Le défit de 40 Tbit/s

Analyse de données structurées en temps réel pour l'expérience LHCb au CERN

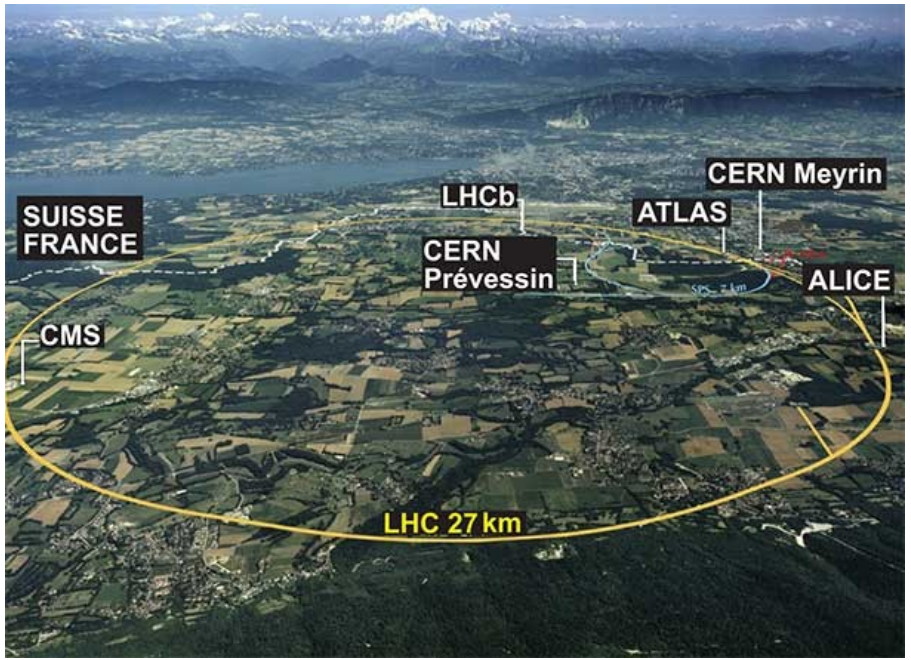
Dorothea vom Bruch
CPPM Marseille

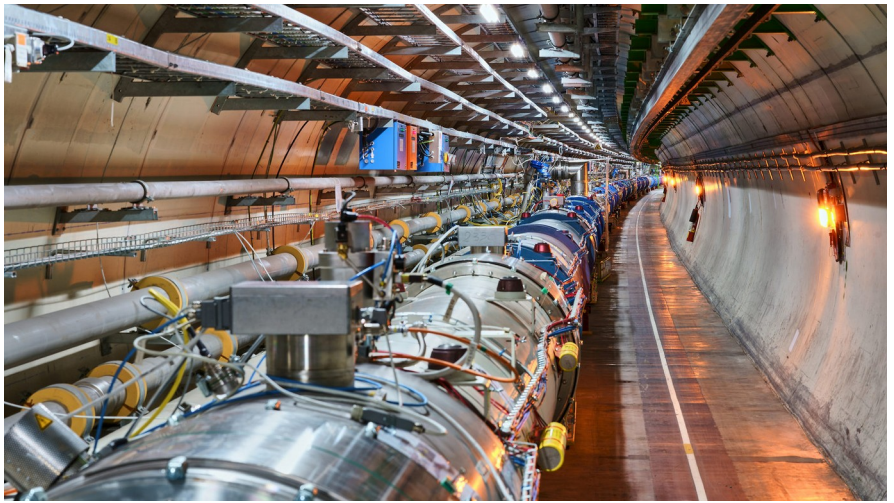
Journée thématique CEDRE
Décembre 1er 2022
Campus Saint-Jérôme

Outline

- Real-time data analysis in particle physics
- Trend towards heterogeneous computing systems
- Computing challenge: Analyze 40 Tbit/s of data in real-time at the LHCb experiment @ CERN
  - Analyze data on Graphics Processing Units (GPUs)
  - Data structure, algorithms processed in real-time
  - Parallelization strategy
  - Commissioning of system in 2022

LHC @ CERN

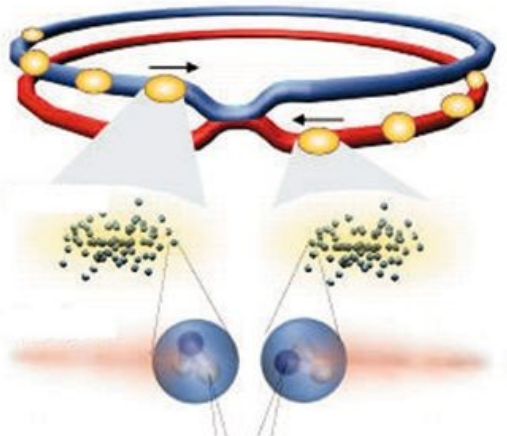

SUISSE
FRANCE
LHCb
CERN Meyrin
ATLAS
CERN
Prévessin
ALICE
CMS
LHC 27 km

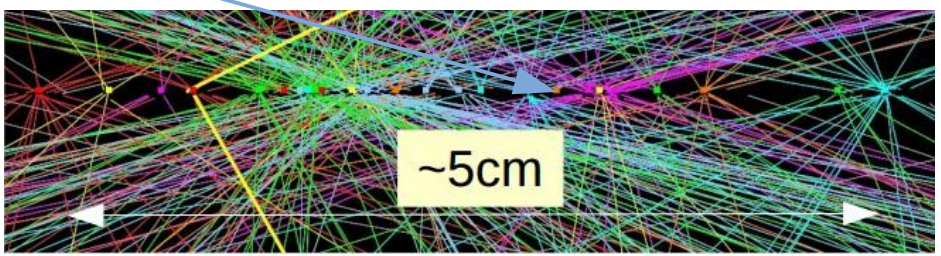


Particle collisions

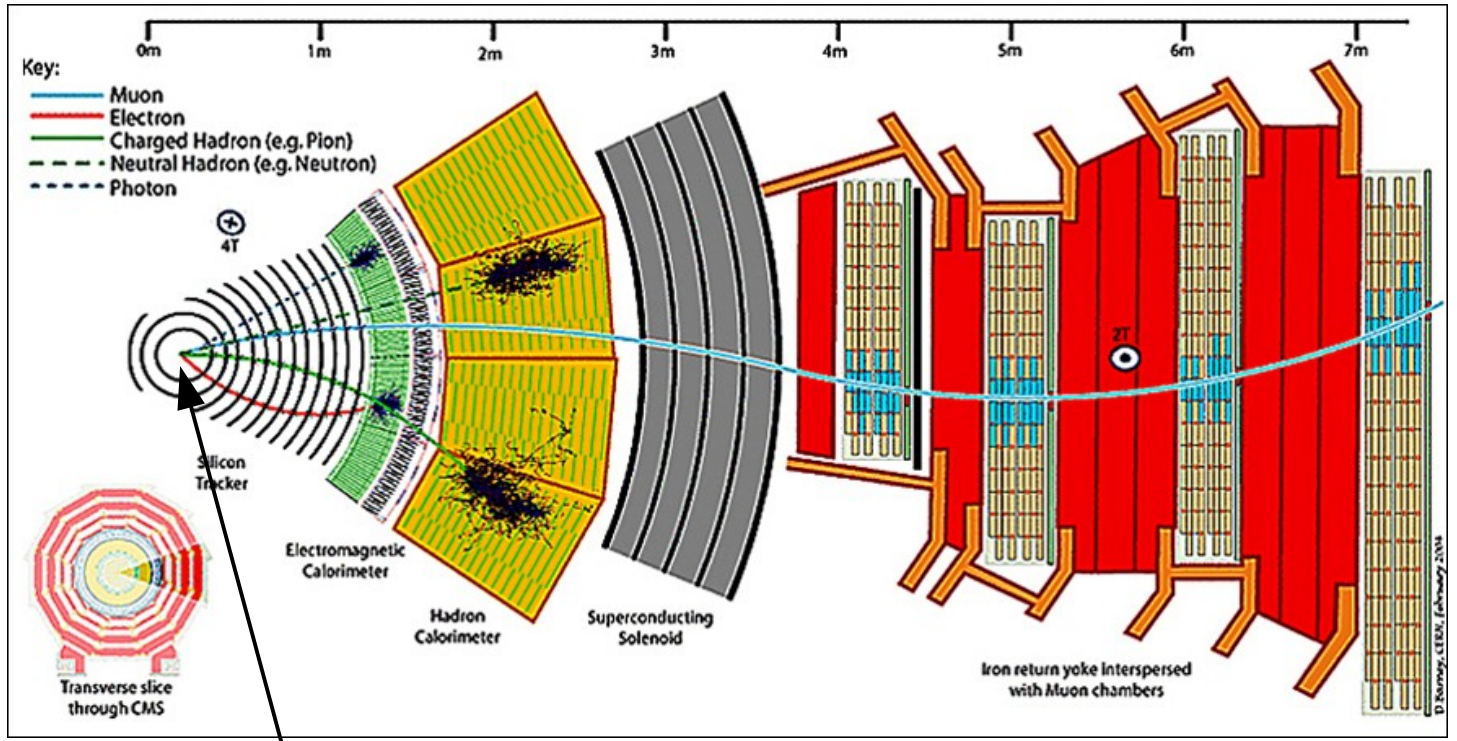
- Two beams of proton bunches in opposite directions
- One bunch crossing of the two beams every 25 ns at the four large LHC experiments

  → “Event”
- The proton-proton collisions occur in a region spread along the beamline
- The position of one proton-proton collision is called primary vertex (PV)





