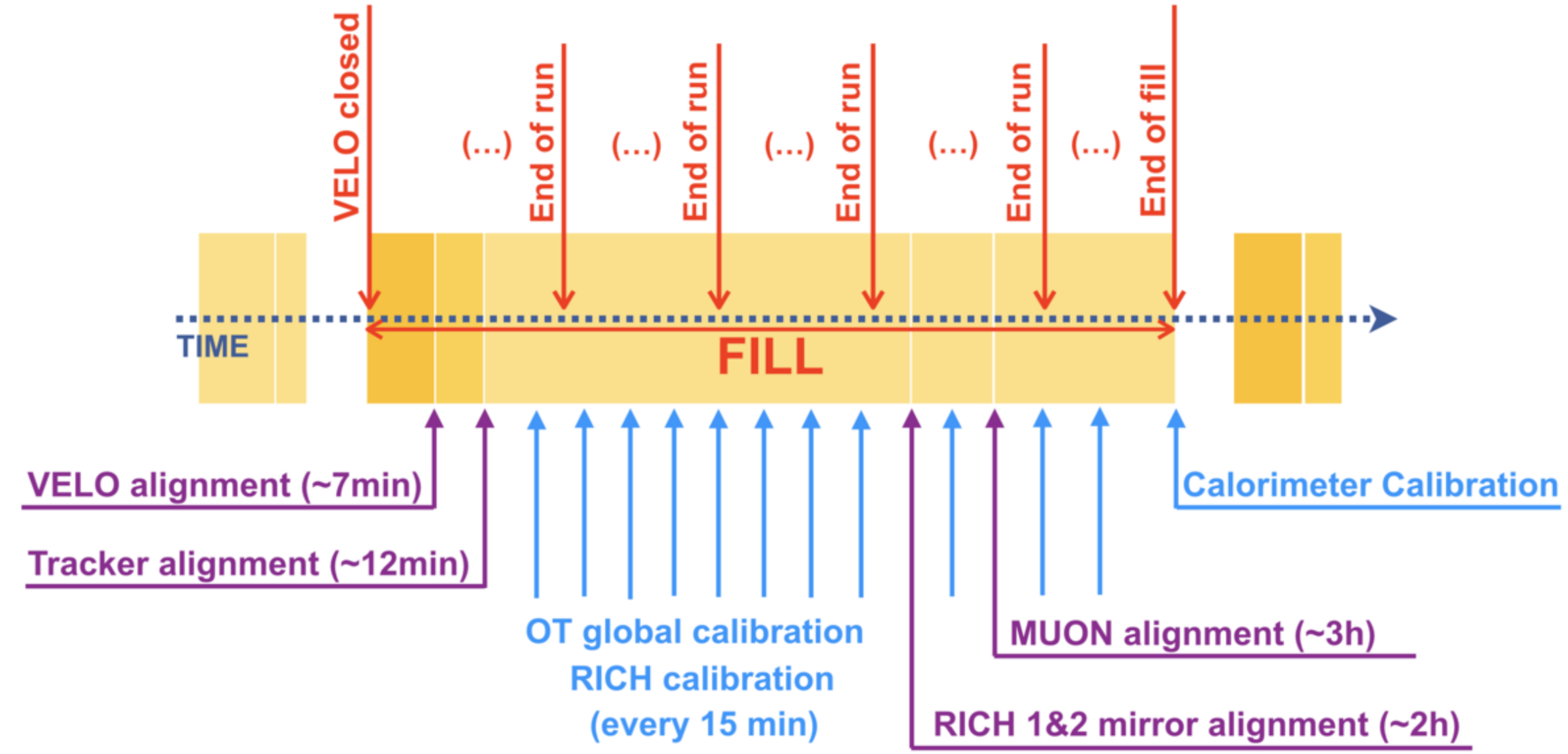
Data driven efficiency calibration key to precision physics

