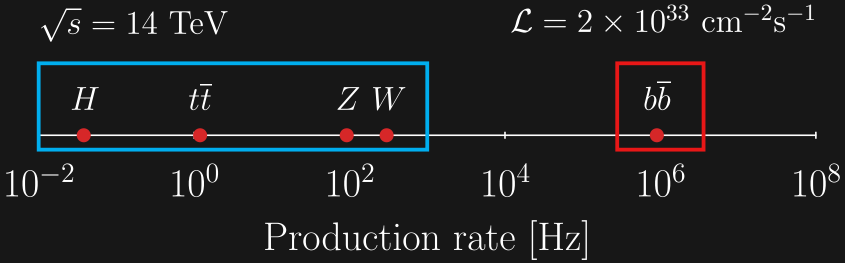
Allen: A High Level Trigger on GPUs for LHCb

Thomas Boettcher

on behalf of the LHCb Real Time Analysis project

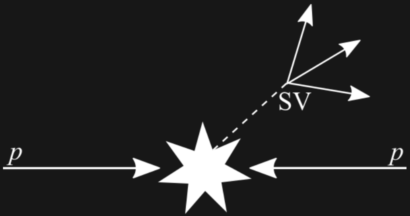
Connecting The Dots
April 20, 2020

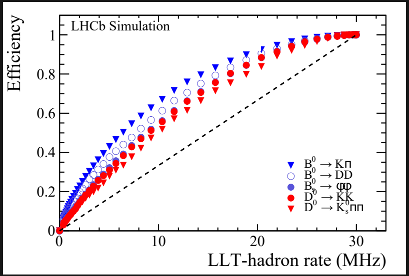
Triggering on MHz Signals



- **General Purpose Detectors:** Can trigger efficiently at ~ 100 kHz with single detector systems (e.g. high E_{T} calorimeter clusters)
- **LHCb:** The $b\bar{b}$ and $c\bar{c}$ rate will exceed a MHz, and final state particles can have $p_{\mathrm{T}} \lesssim 1$ GeV

