Energy efficiency of a GPU-based computing system in HEP

Jiahui Zhuo on behalf of the IFIC HIGH-LOW group

(IFIC, U.V. - CSIC)

Sustainable HEP 2026 — 5th Edition (8-10 July 2026)

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

Outline

- The HIGH-LOW project at IFIC at Valencia
- LHCb experiment
- Energy efficiency of a GPU system
- Summary

The HIGH-LOW project at Valencia

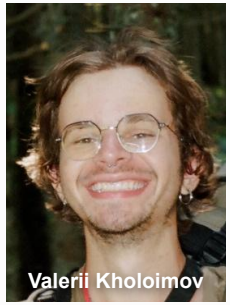
Design of HIGH performance algorithms for LOW power sustainable hardware for LHC experiments and their upgrades

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


Arantza Oyanguren


Jiahui Zhuo


Álvaro Fernández

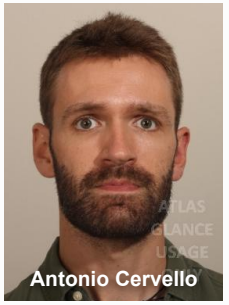

Valerii Kholoimov


Miriam Lucio


Miranda Carou


Luca Fiorini


Alberto Valero


Antonio Cervello


Rui Wang


Héctor Gutiérrez

Francesco Curcio

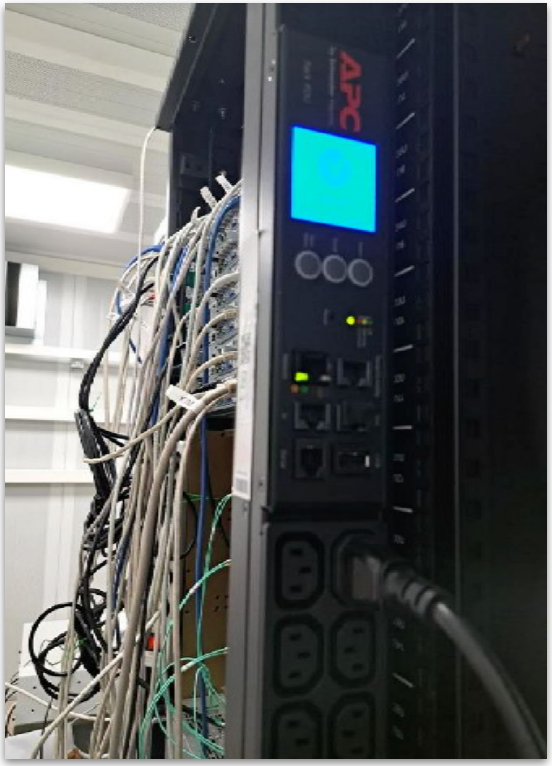
The HIGH-LOW project at Valencia

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

HL

Quantifying energy consumption is important to take decisions!

The HIGH-LOW project at Valencia



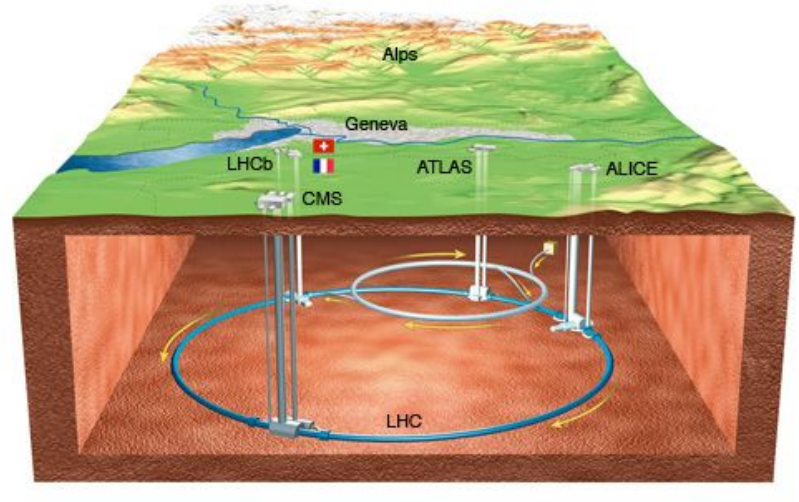
HL01
2 x Intel Xeon Gold 5318Y,
256GB DDR4 RAM,
80TB storage,
NVIDIA RTX A5000 x1,
NVIDIA RTX A6000 Ada x1

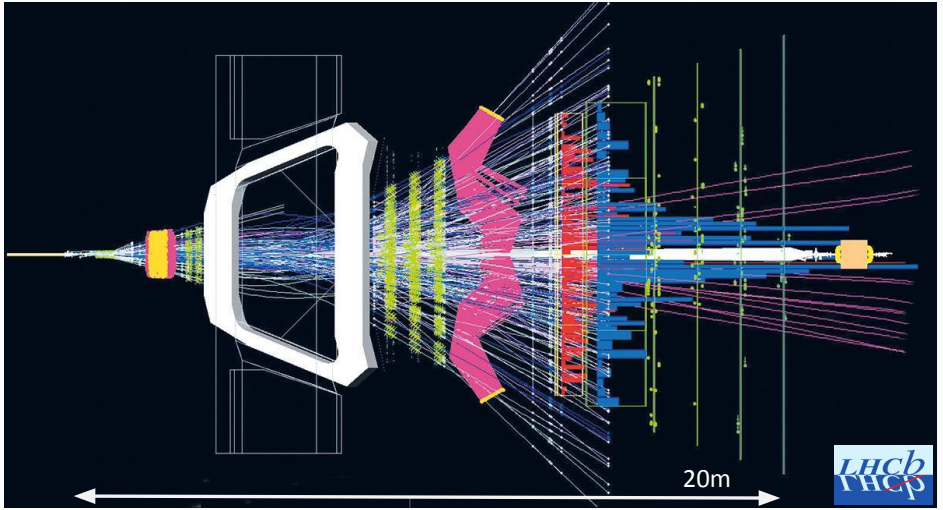
HL02
AMD EPYC 9474F,
768GB DDR5 RAM,
160TB storage,
NVIDIA RTX A6000 Ada x1,
NVIDIA H100 NVL x2

