



AI and the Future of the HEP Computing Model (?)

Hardware costs, data processing, ML and AI at the HL-LHC scale

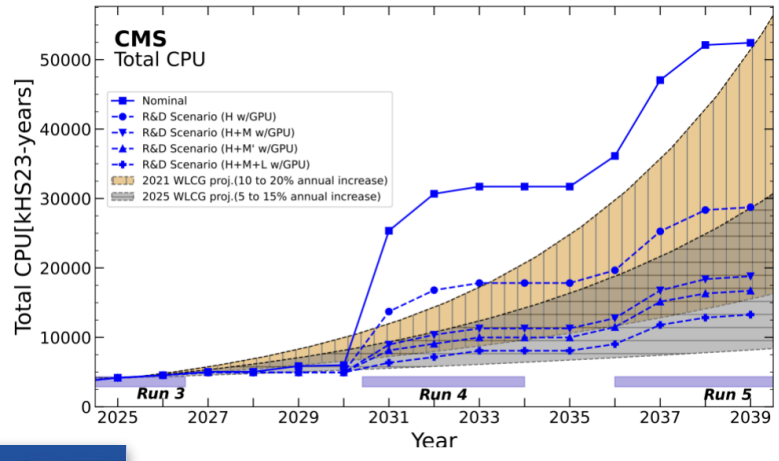
by Markus Elsing, CERN

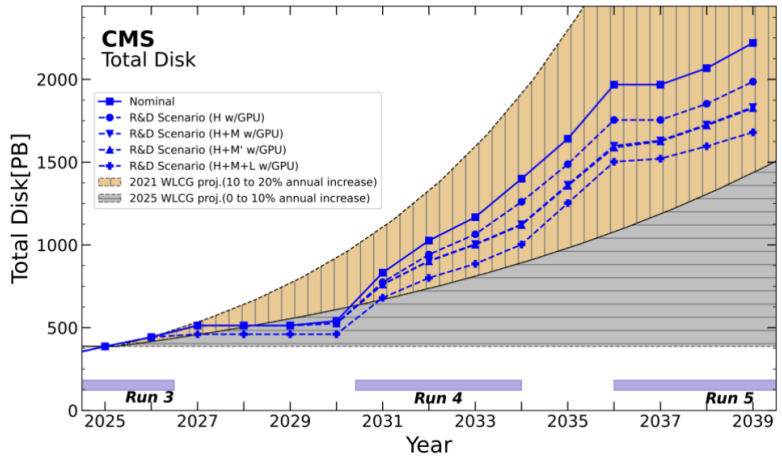
talk written using Agentic AI assistance

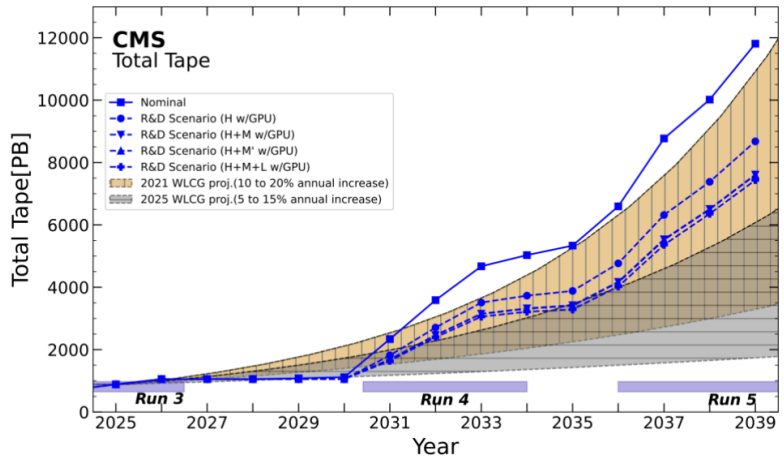
Computing needs vs. expected Resources ?

- Recent update of **HL-LHC needs** of **CMS** up to 2040: **CPU**, **disk** and **tape**
  - ➡ **ATLAS** is as well in the process of updating its needs for the upcoming HL-LHC computing TDR (overall very similar)
- WLCG **constant budget** extrapolation was corrected down for the **2025 update**
  - ➡ still **extremely optimistic** as of today, assumed **-10% / year** cost for CPU and tape, **-5% / year** for disk
  - ➡ even with the most ambitious **R&D scenarios** the needs **are at or above** the upper band (faster simulation and reconstruction, GPU usage where possible, smaller formats and less replicas)
  - ➡ **tape** and **disk** are the most critical, **GPU** needs would come on top

