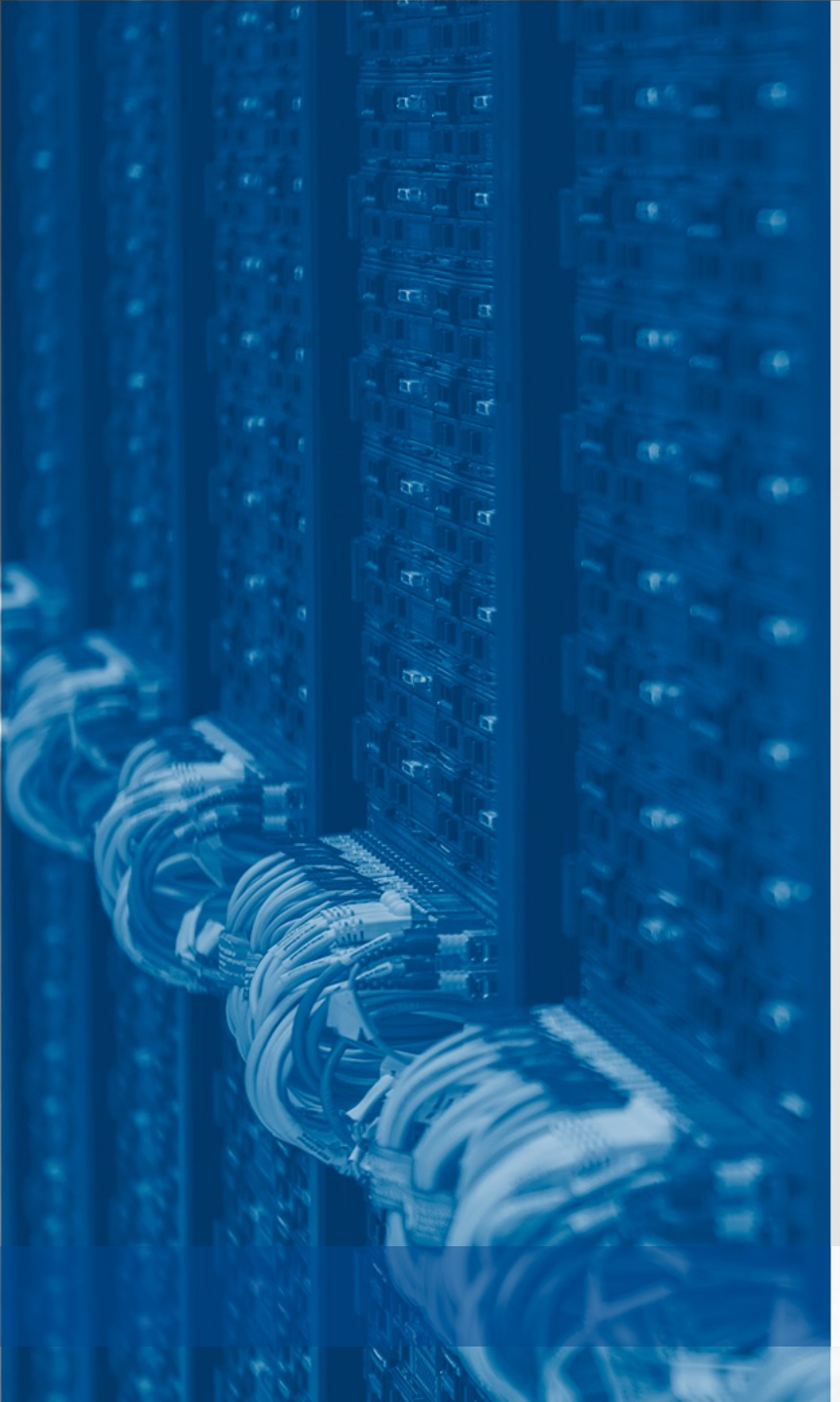


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

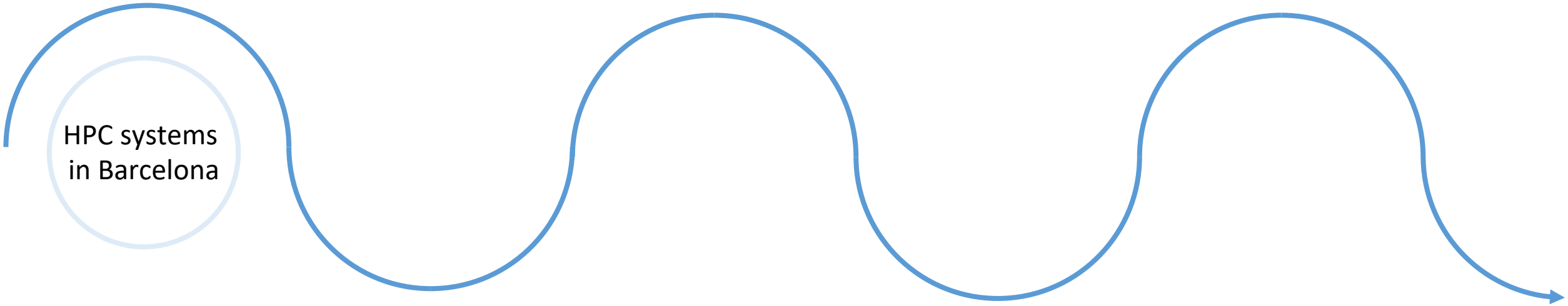
Hybrid Quantum–HPC Integration at BSC:

Architectures, Workflows, and Emerging Capabilities

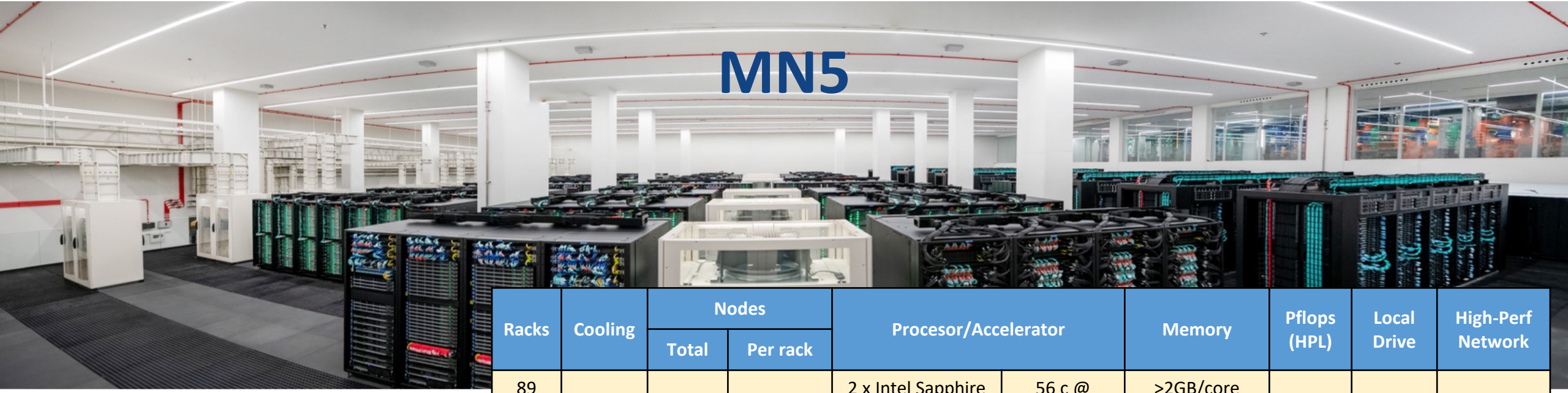
Mar Tejedor

ISC 2nd QRUCH workshop, 26th June 2026

Presentation roadmap



MN5



- Federated slurm:
  - Hybrid allocations enabled
- Shared GPFS

GPP

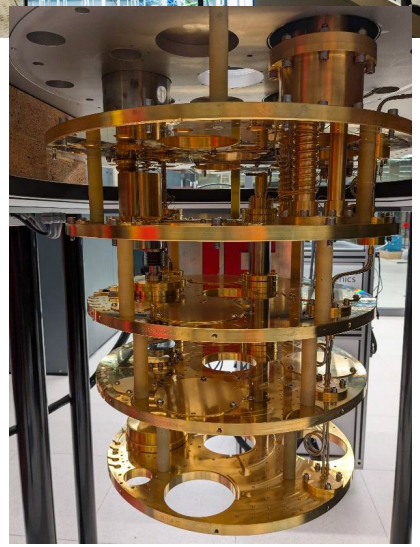
Racks	Cooling	Nodes		Procesor/Accelerator		Memory	Pflops (HPL)	Local Drive	High-Perf Network
		Total	Per rack						
89	DCL + RDHX	6192	72 (6 x 6 x 2)	2 x Intel Sapphire Rapids 8480+	56 c @ 2GHz	>2GB/core 256 GB DDRS	40.10	960 GB NVMe	1 x NDR200 Shared between 2 nodes
		216				>8GB/core 1024 GB DDRS			
1		72		2 x Intel Sapphire Rapids 9480	56 c @ 1.9 GHz	> 0.5 GB HBM/core 128 GB HBM + 32 GB DDRS	0.34		

