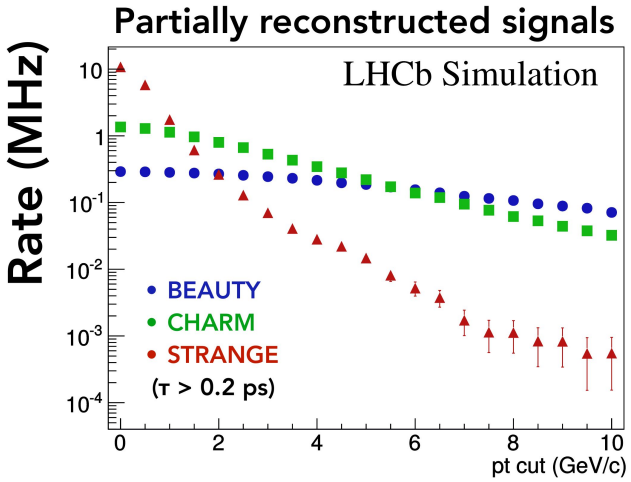
Allen: Processing 4 TB/s of Streaming Data From the LHCb Experiment on GPUs.

Roel Aaij

October 19th, 2021

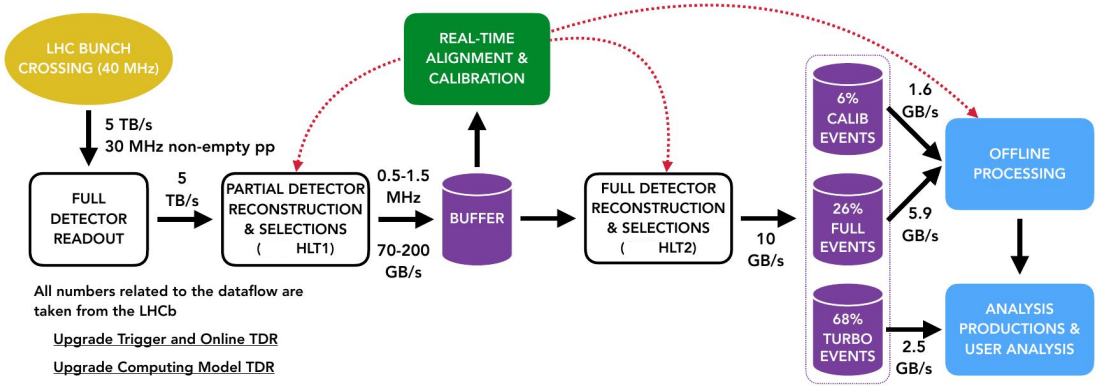
Nikhef

LHCb Upgrade in a Single Slide



30 MHz (4 TB/s) of input contains a MHz of signal, while we can only store 10 GB/s long-term

LHCb Upgrade Dataflow



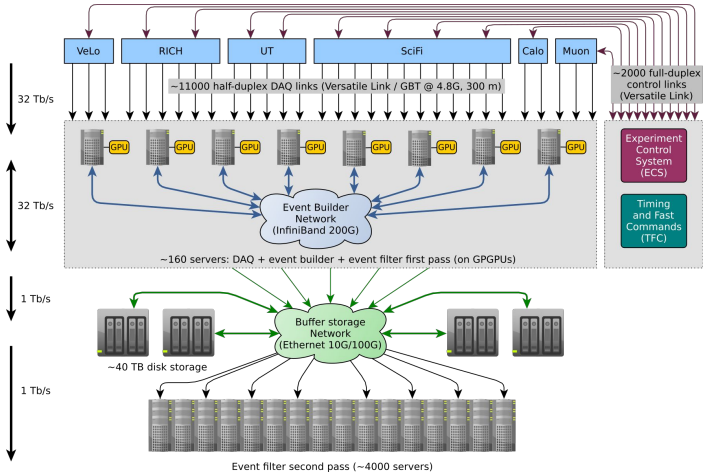
All numbers related to the dataflow are taken from the LHCb

Upgrade Trigger and Online TDR

Upgrade Computing Model TDR

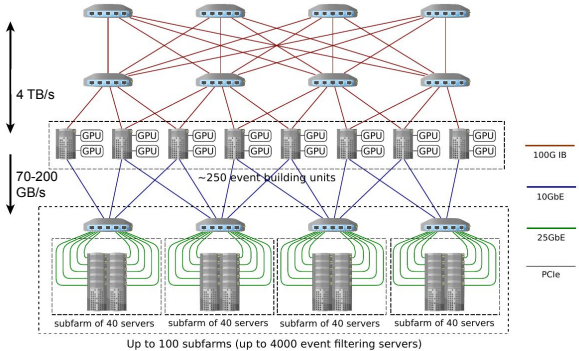
HLT1 challenge: reduce 5 TB/s to 70-200 GB/s in real-time with high physics efficiency

