Enhancing Scalability of Quantum Circuits through Gate Cutting

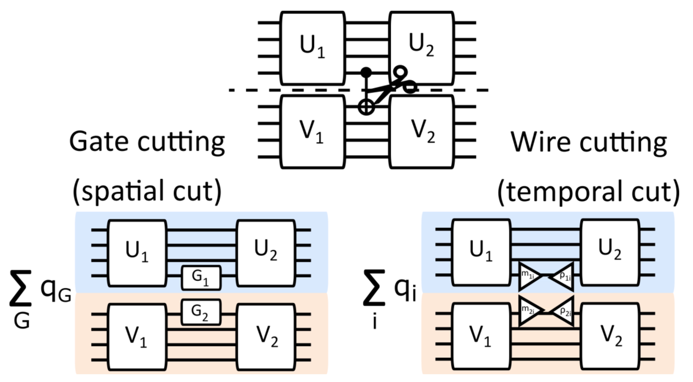
G. Díaz Camacho, C. Rodríguez Ramos, M. Mussa Juane, A. Gómez Tato
Galicia Supercomputing Center (CESGA), Avda. de Vigo, 15705 Santiago de Compostela, Spain.

In the NISQ era, quantum algorithms are limited to circuits with reduced width and depth. Hybrid classical-quantum algorithms, such as Variational Quantum Algorithms (VQAs), aim to solve the depth bottleneck problem by repeatedly running shallow parameterized circuits. However, the number of qubits in available QPUs and the memory in classical computers still limit VQAs' applicability. With the aim of building a High-Performance Quantum Computing environment, we combine HPC techniques with gate cutting to enhance scalability. This way, we can sequentially execute parts of a quantum circuit with fewer qubits or in parallel in separate computers. Here, we simulate two-qubit gates using only local gates through quasi-probabilistic decomposition, both for toy models and VQAs. While this method introduces an overhead in the number of required executions, this cost can be reasonable for low-depth quantum circuits, such as Variational Quantum Eigensolver (VQE) circuits. We explore the potential of gate cutting in VQE problems to reduce, first, the effect of noise on the ground state energy and, second, simulation resources.

1. Circuit partitioning

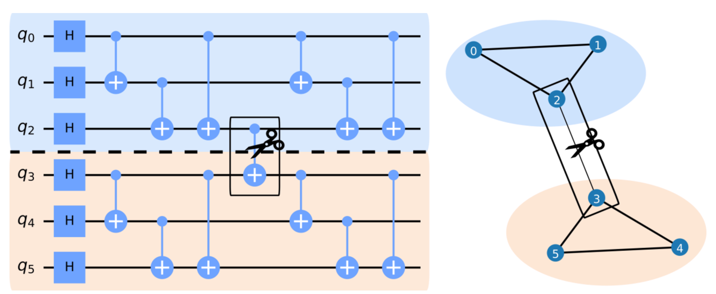
To partition circuits in two or more subcircuits, we use **Gate Cutting** of 2-qubit gates (CNOT, CZ):

- Simulate 2-qubit gates using single qubit ones
- Quasi Probabilistic Decomposition of quantum channels



Finding optimal circuit division: minimize N_{cuts}

- Equivalent to solving a graph problem: Minimum-k cut



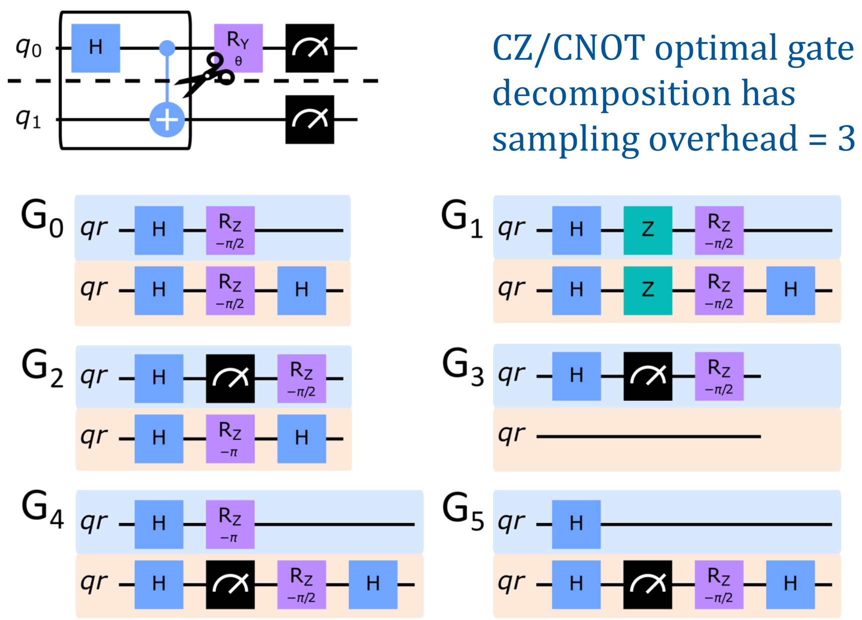
2. Quasi Probabilistic Decomposition (QPD) of Bell state