CMS HL-LHC CDR 2026

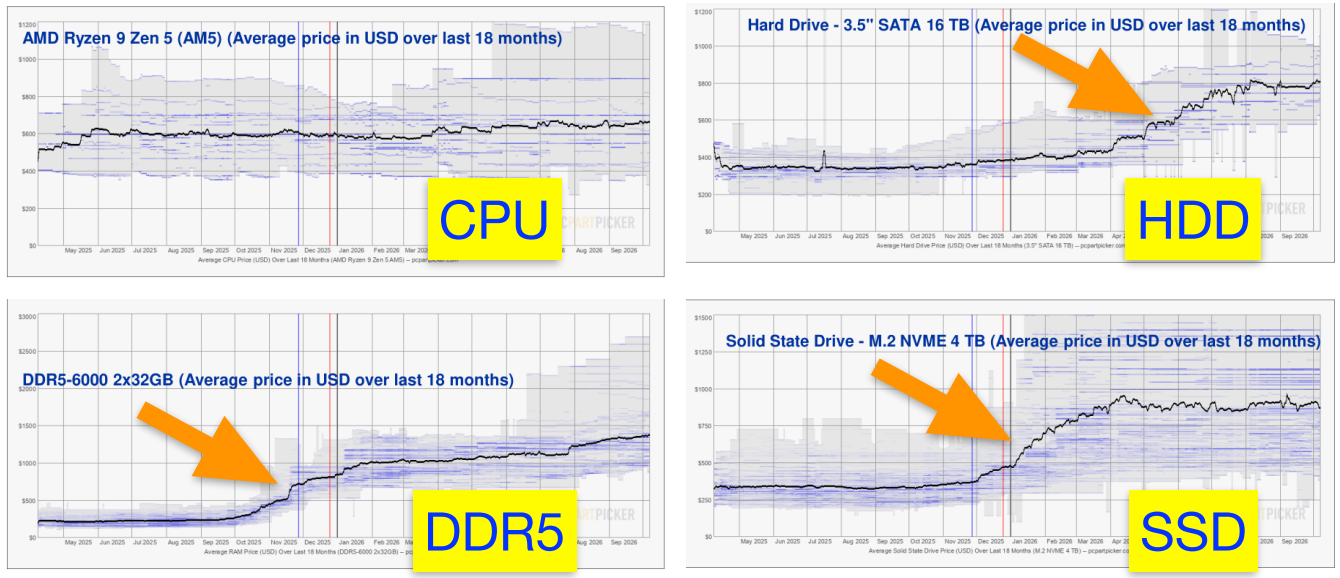






Hardware prices: driven by the AI industry

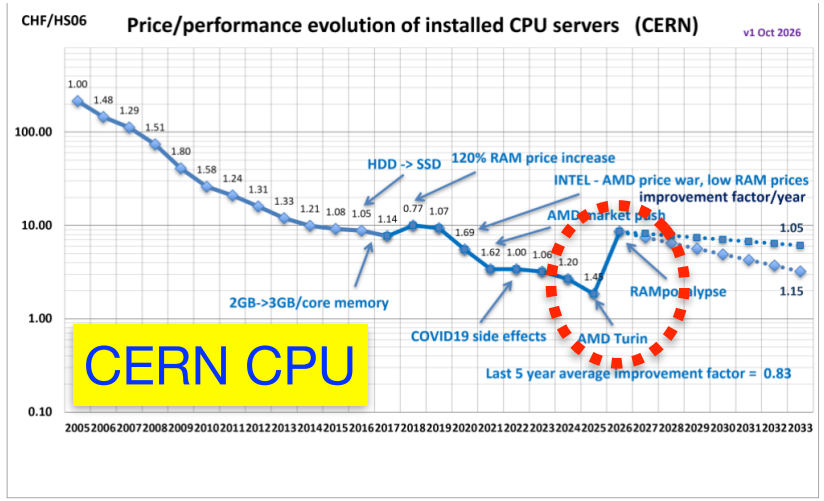
- **CPU prices flat** so far, but...
  - ➨ **memory prices** are the real killer:
    DDR5 up **> 5x** in a year
  - ➨ hyperscalers bought nearly all 2027 DRAM
- **HDD prices have more than doubled**, flash is following
  - ➨ demand from the AI sector, no new capacity soon
- **Tape** will equally become more expensive
- No relief expected before **2028 (?)**

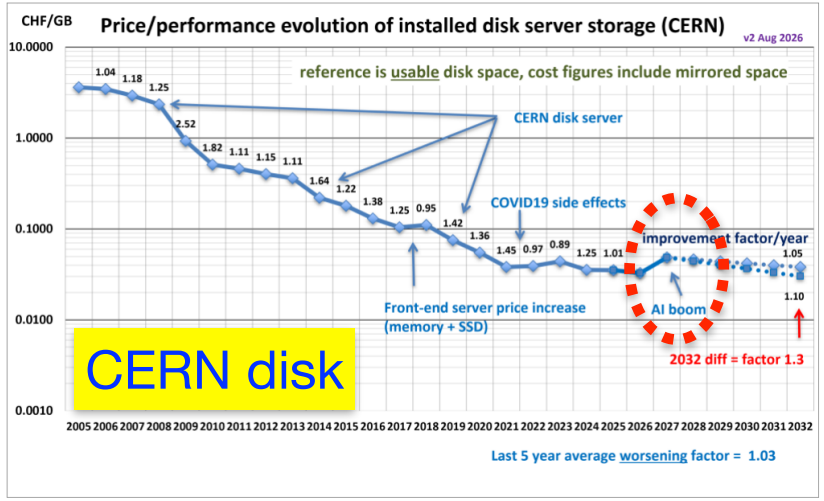


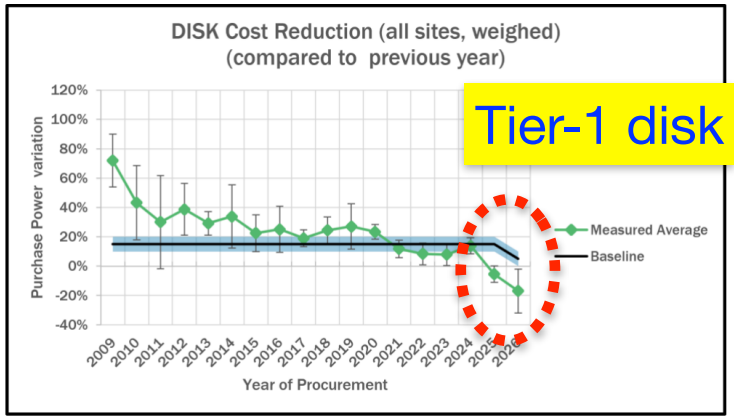
Price evolution of CPU, DRAM, HDD, SSD (A. Sciaba, ATLAS S&C week 85)

