Valencia High-Low activities

L. Fiorini (IFIC-Valencia)
on behalf of the High-Low Team

COMCHA Session

COMCHA
Computing challenges

19/11/2025
CPAN Workshop

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
AGENCIA ESTATAL DE INVESTIGACIÓN
CPAN Ingenio 2010

HL

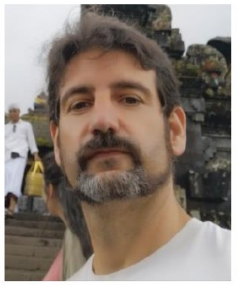
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
CSIC Consejo Superior de Investigaciones Científicas
Universitat de València
IFIC Institut de Física Corpuscular
Excelencia Severo Ochoa

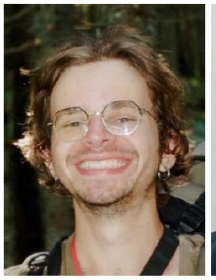
High-Low Team

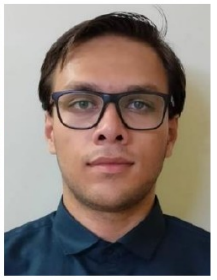
- Team is formed by LHCb and ATLAS members


Arantza Oyanguren


Jiahui Zhuo


Álvaro Fernández


Valerii Kholoimov

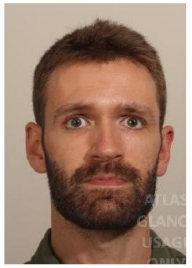

Karan Singh

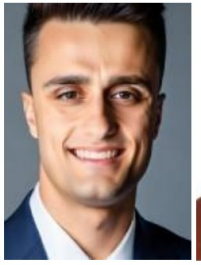

Miranda Carou


Miriam Lucio


Luca Fiorini


Alberto Valero

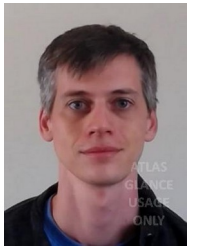

Antonio Cervello


Francisco Hervás


Héctor Gutierrez


Rui Wang


Francesco Curcio


Fernando Carrió


Arantxa Ruiz


Ximo Poveda

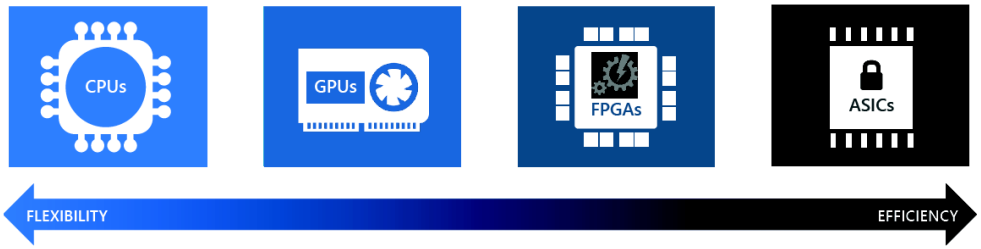
High-Low Infrastructure

- **HIGH-LOW: "High-Performance Algorithms for Low Power Sustainable Hardware for LHC experiments and their upgrades"**

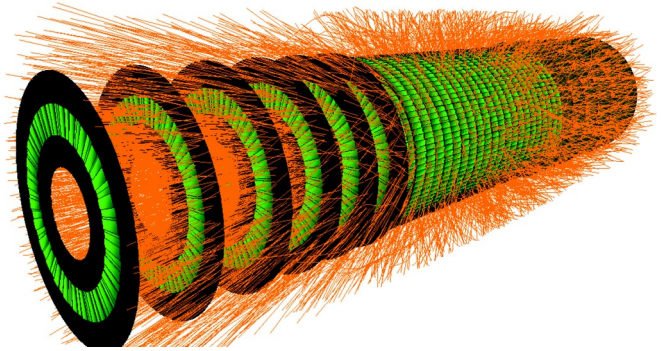
- NG8 Dual Epyc 9004-8B
  - 2 x NVIDIA H100 NVL 94GB
  - 1 x NVIDIA RTX A6000 Ada Generation 48GB DDR6
- T10G Dual Xeon Scalable
  - 1 x NVIDIA RTX A5000 24GB GDDR6
  - 1 x NVIDIA RTX A6000 Ada Generation 48GB DDR6
- ARM Ampere Altra
- APC Metered Rack PDU ZeroU 2G AP8

Motivations

- Research is evolving towards new computational technologies to face the expected requirements for many of the upcoming projects.