ACC

Racks	Cooling	Nodes		Procesor/Accelerator		Memory	Pflops (HPL)	Local Drive	High-Perf Network
		Total	Per rack						
89	DLC	1120	32	2 x Intel Sapphire Rapids 8460Y+	40c @ 2 GHZ	512 GB	175.3	480 GB NVMe	4 x NDR200
				4 Nvidia Hopper 64GB HBM					

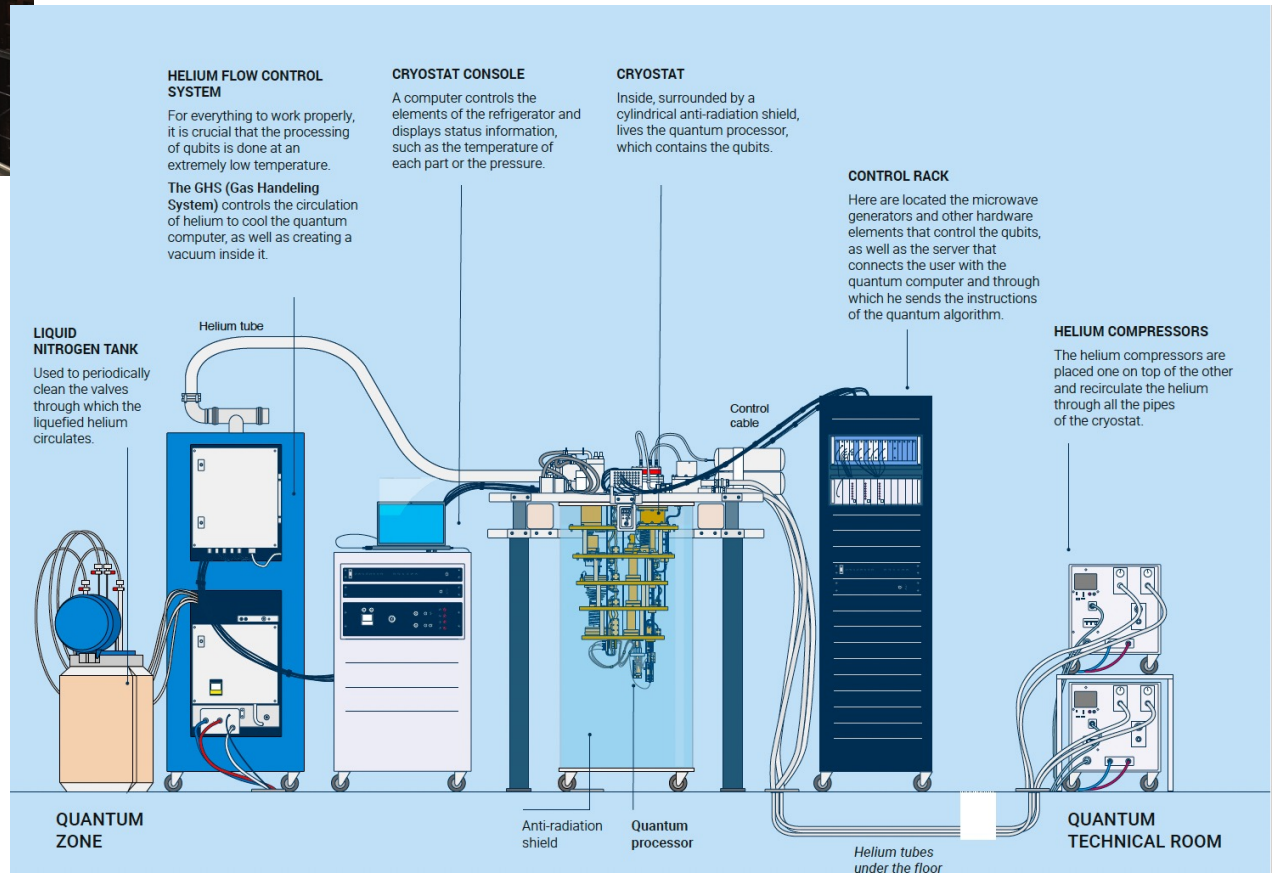
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MN5 - Ona



- 2 coolers for the various quantum chips installation
- Shared GPFS
- Multi-cluster SLURM

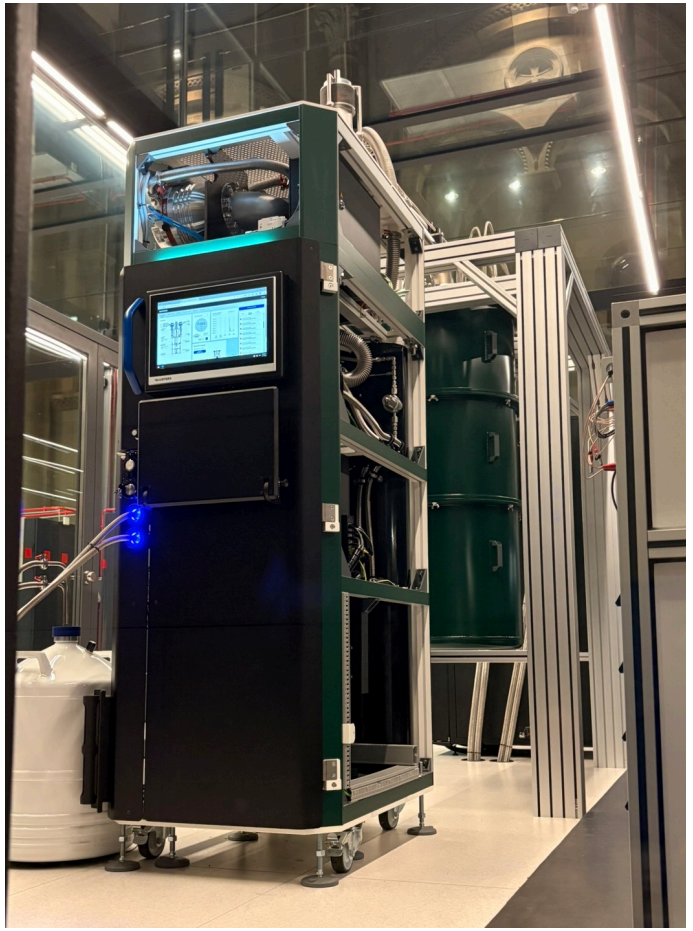
- Digital quantum computer (Qilimanjaro Quantum Tech)
  - Red 35 qubits
  - Blue 25 qubits



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EuroQCS-Spain

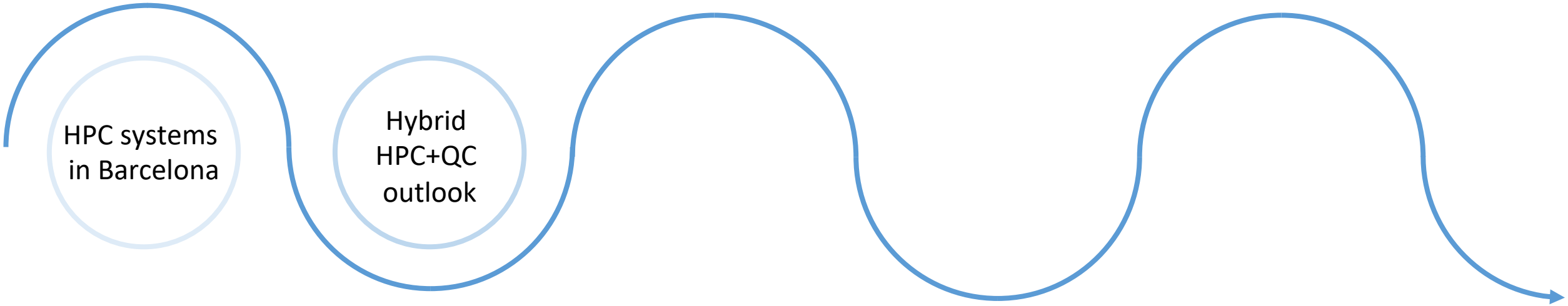
- Analogue quantum computer in the form of a **quantum annealer**
- Effective in solving certain types of optimization problems
- Based on the concept of adiabatic evolution, where a system is slowly transformed from an initial state to a final state, that represents the optimal solution
- Quantum annealing is particularly effective for solving problems with a large number of local optima
- Used in various fields, including machine learning, finance and logistics
- 10 -> 15 -> 25 qubits
- From Qilimanjaro Quantum Tech - Spain