AMPERE ALTRA Q64-22
(ARM Neoverse N1)

APC Metered Rack PDU ZeroU 2G AP8

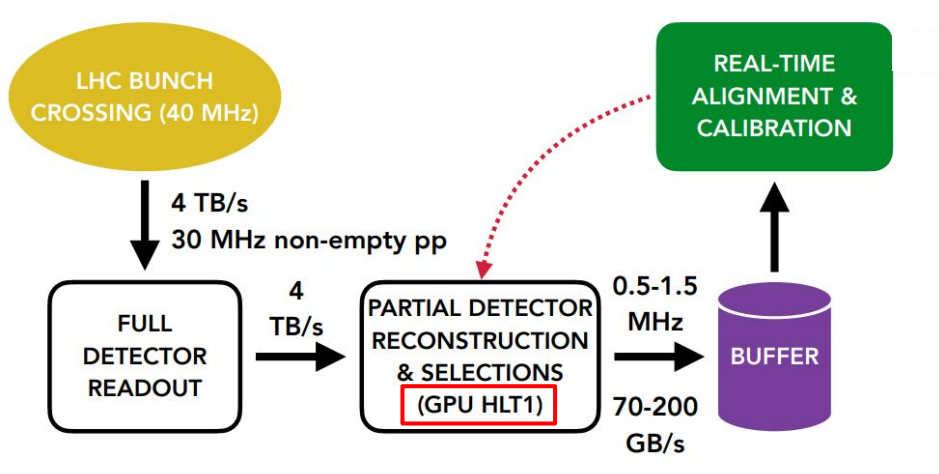
LHCb experiment

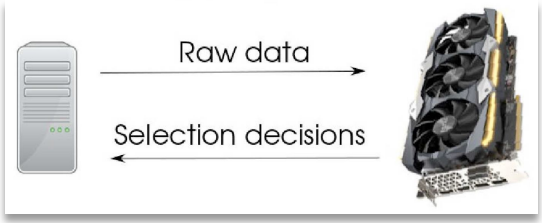




	LHCb	ATLAS	CMS	ALICE
$\mathcal{L}\ [cm^{-2}s^{-1}]$	2×10^{33}	2×10^{34}	2×10^{34}	6×10^{29}
pile-up	5	60	60	1
reconstruction rate	30 MHz	100 kHz	100 kHz	50 kHz
reconstructed tracks/s	1800 M	90 M	90 M	10 M

LHCb experiment





- First LHC experiment that uses GPU-based system as first level trigger
- About 300 algorithms
- ~500 NVIDIA RTX A5000

Allen

LHCb / Allen · GitLab

Energy efficiency of a GPU system

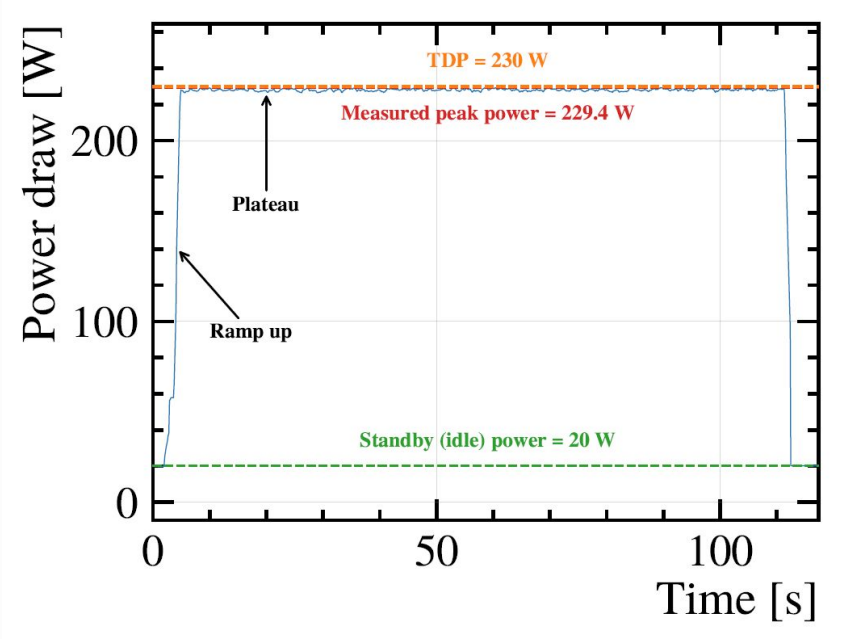
LHCb HLT1 is GPU-bound

With the same GPU, CPU choice does not affect performance.

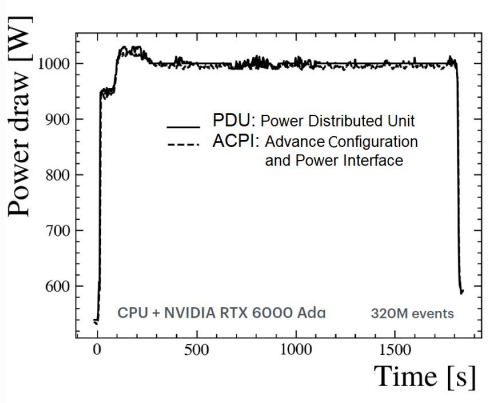
	GPU	Architecture	CUDA cores	Boost clock [MHz]	Memory BW [GB/s]	TDP [W]
Samsung 8nm	RTX A5000	Ampere	8192	1695	768	230
	RTX 3090	Ampere	10496	1695	936	390
TSMC 4nm	RTX 4070 Ti Super	Ada Lovelace	8448	2610	672	285
	RTX 4080 Super	Ada Lovelace	10240	2550	736	320
	RTX 5000 Ada	Ada Lovelace	12800	2550	576	250
	RTX 4090	Ada Lovelace	16384	2520	1008	450
	RTX 6000 Ada	Ada Lovelace	18176	2505	960	300
	RTX PRO 4000	Blackwell	8960	2617	672	145
	RTX PRO 6000	Blackwell	24064	2617	1792	600
	H100 NVL	Hopper	16896	1785	3938	400

