Malleability for HPC-QC jobs scheduling on Leonardo

26 June 2026

Sara Marzella

CINECA QUANTUM COMPUTING LAB

CINECA: Not-For-Profit Consortium

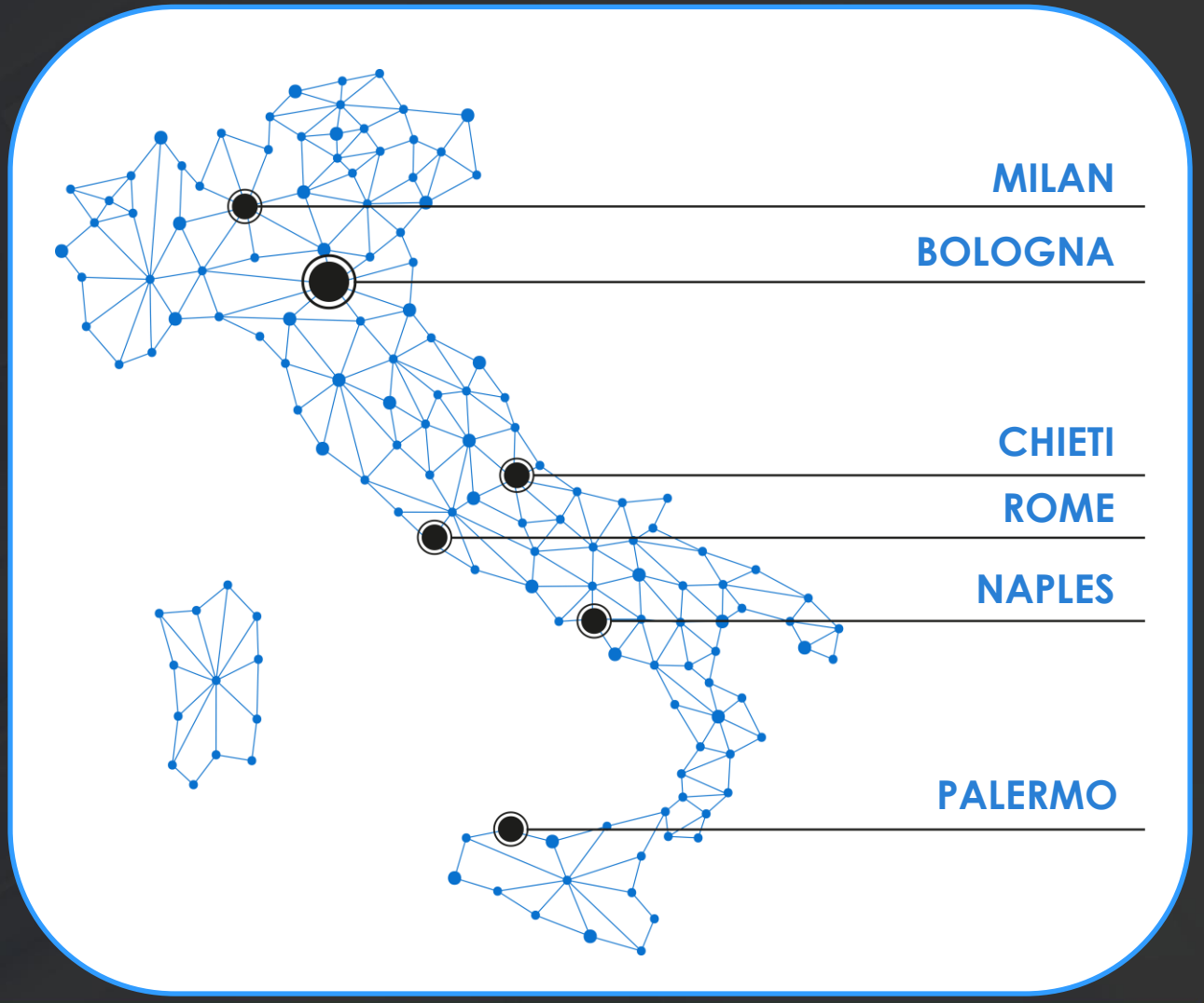
Since 1969 Cineca supports the Italian Academic System

122 MEMBERS
2 Ministries, 71 Universities,
49 Academic and Research Institutions

6 OFFICES
Bologna, Milan, Rome, Naples, Chieti, Palermo

≃1100
Employees

≃ 130 MLN €
Yearly Revenue in 2024



Supercomputer Leonardo

TOP 500 The List.