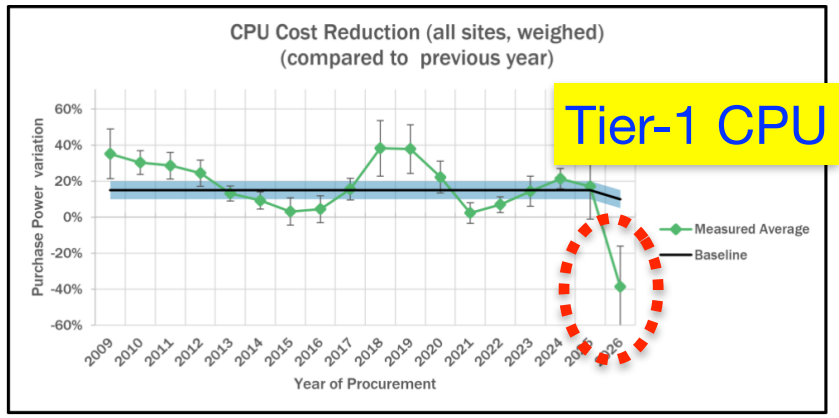
Costs at CERN and in WLCG: the trend has reversed

- No CPU procurement at CERN in 2026
  - ➡ same machine would cost **5x** more with half the RAM
- **Disk 50% more expensive**
  - ➡ after 6 years of falling costs
- WLCG Tier-1s see the same: CPU **+180%**, disk **+80%**
  - ➡ sites avoid buying hardware, prices change after signing contracts
- **… and this is not going to get better**





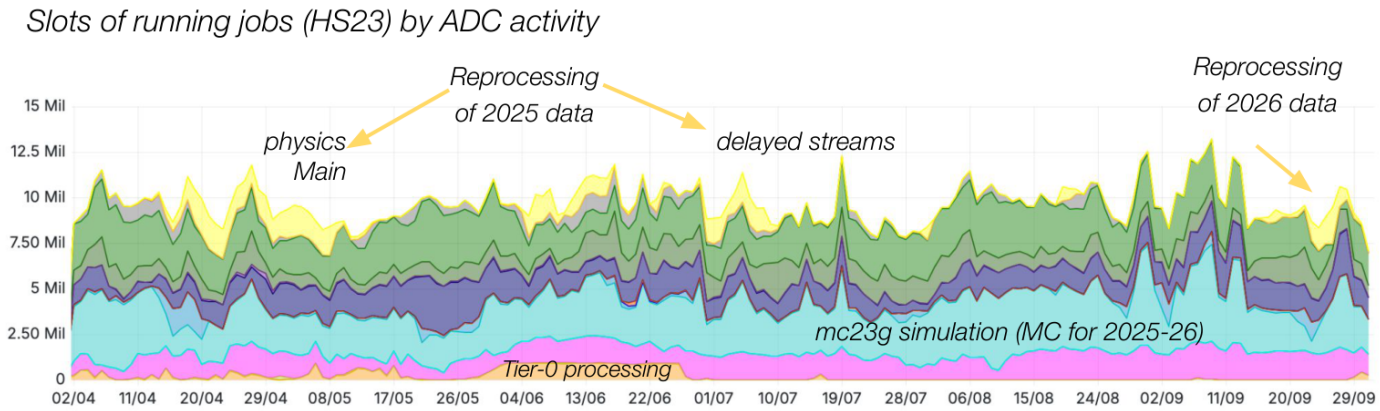


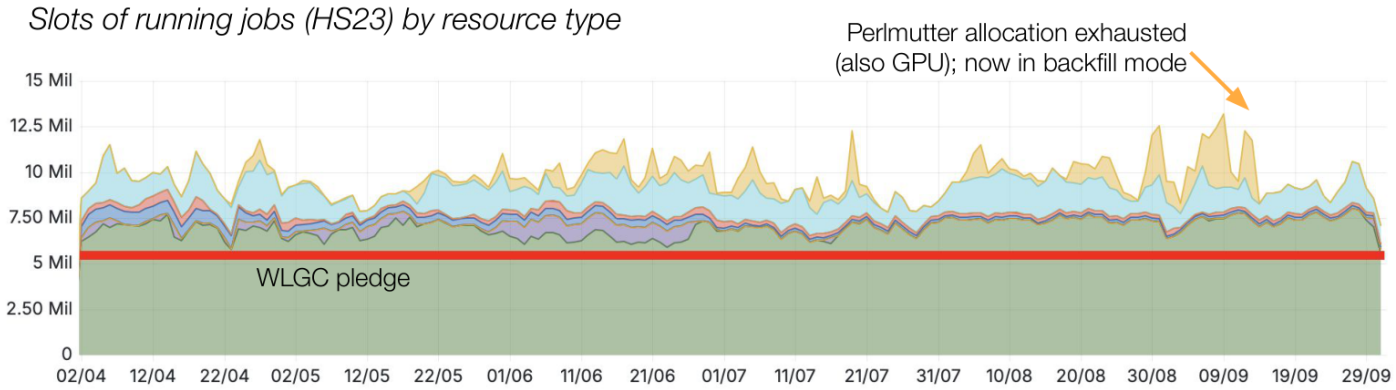


CERN IT price / performance and WLCG procurement cost variation (A. Sciaba)

HEP also relies on opportunistic CPU

- ATLAS runs on about **10M slots**, but the **WLCG pledge** is only **~ 5.5M**
  - ➡ the difference is **opportunistic**: **HPC**, HLT farm, BOINC, Tier-0 spare cycles
  - ➡ allocations are finite, e.g. **Perlmutter** exhausted, now in backfill mode
- The **baseline pledge** is becoming very expensive
  - ➡ we risk to **lose** the additional resources we get on top
  - ➡ opportunistic resources need **flexible workflows**: fast, preemptable, small footprint





last 6 months of ATLAS compute (HS23 slots)

What it means for the projected Computing Costs

- Typical Run 4 resources for **one** big LHC experiment:
  - ➨ 20 MHS23 CPU, 1.5 EB disk, 5 EB tape
  - ➨ **400 MCHF** at current cost
- A flat budget until 2034 buys **~ 1/4** of it under the new price regime
- Total resource cost up by a factor **~ 4** (ca. +300 MCHF)
  - ➨ if we are lucky, half of that
- We have to **revisit the computing model** and the hardware requirements

