Sustainability of real time analysis at 5 TB/s data rate

Arantza Oyanguren (IFIC – Valencia)

FH- Sustainability forum @ DESY 12th May 2025

IFIC INSTITUT DE FÍSICA CORPUSCULAR
Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Plan de Recuperación, Transformación y Resiliencia
GENERALITAT VALENCIANA Conselleria de Educación, Universidades y Empleo
GVANEXT Fondos Next Generation en la Comunitat Valenciana
CERN

Outline

- The HIGH-LOW project at IFIC - Valencia
- Real Time Analysis at LHCb
- Impact of hardware on power consumption
- Hardware utilisation
- Software optimization
- Conclusions & prospects

About me

Professor at the University of Valencia (Spain), researcher at IFIC.
https://webific.ific.uv.es/web/
IFIC team leader of the LHCb experiment at CERN.
https://lhcb.web.cern.ch/
Expert on heavy flavor physics (DELPHI, BaBar, Belle II, LHCb)
Hardware and physics analyses.
Strong interest on the use of new hardware architectures and AI for HEP, in particular for trigger systems (aiming to detect long-lived particles).
Previous coordinator of the hw accelerators WP of the Real Time Analysis project at LHCb.
Founder of the Computing Challenges (COMCHA) network in Spain https://comcha.es/
Executive member of the AIHUB-CSIC. https://aihub.csic.es/en/
Previous executive member of ARTEMISA committee at IFIC, a platform based on GPUs for Machine Learning. https://artemisa.ific.uv.es/web/

Working with an amazing group of people from ATLAS and LHCb !!

The HIGH-LOW project at Valencia

Arantza Oyanguren

Jiahui Zhuo

Volodymyr Svintozelskyi

Valerii Kholoimov

Álvaro Fernández

Miriam Lucio

Luca Fiorini

Alberto Valero

Antonio Cervello

Francisco Hervás

Héctor Gutierrez

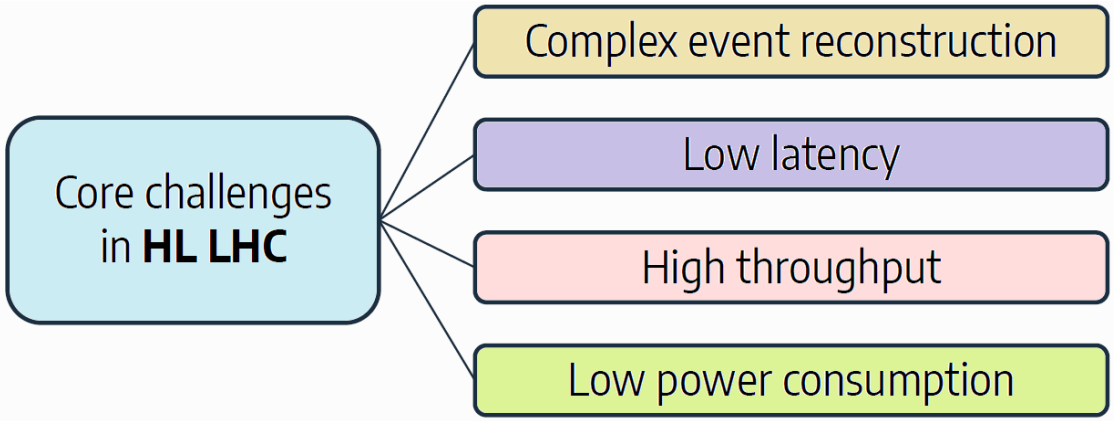
Francesco Curcio

The HIGH-LOW project at Valencia

Design of HIGH performance algorithms for LOW power sustainable hardware for LHC experiments and their upgrades

PIs: L. Fiorioni (ATLAS) & A. Oyanguren (LHCb)

Funded by the Spanish Ministry of Science and Innovation (TED2021-130852B-I00)

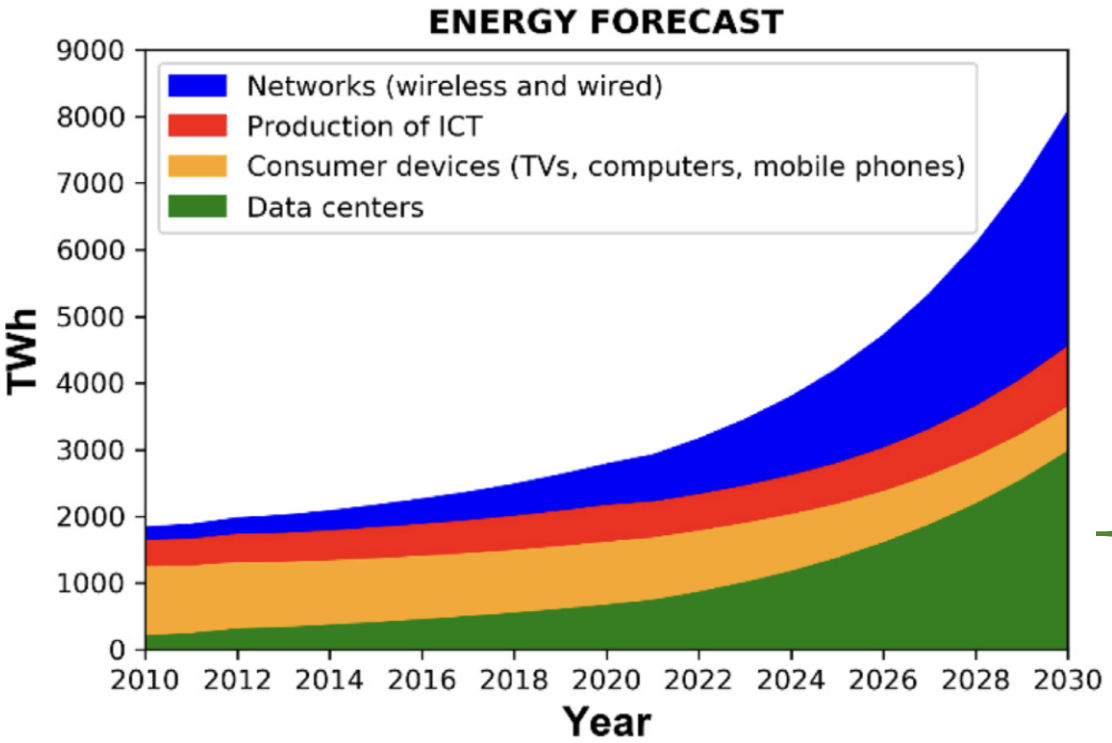


Designing **AI-driven**, **high-performance** algorithms optimized for **low-power**, sustainable hardware in LHC upgrades.

→ Extensible to other HEP (and non-HEP) experiments which rely on CPU architectures (lack scalability, significant energy constraints).

The HIGH-LOW project at Valencia

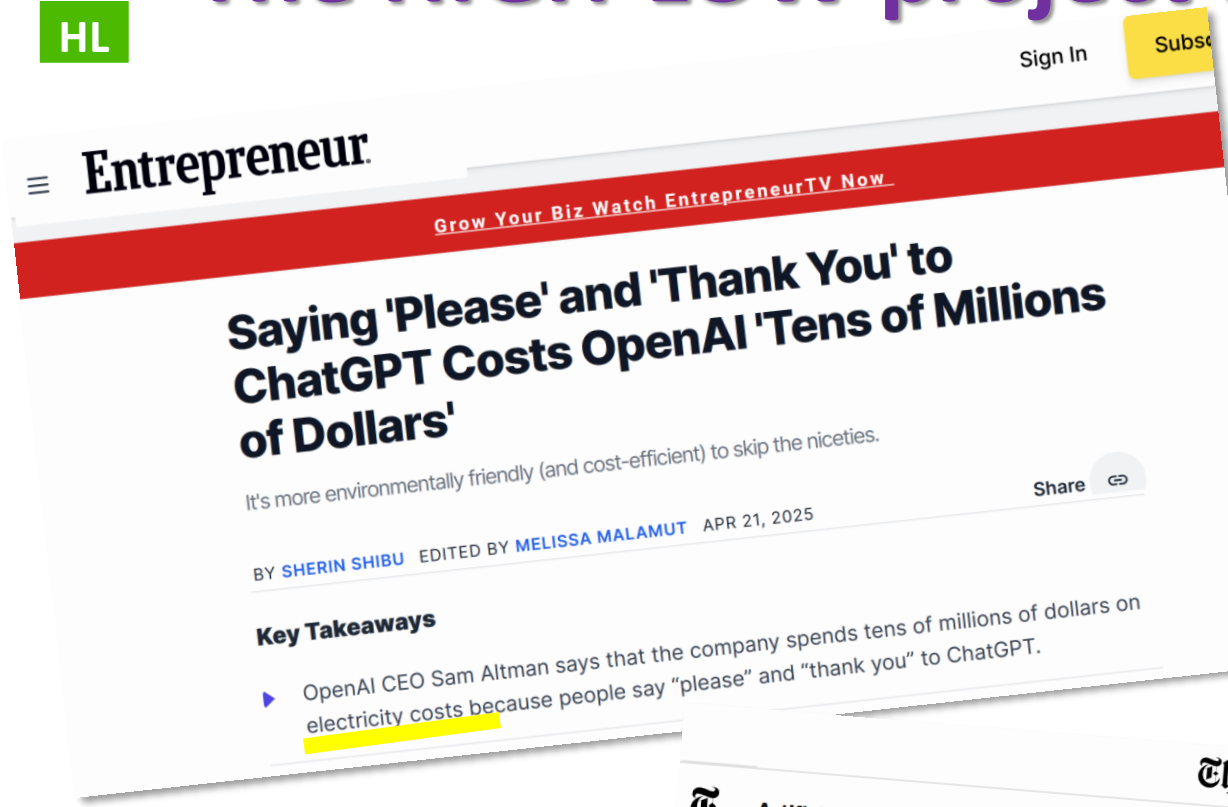
The energy consumption in data centers is rising significantly due to AI utilization and other digital technologies:



- 50-60% IT equipment
- 35-45% Cooling systems (HVAC, liquid)
- Few %: lighting, backup generators, power supplies, security and monitoring systems, etc…

Can we make a difference using specific hardware and optimized software?

The HIGH-LOW project at Valencia



(Each ChatGPT-4 query requires about 2.9 watt-hours i.e, X 10 Google search)

The HIGH-LOW project at Valencia

Quantifying is important to take decisions…. we try to do that at IFIC

Noise isolated small room at the IFIC experimental lab area with racks

The HIGH-LOW project at Valencia

The HIGH-LOW project at Valencia

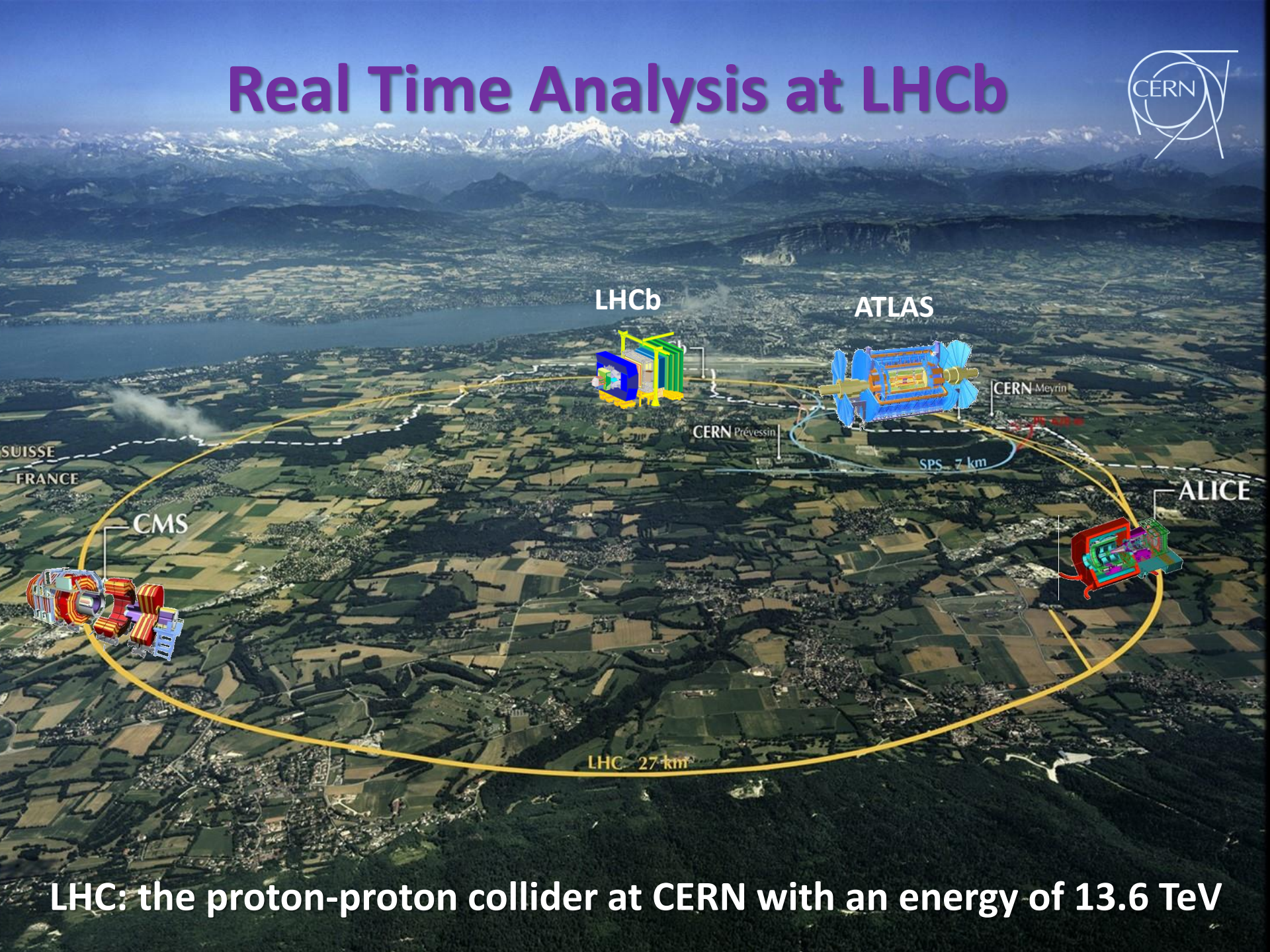
Rack K RETEX LOGIC-2 A600 42U F1000 PH
APC Metered Rack PDU ZeroU 2G AP8
SWITCH D-LINK DXS-1210-28T 24x 10GB