Typical particle detector



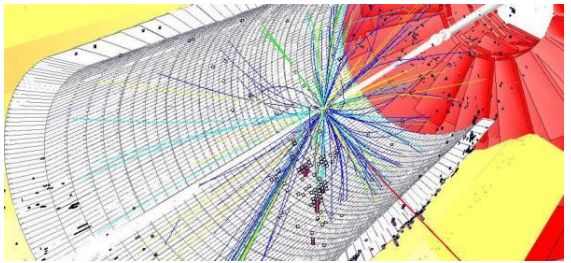
Data challenges in particle physics

Detector

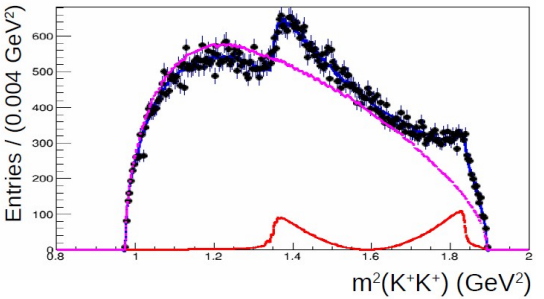
Selection

Storage

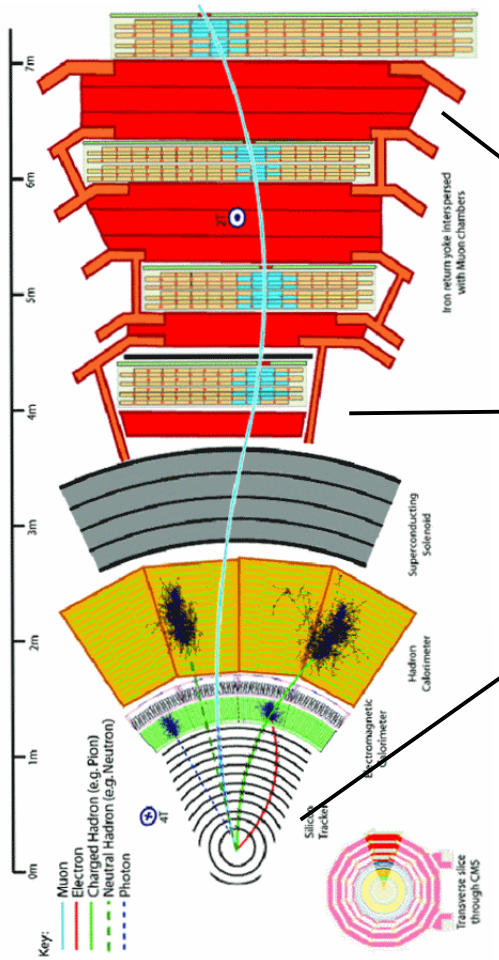
Simulation



Data analysis

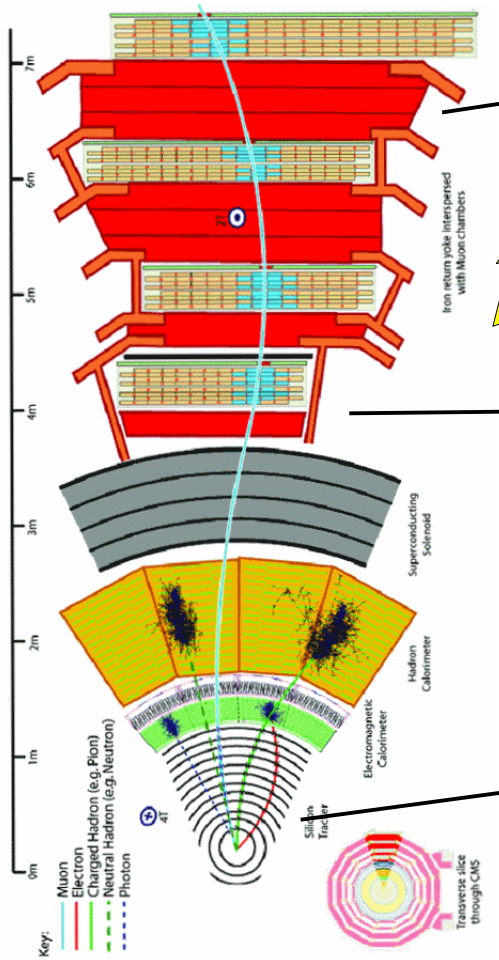


“Trigger”: Real-time data analysis and reduction



accumulate → analyze → reduce

“Trigger”: Real-time data analysis and reduction



analyze → reduce

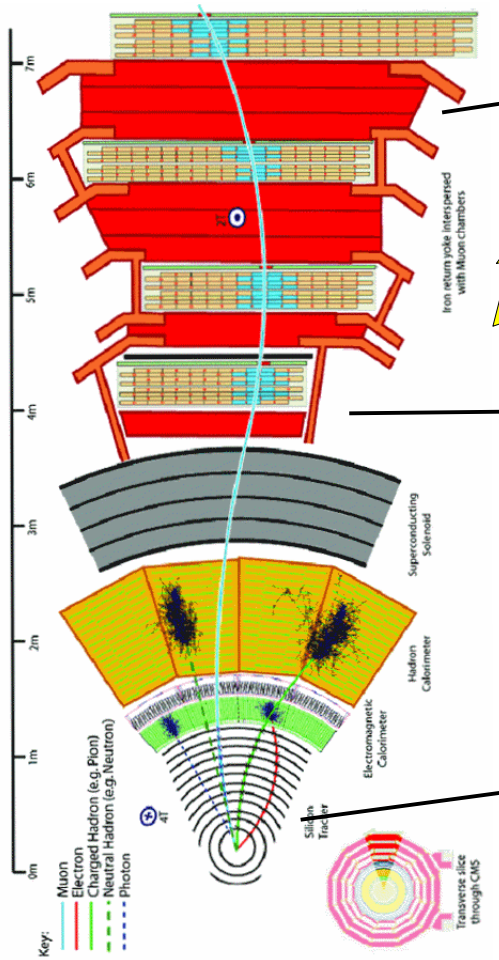
Large data rate

accumulate

analyze → reduce

analyze → reduce

“Trigger”: Real-time data analysis and reduction



Large data rate

analyze → reduce

accumulate

analyze → reduce

analyze → reduce

First: Hardware trigger

- Data obtained directly from detector
- Decision taken in fixed time, low latency
- Based on local information from a subdetector
- Chip constraints → not too complex calculations

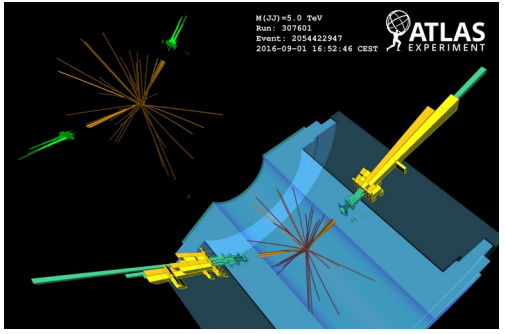
Second: Software trigger

- Data already transferred to a server
- Decision taken with medium latency
- Based on information from several subdetectors
- Processor constraints less stringent

When to use hardware versus software trigger?

Hardware trigger

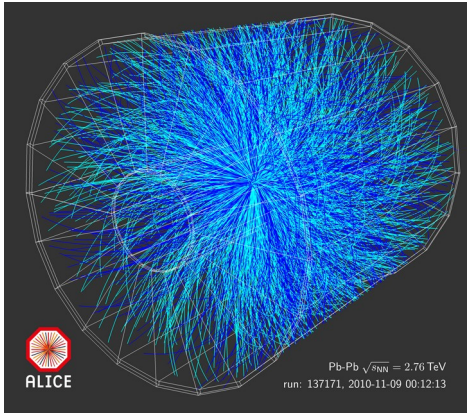
Local characteristic signature,
For example high energy / pt particle



Selection

Software trigger

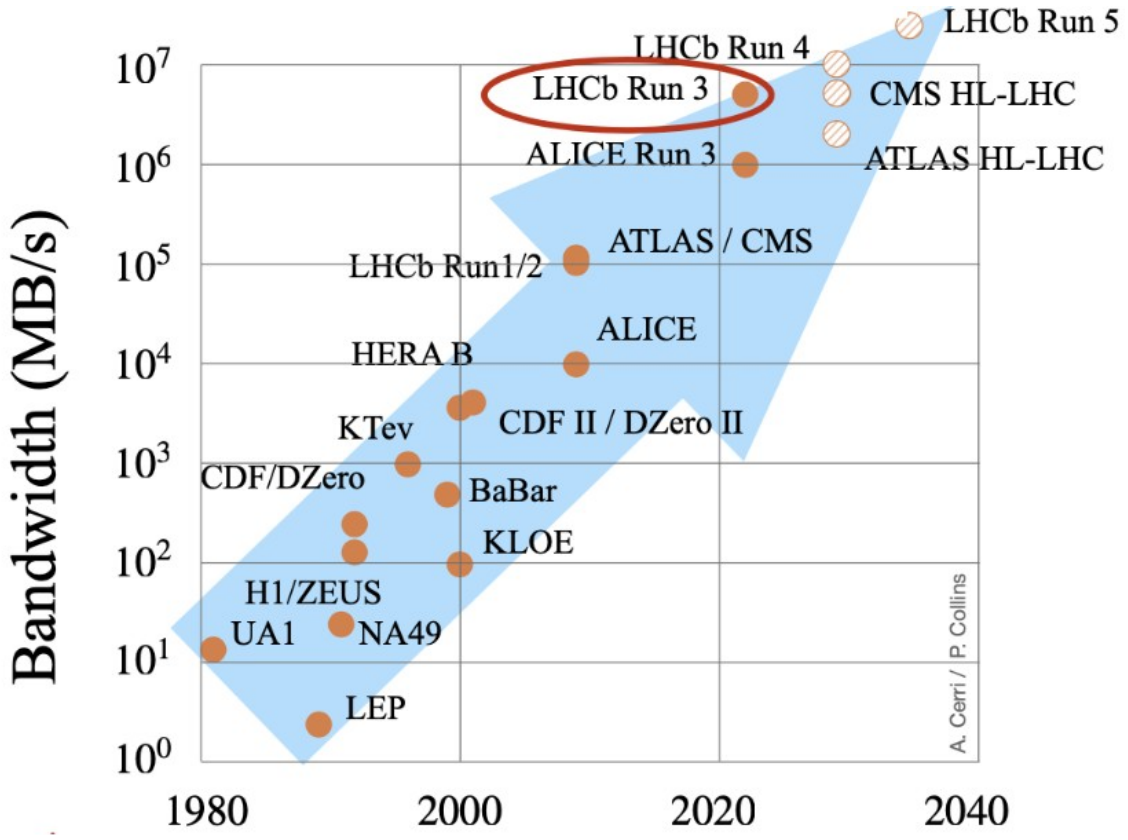
Analysis of whole event required
→ reconstruct all trajectories