Triggering on Heavy Flavor Decays

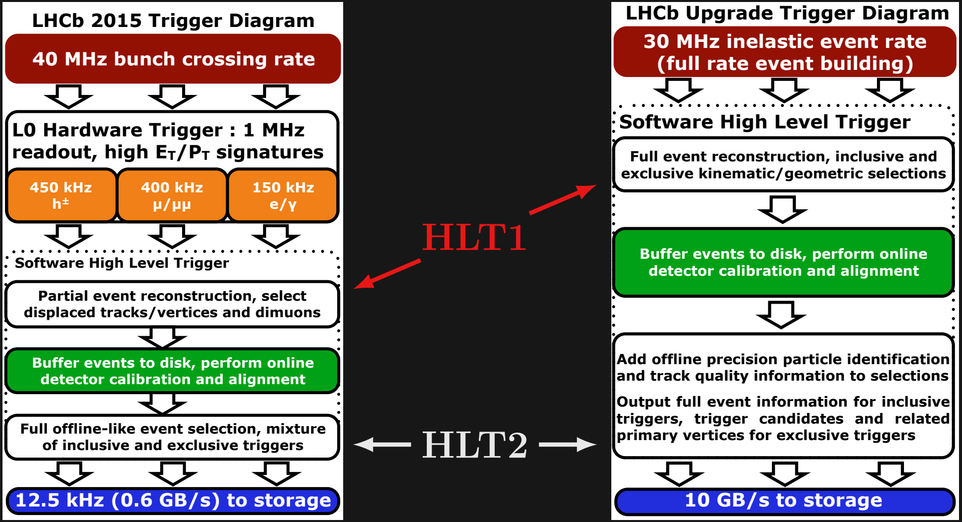




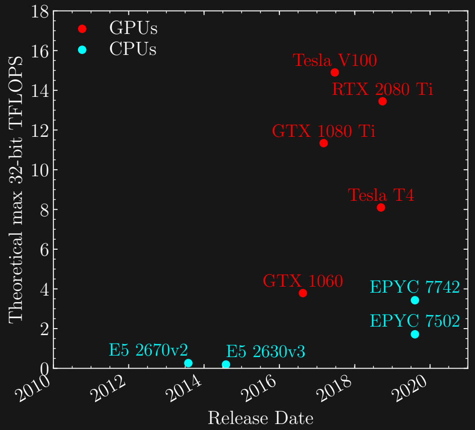
LHCb-TDR-016

- Heavy flavor decays produce displaced low-p_T tracks
- Characteristic signal is a displaced secondary vertex
- Requires information from the entire tracking system
- Solution: read out the full detector at 40 MHz in Run III

The Evolution of the LHCb Trigger



Why make a GPU trigger?



- GPUs offer more theoretical FLOPS* in a compact package
- Lower cost per theoretical FLOPS
- Many HLT1 tasks are inherently parallel

* FLOPS aren't everything. LHCb also has a viable CPU HLT1 for Run III. See Louis Henry's talk: A 30 MHz software trigger and reconstruction for the LHCb upgrade

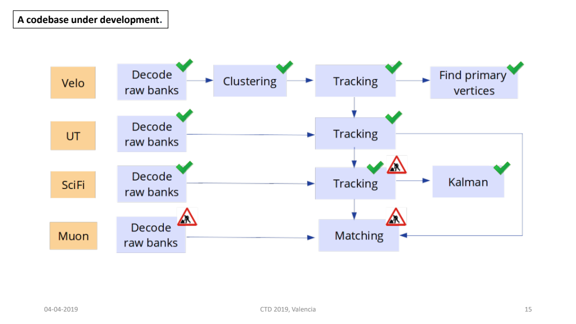
The Allen Project

Frances E. Allen

- Project began in February 2018: gitlab.cern.ch/lhcb/Allen
- Standalone application requiring only C++17 and CUDA v10.2
- First publication accepted: arxiv:1912.09161

What's new since CTD2019?

Allen in April 2019



Brij Kishor Jashal's talk from CTD2019

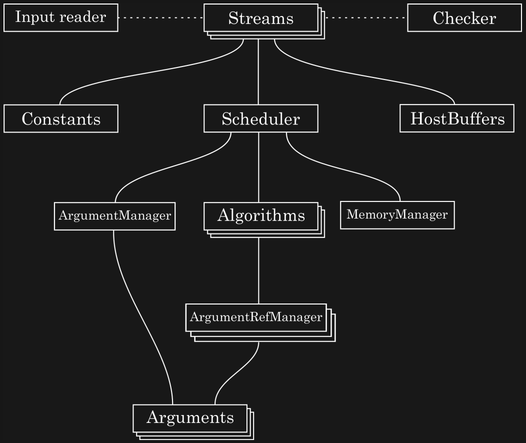
Since then...

- All reconstruction algorithms completed
- Added trigger selections and output
- Huge gains in throughput
- Improved scalability and configurability

We have a complete HLT1 on GPUs!

- Allen reviewed as a viable option for LHCb's HLT1 in Run 3
- HLT1 technology decision in progress

Allen is for Everyone



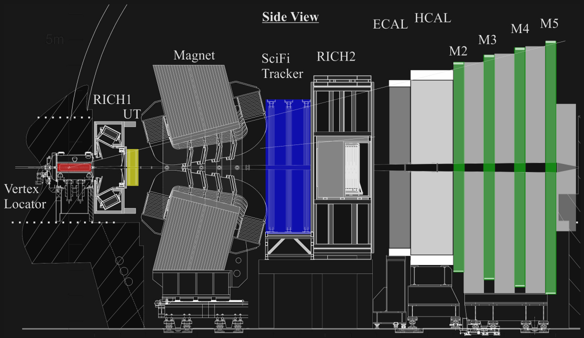
Allen isn't just for GPU experts

- Custom memory manager and scheduler hide some tricky parts of CUDA development
- Can be compiled for CPU or GPU
- Most of the ~ 15 Allen developers are students

