OPPORTUNITIES FOR A HETEROGENEOUS FUTURE

Sofia Vallecorsa

Coordinator - Quantum Technology Initiative

CERN

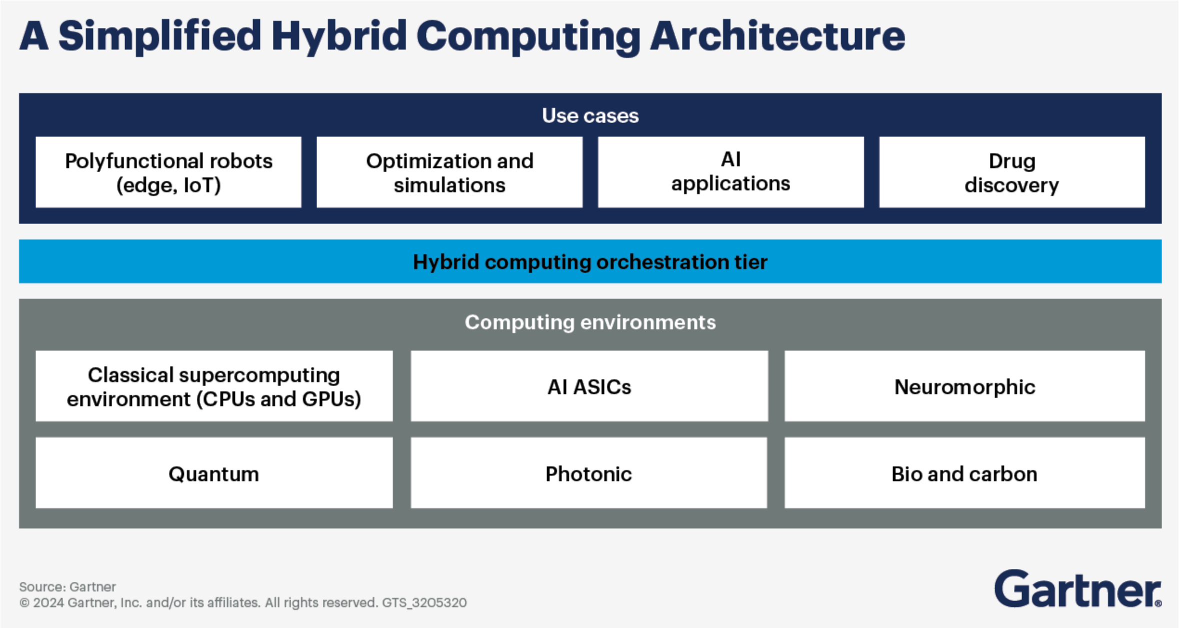
WLCG 20th Anniversary

December 8th, 2025

INTRODUCTION

*"The next era of computing will **combine conventional computing technologies with quantum, neuromorphic and other computing mechanisms** to help organizations explore and solve problems»*

Gartner, October 2024



HEP has a long history of **exploiting cutting edge technologies**

- Some had a fundamental impact, e.g. full **SW triggers** in LHCb, **GPU based HLT** in CMS, **GPU-based continuous readout** in Alice, **AI** for data processing/analysis

Today we study interesting approaches such as e.g. **Quantum Computing, Neuromorphic…**

- Others (FPGA, GPUs, …) **are not «new», but their use in HEP is still «relatively» limited.**

Technology now evolves in months, not years as it did decades ago: **we need flexibility in how we approach, combine, and utilize resources**

Efficiently integrating new technologies requires a roadmap and time !

INTEGRATING NEW TECHNOLOGIES PRESENTS MULTIPLE CHALLENGES

General Challenges

- Advanced scheduling techniques
- Standardization and interoperability
- Agnostic scalability
- Data management
- Reliability and fault tolerance

Specific to Scientific Computing

- Portability of scientific software
- Precision and numerical stability
- Ensuring performance across heterogeneous resources
- Workflow complexity
- Specific libraries

When we get to establish strategies for integration we need to have **a very good understanding of algorithms and expected performance**

DO WE HAVE **PROBLEMS** THAT REQUIRE HETEROGENEOUS INFRASTRUCTURE ?

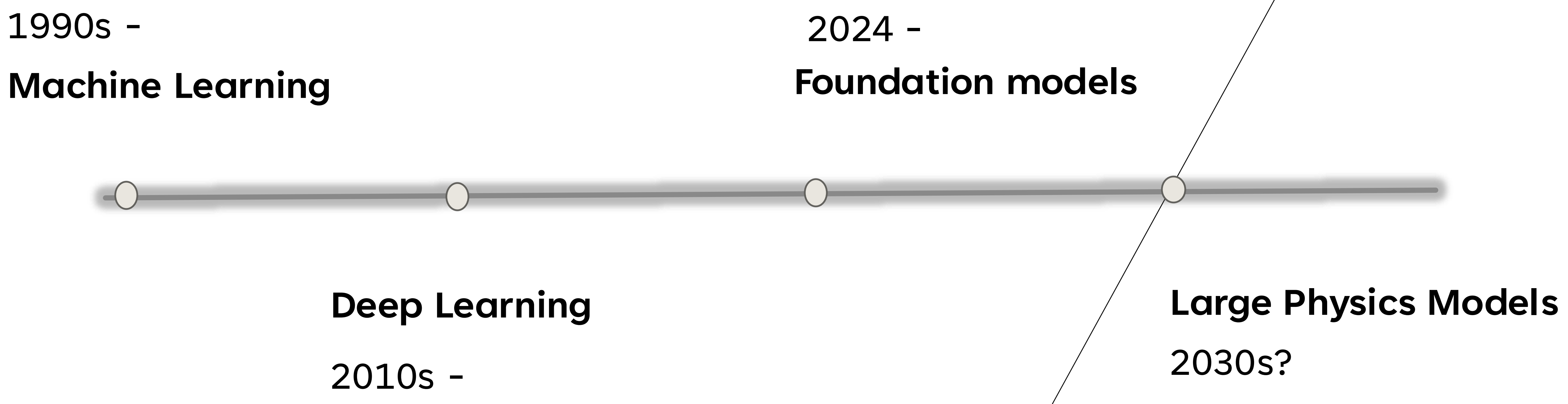
DO WE HAVE **ALGORITHMS** THAT CAN EFFECTIVELY USE IT?

DO WE HAVE **PROBLEMS** THAT REQUIRE HETEROGENEOUS INFRASTRUCTURE ?