Under deployment

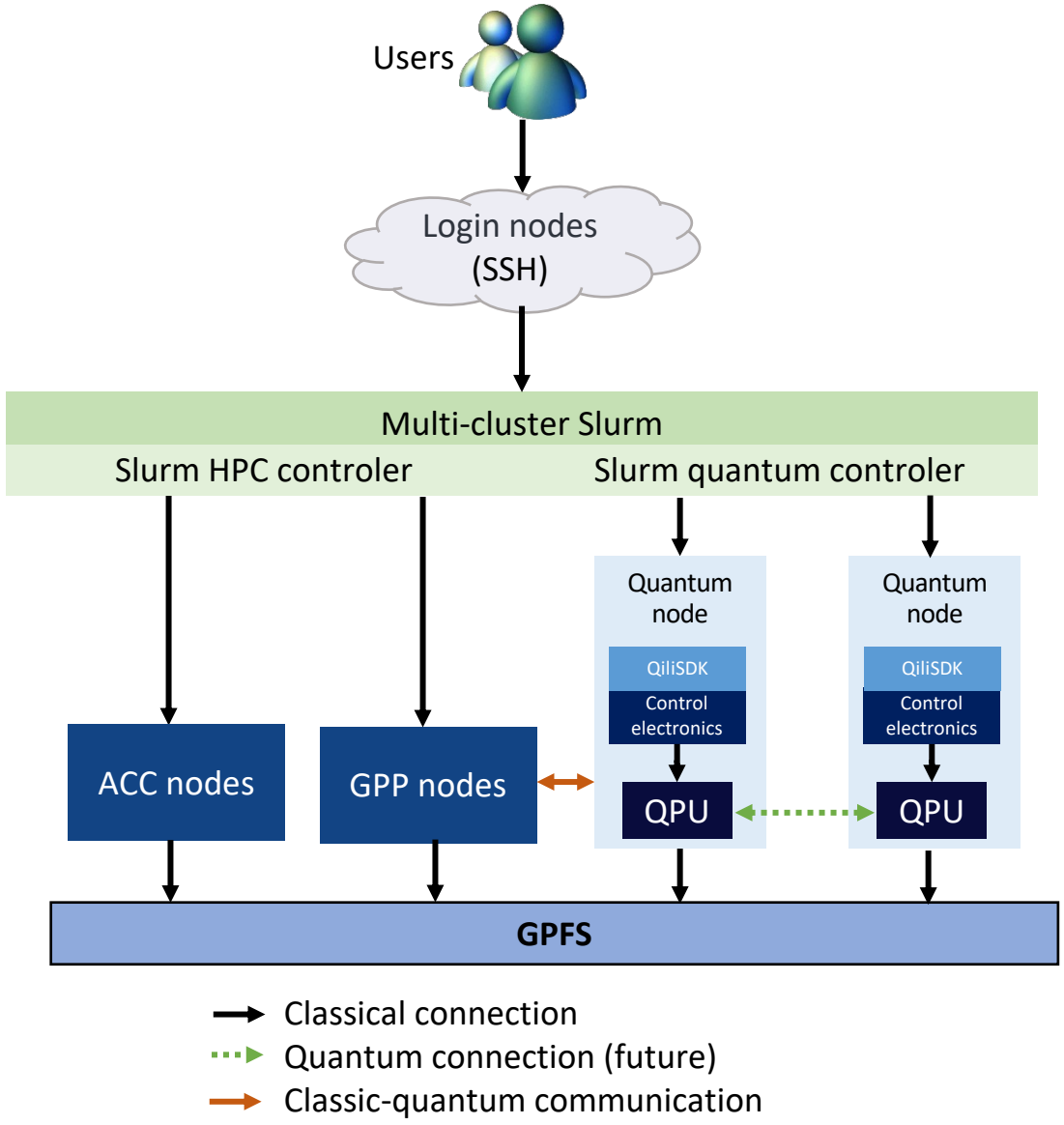
https://www.eurohpc-ju.europa.eu/eurohpc-quantum-computers/our-quantum-computers_en#euroqcs-spain