Change in trigger paradigm

Access as much information about the collision as early as possible

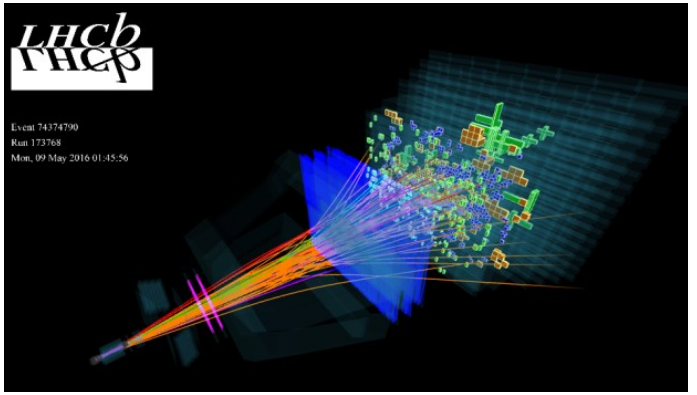
Real-time software challenges



… in the global context

Largest single internet exchange point:
14 Tbit/s

LHCb experiment @ CERN
40 Tbit/s



Computing performance challenge @ CERN

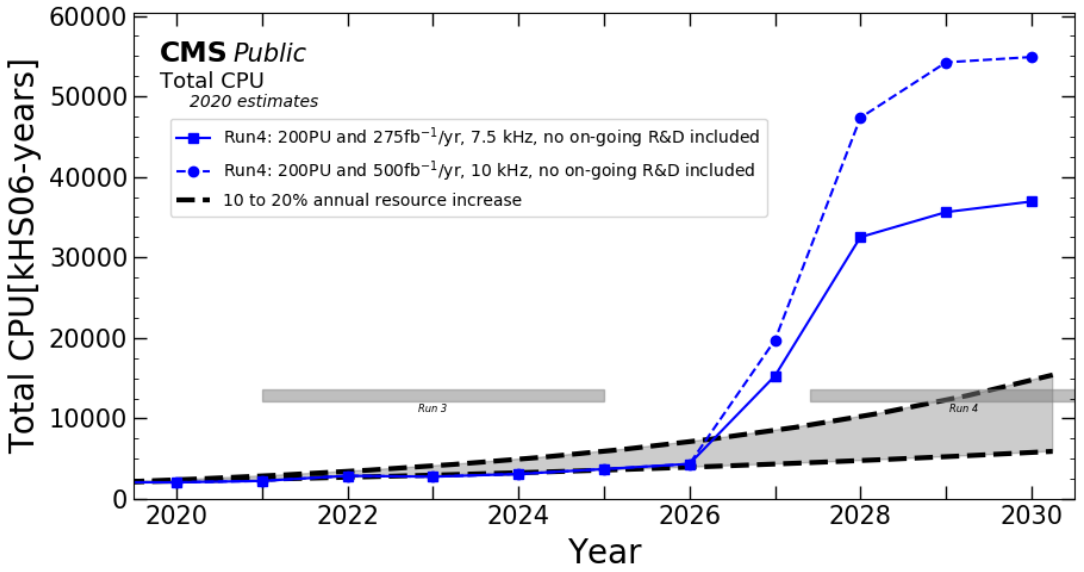
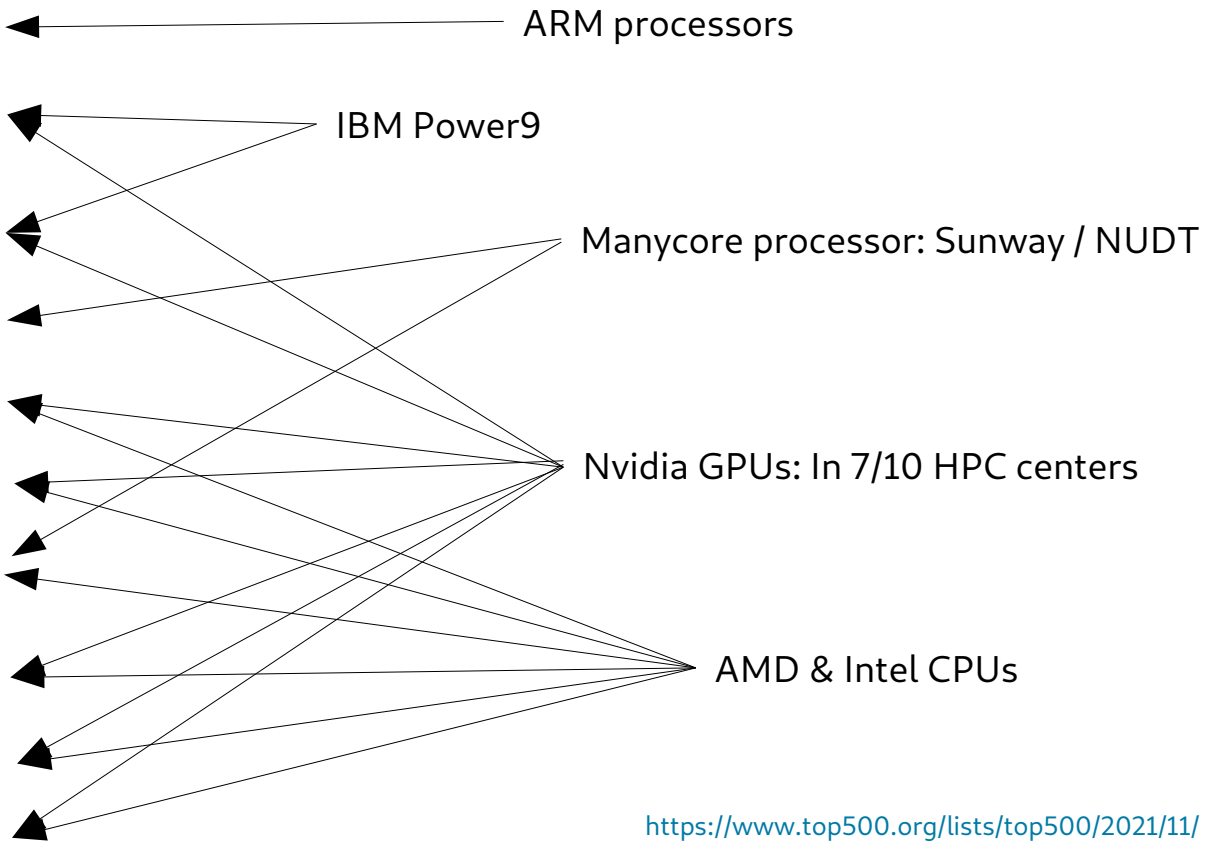


Image source

- In high energy physics, usually assume flat budget for computing cost estimation
- Estimated improvement increase: 10-15% per year for the same budget
- Can no longer count on a stable increase for CPU servers

Trend towards heterogeneous solutions: TOP500

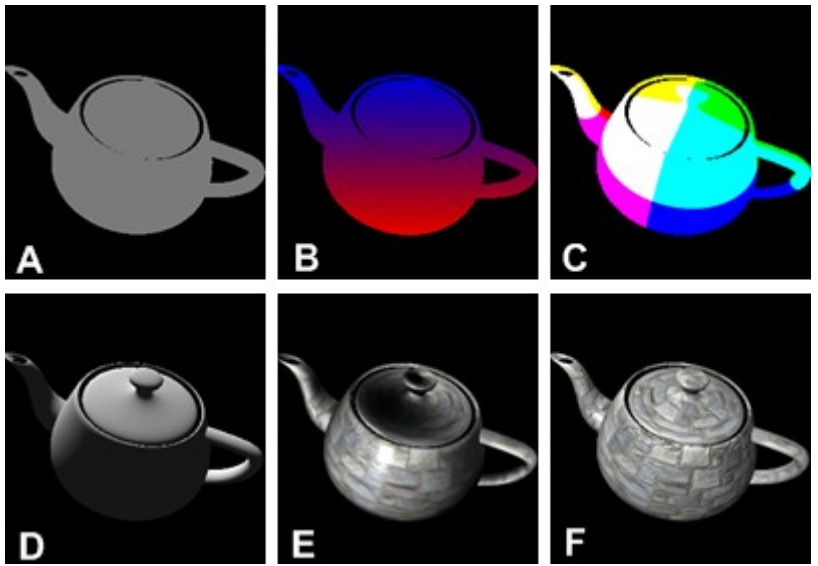
Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, **Fujitsu** RIKEN Center for Computational Science Japan	7,630,848	442,010.0	537,212.0	29,899
2	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, **IBM** DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
3	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, **IBM / NVIDIA / Mellanox** DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
4	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, **NRCPC** National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
5	**Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, **HPE** DOE/SC/LBNL/NERSC United States	761,856	70,870.0	93,750.0	2,589
6	**Selene** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, **Nvidia** NVIDIA Corporation United States	555,520	63,460.0	79,215.0	2,646
7	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000, **NUDT** National Super Computer Center in Guangzhou China	4,981,760	61,444.5	100,678.7	18,482
8	**JUWELS Booster Module** - Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite, **Atos** Forschungszentrum Juelich (FZJ) Germany	449,280	44,120.0	70,980.0	1,764
9	**HPC5** - PowerEdge C4140, Xeon Gold 6252 24C 2.1GHz, NVIDIA Tesla V100, Mellanox HDR Infiniband, **DELL EMC** Eni S.p.A. Italy	669,760	35,450.0	51,720.8	2,252
10	**Voyager-EUS2** - ND96amsr_A100_v4, AMD EPYC 7V12 48C 2.45GHz, NVIDIA A100 80GB, Mellanox HDR Infiniband, **Microsoft Azure** Azure East US 2 United States	253,440	30,050.0	39,531.2	