DO WE HAVE **ALGORITHMS** THAT CAN EFFECTIVELY USE IT?

YES, WE DO

AI: EXAMPLE OF A NEW PARADIGM



AI: EXAMPLE OF A NEW PARADIGM

Simple representations
Important preprocessing

1990s -

Machine Learning

Deep Learning

2010s -

Complex representations
Learn specific tasks

"Learn the data"
Capable of multiple tasks

2024 -

Foundation models

Large Physics Models

2030s?

Full multi-modality
Physics "understanding"

AI: EXAMPLE OF A NEW PARADIGM

Simple computational graphs

1990s -

Machine Learning

Deep Learning

2010s -

Depth
Hierarchical abstraction

Complex computational graphs
Attention

2024 -

Foundation models

Large Physics Models

2030s?

??

AI: EXAMPLE OF A NEW PARADIGM

Simple computational graphs

1990s -

Machine Learning

CPUs

Deep Learning

2010s -

Depth
Hierarchical abstraction

GPUs/accelerators

Complex computational graphs
Attention

2024 -

Foundation models

GPU/accelerators clusters

Large Physics Models

2030s?

??

AI factories

Large collaborative projects

DEEP LEARNING

...in production for Run 3

Highly optimized

Out of the box models rarely work

A broad range of applications

Different computing requirements

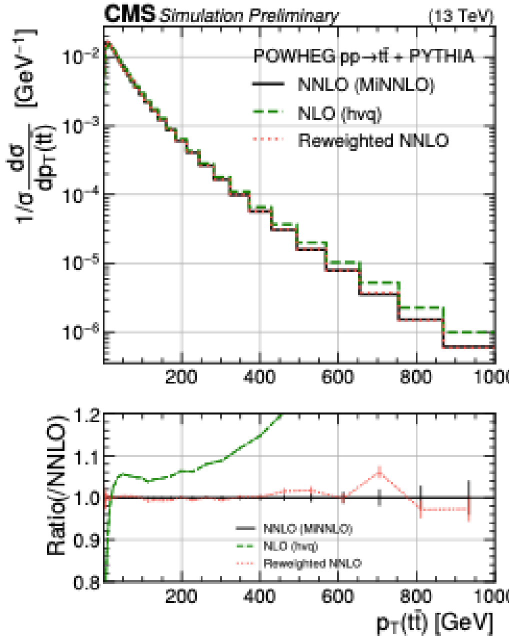
Ex. ML in Real Time environment

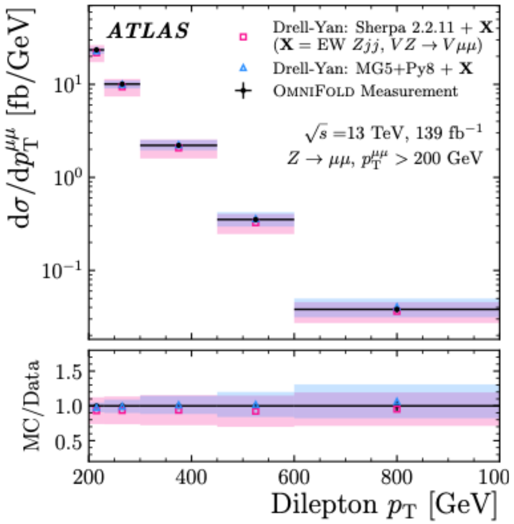
Constraints on Latency

Constraints on model complexity

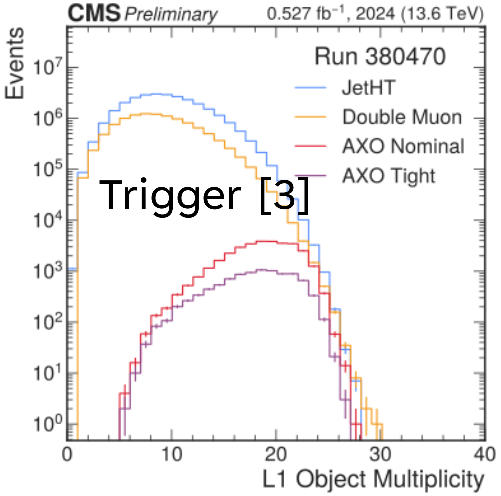
Constraints from the quality of data available

MC reweighting [1]



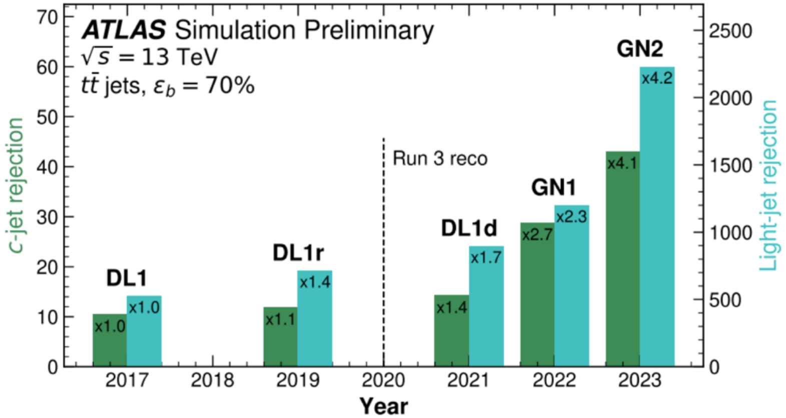


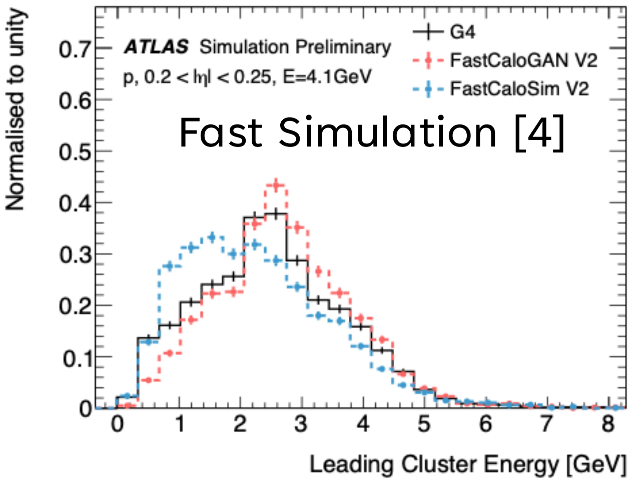
Unfolding [2]