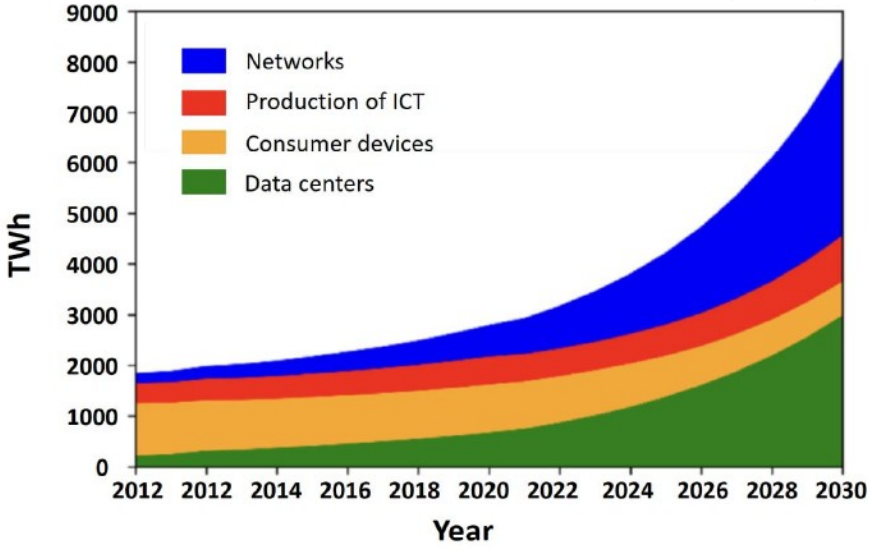


- The energy consumption in data centers is rising significantly,



- **New hardware solutions** (GPUs, FPGAs, …) make a difference?

[Challenges 6, 117 (2015)]

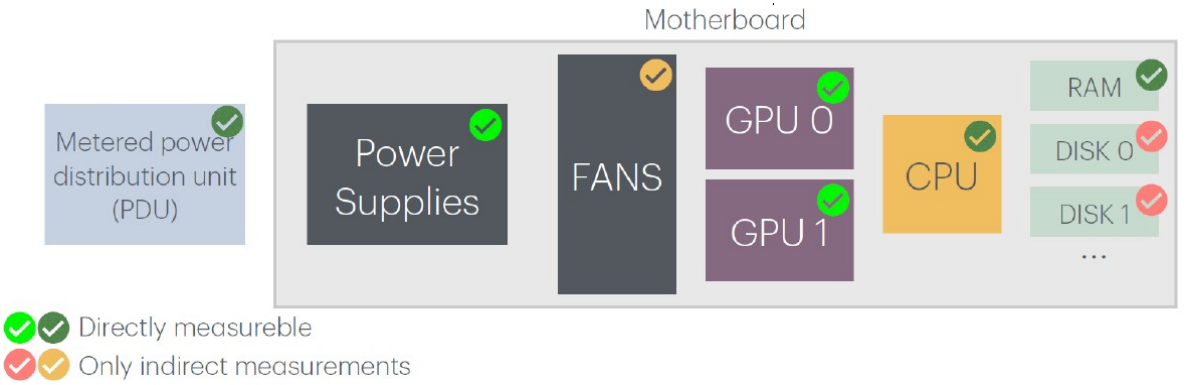


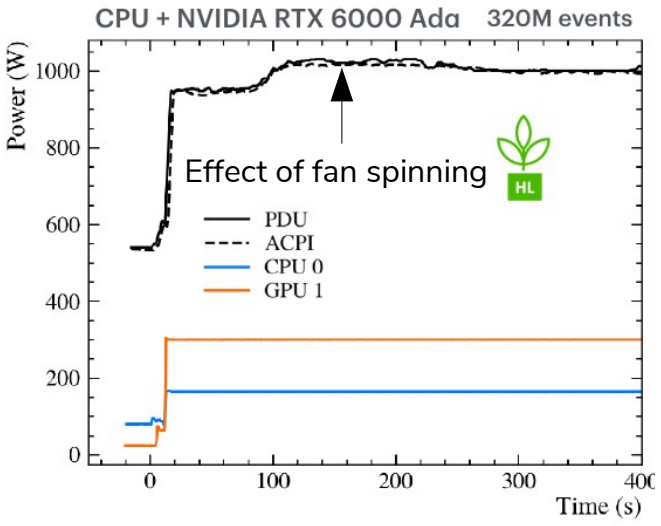
- **Data Centers consumption:**
  - 50-60% IT equipment
  - 35-45% Cooling systems
  - Rest is a few % (lighting, power supplies, monitoring systems, etc.)

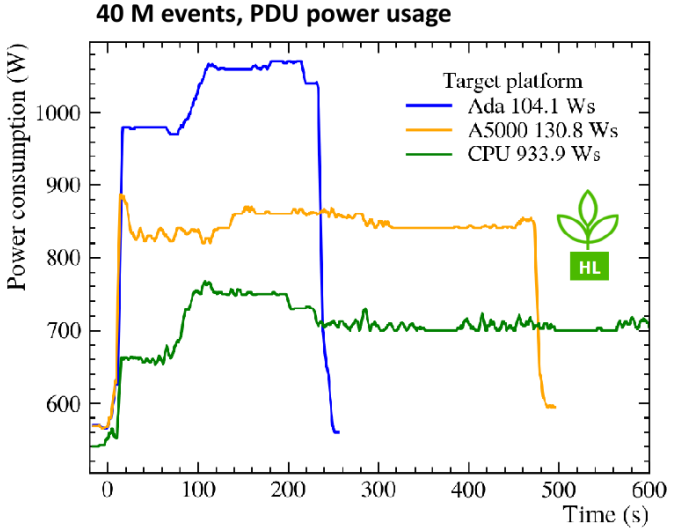
Real Time Analysis @ LHCb

- **Allen**: the LHCb high-level trigger 1 (HLT1) application on GPUs. [LHCB-TDR-021] Fast detector reconstruction in O(500) Nvidia RTX A5000.
- Power consumption measurements performed using dedicated external hardware and specific software to access the built-in sensors.

