Innovative developments for real-time data processing in particle physics within IN2P3

Dorothea vom Bruch
CPPM Marseille

On behalf of the ATLAS-IN2P3, CMS France, LHCb France communities
and the OWEN and THINK projects

Conseil scientifique calcul et données
June 23rd 2022
Paris, France

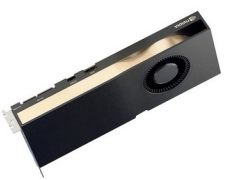
Outline

- Requirements and constraints for real-time data analysis

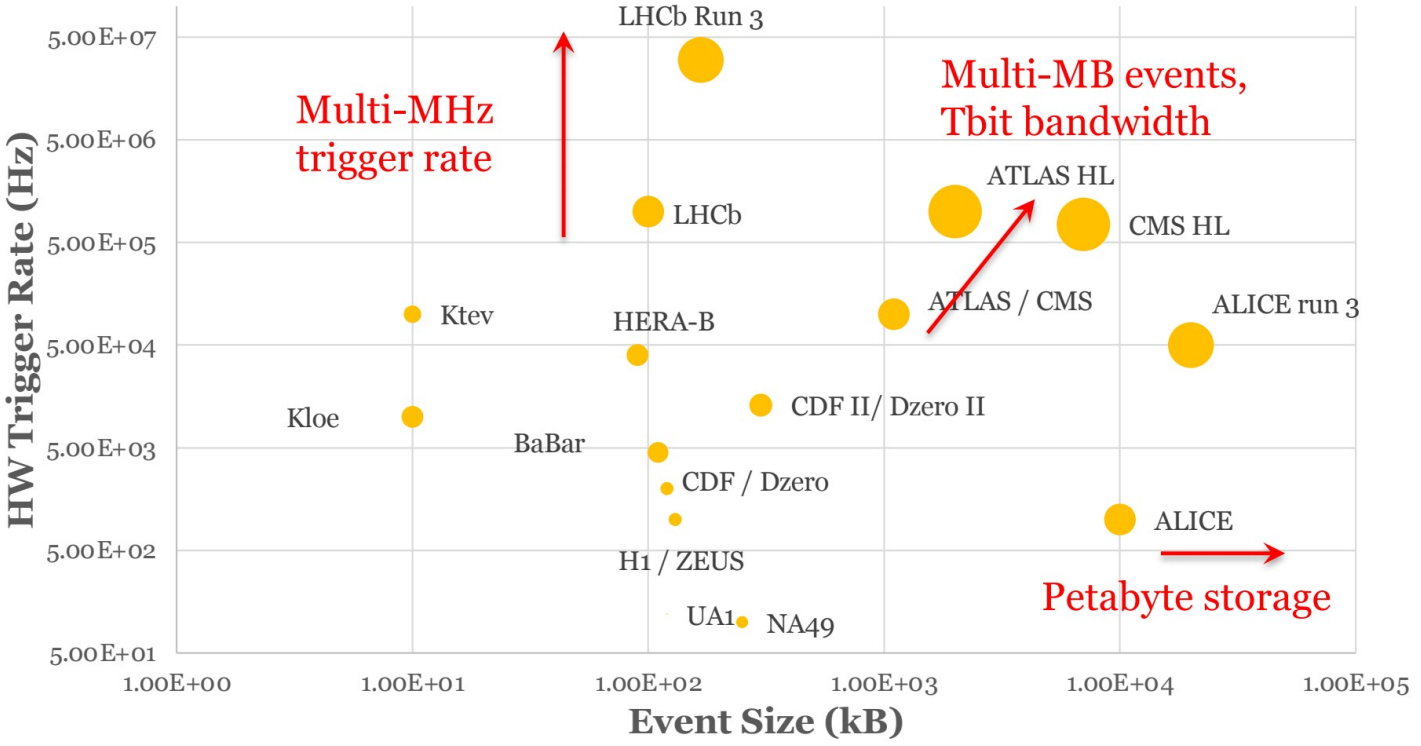
- Efficient hardware trigger selection: neural network implementations on FPGAs
  - ATLAS, CMS, OWEN, THINK

- Efficient software trigger selection: heterogeneous systems with GPUs
  - LHCb

- Summary



Real-time data challenges



Courtesy Alex Cerri, LHCP 2022

LHC Run 3 (2022)

LHCb: pp collisions at 30 MHz,
→ 5 TB/s → processed in software

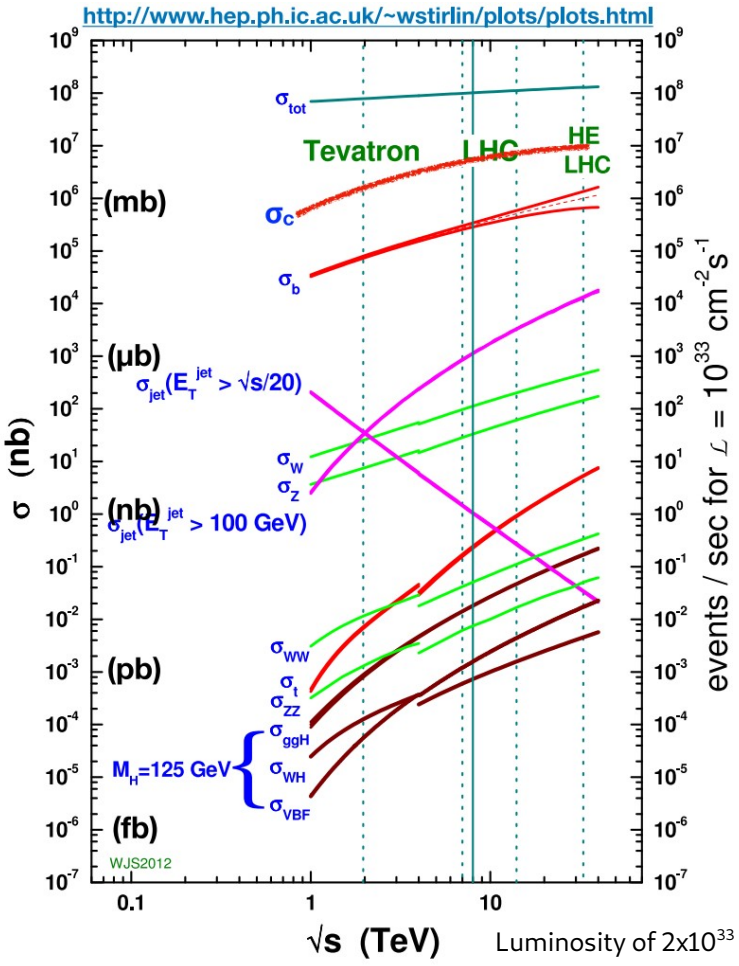
LHC Run 4 (~2029)