Trigger [3]

Tagging [7]





Fast Simulation [4]

PID in ALICE O2 [5]

CMS Track selection [6]

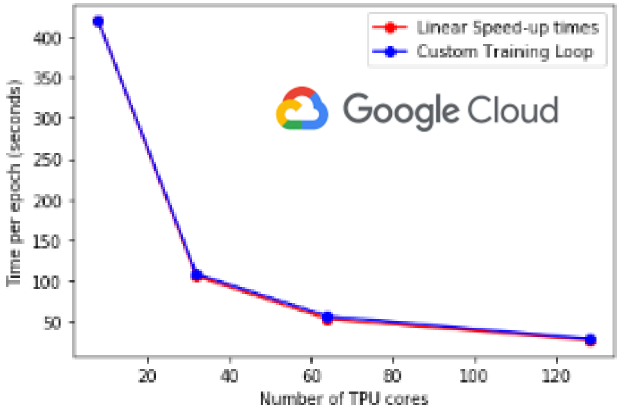
...

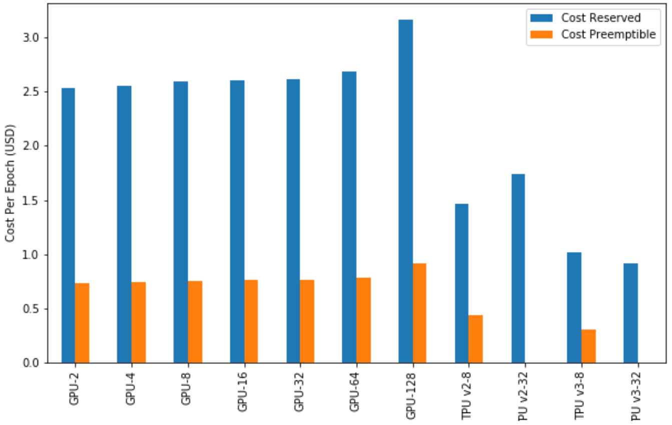
WLCG HETEROGENEOUS ARCHS. WORKSHOP: https://indico.cern.ch/event/1526077/

FROM INFRASTRUCTURE TO SERVICES

BACK IN 2021: GENERATIVE MODELS ON PUBLIC CLOUD

Total time:
1 hour on 128 v3 Google TPU cores

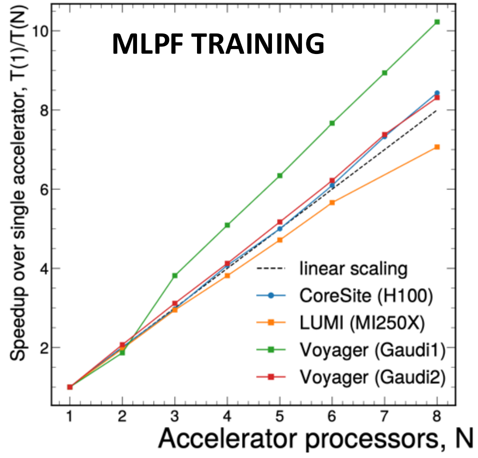


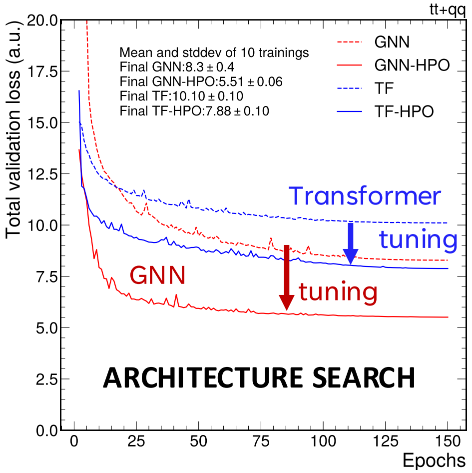


Total time:
1 hour on 128 V100 GPUs

Cardoso, R., et al. "Accelerating GAN training using highly parallel hardware on public cloud." EPJ Web of Conferences. Vol. 251. EDP Sciences, 2021.

TODAY: TRAINING ON HPC CENTERS

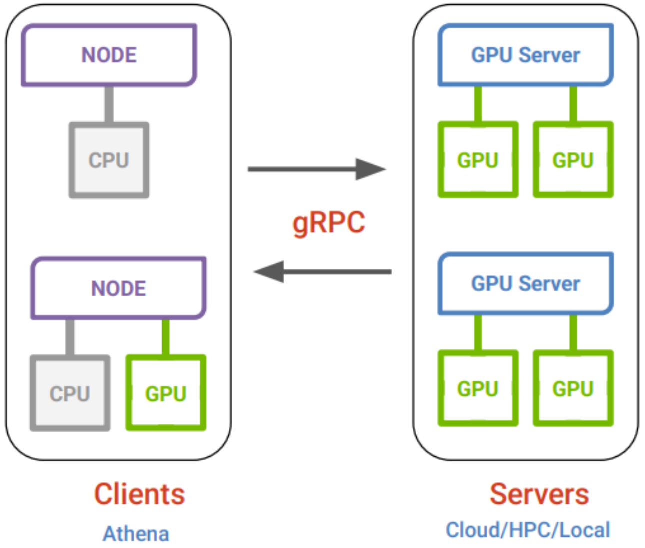




Pata, J., Wulff, E., Mokhtar, F. et al. Improved particle-flow event reconstruction with scalable neural networks for current and future particle detectors. Commun Phys 7, 124 (2024).

