

The logo consists of the letters 'E4' in a bold, white, sans-serif font. The 'E' and '4' are connected at the top. Below the letters is a thin white horizontal line.

E4

COMPUTER
ENGINEERING

How not to Decelerate your HPC Workloads with Quantum Computing

Gabriella Bettonte & Roberto Rocco– E4 Computer Engineering

23 rd April 2026

WHO IS E4

BASED IN ITALY IN THE DATA VALLEY AND ACTIVE SINCE 2002, WE HAVE BEEN DESIGNING AND PROVIDING
END-2-END SOLUTIONS FOR HPC, ARTIFICIAL INTELLIGENCE AND QUANTUM COMPUTING.

WITH A STRONG PRESENCE IN THE ACADEMIC AND ENTERPRISE MARKETS, WE HAVE BEEN COLLABORATING FOR YEARS WITH THE
MAIN RESEARCH CENTERS AT NATIONAL AND INTERNATIONAL LEVEL.



HPC

For more than 20 years, E4 has been innovating and delivering HPC hardware and software solutions for hundreds of customers from the enterprise and the research centers, such as CINECA, CERN, ECMWF, LEONARDO



AI

Building of the infrastructures for the raising AI technologies, helping customers to face and drive the convergence between HPC & AI.
Development of vertical GenAI solutions for the secure on-premise scenarios

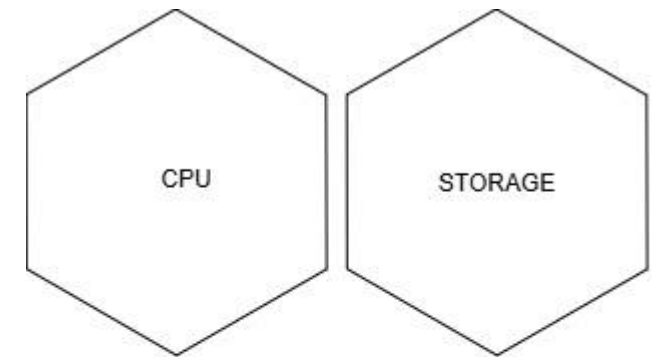


QUANTUM

Pioneering the Quantum Computing technologies, we are involved in several research projects by the European Union, offering our customers expertise for emulation and algorithm optimization

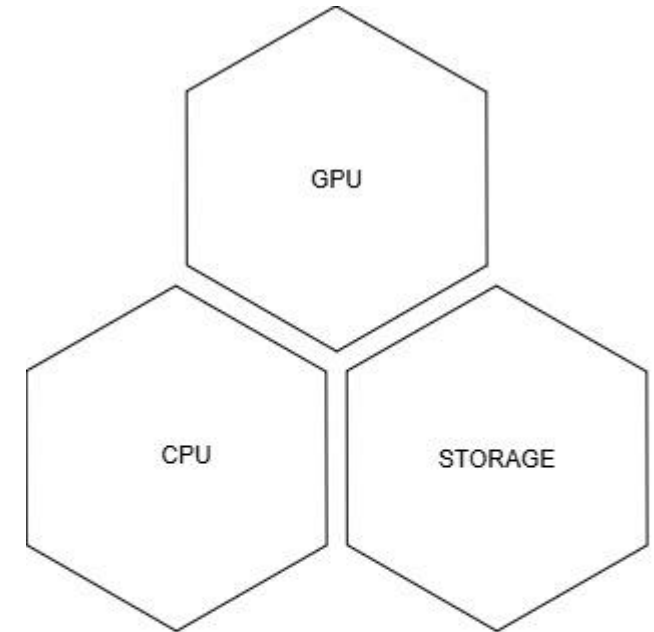
THE FUTURE IS HETEROGENEOUS

Heterogeneous computing refers to the use of multiple types of processors within a single system.



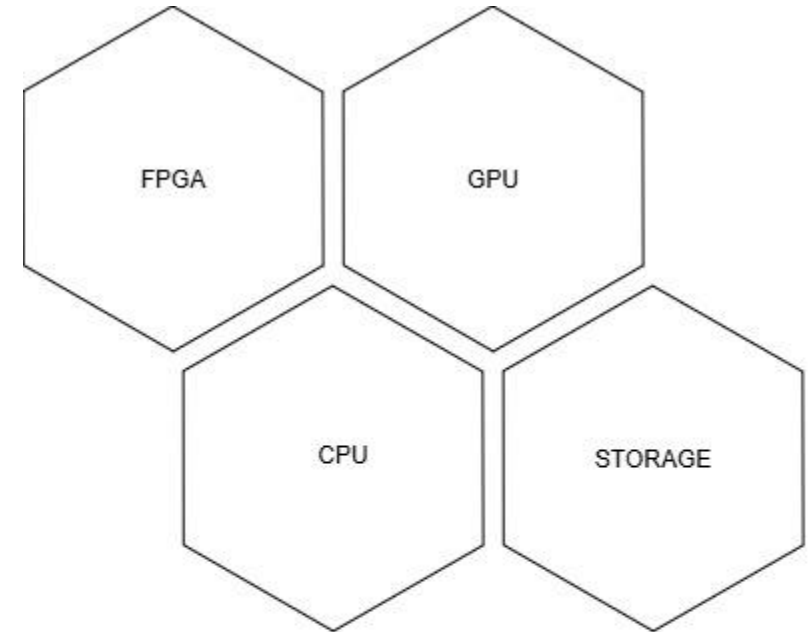
THE FUTURE IS HETEROGENEOUS

Heterogeneous computing refers to the use of multiple types of processors within a single system.



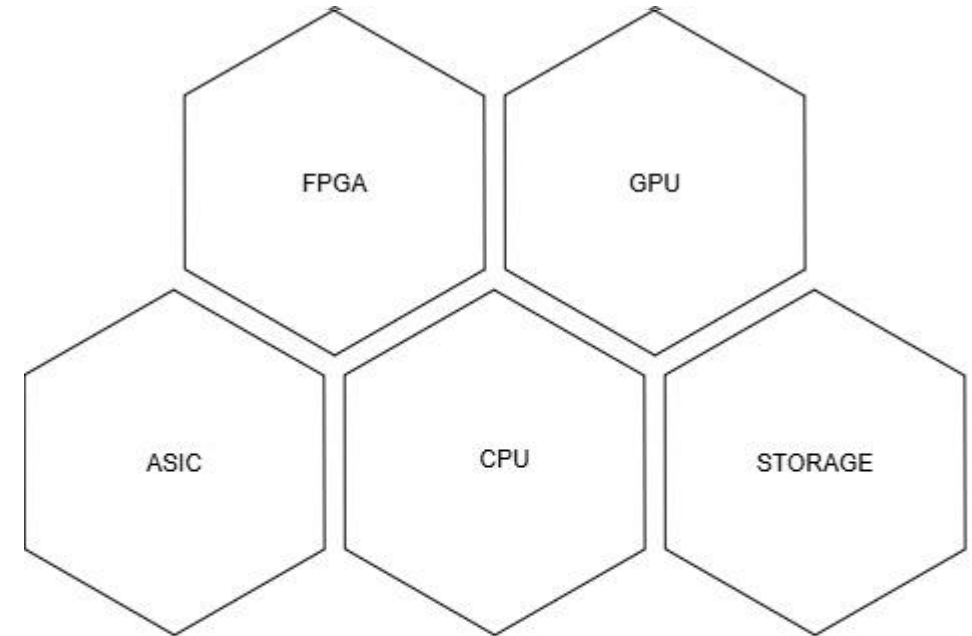
THE FUTURE IS HETEROGENEOUS

Heterogeneous computing refers to the use of multiple types of processors within a single system.



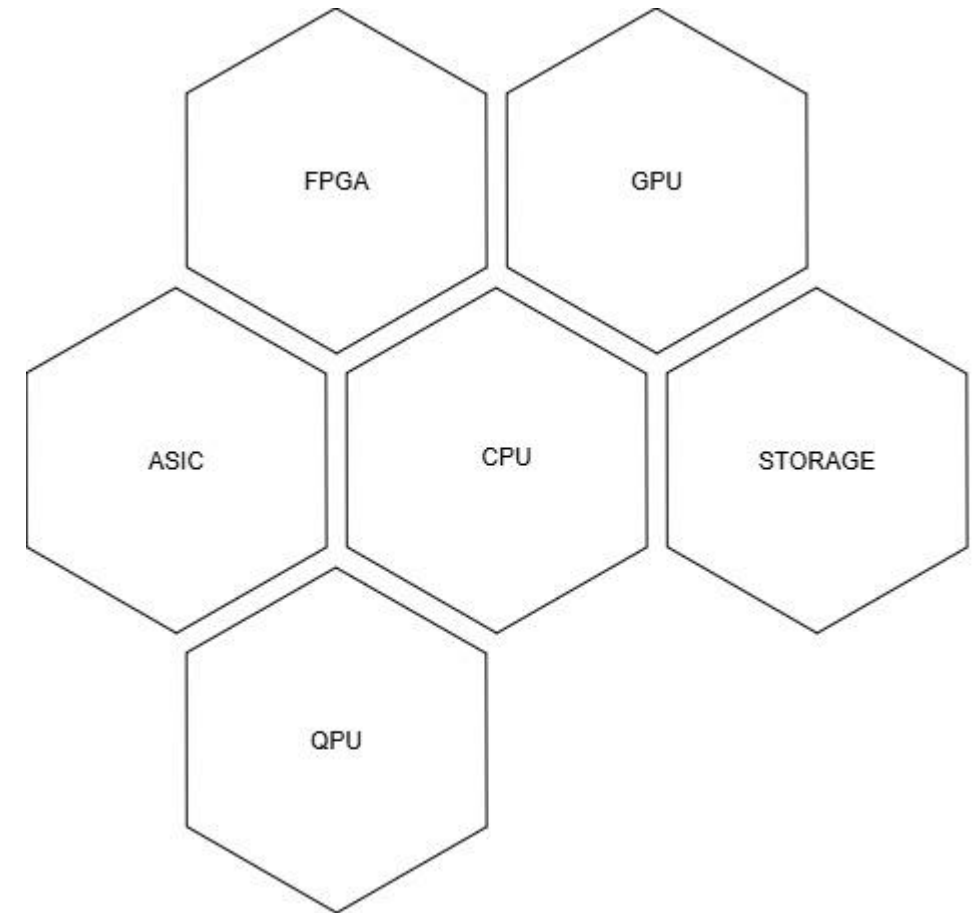
THE FUTURE IS HETEROGENEOUS

Heterogeneous computing refers to the use of multiple types of processors within a single system.



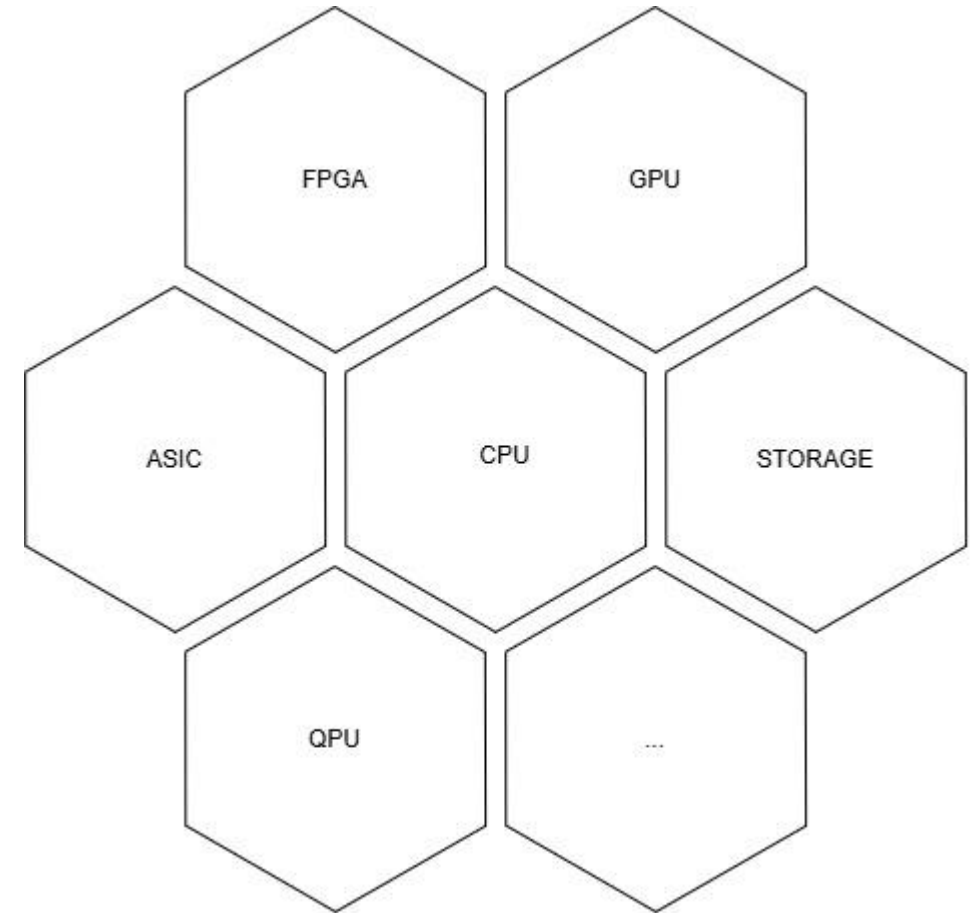
THE FUTURE IS HETEROGENEOUS

Heterogeneous computing refers to the use of multiple types of processors within a single system.