CMS & ATLAS
pp collisions at 40 MHz,
Hardware trigger rate increased:
100 kHz → 1 MHz
→ 6 GB/s processed in software

LHC Run 5 (~2035)

LHCb undergoes Upgrade II
25 TB/s processed in software

Efficient signal selection

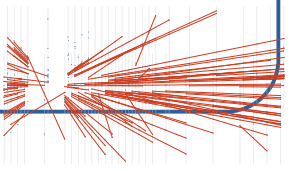


Luminosity of 2×10^{33} cm^{-2}s^{-1}

LHCb: Mainly beauty and charm

- Signal rates at MHz level
- No local criteria for selection → Efficient hardware trigger not possible
- Readout full detector

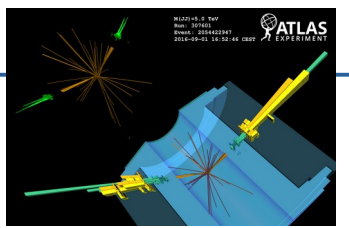
  → **Real-time analysis and selection fully in software**



ATLAS & CMS: Mainly Higgs properties, high p_T new phenomena

- Local criteria for selection → Efficient hardware trigger possible
- Hardware trigger necessary (cannot read out full data stream)

  → **First reduction in hardware to manageable level, second reduction in software**



Which co-processor is best for which workload?

Software level triger

- High-bandwidth processing power
- No strict latency requirement
- Data obtained from server

Hardware level triger

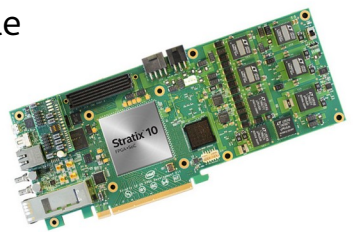
- Fixed, low-latency
- Data obtained from detector

Graphics Processing Units (GPUs)

- Higher latency
- Connection via PCIe → bandwidth limited
- Very good floating point performance
- Low engineering cost
- Backward / forward compatibility

Field Programmable Gate Arrays (FPGAs)

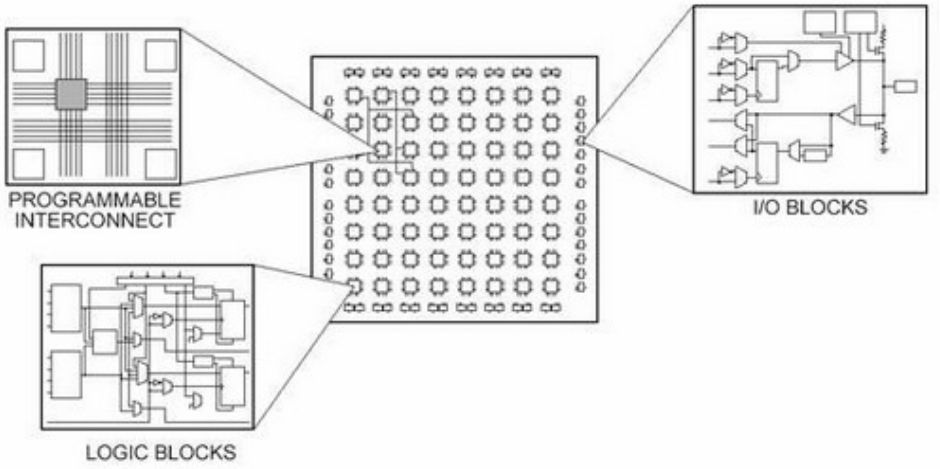
- Low & deterministic latency
- Connectivity to any data source → high bandwidth
- Intermediate floating point performance
- High engineering cost
- Not so easily backward compatible