DAQ Architecture

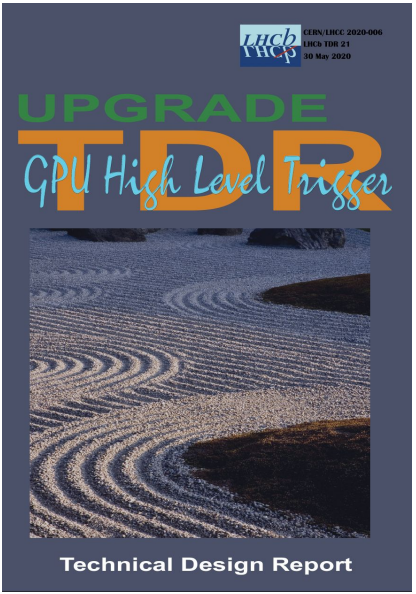


30 MHz (5 TB/s) of event building and processing in a data center

HLT1 on GPUs: Allen



GPU solution (Allen) selected as baseline HLT1; up to 3 GPUs installed in each event builder server



GPU HLT1 TDR

HLT1 on GPUs: Allen

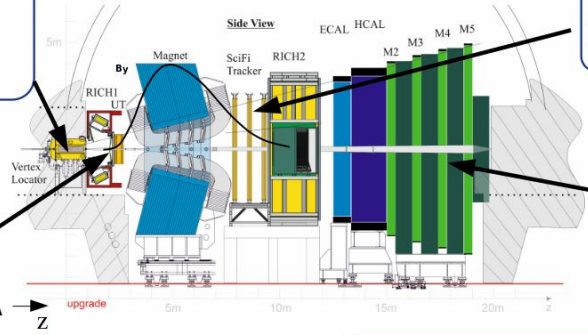
- Fully standalone software project: https://gitlab.cern.ch/lhcb/Allen
- Dependencies: C++17 compliant compiler, boost, ZeroMQ, cppgsl
- Built-in physics validation
- Configurable algorithm sequence, custom memory manager
- Cross-architecture compatibility (CPU, CUDA, HIP)
- Approximately 100k LOC, 90% written from scratch
- Integrated with LHCb stack and LHCb DAQ and control system

- Project started in February 2018
- After 15 months of development time:
- project reviewed as viable solution for Run 3 (starting in 2022)
- Accepted as baseline solution in May 2020

Reconstruction and Selection

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

UT
- Decode raw data
- Track reconstruction



SciFi
- Decode raw data
- Track reconstruction

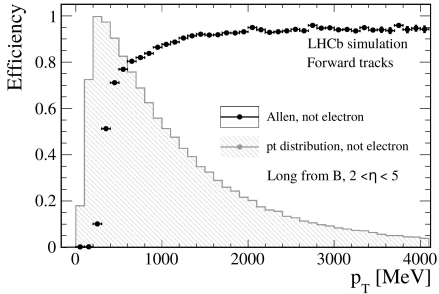
Muons
- Decode raw data
- Match hits to tracks

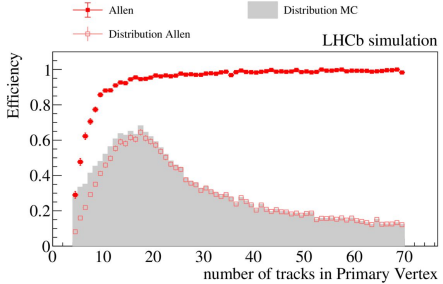
Track fit: Kalman filter

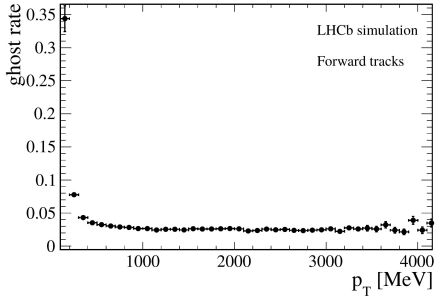
Find secondary vertices

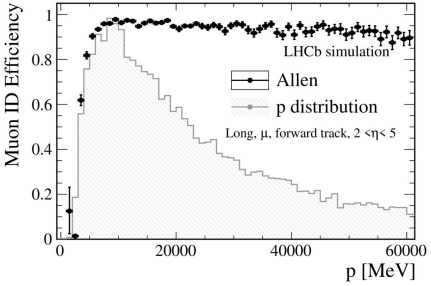
Selections
- 1-track selection
- 2-track selection
- Based on p, p_t, displacement, vertex criteria and muon identification