TODAY: INFERENCE AS A SERVICE

Services for Optimized Network Inference on Coprocessors (SONIC)



INFERENCE ON FPGAs

hls4ml

See Anna Hallin, ML4Jet, June 2025

A NEW APPROACH : FOUNDATION MODELS

Learn optimized data representation and adapt to multiple tasks via fine tuning
Trained on large scale, multi-modal data sets.
Enable a more efficient, sustainable and streamlined approach

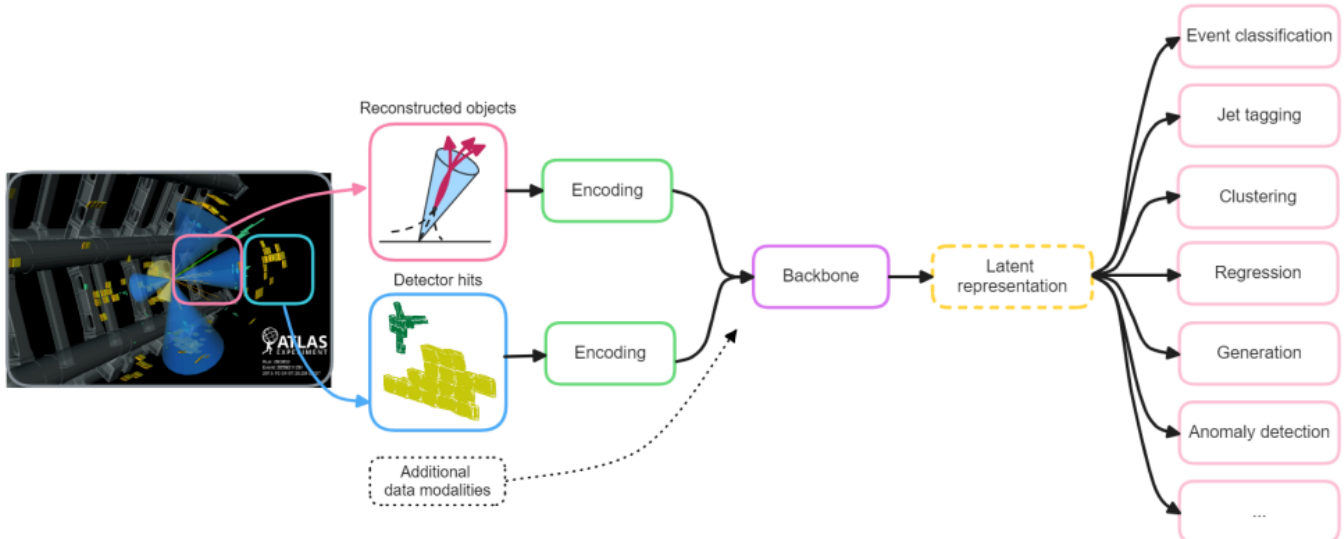
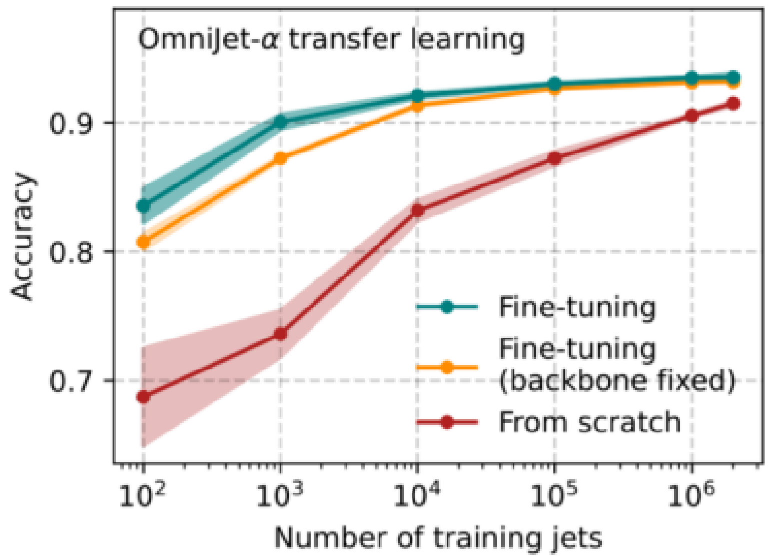


Image credit: J. Birk



Conventional training process becomes «multi-step»:

- **Initial pre-training**: large datasets, GPU cluster (Possibly centralized?)
- **Model adaptation**, fine-tuning, etc: smaller data sets, smaller computing requirements (but large models could have large memory requirements)

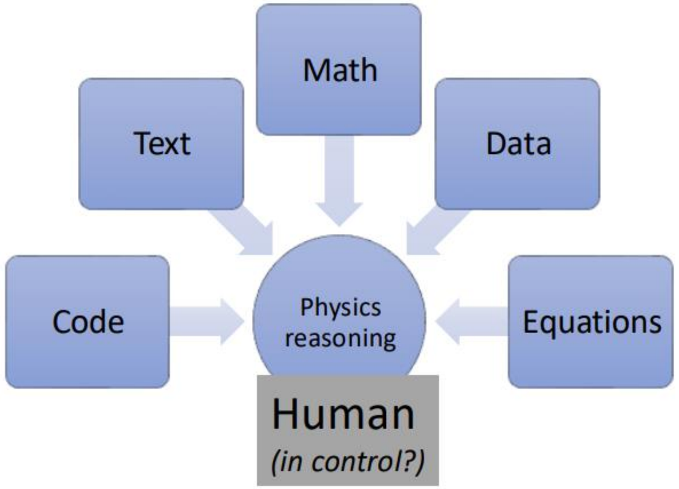
Sascha Caron, Nikhef

WLCG HETEROGENEOUS ARCHS. WORKSHOP:
https://indico.cern.ch/event/1526077/

LARGE PHYSICS MODELS... ...THE FUTURE?

Advances in **architectures,** transformer based and beyond, and **hybrid training methods**, combining self-supervision and reinforcement learning, **extend LLM concepts into scientific domains**

Large scale models trained on physics data, capable of multiple tasks, integrating symbolic reasoning, knowledge generation, ...



LPMs are built for physics reasoning and integration across our modalities.

- **End-to-end approach**
- **Faster simulations** and experiment design
- Symbolic reasoning for **improved interpretability**
- **Improved knowledge discovery** to generate new hypotheses across large datasets.
- **Scalability and efficiency via adaptation**
- **Democratization** (accessible to smaller teams)

Sascha Caron, Nikhef

WLCG HETEROGENEOUS ARCHS. WORKSHOP
https://indico.cern.ch/event/1526077/

SHIFTING AI PARADIGMS REQUIRE NEW SCALABLE APPROACHES

…collaborations and infrastructure

Large Physics Models: Towards a collaborative approach with Large Language Models and Foundation Models

Kristian G. Barman[*1], Sascha Caron[*2], Emily Sullivan[3], Henk W. de Regt[4], Roberto Ruiz de Austri[5], Mieke Boon[6], Michael Färber[7], Stefan Fröse[8], Faegheh Hasibi[9], Andreas Ipp[10], Rukshak Kapoor[11], Gregor Kasieczka[12], Daniel Kostić[13], Michael Krämer[14], Tobias Golling[15], Luis G. Lopez[16], Jesus Marco[17], Sydney Otten[18,19], Pawel Pawlowski[1], Pietro Vischia[20], Erik Weber[1], and Christoph Weniger[21]

https://arxiv.org/pdf/2501.05382

- Building large AI models would require a **collaborative effort within HEP and beyond**
  - **At a scale comparable only to the inception of WLCG!**
  - **Stable access to AI factories for training**
  - **Combining data from different sources (experiments?)**
- Initial ideas for **integrating a Federated Learning approach** (M. Pierini, L. Serio @CERN) as a path towards efficient use of distributed resources
  - Train large models on multiple experiments data while adhering to specific data policies

IN THE MEAN TIME... QUANTUM COMPUTING

Not a plug-and-play technology.