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

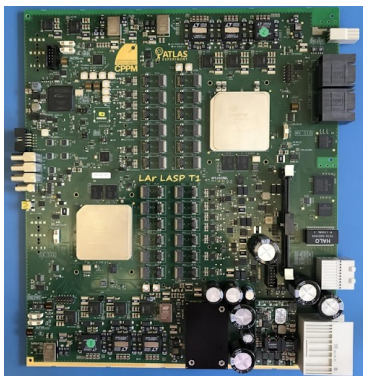


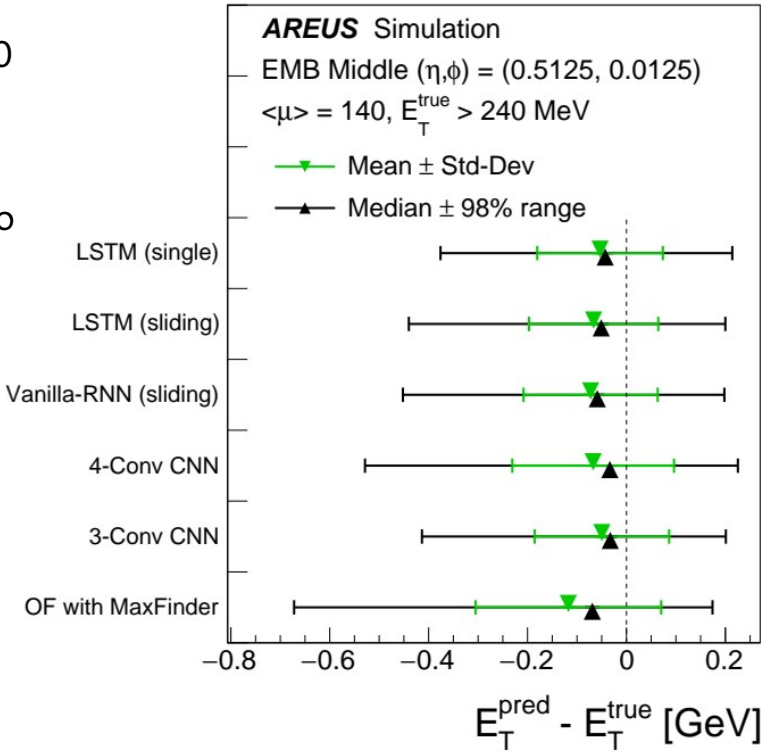
Source: National Instruments

ATLAS: LAr calorimeter energy calculation

- Exchange of full readout electronics of LAr calorimeter for Run 4

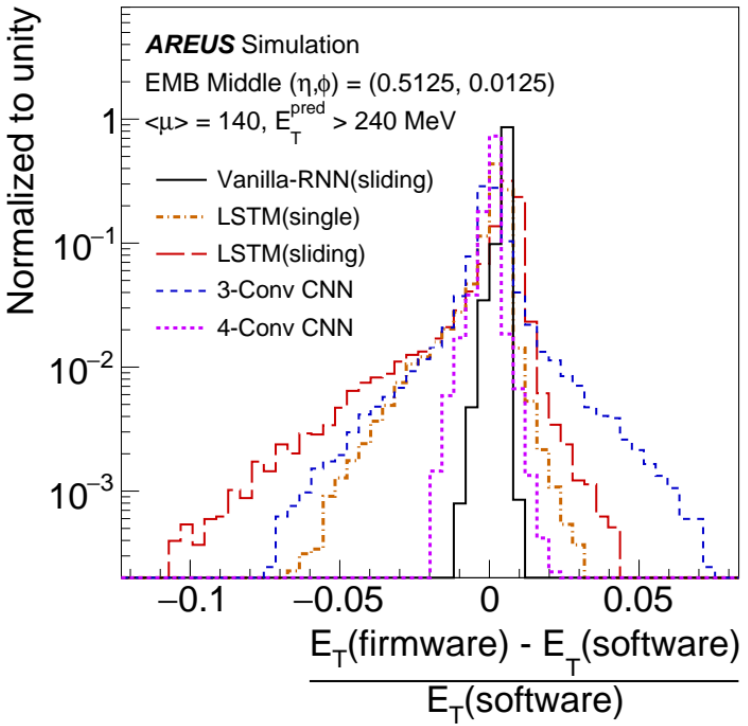
  → Cope with higher pileup and level 1 trigger rate increase of factor 10
- Off-detector board (LASP) responsible for computation of energy, designed by CPPM
- Challenge: Recover filter performance for increased pileup and fit into O(100 ns) timing constraint
- Recursive Neural Networks (RNN) used for energy reconstruction
- Challenge: fit into FPGA resources and implement in HLS language





Comp. And Soft. For Big Science 5, 19 (2021)

ATLAS: RNNs on FPGAs



Comp. And Soft. For Big Science 5, 19 (2021)

- Demonstrated for the first time that
  - Energy reconstruction of RNNs is more performant than filtering algorithms
  - RNNs can fit within resource usage and latency requirements of FPGA
- Demonstrator using Stratix 10 works successfully
- RNN implementation on Intel FPGAs added to HLS4ML toolkit
- Currently producing first prototype with next generation FPGAs: AGILEX from Intel

ATLAS: Resources & Visibility

- Close collaboration between CPPM and Dresden University to use Neural Networks for energy calculation
- LAPP designed and produced main off-detector board for Run 3 upgrade
- CPPM & LAPP coordinate ATLAS group responsible of firmware design of the board

- CPPM team visible within LAr collaboration, especially after quick proof-of-concept of RNN usage
- Joined HLS4ML developers, implementations & optimizations very well received

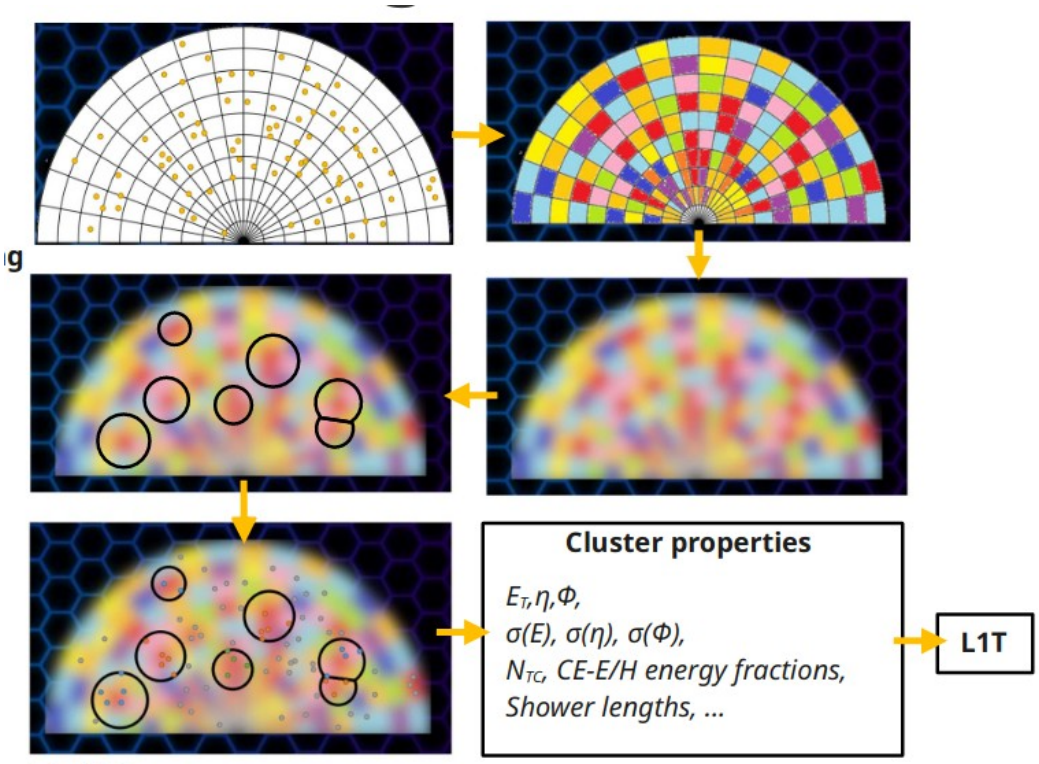
- Project-funded:
  - AIDAQ project from AMIDEX since 2019 in collaboration with Dresden
  - ANR JCJC since 2021
- Lack of permanent resources can lead to loss of engineering knowledge

Physicists	G. Aad, E. Monnier
Postdoctoral researchers	T. Calvet (2020-2021), N. Sur (2021-2024)
Electronic engineers	R. Faure (2022)
Doctoral students	N. Chiedde (2020-2023), E. Fortin (2020-2022), L. Laatu (2020-2023)

People at CPPM involved in NN developments and corresponding firmware (not LASP board)