Fan speed measurement device







EPJ Web of Conferences 337, 01272 (2025)

More powerful devices with faster execution time usually exhibit less power consumption:
(Throughput ↑ Energy ↓)

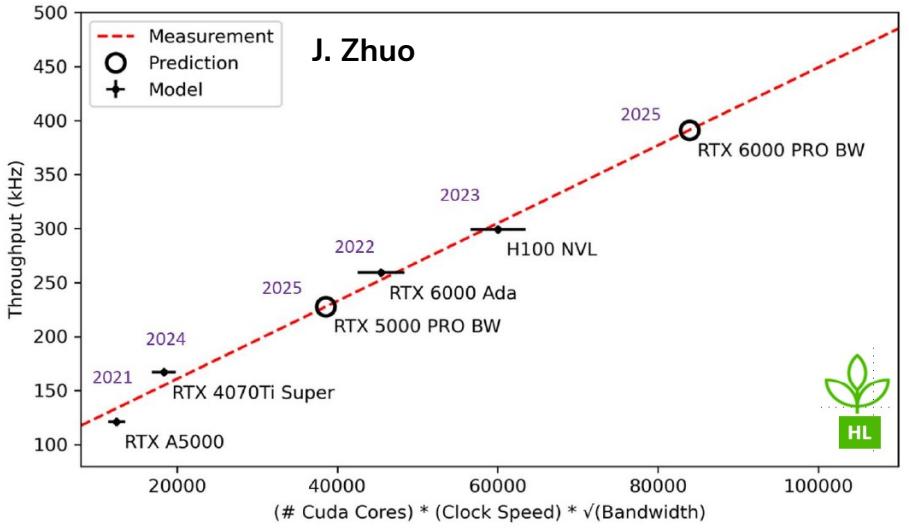
Hardware choice optimization

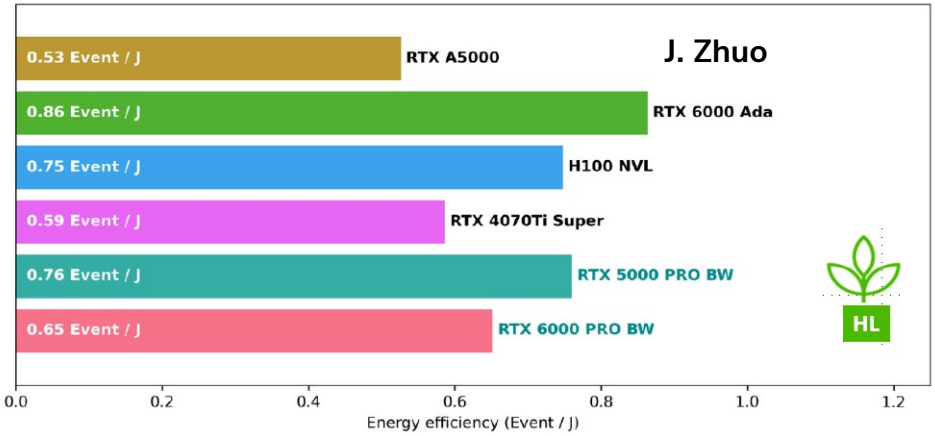
- **Recent Results!**
  - Work in Progress

- Define **Energy Efficiency (EE)** for a given task:
- **EE** = Throughput / TDP (#events/J)
- TDP (W) in a GPU is the Thermal Design Power: the maximum amount of heat the cooling system is designed to dissipate during normal operation.

- Measured throughput (Allen sequence, 5M evts) vs Predicted one from GPU parameters (available in data sheets)

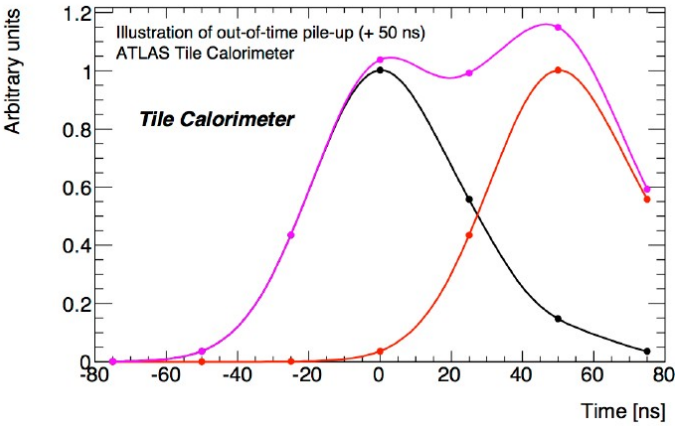
Evaluating the Energy Efficiency (*EE*) of a GPU:





ATLAS TileCal Signal Reconstruction

- During HL-LHC, ATLAS Tile Calorimeter signals will be processed in real-time at 40 MHz with fixed latency before sending it to the L0 trigger.
- Higher pileup and rates require better algorithms than linear filters used for LHC
- Signal processed in Xilinx Kintex Ultrascale KU115
- Input per FPGA is 2x77 channels = 154