Allen isn't just for LHCb

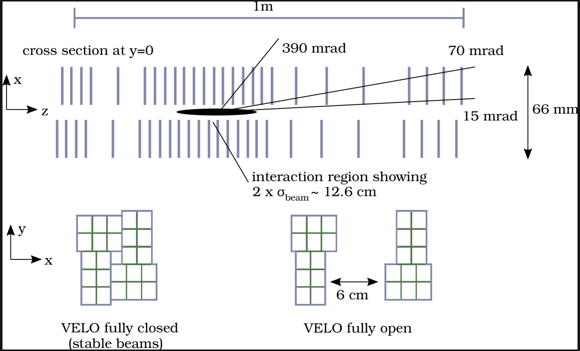
- Allen could easily host non-LHCb algorithms
- Could serve as a platform for other high-throughput GPU applications

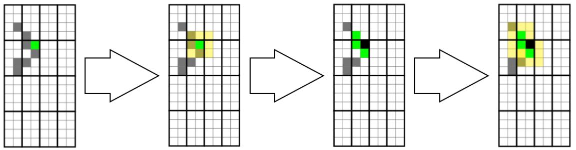
Reconstruction in HLT1



- Decode data from the **VELO**, **UT**, **SciFi**, and **Muon** systems
- Cluster detector data into "hits"
- Build tracks (**VELO**, **UT**, and **SciFi**)
- Find primary vertices (PVs) (**VELO**)
- Match tracks to **Muon** hits
- Fit tracks with a (fast) Kalman Filter
- Make 2-track secondary vertices
- Perform trigger selections

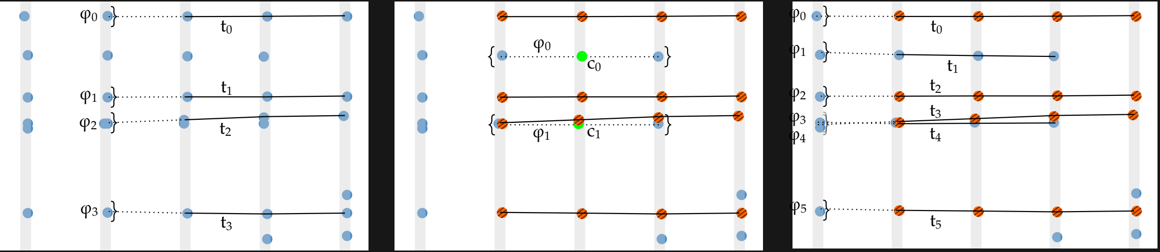
The VELO Detector

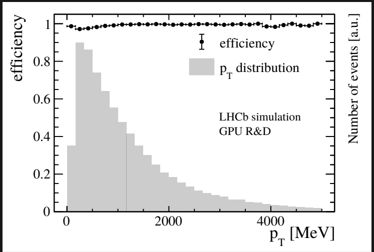




- 26 layers of silicon pixel detectors
- Crucial for primary and secondary vertex finding
- Cluster in constant time using bit masks