The LHCb detector at the LHC



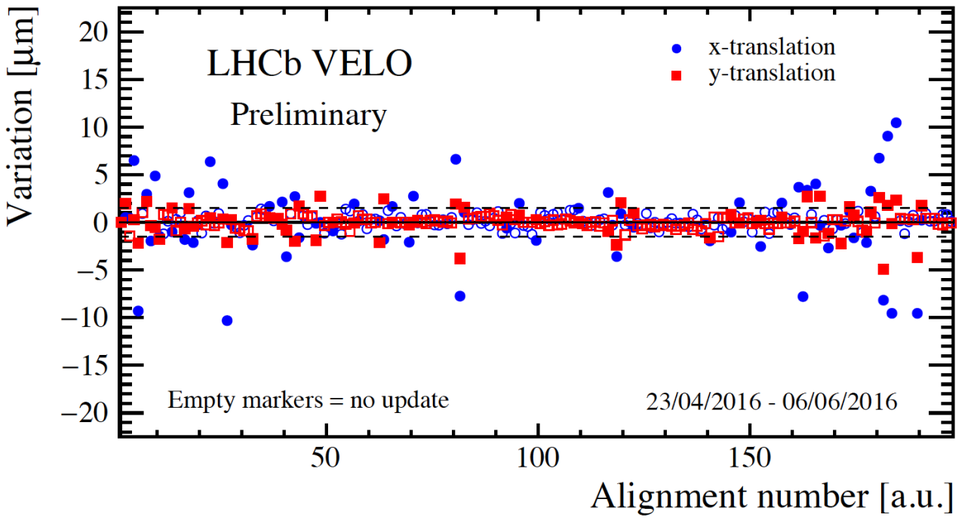
Forward spectrometer optimized for precision physics

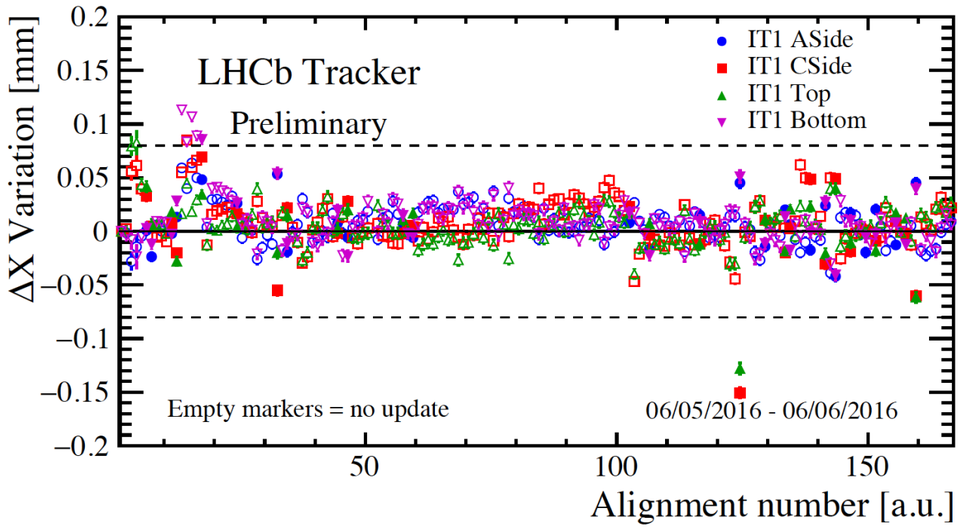
We also need to align and calibrate our detector in real time

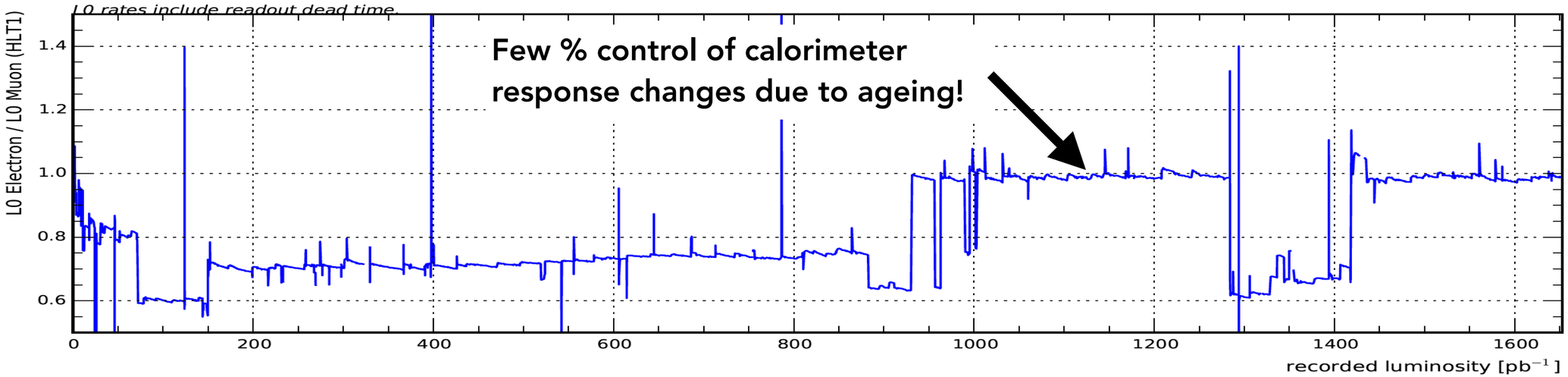


((~7min),(~12min),(~3h),(~2h)) - time needed for both data accumulation and running the task

So we did!