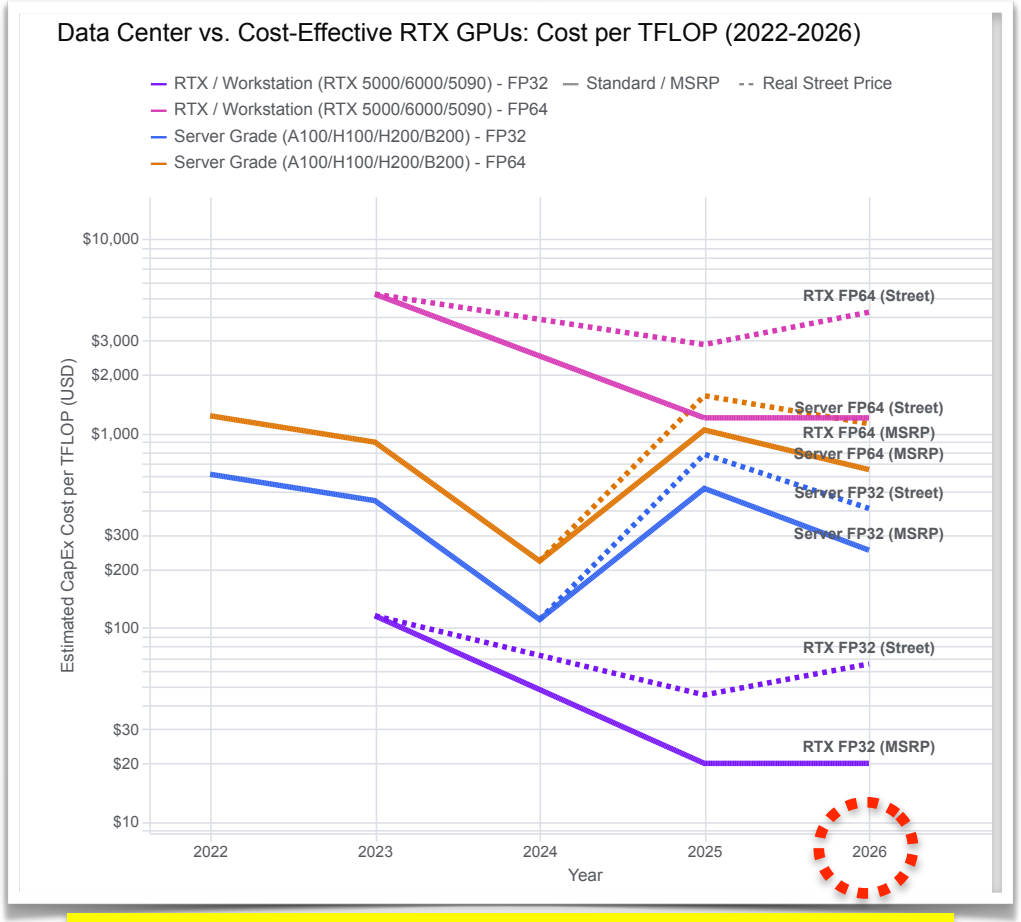
Resource	Current cost (MCHF)	Cost in 2034 (flat budget again)	Cost in 2034 (flat budget as expected in 2025)
20 MHS23 of CPU	200	93	30
1.5 EB of disk	150	102	48
5 EB of tape	50	23	21
Total	400	218	99

Cost	Current	In 2034 (flat budget again)	In 2034 (flat budget as expected in 2025)
CPU (CHF/HS23)	10	4.7	1.5
Disk (CHF/TB)	100	67.7	32.2
Tape (CHF/TB)	10	4.7	4.2

A. Sciaba, ATLAS S&C week 85

GPUs: the other Resource

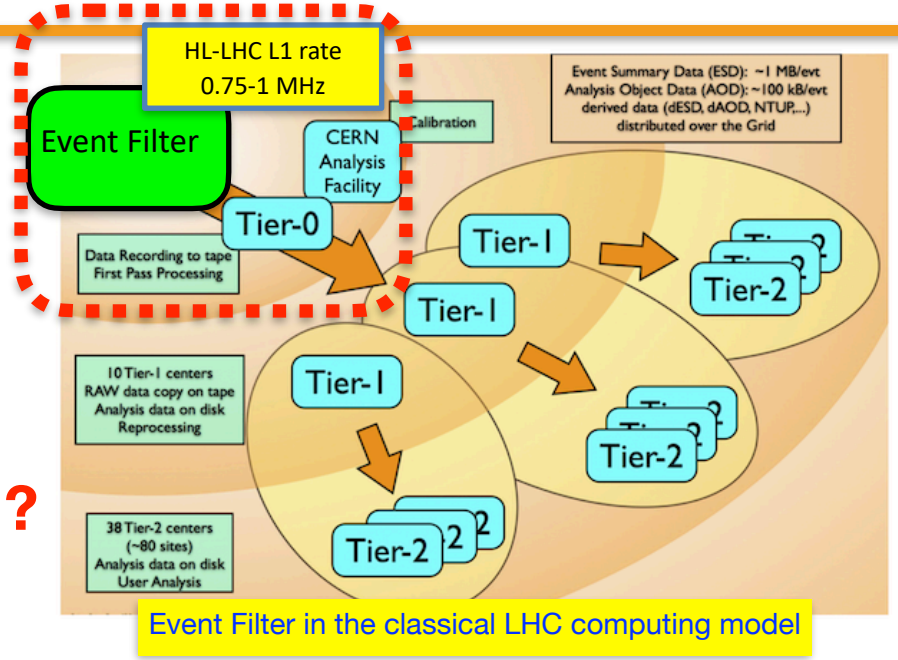
- **GPU needs** of the experiments are today **not** for offline computing
  - ➡ work ongoing to **migrate HEP workloads** to GPUs: generators, detector simulation, reconstruction
  - ➡ ATLAS and CMS plan to use (GPU+CPU) **heterogeneous Event Filter** trigger farms for the online event selection
- GPUs are not yet part of **WLCG pledges**
  - ➡ HEP makes extensive use of **HPC** resources
  - ➡ GPUs are only useful for **workflows** can really exploit them
  - ➡ **ML training** rapidly gaining importance (~million GPU hours)
  - ➡ **LLM hosting,** R&D on LLMs and **foundation models** (Mistal Flare)
- **Market for GPUs** even more under pressure !
  - ➡ NVIDIA GPUs move to **low precision AI** use-cases
  - ➡ **street prices** now significantly higher than nominal (MRSP)