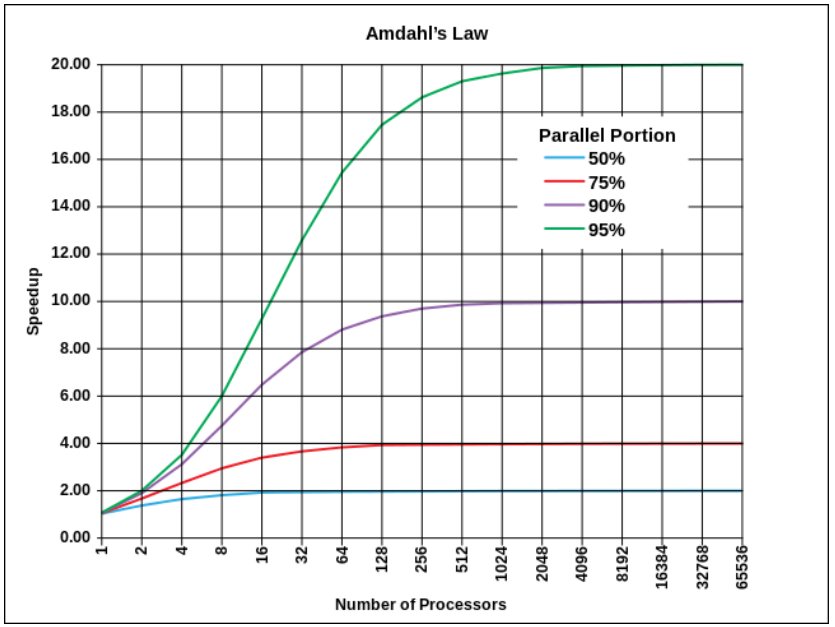
https://www.top500.org/lists/top500/2021/11/

Graphics Processing Unit (GPU)

Developed for graphics-oriented workloads



When to go parallel?



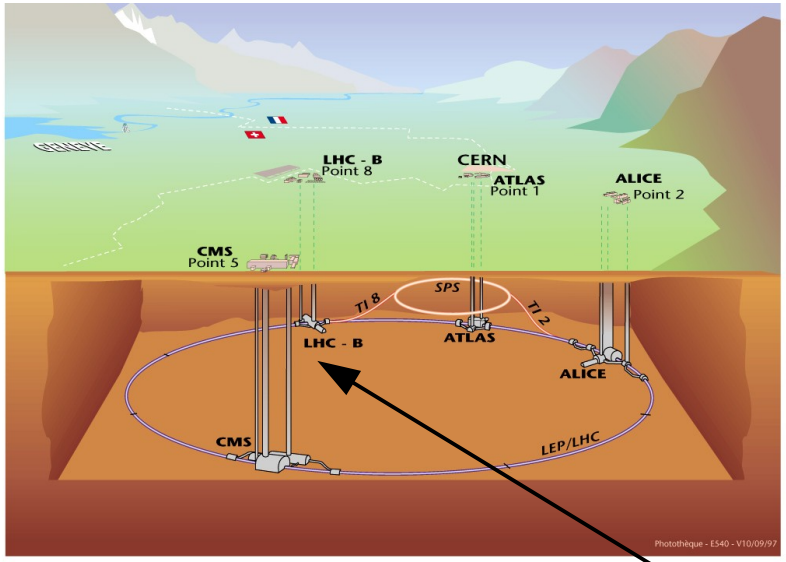
Parallel

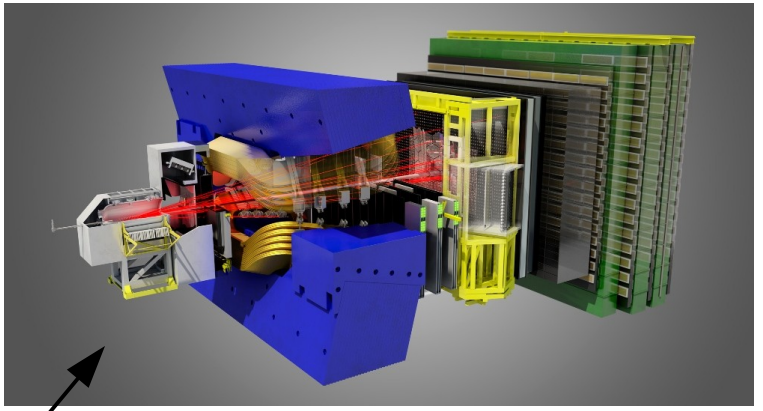
Sequential

Consider how much of the problem can actually be parallelized!

The LHCb experiment at CERN

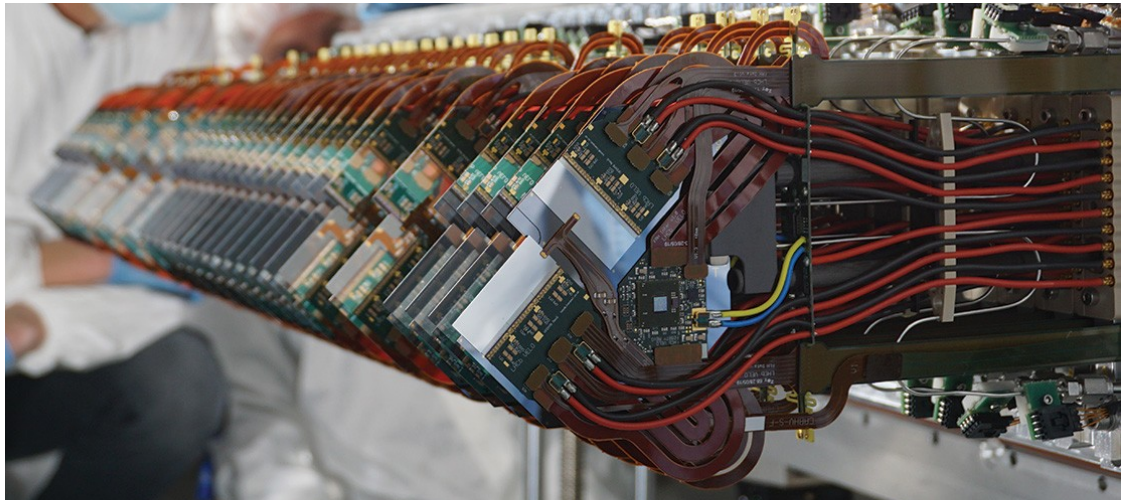
LHC @ CERN





LHCb produces structured data

Example: Vertex locator detector



Data produced by every sensor of the detector looks like this:

```
Sensor number,
layer number,
number of fired
pixels in sensor

Pixel 4, Pixel
48, Pixel 153
```

Header

Payload

Recurrent tasks in real-time data analysis

Raw data decoding

- Transform binary payload from subdetector raw banks into collections of hits (x,y,z) in LHCb coordinate system

Track reconstruction

- Consists of two steps:
  - Pattern recognition: Which hits were produced by the same particle? → "Track"

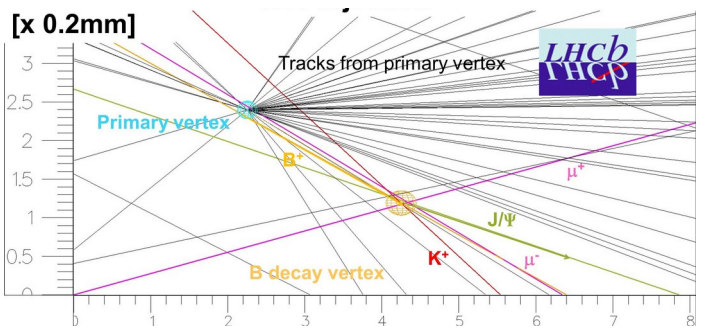
    → Huge combinatorics when testing different combinations of hits
  - Track fitting: Describe track with mathematical model

Vertex finding

- Where did proton-proton collisions take place?
- Where did particles decay within the detector volume?

Calorimeter / muon detector reconstruction

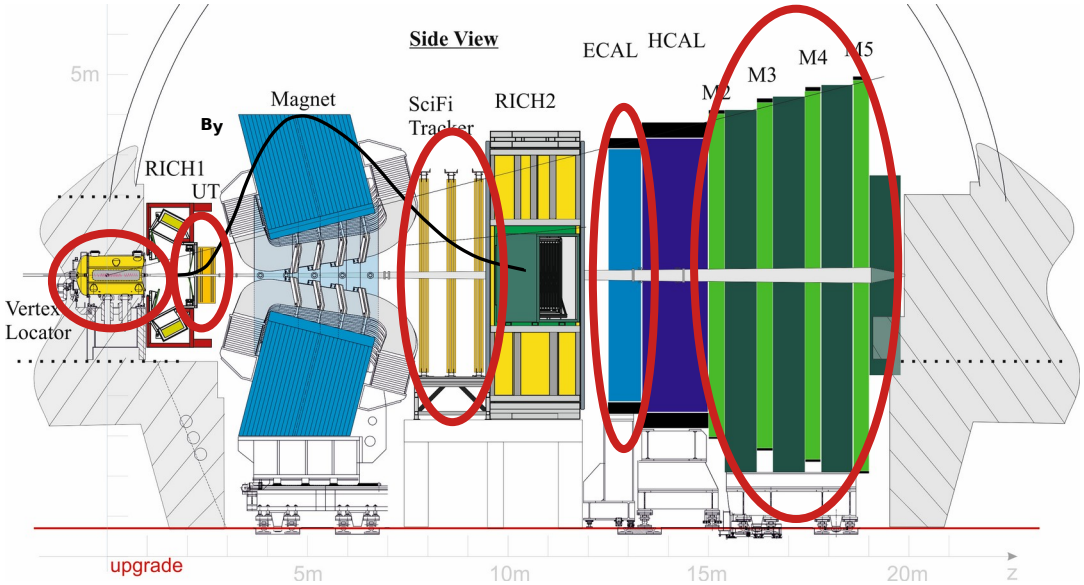
- Reconstruct clusters in the calorimeter / muon detectors
- Match tracks to clusters