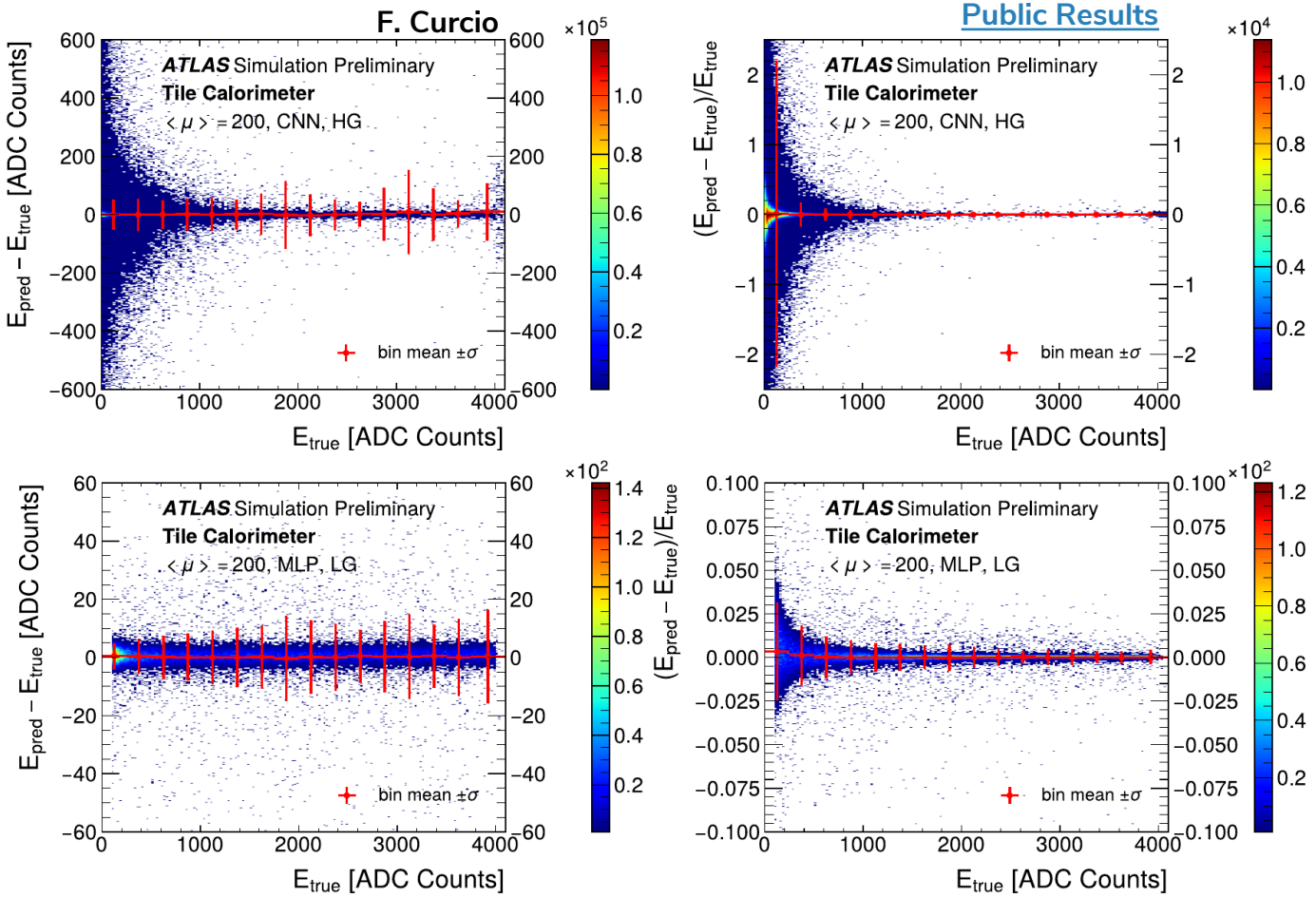
Compact Processing Module (CPM)

- Several Machine Learning algorithms based on MLP, CNN and RNN have been tested.
- Number of parameters is a bottleneck (FPGA resources)

Firmware Block	Latency
Uplink	50 ns, 2 BC
Data Decoder	12.5 ns, 0.5 BC
Energy reconstrucion + sample delay	225 ns, 9 BC
Trigger Packer	12.5 ns, 0.5 BC
Trigger Interface	25 ns, 1 BC
Total	325 ns, 13 BC

ATLAS TileCal Signal Reconstruction

- Input dataset based on specific simulation replicating HL-LHC conditions.
- Algorithms trained with PyTorch.
- CNN trained with hybrid loss (0.5 · MAE + 0.5 · RMSE) gives best trade-off.