HL01
2 x Intel Xeon Gold 5318Y (24 cores), 256GB DDR4 RAM, 80TB storage with NVIDIA GPUs RTX A5000 and RTX A6000 Ada

HL02
AMD EPYC 9474F, 768GB DDR5 RAM, 160TB storage with NVIDIA GPUs RTX A6000 Ada and 2 H100 NVL

Other small devices to measure fan speed, temperature, etc…

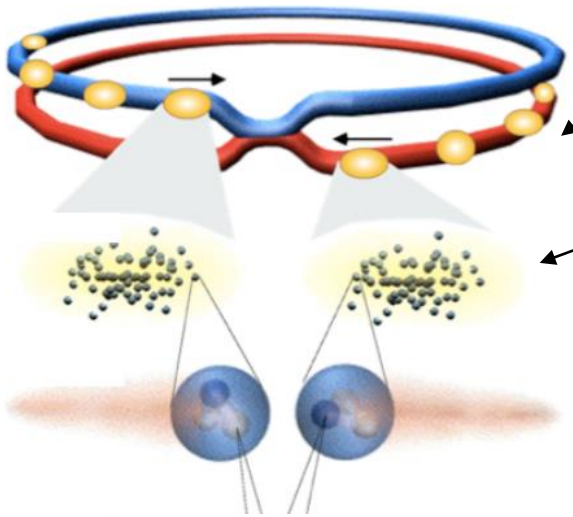
Real Time Analysis at LHCb



LHC: the proton-proton collider at CERN with an energy of 13.6 TeV

Real Time Analysis at LHCb

Proton-proton collision



2028 bunches of protons per beam

Beam energy of 7 TeV

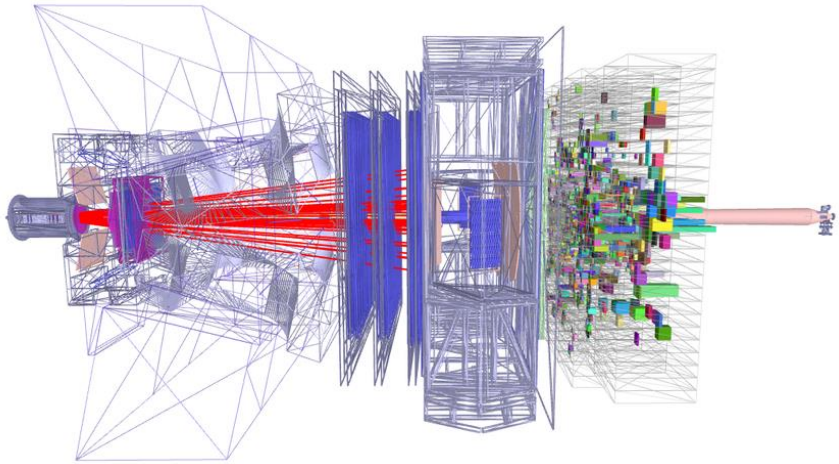
Luminosity 10^{34} cm^{-2} s^{-1}

10^{11} protons per bunch

Crossing rate 40 MHz, i.e. 40 M collisions/s

and ~ 1 MB data from detectors

→ 5 - 40 TB/s

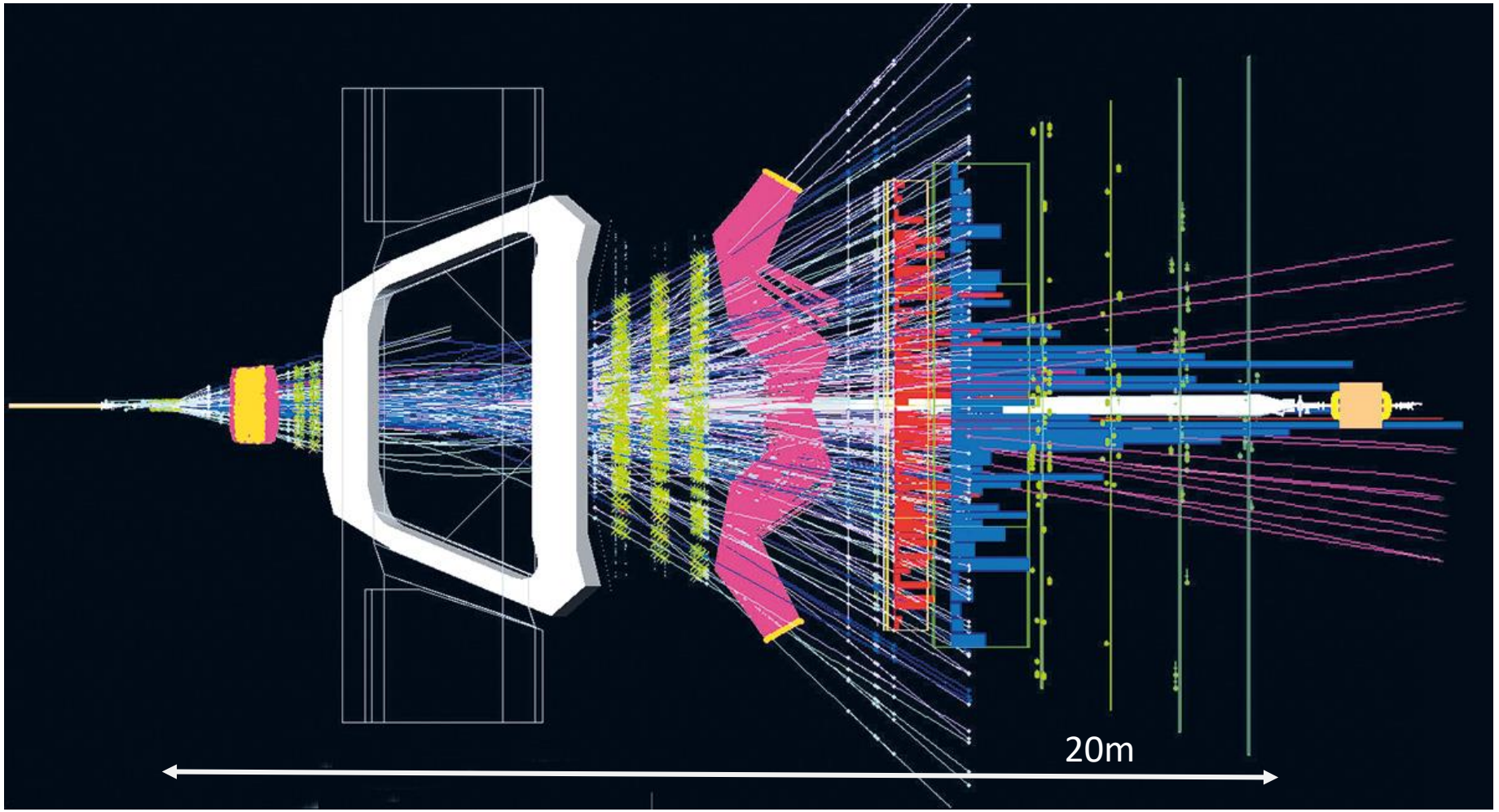


We need to select the events to be persisted!

Google+ + +

~ 0.3 TB/s

Real Time Analysis at LHCb



Tracking, particle identification (RICH, muons) and calorimeter systems + magnet

(millions of readout channels)

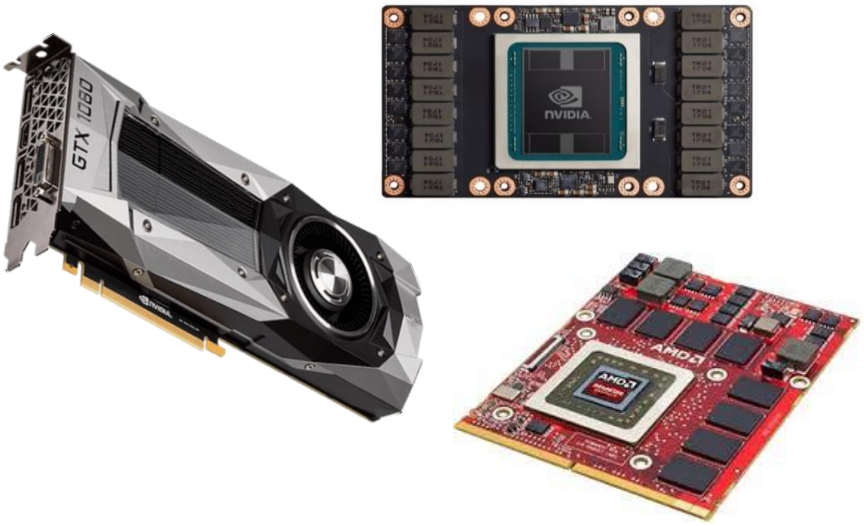
Real Time Analysis at LHCb

Exploiting hardware accelerator technologies in event reconstruction:

→ Use more than one kind of processor or cores to maximize performance and energy efficiency.

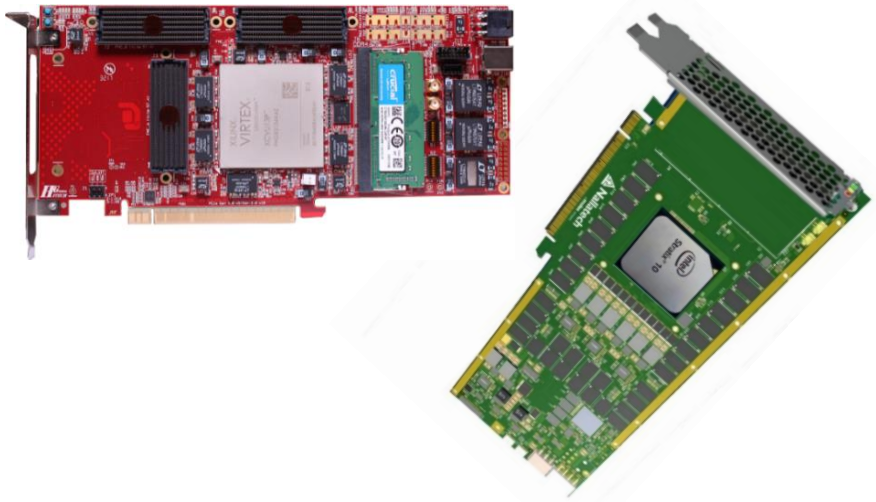
→ Exploit the high level of parallelism to handle particular tasks.