CMS: HGCAL reconstruction

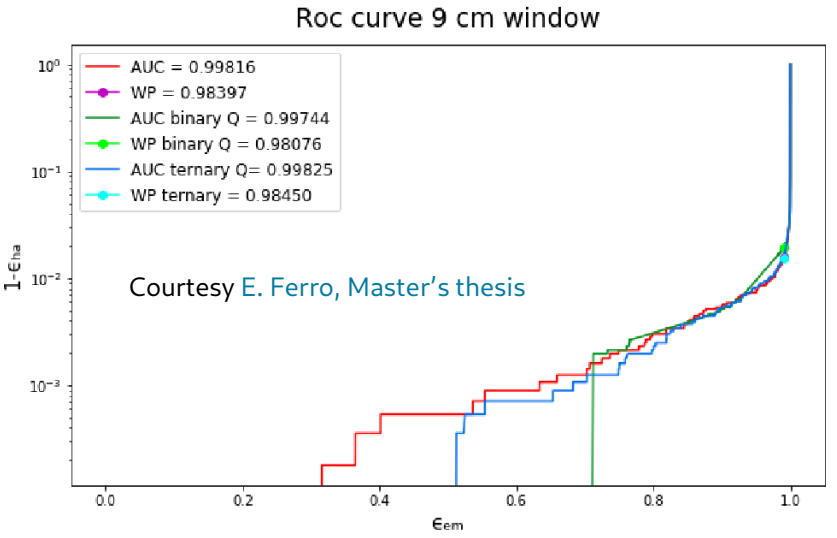
- High Granularity Calorimeter (HGCAL) for Run 4
- 5D detector: position, energy, timing
- 1M channels used for L1 trigger
- Latency constraints of a few micro seconds
- Reconstruct 3D clusters of energy as input for L1
- Used in central L1 system to build electrons, photons, hadronic taus and jets

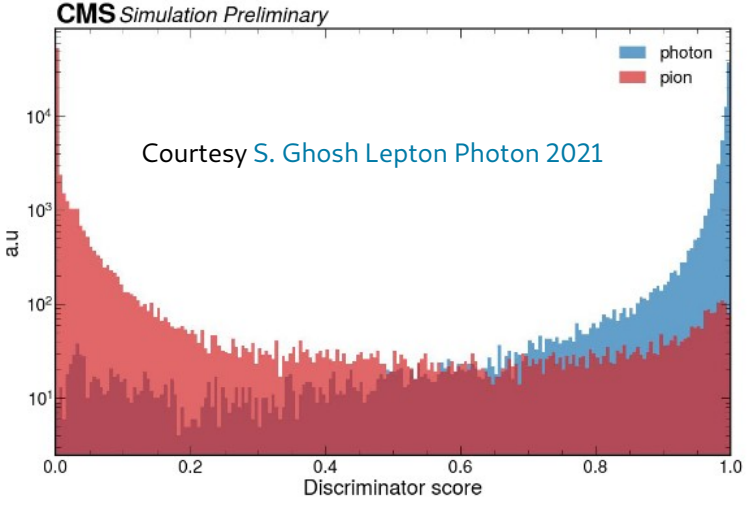


Courtesy L. Portales CALOR 2022

CMS: Energy reconstruction with machine learning

- Run 1: Boosted Decision Trees (BDTs) in lookup tables
- Limitation: Number of block RAM on FPGA
- Study implementation of actual models such as BDTs and NNs
  - BDTs & fully connected networks can be implemented with Conifer
  - Fully connected NNs and CNNs can be implemented with HLS4ML





- Looking into Graph Neural Networks (GNNS) due to irregular geometry of HGCAL
- Mostly applied in offline reconstruction so far

  → Move to real-time in L1 trigger next
- Tested as isolated components

  → Move to real hardware system, connect with rest of firmware

CMS: Resources & Visibility

- LLR has been protagonist in CMS L1 trigger development
  - Project manager of L1 trigger system: A. Zabi (LLR)
  - Coordinator of algorithm developments for HGCAL trigger primitives generation: J.-B. Sauvan
- Long-standing involvement and strong collaboration with University of Split (Croatia), Imperial College (London, UK)

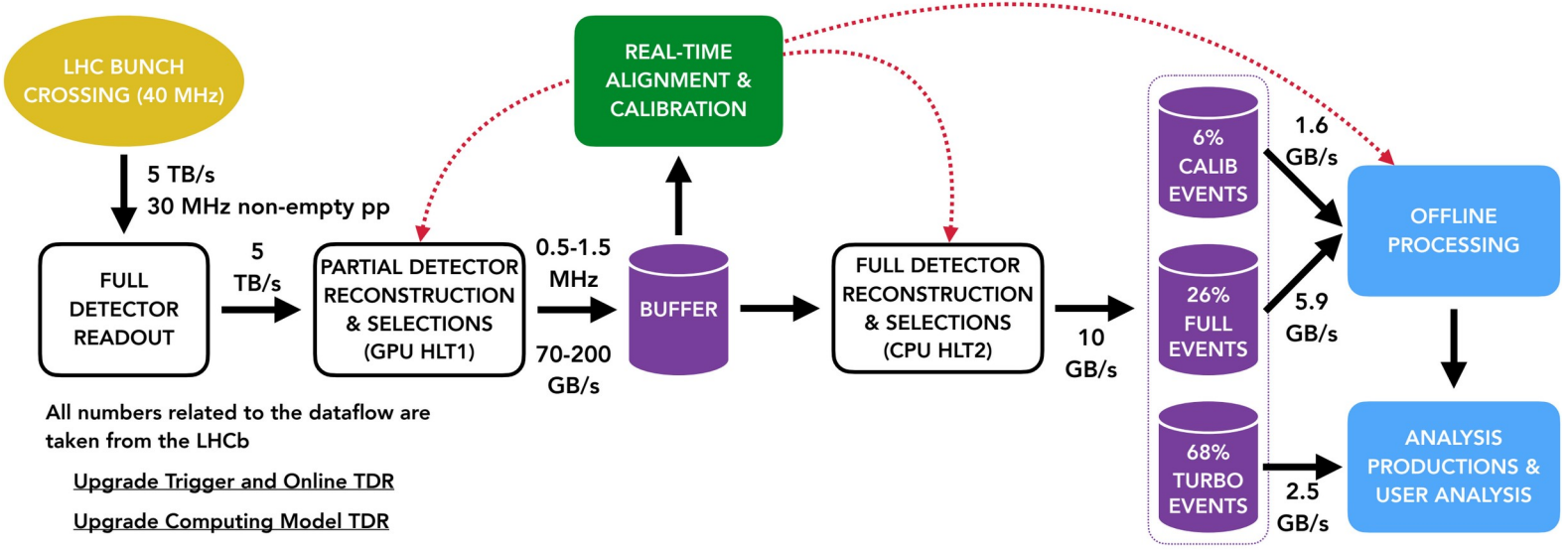
- GPU (for training) & FPGA platforms funded by P2IO Labex, ANR, TGI HL-LHC
- Project funded:
  - ANR HiGranTS since 2018 (PI: J.-B. Sauvan)
  - ANR OGCID since 2021 (PI: F. Magniette)
  - Joint international project with Imperial Collegee