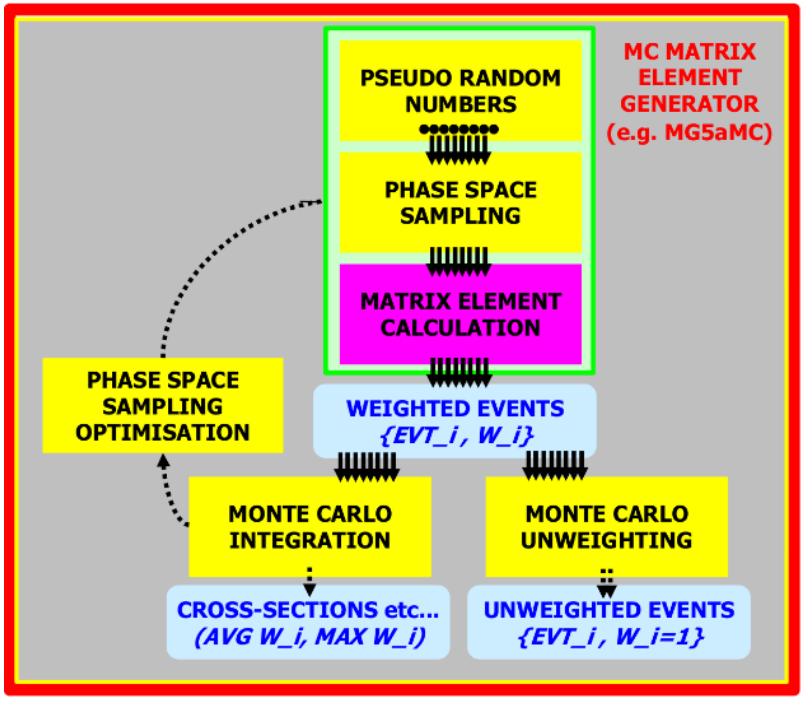
σ_{avg} = 72.1 ADC

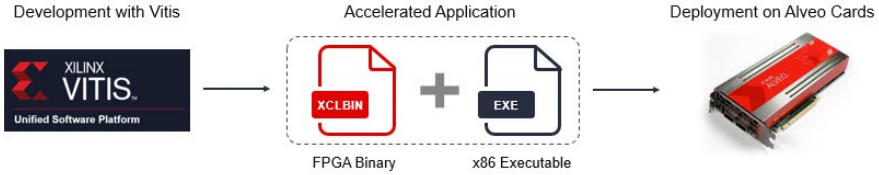
σ_{avg} = 8.4 ADC

Porting MADGRAPH to FPGA using HLS

The Computational Challenge

- Event generation in HEP consumes significant CPU resources.
- As the event generation becomes more and more precise (N^{x}LO) the Matrix Element calculation becomes a bottleneck for HEP computing.
- Explore new architectures (GPUs, FPGAs) to accelerate these calculations in Event Generators like MADGRAPH.
- Projects like Madgraph4GPU already showed nice progress.
- We are testing a further step by porting the ME calculation to an FPGA (Alveo U250) using HLS for several processes
  - $e^+e^- \rightarrow \mu^+\mu^-$ process.
  - gg → ttbar+jets

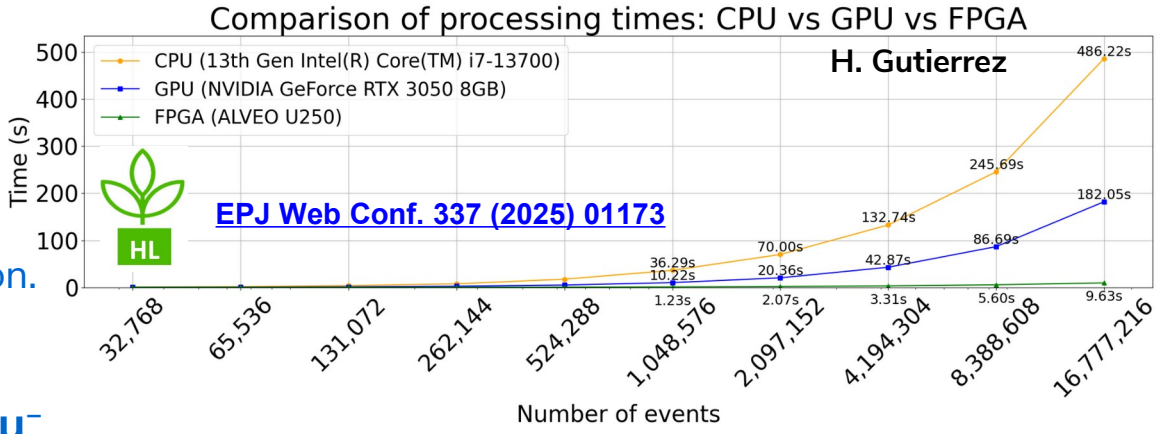




FPGA Performance

Platform and Tools

- **Hardware:** AMD Xilinx Alveo U250 FPGA operating at 120 MHz.
- **Software:** Vitis HLS (C/C++) and Xilinx Runtime (XRT) for host-FPGA communication.
- HLS Adaptation (from GPU to FPGA)



Performance tested on 17M $e^+e^- \rightarrow \mu^+\mu^-$ generated events.

- The FPGA (Alveo U250) completed the simulation in **9.63 s,** 1765 ·10^3 ev/s.
- **~19x faster** than the GPU (NVIDIA RTX 3050 8 GB) 93 ·10^3 ev/s.
- **~51x faster** than the CPU (Intel i7-13700), 35 ·10^3 ev/s.