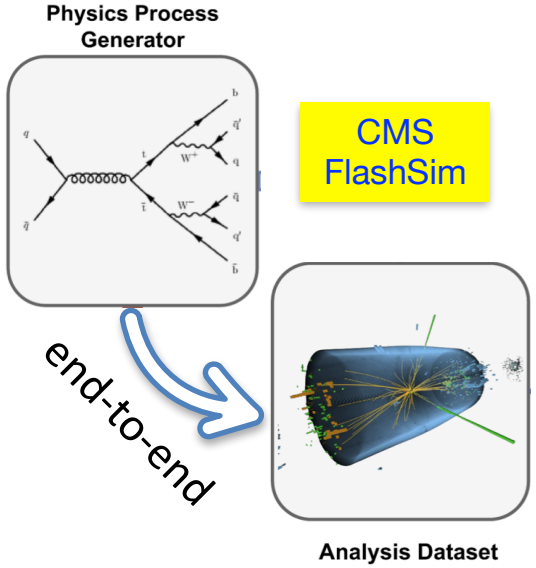
market summary obtained using Google Gemini

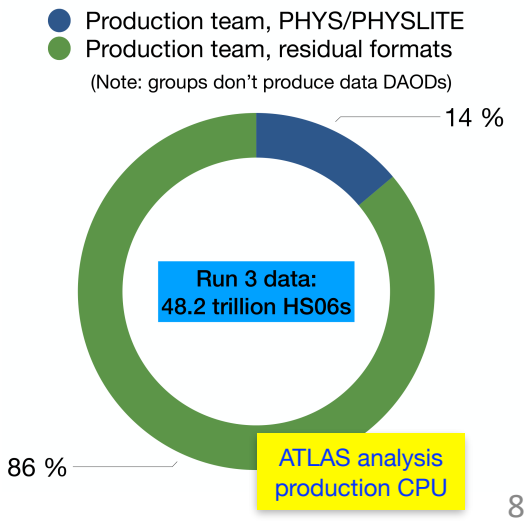
Consequences for LH-LHC Computing Models ?

- Experiments are facing a **computing crisis**
  - ➡ **normal R&D** to improve upon Run-3 model **not sufficient**, but helps:
    - drop group analysis formats, only **Nano** (CMS) and **PhysLite** (ATLAS) ?
    - increasing **MT count** for ATLAS production allows for **< 1GB/core**
  - ➡ need to consider **more dramatic measures**, e.g.:
    - drop Geant4 full simulation, only **ML based FlashSim / FastChain** ?
- Need to rethink parts of the **computing architecture ?**
  - ➡ size of **Event Filter** trigger farms equal to sum of a number of Tier-1 sites
  - ➡ Event Filter and **Tier-0** farms have both significant **disk buffers**
  - ➡ **merging** trigger selection and prompt Tier-0 processing to reduce combined resource needs ?
- Role of **AI and ML** beyond FlashSim
  - ➡ **agentic AI** for coding has taken over, leading to **faster R&D**
    - since 3 months **Claude is biggest contributor** to ACTS project
  - ➡ fast **ML inference** (GPUs) replacing classical **reconstruction**, etc.



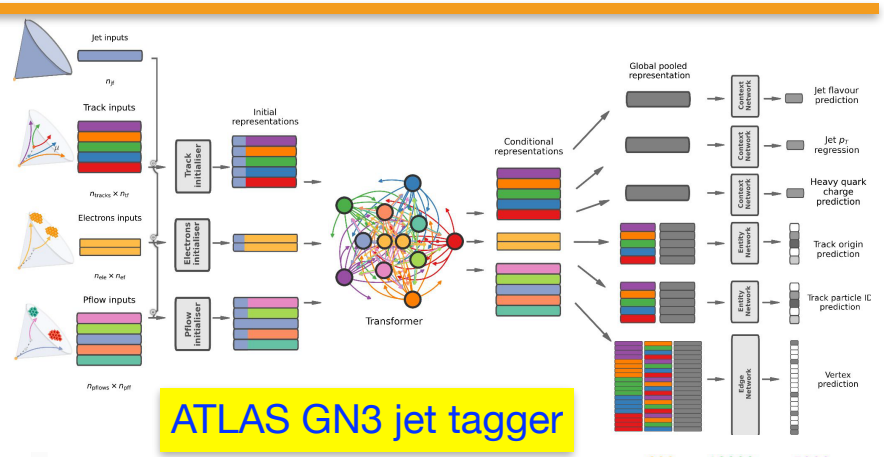
Event Filter in the classical LHC computing model