VELO Tracking

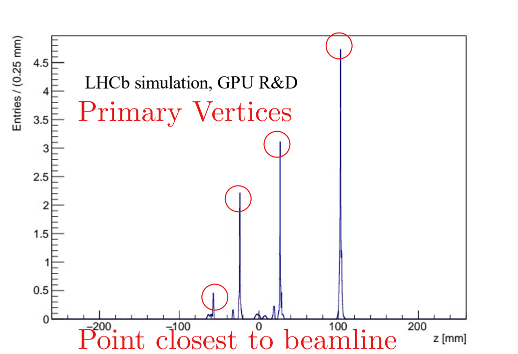


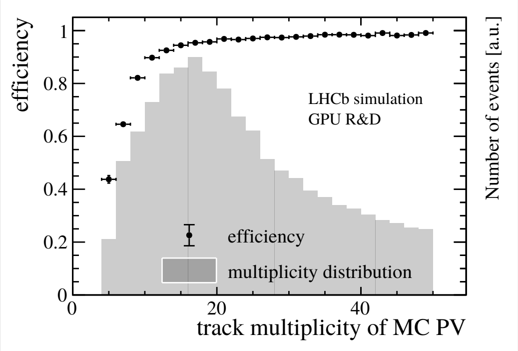


- Sort hits by ϕ
- Create triplets → forward triplets → repeat

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

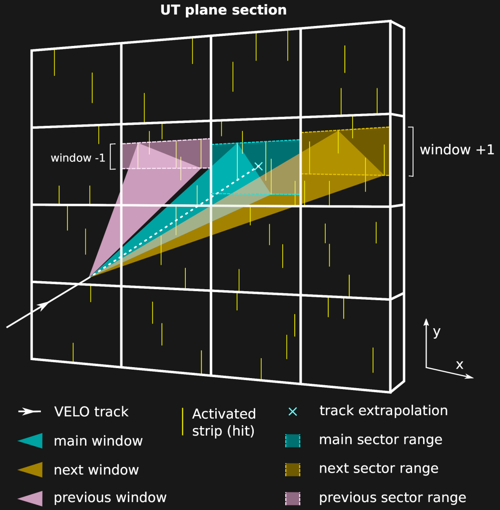
PV Finding



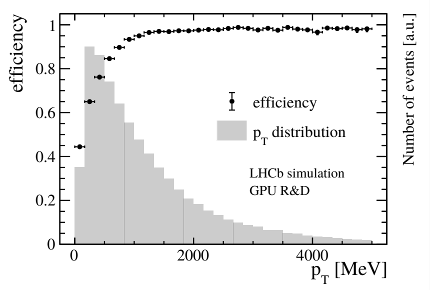


- See Florian Reiss's talk for more info: Fast parallel Primary Vertex reconstruction for the LHCb Upgrade
- See Marian Stahl's talk for more info on a deep learning approach: An updated hybrid deep learning algorithm for identifying and locating primary vertices

UT Tracking

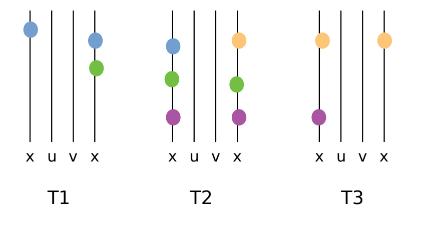


- 4 layers of silicon strips
- Use extrapolated VELO tracks to determine search regions
- Provides initial momentum estimate for extrapolating to SciFi

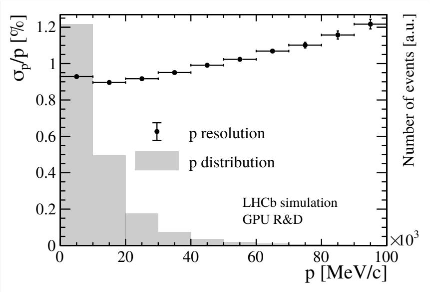


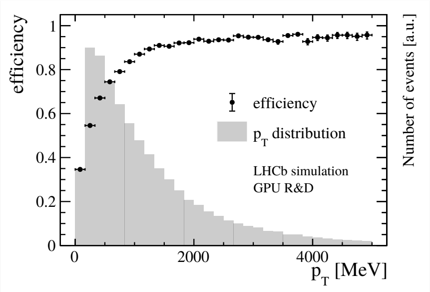
Fernandez Declara, D. Campora Perez, J. Garcia-Blas, D. vom Bruch, J. Daniel Garca, N. Neufeld , IEEE Access 7 (2019)