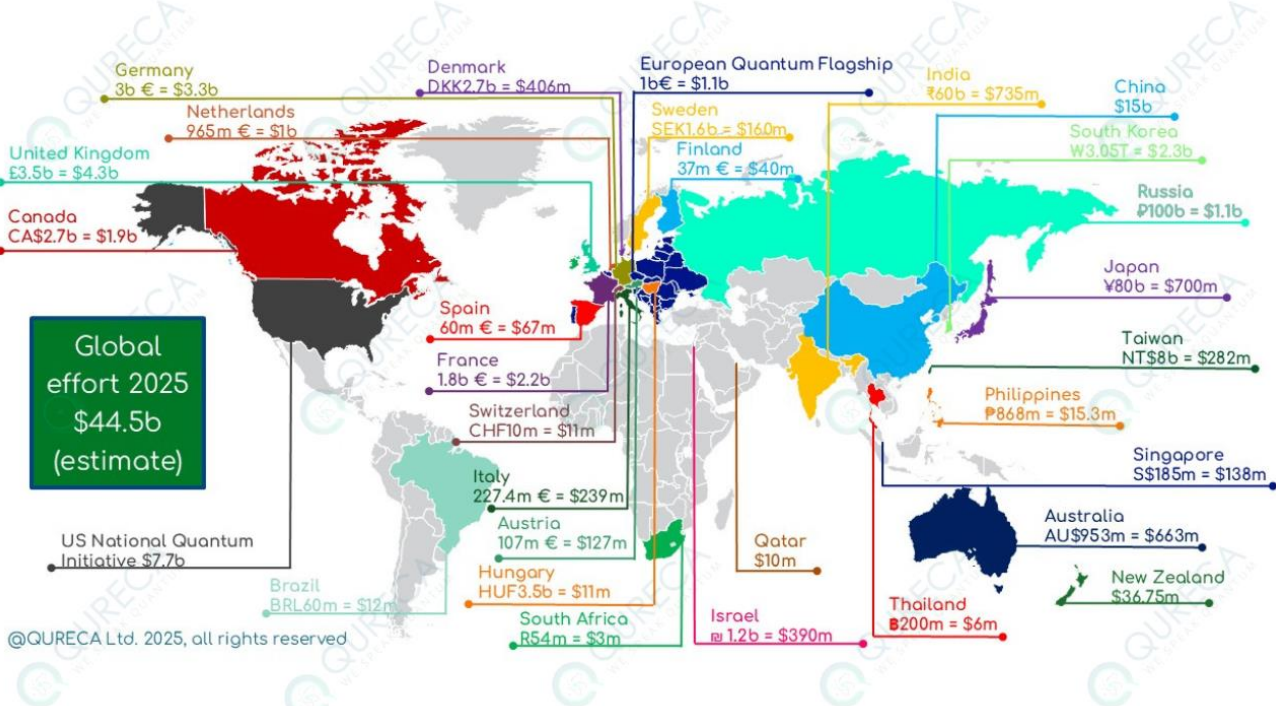
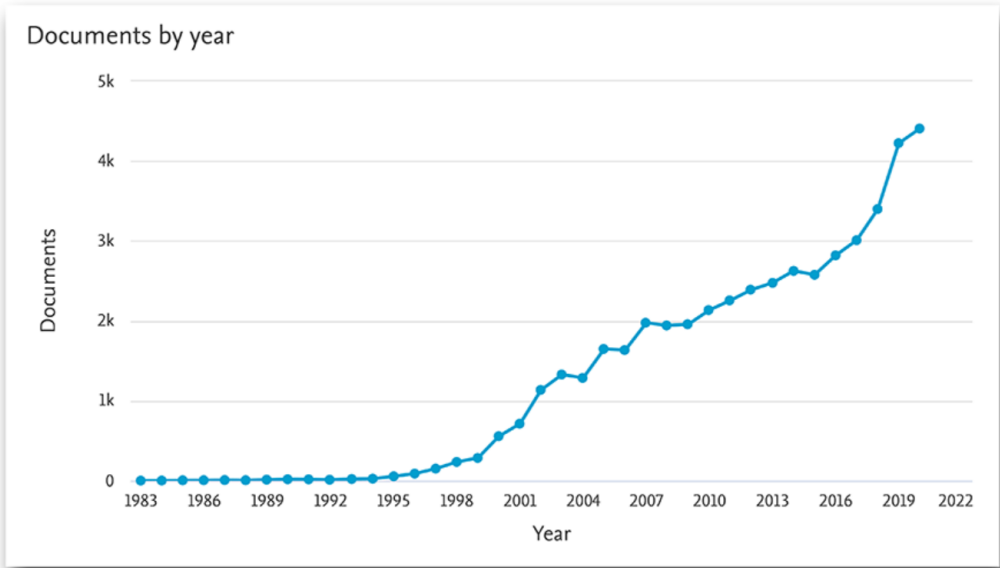
THE FUTURE IS HETEROGENEOUS

- Each processing unit type is **optimized for specific tasks**, allowing workloads to be assigned to the most suitable processor.
- This approach enhances both **performance** and **energy efficiency**. Different processors work simultaneously on various parts of a task, significantly reducing computation time.



QUANTUM EFFORT WORLDWIDE IS INCREASING

Publication rate on quantum computing has steeply increased



<https://www.elsevier.com/solutions/scopus/research-and-development/quantum-computing-report>

<https://www.quireca.com/quantum-initiatives-worldwide/>

QUANTUM EFFORT WORLDWIDE IS INCREASING

EDITOR'S NOTE



BY HARRY GOLDSTEIN

The Coming Quantum Boom

A century after quantum mechanics was described, a vibrant industry blooms

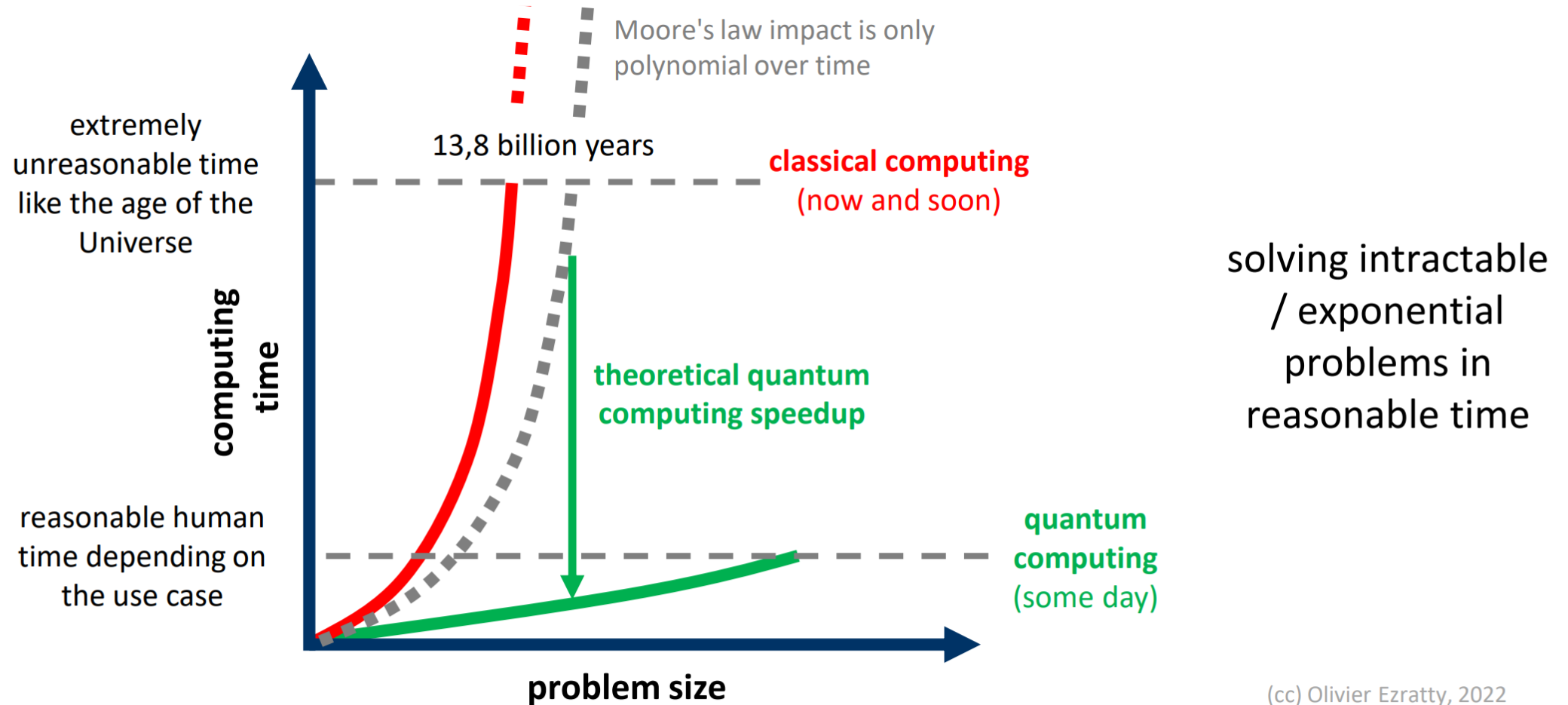


$\hat{H}\psi(\vec{r}, t) = E\psi(\vec{r}, t) = \sum_n c_n e^{-i\omega_n t} \psi_n(\vec{r})$

$\psi(\vec{r}, t) = \sum_n c_n \exp\left(\frac{i\hat{H}t_0}{\hbar}\right) \psi_n(\vec{r})$

$\psi(\vec{r}, t) = T_0 \exp(-i\omega t) \psi_0(\vec{r}) + \psi_2(\vec{r}) + 2\text{Re}[\psi_1(\vec{r})] + c_2 \phi_2(\vec{r}) + c_1^* c_2 \phi_1(\vec{r})$

WHY QUANTUM COMPUTING?



(cc) Olivier Ezratty, 2022

Figure 6: simplified view of the quantum computing theoretical promise. Before delivering this promise, quantum computers may bring other benefits like producing better and more accurate results and/or doing this with a smaller energy footprint. (cc) Olivier Ezratty, 2022.

QUANTIZED COLUMNS

What Makes Quantum Computing So Hard to Explain?

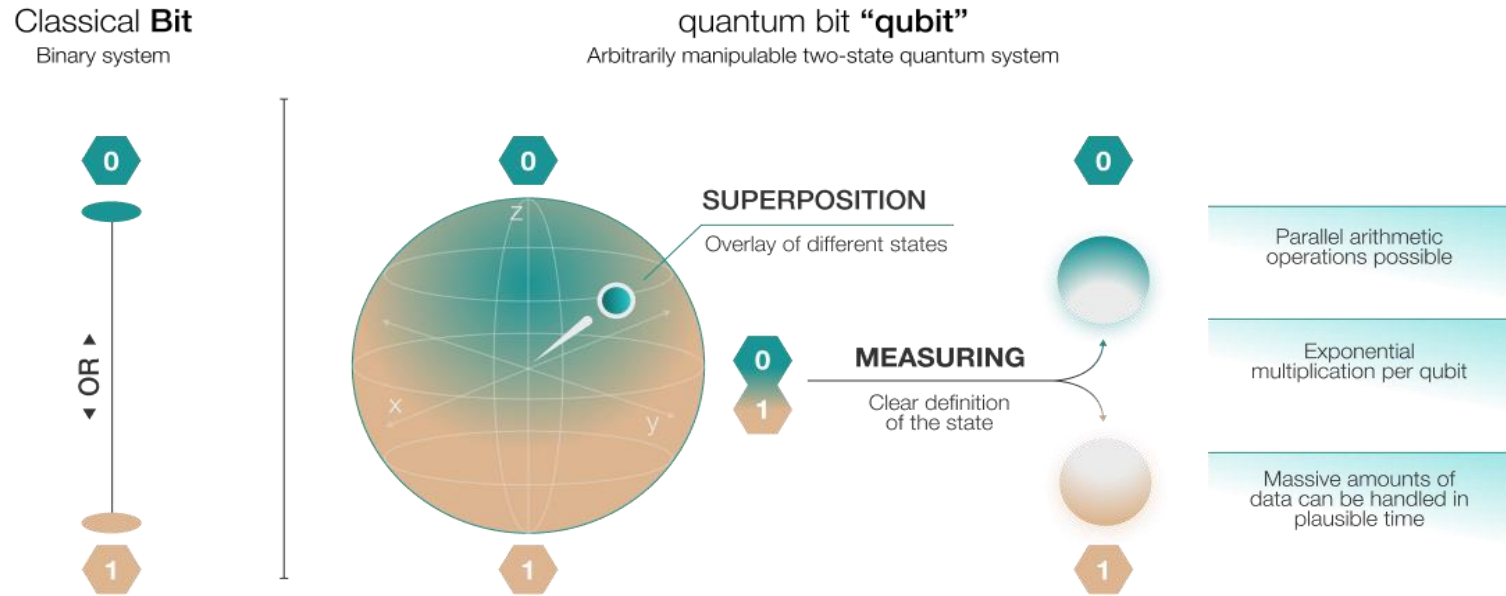
 22 | 

To understand what quantum computers can do — and what they can't — avoid falling for overly simple explanations.

“Quantum computers aren't the next generation of supercomputers — **they're something else entirely.**” Scott Aaronson

→ we need to understand the fundamental physics that drives the theory of quantum computing.

HOW IT IS DIFFERENT - QUBITS & SUPERPOSITION



A qubit is the basic unit of information used to encode data in quantum computing.

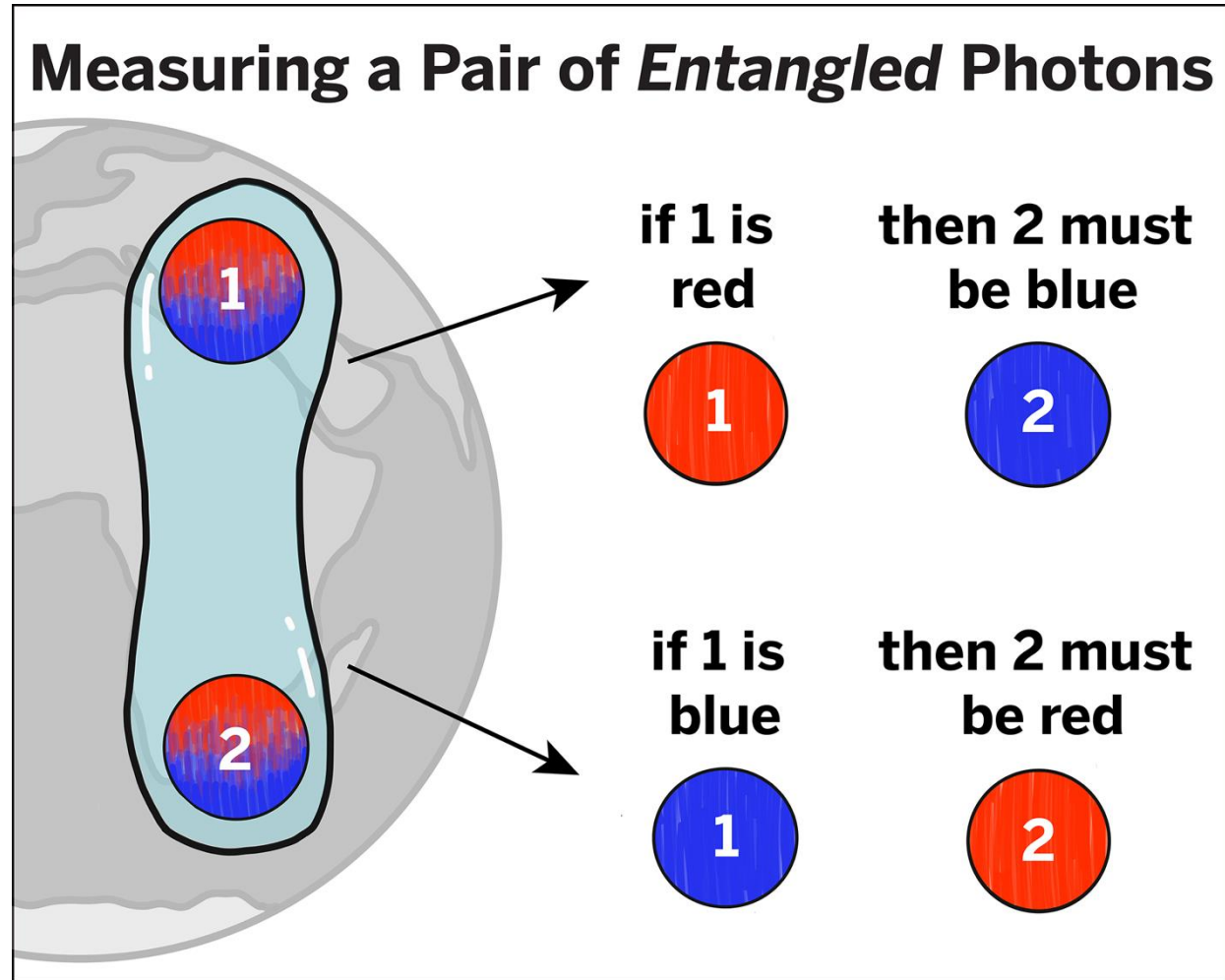
Each qubit has two pure state vectors $|0\rangle$ and $|1\rangle$.