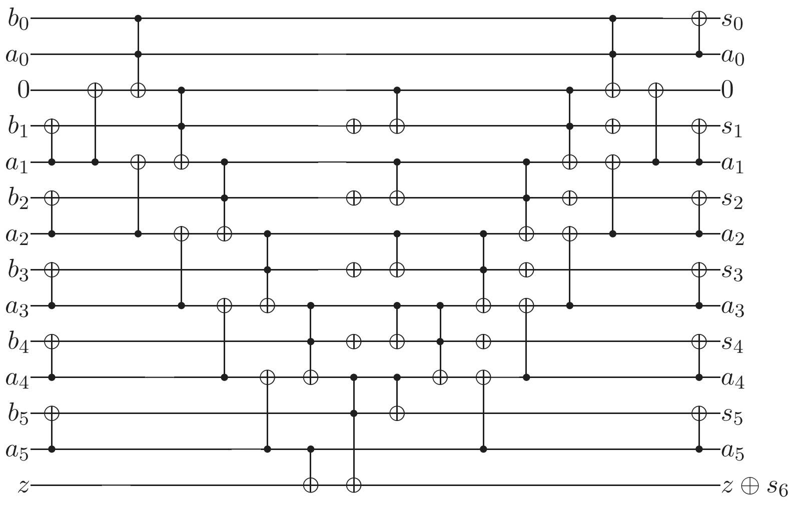
To manage expectations, **focus on incremental integration and R&D**

When the technology matures, we will be **prepared to adopt it seamlessly**.

https://quantumcomputingreport.com/the-quantum-hype-curve-balancing-excitement-and-realism-in-tech-communications/

MAIN QUANTUM COMPUTING PARADIGMS

Gate-based quantum computers

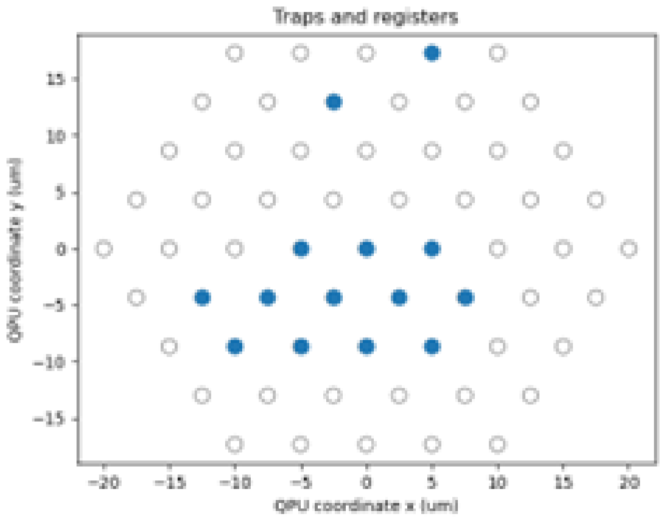


The ripple-carry adder for $n = 6$.

Algorithms are built as series of quantum gates (unitary transformation)

IBM QUANDELA

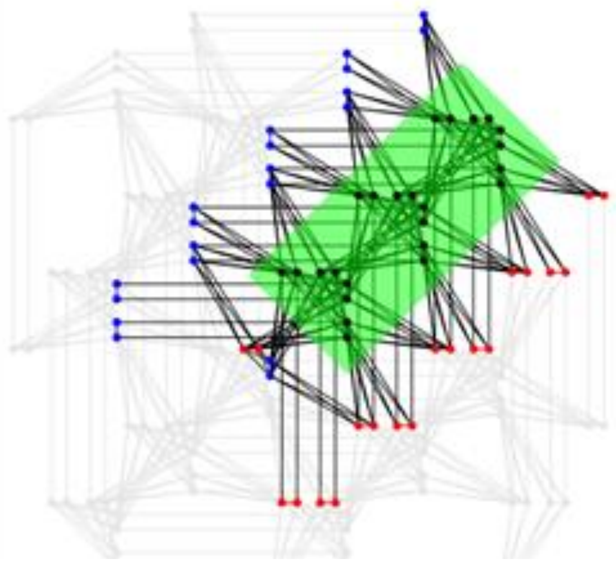
Analog quantum simulators



Problems embedded as graphs, solved as Ising or QUBO, using dynamic qubit positioning but no or poor local qubit control