Graphic Processor Units (GPUs)



- Multicore processors, highly commercial
- High throughput (# processed events / time)
- Ideal for data –intensive parallelizable applications

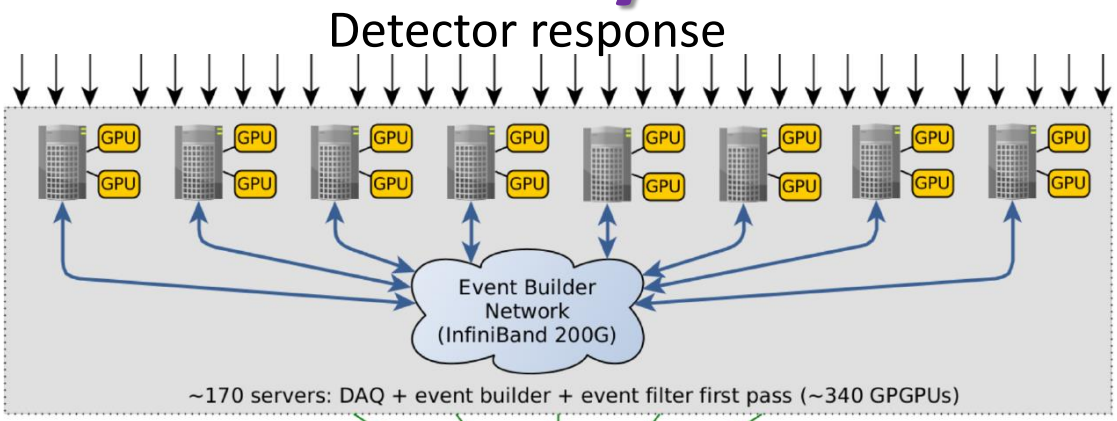
Field Programmable Gate Arrays (FPGAs)



- Programmable and flexible devices
- Low latency
- Low power consumption
- Ideal for compute- and data-intensive workloads

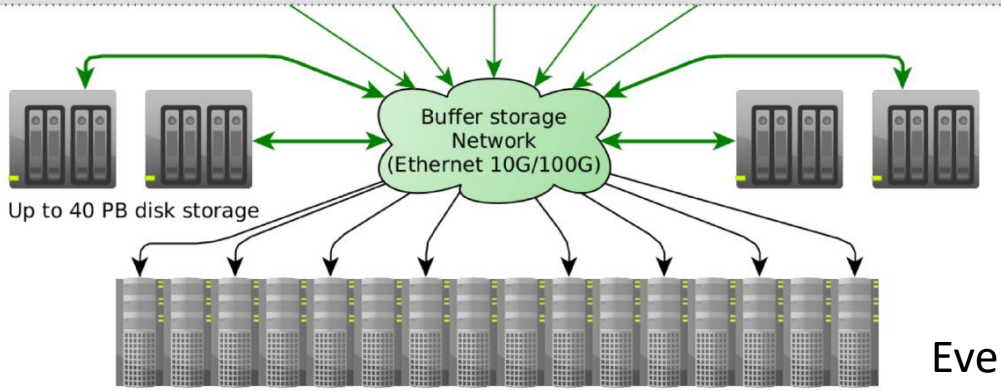
Real Time Analysis at LHCb

30 MHz
(non-empty pp)
5 TB/s



Event builder
& event filter 1

1 MHz
100 GB/s



100 KHz
10 GB/s

Event filter 2

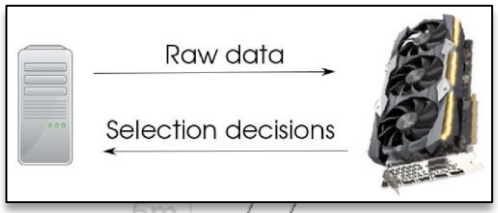
Event filter second pass (up to 4000 servers)

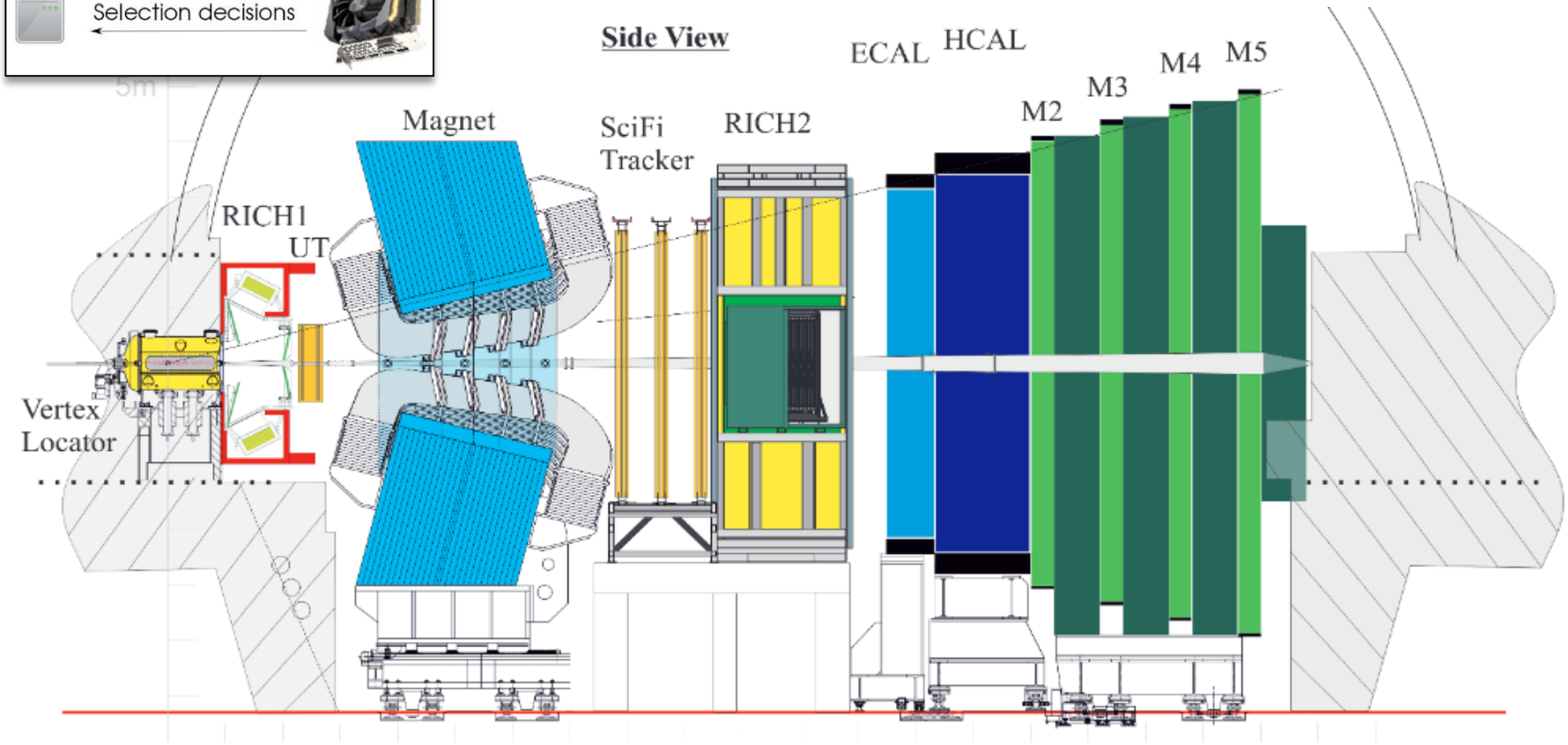
LHCb data
center at Pit 8

Real Time Analysis at LHCb

Allen: the LHCb high-level trigger 1 (HLT1) application on GPUs.

[LHCB-TDR-021] → Fast detector reconstruction in O(500) Nvidia RTX A5000





[Com. Softw Big Sci 4, 7 (2020)]

https://allen-doc.docs.cern.ch/

Real Time Analysis at LHCb

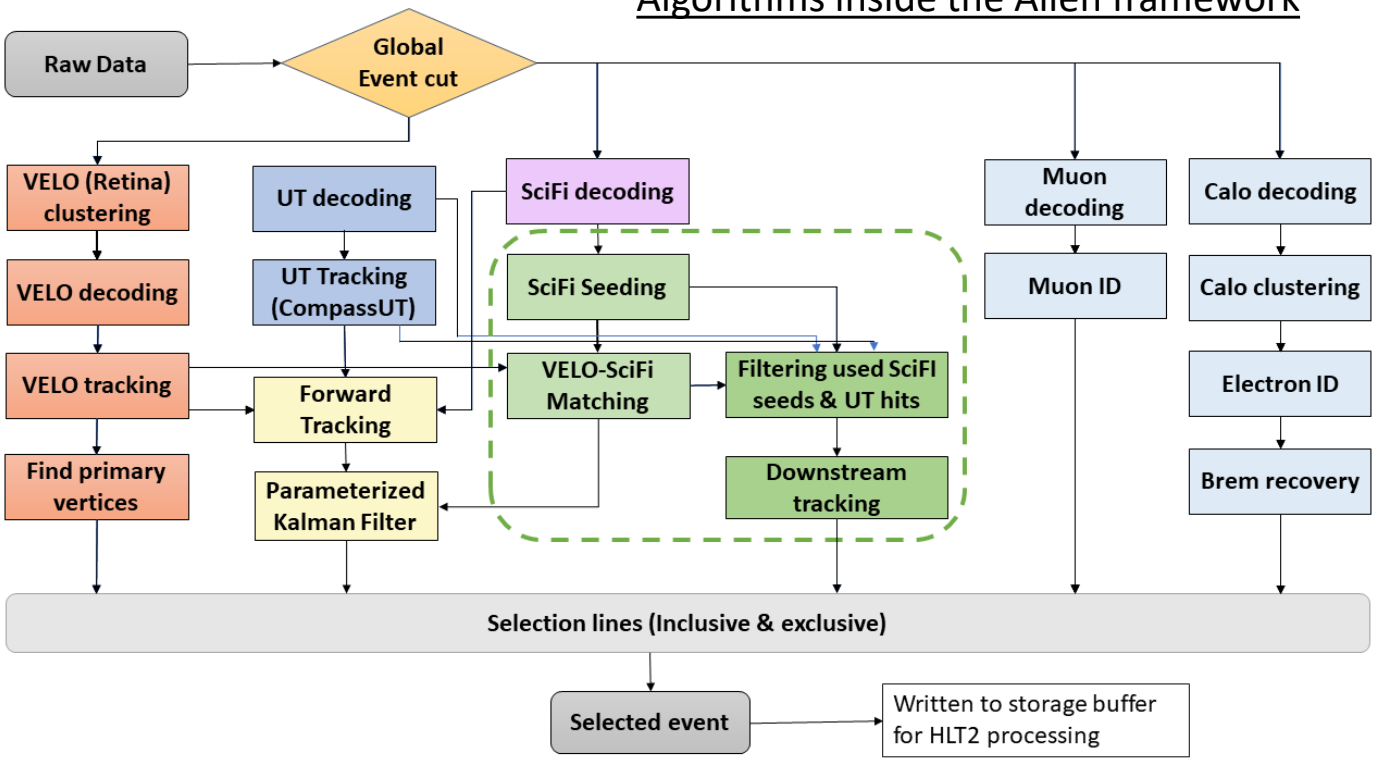
[*Allen project*]

Open-source!