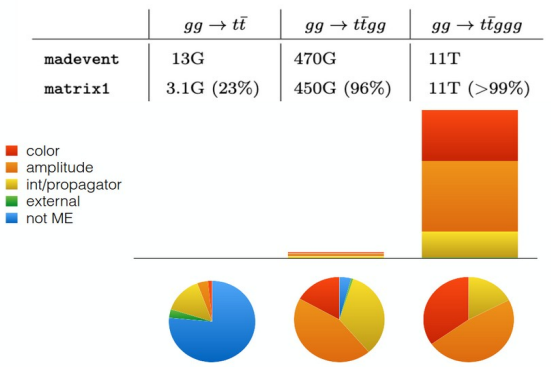
Energy Measurements:

- CPU: 2.287 mJ/s
- GPU: 1.553 mJ/s
- FPGA: 0.016 mJ/s

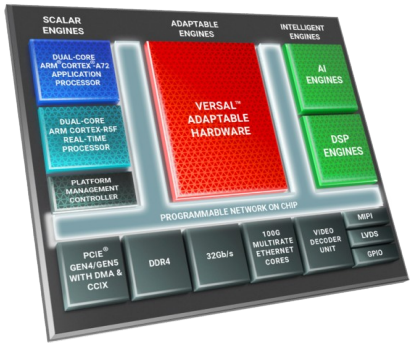
Amplitude and Color in hadronic processes

Platform and Tools

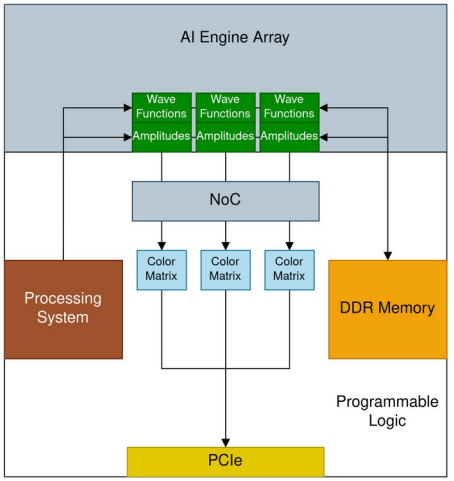
- For complex hadronic processes, time spent by color recombination grows exponentially
- Use Versal AI engines for computation of the color part

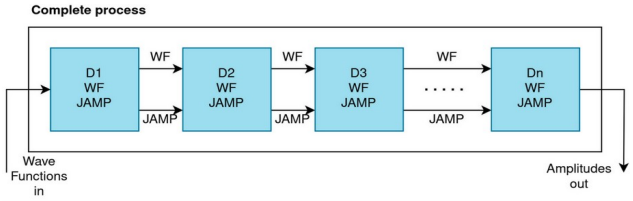


	$gg \to t\bar{t}$	$gg \to t\bar{t}gg$	$gg \to t\bar{t}ggg$
madevent	13G	470G	11T
matrix1	3.1G (23%)	450G (96%)	11T (>99%)



VCK190





Color Matrix (FPGA-HLS)

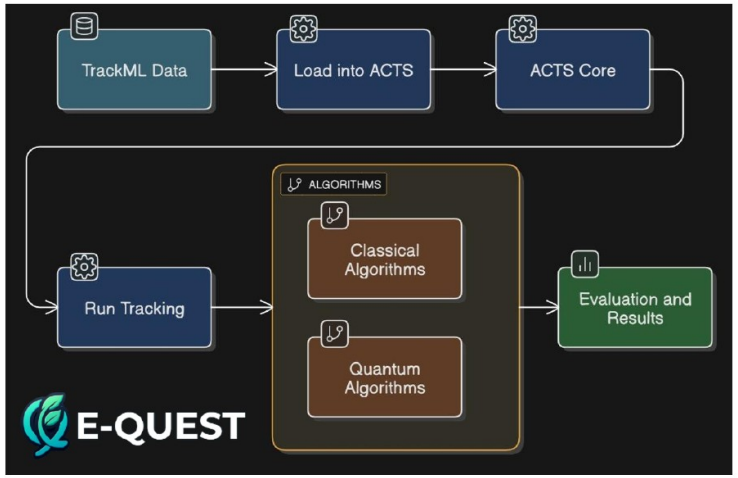
gg -> tt~g

Pipeline Level	AIE HWsim [ns/event]	Max. Cores	* Aggregated HWsim [ns/event/max_cores]	CPU [ns/event]	Speedup Factor
17	1300	5	260	600	x2.3
10	1300	10	130		x4.6
7	1750	14	125		x4.8

Process	t_CPU/evt (ns)	t_FPGA/evt (ns)
gg -> tt~g	24	13
gg -> tt~gg	120	16
gg -> tt~ggg	4400	2500