Pasqal QuEra

Quantum annealers

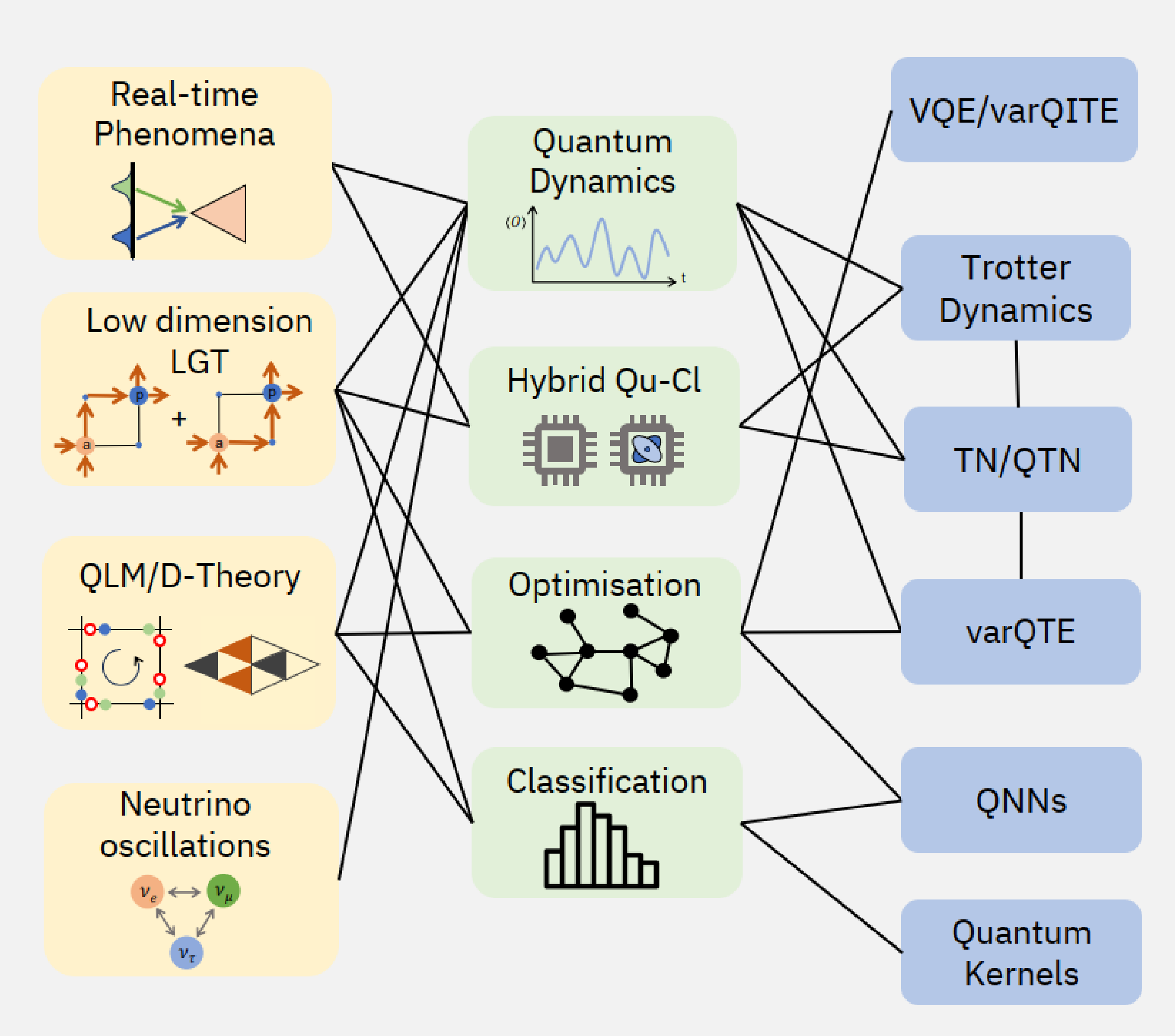


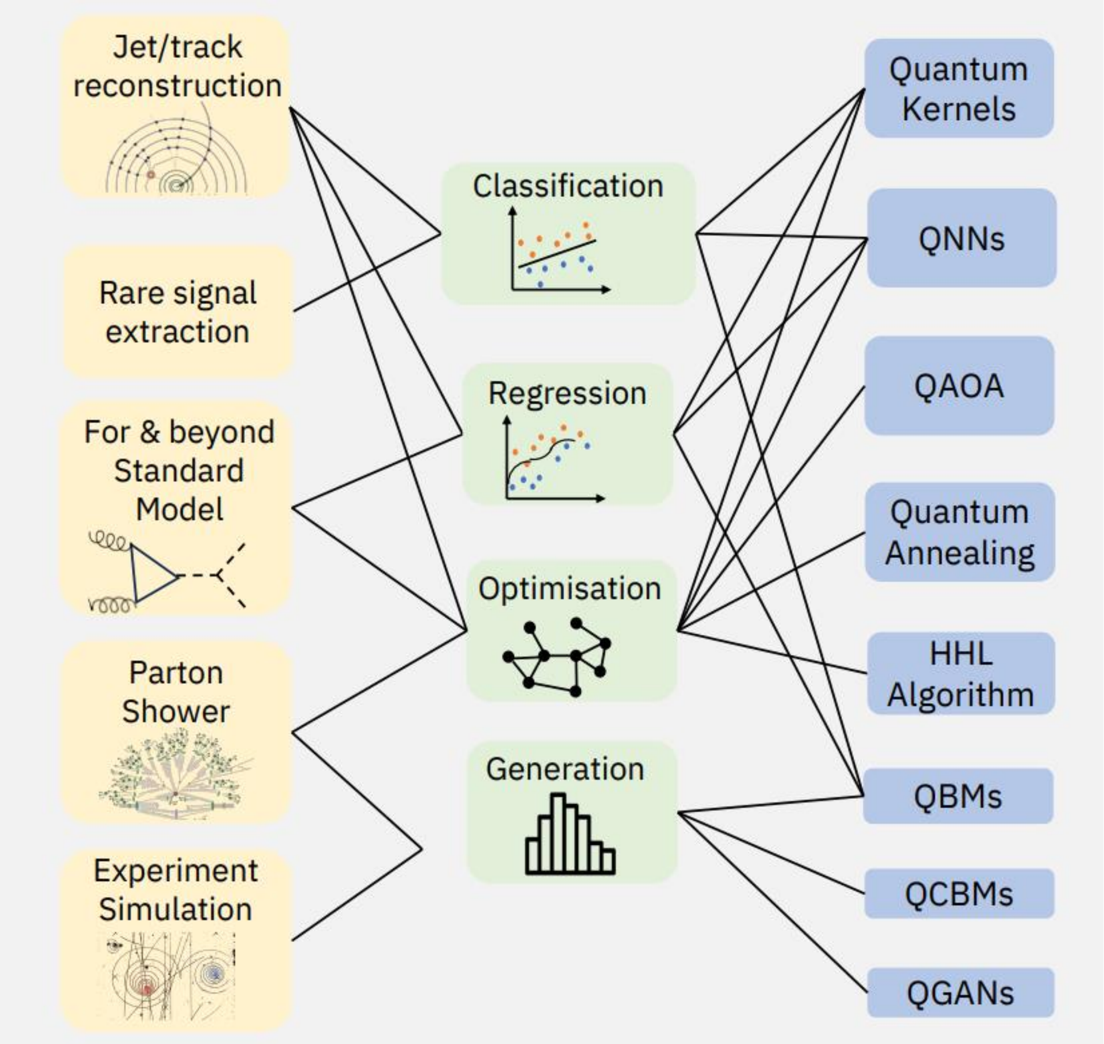
Problem embedded in a BQM model to solve Ising or QUBO problems, using static qubit connectivity and local control