12th

CINECA QUANTUM COMPUTING LAB

Quantum Computing Lab

- PASQAL Orion Beta 140 qubits
  Analog Quantum Computer
- IQM Radiance Machine 54 qubits
  Digital Quantum Computer
- Cloud Resources via ISCRA-C
  D-Wave Quantum Annealers
  Pasqal 100 qubits Quantum Simulator

SOL

NOX

Pasqal

IQM

D-Wave
The Quantum Computing Company™

European and National projects

QEC-QEA

EuroQHPC-I

Q-SUD

QUANTUM COMPUTING AND SIMULATION CENTER

Fondazione ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

HPC|QS

HPC4UPC

MoSeGad

SmartHPC.QC

CINECA

QUANTUM COMPUTING LAB

Cineca Quantum Computing Resources

Hybrid HPC-QC System

NOX
IQM

SOL
PASQAL

Emulators

QUANTUM COMPUTING LAB

Cloud QC

PASQAL

D-Wave
The Quantum Computing Company™

Quantum «EDU» - Nox

- Superconducting Digital Quantum Computer
- 54 qubits
- 90 couplers
- Installed, integration process in progress

Cineca to house Italy's most powerful quantum computer IQM Radiance 54

17/03/2025 3 min. read

- IQM Radiance 54-qubit full-stack superconducting quantum computer will be integrated into Leonardo, one of the world's fastest supercomputers in Bologna, Italy.
- Cineca intends to use the system for optimisation of quantum applications, quantum cryptography, quantum communication and artificial intelligence quantum algorithms.

IQM CINECA

CINECA QUANTUM COMPUTING LAB

EuroQHPC-I EuroQCS-Italy - Sol

SOL

- Installation in progress
- 140 qubits analog qc
- Upgrade with local addressability in 2027

PASQAL

EuroQHPC-I

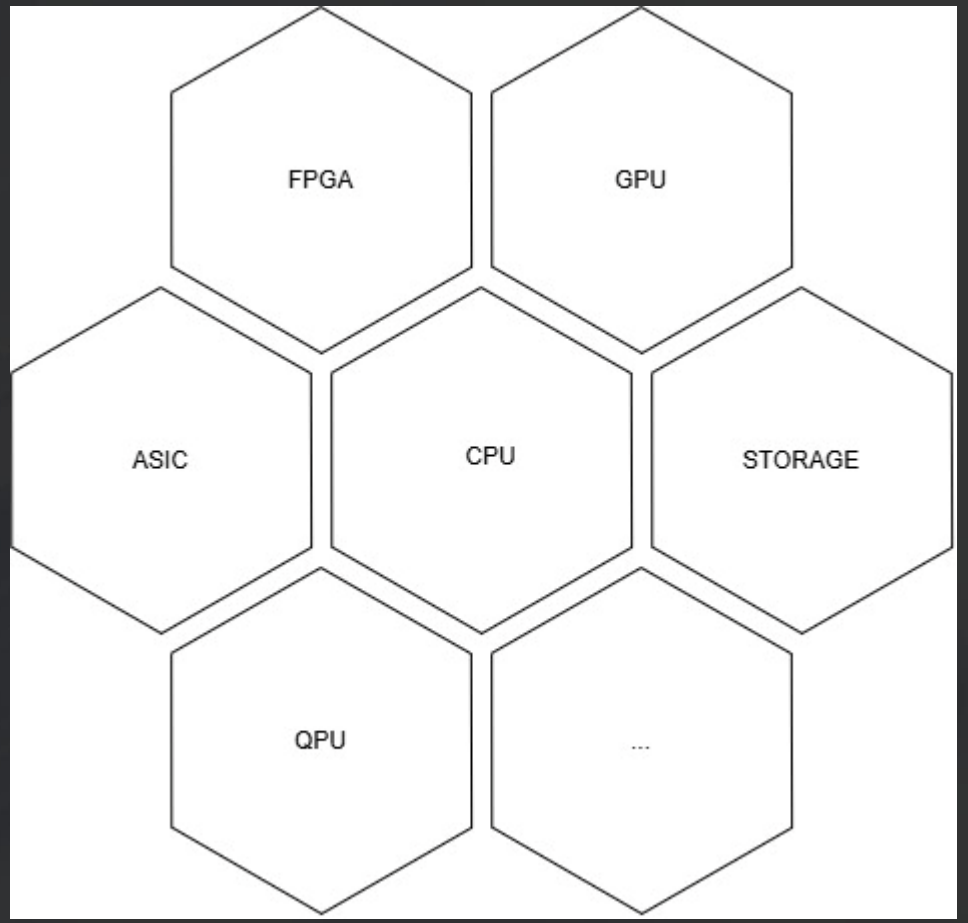
CINECA

QUANTUM COMPUTING LAB

The future is Hetereogeneous

- 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.



QPUs in HPC today: resource or bottleneck?