LHCb's first level real-time analysis on GPUs

High Level Trigger 1 (HLT1) tasks

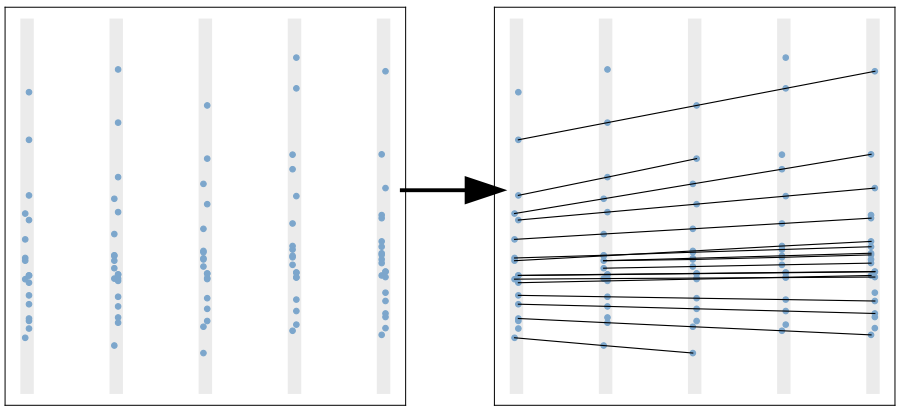
- Decode binary payload of five sub-detectors
- Reconstruct charged particle trajectories
- Identify particle types
- Reconstruct particle decay vertices
- Select pp-bunch collisions to store



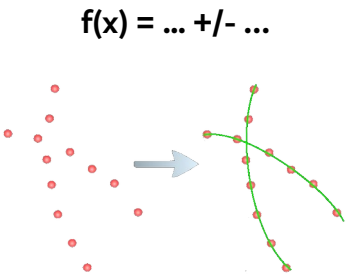
- Manageable amount of algorithms with highly parallelizable tasks
- Ideally suited for parallel architecture of GPUs

Main task: particle trajectory reconstruction

Pattern recognition



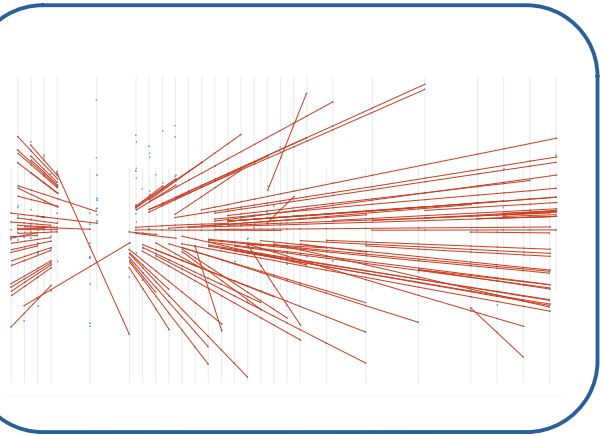
Track fit



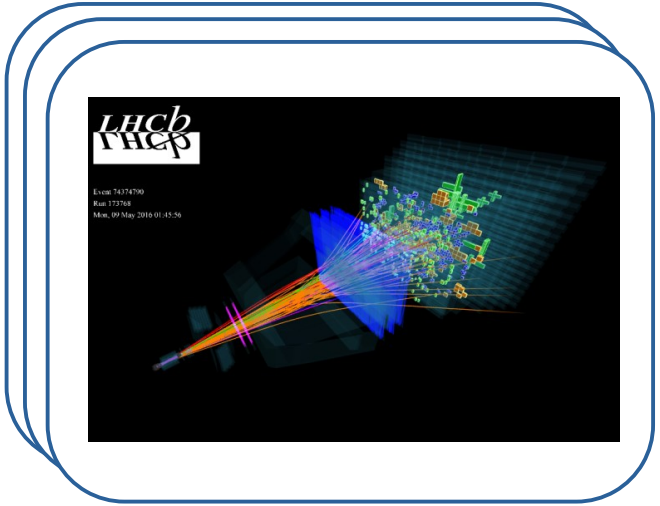
Huge computing challenge for 10^9 – 10^{10} tracks / second

Three levels of parallelization

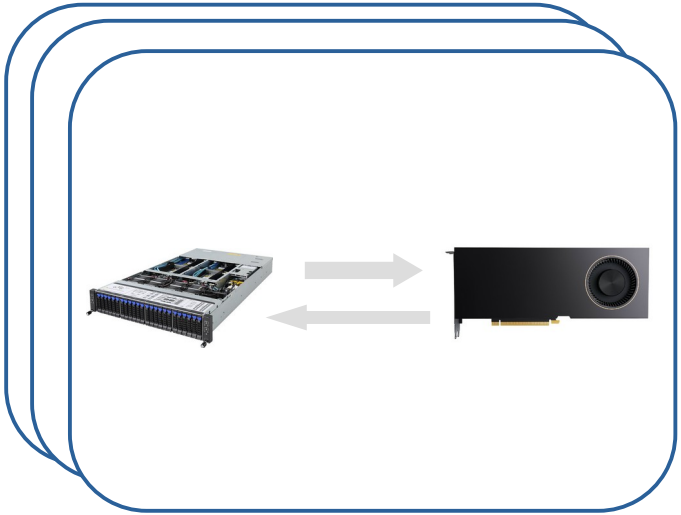
Intra-collision: Tracks, vertices, …



Proton collisions



Collision batches



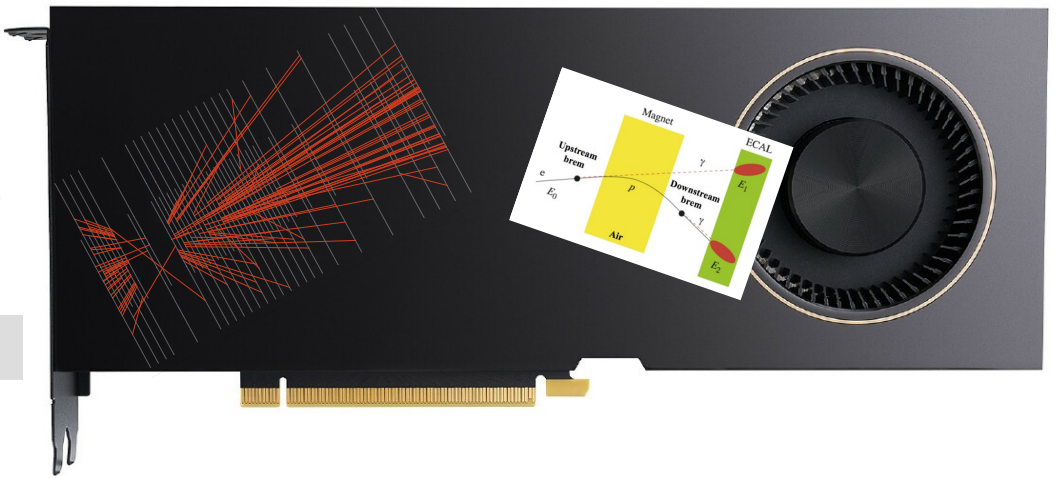
Minimize copies to / from GPU

Server

Raw data

Selection decisions

GPU



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 software project

- Named after Frances E. Allen

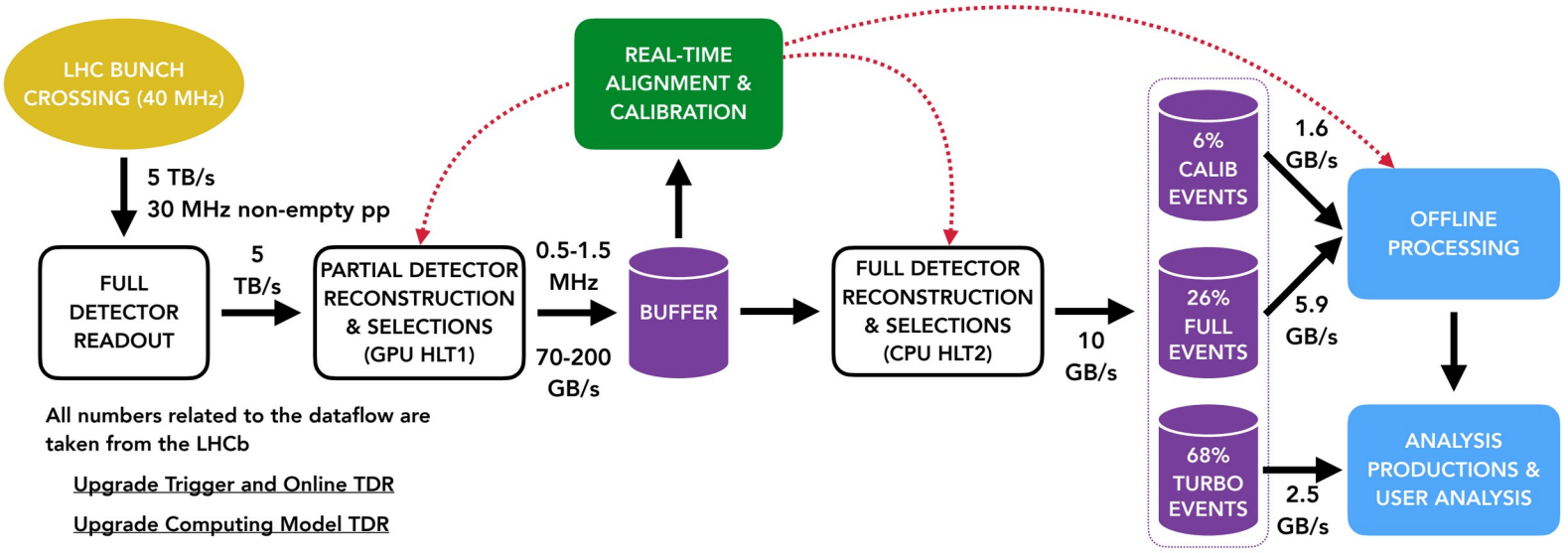
- Fully standalone software project: https://gitlab.cern.ch/lhcb/Allen, Sphinx documentation
- Framework developed for processing LHCb´s first real-time selection stage (HLT1) on GPUs