Presentation roadmap



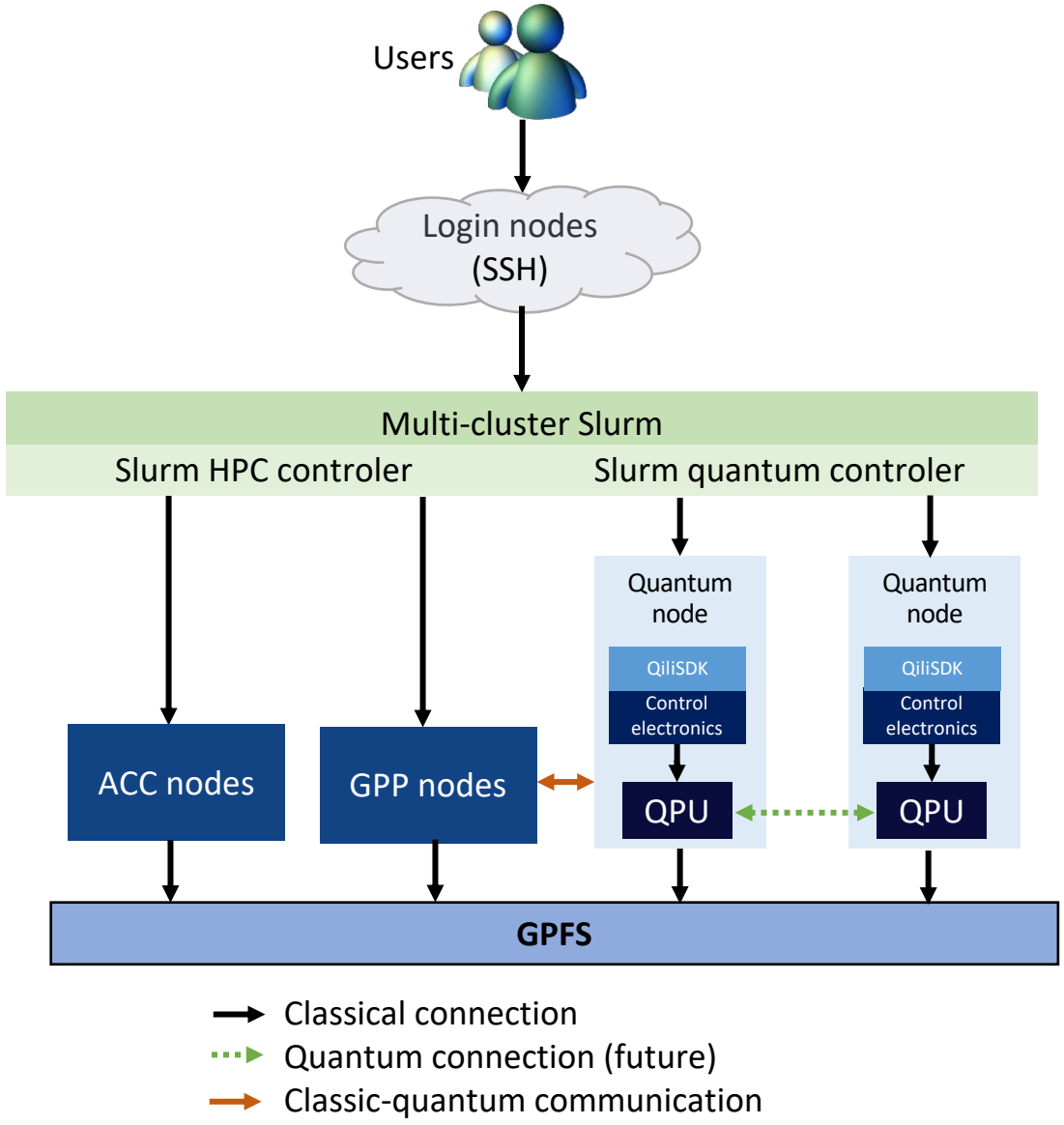
HPC+QC setting at BSC

HPC+QC setting at BSC



HPC+QC setting at BSC

All systems share GPFS

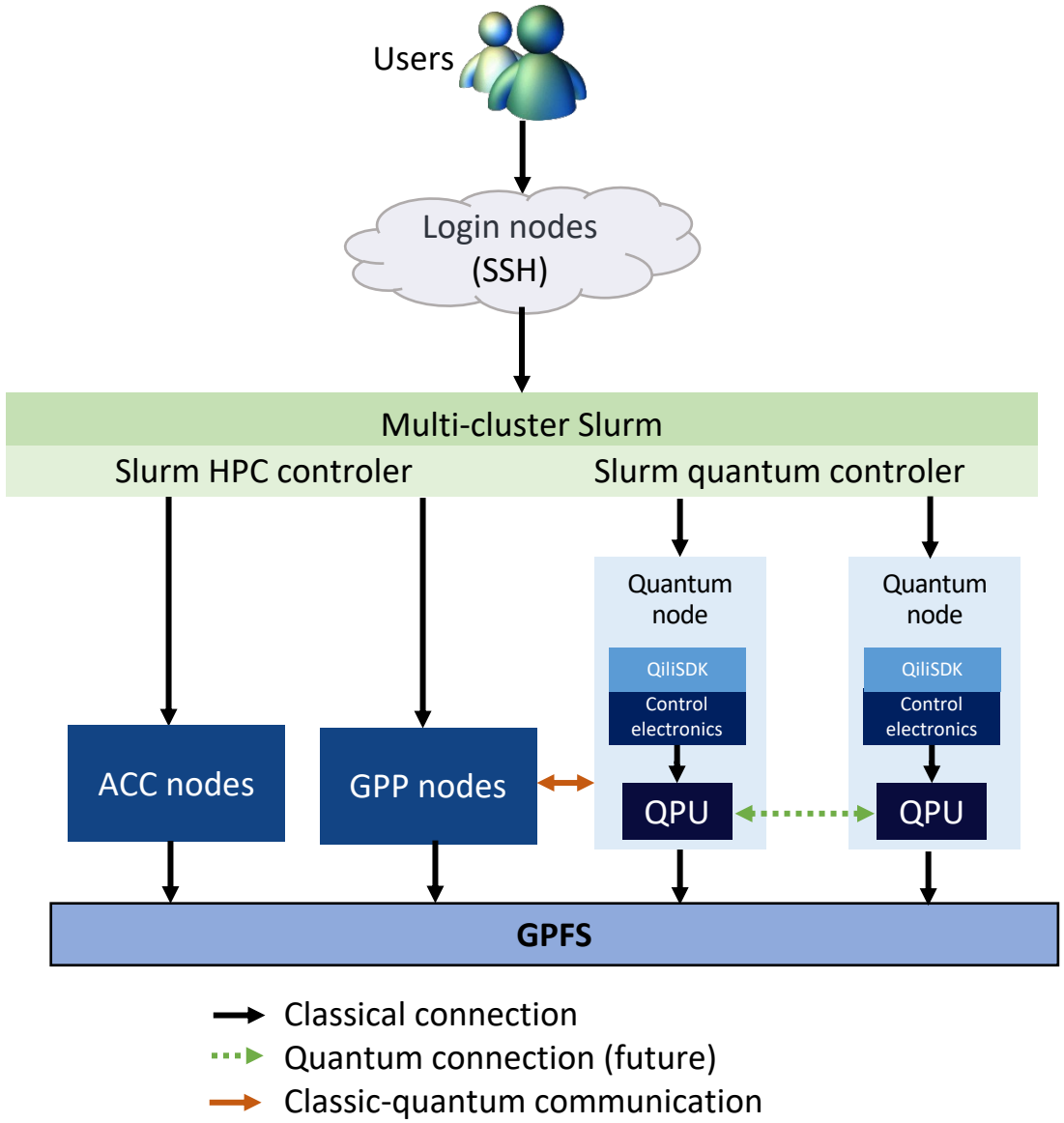


HPC+QC setting at BSC

All systems share GPFS

Multi-cluster slurm (not yet...)

- Support for hybrid jobs HPC-QC
- Possibility to request more than one quantum node

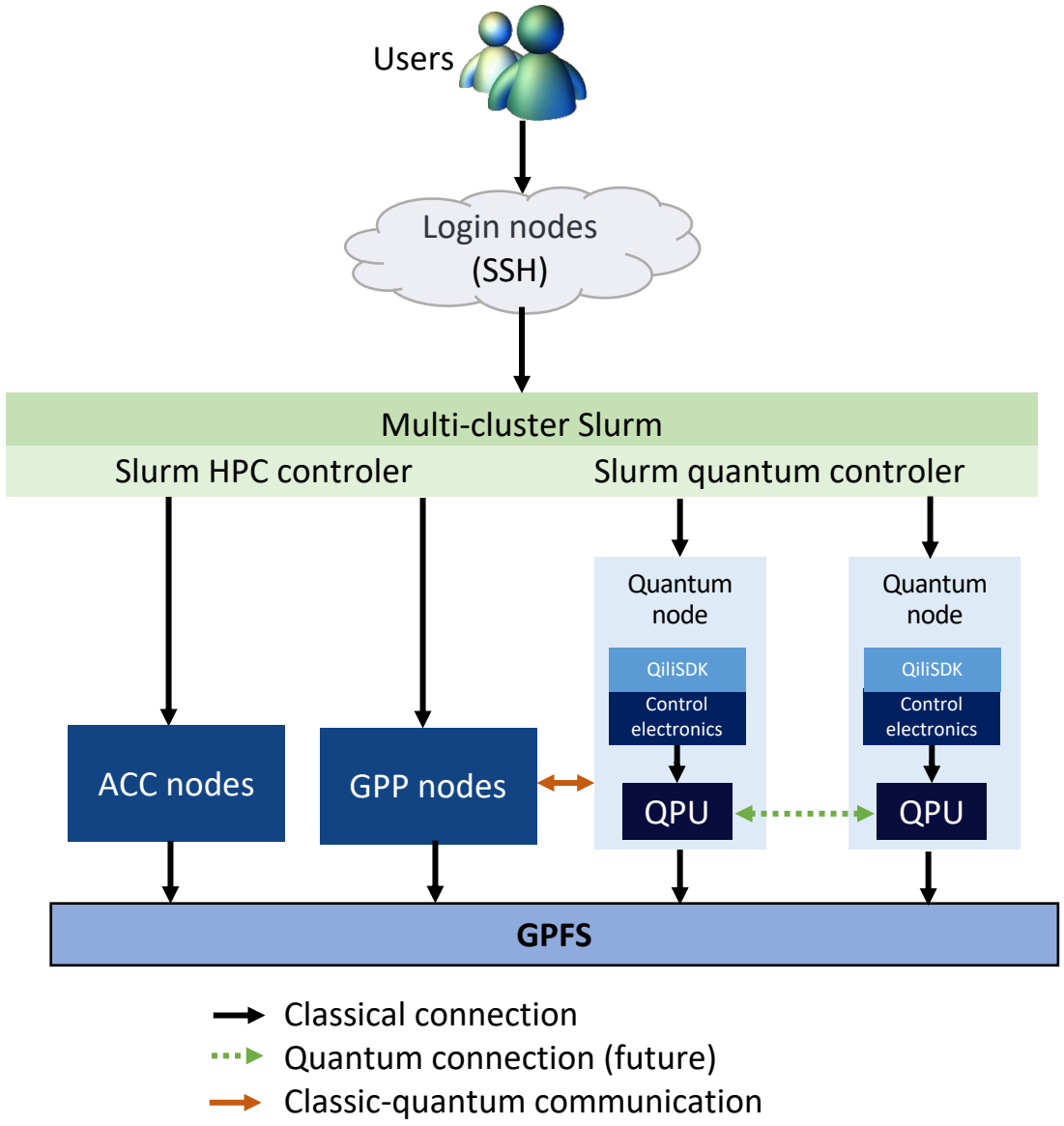


HPC+QC setting at BSC

All systems share GPFS

Multi-cluster slurm (not yet...)
- Support for hybrid jobs HPC-QC
- Possibility to request more than one quantum node

Communication Compute nodes with quantum nodes
- Through the storage network
- Same latency as network for moving data (10Gb)
  - Size of data depends on circuit size
  - Smaller bandwidth inside the quantum node
    - Qblox and control electronics – 1Gb