LLR	Physicists	F. Beaudette, J.-B. Sauvan, A. Zabi
	Research engineers	E. Becheva, F. Magniette
	Doctoral students	A. Hakimi, J. Motta

LHCb: Full detector readout in Run 3

- Luminosity increase of factor 5 in Run 3 → hardware trigger no longer efficient due to signal saturation
- Two challenges:
  1) Connect sub-detectors to server-farm → FPGA card (PCIe40 card developed by CPPM)
  2) Use best suited computing architecture for reconstruction of particle collisions at 30 MHz

     → Partial reconstruction fully implemented on GPUs (Allen project co-led by LPNHE & CPPM)



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

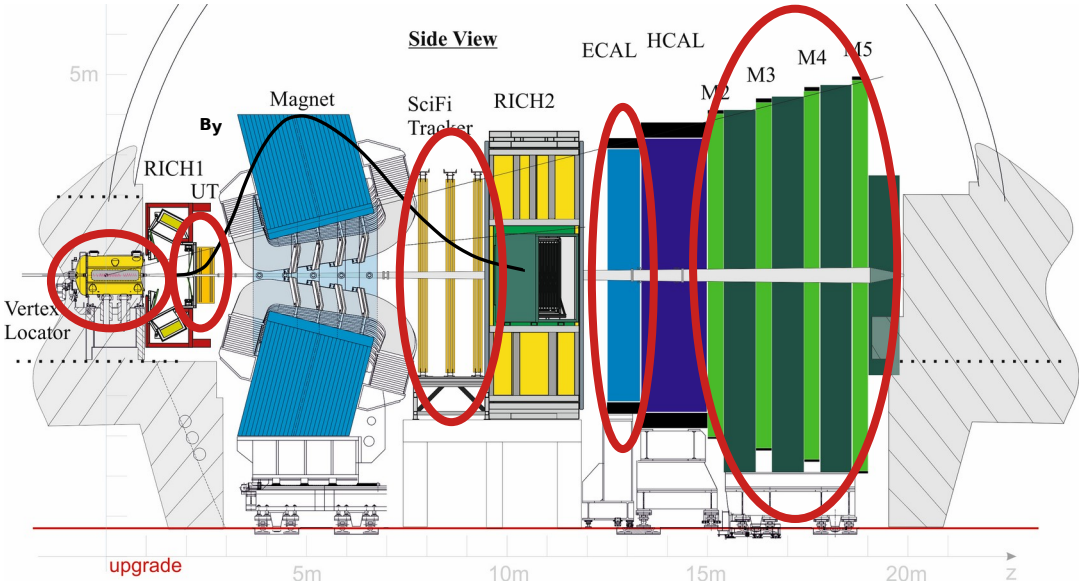
LHCb PCIe40/400 Resources & Visibility

- PCIe40 developed at CPPM
  - R. Le Gac: Scientific project leader
  - J.-P. Cachemiche: Technological project leader
- PCIe400 card developed within R&T Project PCIe400 by CPPM, CENBG, IJCLab, LAPP, LPC Caen
- Interest of various laboratories in innovative technology
- Might be used by upgraded ALICE and Belle-II
- CPPM team is natural candidate to lead next generation board project, but suffers from retirement of key engineers

CPPM	Physicists	R. Le Gac
	Research engineers	K. Arnaud, P. Bibron, J.-P. Cachemiche, J. Langouet
LAPP	Physicists	S. T'Jampens
	Research engineers	G. Vouters

LHCb: High Level Trigger 1 on GPUs

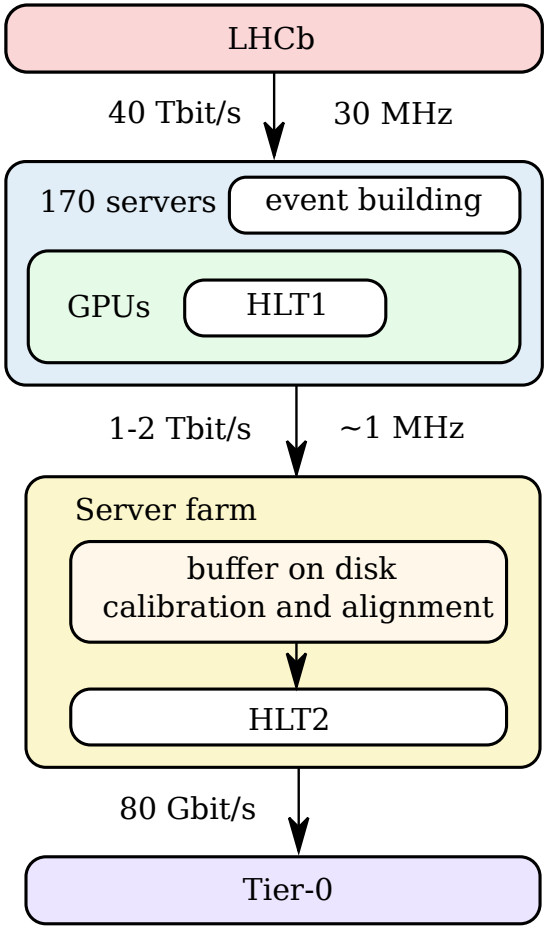
- Decode binary payload of five sub-detectors
- Reconstruct charged particle trajectories
- Identify muons and electrons
- Reconstruct primary and secondary decay vertices
- Select pp-bunch collisions based on
  - Single-track properties
  - Secondary vertex properties