Energy efficiency of a GPU system

Measurement strategy



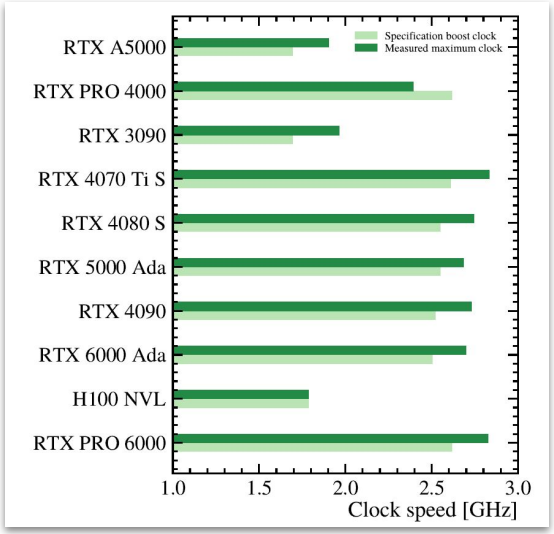
- Use `nvidia-smi` to monitor GPU parameters.
- Measure **power consumption**, **CUDA core clock speed**, and **memory frequency** at the plateau.



The power measurement is later **double-checked** with **PDU** (external) and **ACPI** (motherboard)

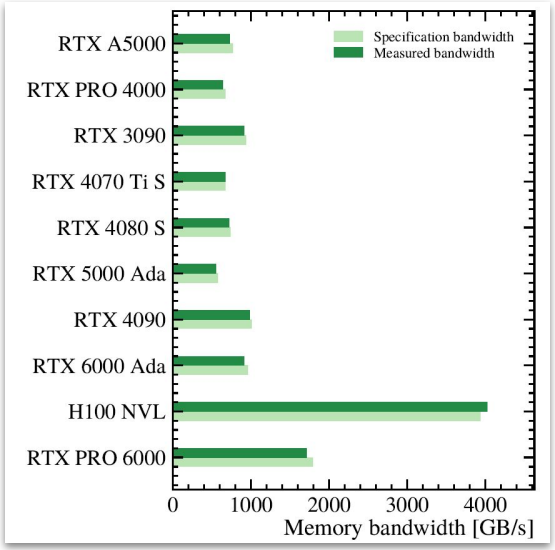
Energy efficiency of a GPU system

Clock speed



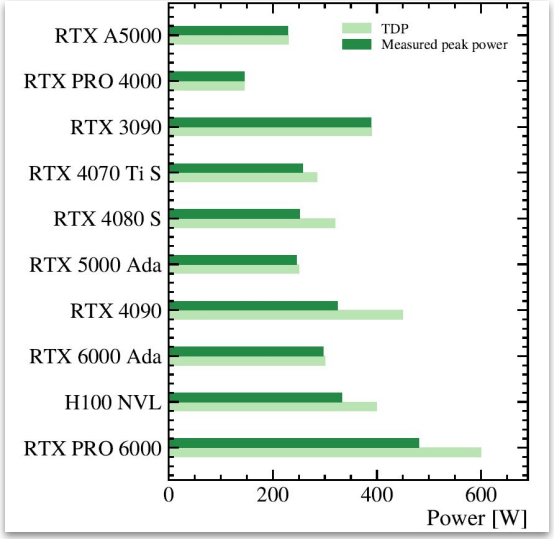
Run **5–16% above** specification (GPU boost mechanism)

Memory bandwidth



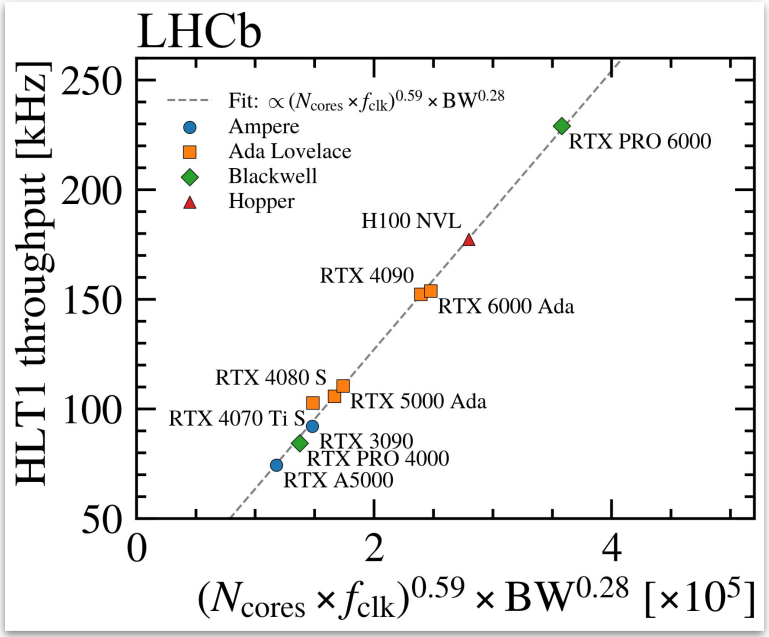
Good agreement

Power

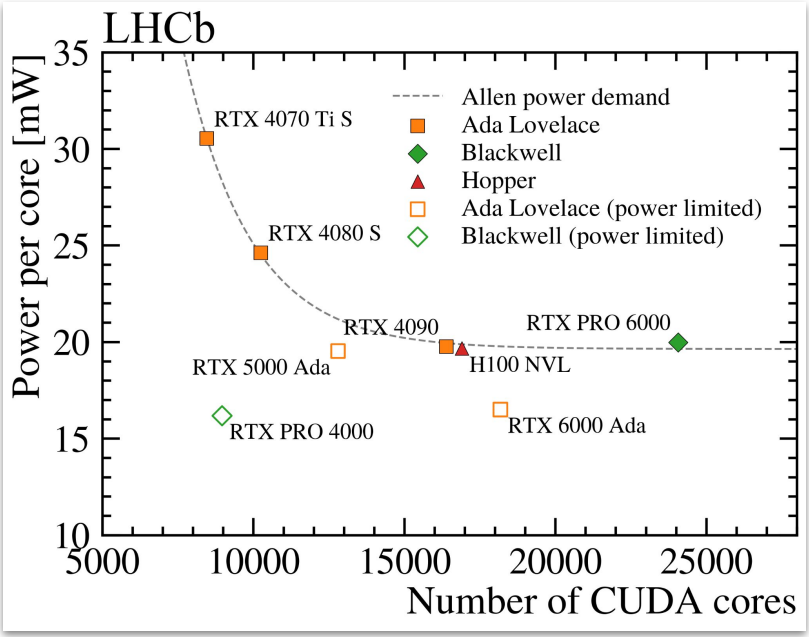


Power-limited GPU:
~100% TDP
Non-power-limited GPU:
~71-90% below TDP

Energy efficiency of a GPU system



$$\text{Throughput} = \text{k} \times (\text{N}_{\text{cores}} \times \text{f}_{\text{clock}})^{\text{a}} \times \text{Bandwidth}^{\text{b}}$$