Qubits can also occupy a third state known as a **superposition** → all the possible states between 0 and 1 (probability of the qubit's state) → MORE INFORMATION

Qubits are created by manipulating and measuring quantum systems, for instance: photons, electrons, trapped ions, superconducting circuits, and atoms.

HOW IT IS DIFFERENT - ENTANGLEMENT

two qubits are intertwined in such a way that the state of one particle cannot be described independently of the state of the other, regardless of the distance between them



QUANTUM COMPUTERS

There is not just one category of quantum computers, but many!

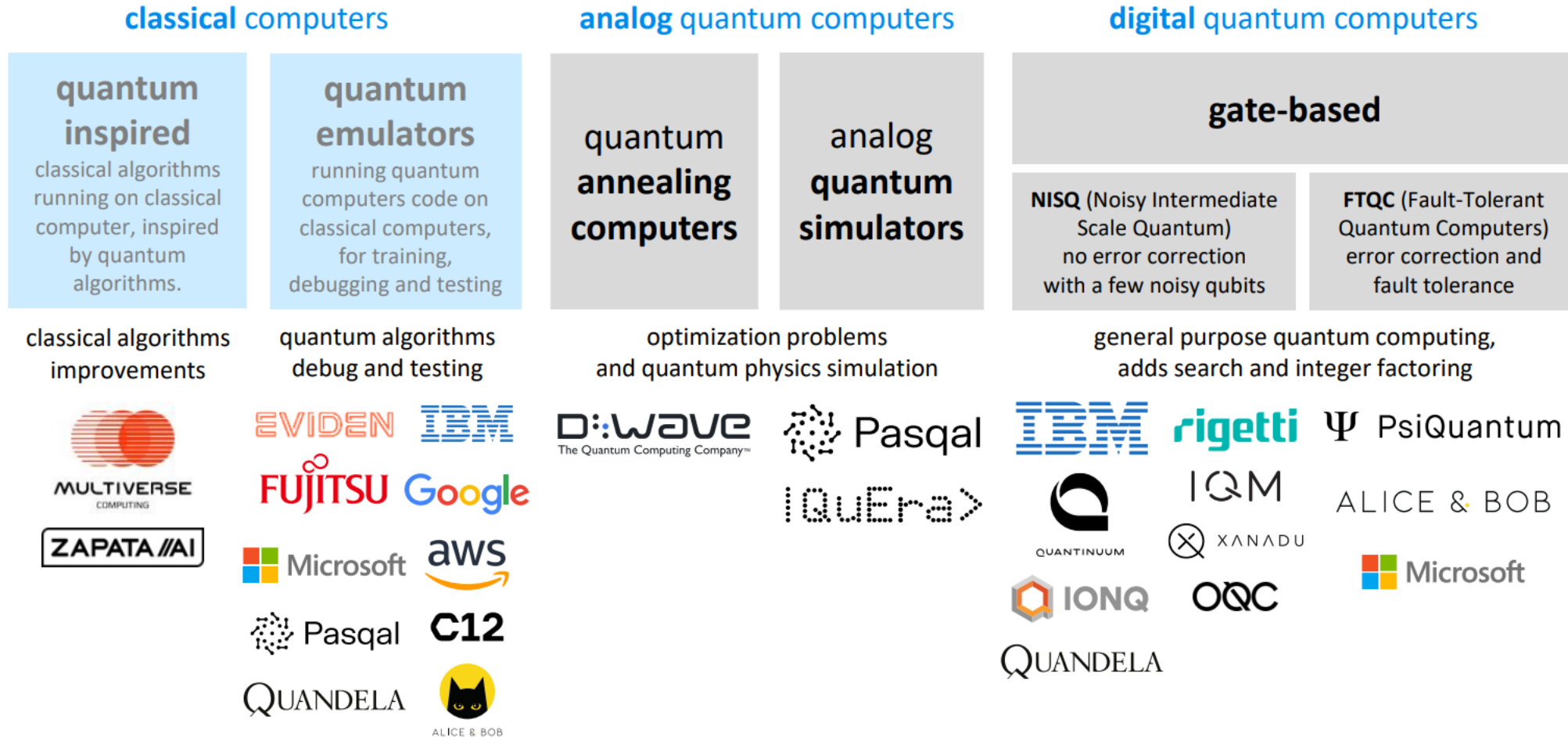


Figure 224: the different computing paradigms with quantum systems, hybrid systems and classical systems. (cc) Olivier Ezratty, 2022-2024.

EMULATING QUANTUM COMPUTERS

JUPITER has about 6000 nodes (24 000 superchips). The largest number of nodes that can actually be used for JUQCS is 4096 (16 384 superchips) due to the powers-of-two restriction. Each of the 16 384 superchips is equipped with 96 GiB of HBM3 (device) and 120 GiB of LPDDR5 (host) memory; 216 GiB in total per GH200 superchip.

→ reach the emulation 50-qubit on
JUPITER

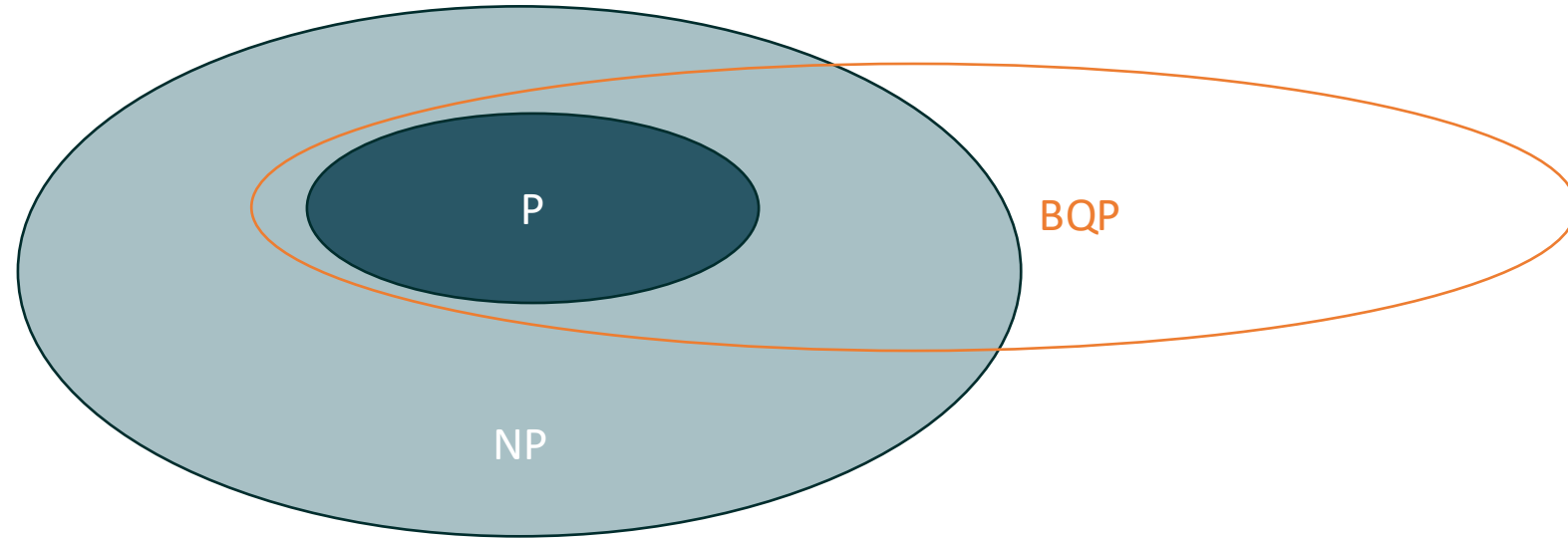


ONLY CERTAIN CLASSES OF PROBLEMS ARE FOR QC

P: Polynomial time

NP: Non deterministic
polynomial time

BQP: Bounded error
quantum polynomial



The class of problems that can be solved efficiently by a quantum computer are BQP.

To this day, researchers have designed only a few quantum algorithms that provide a speed-up from exponential to polynomial time for a problem.

QUANTUM EVOLUTION AND MEASUREMENTS

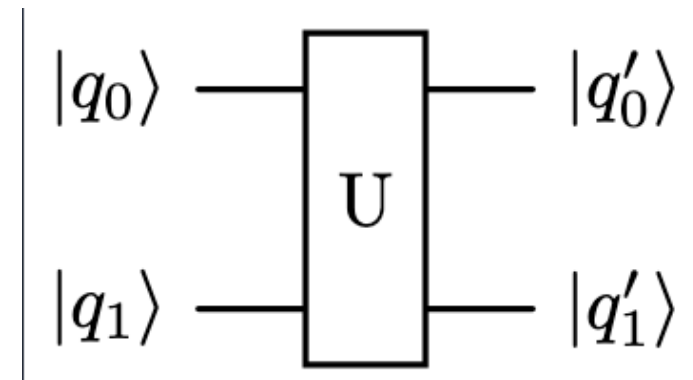
INITIALIZATION In a Quantum Algorithm, the input data of the algorithm is loaded on the quantum register

Then this initial state is **evolved under unitary gates** (U s.t. $U^\dagger U = I$).

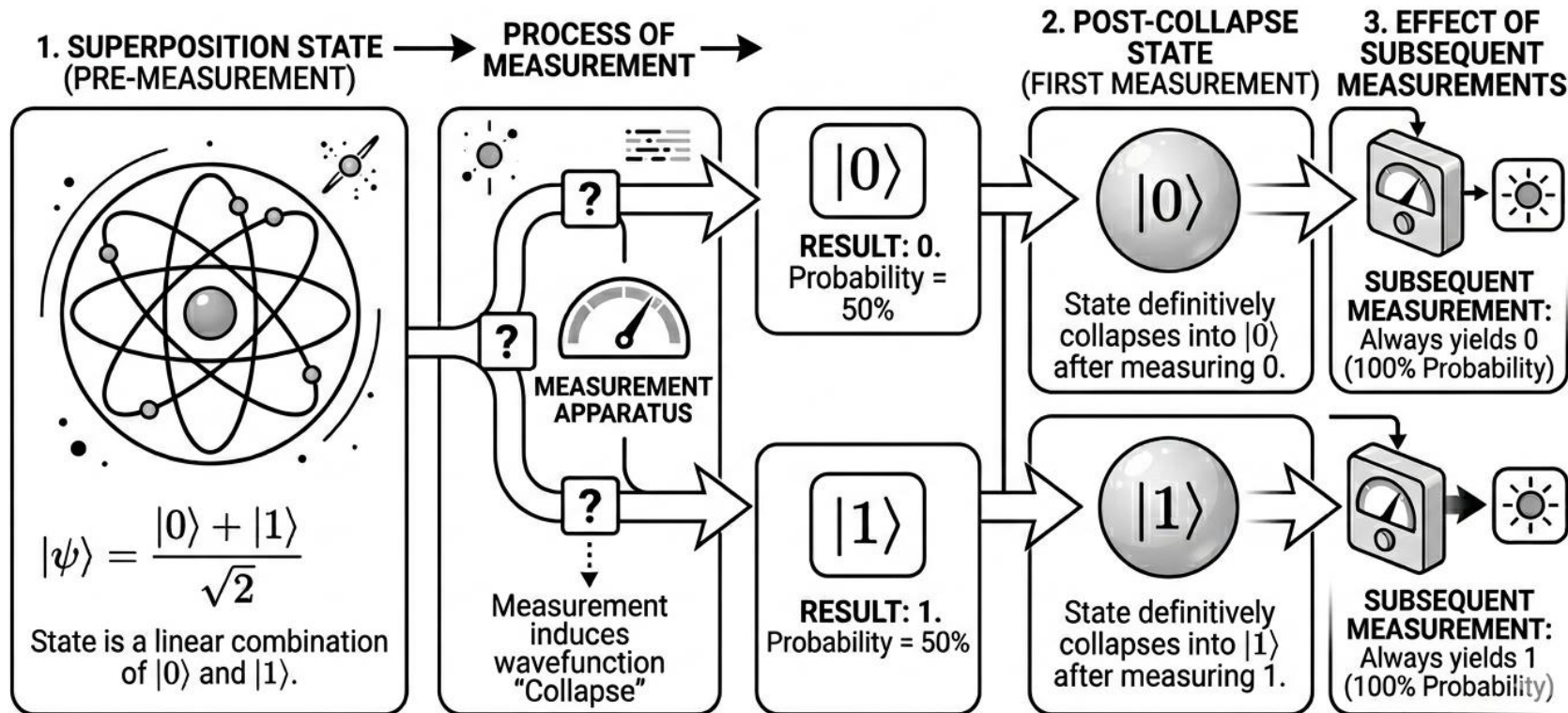
After the evolution, the qubits are measured.

Measurement is probabilistic.

Suppose to have a qubit in superposition $|\psi\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$.



EXPLANATION OF QUANTUM STATE MEASUREMENT AND WAVEFUNCTION COLLAPSE



Probabilistic measures imply that one has to run the quantum evolution multiple times to extract a probability distribution.

In practice the number of samples one has to do could be a drawback of a quantum algorithm.

QUANTUM ERRORS

- Quantum states are affected by *quantum decoherence*. Over time, a quantum state can lose information (entanglement, superposition).
- This effect, in a quantum circuit, produces *quantum errors*, with the effect that a quantum computation is not fully reliable.
- This problem exists also in classical computation, where bit flip errors may occur. Classical Error Correction relies on redundancy, namely copying the bits. This cannot be done in the quantum realm, where copying unknown states is not allowed by the *No Cloning Theorem*.
- *Quantum Error Correction* studies how to correct such errors, however introducing an overhead in the number of *physical qubits* needed to construct a *logical (error free) qubit*.