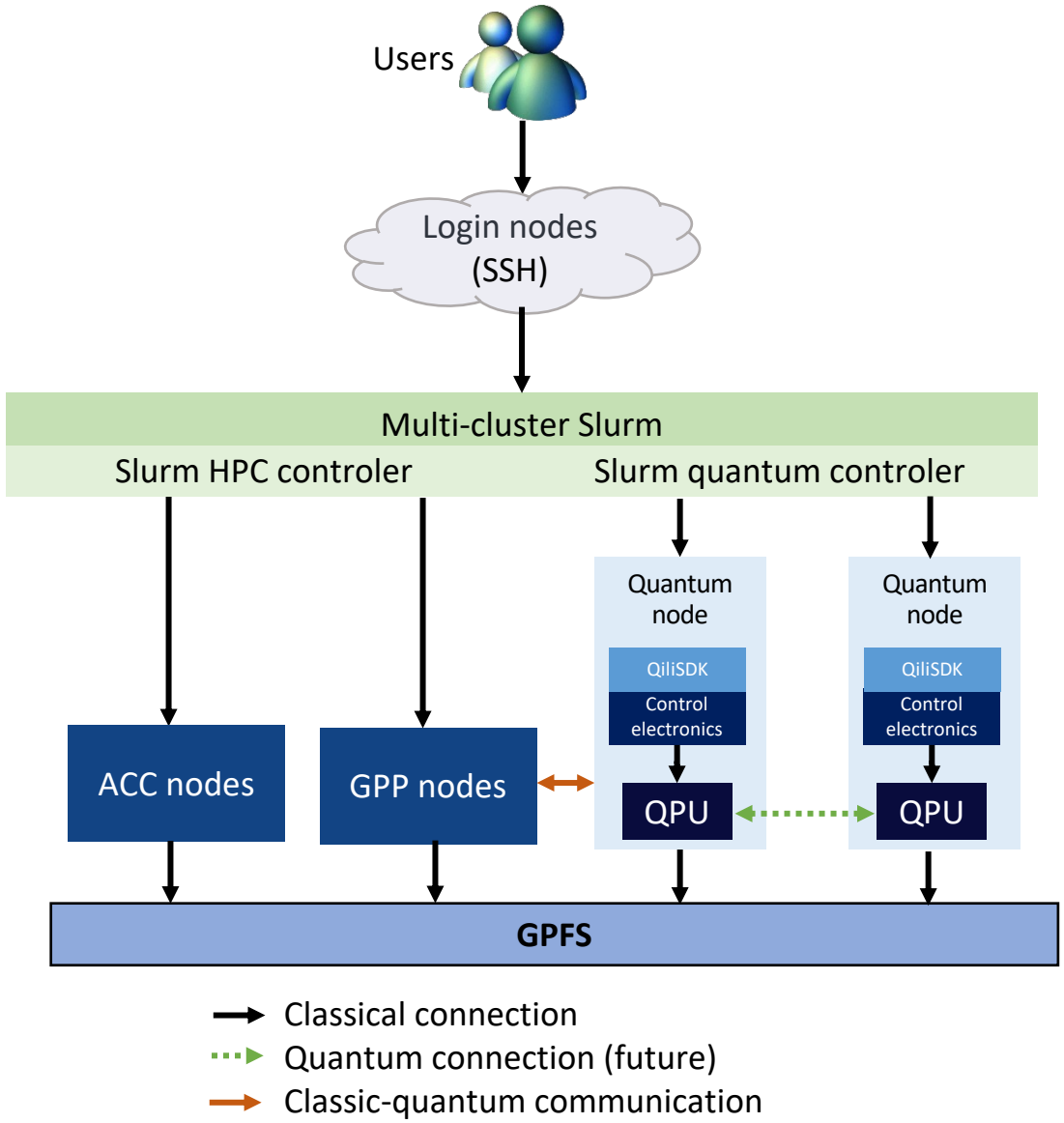
HPC+QC setting at BSC

All systems share GPFS

Multi-cluster slurm (not yet...)
- Support for hybrid jobs HPC-QC
- Possibility to request more than one quantum node

Communication Compute nodes with quantum nodes
- Through the storage network
- Same latency as network for moving data (10Gb)
  - Size of data depends on circuit size
  - Smaller bandwidth inside the quantum node
    - Qblox and control electronics – 1Gb

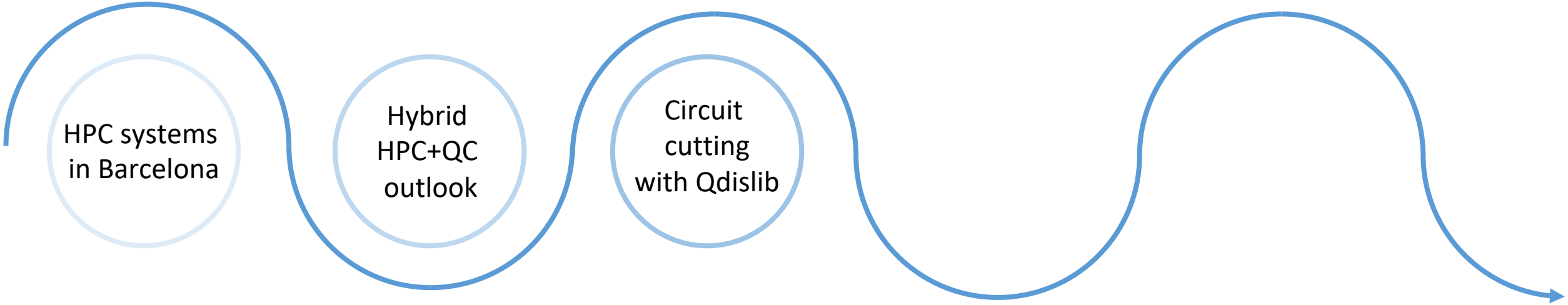
Goal: support for developing hybrid HPC-QC workflows in this setting



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Presentation roadmap

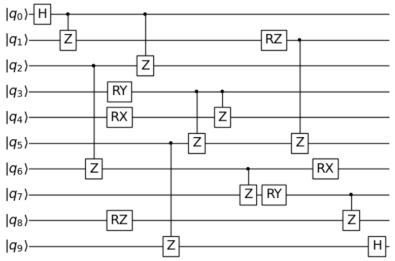
QDISLIB
QUANTUM DISTRIBUTED LIBRARY



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

A problem we wanted to solve

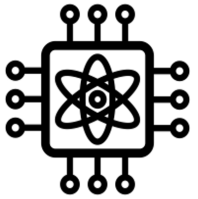
- In our search for hybrid use cases we run into the circuit cutting problem:



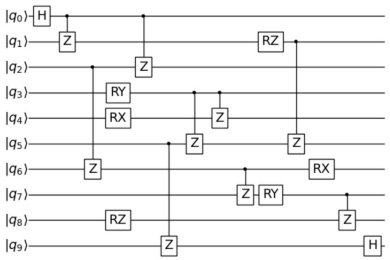
10 qubit circuit

?

5 qubit
QC chip



QPU

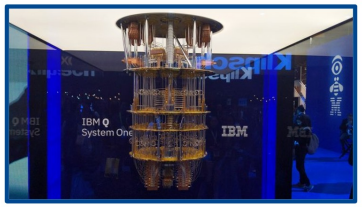


?

How can we take full advantage of all of them?

MN ONA

MN5

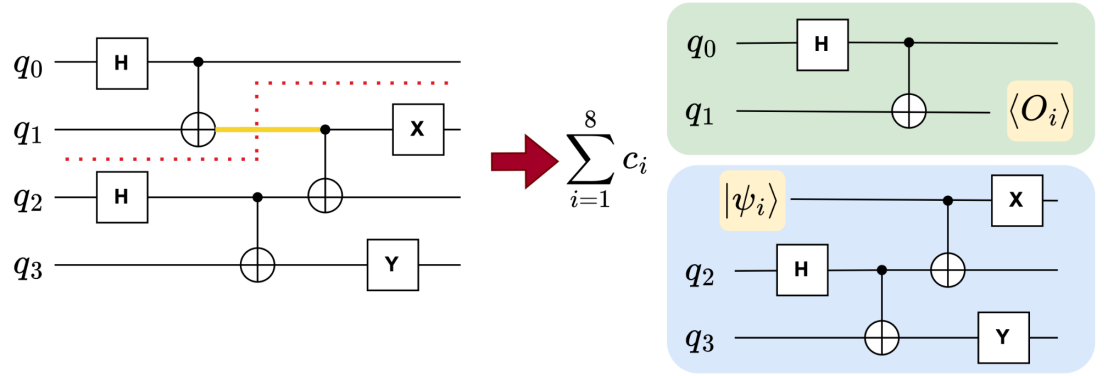


IBMQ

Circuit Cutting

Wire cutting

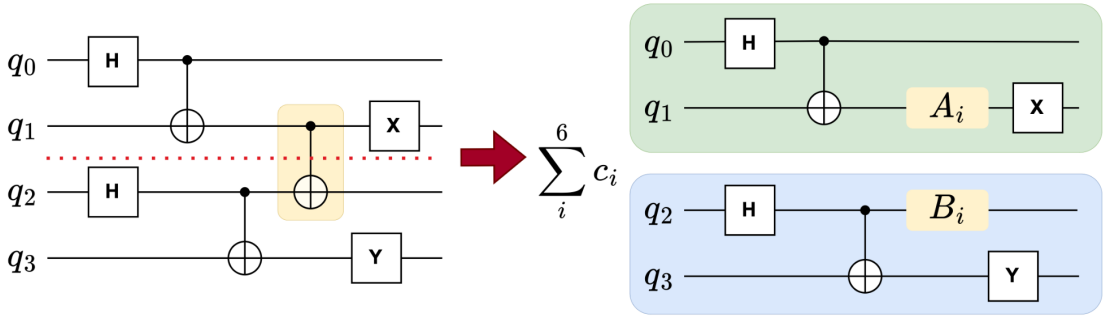
Number of subcircuits scales exponentially within the number of cuts $\mathbf{8^k}$