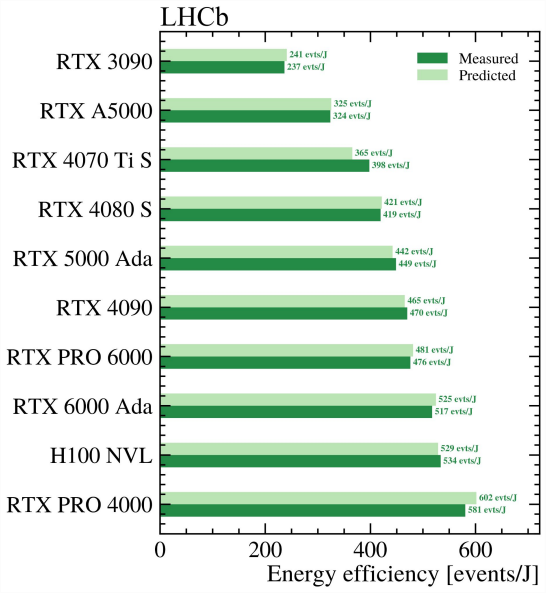
$$\text{Power} = \min(\text{Power}_{\text{cores}} \times \text{N}_{\text{cores}}, \text{TDP})$$

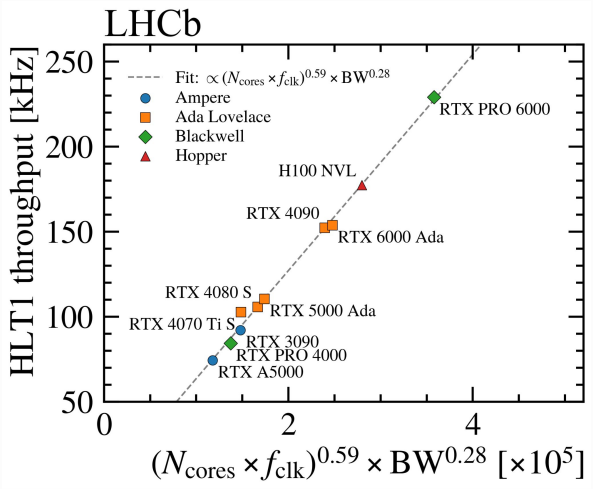
$$\text{P}_{\text{cores}} = \text{P}_{\text{floor}} + \text{A} \times \exp(-\frac{\text{N}_{\text{cores}}}{\tau})$$

Energy efficiency of a GPU system

$$\text{Energy Efficiency} = \frac{\text{Throughput}}{\text{Power consumption}} = \frac{\#\ \text{Event}}{\text{Joule}}$$

Our model can predict power efficiency from GPU specifications.





- Fastest GPU ≠ most energy-efficient.
- Being power-limited does not determine energy efficiency.
- Energy efficiency is an independent metric.
- Similar throughput can imply very different energy efficiency.

Summary

- High-Low project at IFIC (Valencia) for sustainability studies at HL-LHC.
- Energy efficiency is an independent metric.
- The model can predict power consumption from GPU specification.
- The methodology is not specific to LHCb HLT1: it applies to any GPU-based application.
- The related paper is on arXiv, the journal publication is in progress.

Energy efficiency of a GPU-based computing system for High Energy Physics experiments

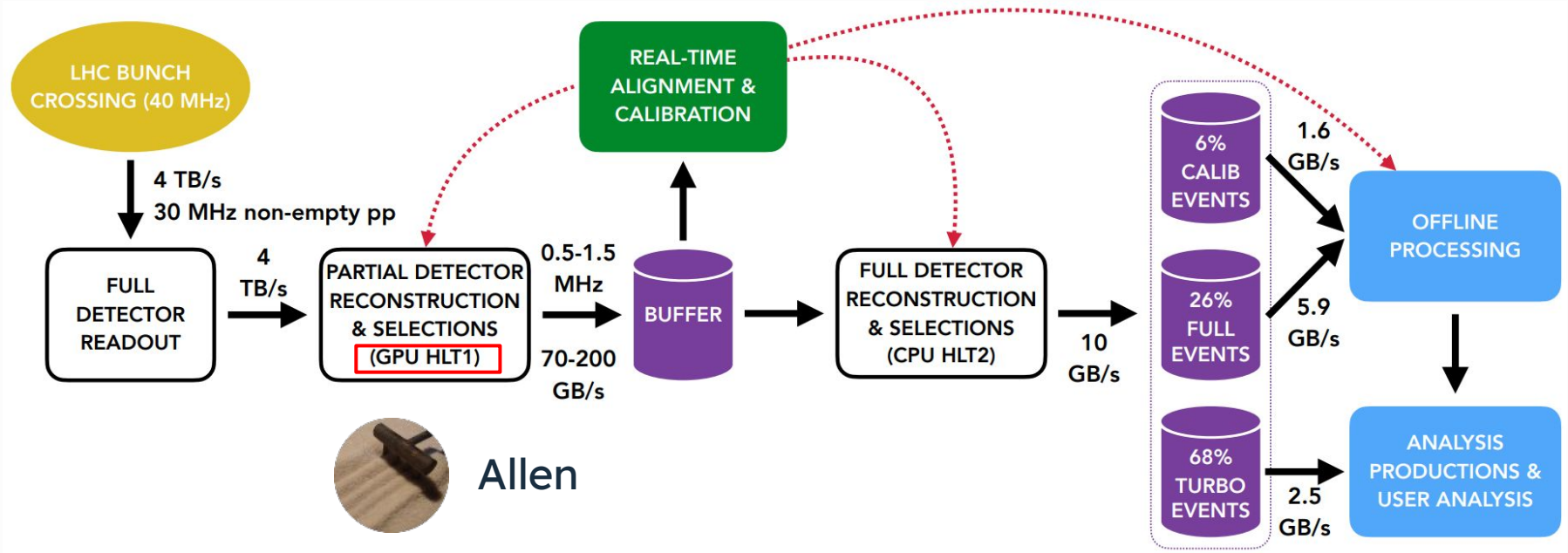
Jiahui Zhuo, Arantza Oyanguren, Álvaro Fernández Casani, Luca Fiorini, Valerii Kholoimov