D-Wave The Quantum Computing Company™

Quantum inspired : classical methodologies inspired by quantum system behavior

A WIDE RANGE OF ALGORITHMS USEFUL FOR HEP

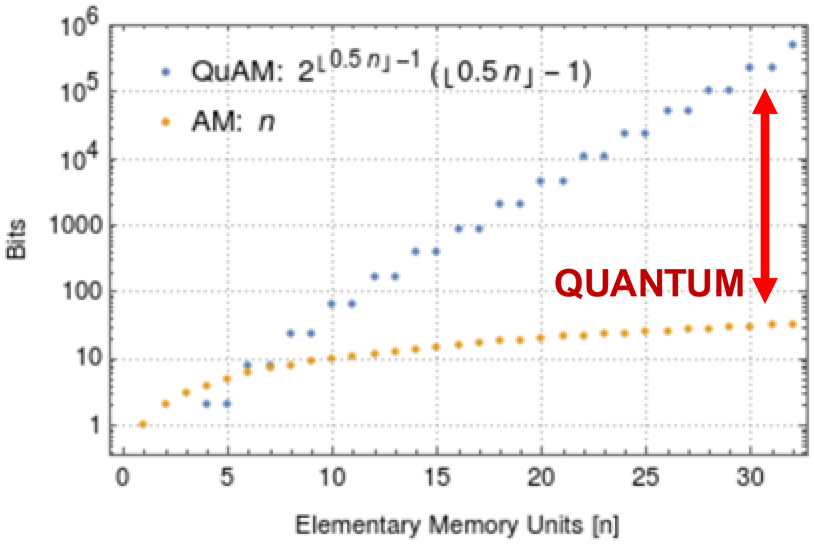




"Quantum computing for high-energy physics: State of the art and challenges." PRX Quantum 5.3 (2024): 037001

WHAT DOES «USEFUL» MEAN FOR US?

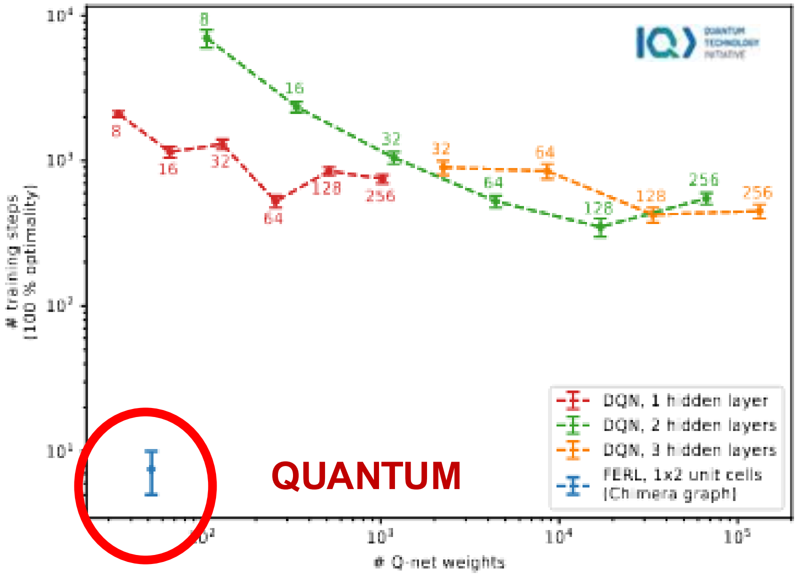
Extend the **definition of quantum advantage** towards **proving usability for realistic problems.**



Exponential Memory

(Quantum **associative memory** for particle trajectory reconstruction)

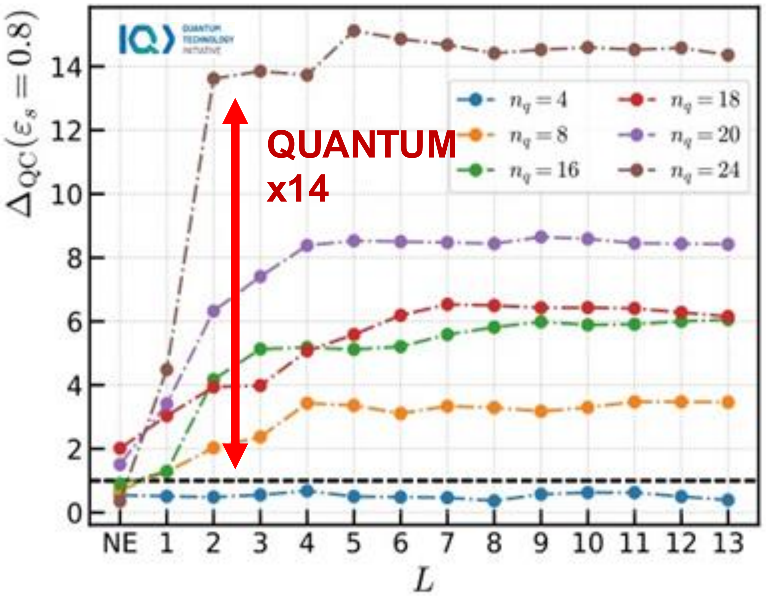
arXiv:1902.00498 [hep-ex]



Resource efficiency

(Quantum **reinforcement learning** for beam steering in particle accelerators)

Quantum Science and Technology 9.2 (2024): 025012.