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

Raw data → **GPU** → Selected events

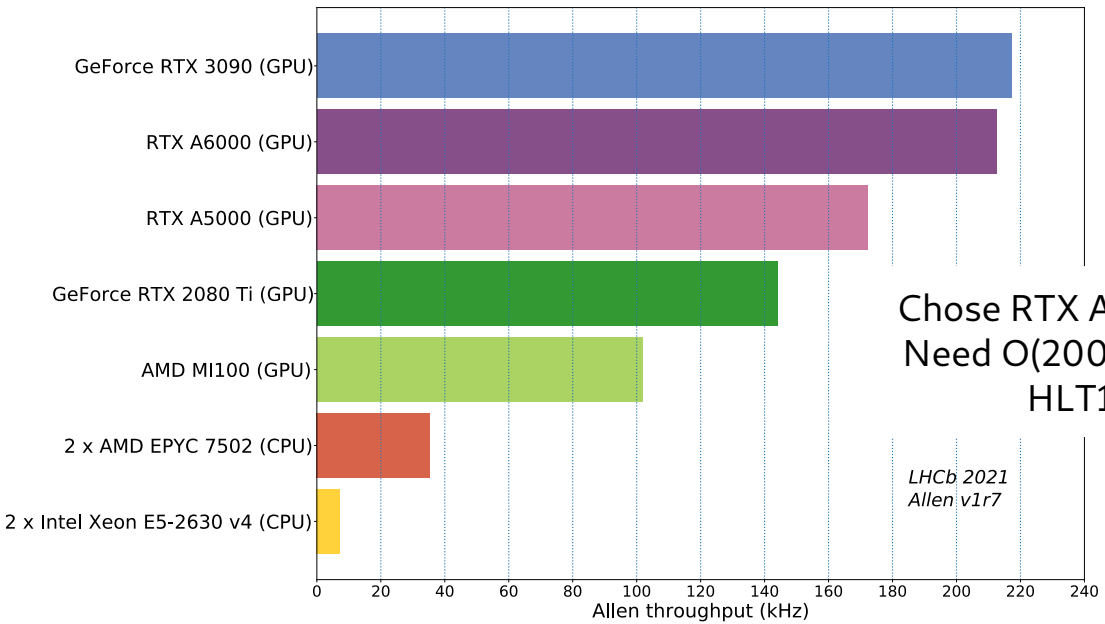
LHCb: High Level Trigger 1 on GPUs ("Allen")



CERN-LHCC-2020-006

Cost saving

- Originally planned CPU implementation of HLT1
- GPU trigger saves O(1M) Euros compared to CPU option
- Saving on network between the server farms and processor cost
- Comparison of CPU & GPU option: arXiv:2105.04031



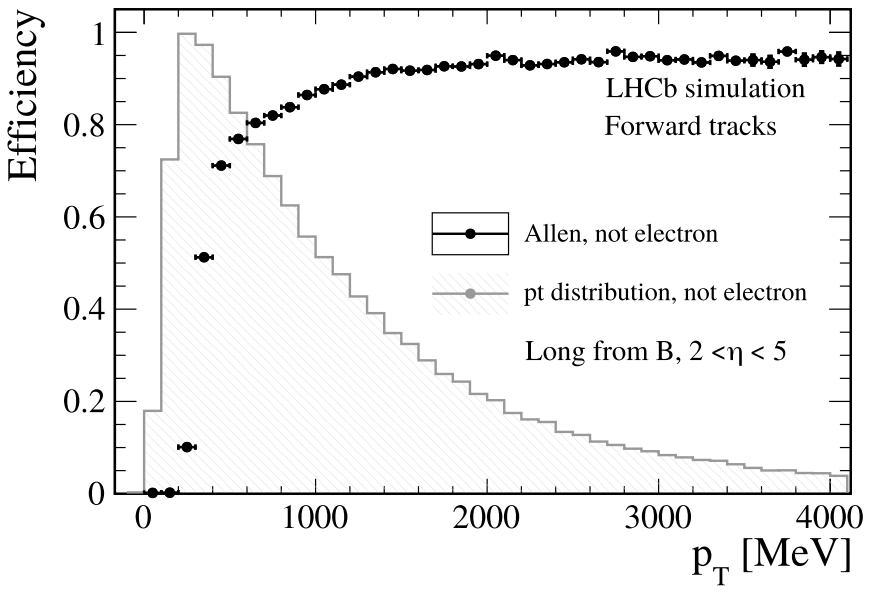
Chose RTX A5000 for Run 2022
Need O(200) GPUs to procoess
HLT1 @ 30 MHz

D. vom Bruch

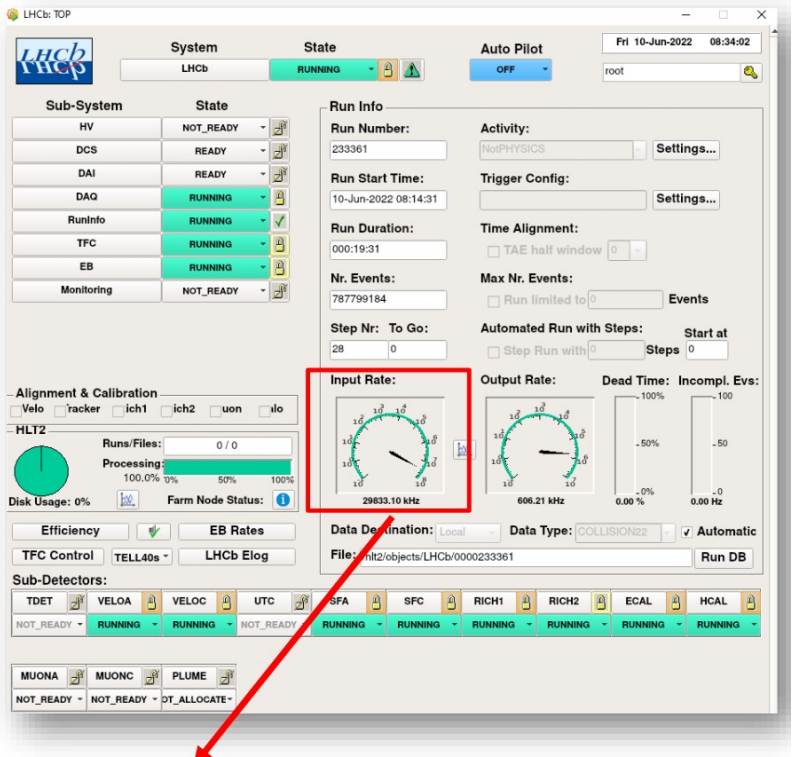
Update of LHCb-FIGURE-2020-014

LHCb: HLT1 on GPUs performance

Track reconstruction efficiency on simulated data



Ran GPU trigger at 30 MHz on real data for the first time last week



*SCIFI and VELO using internal generators for this

LHCb-FIGURE-2020-014

LHCb: Real-Time Analysis Resources & Visibility

- RTA in LHCb has received wide recognition in the community both for Run 2 and Run 3 RTA system
- Project leader of LHCb RTA: V. Gligorov (LPNHE)
- Co-leading Allen (HLT1 on GPUs): V.V. Gligorov (LPNHE), D. vom Bruch (CPPM)
- Accomplished by combining competences of several laboratories in terms of reconstruction software, trigger system development, DAQ integration: CPPM, IJCLab, LAPP, LPNHE
- Funded largely by
  - Two ERC grants: RECEPT (PI: V.V. Gligorov), ALPaCA (from 2022, PI: D. vom Bruch)
  - Two ANR grants: BACH (PI: Y. Amhis), ANN4Europe (PI: V. Gligorov)
- Allen provides a natively cross-architecture framwork
  - Long-term maintenance and continuous support from engineers and applied physicists crucial to support wide range of use-cases