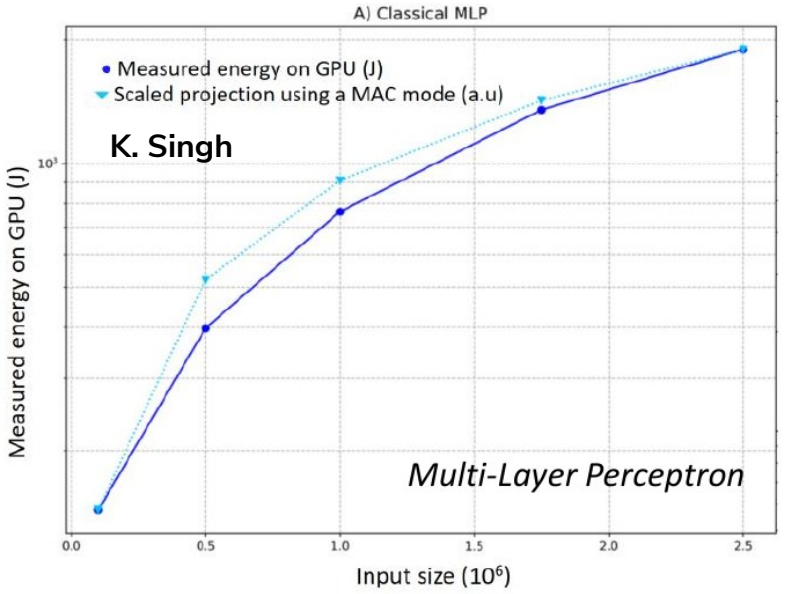
Sustainability in Quantum Computing

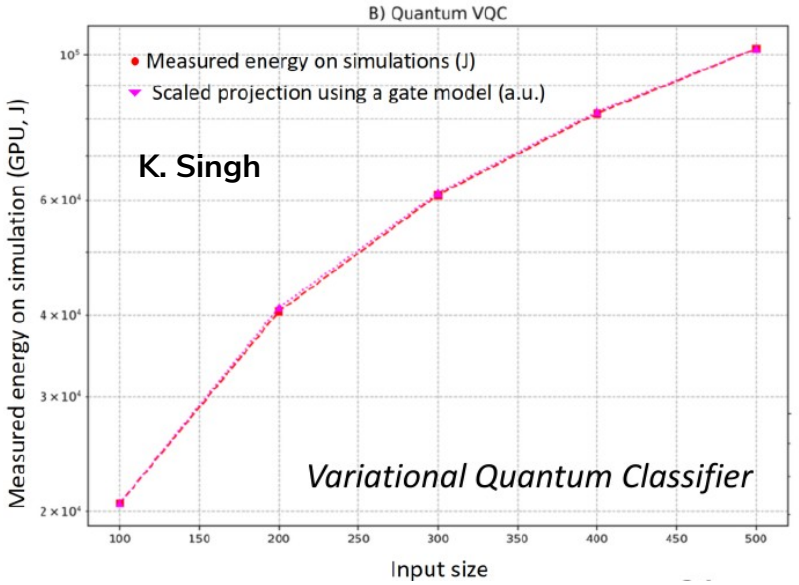
- Quantum computing is an emerging technology.
- Testing quantum computing on complex reconstruction tasks, like tracking.



- General framework for studying the energy consumption of quantum and classical computation being established to compare power consumption with classic algorithms.

[PRX Energy 4 (2025) 2, 023008]





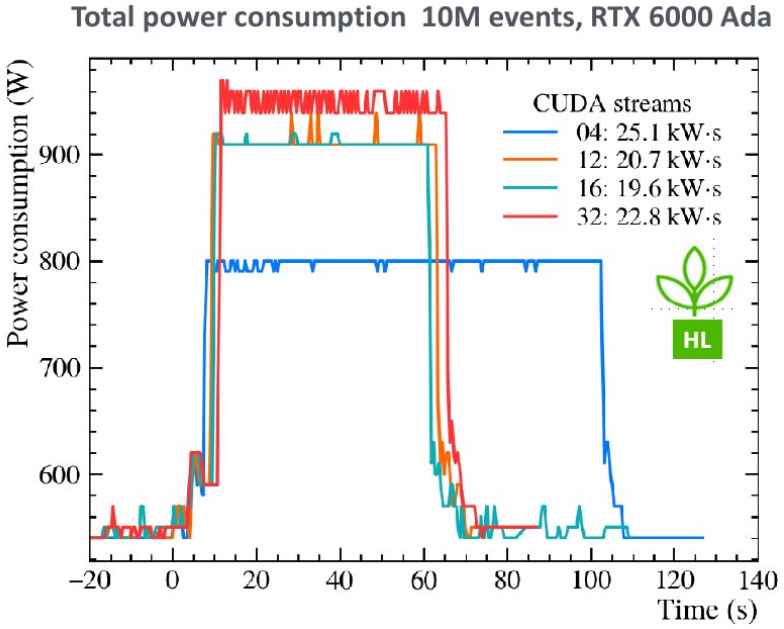
Conclusions

- Computing in HEP is evolving fast.
- New computational technologies are required to face many of the upcoming projects.
- Use the best and more efficient available hardware.
  - Optimize the utilization of the hardware.
  - Optimize the software design to increase throughput.
- Machine Learning and complex tasks can be ported to low level architectures (FPGA).
- Heterogeneous computing systems allows to optimize the resources and power consumption.
- Quantum Computing is an emerging technology, power consumption efficiency is under study.

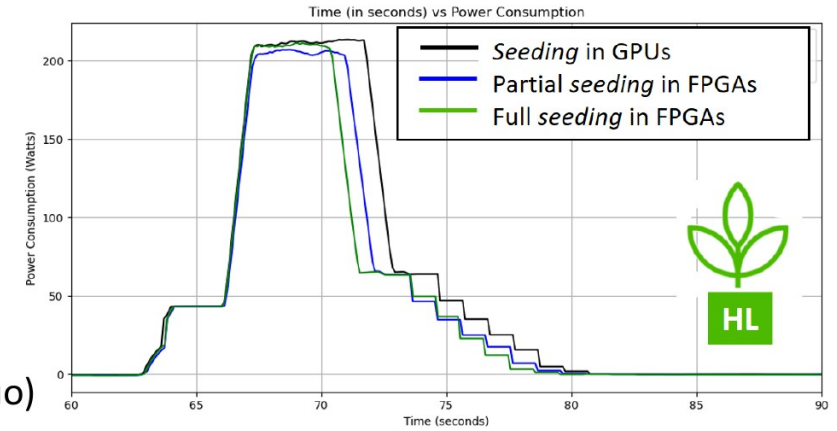
Backup

LHCb Hardware optimization

- Optimization of hardware utilization
- GPU parallelisation (# CUDA streams)
- Optimize the GPU usage for higher performance