QPUs in the future:
Many qubits
Hopefully, fault tolerant
Directly attached to CPUs with high-speed connections, similarly to GPUs
One interface independent from technology

QPUs as of today:
Limited amount of qubits
Limited reliability, need for fault-handling mechanisms
Noise-sensitive
Attached via ethernet
Every QPU has its own features and interfaces
Small amount of quantum computers compared to the number of HPC nodes

SmartHPC-QC

Project funded by the European Union

NextGenerationEU, ICSC National Centre, CN00000013, MUR Act n. 1031 - 17/06/2022

ICSC
Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

E4 COMPUTER ENGINEERING

FONDAZIONE links PASSION FOR INNOVATION

UNIVERSITÀ DI TORINO

Politecnico di Torino

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

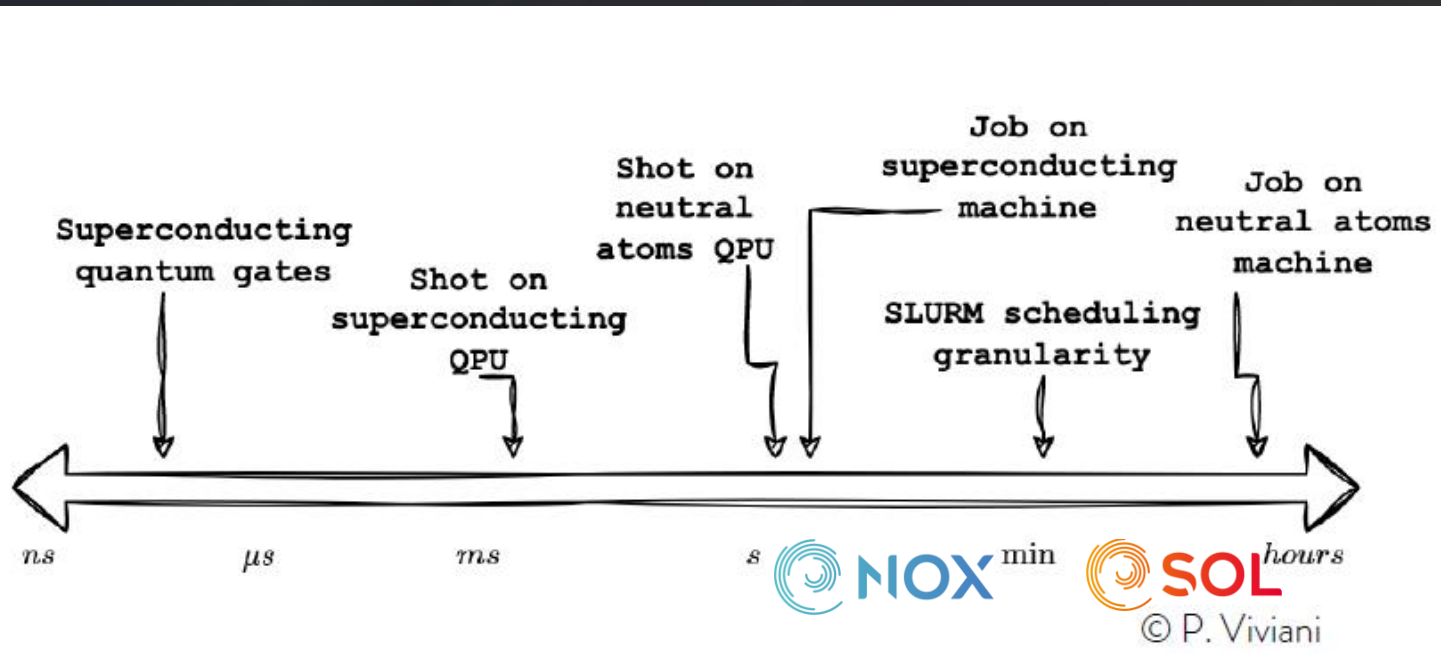
CINECA

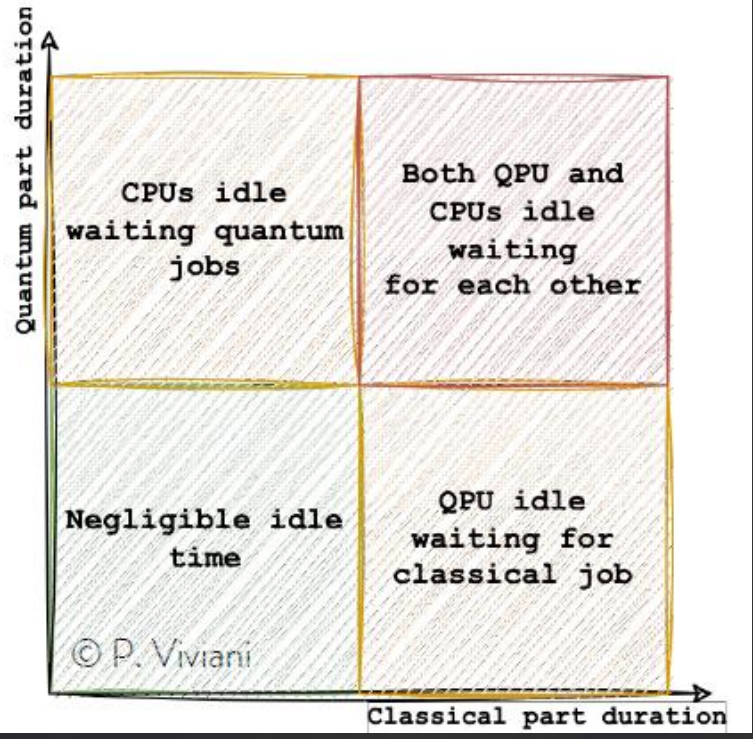
Target:

Integration of a Quantum Computer in an HPC environment

- Heterogeneous Computation
- Effective Resource Management
- Malleability
- Scheduling: Virtual QPU

Different QPUs have different execution times





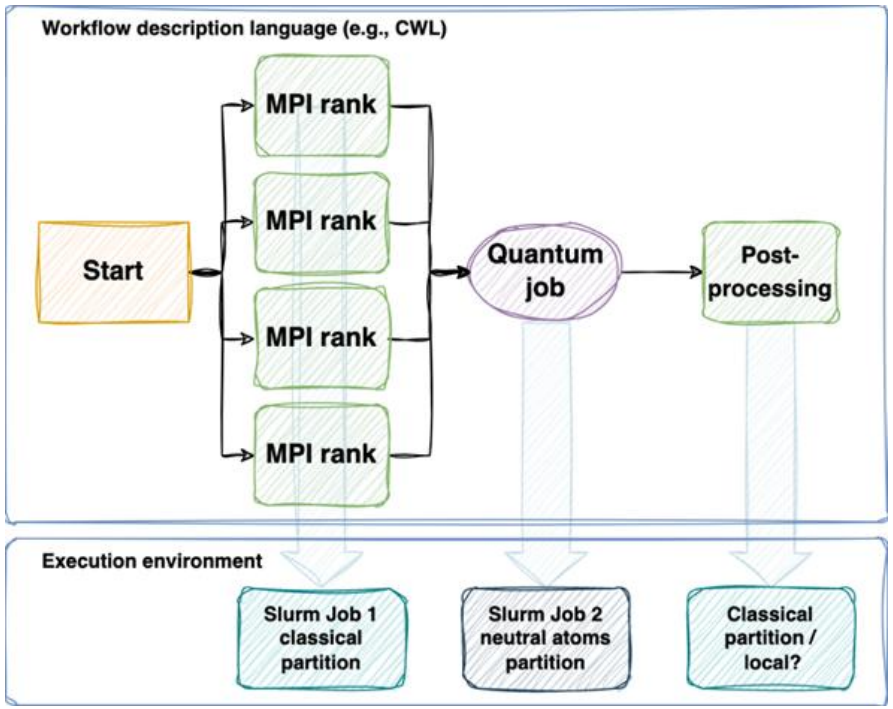
PROBLEM: Small amount of quantum computers compared to the number of HPC nodes + different QPUs have different execution time