THE (BRIGHT) PRESENT - HARDWARE

~ 5000 qubits



~ 150 qubits



~ 50 qubits



NOW

'90

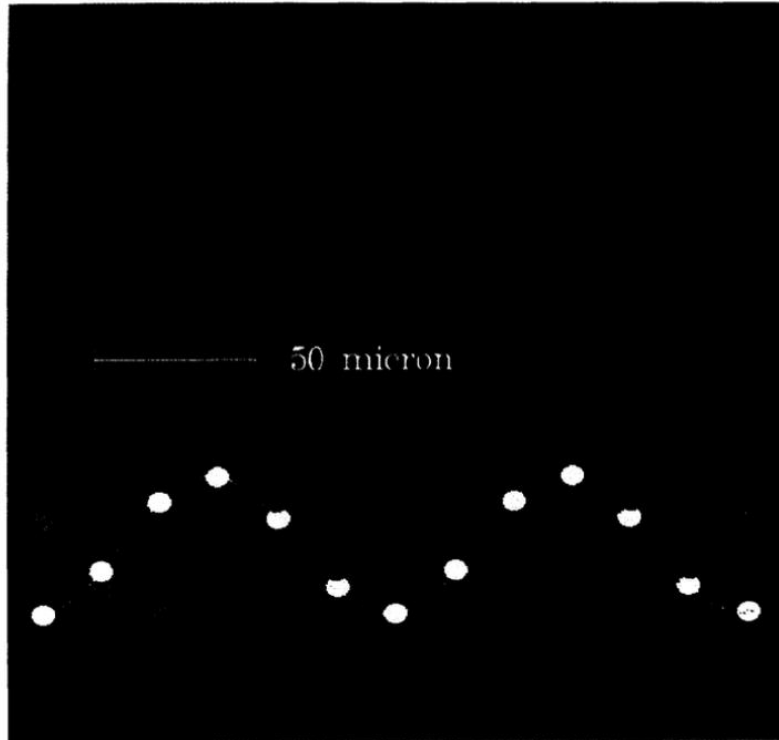
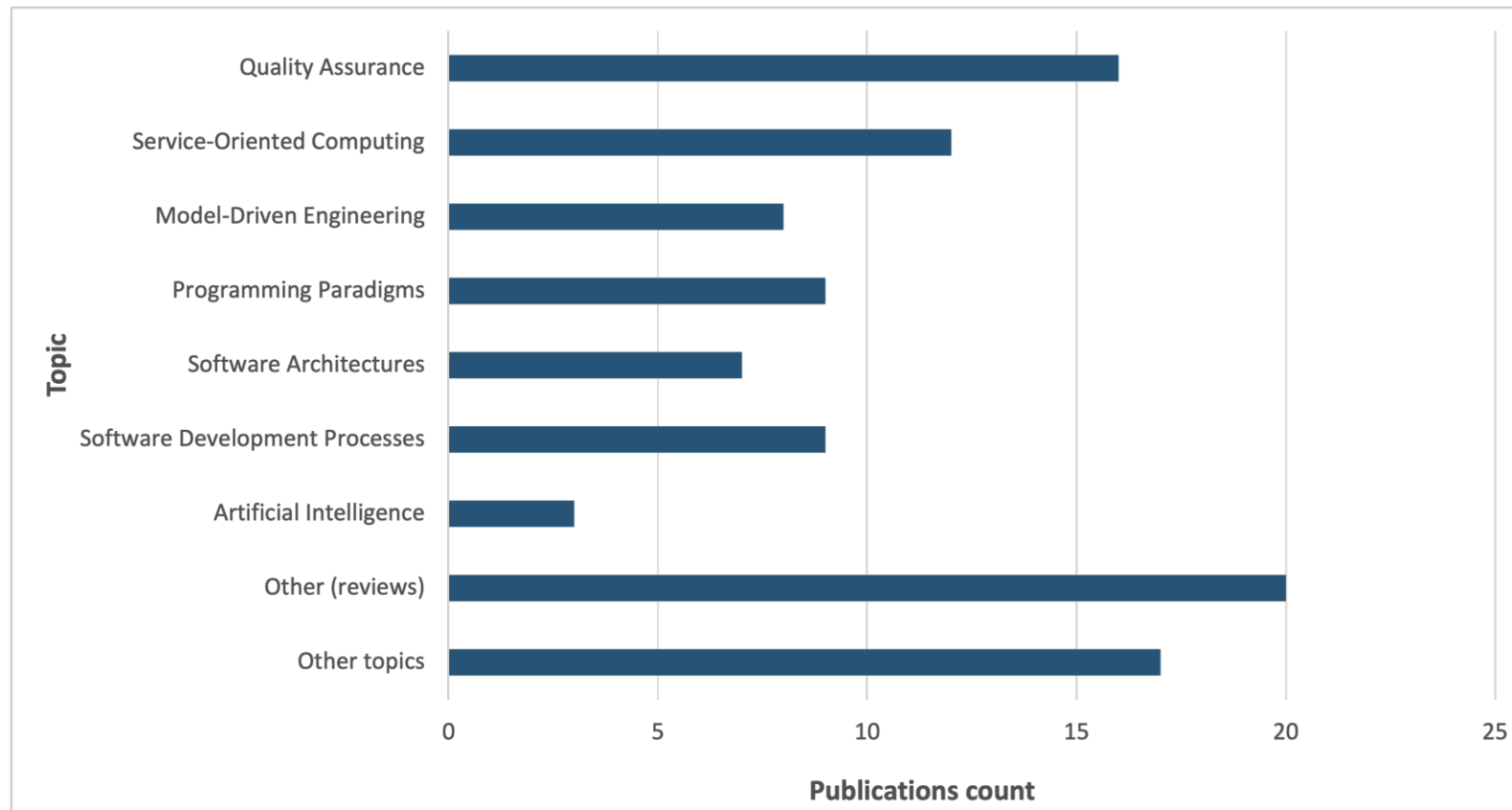


FIG. 6. Helical structure of $^{24}\text{Mg}^+$ ions with a diameter of $63 \pm 2 \mu\text{m}$. The experimental image (top) corresponds to three interwoven helices (shown in different colors, bottom). The closely appearing pairs of ions are sitting on opposite sites, resulting in twice the intensity at those positions ($\psi_0 = 1.1 \text{ eV}$).

Observation of Ordered Structures of Laser-Cooled Ions in a Quadrupole Storage Ring, I. Waki et al, 1992

THE (BRIGHT) PRESENT - SOFTWARE

Many research initiatives on software to ensure this software is efficient, maintainable, reusable, and cost-effective → **Quantum Software Engineering as a distinct field**



THE (BRIGHT) PRESENT - ALGORITHMS

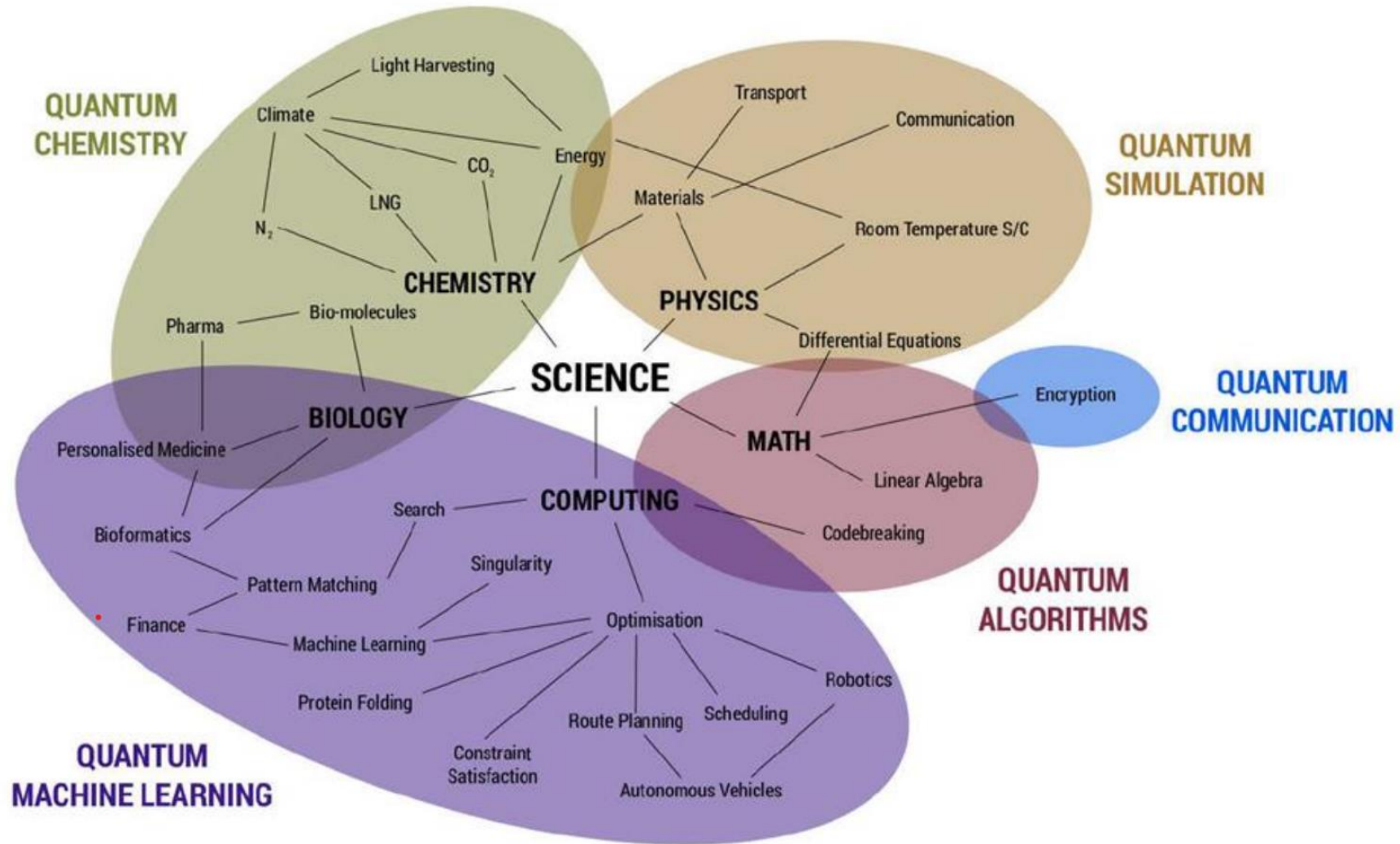


Figure 551: the breadth of science domains covered by quantum algorithms. Source: [Silicon Photonic Quantum Computing](#) by Syrus Ziai, PsiQuantum, 2018 (72 slides).

IS THE FUTURE OF QC MODULAR?

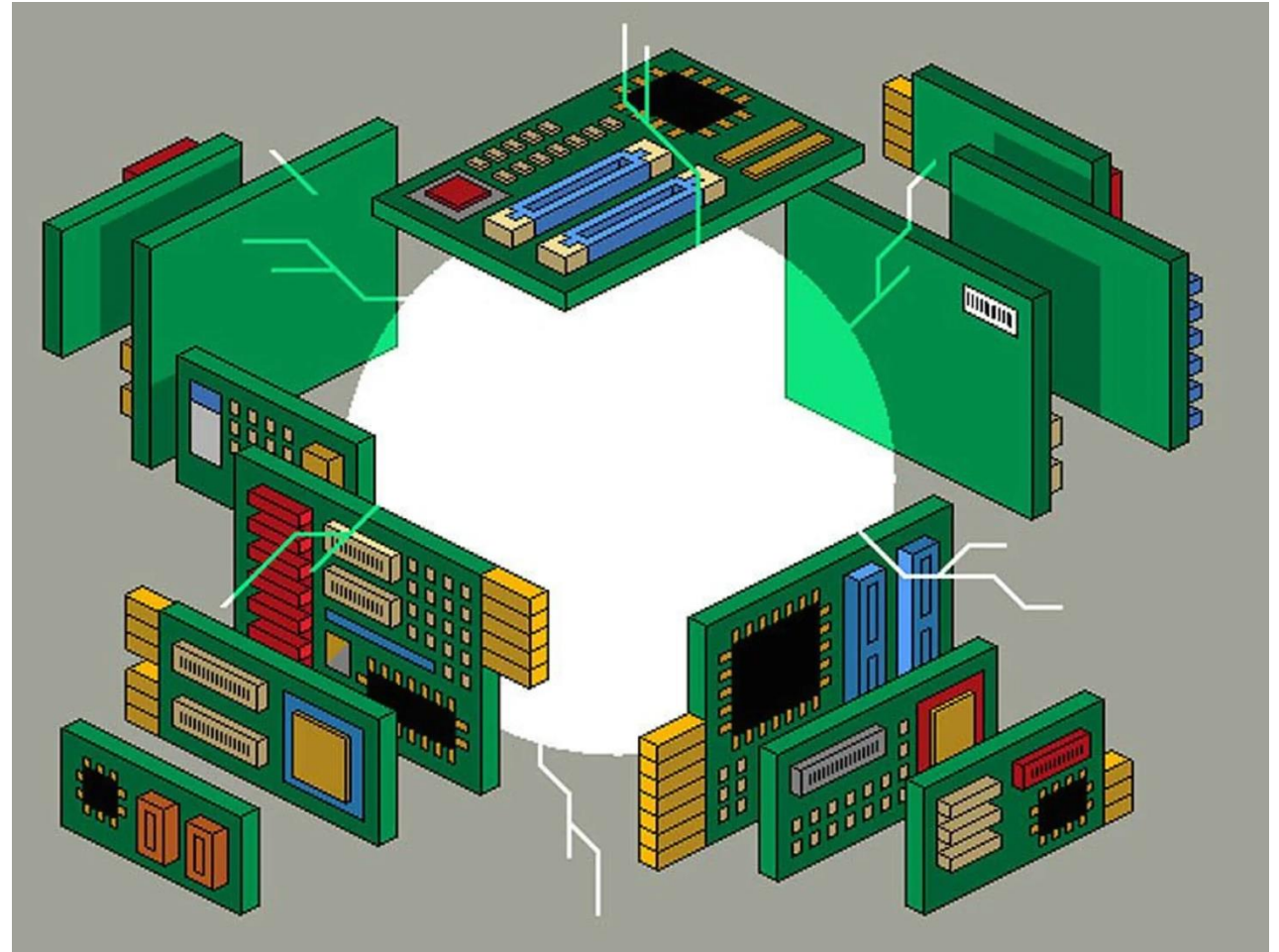
Quantum computing companies have been competing for years to squeeze the most qubits onto a chip. There are limits to the chip size. As they grow larger, wiring up the control electronics becomes increasingly challenging.

PRO: Computers with smaller, testable, and replaceable components simplifies manufacturing and maintenance.

The focus is now shifting to linking multiple quantum processors.