- Cross-architecture compatibility via macros & few coding guide lines
  - GPU code written in CUDA, runs on CPUs, Nvidia GPUs (CUDA), AMD GPUs (HIP)
- Algorithm sequences defined in python and generated at run-time
- Multi-event processing with dedicated scheduler
- Memory manager allocates large chunk of GPU memory at start-up
- Reconstruction algorithms re-designed for parallelism and low memory usage: O(MB) per core

LHCb: Software-only real-time analysis since 2022

- Two challenges:
  1) Connect sub-detectors to server-farm → FPGA card
  2) Use best suited computing architecture for reconstruction of particle collisions at 30 MHz

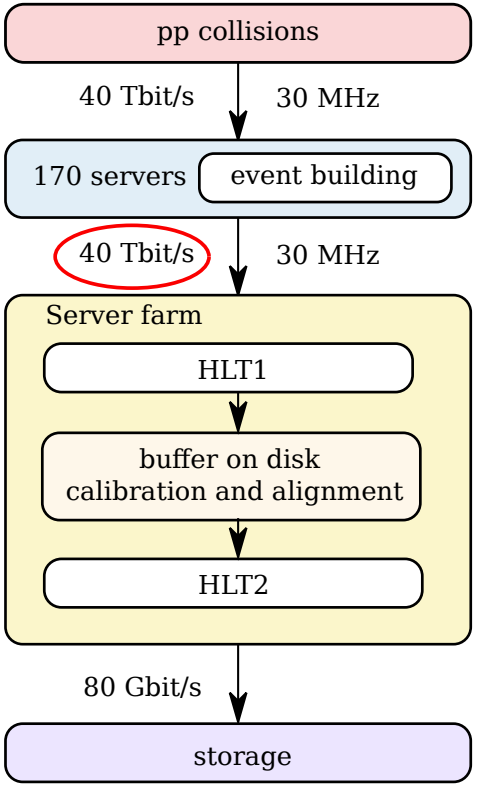
     → Partial reconstruction fully implemented on GPUs



History: HLT1 architecture choice

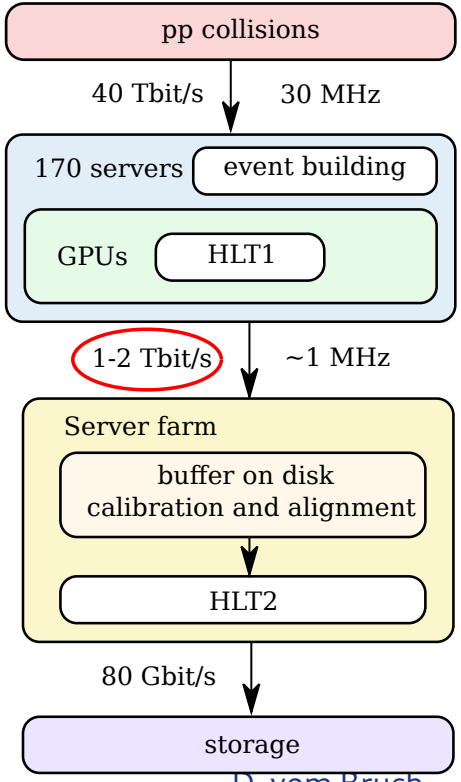
Proposal in TDR (2014)

CERN-LHCC-2014-016

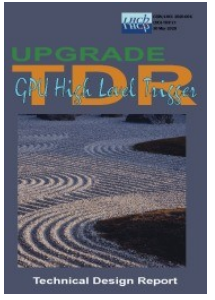


Updated strategy (as of 5/2020)

CERN-LHCC-2020-006

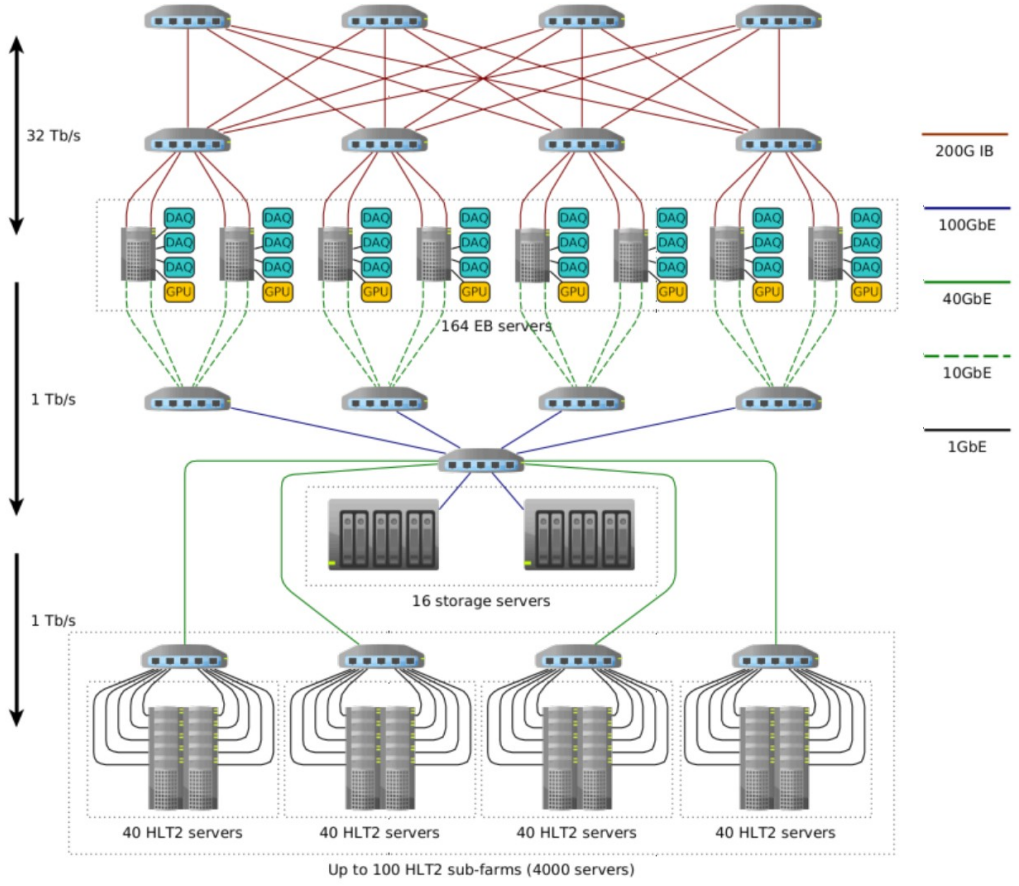


- Developed two solutions simultaneously
- Both the multi-threaded CPU & the GPU HLT1 fulfilled the requirements from the 2014 TDR
- Detailed cost benefit analysis (arXiv:2105.04031)
- GPU solution leads to cost savings on processors and the network
- Throughput headroom for additional features
- Decision: A GPU-based software trigger will allow LHCb to expand its physics reach in Run 3 and beyond.