SciFi Tracking

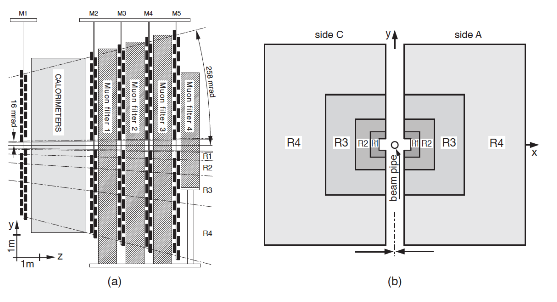


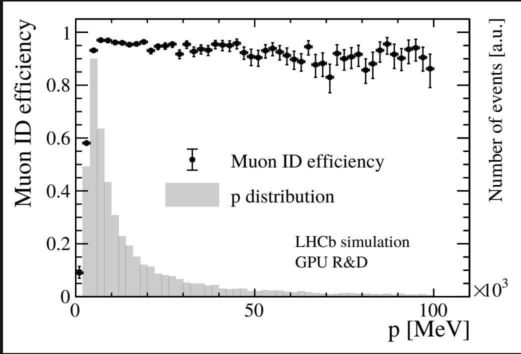
- 12 layers of scintillating fibers
- Reconstructs tracks with $p > 3$ GeV (minimum required for muon ID)
- No p_T requirement ($p_T > 500$ MeV threshold used in Run 2)





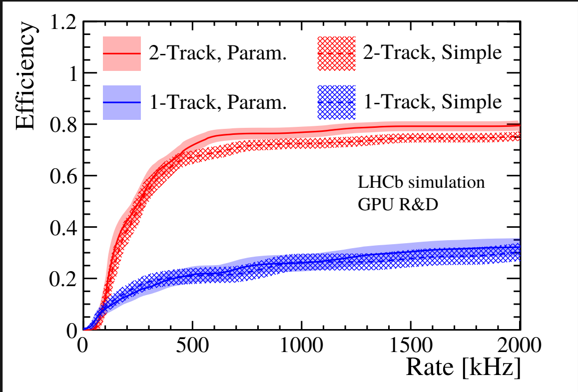
Muon Matching





- Match forward tracks to hits in Muon stations
- Same algorithm LHCb has used since Run I. See here for more information

Kalman Filter



- Simple: No momentum information
- Param.: Uses momentum from forward tracking in noise calculation

- Fast VELO-only Kalman Filter
- Improves track description at position closest to beamline
- Better impact parameter (IP) resolution
- Better descrimination between prompt and displaced tracks
- Takes $\mathcal{O}(1\%)$ of the total sequence time

Selections

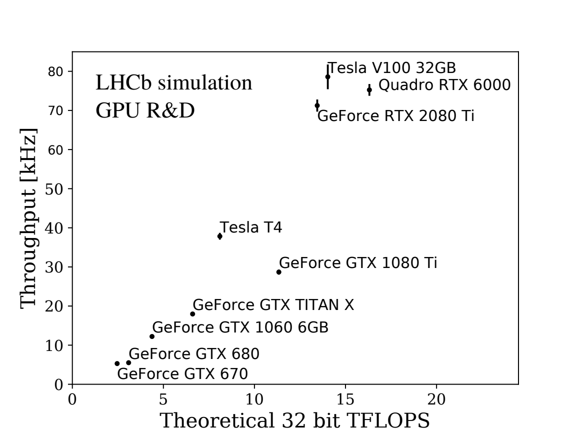
Trigger	Rate [kHz]
1-Track	215 ± 18
2-Track	659 ± 31
High-p_T muon	5 ± 3
Displaced dimuon	74 ± 10
High-mass dimuon	134 ± 14
Total	999 ± 38

- Trigger on 1- and 2-track candidates
- Prototype selections cover most LHCb physics
- Replacing cut-based selections with machine learning models will reduce rates
- Allen can handle $\mathcal{O}(100)$ selections with minimal impact on throughput

Signal	GEC	TIS -OR- TOS	TOS	GEC $\times$ 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