- Portable: executed on several architectures: CPU, GPU
- Modular: design allows various execution sequences
- Total of approx. 250 algorithms used in data-taking
- It has to reduce in real time 40 Tbits/s by a factor 50
- LHCb simulation samples available

Case study for power consumption measurements

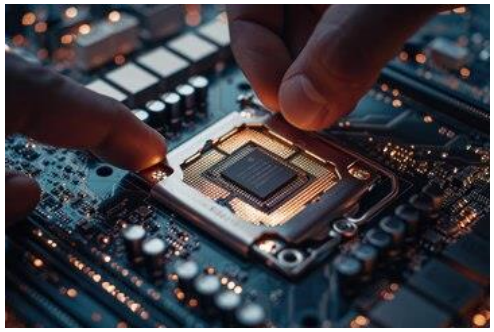
Algorithms inside the Allen framework

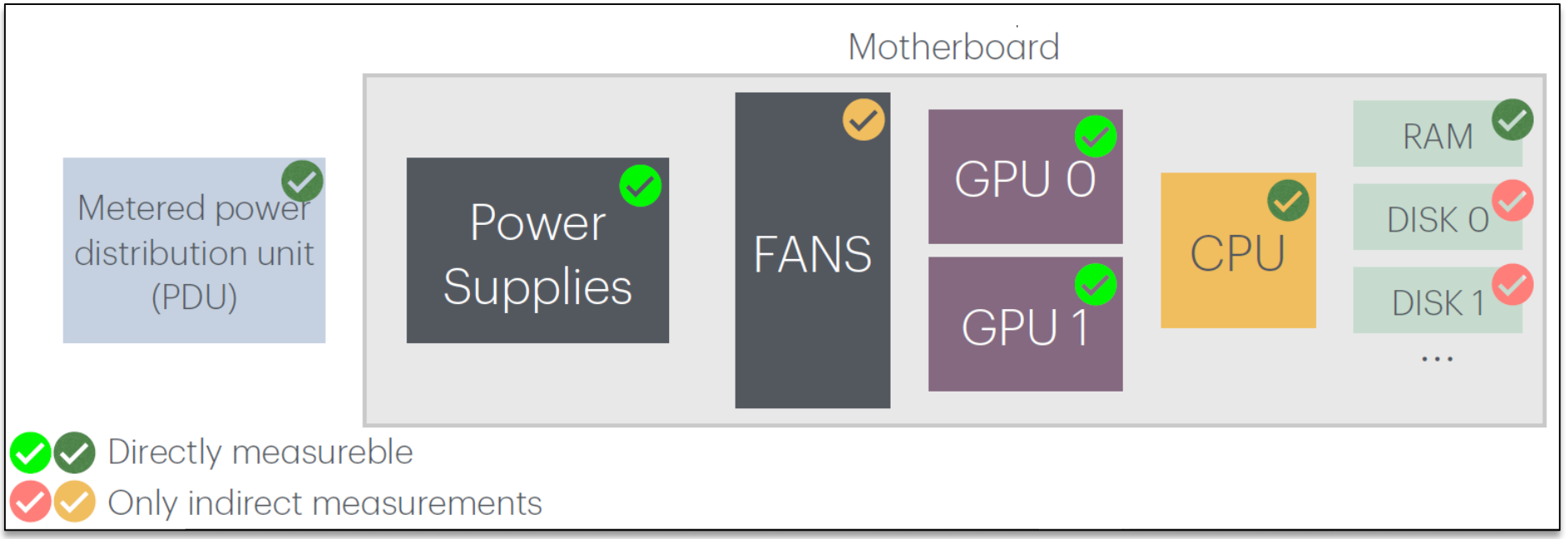


Impact of hardware on power consumption

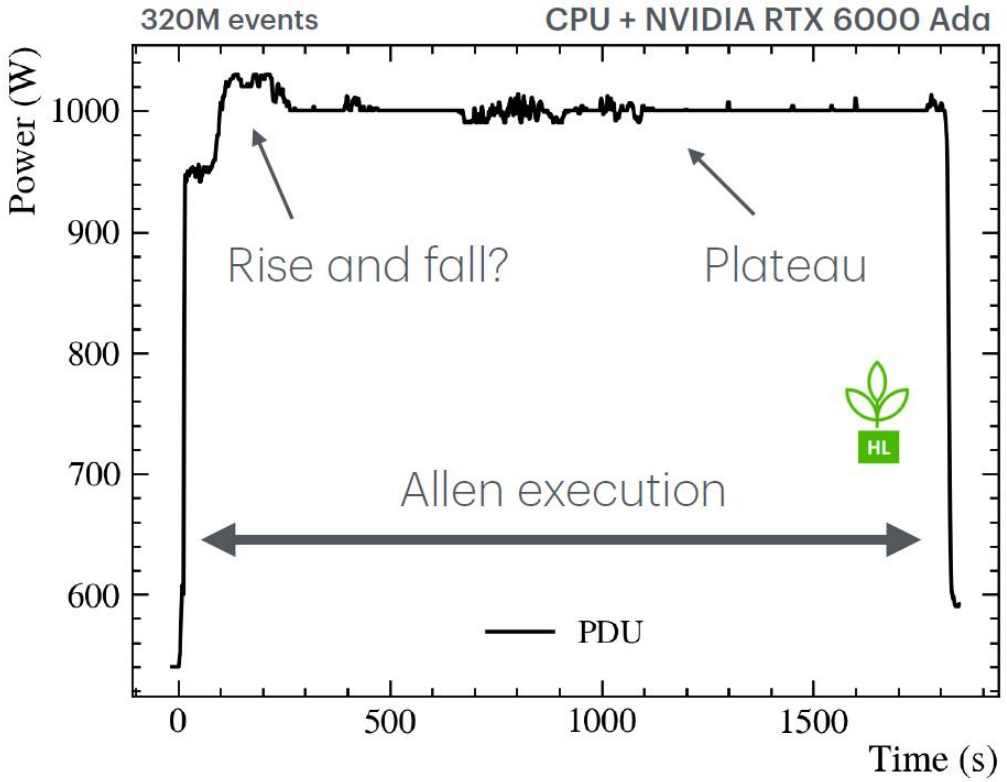
Power consumption measurements can be performed using dedicated external hardware or using specific software to access the buil-in sensors.
In our case, we are based on:

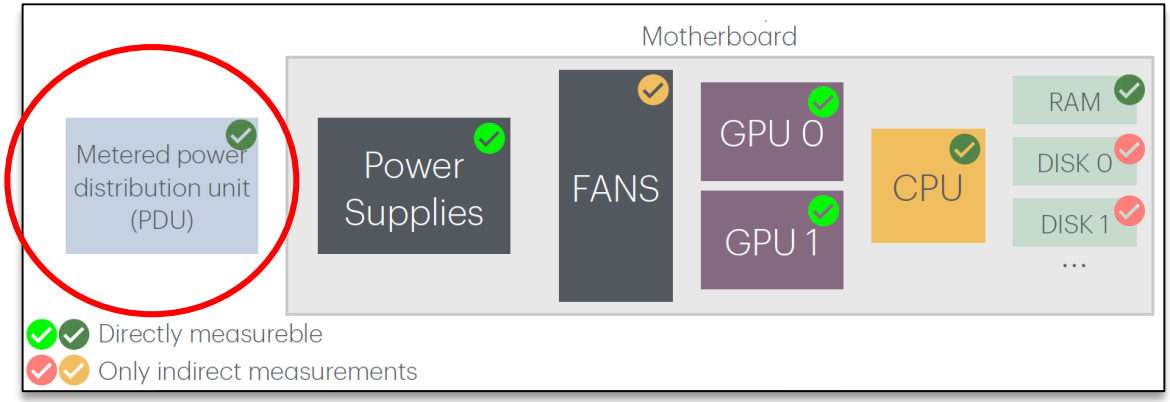
- A metered power distribution unit (PDU)
- Relying on device drivers (*Nvidia DCGM*)
- Reading of CPU performance counters (ACPI, RAPL)





Impact of hardware on power consumption



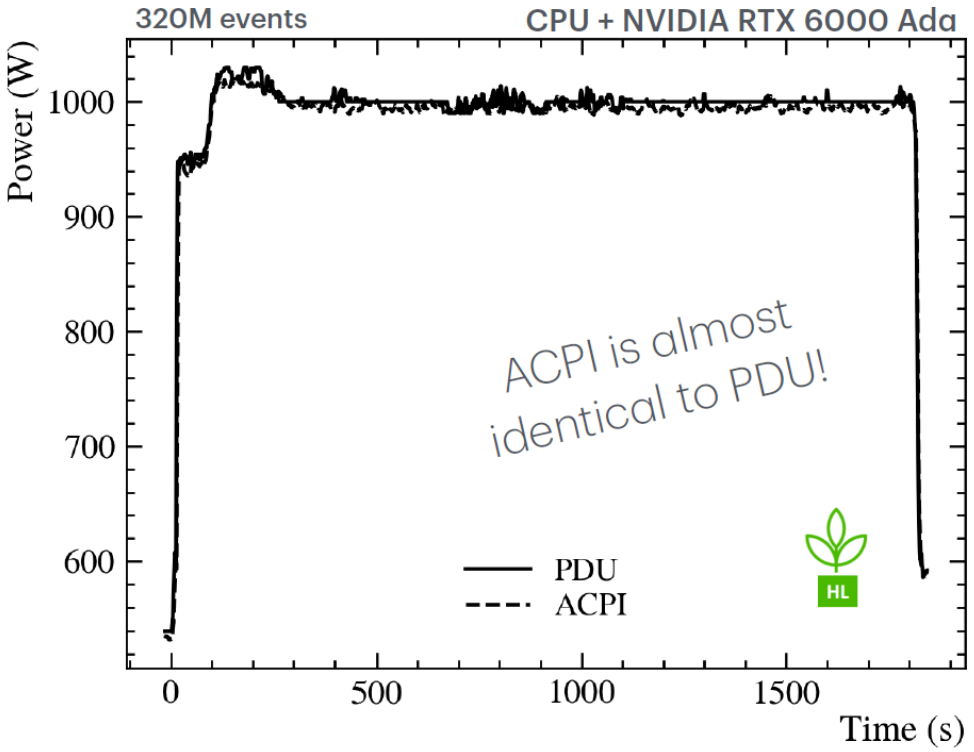


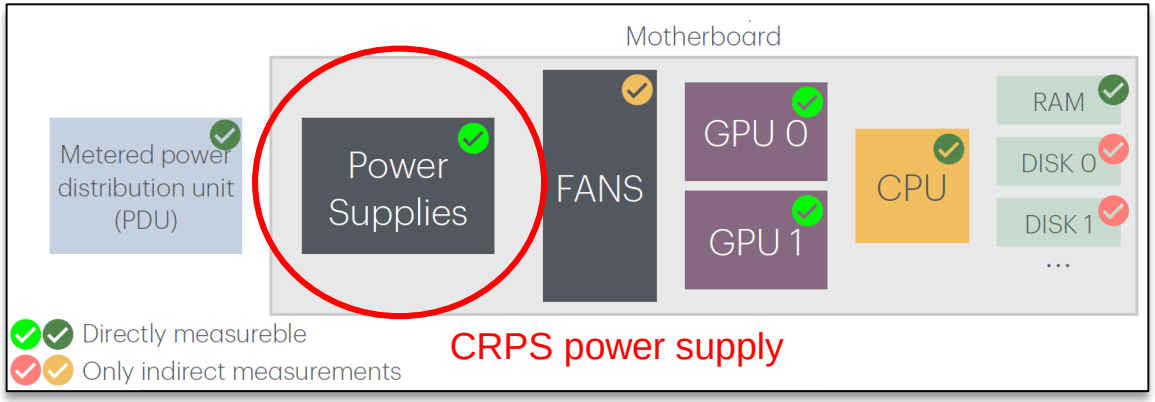
APC PDU ZeroU
2G APS

Impact of hardware on power consumption

Advanced Configuration and Power Interface (ACPI)

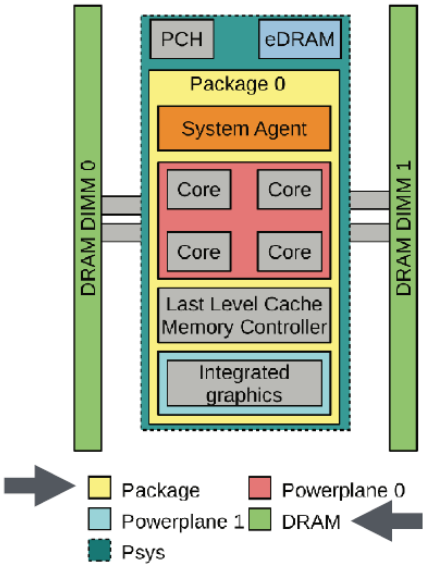
Hardware abstraction interfaces between the hardware and the OS