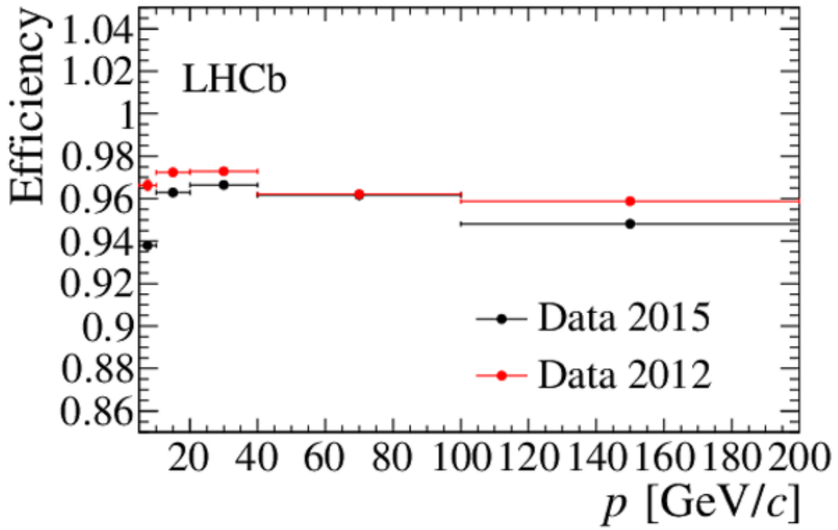


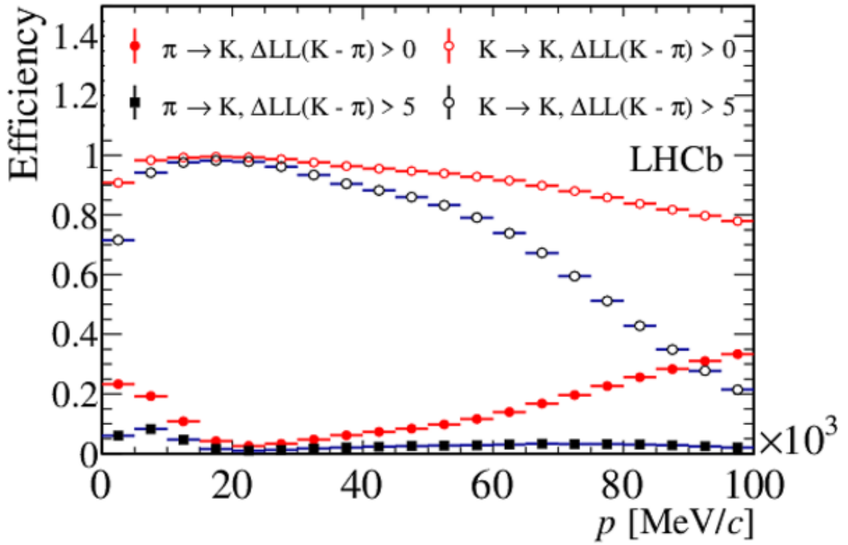


Implemented for the first time in Run 2 with offline like quality from very early in 2015. Not only tracker but also RICH and calorimeter. For me this is the most impressive aspect of LHCb's Run 2 and required a huge team effort across projects and working groups.

We also need to measure our efficiencies in real-time!