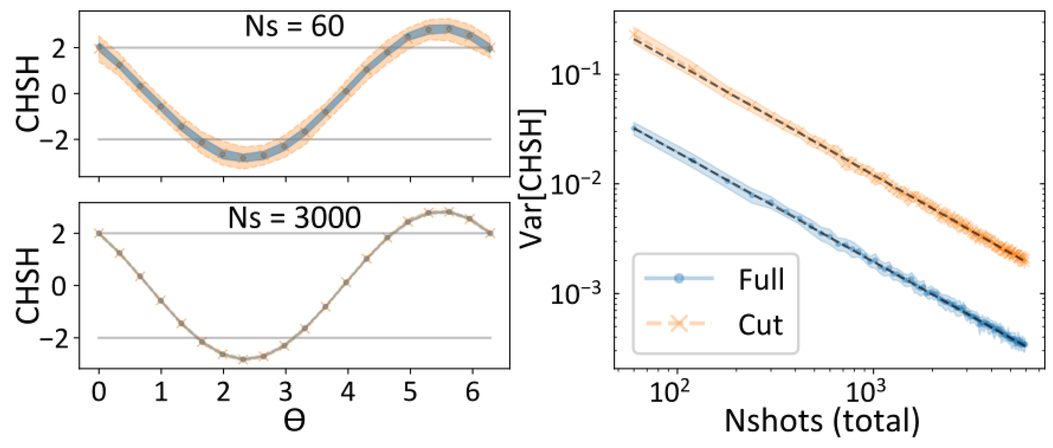
As a toy example, the outcome of a **Bell circuit** can be simulated with six weighted sets of subcircuits, which can be executed asynchronously, in smaller QPUs **[1,2].**

- Execute separately the subcircuits.
- Combine the weighted result of the subcircuits.
- Intermediate measurements affect the weight of the subcircuit.

The QPD is an unbiased estimator but incurs in a larger variance. Shots are spread between subcircuits. To recover the original precision, we need more shots/preparations: **Sampling overhead.**

- The sampling overhead is dependent on the specific state and observable of choice of the problem.
- The bound for the overhead of the original circuit/unitary is related to its Robustness of Entanglement (RoE) **[3].**
- Overhead is still exponential with $\mathbf{N_{cut}}$. Only viable for sparse, low depth circuits (such as VQAs).