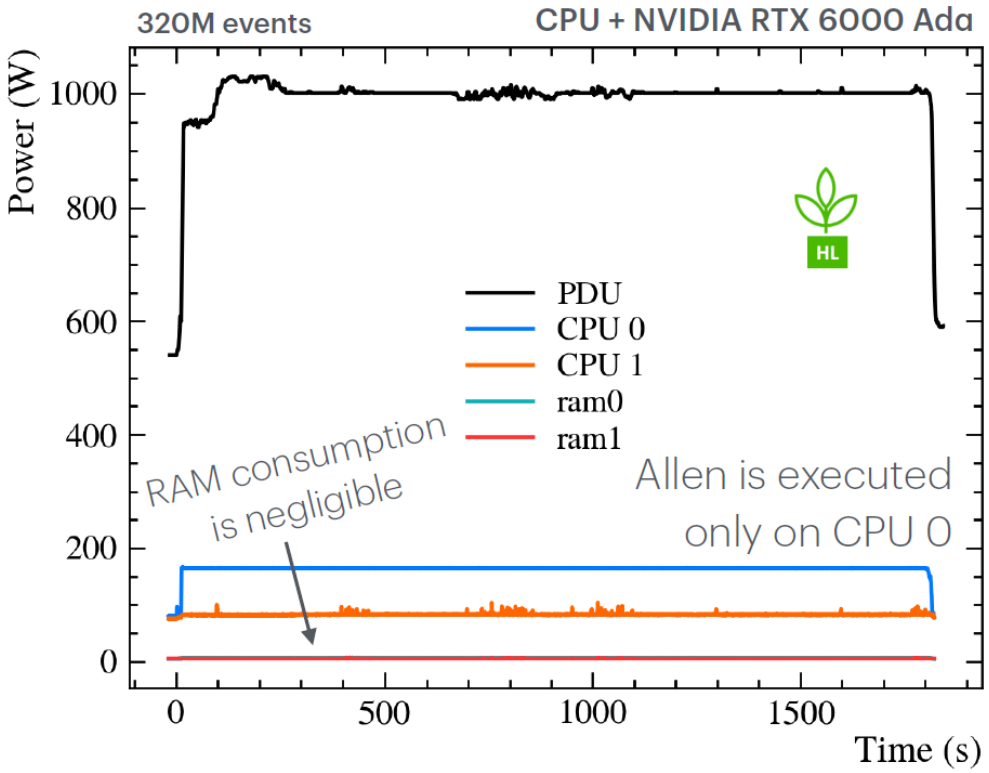


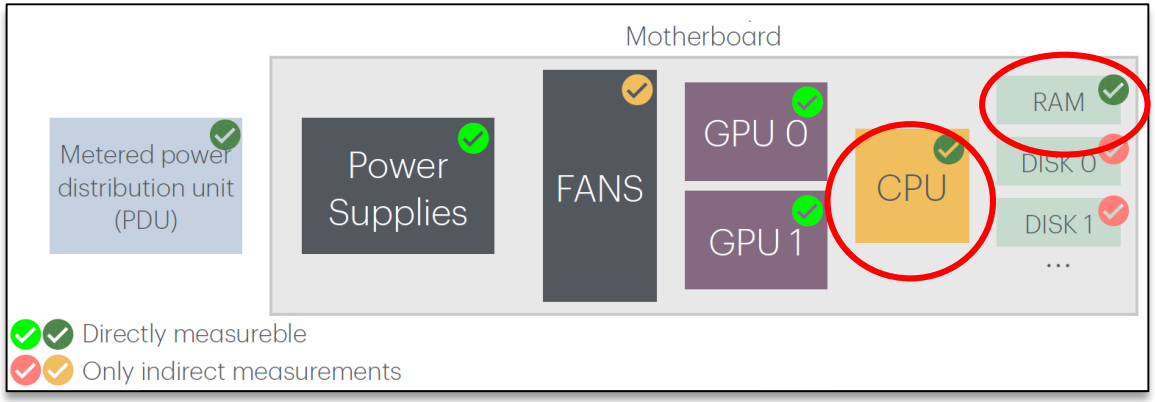


Impact of hardware on power consumption

Running average power limit (RAPL) based on CPU performance counters

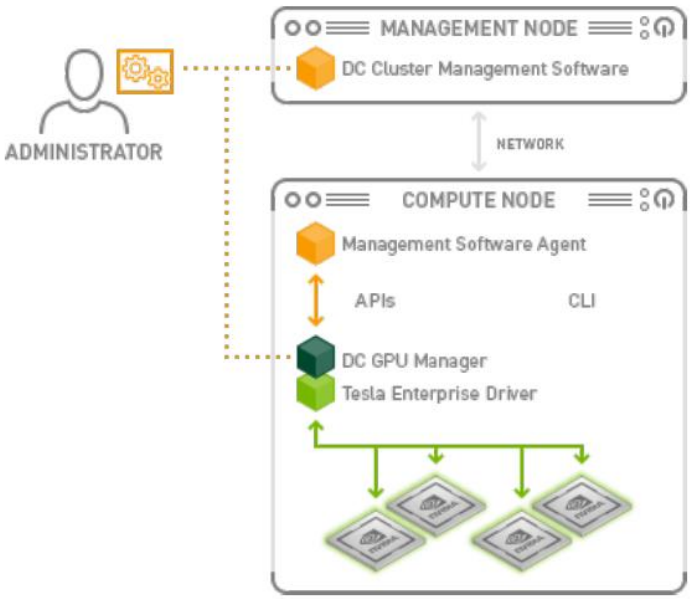


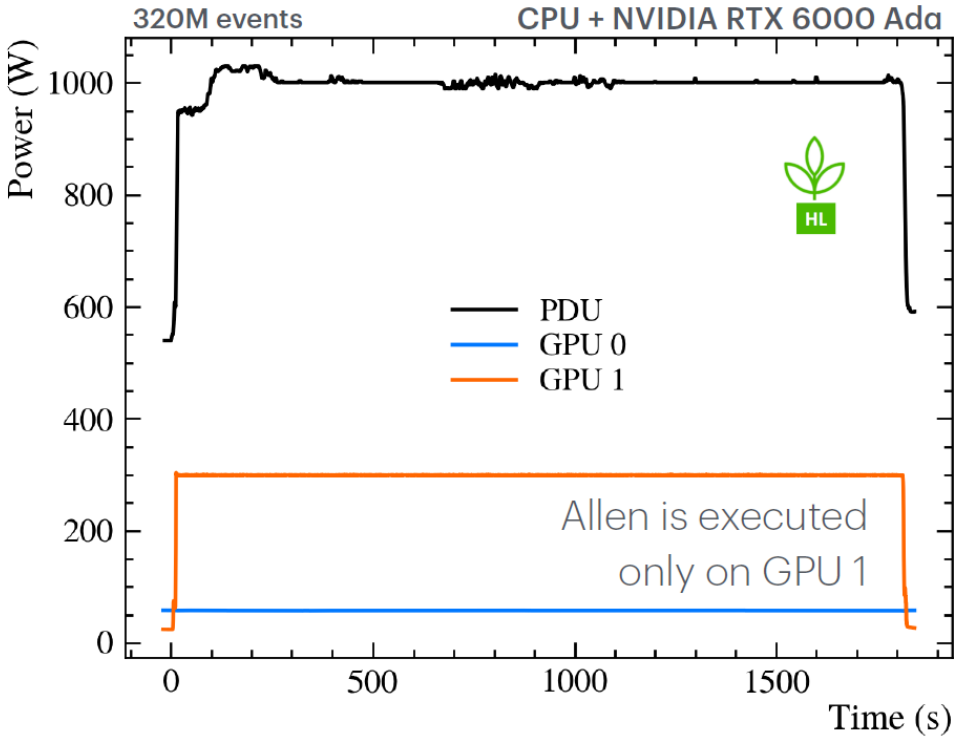


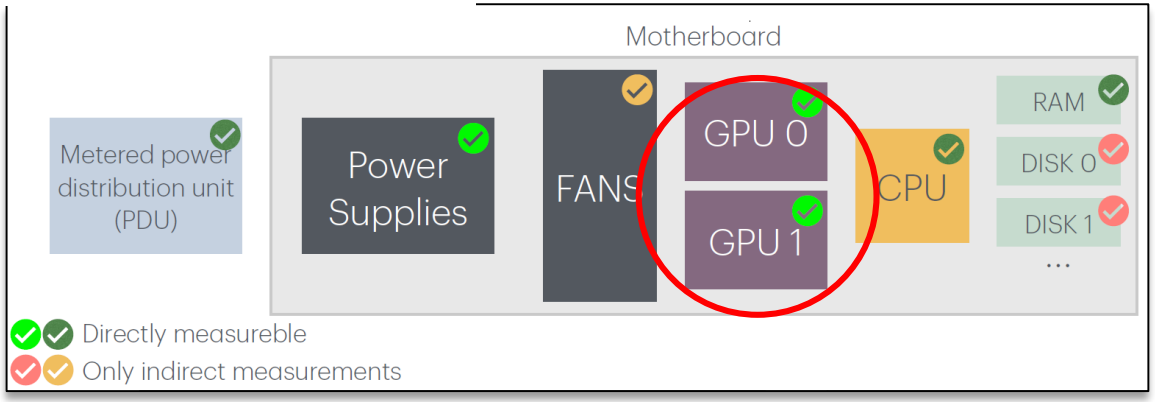


Impact of hardware on power consumption

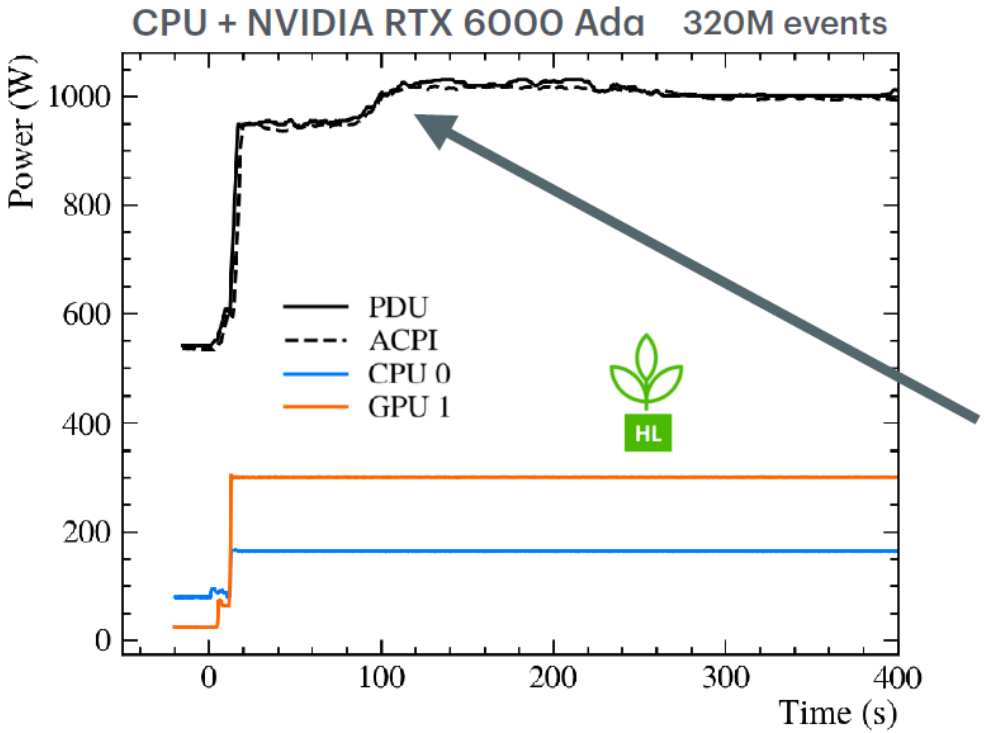
NVIDIA Data Center GPU Manager (DCGM) for power consumption measurements