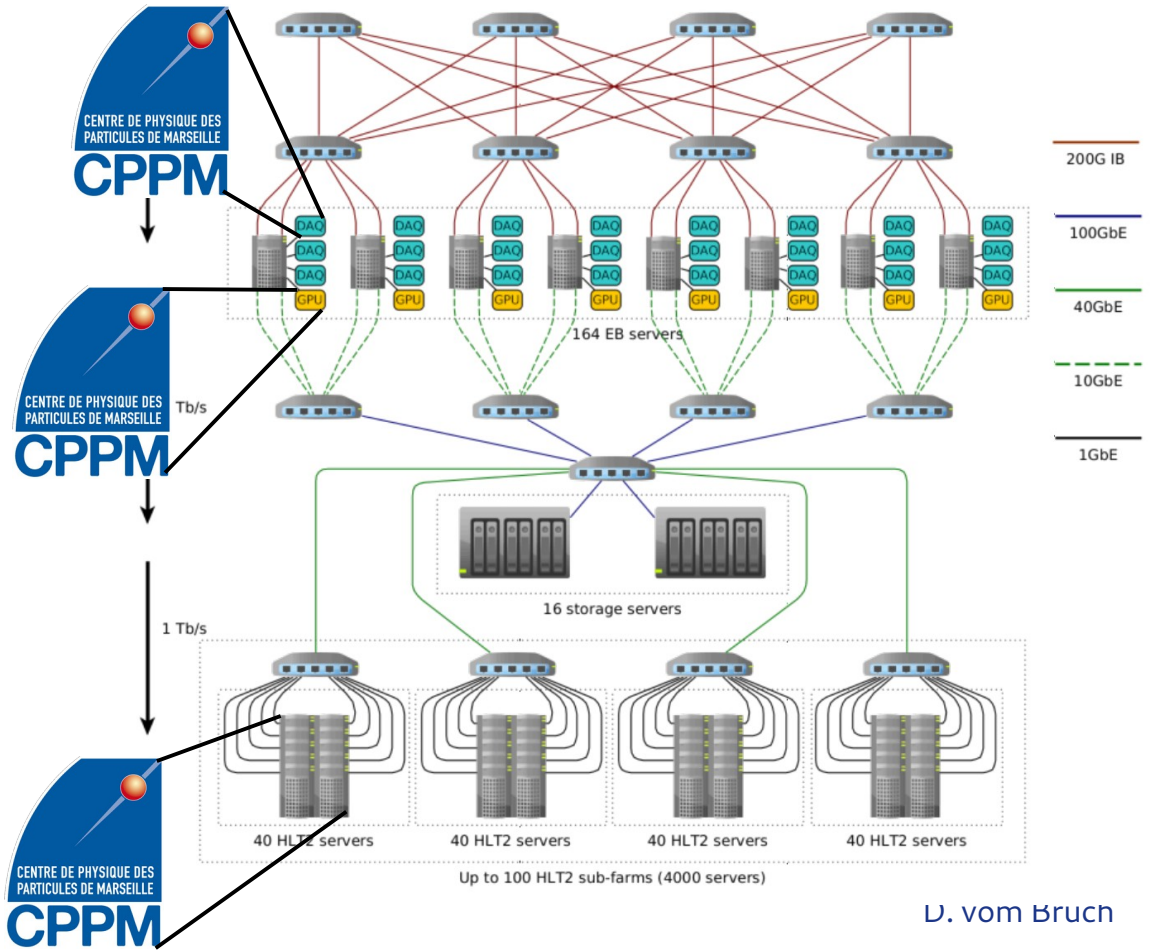


See also arXiv:2106.07701 on LHCb's energy efficiency with a CPU and GPU HLT1

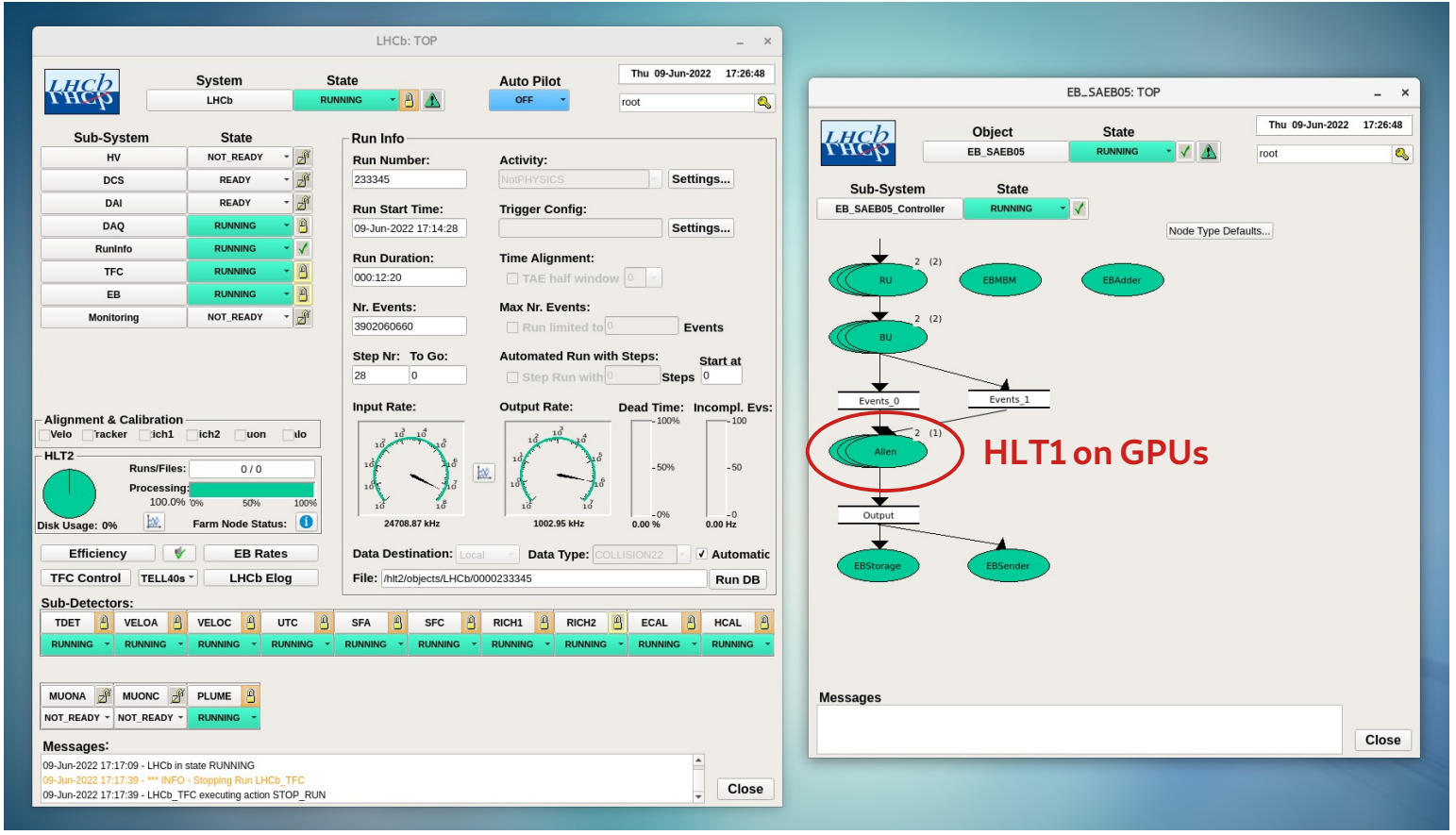
GPU HLT1 within data acquisition system



GPU HLT1 within data acquisition system

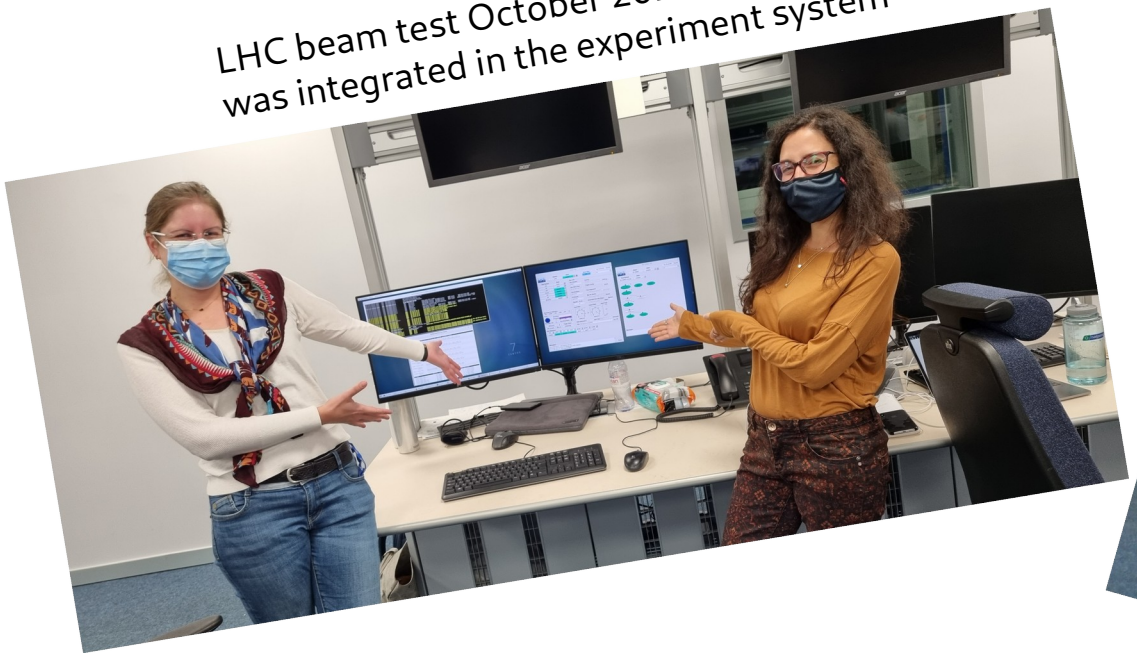


HLT1 commissioning: Allen within the DAQ system



HLT1 commissioning: Towards first collisions

LHC beam test October 2021: First time Allen was integrated in the experiment system



May 2022: First time Allen ran at 25 MHz input rate

HLT1 commissioning: Towards first collisions

July 2022: First collisions @ 13.6 TeV at the LHC
Happy trigger commissioning team

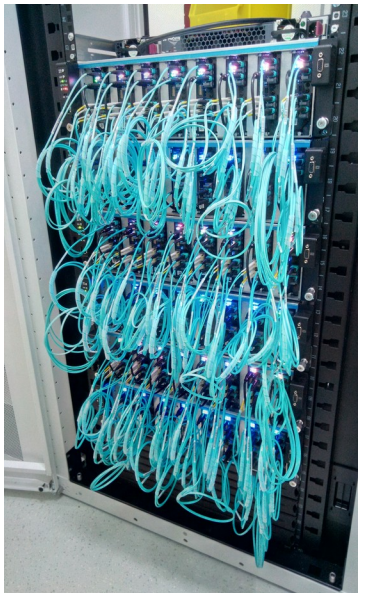
LHC beam
Allen was i

Allen ran at 25 MHz input rate

Summary

- Particle physics experiments real time analysis systems are entering the exascale computing era
- Need to exploit modern computing techonolgies to face this challenge
- LHCb experiment is commissioning a real-time analysis system full implemented in software in 2022
- First time in particle phyiscs to process 30 million proton-proton collisions per second on GPUs
- Developed Allen: a heterogeneous software framework for multi-event processing
- Gain expertise in heterogeneous DAQ systems

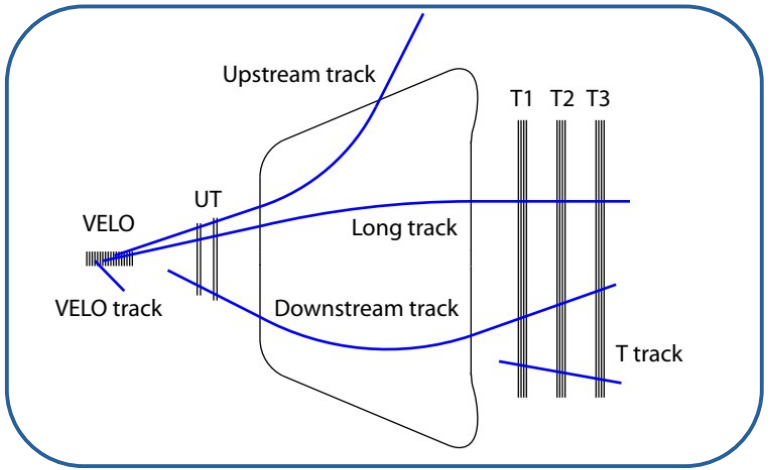
  → Preparing to exploit emerging new architectures entering the market



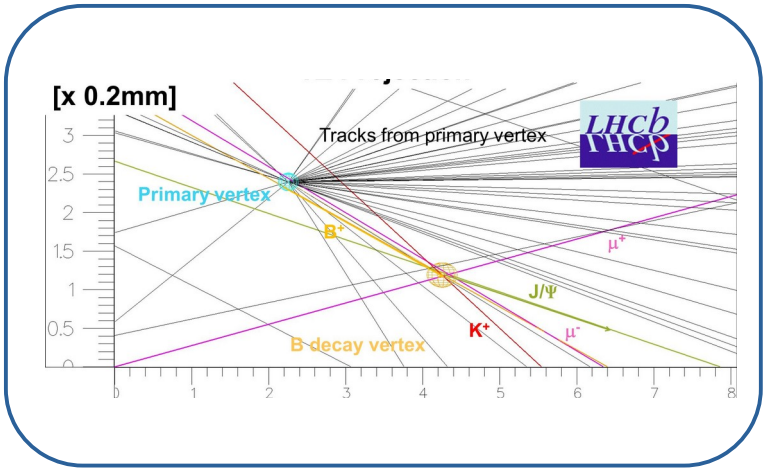
Backup

What do we reconstruct at LHCb?