Improved discovery capability

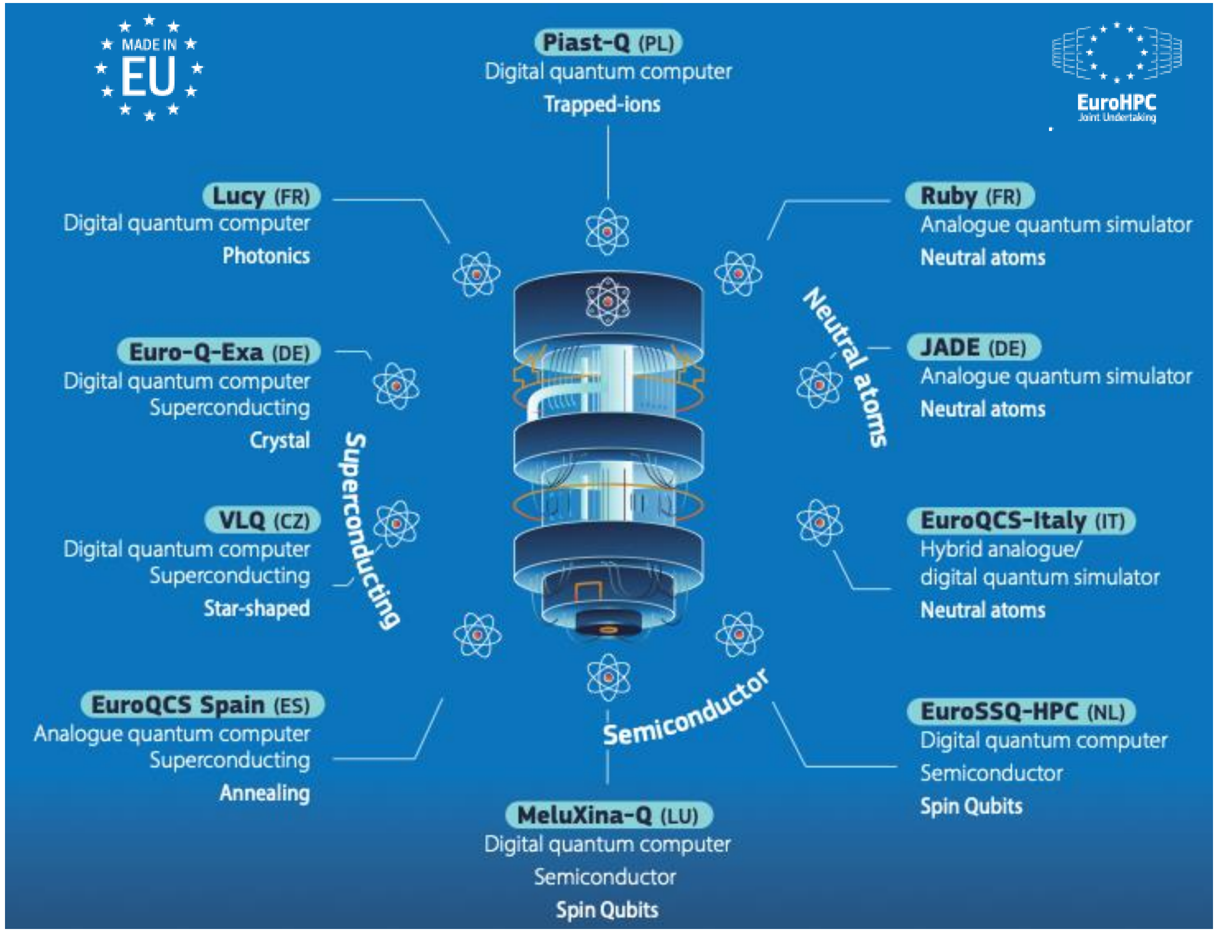
(Quantum **anomaly detection**)

Communications Physics 7.1 (2024): 334.

GLOBAL EFFORTS TOWARDS QUANTUM CENTRIC SUPERCOMPUTERS

- **Hybrid quantum algorithms extend the reach** of near-term quantum infrastructure
- **Advanced workload management systems control hybrid quantum-classical** jobs for realistic applications

See Robledo-Moreno J. *et al.* "**Chemistry beyond the scale of exact diagonalization on a quantum-centric supercomputer**." Science Advances 11.25 (2025): eadu9991.



Use these systems to study scaling towards fault tolerant era *(quantum computers allowing full error correction at a sizeable scale)*

TOWARDS USABLE SYSTEMS?

Major companies’ roadmaps **promise hundreds of logical qubits** and **hundred million quantum gates** by early 2030s (pre-fault tolerant era) *(see https://www.arxiv.org/abs/2506.03094)*

Initial experiments on **distributed quantum computing** are underway

With such systems, we could .... probably:

- Implement initial quantum associative memory algorithms
- Implement symmetry aware Quantum Machine Learning
- Devise new training strategies for generative models
- Initial demonstrators for event generators and MonteCarlo integrations

Source:IBM

INTEGRATING EMERGING TECHNOLOGIES IN SCIENTIFIC COMPUTING

Careful **evolution and phased integration** of new computing paradigms is essential to manage rapid technological change.

We have a unique opportunity to **embed sustainability principles** from the outset in the design and adoption of new technologies.

A strategic change **from domain-specific infrastructures to cross-domain**, shared facilities is already ongoing for European research.

Managing expectations and technological hype is essential to ensure realistic roadmaps and long-term impact.

HEP's success is rooted in its result-oriented, bottom-up culture.

Strategies for future heterogeneous infrastructure must be developed

inclusively, ensuring full credibility, leveraging the expertise of the community

THANK YOU

Sofia Vallecorsa

Sofia.Vallecorsa@cern.ch

PATHS TOWARD ENERGY EFFICIENT AI

STRATEGIC

Optimise **use case definition**

Optimise **integration** with existing software

Estimate classical tools **replacement savings**

Actively contribute to **existing green intiatives beyond HEP**

HARDWARE

Improve usage efficiency of available h/w

AI models are based on a few frameworks: optimising them impacts all use cases

Introduce **new h/w technologies** (dedicated accelerators, Quantum Computing…)

DEPLOYMENT

Optimise across the **AI lifecycle**

Optimise workloads **definition, scheduling, …**

Data centers choice: centralisation for better resource managament

AI ARCHITECTURES

Improved/ compactified **data representation and computational graphs**

Foundation models

New **approaches to training**

Neural Architecture Search (NAS)

HEP REFERENCES

[1] CMS Collaboration, “Reweighting of simulated events using machine learning techniques in CMS”, CMS Physics Analysis Summary CMS-PAS-MLG-24-001, 2024.
[2] ATLAS Collaboration, “A simultaneous unbinned differential cross section measurement of twenty-four Z+jets kinematic observables with the ATLAS detector”, 2024. arXiv:2405.20041. Submitted to *Phys. Rev. Lett.*
[3] CMS Collaboration, “2024 Data Collected with AXOL1TL Anomaly Detection at the CMS Level-1 Trigger”, CMS Detector Performance Note CMS-DP-2024-059, 2024.
[4] ATLAS-SIM-2023-004
[5] EPJ Web of Conferences 295, 09029 (2024)
[6] EPJ Web of Conferences 295, 03001 (2024)
[7] ATL-PHYS-PROC-2024-081