- Checking other hardware architectures in tracking reconstruction:
- Real-time reconstruction on FPGAs with the “artificial retina”architecture
- Clustering of the VELO detector already running for Run3 in FPGAs
- SciFi tracking in development for Run4 (2030) [CERN-LHCC-2024-001]



(J. Zhuo)

[https://cds.cern.ch/record/2888549]

Activity Example: DNN Signal Reco

arXiv.org > physics > arXiv:1903.02439

Search or Article

(Help | Advanced se

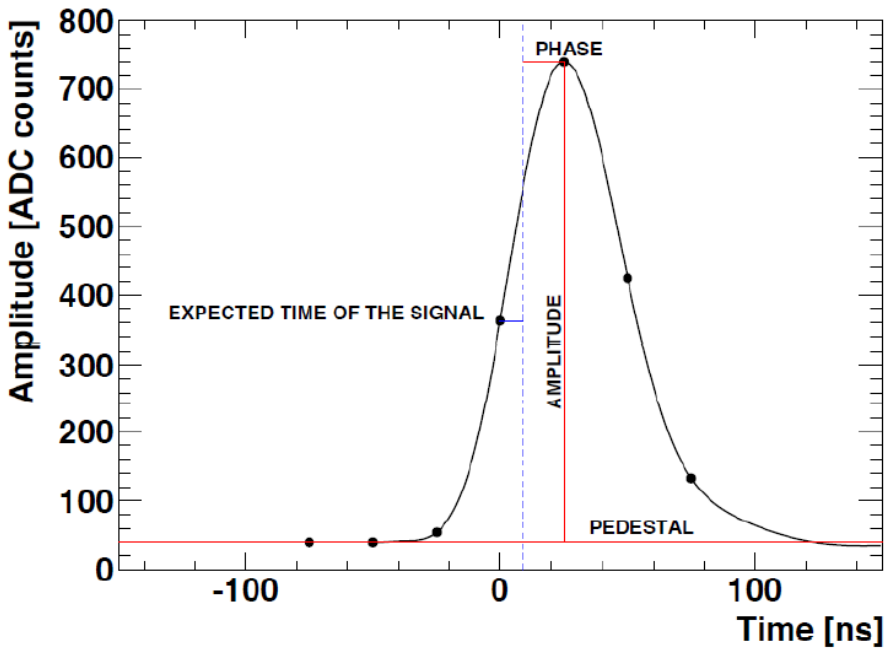
Physics > Instrumentation and Detectors

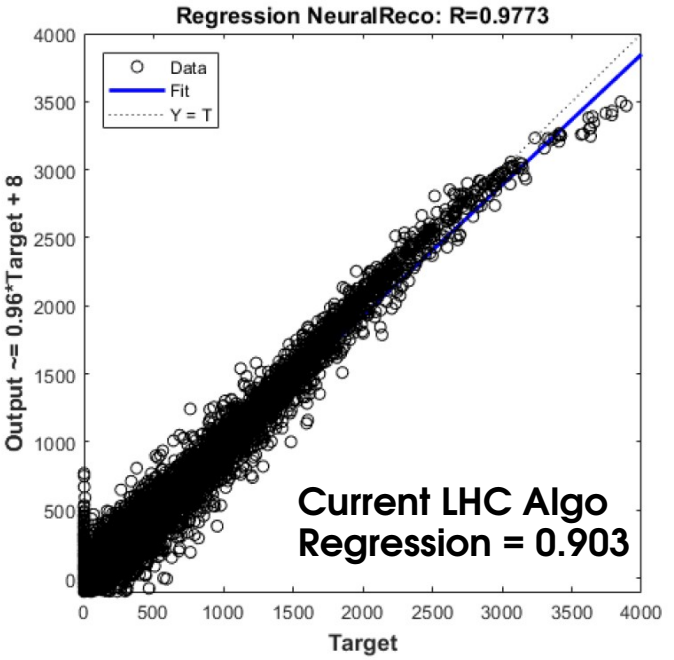
FPGA implementation of a deep learning algorithm for real-time signal reconstruction in radiation detectors under high pile-up conditions

J. L. Ortiz, F. Carrió, A. Valero

(Submitted on 6 Mar 2019)

- **Deep NN for real-time signal reconstruction** at the HL-LHC.
- HL-LHC pileup degrades the pulse quality, LHC algos performance deteriorates.
- 128 FPGAs to process the full ATLAS Tile Calorimeter.
- Each FPGA process 96 different signals with 40 MHz rate





Activity Example: Machine & Deep Learning

- Application of **Machine Learning** in LHC searches for new physics.

- **Computer Vision** (ConvNets) Network for event classification, muon tomography and search for long-lived particles
  - Application as well for satellite imaging, medical diagnosis, etc.