In this paper we introduce the energy efficiency as a new metric for evaluating both hardware platforms based on Graphic Processor Units (GPU), and algorithm optimisations at High Energy Physics (HEP) experiments. We develop a method to compute the energy efficiency for the case of the first high level trigger (HLT1) of the LHCb experiment, relating the throughput with GPU specifications such as the number of cores, clock frequency, memory bandwidth and thermal design power. The model can be extended to other HEP experiments to make decisions and reach sustainable computing ecosystems.

Comments: Submitting to Frontiers in Physics - High-Energy and Astroparticle Physics
Subjects: **High Energy Physics - Experiment (hep-ex)**
Cite as: arXiv:2604.27523 **[hep-ex]**
(or arXiv:2604.27523v1 **[hep-ex]** for this version)
https://doi.org/10.48550/arXiv.2604.27523

Backup

Introduction



We want to study the throughput and energy efficiency of Allen (HLT1)

Study setup

Allen

```
Version: V7r10p1
Sequence: hlt1_pp_default
Input: MEP_2025_pp_pD2_bu_321834_LHCb_ECEB01_BU_0
```

	GPU	Architecture	CUDA cores	Boost clock [MHz]	Memory BW [GB/s]	TDP [W]
Samsung 8nm	RTX A5000	Ampere	8192	1695	768	230
	RTX 3090	Ampere	10496	1695	936	390
TSMC 4nm	RTX 4070 Ti Super	Ada Lovelace	8448	2610	672	285
	RTX 4080 Super	Ada Lovelace	10240	2550	736	320
	RTX 5000 Ada	Ada Lovelace	12800	2550	576	250
	RTX 4090	Ada Lovelace	16384	2520	1008	450
	RTX 6000 Ada	Ada Lovelace	18176	2505	960	300
	RTX PRO 4000	Blackwell	8960	2617	672	145
	RTX PRO 6000	Blackwell	24064	2617	1792	600
	H100 NVL	Hopper	16896	1785	3938	400

Study setup

Version: V7p10p1

The HIGH-LOW project at Valencia

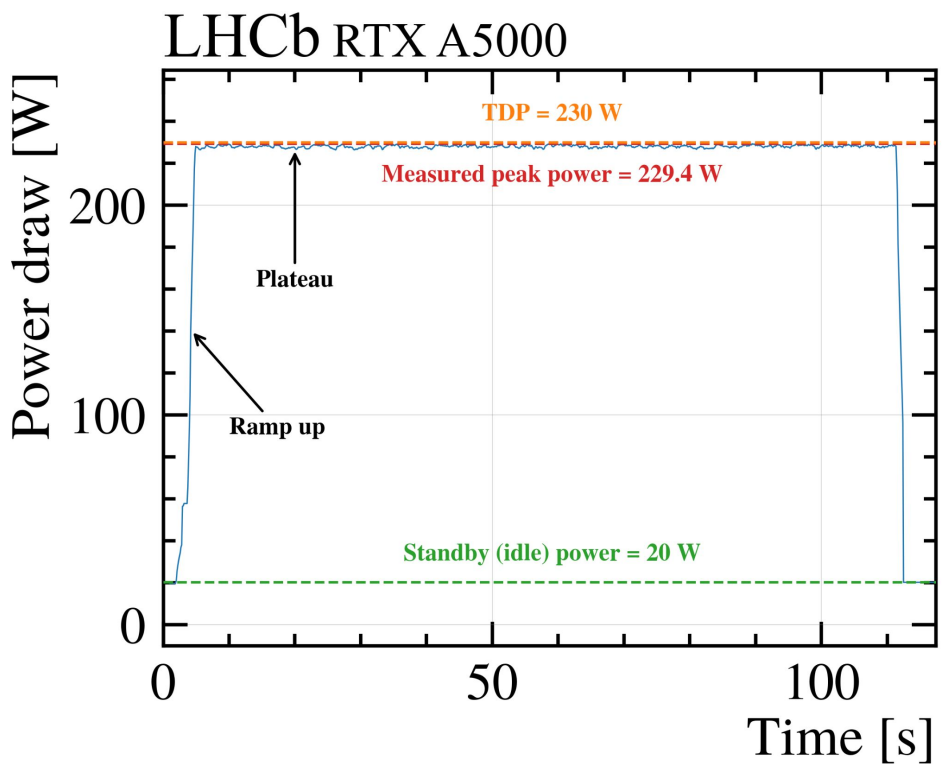
Design of HIGH performance algorithms for LOW power sustainable hardware for LHC experiments and their upgrades