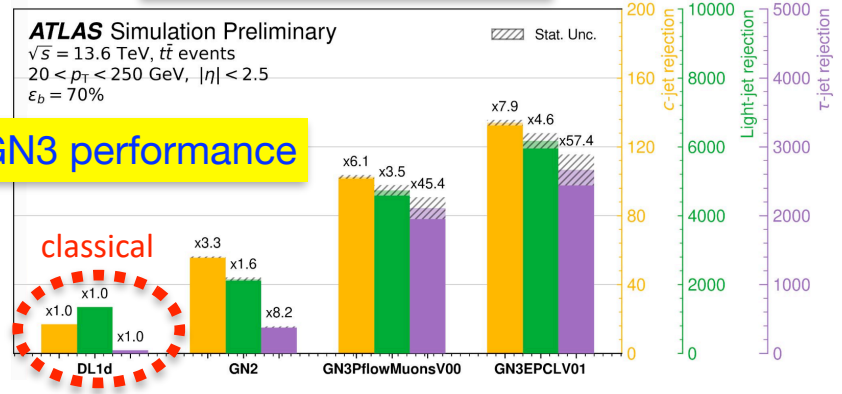


ML as a Opportunity and Compute Workload

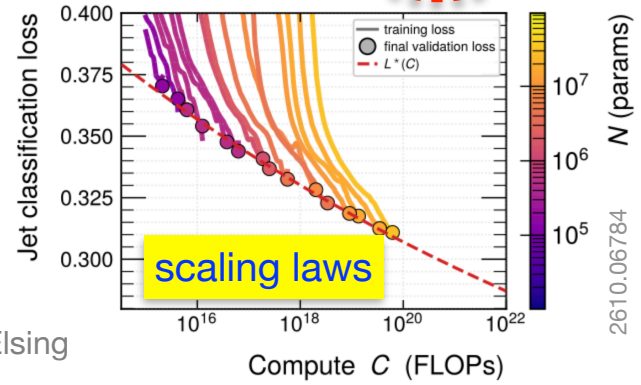
- **ML revolution** for (combined) event reconstruction
  - ➡ complements efforts to bring classical algorithmic code to GPUs
  - ➡ CMS **particle flow**, **jet tagging** (ATLAS/CMS), **tau-ID**,…
- **end-to-end** approaches boost **performance**
  - ➡ typically **transformer**, GNN or hypergraph based approaches
  - ➡ first generation models start from **low level inputs** (hits, tracks, clusters,…)
  - ➡ rapidly developing field to more complex taggers with **multiple targets**
- compute requirements **scaling fast**
  - ➡ ATLAS GN3 jet tagger **training** request for **1 Million GPU hours** (!)
  - ➡ **model size** rapidly increasing for latest model
  - ➡ **inference** still relatively fast, but concern for analysis data **production**
- R&D trend towards "**foundation**" models
  - ➡ model **combines** jet tagging with tau-ID, electrons, …
  - ➡ could reduce combined **memory** and **compute** footprint



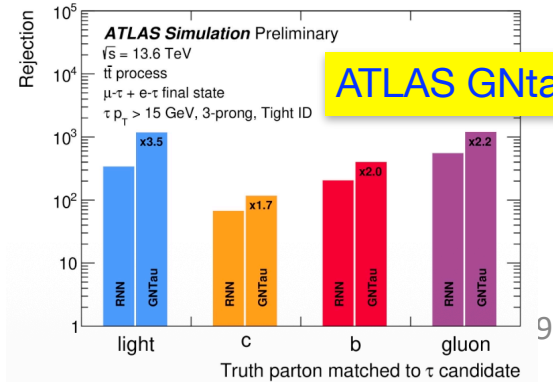
ATLAS GN3 jet tagger



GN3 performance



scaling laws

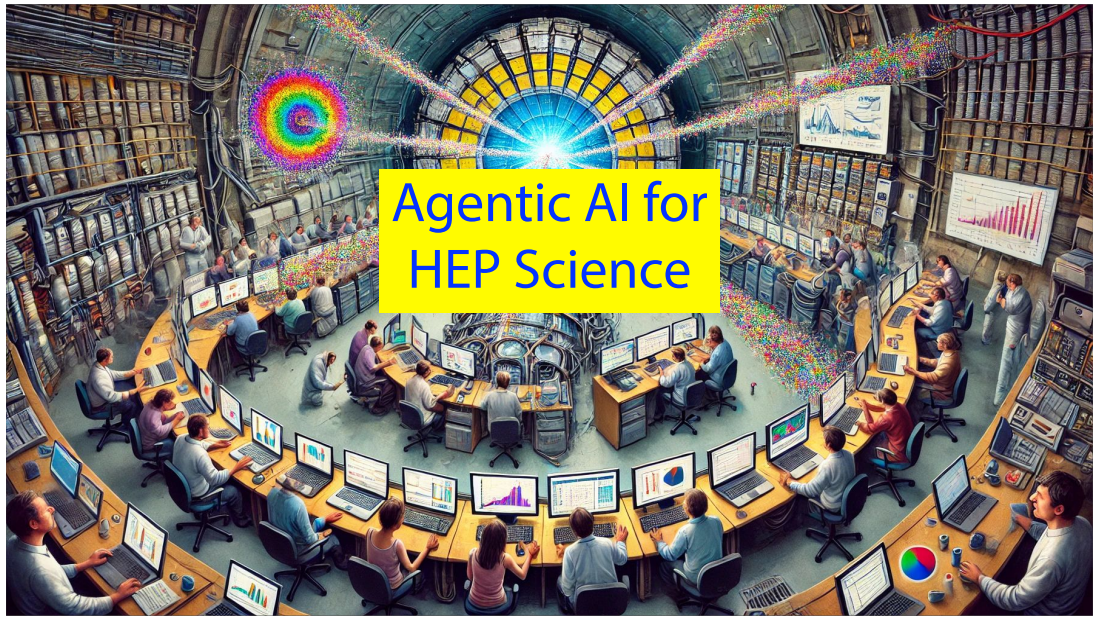


ATLAS GNtau

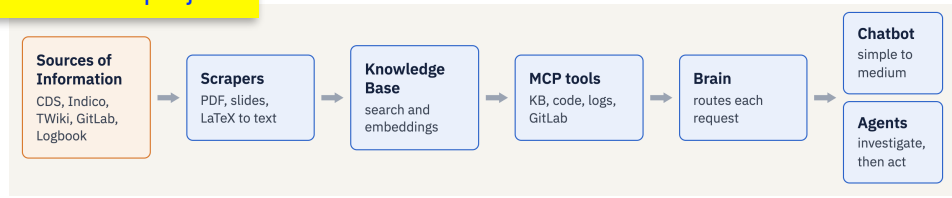
Agentic AI for HEP

- Agentic AI is a **transformative** technology

- Affects all areas of work:
  - ➨ agentic AI **coding** and software review, whole software projects now fully done via AI
  - ➨ **automatising** all workflows, from central production to students doing physics analysis
  - ➨ **theoretical calculations** shifting to agentic AI, pushing the limits in higher order calculations
  - ➨ AI assisted documentation and **publishing**
  - ➨ agentic AI helping to **review literature** and advising experts for their R&D
  - ➨ R&D on **agentic AI centric** collaborative projects