Wire Cutting

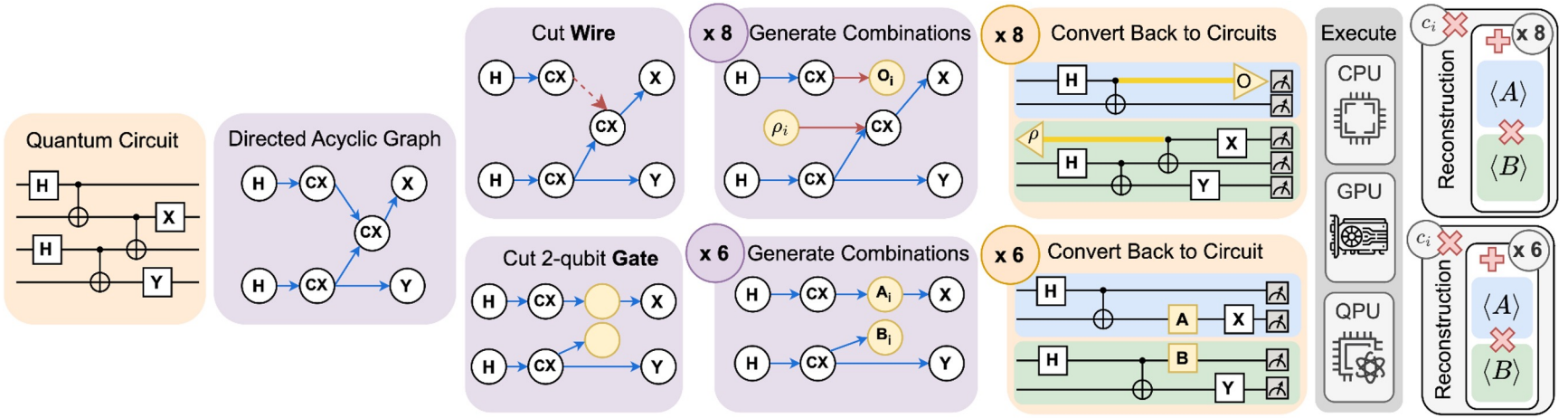
Gate cutting

Number of subcircuits scales exponentially within the number of cuts $\mathbf{6^k}$



Gate Cutting

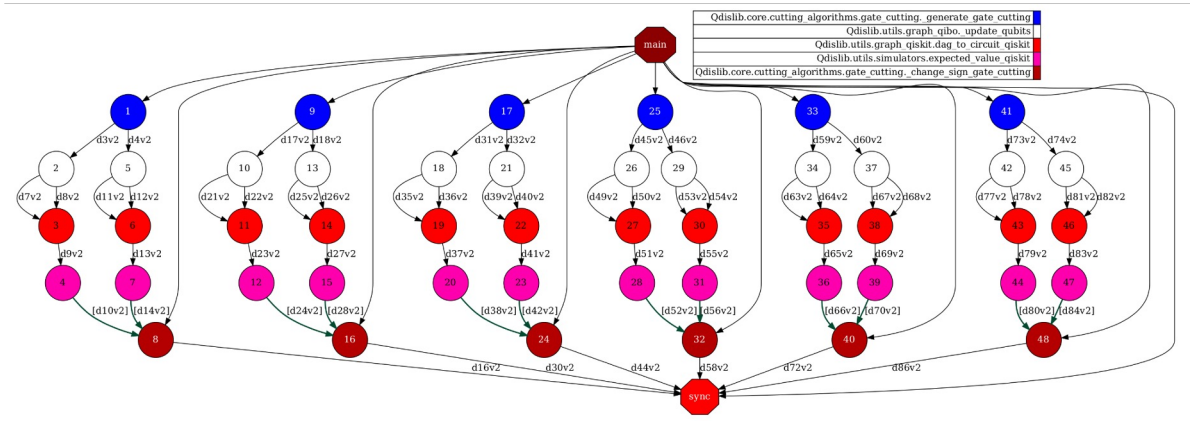
Qdislib Circuit Cutting Workflow



QDISLIB
QUANTUM DISTRIBUTED LIBRARY

ACM DL

- Portability: agnostic of the quantum API - Qibo, Qiskit, CUDAQ
- Parallelized with PyCOMPSs



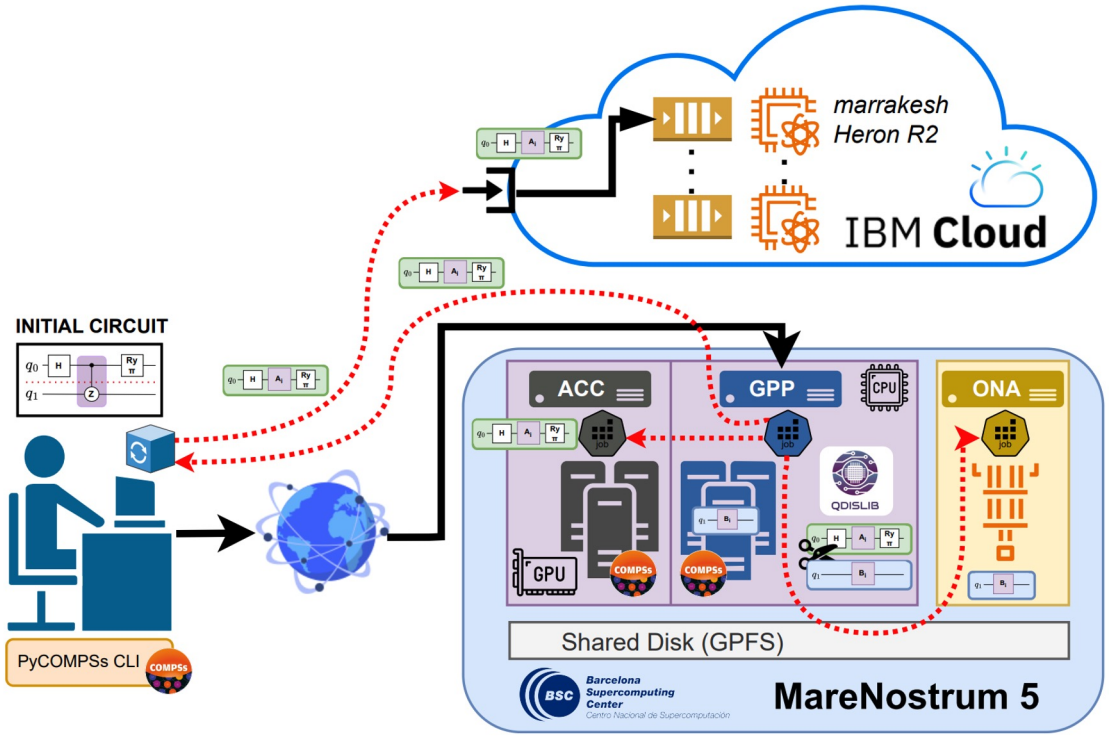
One-cut task-graph:
Number of tasks grow exponentially with cut number

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Hybrid Executions

- Marenostrum 5 GPP and ACC
- Marenostrum Ona
- IBM Quantum Cloud

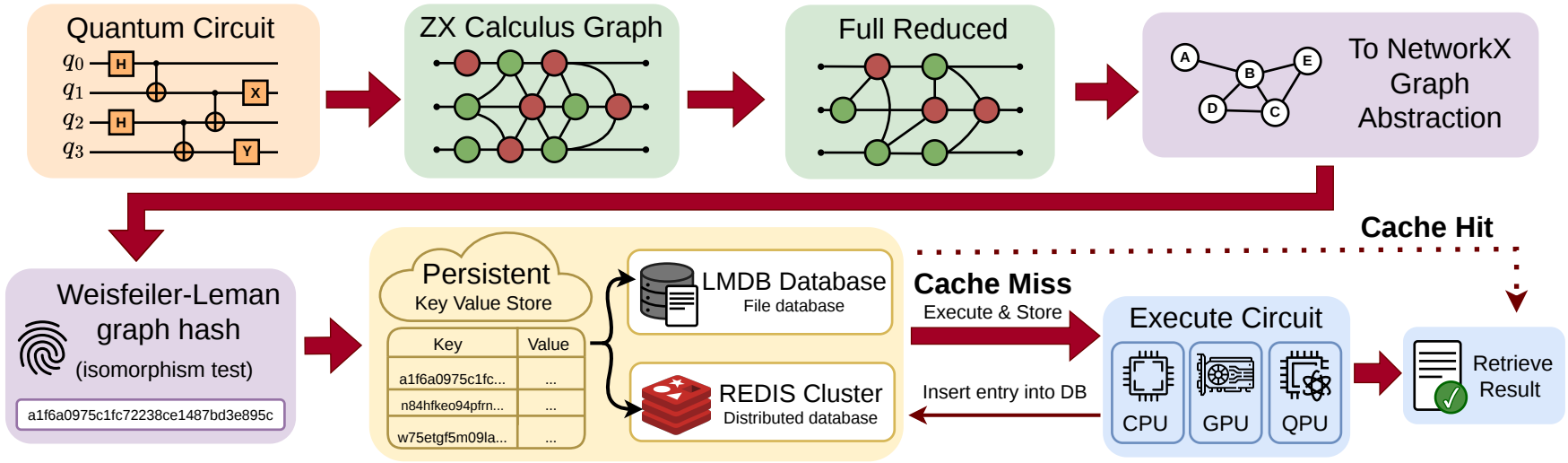
	Qubits	Cuts	CPUs ncores	GPUs ngpus	QPU nqubits	Cloud QPU nqubits	Time (s)
HEA	10	2	112	–	–	–	7.1
HEA	10	2	–	4	–	–	31.1
HEA	10	2	–	–	5	–	992
HEA	10	2	–	–	–	5	1324
HEA	32	3	112	–	5	5	2061
HEA	32	3	80	4	5	5	1597
HEA	64	2	–	4	–	39	786
HEA	96	2	–	4	–	71	803
HEA	128	2	–	4	–	103	826
RC	36	5	80	4	–	–	19347
RC	36	3	–	–	3	33	937
RC	30	3	112	–	–	18	1318
RC	30	3	–	4	–	18	1251
RC	30	3	80	4	5	5	1636
RC	30	3	–	–	5	25	854