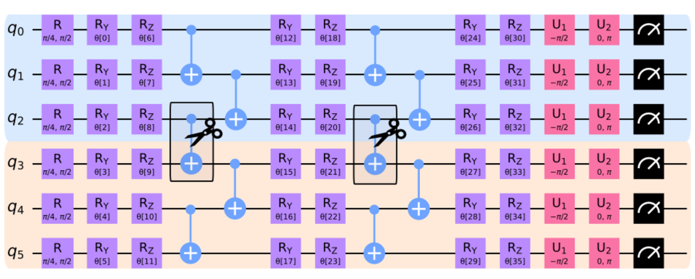




4. VQE for Ising Model

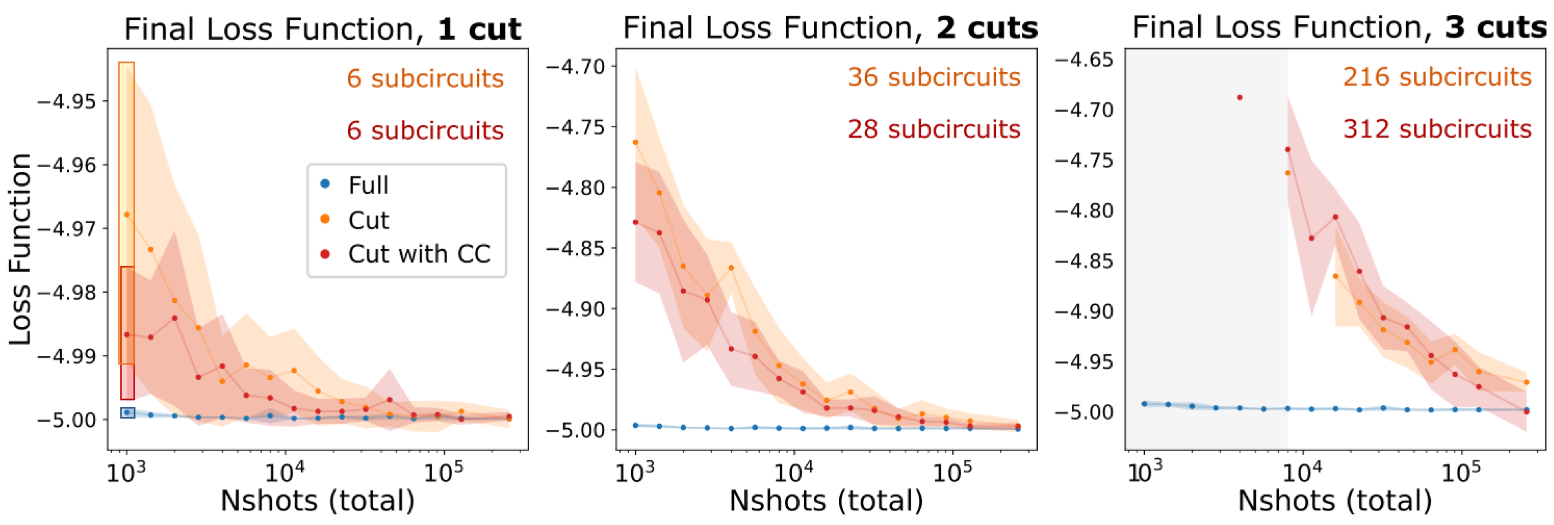
Simple optimization problem: finding ground state energy of **1D-Ising model** without external magnetic field using the **Variational Quantum Eigensolver (VQE)** algorithm **[5]**

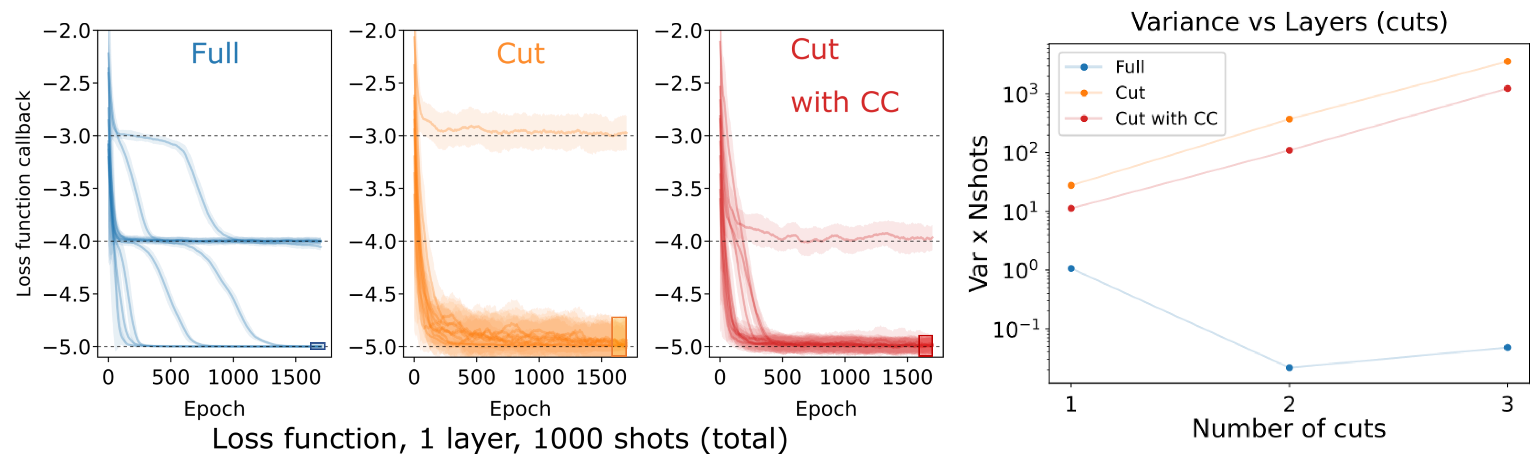
- 6 qubits, 1-3 layers of linear entanglement.
- Cut in two halves of 3 qubits: one cut per layer.