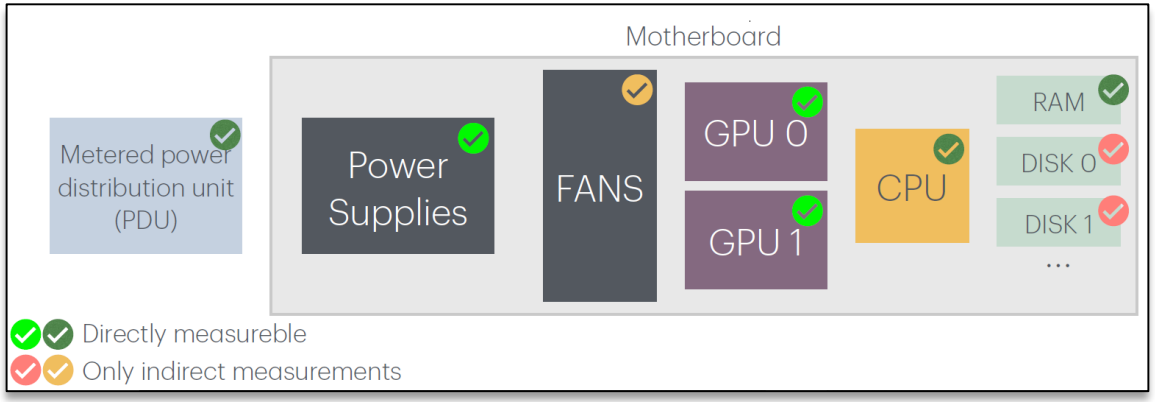


Impact of hardware on power consumption

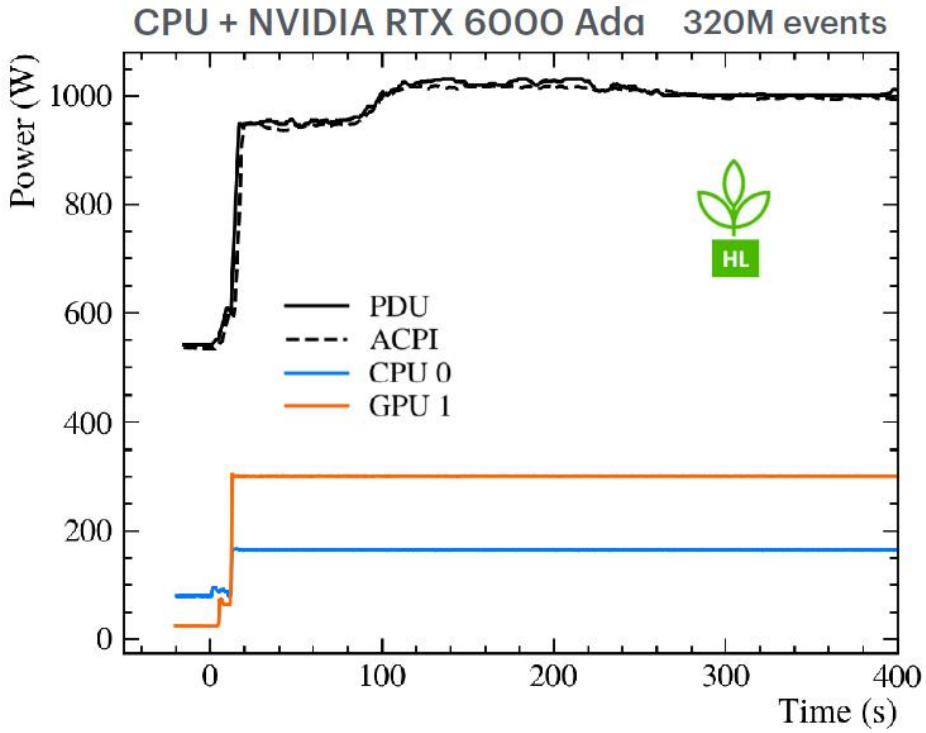


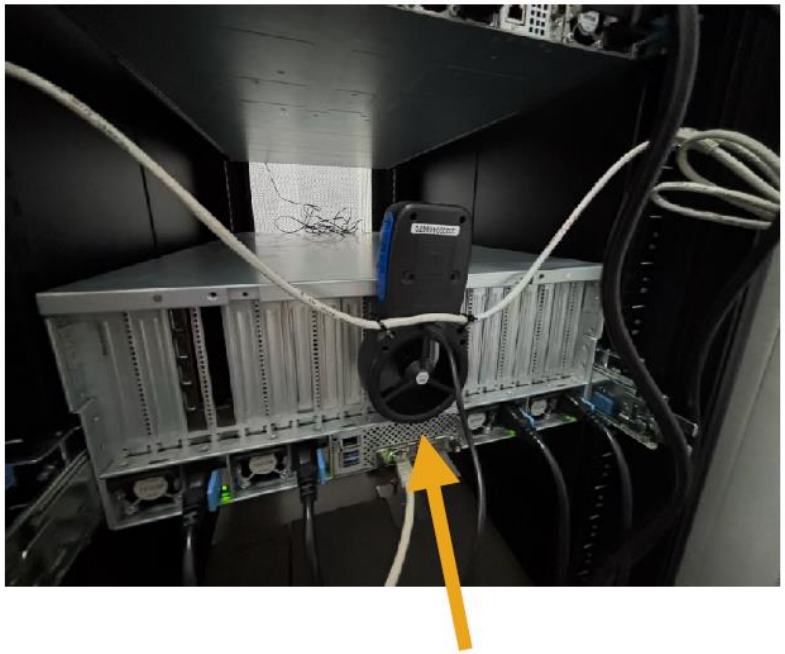
Main components have constant power consumption

What is the cause of the power consumption rise?