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

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

RTX PRO 6000	Blackwell	24064	2617	1792	600
H100 NVL	Hopper	16896	1785	3938	400

Measurement

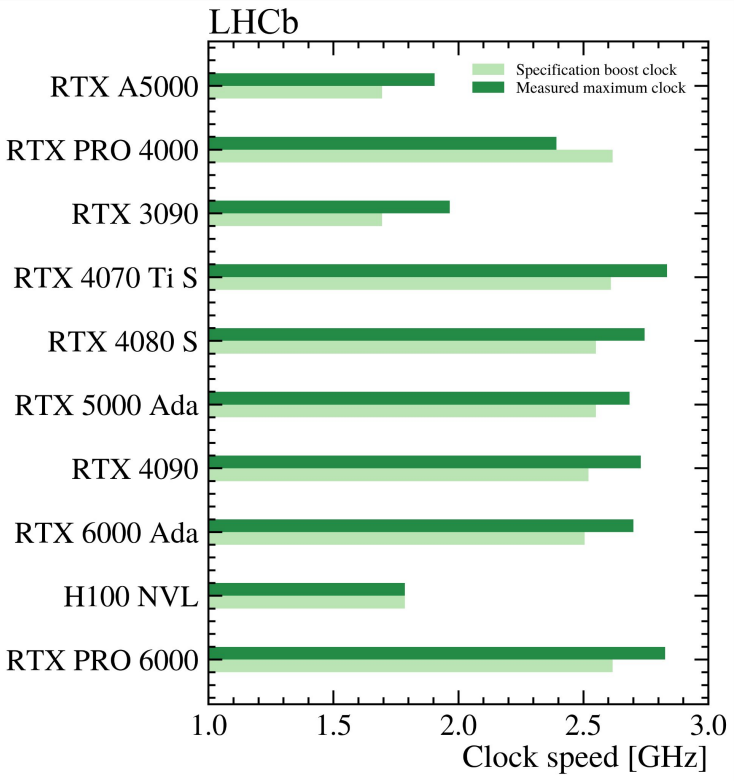


Measurement strategy

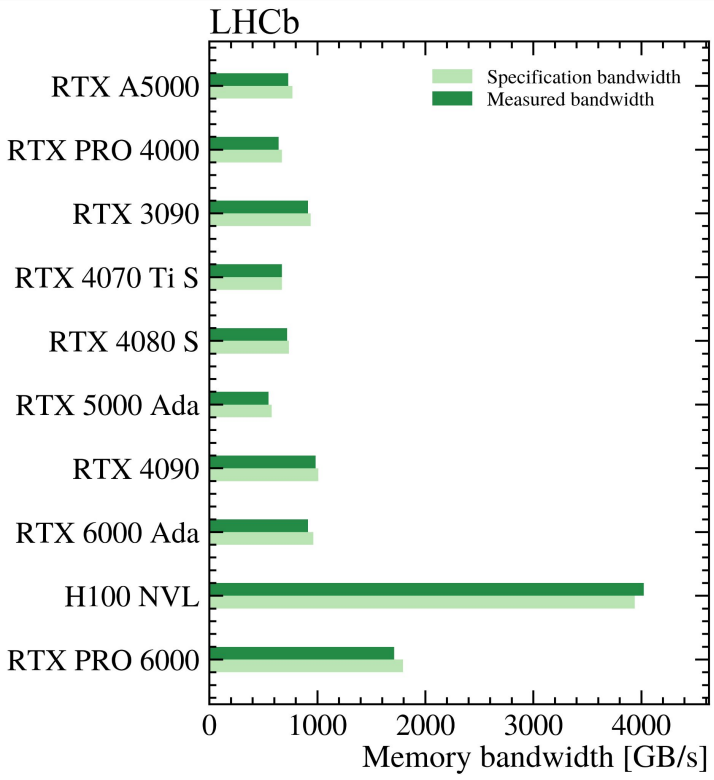
- Use `nvidia-smi` to monitor GPU parameters while running Allen.
- Measure the **power consumption**, **CUDA core clock speed**, and **memory frequency** at the plateau.

Measurement

- Most GPUs run **5–16% above** their specification boost clock.
- The specification boost clock is a **guaranteed minimum**, **not a maximum**.
- Modern NVIDIA GPUs automatically increase their clock frequency when there is **enough thermal** and **power margin**, a behaviour known as GPU Boost.