BSC specific case

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Optimizing the performance with a circuit cache



- **Motivation**: circuit cutting generates a combinatorial number of reconstruction terms that often differ only in classical coefficients, but the quantum subcircuits are identical
- **Goal**: Reduce the number of simulations by storing previous results in a software cache
- Generally applicable, not only for circuit cutting

https://arxiv.org/abs/2604.26788

Evaluation

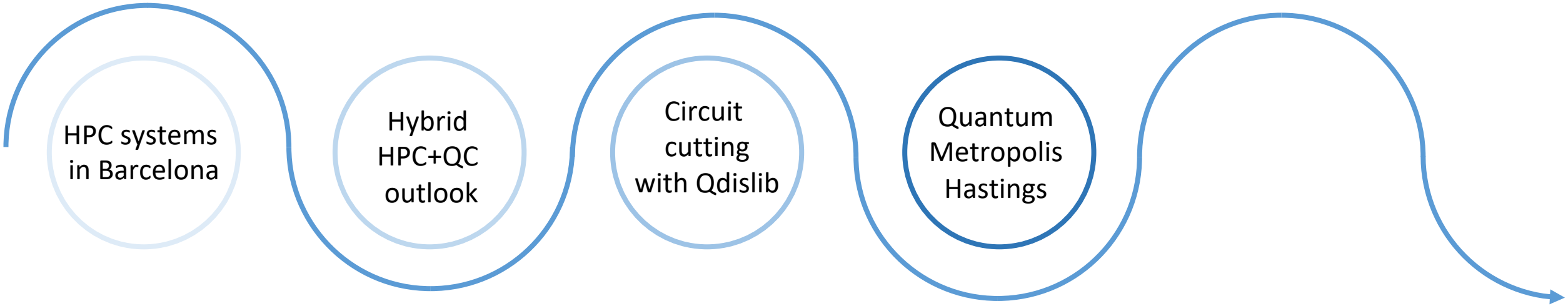
- Validation in 35-qubit MareNostrum Ona red chip
- **The circuits in the QPU are executed sequential**
- Huge speedup when using the Quantum Circuit Cache

TABLE III: Quantum hardware validation results on MareNostrum Ona, a 35-qubit superconducting quantum processor.

Configuration	Unique Circuits	Total Circuits	QPU Time (with cache)	QPU Time (without cache)	Speedup
4 cuts (HEA, 2 layers)	648	8192	1.83 hours	~20.5 hours*	11.2×
2 cuts (HEA, 1 layer)	36	128	5.86 min	17.5 min	2.98×

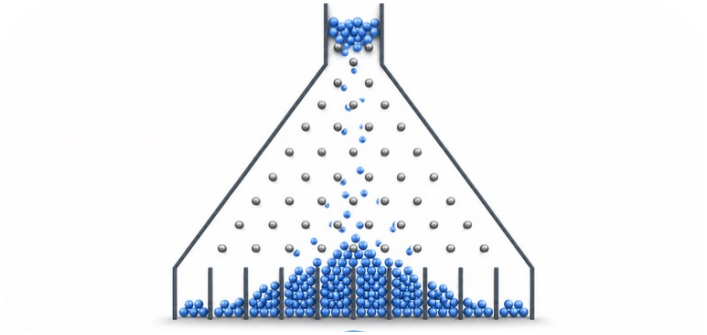
*Theoretical estimate based on simulation (8,192 circuits × 9s). The 2-cuts without cache was measured directly on QPU.

Presentation roadmap

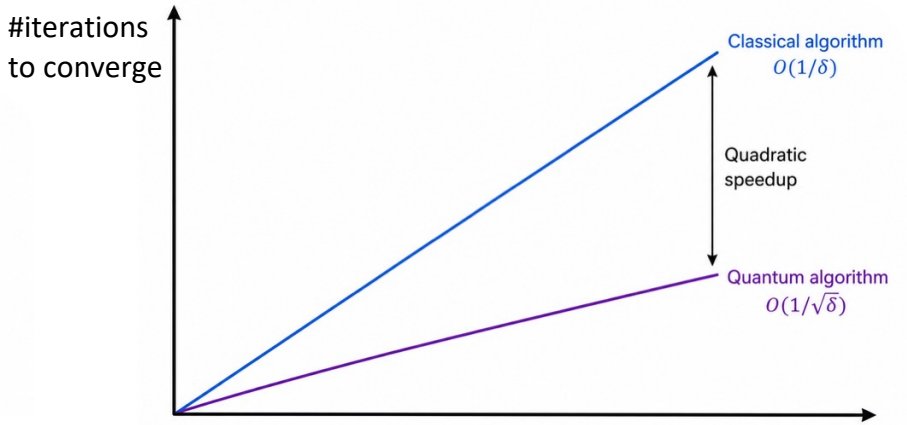


Quantum Metropolis Hastings (MH)

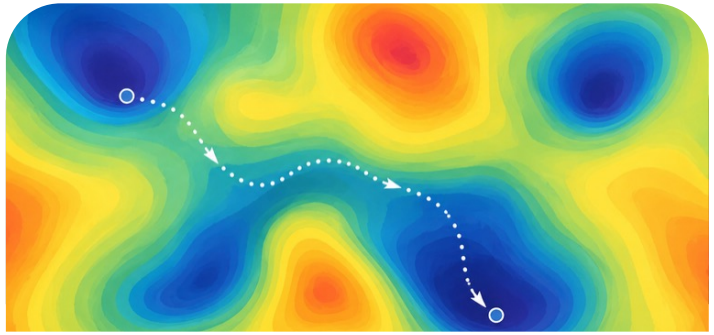
Characteristics



Fault-tolerant Markov chain Monte Carlo (MCMC) algorithm
Generally used for sampling multi-dimensional distributions, especially when the number of dimensions is high



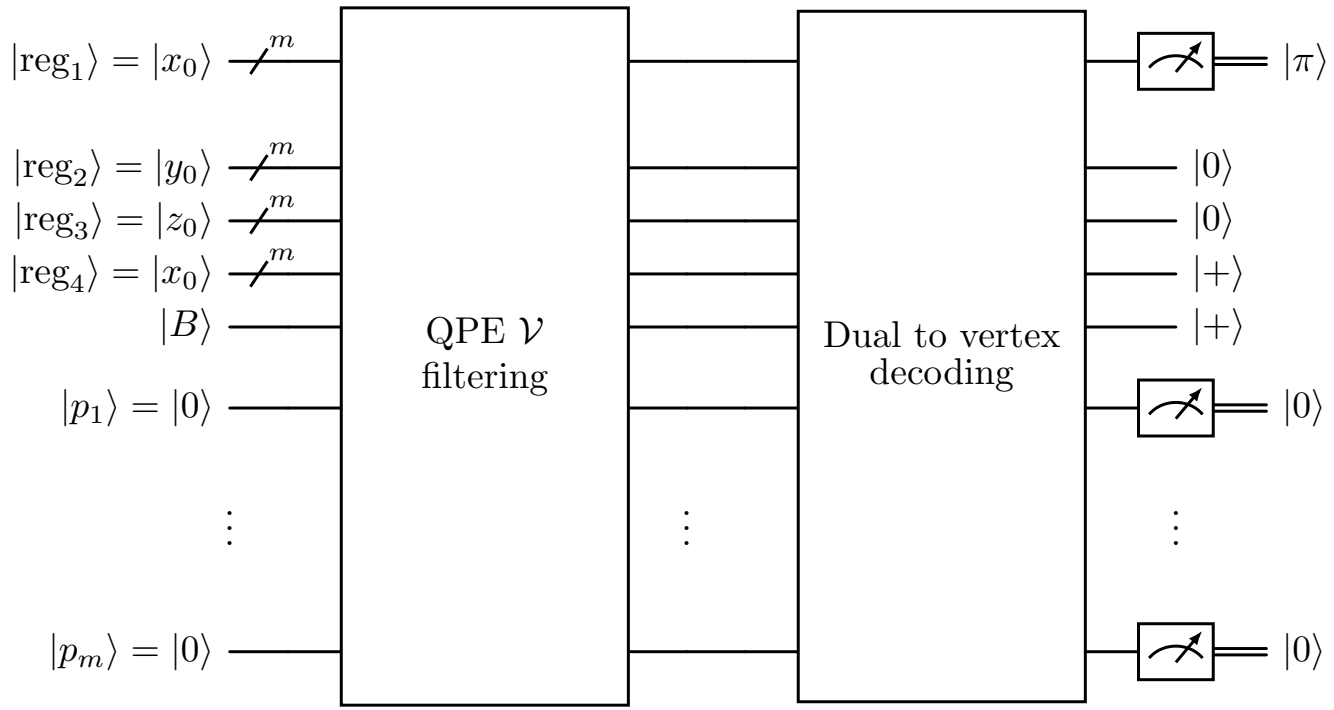
Uses quantum walks formalism to achieve a quadratic speedup converging with less iterations