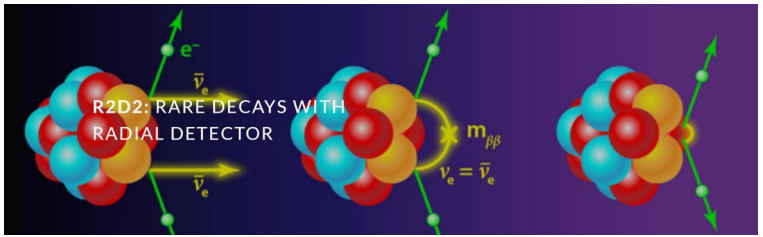
LPNHE	Physicists	V. Gligorov, F. Polci
	Postdoctoral researchers	M. Fontana, C. Agapopulou
	Software engineers	A. Bailly-Reyre, N. Garroum
	Doctoral students	A. Scarabotto, T. Fulghesu
IJCLab	Physicists	Y. Amhis, F. Machefert, P. Robbe
	Doctoral students	F. Volle, V. Vayeroshenko
CPPM	Physicists	A. Poluektov, R. Le Gac, D. vom Bruch
	Doctoral students	V. Dedu

OWEN (Optimized Waveform for Electronic Nodes)

- Proposed new experiment to search for neutrinoless double beta decay
- High pressure gaseous Time Projection Chamber (TPC)
- For the first time process raw signal just after charge amplifier

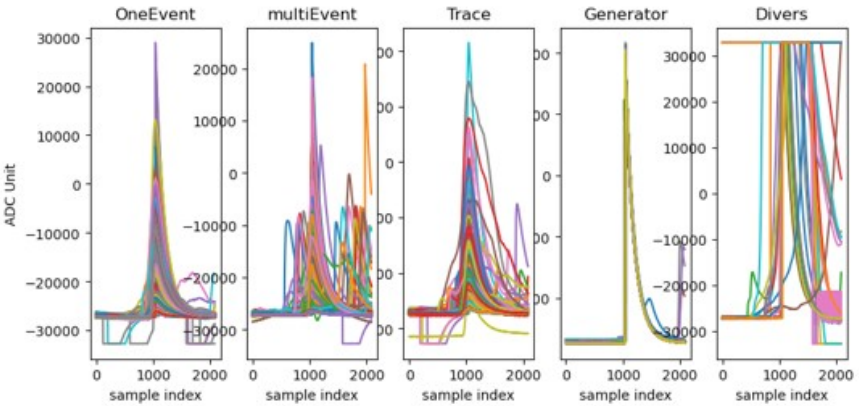
  → classify signal versus background with neural network
- OWEN project:
  - Develop versatile charge amplifier for low capacitance detector
  - Use AI in embedded system for real-time signal selection
  - Change control & command system based on user's experience

Requirements for real-time analysis
- Fixed, low-latency
- Data obtained from detector



OWEN: Developments, Resources, Visibility

- Determine signal waveform parameters and energy, perform particle identification in real-time
- Use machine learning approaches for this inverse problem
- Curent status:
  - Low noise charge amplifier developed
  - Embeeded data acquisition system built tagging events to test AI algorithms
- Future plans:
  - Single DAQ board with more powerful FPGA performing all digital data processing steps
  - Open source framework for embedded neural network models
- Financed by IdEx Programme Emergence (2019-2022)

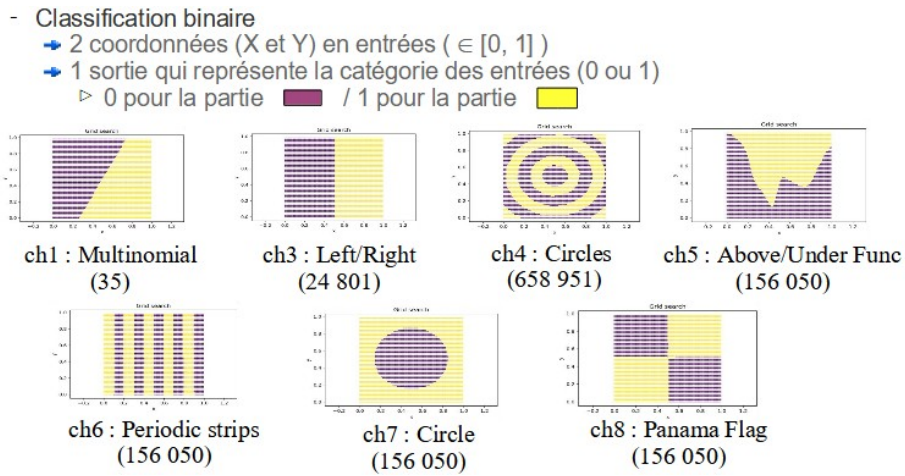


Courtesy F. Druillole

LP2I Bordeaux	Physicists	A. Meregaglia, C. Jollet, F. Piquemal
	Research engineers	F. Duillole, P. Hellmuth, A. Rebii
	Electrical engineers	R. Bouet
	Doctoral students	P. Charpentier
Subatech	Doctoral students	V. Cecchini

THINK: Testbed of various processor types

- THINK project combines use cases and inputs from different experiments
- Project length: 2020 – 2023, to be extended
- Provide tools to select best suited hardware for given problem without conducting costly studies
- First step: Compare instantiation of Neural Networks on several hardware architectures:
  FPGAs, GPUs, Tensor Processing Unit (TPU) engines, neuromorphic chips, embedded tensor blocks
- Metrics: Computing performance, cost, manufacturer information, learning curve, speed of implementation
- Synthesis of comparison will soon arrive on the THINK website
- Labs involved:
  LPC Caen, LAPP, LPNHE, CENGB, IRFU/AIM, LLR, CPPM



Courtesy F. Magniette

Summary

- Strong historical involvement of French groups real-time data analysis
- Held various roles of responsibility and made major contributions to both hardware and software-level triggers
- Incremental changes and simple extrapolation of methods not sufficient for computing demands of future experiments
- Move towards heterogeneous computing solutions: In line with European Strategy for particle physics and Roadmap of the HEP Software Foundation
  - Full HLT on GPUs (LHCb)
  - Processing AI algorithms on FPGAs (ATLAS, CMS, OWEN)
- Two of the priorities of the "Calcul, algorithmes et données" working group of IN2P3 Prospectives:
  - Efficient usage of accelerators like GPUs and FPGAs
  - Lead in the domain of AI of IN2P3 related science domains

To continue the success story of computing and data science at IN2P3, we rely on the ability to attract and train experts in the field to build long-term teams of both physicists and engineers.