Measurement



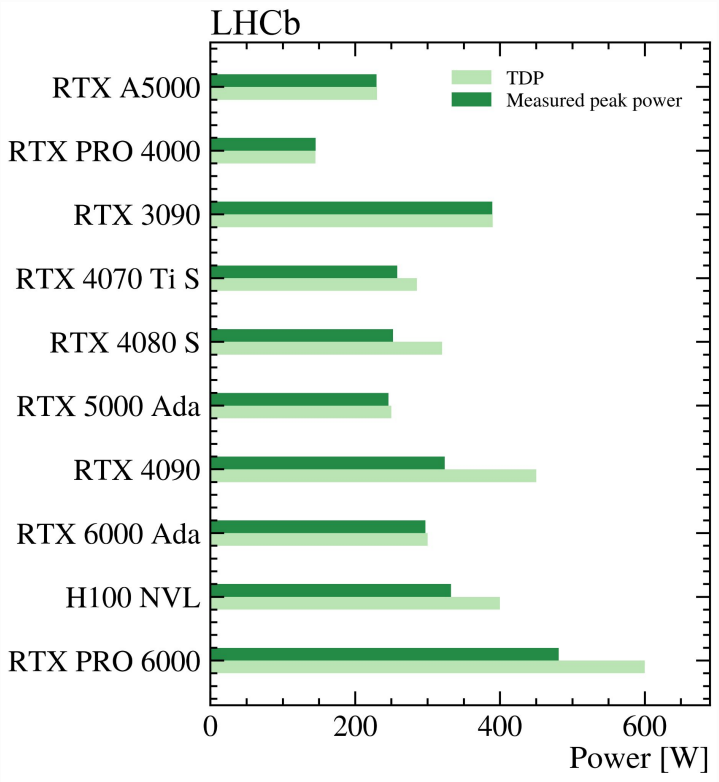
- Memory bandwidth is

  Memory frequency ✖ Memory bus width

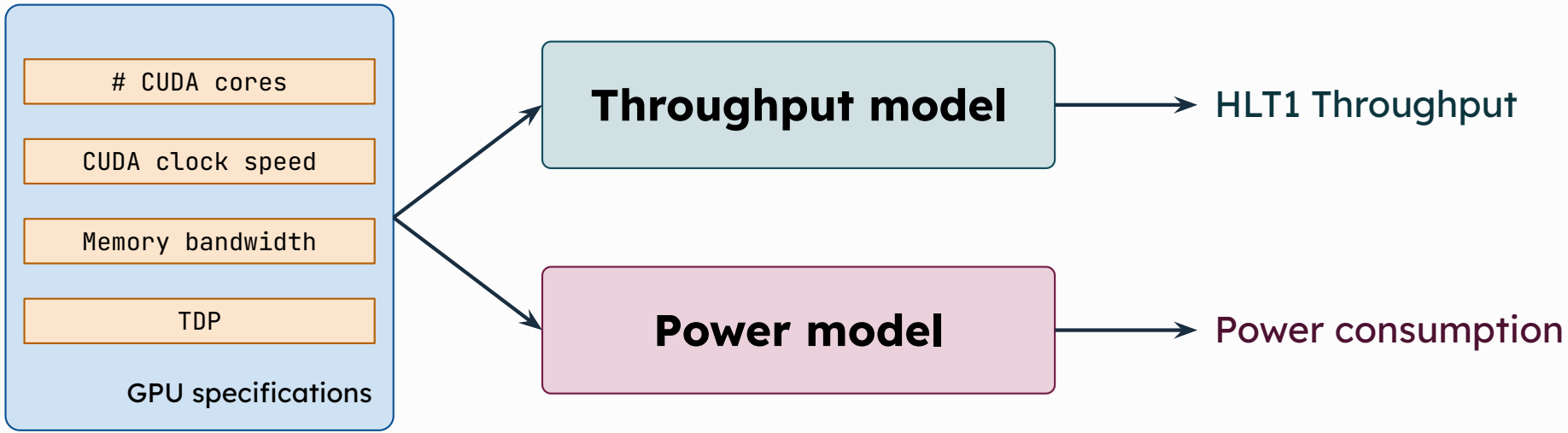
- The memory frequency is measured at the plateau, while the bus width is taken from the GPU specification.
- The measured memory bandwidth closely **matches** the specification for **all GPUs**.
- The memory clock does not have a **boost mechanism**.

Measurement

- Some GPUs reach very **close to their TDP** (99–100%), meaning the HLT1 workload **consumes the entire power budget**.
- These GPUs are referred to as **power-limited**.
- Other GPUs run **well below their TDP (72–91%)**, meaning the power budget is **not** the **bottleneck**.
- These GPUs are referred to as **non-power-limited**.



What do we want?

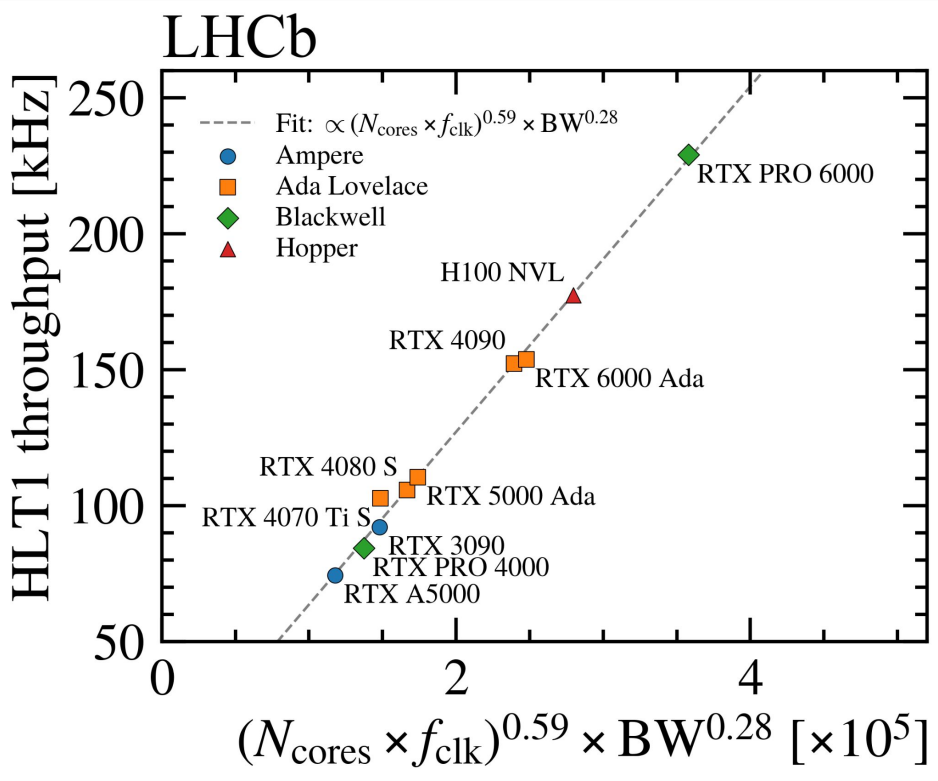


Throughput model

$$\text{Throughput} = k \times (N_{\text{cores}} \times f_{\text{clock}})^{a} \times \text{Bandwidth}^{b}$$

- **k** is a normalization constant.
- **a** is the exponent on the compute capacity **(N_{cores} × f_{clock})**.
- **b** is the exponent on the memory **bandwidth**.
- **Synchronisation overhead** and **limited parallelism** lead to **a < 1**.
- **Cache hits** reduce reliance on global memory bandwidth, leading to **b < 1**.

Throughput model



$$a = 0.59\,, \qquad b = 0.28\,, \qquad k = 0.64\,,$$

- **a > b**: HLT1 is more **compute-bound** than **memory-bound**.
- **a = 0.59**: doubling the compute capacity increases the throughput by a factor of $2^{0.59} \approx$ **1.5**, not 2.
- The model is fitted with **measured parameters** instead of specifications, since the measured clock speed may differ from the specification.