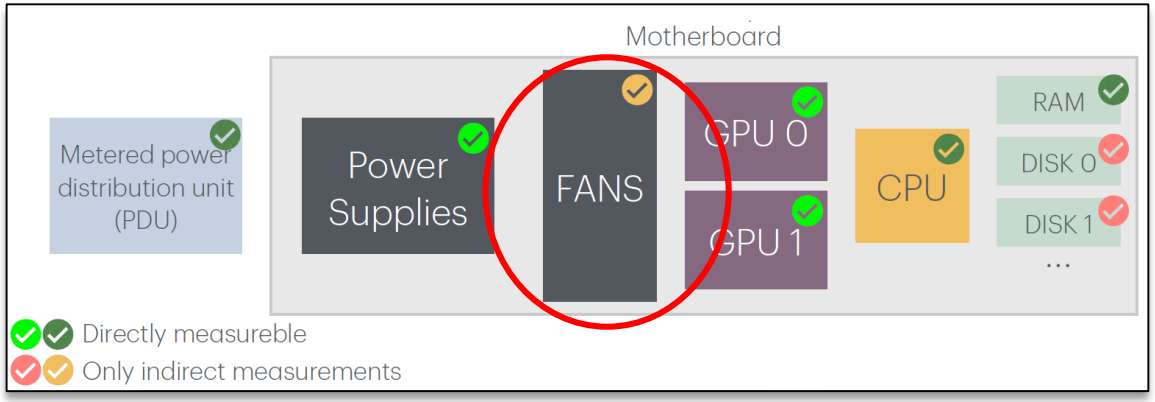


Impact of hardware on power consumption

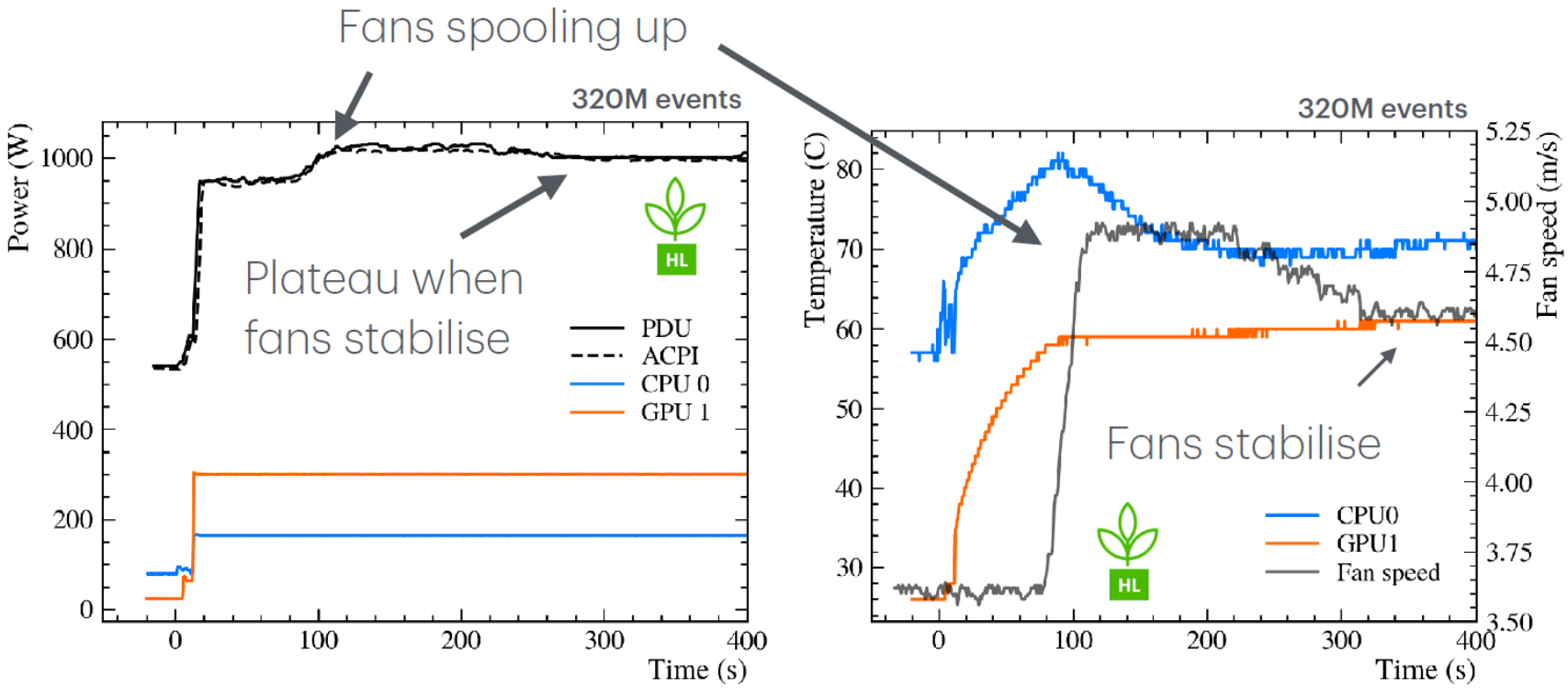




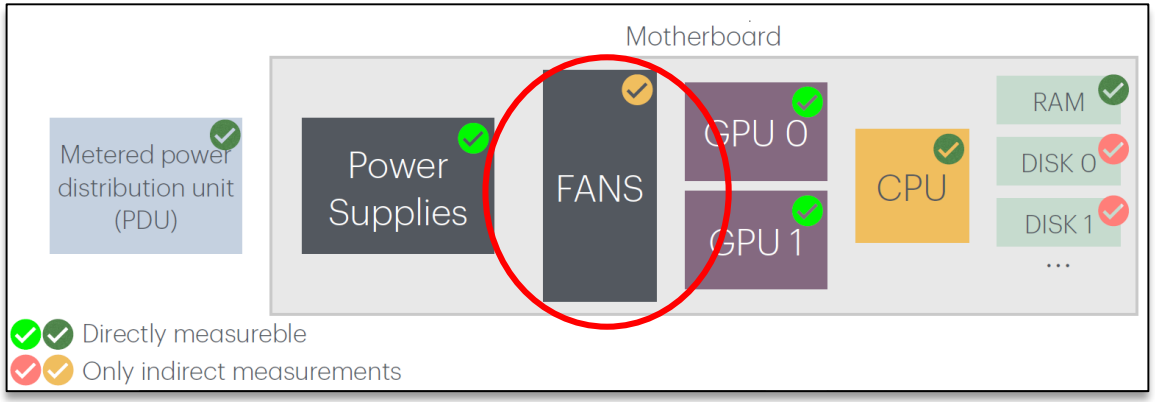
Fan speed measurement device



Impact of hardware on power consumption



Fans activation is the responsible of the increase in power consumption

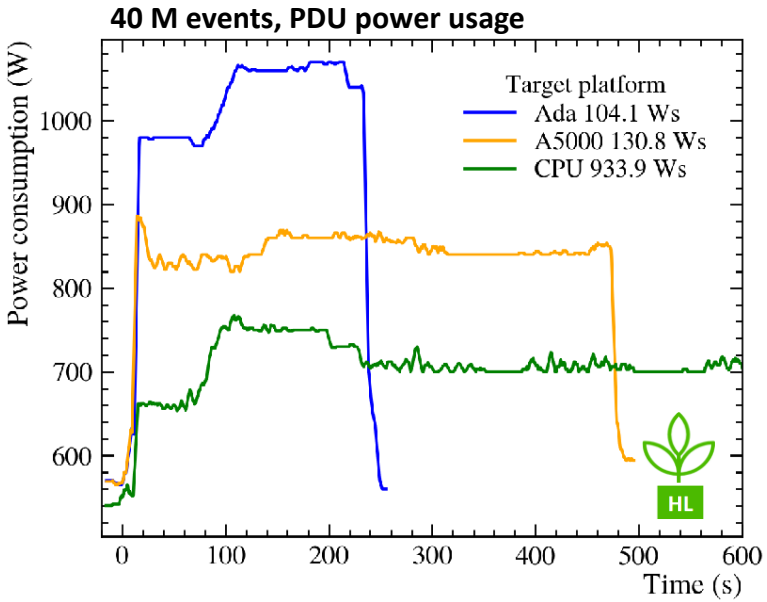


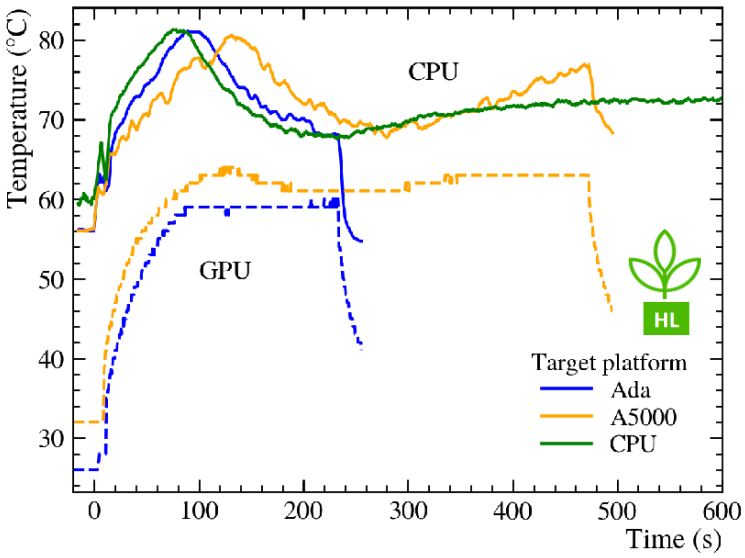
Impact of hardware on power consumption

The hardware option may affect the power consumption

Tests with CPU and two different GPUs

Intel(R) Xeon(R) Gold 5318Y
NVIDIA RTX A5000
NVIDIA RTX 6000 Ada

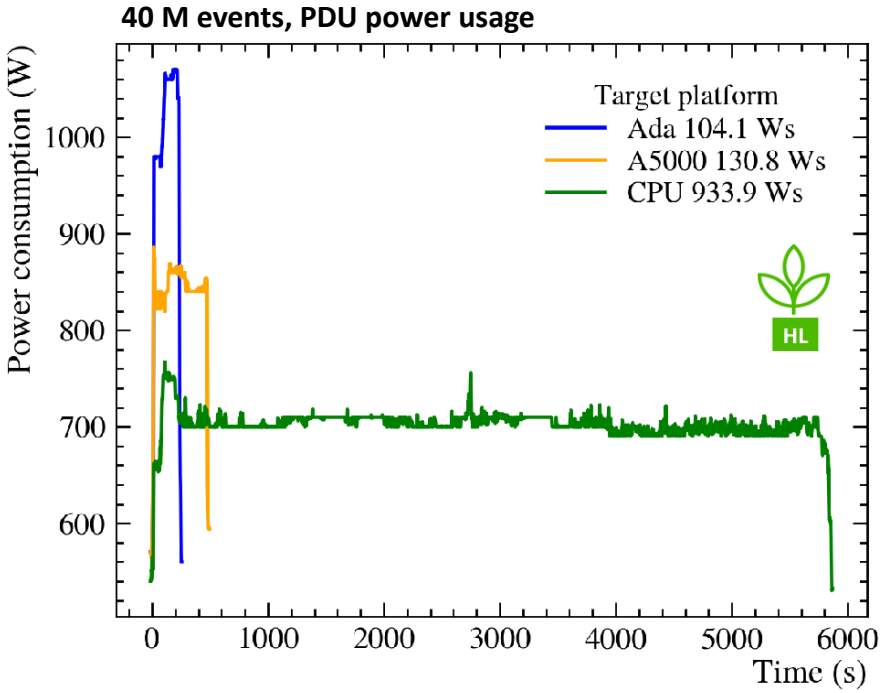


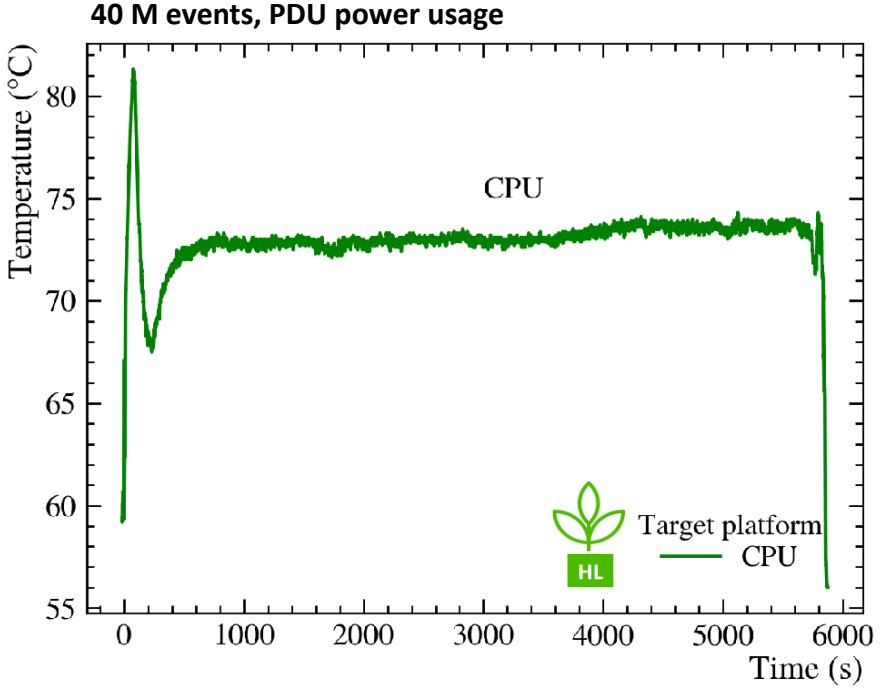


In general, more performant devices with faster execution time means less power consumption (inverse correlation with throughput)

Impact of hardware on power consumption

Allen is optimised for GPUs: execution on the CPU leads to low instantaneous power consumption but in total consumes a lot of energy due to a slow and long process

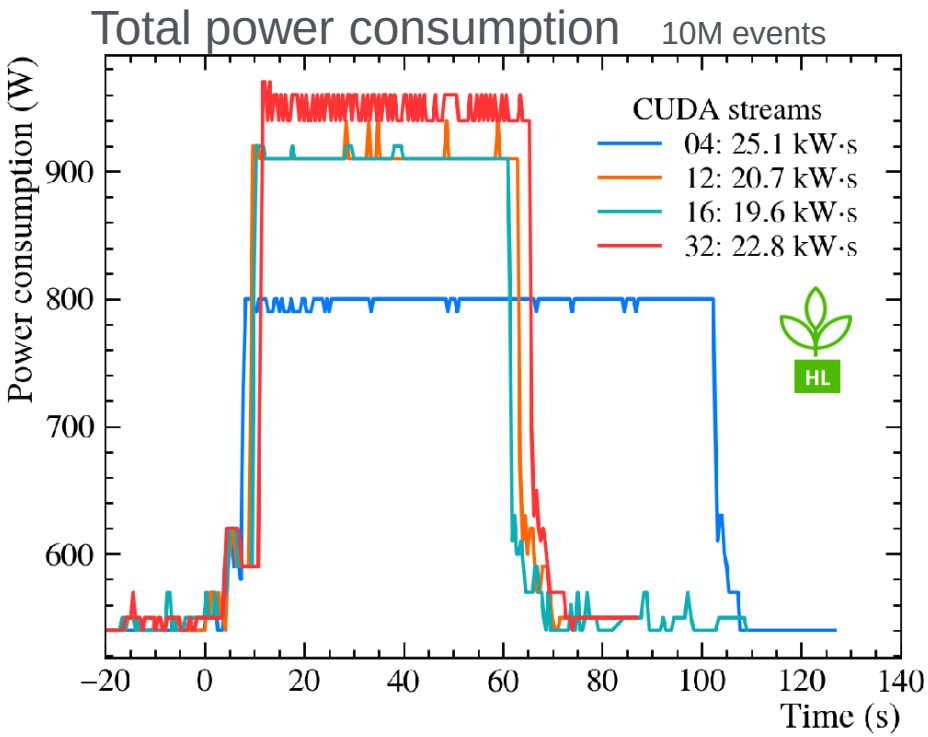


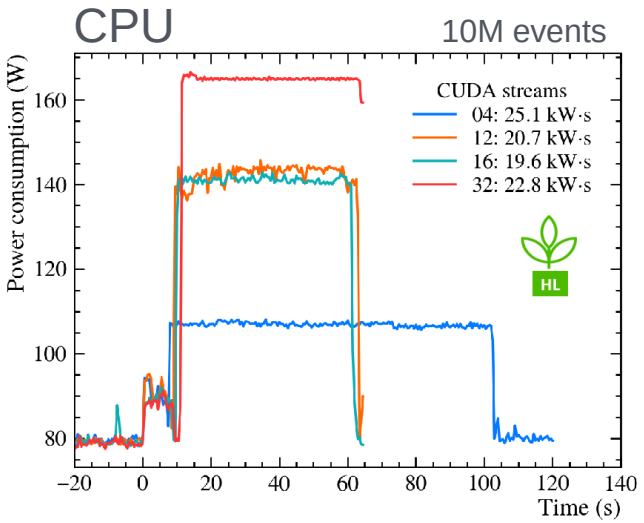


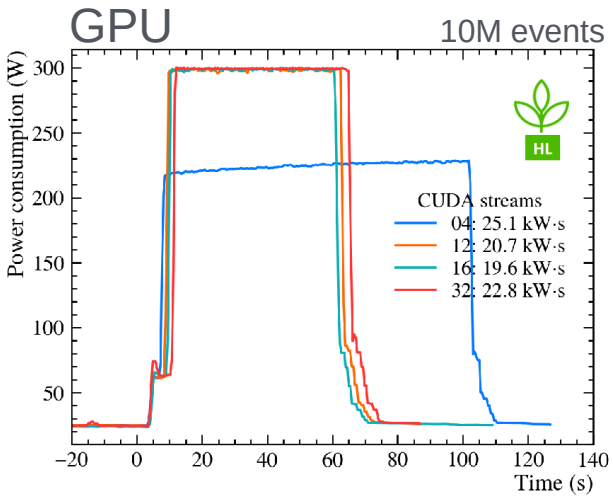
Hardware and software developments should work hand in hand!

Hardware utilisation

Dependence with the hardware utilization:
GPU parallelisation procedure (# CUDA streams)







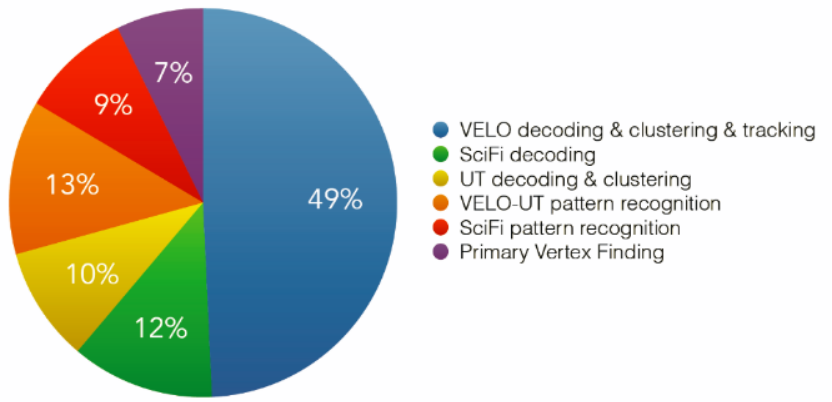
A non-proper use of the hosting hardware configuration leads to a slow execution, i.e., larger time → increase of the energy consumption

Impact of hardware on power consumption

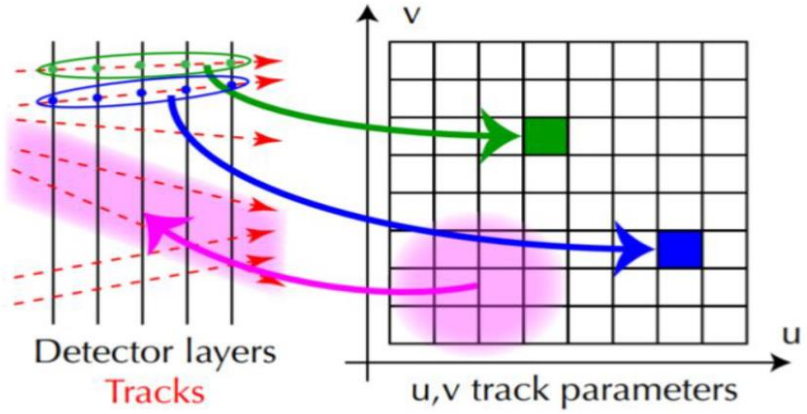
Checking other hardware architectures:

Offloading some tasks to FPGAs (RETINA framework – Pisa group) Stratix 10 (1SG280HN2)
[Framework TDR for the LHCb Upgrade II, CERN-LHCC-2021-012]