Reconstruction Performance









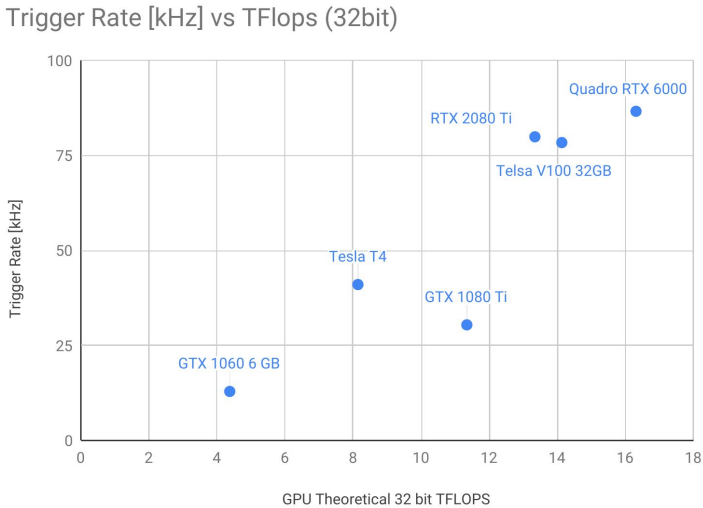
LHCb-FIGURE-2019-009

Selection 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

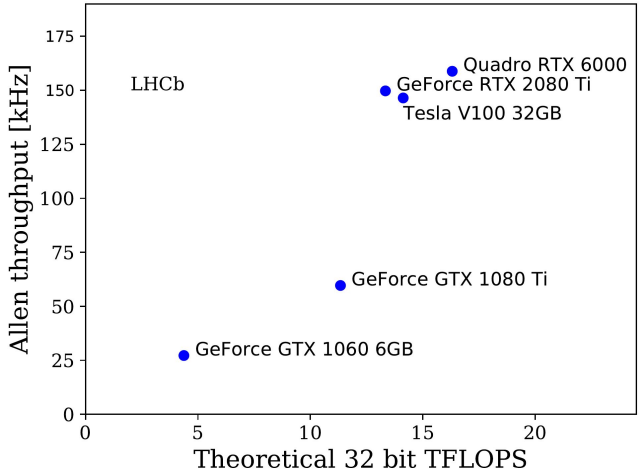
Throughput Evolution Since TDR



GPU HLT1 TDR

Performance at the time of the TDR;
Approximately linear scaling with theoretical FLOPS