Challenges:

- maintaining entanglement across QPUs
- Physical networking between QPUs
- Running quantum algorithm across QPUs (compilers)



WHY HPC-QC INTEGRATION

HPC pursues performance and computation efficiency

QC delivers better results for some problems, and consumes way less than a GPU powered cluster

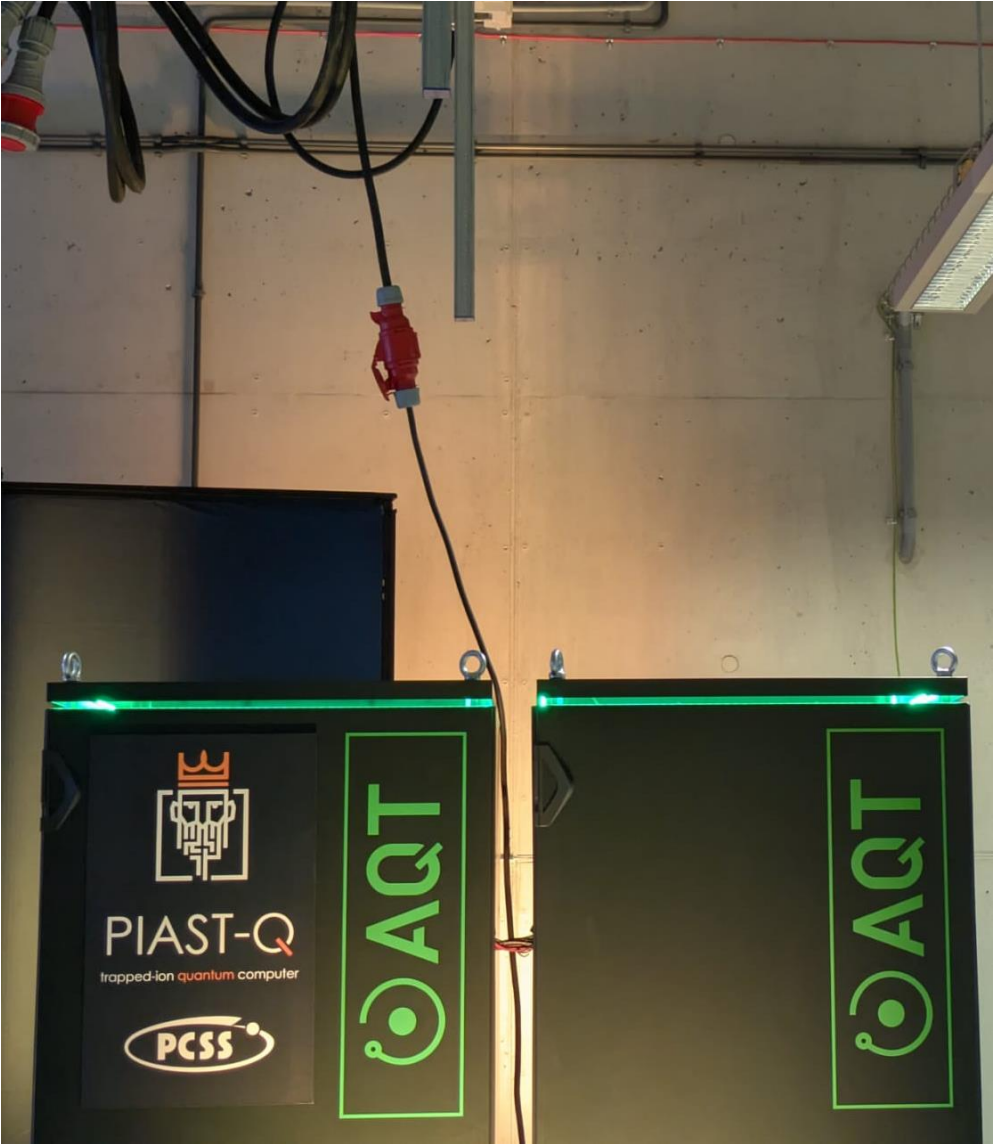
Additionally, QC may need HPC resources with low latency to perform Quantum Error Correction

Why not join forces? Why not having HPC and QC resources together?

- No clash between HPC and QC
- Cooperation to achieve best performance and efficiency
- High speed interconnections to ensure efficient data transfer
- High Performance hardware available for QEC

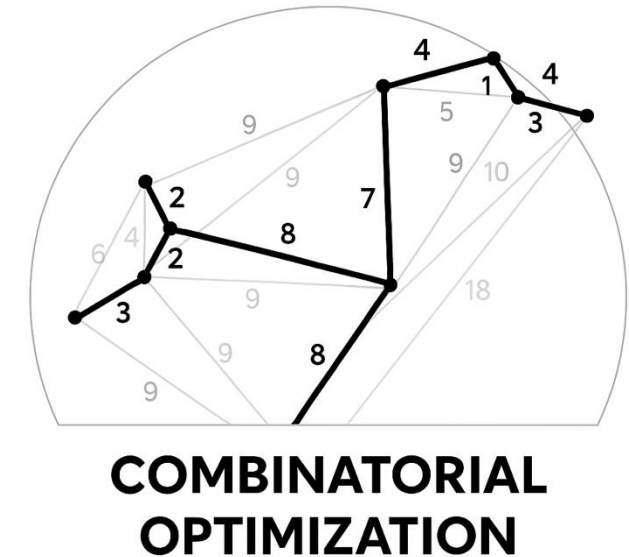
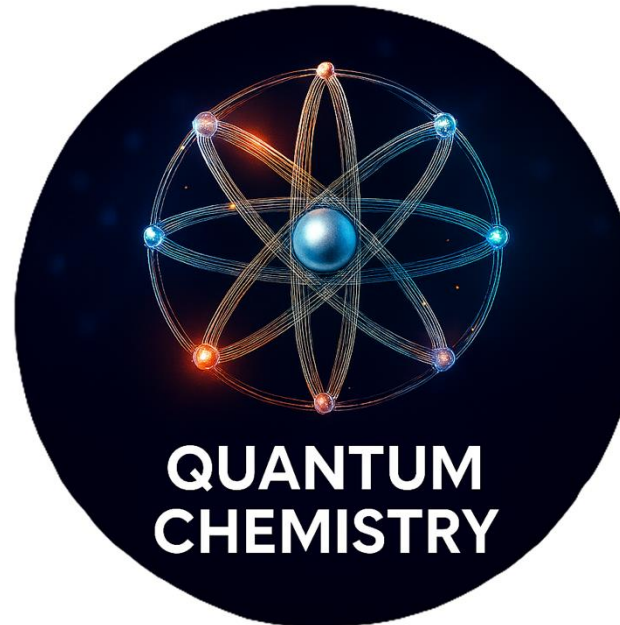
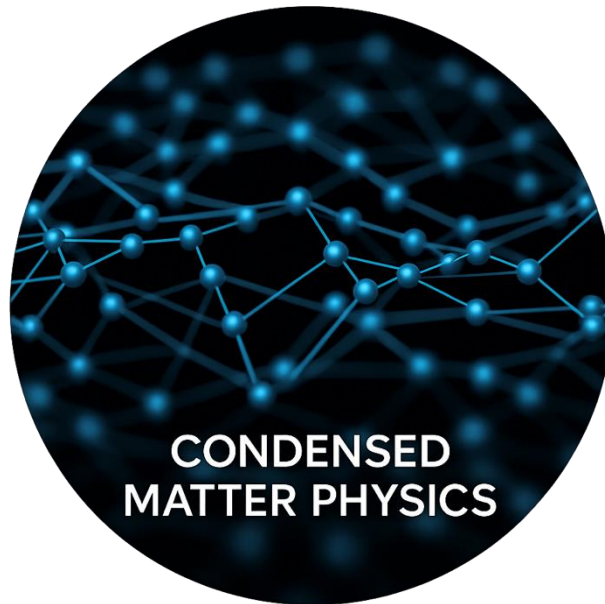
This solution is promising, but we need to clarify roles and bridge the gap between classical and quantum world

ENERGY CONSUMPTION VISUALISED



APPLICATIONS

The synergy between Quantum Computing and HPC holds significant potential to accelerate progress across multiple scientific domains.



The literature already reports concrete implementations of hybrid algorithms evaluated both on large-scale HPC simulators and on early HPC–QC prototype platforms → **STILL, MANY OPEN QUESTIONS!**

QUANTUM COMPUTERS AS ACCELERATORS

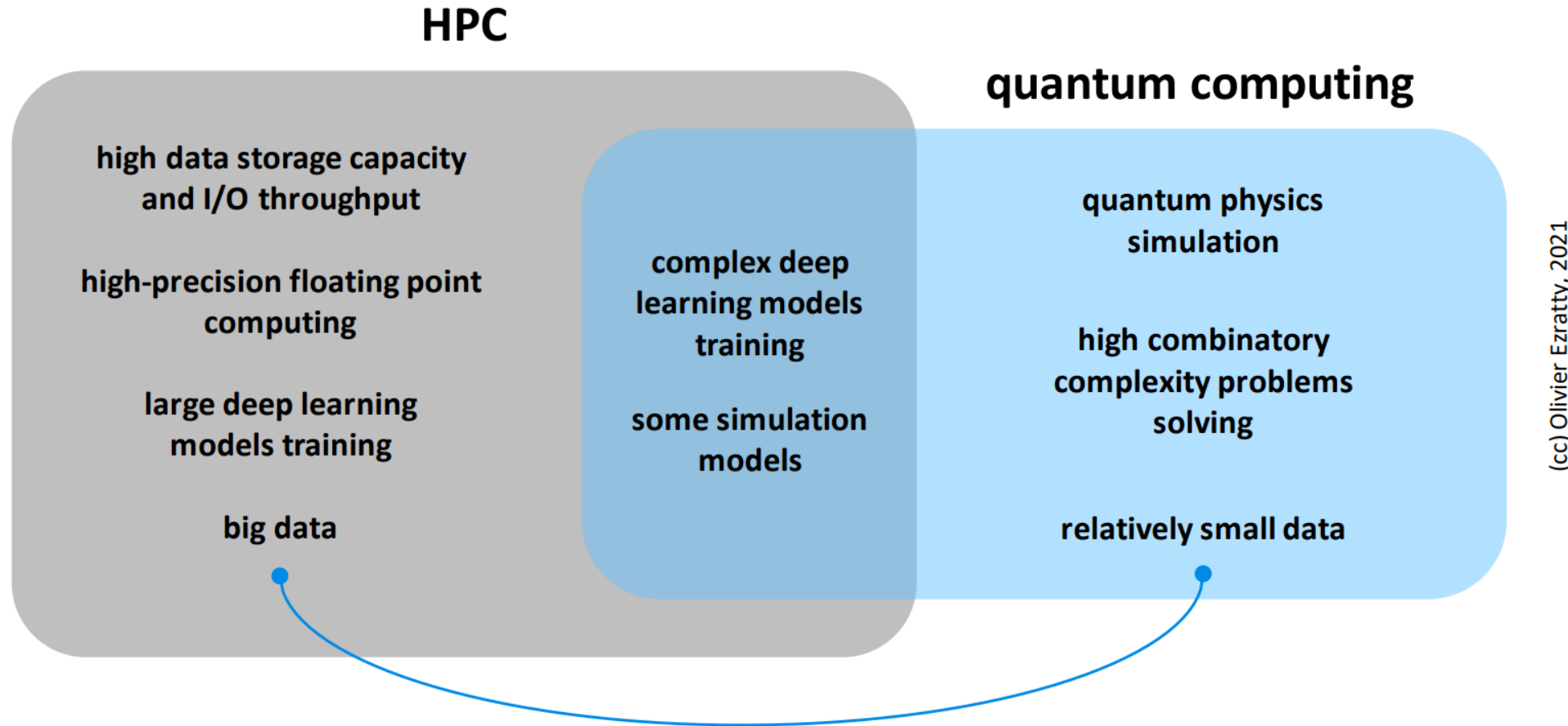


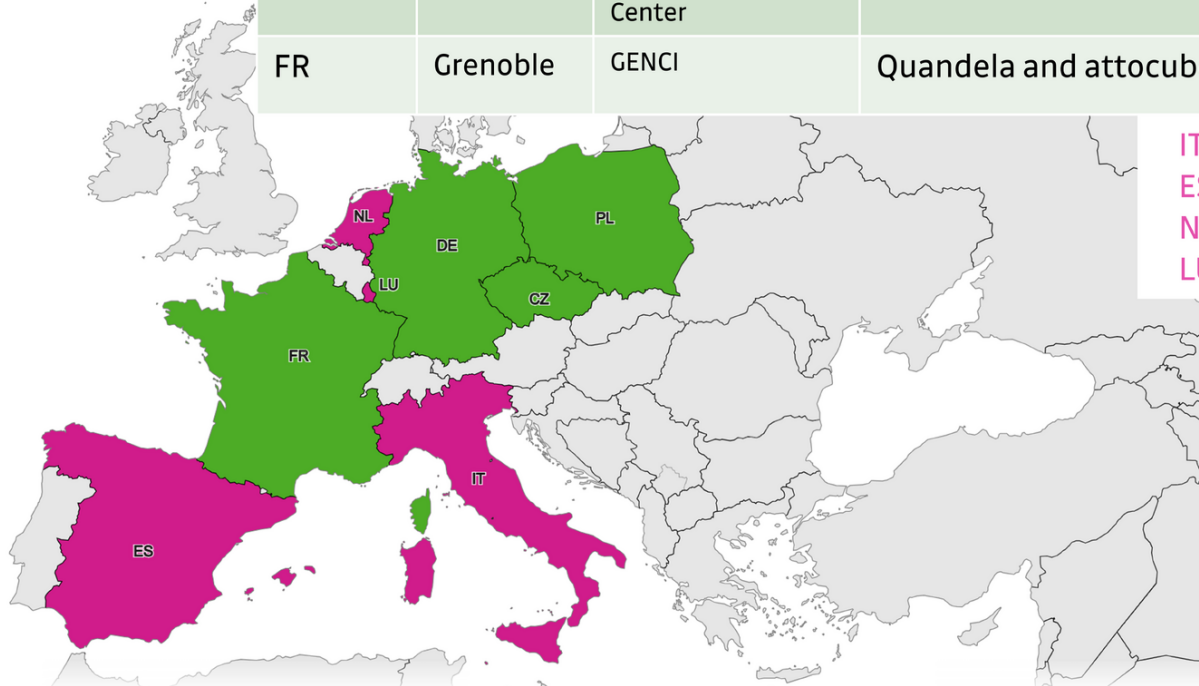
Figure 590: how to position HPCs vs (scalable) quantum computers. HPCs are for big data and high-precision computing. Quantum computing will be adapted to high complexity problems but with relatively reasonable amounts of data. There's some cross-over between both systems and they will work in sync in many cases. (cc) Olivier Ezratty, 2021.