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

SmartHPC.QC CINECA QUANTUM COMPUTING LAB

Solutions

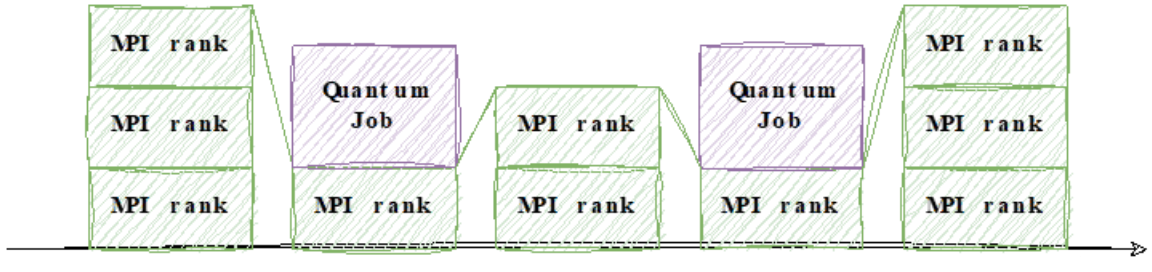
Workflows



© 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



© R. Rocco

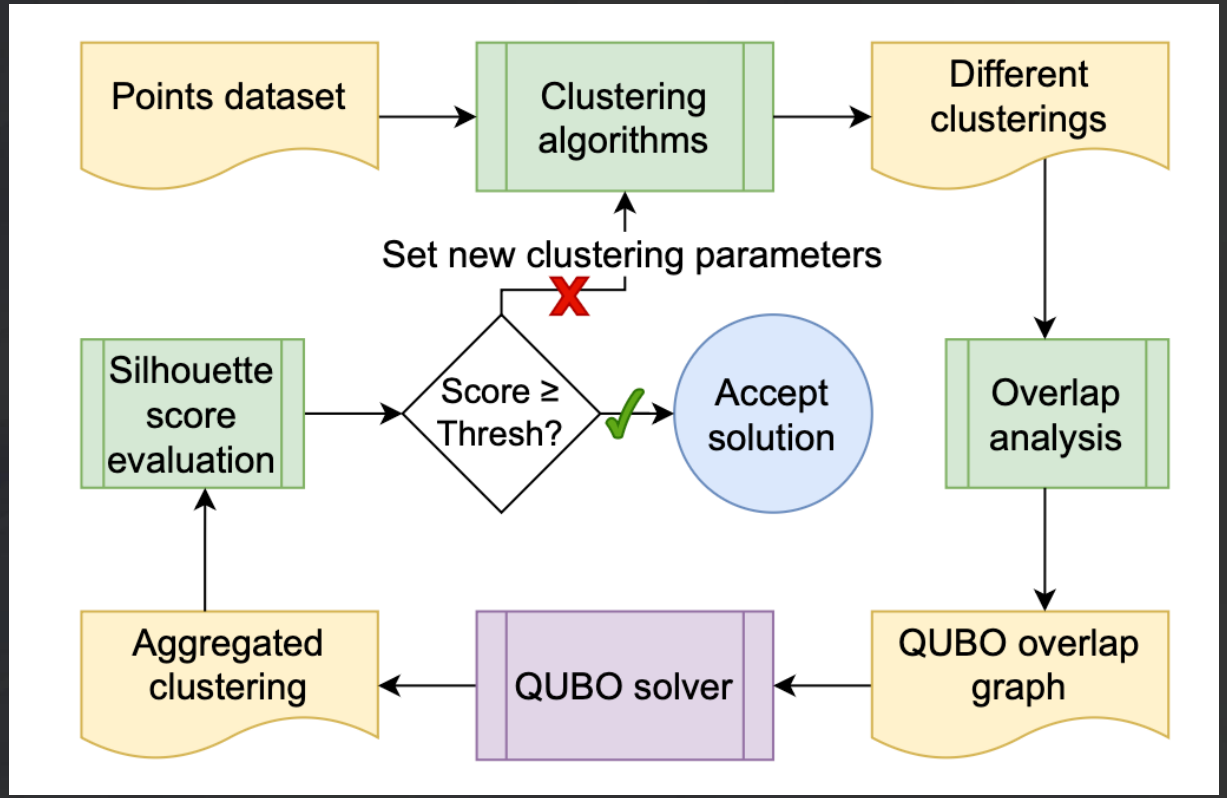
- Ideal when classical and quantum parts of a hybrid job have similar duration
- Allow for varying at runtime number of resources allocated for a specific job
- Could improve energy efficiency and allocation inefficiency

SmartHPC.QC CINECA QUANTUM COMPUTING LAB

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



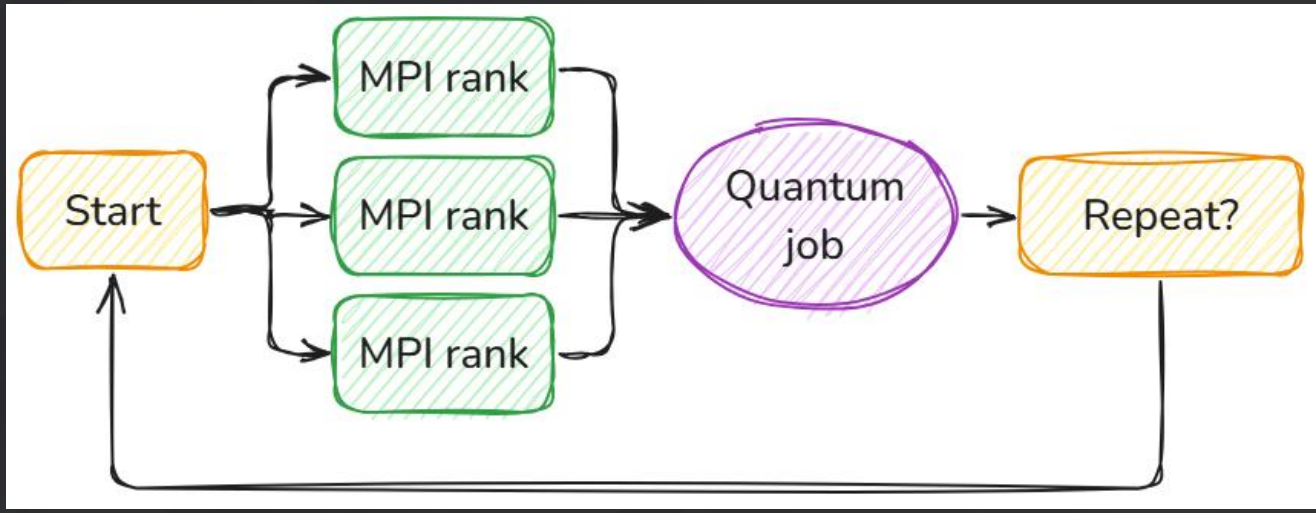
[1] "Clustering Aggregation as Maximum-Weight Independent Set", Li et al., NIPS 2012,
[2] "A clustering aggregation algorithm on neutral-atoms and annealing quantum processors", Scotti et al., arXiv:2412.07558