Thanks to…

- G. Aad (CPPM)
- J.-P. Cachemiche (CPPM)
- F. Druillole (LP2I Bordeaux)
- R. Le Gac (CPPM)
- V. V. Gligorov (LPNHE)
- J.-F. Marchand (LAPP)
- J.-B. Sauvan (LLR)

Backup

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"

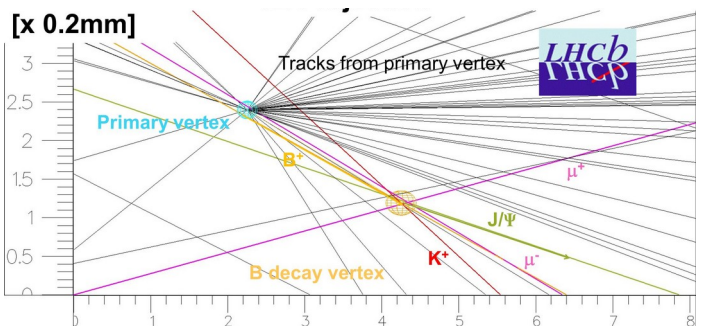
    → Huge combinatorics when testing different combinations of hits
  - Track fitting: Describe track with mathematical model

Vertex finding

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

Calorimeter / muon detector reconstruction

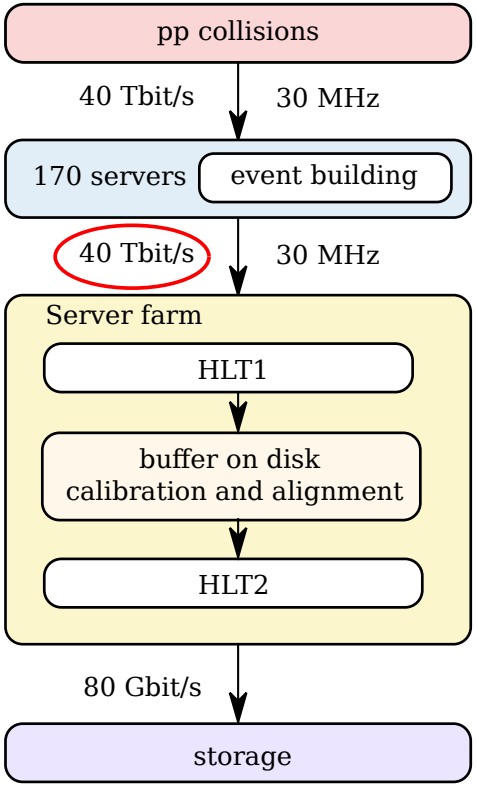
- Reconstruct clusters in the calorimeter / muon detectors
- Match tracks to clusters



History: HLT1 architecture choice

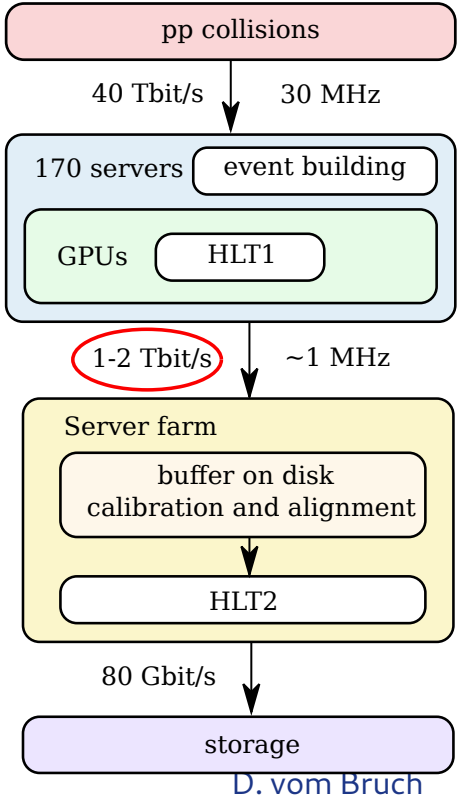
Proposal in TDR (2014)

CERN-LHCC-2014-016

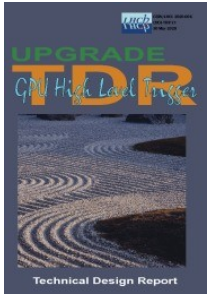


Updated strategy (as of 5/2020)

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

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

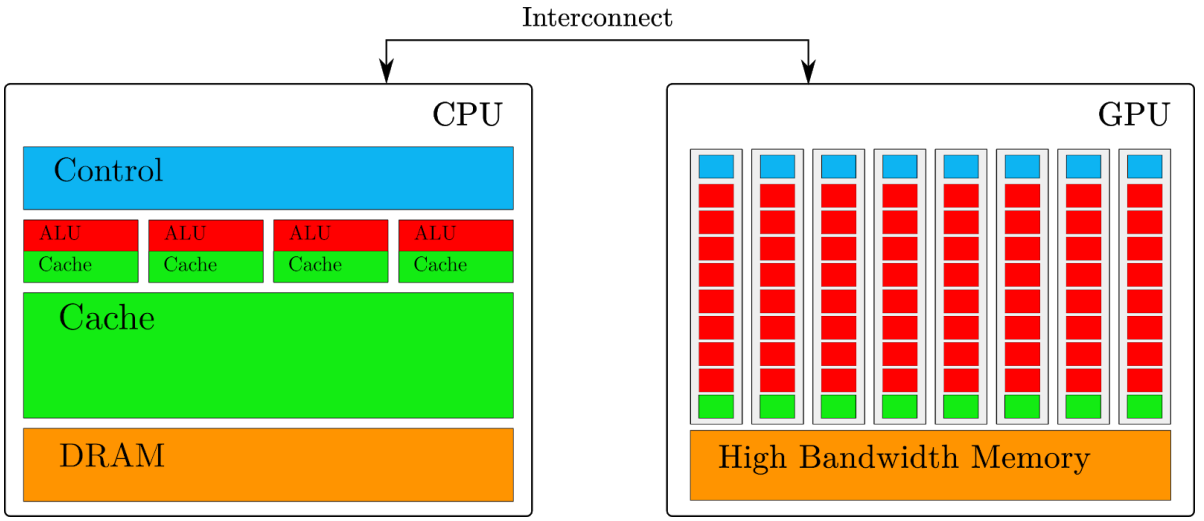
CPU – GPU - FPGA

	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**