Uses large structural information about the geometry of the distribution, obtaining meaningful information even when the data or sampling resolution is limited

Quantum Metropolis Hastings (MH)

- All previous implementation on MH were approximations and were not reversible
- Our work is based on a mathematical formulation by Claudon et al.
  - However, no implementation details were provided in the original article
- **First implementation of pure Quantum Metropolis Hastings algorithm**



https://arxiv.org/abs/2604.15179

Hybrid Markov Chain Monte Carlo workflow

The best of both worlds

Classical MCMC

- Computationally cheap
- Effective local exploration
- May get stuck in local minima

Quantum MCMC

- Enables global exploration
- Effective structural exploration of the whole space
- No fault-tolerant in current systems

Parallel solution

- Classical MCMC can run independent chains in parallel
- Quantum MH can run each shot independently in different QPUs

Hybrid workflow: Gravitational Wave (GW)

Collision of two massive black-holes using Bilby

1. Create GW Problem
Find out masses of two colliding black holes

→ **2.** Build a coarse model (discrete 16x16 simplification)

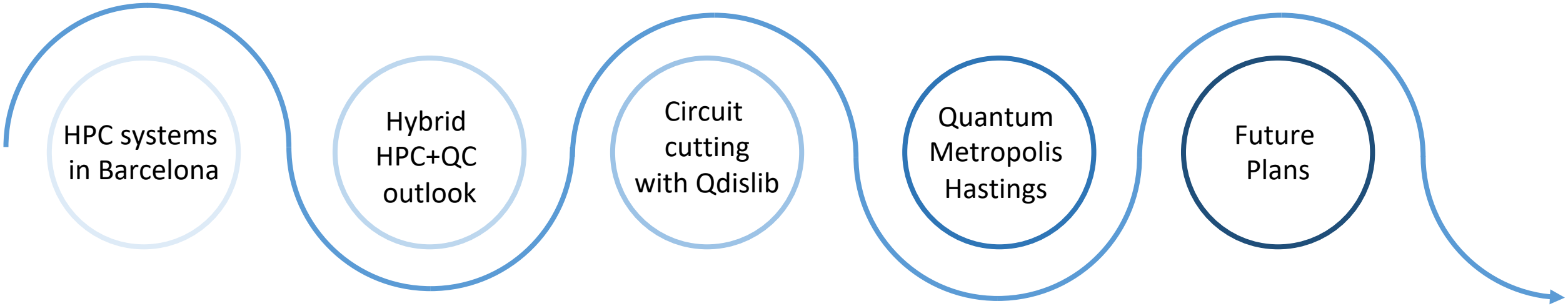
→ **3.** Run Quantum Metropolis Hastings

→ **4.** Lift back to continuous space

→ **5.** Run classical MCMC (fine refinement)

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Presentation roadmap



Future and ongoing plans

- **Quantum Circuit Optimization in Qdislib**
  - Improve circuit cutting efficiency and educe simulation overhead
  - Add new Circuit Cutting Algorithms
  - Add CC algorithms with Classical Communication

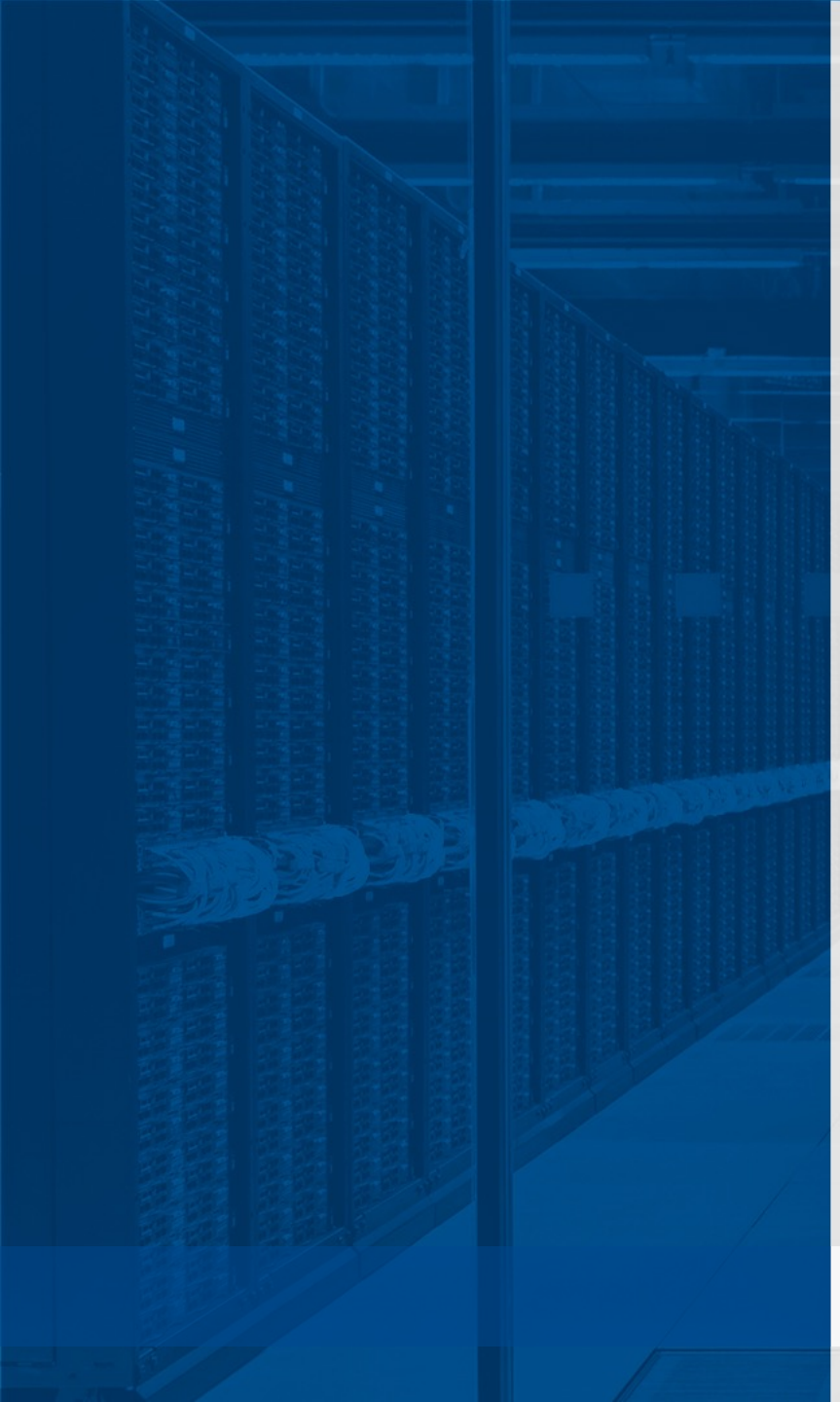
- **Implementation and hybrid workflows use cases**
  - Implement other use cases with the Metropolis Hastings
  - Workflow integrating both Digital and Annealer QC

- **Extend environment to support hybrid HPC-QC workflows**
  - Support for Distributed Quantum Computing

- **Other**
  - Initial research on QEC and its codesign with RISC-V

QDISLIB
QUANTUM DISTRIBUTED LIBRARY

Qdislib



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Thank you!

mar.tejedor@bsc.es