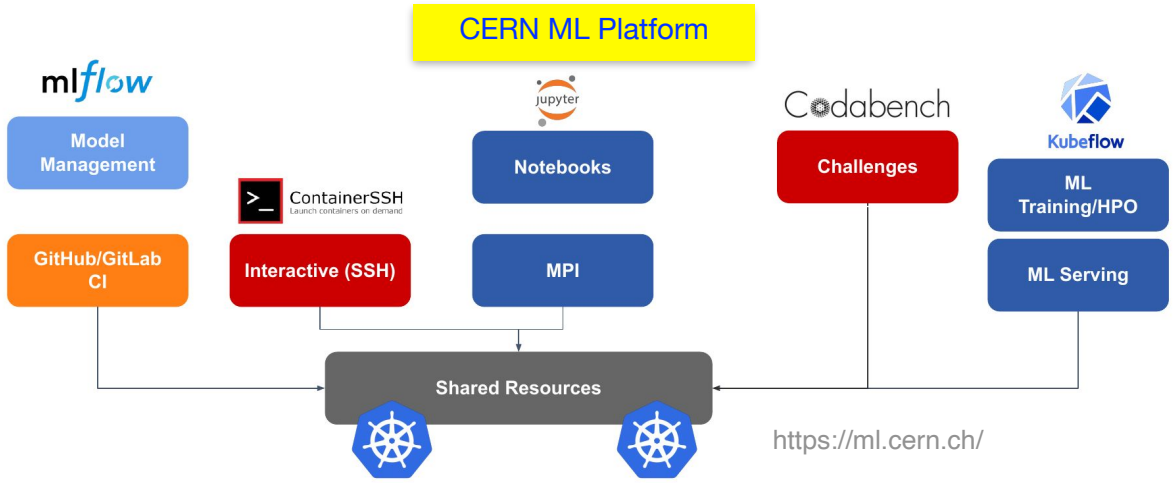
LHCb Brain project



- HEP community just starting
  - ➨ **institutes** using mix of scientific licences, private **subscriptions** and institute (**commercial**) licences
  - ➨ **cost** evolution is a worry, as is the question about **data privacy** / **data protection**

CERN Platform for ML and AI Infrastructure

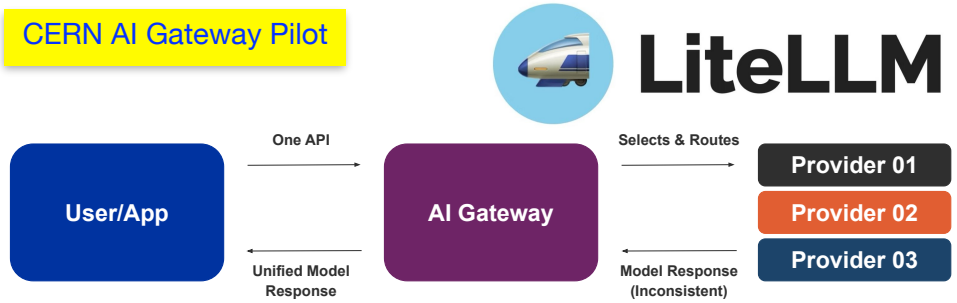
- CERN launched **ML platform**
  - ➡ starting point was ML farm for NextGen Triggers project
    - extensive use for training and R&D (initially NGT)
  - ➡ growing in size and functionality, Codabench for challenges



https://ml.cern.ch/

- CERN **AI infrastructure** pilot
  - ➡ task force within the framework of CERNs AI Strategy
  - ➡ Pilot 1: **commercial subscriptions** test with limited users
    - pay by token, data protection, cost for power user easily **7.5K$/month**
  - ➡ Pilot 2: **LLM hosting prototype** using **open weights** models
    - **AI Gateway** pilot with on premises and cloud resources
    - **Kimi-K3** (8×B300, Flatiron Institute), **MiMo 2.6 Pro** (4×MI300X, NGT)