IN EUROPE

HPC-QC integration is an active research topic, it is being supported by the European Commission through EuroHPC

EuroHPC JU Quantum Computers

Country	City	Hosting site	Vendor	Qubits
DE	Garching	Leibniz Supercomputing Centre	IQM	54 + 150
CZ	Ostrava	IT4Innovations	IQM	24
PL	Poznan	Poznan Supercomputing and Networking Center	AQT	20
FR	Grenoble	GENCI	Quandela and attocube	12



HPC

- Established technology, with standards and a consolidated methodology.
- Performance through parallelism and hardware control
- Heterogeneity (accelerators with different characteristics)
- Many nodes -> concurrent access

QC

- Novel solution, far from standardisation
- Performance using quantum properties
- Diverse technologies, each with own strenghts
- Rare resource, only few available
- Coherence times, error rates, limited Qubits

GOAL: A SMOOTH INTEGRATION

- Key observation: solutions to the integration challenge should not require HPC centers to overhaul their existing infrastructure or policies.
- HPC providers operate mature software stacks, including workload managers, schedulers, job accounting, and access control frameworks, that have been refined over decades.
- Any viable approach to integrate near-term, scarce, quantum resources within HPC facilities must be deployable within these constraints, augmenting rather than replacing what is already in place.

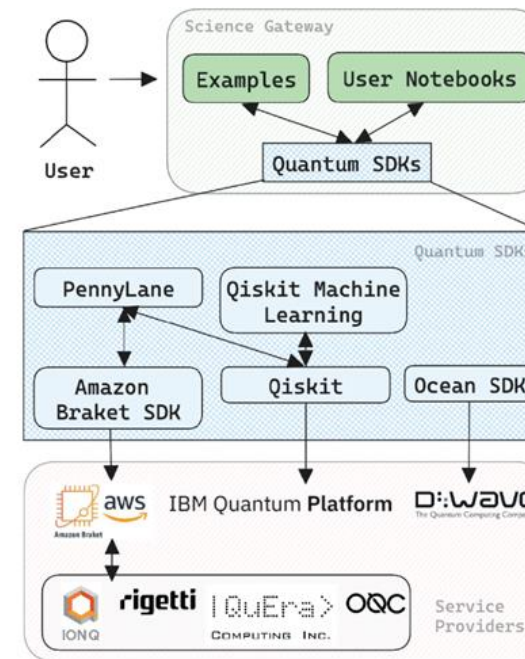
TWO DIRECTIONS

```
1 #!/bin/bash
2 #SBATCH --partition classical
3 #SBATCH --nodes 10
4 #SBATCH --time=01:00:00
5
6 #SBATCH hetjob
7
8 #SBATCH --partition quantum
9 #SBATCH --gres=gpu:1
10 #SBATCH --time=01:00:00
11
12 srun ./hybrid_job
```

HPC Solution: access through **SLURM**

Different timespans
Different availabilities

Co-scheduling blocks resources



© Marosi et al. [1]

Quantum Solution: access through **Cloud**

High Latency
High level integration, no tuning

Lose most benefits of HPC-QC proximity

QUANTUM COMPUTERS AS ACCELERATORS

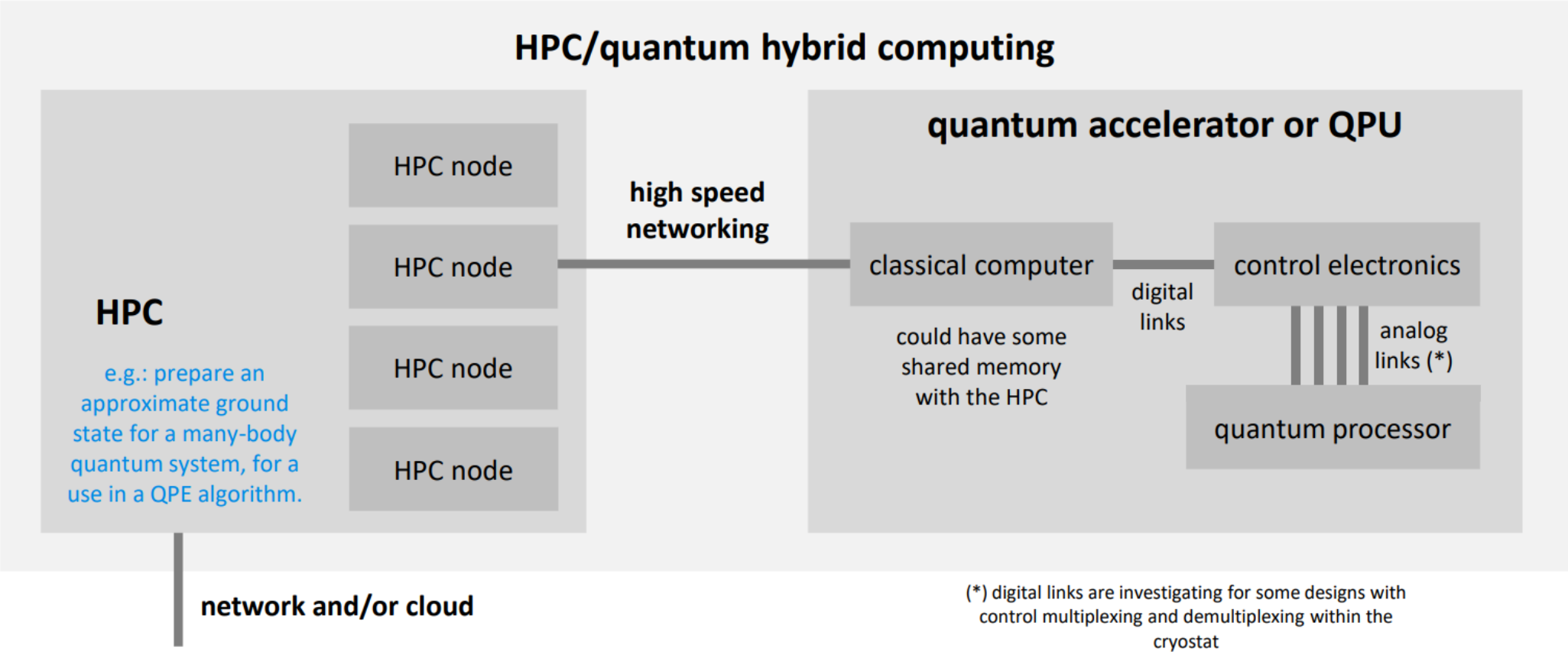


Figure 227: basics of a hybrid classical/quantum computing hardware architecture. (cc) Olivier Ezratty, 2021-2024.

CHALLENGES

Language differences:

- HPC focuses on low level languages: C, C++, Fortran
- QC mainly uses Python with dedicated libraries (e.g., Qiskit, Pulser)
- QC machine interactions happens through REST APIs, sequential execution

Maturity differences

- HPC uses highly tested and validated code, which does not change frequently
- QC TRL is way lower

Availability

- HPC clusters feature hundreds if not thousands of nodes
- QC systems are scarce, having one is a lot nowadays

Research Directions

- Actual interconnection between HPC and QC, analysing best practices
- Quantum resource management in an HPC-QC scenario

THE CHALLENGE OF RESOURCE CONTENTION

The imbalance between Classical and Quantum resources causes pain

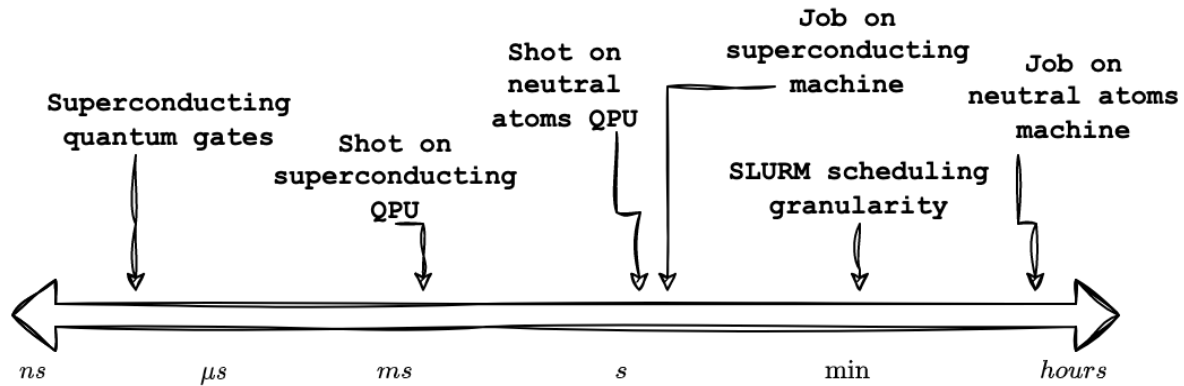
Consider a series of workloads executing part of their load on Classical resources and part on Quantum ones

Can you saturate the available resources?

- If I request both classical and Quantum resources, either is waiting for the other
- Moreover, if I block quantum resources without using them, I am worsening the situation on the bottleneck resource
- Result: bad experience for HPC-QC users, and worse utilisation in the HPC cluster, even for standard HPC jobs

SMARTHPC-QC

→ The SMARTHPC-QC project focus on the resource allocation issue



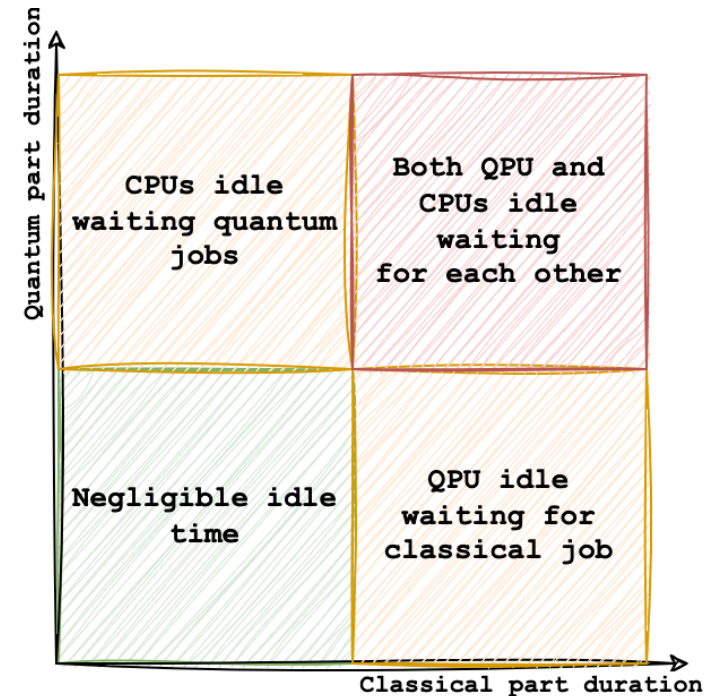
PROBLEM: Small amount of quantum computers compared to the number of HPC nodes + different QPUs have different execution time

OUR CLAIM: simple co-scheduling with exclusive QPU access is inadequate for achieving optimal resource utilization in heterogeneous HPC-QC environments

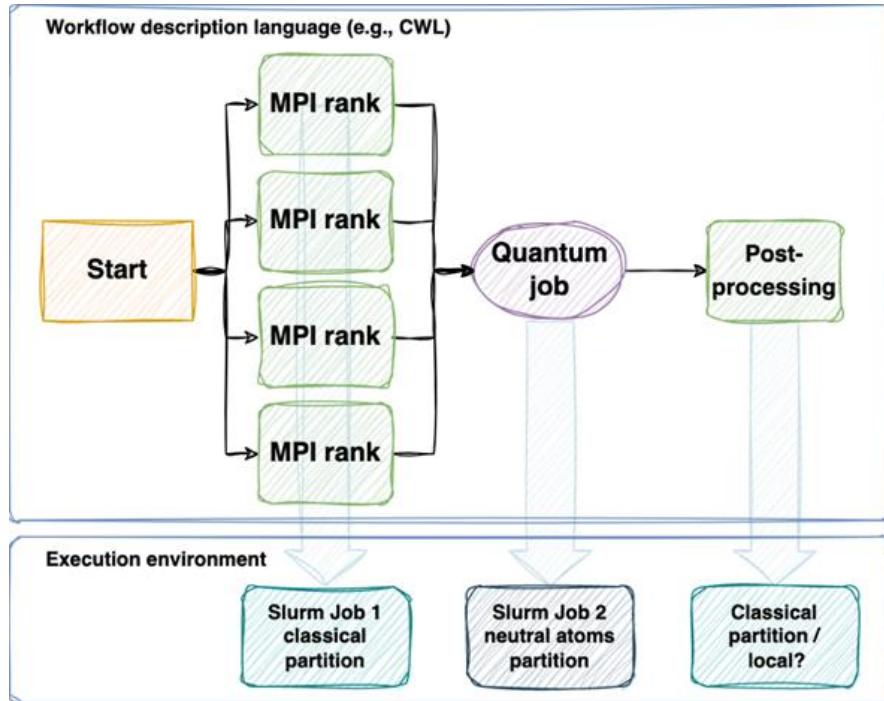
Lead by:



In collaboration with:



Workflows

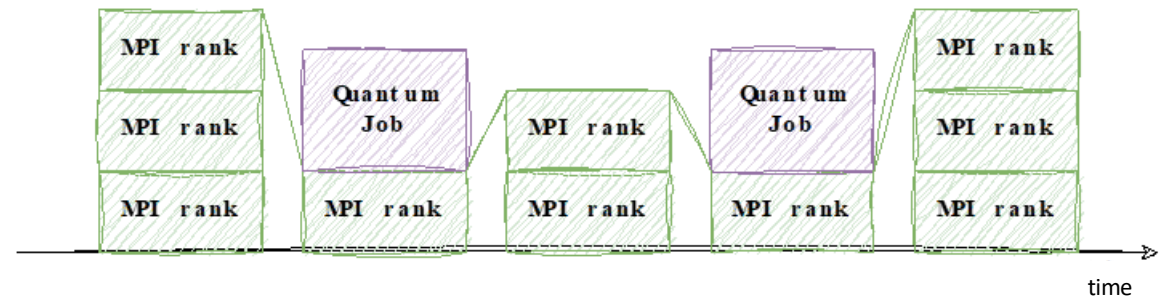


© P. Viviani

- Ideal when quantum portion of a hybrid job lasts long (e.g. > 30 min.)
- Quantum and classical jobs scheduled in an independent way, but with a single workflow
- Workflow managers or batch script could be used

Malleability

y



© R. Rocco

- Ideal when classical and quantum parts of a hybrid job have similar duration
- Allow for varying at runtime number of resources allocated for a specific job
- Could improve energy efficiency and allocation inefficiency