- Real-time reconstruction on FPGAs with the “artificial retina”architecture



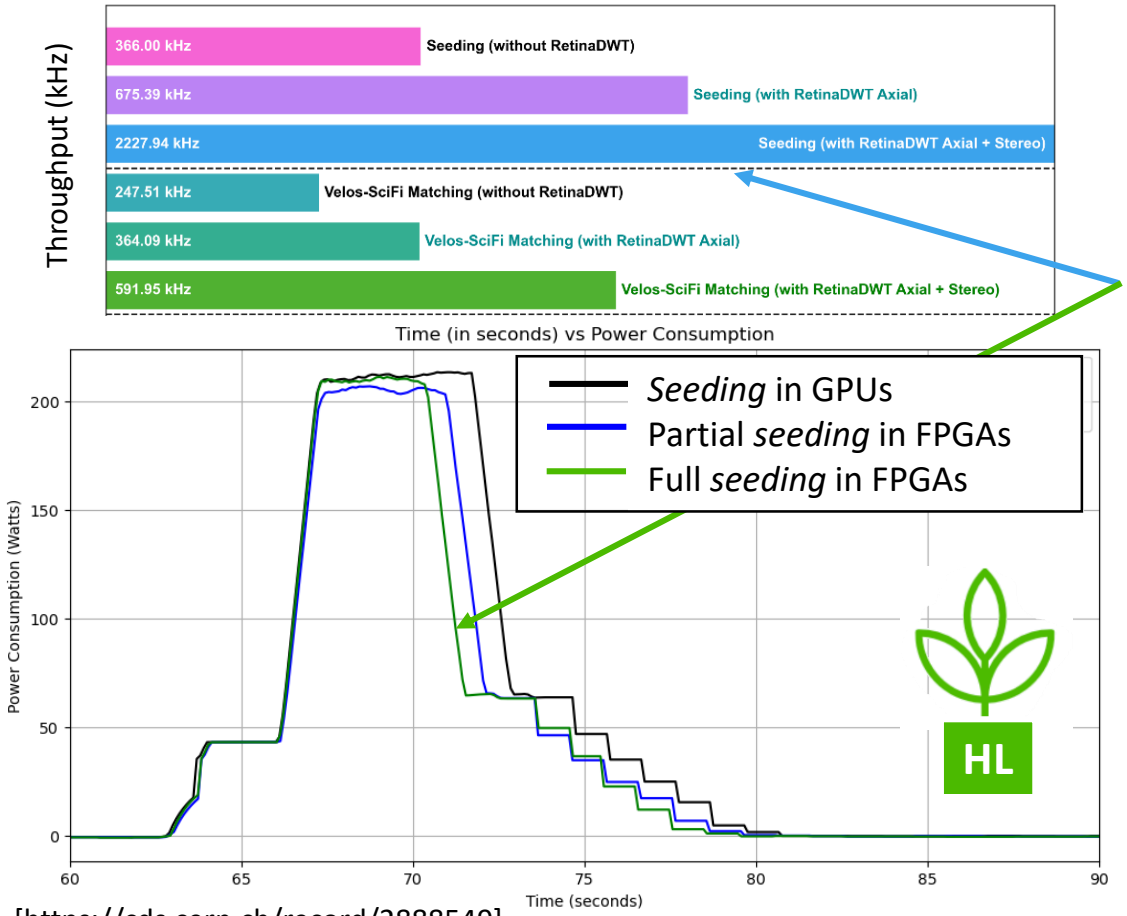
[NIMA 453 (2000) 425-429]



- Clustering of the VELO detector already implemented for Run3 in FPGAs !
- Tracking in development for Run5 (~2030), coprocessor testbed established at CERN for tests in realistic conditions

Impact of hardware on power consumption

Offloading some tasks to FPGAs



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

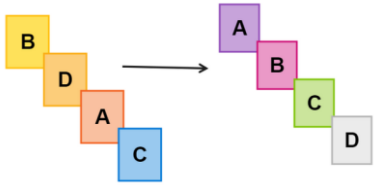
Seeding algorithm for making tracklets in the last LHCb tracker (SciFi) in FPGAs: throughput increases by 30%
→ Saving 6.2 mW·s/event

(for 30 MHz rate: 186kW/s)

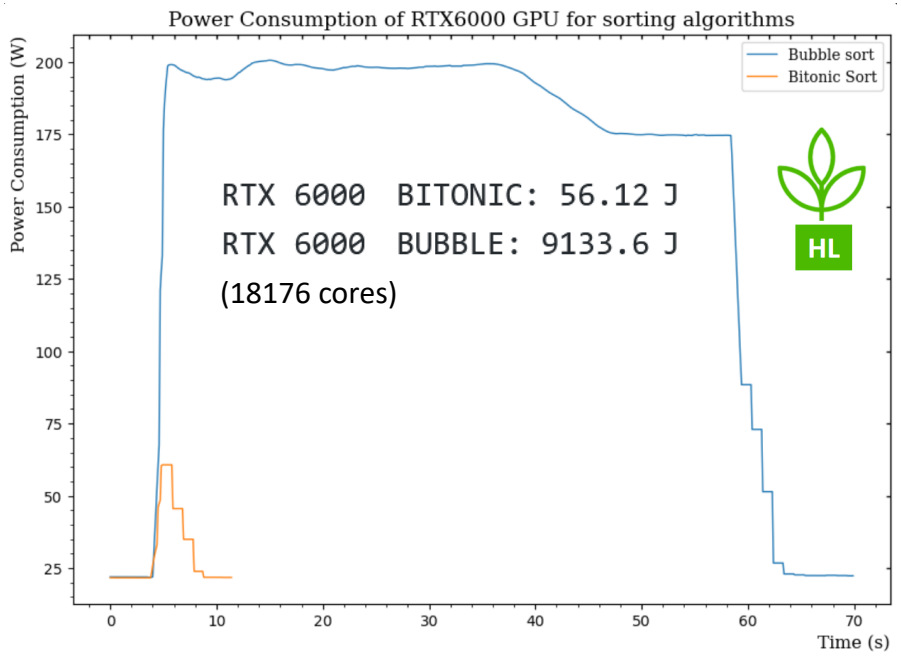
→ Use hybrid systems (FPGAs + GPUs) to take benefits of each one

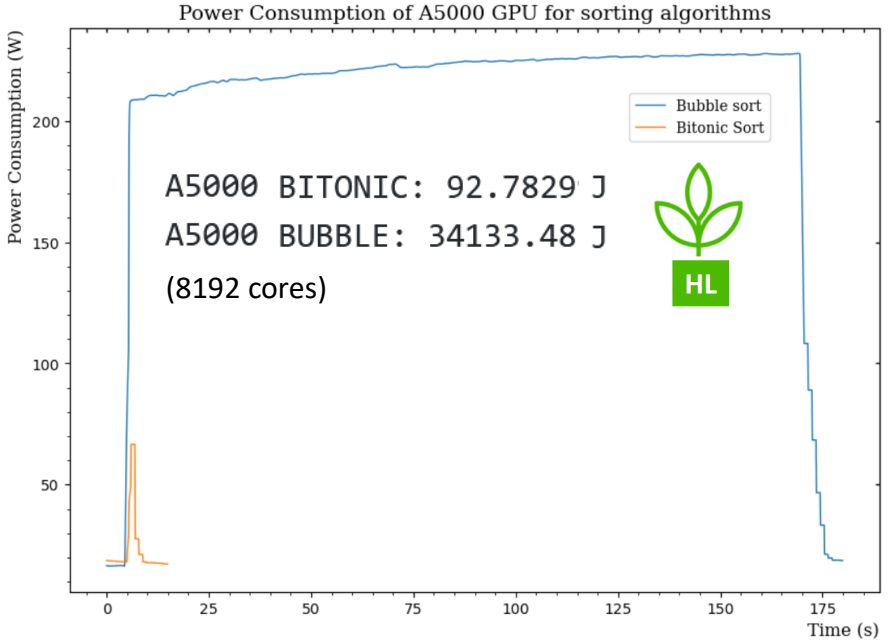
Software optimization

Dependence with the software: how well we program?



A basic known example:
Sorting 4M elements with ***Bubble*** and ***Bitonic*** algorithms:





→ Two orders of magnitude difference in energy consumption, with a high dependence of the hardware utilization!

Software optimization

A **Downstream** and vertexing algorithm for Long Lived Particles (LLP) selection at the first High level trigger (HLT1) of LHCb

V. Kholoimov[1], B. Kishor Jashal[1,2], A. Oyanguren[1], V. Svintozelskyi[1] and J. Zhuo[1]
[1]Instituto de Física Corpuscular (IFIC), University of Valencia- CSIC, Valencia, Spain.
[2]Rutherford Appleton Laboratory (RAL), Oxford, United Kingdom.

Abstract

A new algorithm has been developed at LHCb which is able to reconstruct and select very displaced vertices in real time at the first level of the trigger (HLT1). It makes use of the Upstream Tracker (UT) and the Scintillator Fiber detector (SciFi) of LHCb and it is executed on GPUs the Allen framework. In addition to an optimized strategy, it utilizes a Neural Network to increase the track efficiency and reduce the ghost rates, with very high budget. Besides serving to reconstruct K_S^0 and Λ particles from algorithm and the associated two-track vertexing could detecting long-lived particles during the Run 3.

2.4.5 Power consumption

The effect on the power consumption from the execution of **Downstream** algorithm in the HLT1 sequence is studied in the following and shown in Fig. 17. Several techniques are employed to measure the power consumption including the use of a metered power distribution unit (PDU[1]) within the rack, analysis of device driver outputs (e.g., Nvidia

[1]An APC PDU AP8858EU3 is used in this work.

https://arxiv.org/abs/2503.13092

Including power consumption as figure of merit in addition to efficiency, throughput and physics performance

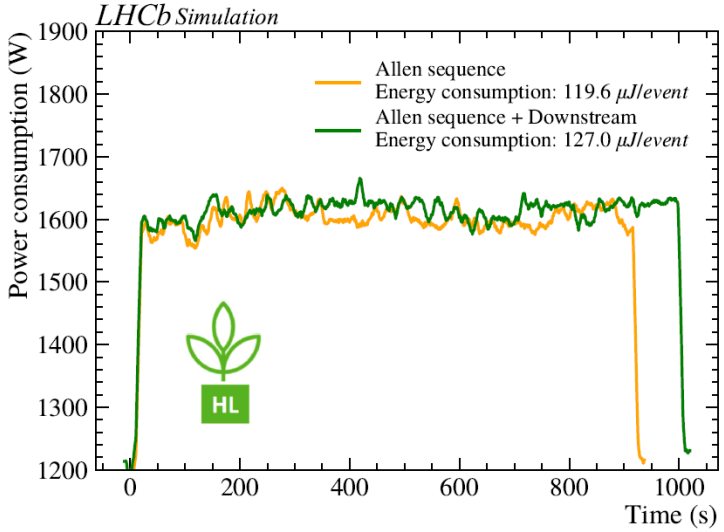


Fig. 17: Power consumption with **Allen** software running over 3.2M $B_s \to \phi\phi$ events without (blue) and with (orange) **Downstream** algorithm. The power consumption is measured using metered rack PDU AP8858EU3 with an average readout frequency of 2 Hz. The moving average filter with window of 20 points is applied. The measurements are obtained using the NVIDIA RTX 6000 Ada Generation GPU card.

Software optimization

In general: **HIGH THROUGPUT ↔ LOW POWER CONSUMPTION**

- Proper parallelization
- Memory usage
- Balanced work distribution
- Instruction-work organization

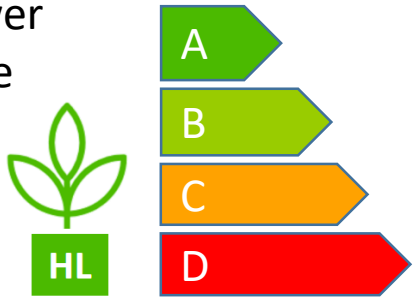
  ….

To take into account:

- *Performance vs energy*

Ex: which is the cost of 0.01% gain in efficiency for a tracking system/clustering algorithm? Per track? Per event? Per year? Vs final physics performance?

- Aiming to develop software tools to check the power consumption of an algorithm in a specific hardware



Conclusions & prospects

- Use the best and more efficient available hardware (vs €)
- Optimize the utilization of the hardware
- Optimize the software design
- Take advantages of the correlations among them
- Be open to hybrid systems
- Quantify the gain! (use of monitoring at software level)

Conclusions & prospects

- Power measurements vs temperature
- Fan speed reduction
- Other architectures (ARM)
- Evaluation of hybrid systems
- Automatized recipes
- Applications (ex: generators, porting MADGRAPH to FPGAs)