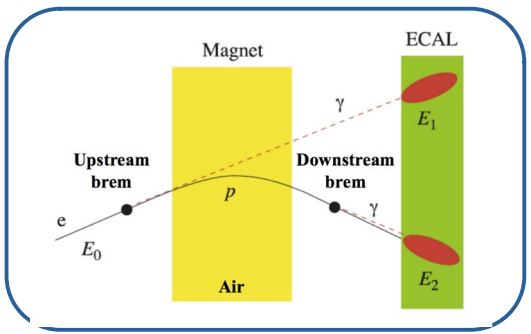
Tracks



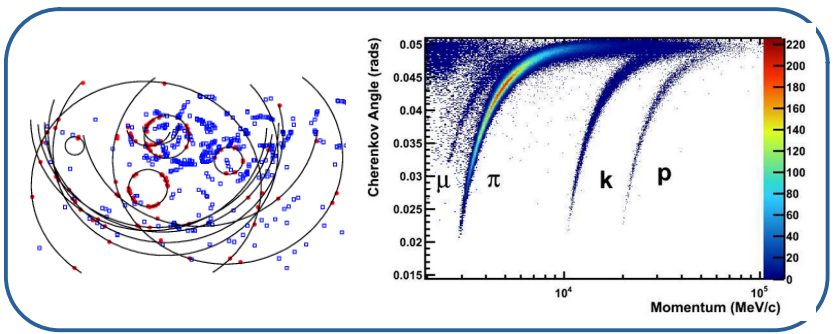
Vertices



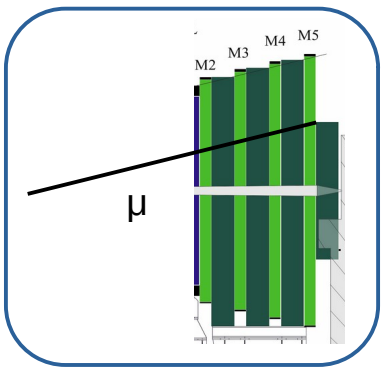
Electrons



Cherenkov rings



Muons

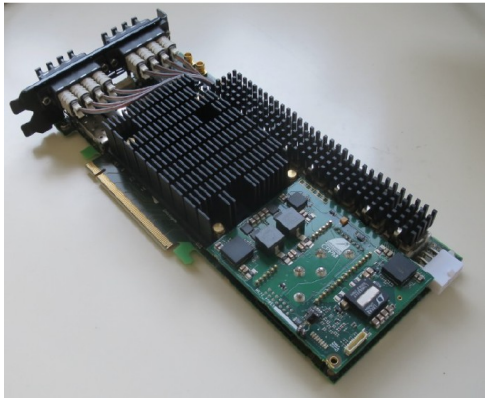


LHCb: Readout board PCIe40/400

Run 3: 40 Tbit/s → PCIe40 card developed

- Receives data from sub-detectors and transfers it to the server memory for event building via PCIe connection
- Local data processing occurs on the card using only the information from the links connected to it
- Card is generic enough to be re-used by other experiments: ALICE, Belle-II, Mu3e

PCIe40 card



- **Towards Run 5: increase bandwidth and processing power by factor 10**
- Run 4: PCIe400 card to transfer 400 Gbit/s via PCIe connection
- Run 5: Transfer 800 Gbit/s via ethernet connection using more powerful FPGA
- Add more local processing to the board in the future to reduce processing load of HLT

Overview of GPU usage in various HEP experiments

Experiment	Main tasks processed on GPU	Event / data rate	Number of GPUs	Deployment date
Mu3e	Track- & vertex reconstruction	20 MHz / 32 Gbit/s	O(10)	2023
CMS	Decoding, clustering, pattern recognition in pixel detector	100 kHz		2022 (tbc)
ALICE	Track reconstruction in three sub-detectors	50 kHz Pb-Pb or < 5 MHz p-p / 30 Tbit/s	O(2000)	2022
LHCb	Decoding, clustering, track reconstruction in three sub-detectors, vertex reconstruction, muon ID, selections	30 MHz/ 40 Tbit/s	O(250)	2022

https://arxiv.org/pdf/2003.11491.pdf

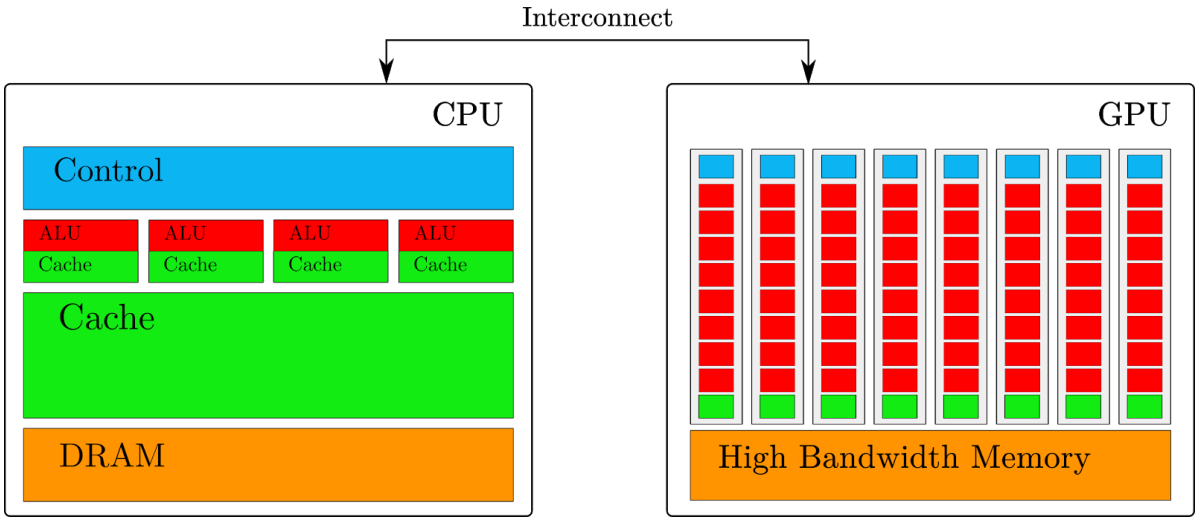
CPU – GPU - FPGA

	Latency	Connection	Engineering cost	FP performance	Serial / parallel	Memory	Backward compatibility
CPU	O(10) µs	Ethernet, USB, PCIe	Low entry level: Programmable with C++, pthon, etc.	O(1-10) TFLOPs	Optimized for serial, increasingly vector processing	O(100) GB RAM	Compatible, except for vector instruction sets
GPU	O(100) µs	PCIe, Nvlink	Low to medium entry level: Programmable with CUDA, OpenCL, etc.	O(10) TFLOPs	Optimized for parallel performance	O(10) GB	Compatible, exept for specific features
FPGA	Fixed O(100) ns	Any connection via PCB	High entry level: traditionally hardware description languages, Some high-level syntax available	Optimized for fixed point performance	Optimized for parallel performance	O(10) MB on the FPGA itself	Not easily backward compatible

https://arxiv.org/pdf/2003.11491.pdf

GPUs

- Developed for graphics pipeline
- General purpose computations possible
- Increasingly used for AI applications
- Hardware specialized in this direction since few years
- Programmed with high-level language



Low core count / powerful ALU

Complex control unit

Large chaches

→ **Latency optimized**

High core count

No complex control unit

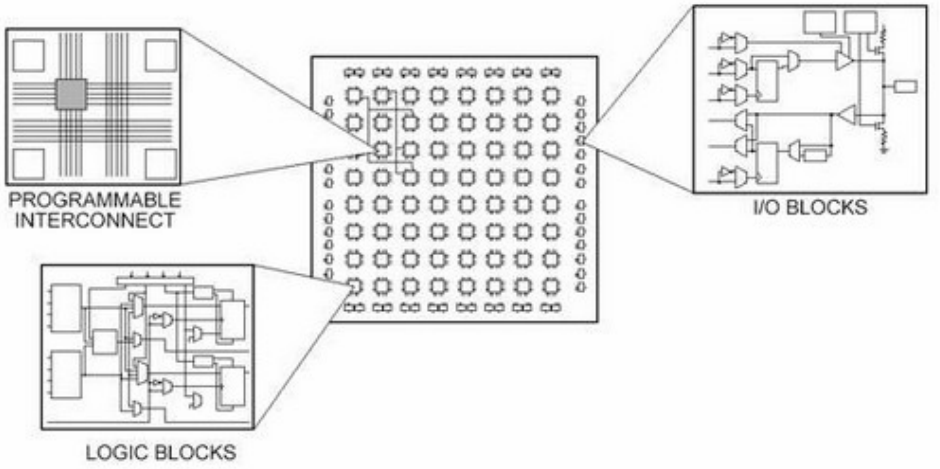
Small chaches

→ **Throughput optimized**

FPGAs – High Level Synthesis for Neural Networks

- Traditionally, programmed with hardware description languages (Verilog, VHDL) → long development time
- Increasingly more high-level languages (HLS) developed
- Challenges:
  - Fit into resource constraints of FPGA
  - Preserve model performance
- Specialized hardware blocks emerging implementing functions for Neural networks such as tensor blocks

FPGA: thousands of logic blocks, I/O blocks, connected via programmable interconnect



Source: National Instruments