SmartHPC.QC CINECA QUANTUM COMPUTING LAB

Our use case: quantum offloading from a parallel job

Classical code runs on a SLURM compute partition executing three widely used clustering algorithms:

- k-means
- DBSCAN
- hierarchical clustering



© S. Rizzo

Each algorithm instance is assigned to a separate HPC node using MPI
➔ it enables parallel execution and performance improvements over the serial baseline

MPI rank 0 computes the silhouette score

Repeat? terminate the loop when we achieve a silhouette score greater than 0.8.
To prevent an infinite loop and excessive quantum resource usage, we also set an upper bound of 10 loop iterations, after which the best is chosen.

SmartHPC.QC CINECA QUANTUM COMPUTING LAB

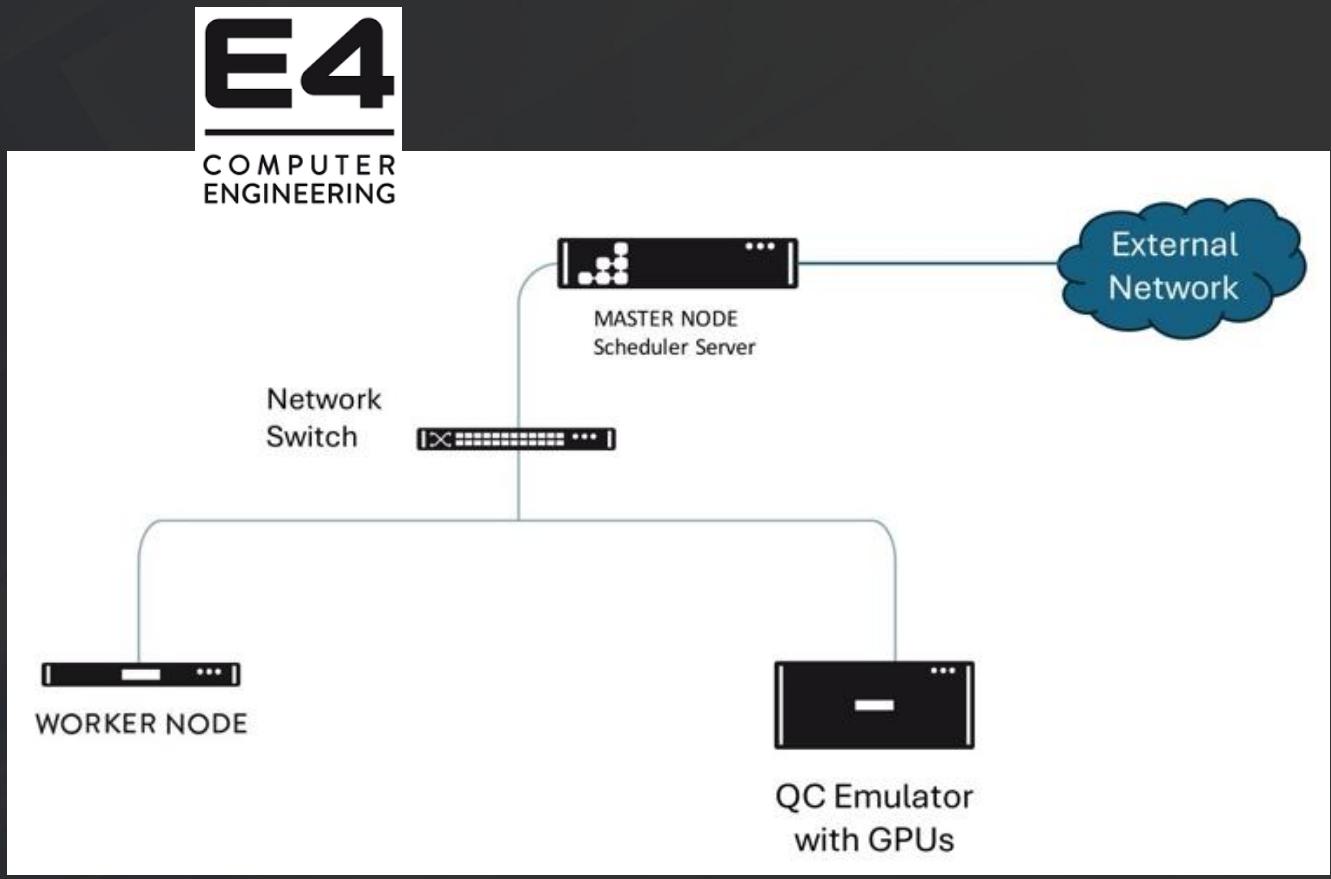
Our testbed platform - QLUSTER

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