Throughput Evolution Since TDR



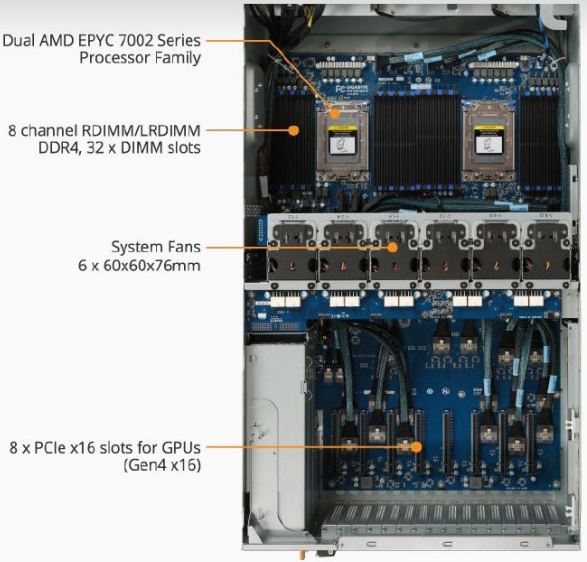
LHCb-FIGURE-2019-009

(More) Recent Improvement

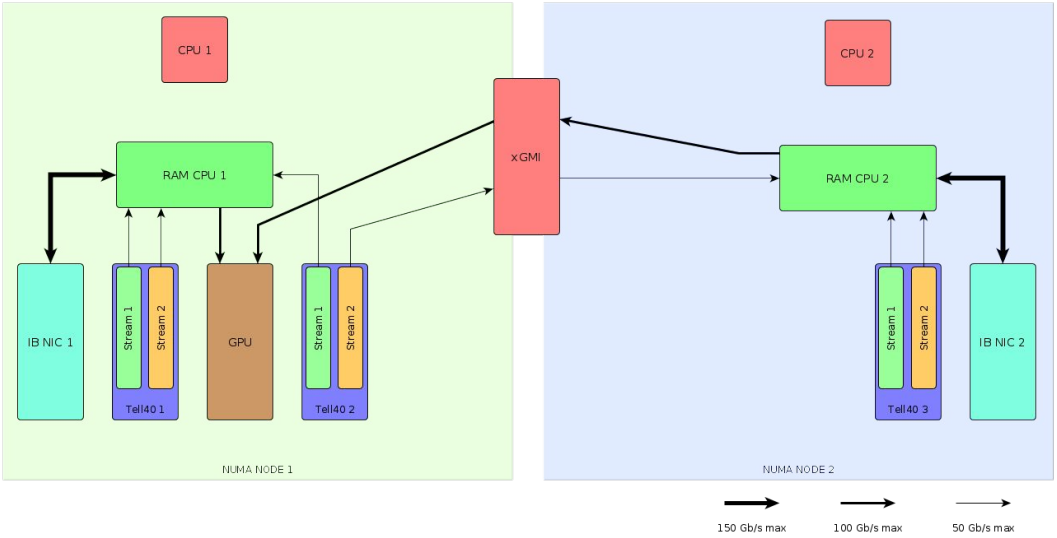
Integration

- DAQ:
  - Data input (event-builder layout)
  - Data output (event-by-event layout)
- Avoid large-scale transposition of data
  - Process data in the layout provided by the event builder
  - Batches of 30k grouped by frontend, not by event
  - Process in batches of 1k events
- Steering by the experiment control system
- Error handling and failover
- Detector information such as geometry and alignment
  - Obtain from “regular” LHCb stack on the fly
  - Deal with changing conditions
- Monitoring for data quality and shifters
- HLT1 hardware and processes share the server with event building: keep a close eye on CPU and memory usage

Event Builder Server (Gigabyte G481-Z51)



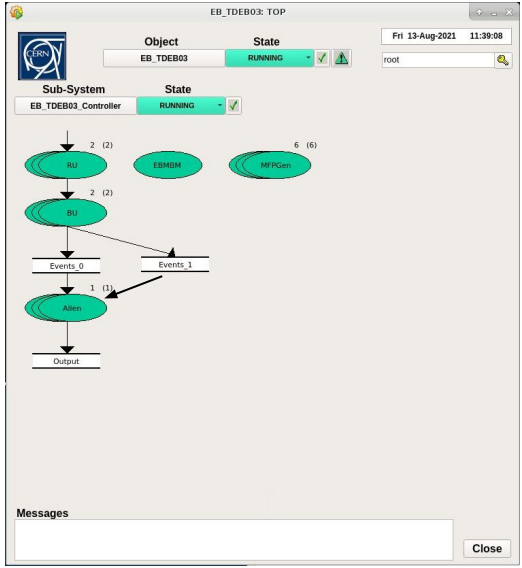
Event Builder Server



Throughput Test

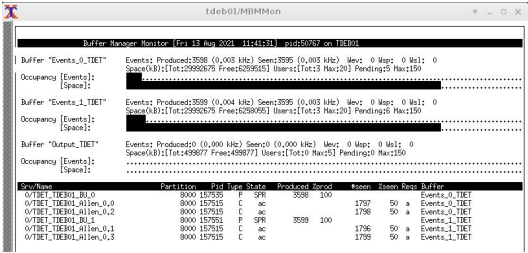
- 10 event builder nodes
- event builder generates data with scaled up events
- HLT1 consumes generated data on (single) GPU-equipped node
- HLT1 without reconstruction, only copy data to and from GPU
- Frontends in data-generator mode

- Goal: verify that data I/O and throughput requirements are met
- Goal: Verify that a single GPU works



Throughput Test

- 10 event builder nodes
- event builder generates data with scaled up events
- HLT1 consumes generated data on (single) GPU-equipped node
- HLT1 without reconstruction, only copy data to and from GPU
- Frontends in data-generator mode



- Test successful: can reach 43 MHz
- Bottleneck: PCIe throughput to GPU
- Important takeaway: NUMA and and (device) runtime considerations very important
- Need more monitoring

Summary

- GPU HLT1 (Allen) selected as the baseline solution for LHCb
- Hardware has been purchased (RTX A5000)
- Work continues to improve performance; both physics and throughput
- Shifting to commissioning mode: focus on integration, consolidation and testing
- Careful consideration of dataflow - memory and networks - crucial to success
- Larger-scale integration tests a success
- Exciting times ahead!