Variance grows exponentially with the number of cuts: Sampling overhead

- Smaller when using Gate Cutting with CC method

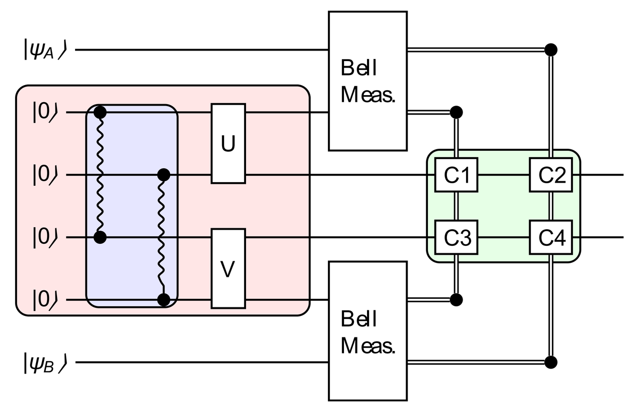




5. Gate Cutting with CC

Classical Communication (CC) can reduce the sampling overhead when cutting multiple gates **[3,4]**.

Optimal strategy: Produce Bell pairs at the beginning of the circuit (**Bell Factory**) and distribute them across the circuit via Gate Teleportation to simulate virtual CNOT/CZ gates.



See **Tino's talk** on **Thursday** for details!

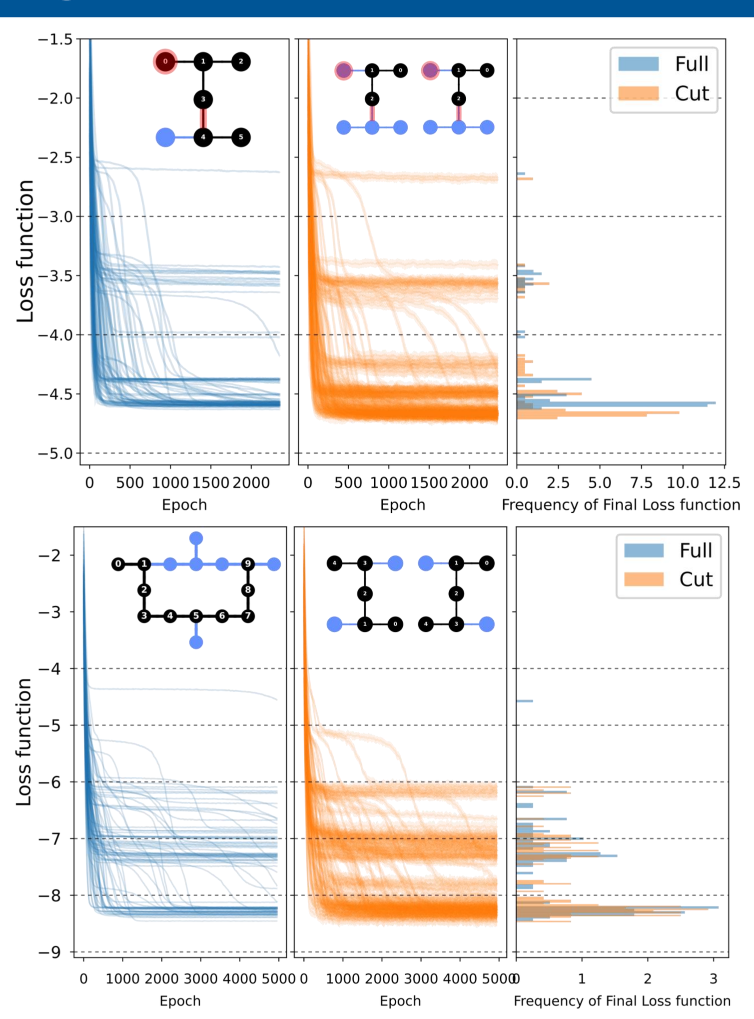
6. Advantages of Gate Cutting with several QPUs

When executing the circuits in **noisy** hardware, we can take advantage of the smaller size of the cut circuits.