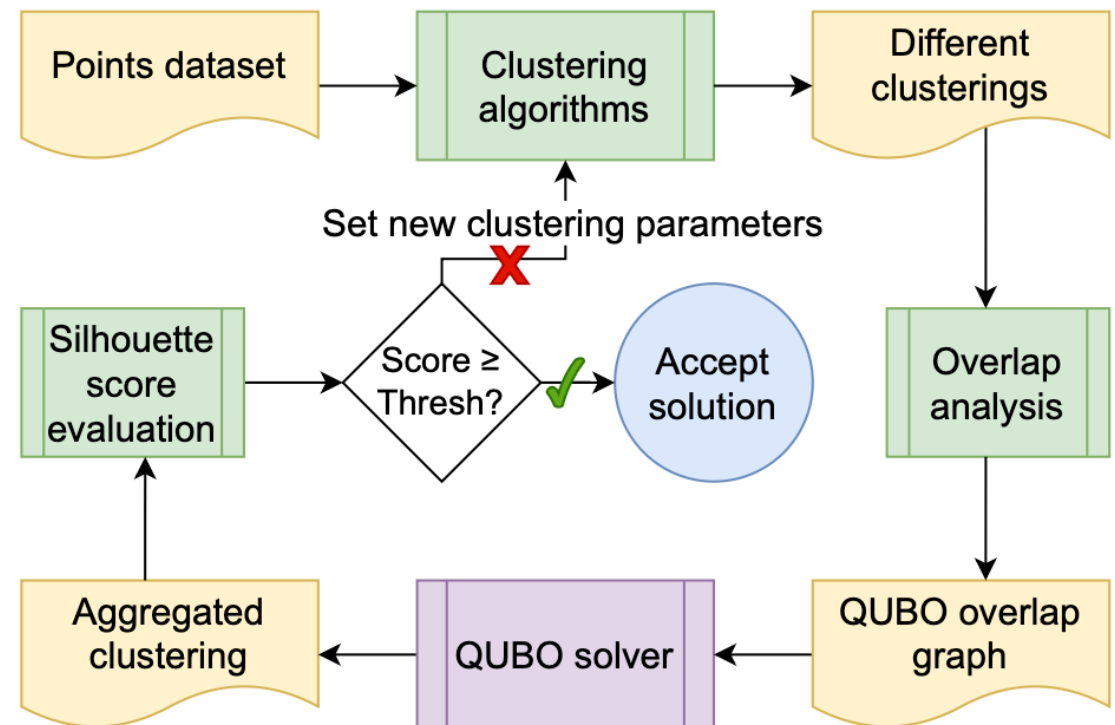
OUR USE CASE: CLUSTERING AGGREGATION

Core idea: map aggregation of multiple clustering methods into a Quadratic Unconstrained Binary Optimization problem and solve it using a QPU. Every algorithm has its pros and cons, the aggregation can improve results [1]

An attractive candidate for investigating dynamic quantum-HPC resource management:

the classical part is highly parallelizable and could effectively capitalize on parallel execution;

Each algorithm instance is assigned to a separate HPC node using MPI



[1] "Clustering Aggregation as Maximum-Weight Independent Set", Li et al., NIPS 2012,

[2] "A clustering aggregation algorithm on neutral-atoms and annealing quantum processors", Scotti et al., arXiv:2412.07558

OUR TESTBED PLATFORM - QCLUSTER

SLURM version 23.02.7.

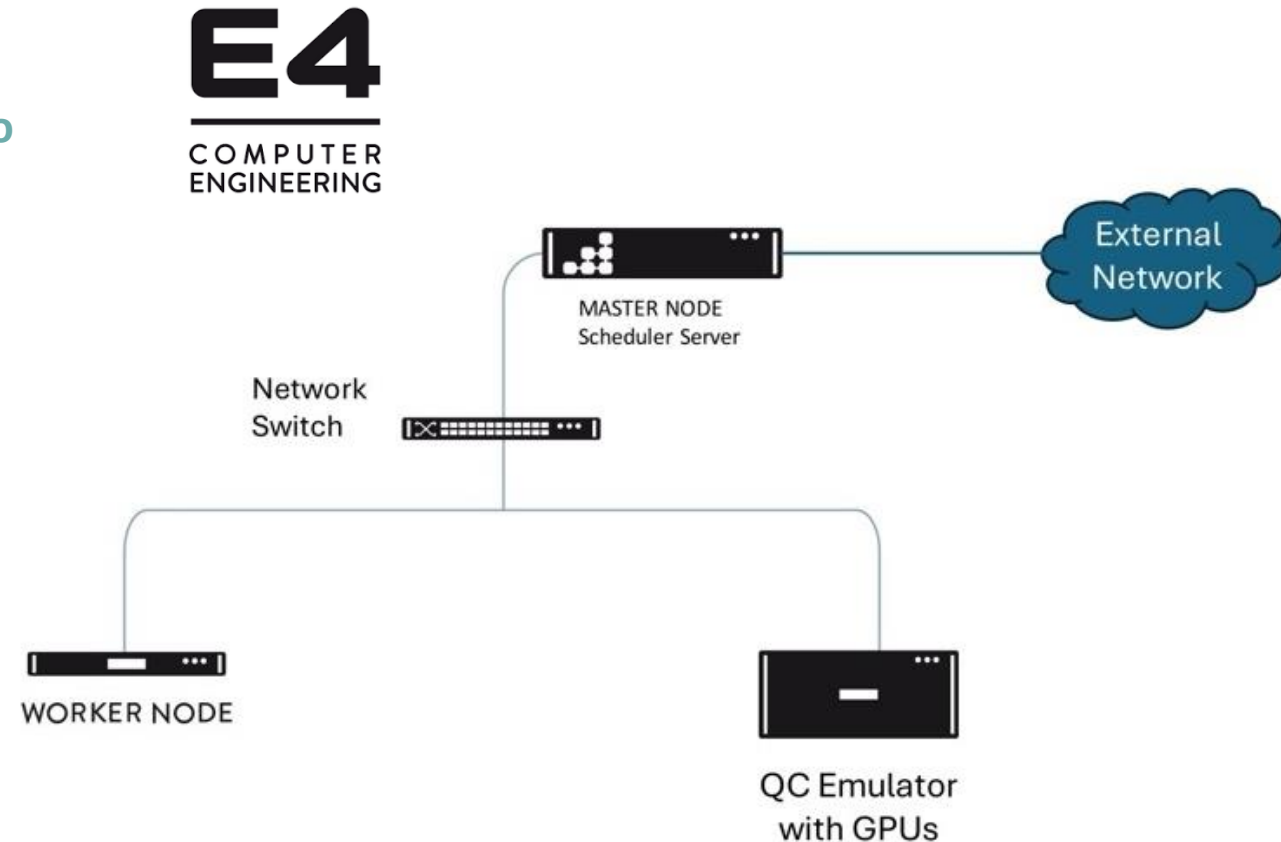
The cluster is a plausible HPC-QC integration scenario at scale, comprising two partitions: a log-in and a master node.

Compute partition: three CPU-only nodes. Each compute node contains two AMD EPYC 7543 CPUs and 256 GB of DDR4 memory.

“Quantum” partition (acting as quantum emulator.): contains two AMD EPYC 7282 CPUs, 512 GB of DDR4 memory.

source code

is publicly available at <https://github.com/E4-Computer-Engineering/clustering-mis>



EXPERIMENTAL RESULTS - QCLUSTER

- no resource contention, i.e. with no other jobs in the cluster queue
- two-minute-long quantum jobs, i.e. reproducing the behavior of a neutral atoms machine
- average the metrics from five runs for each strategy

TABLE I
EXECUTIONS WITH 2 MINUTES LONG QUANTUM JOBS.

Execution Type	Mode	Wall time [seconds]	Resource usage [node-seconds]
Baseline	Single	1019.58 ± 0.85	3058.74 ± 2.56
Workflow	Single	1057.80 ± 6.02	1161.20 ± 6.94
Malleability	Single	1029.06 ± 1.54	1647.75 ± 1.54
Baseline	Double	2038.43 ± 0.96	6115.30 ± 2.89
Workflow	Double	1226.00 ± 1.58	2324.00 ± 3.39
Malleability	Double	1127.65 ± 1.18	2817.73 ± 1.27

RESULTS:

- **BASELINE** : the fastest one, but it is less efficient regarding resource usage.
- **WORKFLOW** : performs poorly in terms of wall time since it asks SLURM for resources at every step, and the overhead of the WMS slows it down. Conversely, it is the best regarding resource usage with minimal node-second consumption.
- **MALLEABILITY** : acts as a compromise between the other two.

→ In the absence of resource contention, both malleability and workflow approaches primarily conserve valuable computational resources with a negligible impact on time-to-solution.

EXPERIMENTAL RESULTS - QCLUSTER

- two concurrent workloads
- under a queue empty from other submissions
- emulating two-minutes-long quantum jobs
- experiment averaged over five runs each.

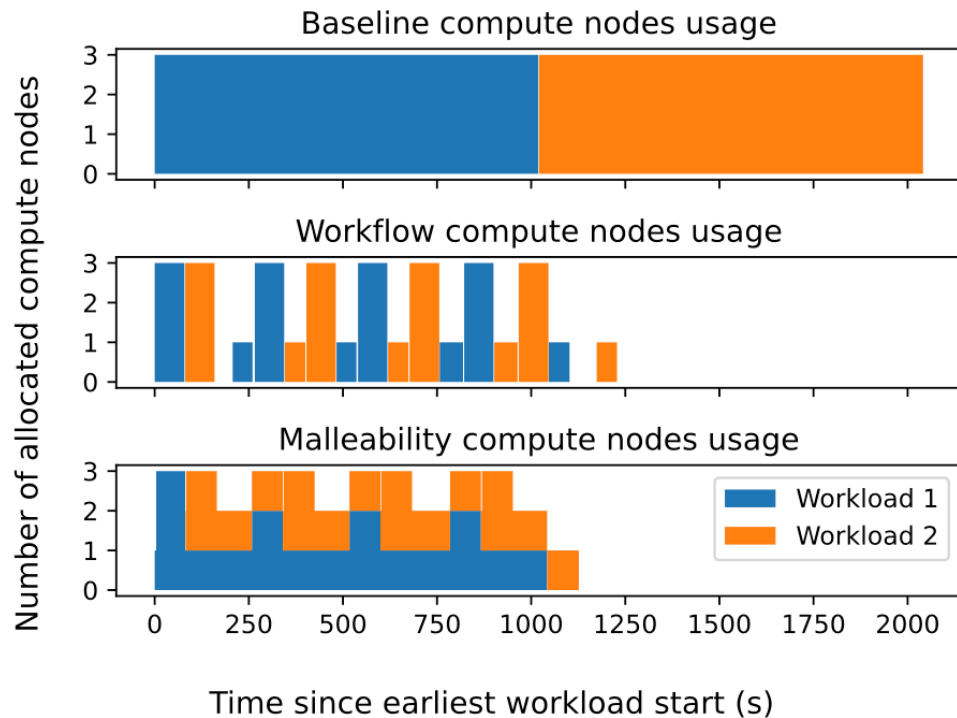


TABLE I
EXECUTIONS WITH 2 MINUTES LONG QUANTUM JOBS.

Execution Type	Mode	Wall time [seconds]	Resource usage [node-seconds]
Baseline	Single	1019.58 ± 0.85	3058.74 ± 2.56
Workflow	Single	1057.80 ± 6.02	1161.20 ± 6.94
Malleability	Single	1029.06 ± 1.54	1647.75 ± 1.54
Baseline	Double	2038.43 ± 0.96	6115.30 ± 2.89
Workflow	Double	1226.00 ± 1.58	2324.00 ± 3.39
Malleability	Double	1127.65 ± 1.18	2817.73 ± 1.27

RESULTS:

- **BASELINE:** the worst-performing one
- **WORKFLOW and MALLEABILITY:** can interleave their execution, finishing earlier and using fewer resources

MALLEABILITY vs WORKFLOW: malleability needs least one MPI process to remain active at all times, even when computations are offloaded to the QPU

→ *ADVANTAGE, not OVERHEAD! The simulation can resume immediately*

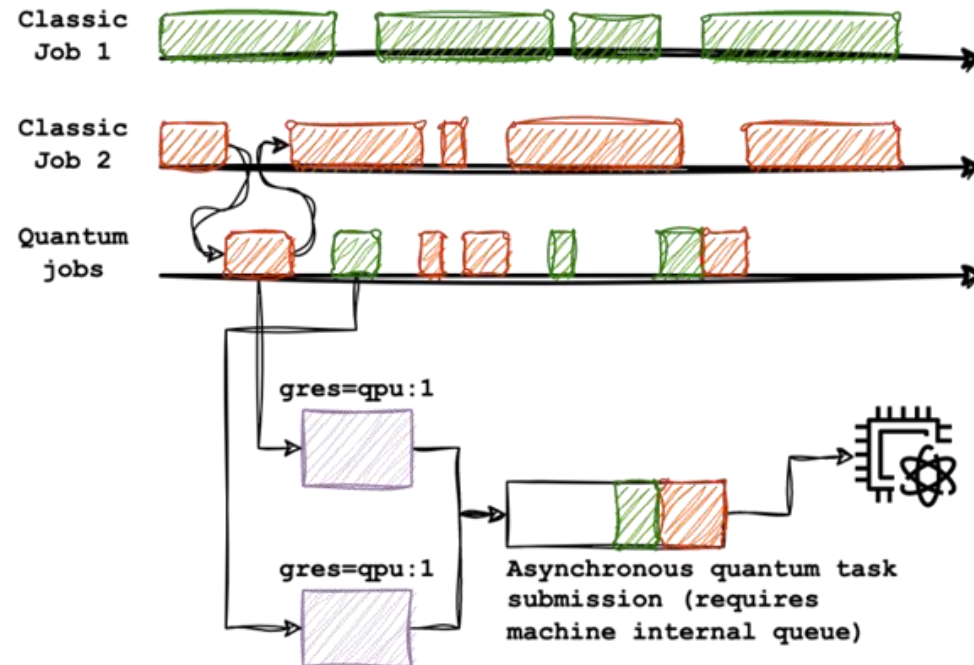
EXPERIMENTAL RESULTS - LEONARDO

LEONARDO: larger-scale execution environment with a complex resource contention scenario

- Both malleability and workflow decomposition substantially reduce classical resource consumption (up to 45.7% and 64% respectively) compared to the static baseline
- workflow achieving the lowest resource usage and variance,
- malleability offering a good balance between wall time and efficiency.