GEC: Global Event Cut, TIS: Trigger Independent of Signal, TOS: Trigger On Signal

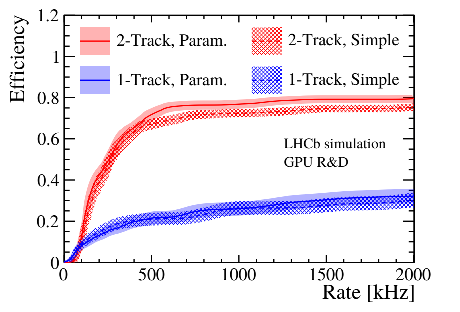
Performance



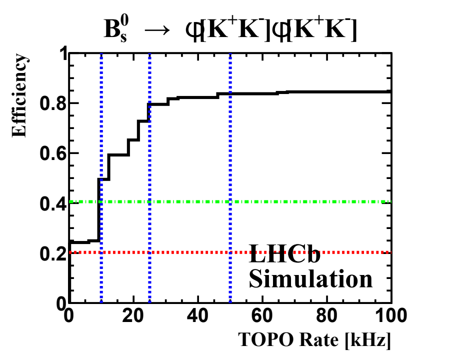
- Can handle the full 30 MHz collision rate with $<$ 500 RTX 2080 Ti GPUs from 2018
- Throughput is approaching results quoted at CTD2019, but those were missing
  - SciFi tracking
  - Muon decoding and matching
  - Kalman filter
  - Trigger selections
- Throughput scales well with theoretical TFLOPs, so Allen will speed up as GPUs improve

Multi-track vertices

- Allen can reconstruct forward tracks with no p_T requirement
- Allows for efficient triggering using 3- and 4-track vertices
- Could lead to totally new trigger strategies



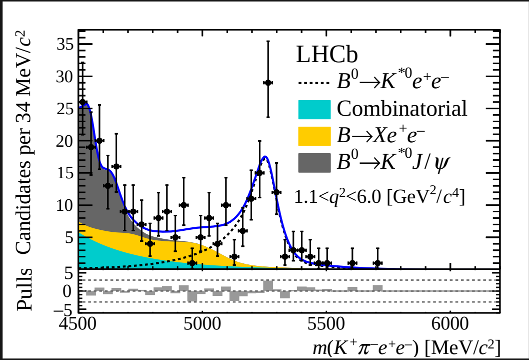
Vs.



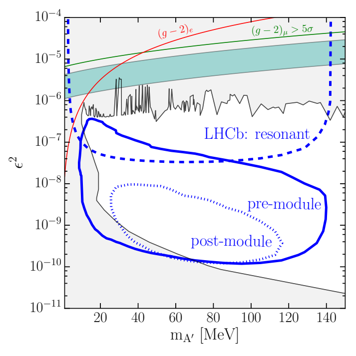
Future Prospects

Calorimeter clustering in HLT1

- Perfect task for GPUs
- Electron identification in HLT1
- Many interesting measurements use electrons, e.g. $R(K^*)$, $A' \to e^+e^-$