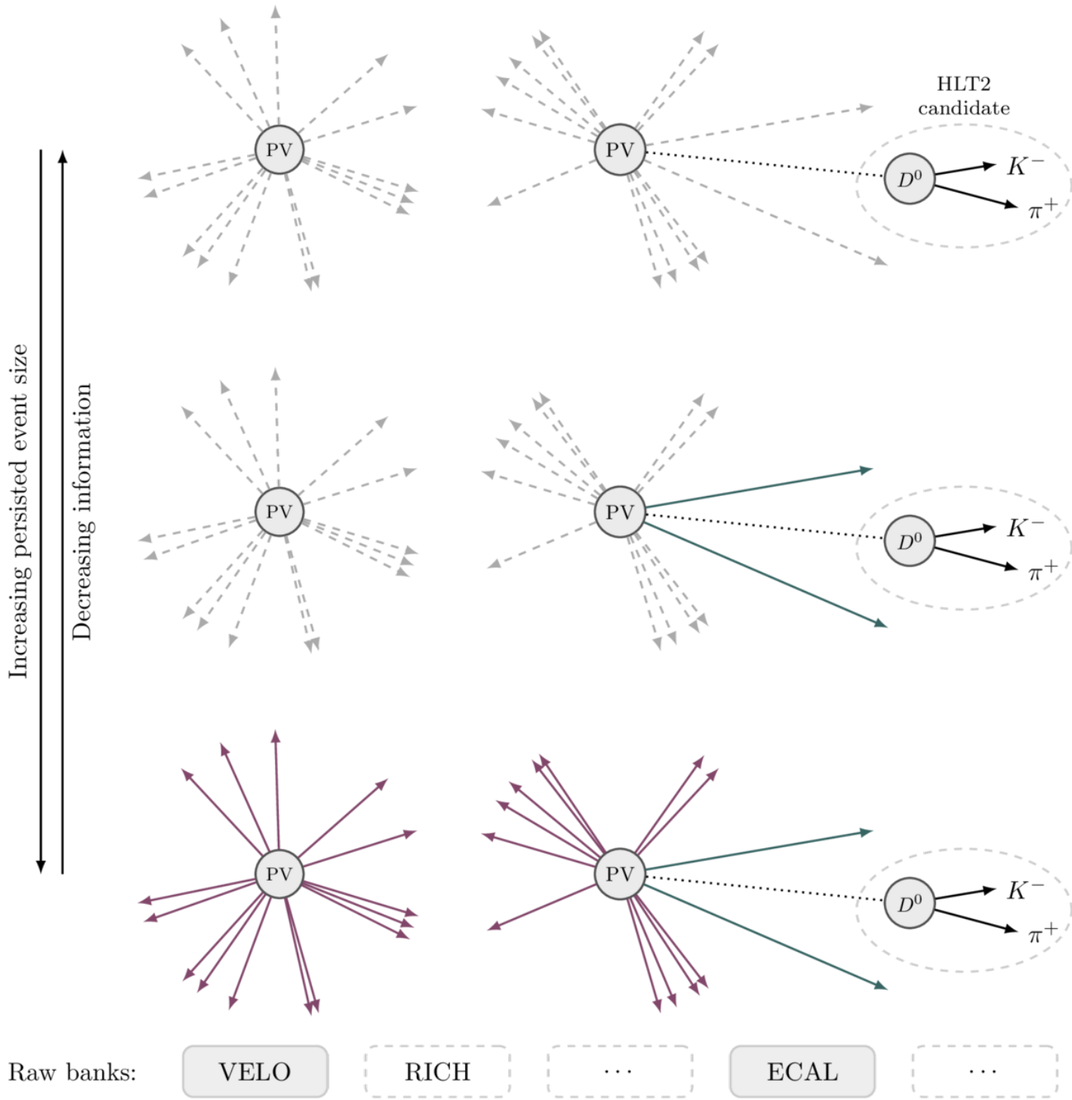
Species	Low momentum	High momentum
$e^{\pm}$	$B^+ \to J/\psi K^+$ with $J/\psi \to e^+e^-$	
$\mu^{\pm}$	$B^+ \to J/\psi K^+$ with $J/\psi \to \mu^+\mu^-$	$J/\psi \to \mu^+\mu^-$
$\pi^{\pm}$	$K^0_S \to \pi^+\pi^-$	$D^{*+} \to D^0\pi^+$ with $D^0 \to K^-\pi^+$
$K^{\pm}$	$D^+_s \to \phi\pi^+$ with $\phi \to K^+K^-$	$D^{*+} \to D^0\pi^+$ with $D^0 \to K^-\pi^+$
p , $\overline{p}$	$\Lambda^0 \to p\pi^-$	$\Lambda^0 \to p\pi^-$; $\Lambda^+_c \to pK^-\pi^+$





Unlike ATLAS and CMS, LHCb must maintain a data-driven permille level control of its efficiency across the kinematic and geometric acceptance of the detector. Requires collecting an extremely wide range of tag-and-probe samples in real time.

Then select signals and associate them to pp collisions



Full flexibility to store "additional" detector information if required by some analyses

Tracks in LHCb