→ **Their benefits become increasingly pronounced as the quantum phase duration grows, making them particularly well suited for hybrid workloads targeting neutral-atom or other longer execution-time quantum technologies**

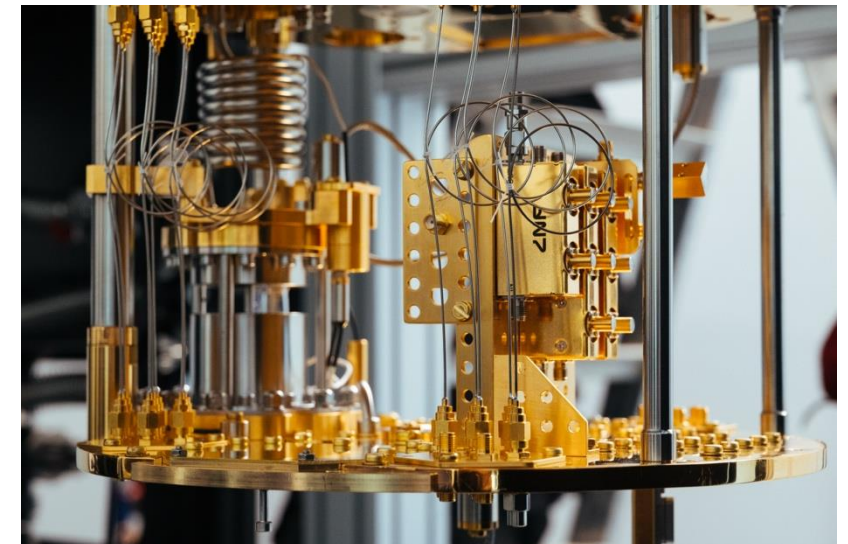
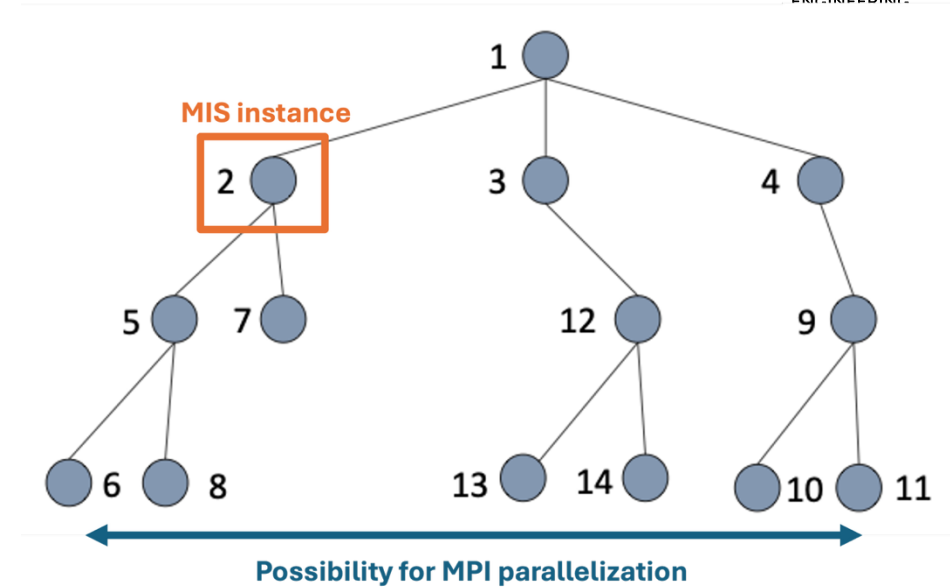




© P. Viviani

- Possible to allocate more QPUs than available with gres
- Internal QPU queue manages the quantum workload
- Maximum number of concurrent quantum/hybrid job submissions must be fixed
- Maximum waiting time for quantum job defined

- We are simulating a multi-user environment by submitting replicas of the application solving a Graph Coloring problem through multiple Maximum Independent Set (MIS) problems formulated as QAOA
- We are also considering the worst-case scenario by submitting all the hybrid jobs simultaneously
- In order to artificially increase the workload imbalance of each job, we set a sleep mechanism to extend the classical computation time
- The experiment was carried out using:
 - Leonardo supercomputer, hosted at the CINECA facilities
 - Lagrange superconducting quantum computer, managed by LINKS foundation



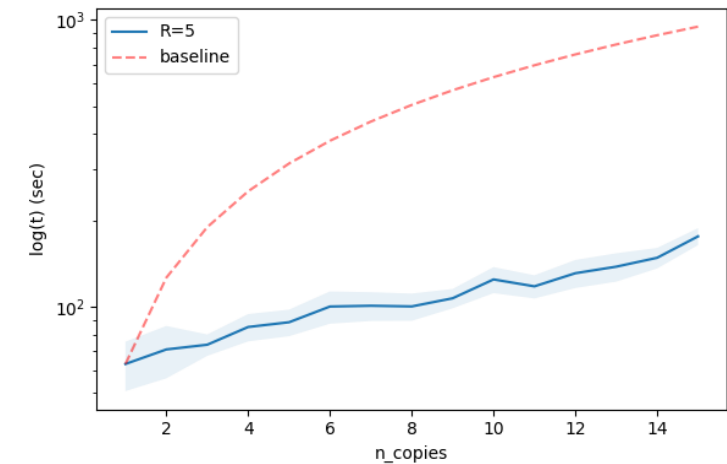
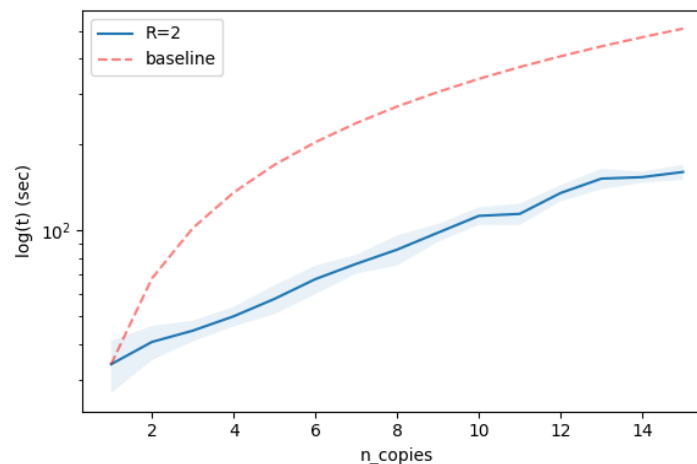
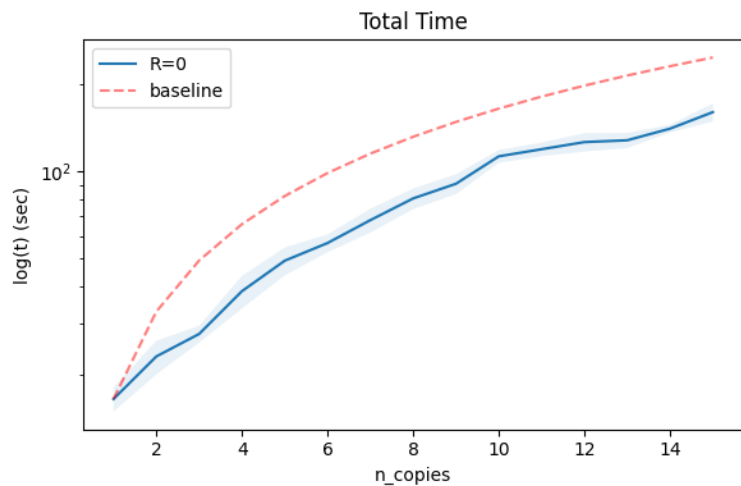
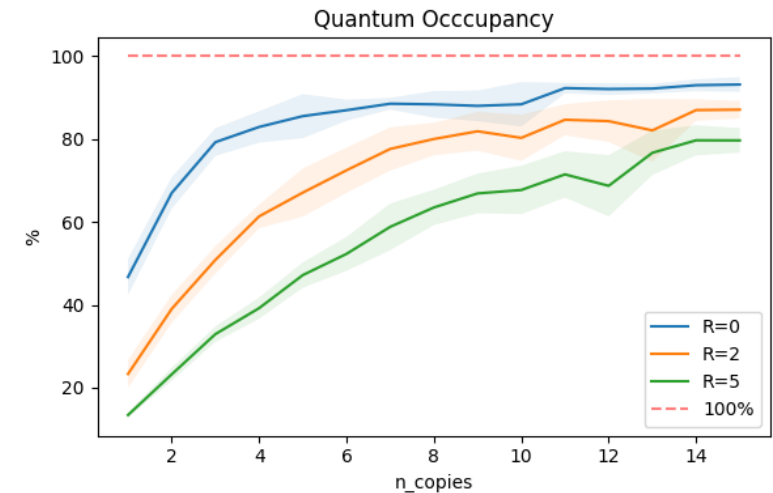
VQPUS EXPERIMENTAL RESULTS

Quantum occupancy is the percentage of occupation of the QPU during the execution of all workflows

- At the increase of the number of workload copies (n_copies), this percentage increases as well, until a saturation level is reached

Total time is the time interleaving between the start of the first workflow and the end of the last one

- We observed a performance gap between the vQPU and the baseline approaches
- By increasing the classical-quantum workload imbalance of each job, the performance gap between the two resources scheduling strategies enlarges as



TAKE HOME MESSAGE

HPC is heterogeneous, and it will include also QC in the future (if not in the present).

Quantum Computers *should* provide benefits for specific problems

Most likely QC will be used as accelerators for HPC clusters, for very specialized tasks.

We must address their presence for resource scheduling problems

→ Many open challenges for the integration ahead!

REFERENCES

Assessing the Elephant in the Room in Scheduling for Current Hybrid HPC-QC Clusters

Paolo Viviani^{1,*} Roberto Rocco², Matteo Barbieri², Gabriella Bettonte², Elisabetta Boella², Marco Cipollini⁴, Jonathan Frassinetti³, Fulvio Ganz², Sara Marzella³, Daniele Ottaviani³, Simone Rizzo², Alberto Scionti¹, Chiara Vercellino¹, Giacomo Vitali^{1,4}, Olivier Terzo¹, Bartolomeo Montrucchio⁴ and Daniele Gregori²

¹LINKS Foundation, Torino, Italy
²E4 Computer Engineering, Scandiano, Italy
³CINECA, Casalecchio di Reno, Italy
⁴Politecnico di Torino, Torino, Italy

Dynamic Solutions for Hybrid Quantum-HPC Resource Allocation

Roberto Rocco^{*§}, Simone Rizzo^{*}, Matteo Barbieri^{*}, Gabriella Bettonte^{*}, Elisabetta Boella^{*}, Fulvio Ganz ^{*}, Sergio Iserte^{††}, Antonio J. Peña^{††}, Petter Sandås^{††}, Alberto Scionti[†], Olivier Terzo[†], Chiara Vercellino^{†¶}, Giacomo Vitali^{†¶}, Paolo Viviani[†], Jonathan Frassinetti [‡], Sara Marzella[‡], Daniele Ottaviani[‡], Iacopo Colonnelli^{**}, Daniele Gregori^{*}

^{*}E4 Computer Engineering, Scandiano, Italy [†]LINKS Foundation, Torino, Italy
^{††}Barcelona Supercomputing Center (BSC-CNS), Barcelona, Spain [‡]CINECA, Casalecchio di Reno, Italy
[¶]Politecnico di Torino, Torino, Italy ^{**}Università di Torino, Torino, Italy
[§]roberto.rocco@e4company.com

<https://ieeexplore.ieee.org/abstract/document/11071588>

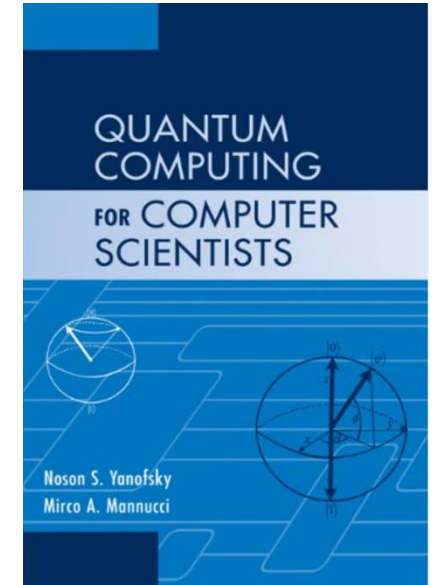
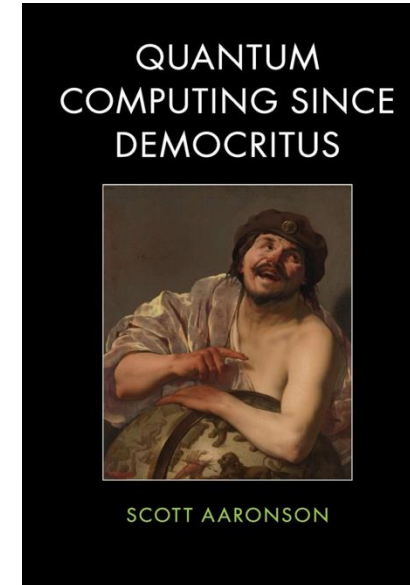
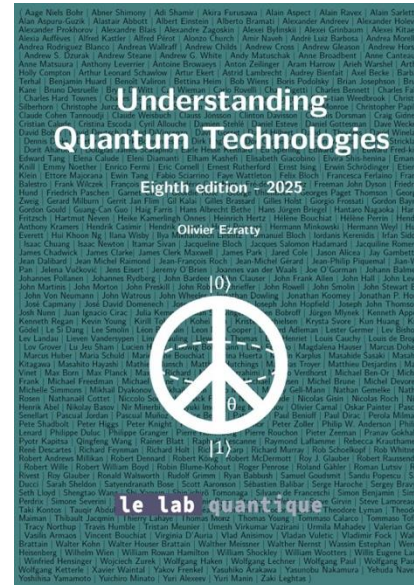
<https://ieeexplore.ieee.org/document/11250123>

Three ways to share a QPU: Scheduling for hybrid Quantum-HPC applications

Marco Cipollini^{a,b,*}, Simone Rizzo^{c,*}, Sergio Iserte^d, Paolo Viviani^{b,**}, Giacomo Vitali^{a,b}, Matteo Barbieri^{c,h,i}, Gabriella Bettonte^c, Elisabetta Boella^c, Fulvio Ganz^c, Roberto Rocco^c, Orazio Spina^{c,8}, Antonio J. Peña^d, Petter Sandås^d, Iacopo Colonnelli^c, Alberto Scionti^b, Chiara Vercellino^{a,b}, Emanuele Dri^b, Jonathan Frassinetti^f, Sara Marzella^f, Andrea Muratori^f, Daniele Ottaviani^c, Olivier Terzo^b, Bartolomeo Montrucchio^d and Daniele Gregori^c

^aPolitecnico di Torino, Torino, Italy
^bLINKS Foundation, Torino, Italy
^cE4 Computer Engineering SpA, Scandiano, Italy
^dBarcelona Supercomputing Center (BSC), Barcelona, Spain
^eUniversità di Torino, Torino, Italy
^fCINECA, Casalecchio di Reno, Italy
^gUniversità di Bologna, Bologna, Italy
^hDipartimento di Fisica e Astronomia "G. Galilei", Università di Padova, Padova, Italy
ⁱIstituto Nazionale di Fisica Nucleare (INFN), Padova, Italy

<https://arxiv.org/pdf/2604.14955>



The logo consists of the letters 'E4' in a bold, white, sans-serif font. The 'E' is slightly larger than the '4'. Below the letters is a thin white horizontal line.

E4

COMPUTER
ENGINEERING

THANK YOU

CONTACTS

gabriella.bettonte@e4company.com

roberto.rocco@e4company.com