SmartHPC.QC CINECA QUANTUM COMPUTING LAB

Experimental results - QLUSTER

- 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.

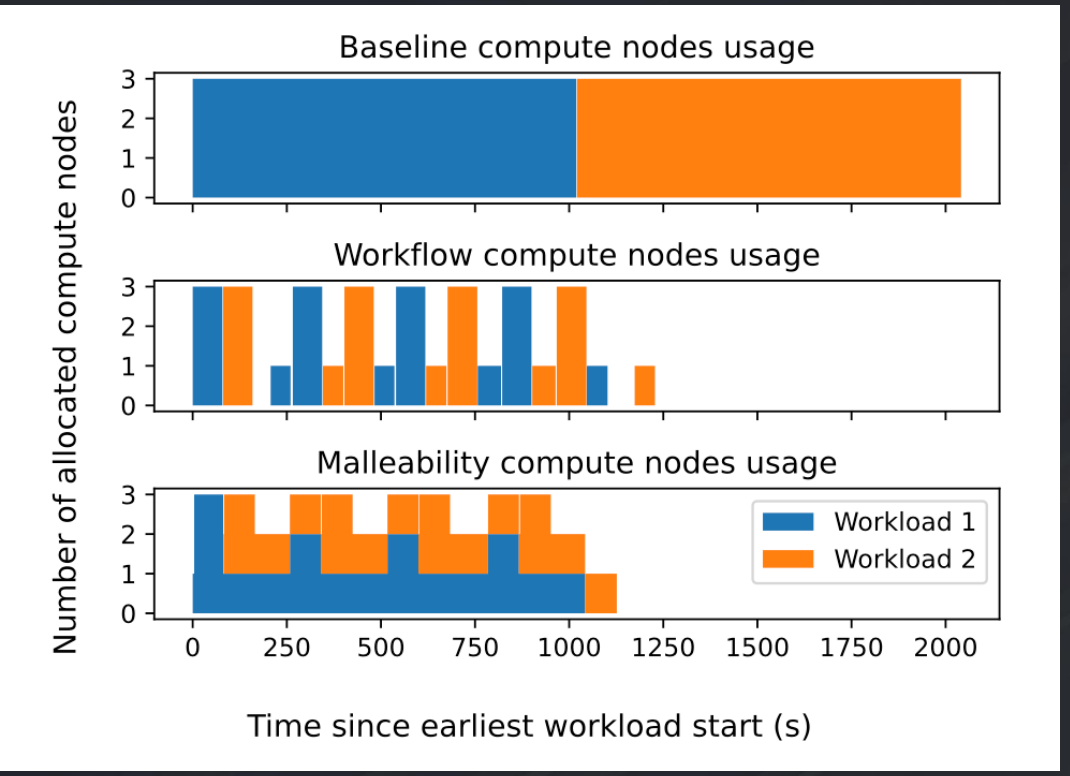
SmartHPC.QC CINECA QUANTUM COMPUTING LAB

Experimental results - QLUSTER

- 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 at least one MPI process to remain active at all times, even when computations are offloaded to the QPU, but **ADVANTAGE, not OVERHEAD!** The simulation can resume immediately

SmartHPC.QC CINECA QUANTUM COMPUTING LAB

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**

Leonardo: executions with two-minute-long quantum jobs.

Execution type	Wall time [seconds]	Resource usage [node-seconds]
Baseline	1126.10 ± 21.20	3378.30 ± 63.59
DMR	1151.50 ± 40.63	1835.60 ± 41.74
Workflow	1127.90 ± 20.93	1216.80 ± 6.55

Leonardo: executions with short (< 1 second) quantum jobs.