Power model

$$\mathrm{Power} = \min(\mathrm{Power}_{\mathrm{cores}} \times \mathrm{N}_{\mathrm{cores}}, \mathrm{TDP})$$

- **$\mathrm{Power}_{\mathrm{cores}}$** is the power demand per CUDA core in HLT1.
- **$\mathrm{N}_{\mathrm{cores}}$** is the number of CUDA cores.
- **$\mathrm{Power}_{\mathrm{cores}}$** is a function of **$\mathrm{N}_{\mathrm{cores}}$**: as more cores are available, the HLT1 workload is spread more thinly and each core consumes less power.

$$\mathrm{Power}_{\mathrm{cores}} = \mathrm{Power}_{\mathrm{cores}}(\mathrm{N}_{\mathrm{cores}})$$

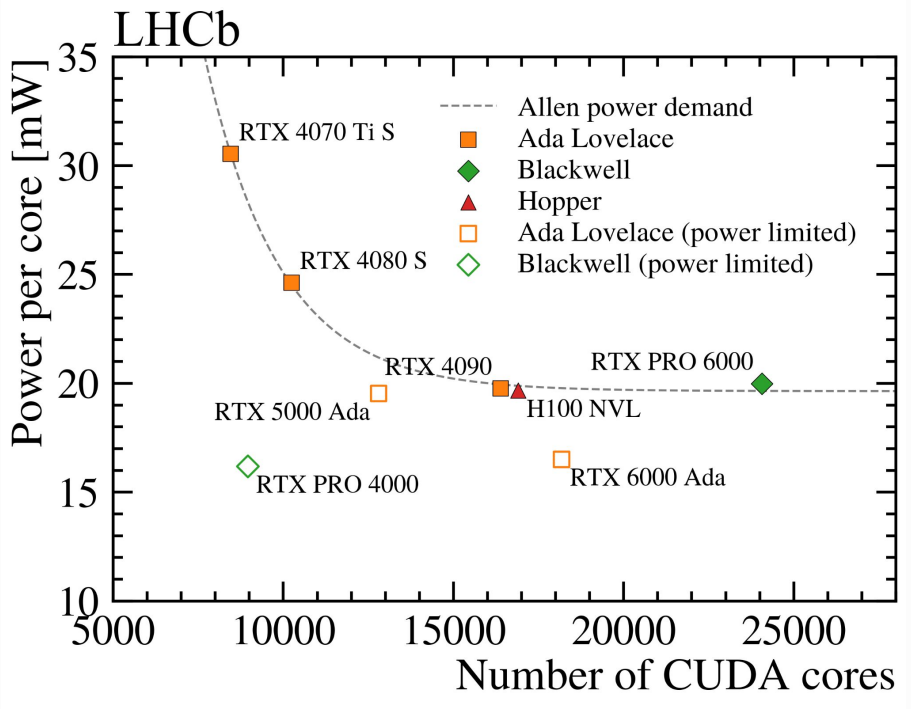
- The power demand per CUDA core in HLT1 can only be measured for **non-power-limited** GPUs.
- It also depends on the **fabrication process**. Here we use **TSMC 4nm** to model the power consumption.

Power model

$$P_{cores} = P_{floor} + A \times \exp(-\frac{N_{cores}}{\tau})$$

$$P_{floor} = 19.6\,\text{mW}, \quad A = 489\,\text{mW}, \quad \tau = 2223\ \text{cores}$$

- At very high N_{cores}, HLT1 parallelisation reaches a limit: each core already handles the smallest unit of work that cannot be subdivided further.
- The power corresponding to this limit is P_{floor}: the per-core power cost of HLT1 when parallelisation is fully exploited on the **TSMC 4 nm** process.
- **Power-limited GPUs** do not have enough P_{cores} to meet the HLT1 demand, so they always sit below the curve.



Energy efficiency

Throughput model

$$\text{Throughput} = k \times (N_{cores} \times f_{clock})^{a} \times \text{Bandwidth}^{b}$$

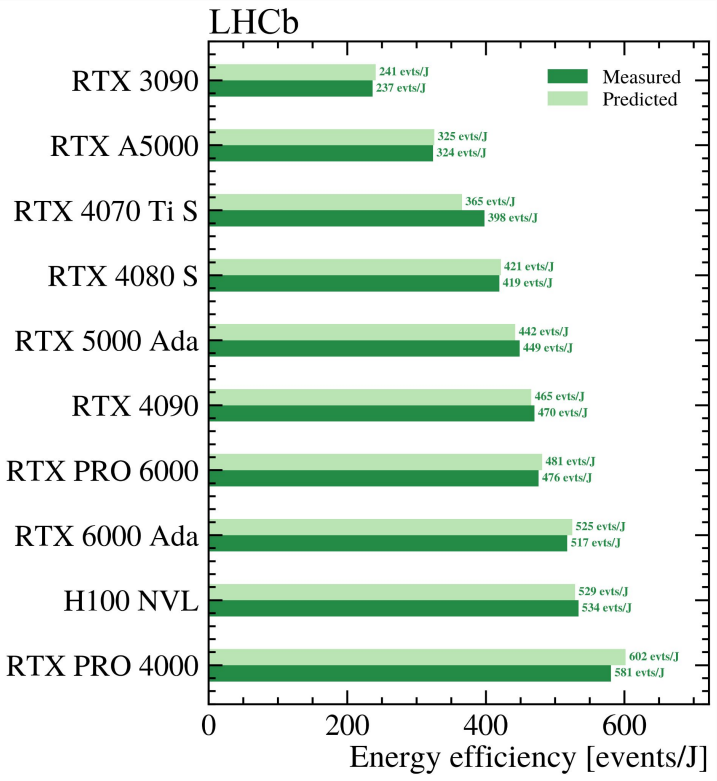
$$\text{Energy Efficiency} = \frac{\text{Throughput}}{\text{Power consumption}} = \frac{\#\ \text{Event}}{\text{Joule}}$$

Power model

$$\text{Power} = \min(\text{Power}_{cores} \times N_{cores}, \text{TDP})$$

$$P_{cores} = P_{floor} + A \times \exp(-\frac{N_{cores}}{\tau})$$

Energy efficiency



- **The fastest GPU is not necessarily the most energy-efficient**: the RTX PRO 6000 has the highest throughput, but is less power-efficient than several other GPUs.
- **Being power-limited does not determine power efficiency**: the RTX PRO 4000 is the most power-efficient GPU despite being power-limited, while the RTX 5000 Ada is also power-limited but less efficient than many others.
- **Energy efficiency is an independent metric** and cannot be simply predicted from throughput.
- **Our model can predict power efficiency from GPU specifications**. Actual performance may differ from the specification due to GPU Boost, but the difference is not large enough to change the energy-efficiency ranking.

Energy efficiency