JHEP 1708 (2017) 055



Phys. Rev. D92 (2015) no. 11, 115017

Conclusions

Allen is the first implementation of a full software trigger stage on GPUs

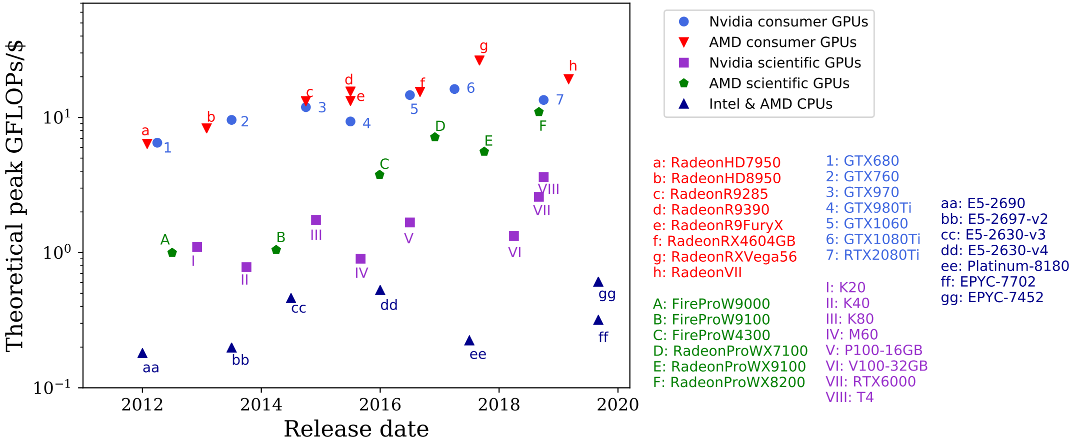
- LHCb's baseline HLT1 has been implemented on GPUs
- Optimizations and improvements continue

Allen could allow LHCb to expand its Run III physics program

- Speeding up HLT1 allows it to handle additional tasks
- Improved algorithms could lead to an overhauled trigger strategy
- GPUs will continue to improve before Run III begins, opening up more possibilities

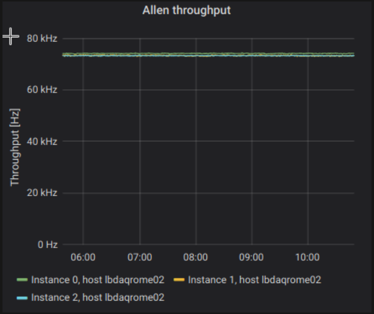
Backup

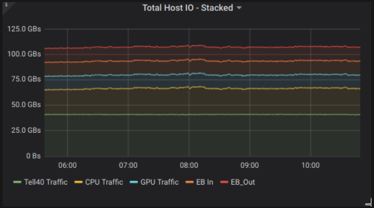
GPU FLOPS/USD

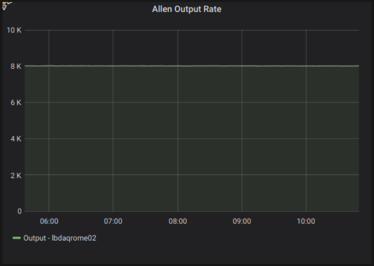


Courtesy of Dorothea vom Bruch, arXiv:2003.11491

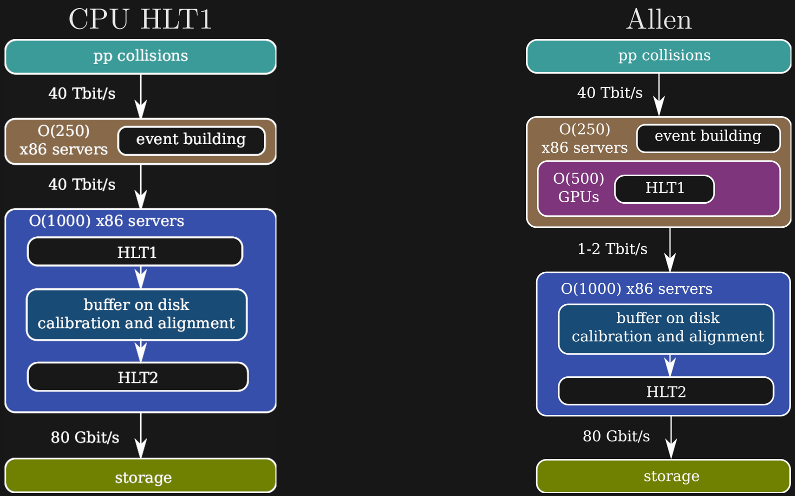
Integration and stability tests







Adding GPUs to the LHCb DAQ



GPUs fit naturally into the LHCb DAQ

Make up cost of GPUs with savings on networking