Execution type	Wall time [seconds]	Resource usage [node-seconds]
Baseline	803.69 ± 516.77	2411.07 ± 1550.30
DMR	844.03 ± 529.89	1517.28 ± 538.12
Workflow	818.48 ± 420.97	1241.34 ± 22.79

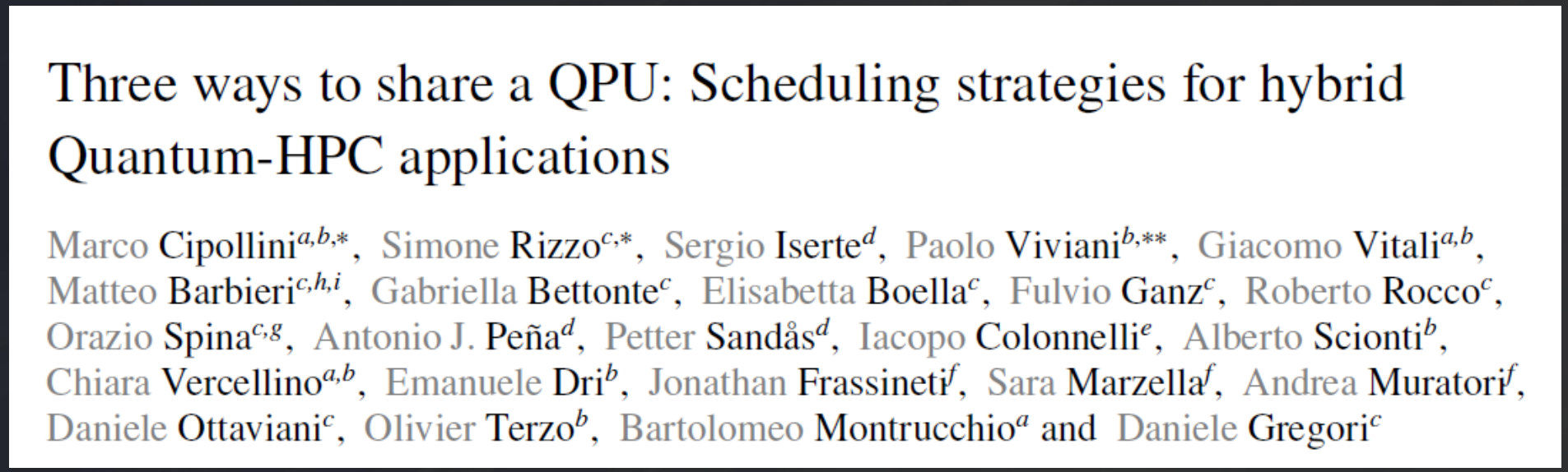
SmartHPC.QC CINECA QUANTUM COMPUTING LAB

Publications

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

"Dynamic Solutions for Hybrid Quantum-HPC Resource Allocation"

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



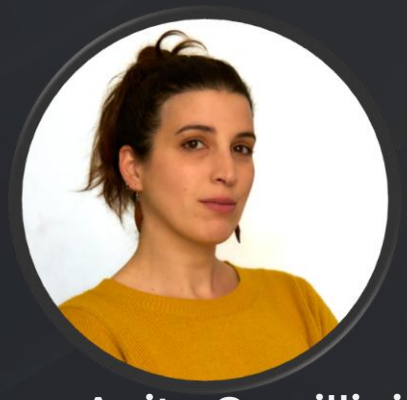
Three ways to share a QPU: Scheduling strategies for hybrid Quantum-HPC applications

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,g], Antonio J. Peña[d], Petter Sandås[d], Iacopo Colonnelli[e], Alberto Scionti[b], Chiara Vercellino[a,b], Emanuele Dri[b], Jonathan Frassineti[f], Sara Marzella[f], Andrea Muratori[f], Daniele Ottaviani[c], Olivier Terzo[b], Bartolomeo Montrucchio[a] and Daniele Gregori[c]

SmartHPC.QC CINECA QUANTUM COMPUTING LAB

Cineca Quantum Computing Lab

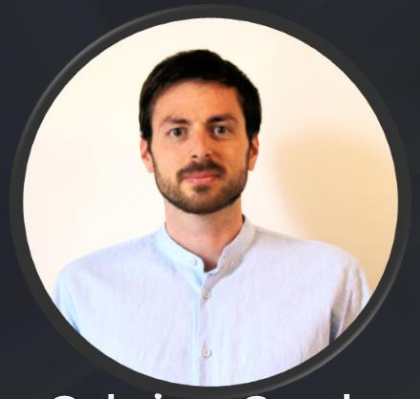
Sara Marzella



Anita Camillini

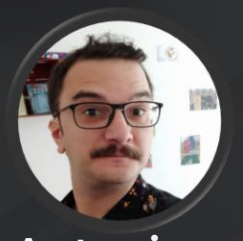
Vito Palmisano

Jonathan Frassineti



Gabriele Spada

Francesca Gebbia



Antonio Costantini

E4 COMPUTER ENGINEERING

FONDAZIONE links PASSION FOR INNOVATION

UNIVERSITÀ DI TORINO

Politecnico di Torino

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

SmartHPC.QC

CINECA

QUANTUM COMPUTING LAB