- Prevent noisy qubits or faulty 2-qubit connections when transpiling to QPU.
- Reduce **swapping overhead**: less SWAP gates to adapt to the QPU topology.

With Gate Cutting we can also execute larger circuits using smaller QPUs.

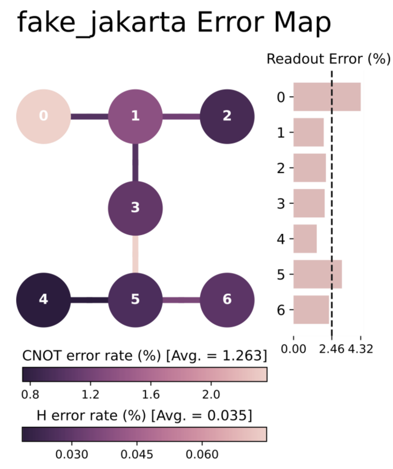
- 10-qubit 1D-Ising model using either a 16-qubit QPU (Full), or a single 7-qubit one (Cut).
- Trade-off between classical and quantum resources!



We run the same 6-qubit 1D-Ising VQE using Qiskit's 7-qubit IBMQ-Jakarta noise model.

- Smaller cut circuit can find a better layout in the QPU, and a better estimation of the ground state energy.
- Different (uncorrelated) noise profiles in separate QPUs might help with optimization.

fake_jakarta Error Map



6. Perspectives

In this work:

- Successful implementation of Gate Cutting techniques in Qiskit, both for toy models and simple VQAs.
- Exploration of Gate Cutting with CC, for future hybrid classical-quantum architecture

Next steps:

- Execute large circuits (>50 qubits) using limited resources (~30 qubits), either in classical emulation or with actual QPUs.
- Benchmark vs other Circuit Partition techniques (Wire Cutting, Entanglement Forging, etc.).
- Implement routines for finding optimal cuts in large circuits, and managing parallelization between QPUs.

References

[1] T. Peng, A. W. Harrow, M. Ozols, and X. Wu, "Simulating large quantum circuits on a small quantum computer," Physical Review Letters, vol. 125, no. 15, p. 150 504, 2020
[2] K. Mitarai and K. Fujii, "Constructing a virtual two-qubit gate by sampling single-qubit operations," New Journal of Physics, vol. 23, no. 2, p. 023 021, 2021
[3] K. Mitarai and K. Fujii, "Overhead for simulating a non-local channel with local channels by quasiprobability sampling," Quantum, vol. 5, p. 388, 2021
[4] C. Piveteau and D. Sutter, "Circuit knitting with classical communication," arXiv preprint arXiv:2205.00016, 2022
[5] Tilly, J et al. "The variational quantum eigensolver: a review of methods and best practices". Physics Reports, 986, 1-128, 2022

Contact Information
gdiaz@cesga.es
REACT-EU Project Senior Technician, CESGA

Acknowledgments
We thank the CESGA Quantum Computing group members for feedback and the stimulating intellectual environment they provide. This work was supported by Axencia Galega de Innovación through the Grant Agreement "Despregamento dunha infraestructura baseada en tecnoloxías cuánticas da información que permita impulsar a I+D+i en Galicia" within the program FEDER Galicia 2014-2020. A. Gómez was supported by MICIN through the European Union NextGenerationEU recovery plan (PRTR-C17.I1), and by the Galician Regional Government through the "Planes Complementarios de I+D+i con las Comunidades Autónomas" in Quantum Communication. Simulations are performed using the Finisterrae III Supercomputer, funded by the project CESGA-01 FINISTERRAE III.

AXENCIA GALEGA DE INNOVACIÓN
Unión Europea
Xacobeo 21·22
CESGA
GALICIA SUPERCOMPUTING CENTER