Paulo, Raulian-Ionut | AIGW: CERN's LLM Service

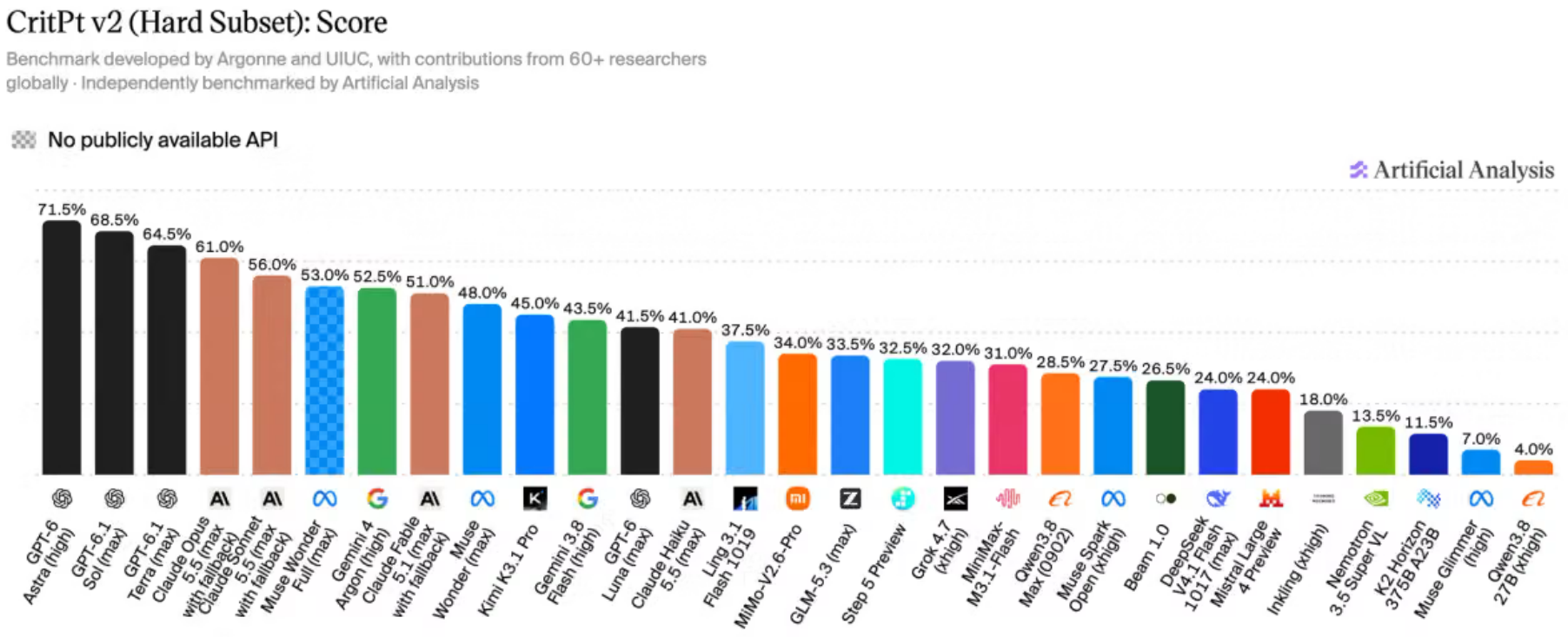


https://aigw.cern.ch/

- Additional resources for both to increase **user base**
  - ➡ CERN alone not able to serve **HEP community** and experiments
  - ➡ extend concept of WLCG to cover ML and AI use-cases (**WLCG for AI**) ?!

How good are Open Models?

- CritPT v2 benchmark : revised version of **71 research-level** physics challenges
  - ➨ best frontier model **GPT-6 Astra (high)** reaches **71.5%** and rapidly getting better
  - ➨ best open model **Kimi K3.1 Pro: 45.0%** or **63%** of the leader
- Open models are **close enough** to be useful
  - ➨ best models come from China: **longevity** of open weights, **training data** not public ? (applies also to other models)
  - ➨ alternative open models: **Muse Spark Open (27.5%)** (Muse Wonder is closed) or **Mistral Large 4 Preview (24.0%)**



CritPT: arXiv:2509.26574
Artificial Analysis v.5, 28.10.2026

The CERN AI Strategy 2026 - 2030

- AI/ML identified as a **core organisational capability**
  - ➨ scientific, operational and impact enabler
- **23.5 MCHF** via the Medium Tern Plan
  - ➨ funding dedicated activities across the CERN sectors
  - ➨ RCS organisation in place with **8 working groups**
- **10 lighthouse** projects
  - ➨ physics foundation model, fast simulation engines, AI agent …
  - ➨ **2030+ lighthouse** projects rely on additional funding
- Outcome needs **experiments** and **community** on board
  - ➨ HEP wide working groups and collaborative projects

RCS (Research Sector) working groups

WG	Topic	Scope
1	Pattern recognition & object reconstruction	tagging, PID, calibration, tracking
2	Global event interpretation, particle flow & foundation models	end-to-end event reconstruction
3	Generation & fast simulation	calorimeter surrogates, flash sim
4	AI for online processing	on-sensor, FPGA, real-time inference
5	AI for system control & operations	DQ, monitoring, root-cause analysis
6	LLM assistants for scientific production	coding, analysis, plotting, writing
7	Software & hardware infrastructure	GPU platform, storage, model registry
8	AI in theory & phenomenology (TH AI node)	AI for TH, TH for AI

Lighthouse Projects

LH	Initiative	Sector	Tier
01	Common AI Infrastructure & Platform	RCS-IT	MTP-funded
02	AskCERN: information discovery	all sectors	MTP-funded
03	Edge AI for real-time data acquisition	RCS	MTP-funded
04	Universal Fast Simulation Engines	RCS	MTP-funded
05	AI-driven operations & equipment support	ATS, RCS	MTP-funded
06	AI-assisted accelerator systems design	ATS	MTP-funded
07	Physics Foundation Model	RCS	MTP-funded
08	CERN AI Agent for Particle Physics	RCS	Horizon 2030+
09	AI-assisted detector design	RCS, ATS	Horizon 2030+
10	AI-driven material discovery	ATS	Horizon 2030+

M. Pierini, CERN AI Strategy (Nov 2026)

The Road Ahead

- **Costs increase** of puts LH-LHC **computing model** at risk
  - ➨ aggressive R&D to use resources **more efficiently** will help, but more **dramatic changes** may be required
- **ML** revolution has major impact on HEP data processing
  - ➨ fast simulation and fast chain, online and offline reconstruction and identification, analysis …
  - ➨ training needs and model sizes are scaling fast (~ 1M GPU-hours for one tagger)
- **Agentic AI** is transforming how science is done
  - ➨ many open questions, including cost, data protection, ethics …
- CERN pilots for **ML training** and **LLM hosting**
  - ➨ serving HEP community may require a WLCG-like federated organisation (**WLCG for AI**)
- **CERN AI strategy** framework needs experiments and community on board

Thank you

Questions ?