A Fidelity-Aware Framework for Joint Circuit Knitting and Error Mitigation

A Fidelity-Aware Framework for Joint Circuit Knitting and Error Mitigation

Matteo Mendula*, Deborah Volpe†, Fabrizio Riente‡, Giovanna Turvani‡, and Carla Fabiana Chiasserini‡
*Centre Tecnològic de Telecomunicacions de Catalunya, Barcelona, Spain
†Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy
‡Department of Electronics and Telecommunications, Politecnico di Torino Italy
mmendula@cttc.es, deborah.volpe@ingv.it, fabrizio.riente@polito.it, giovanna.turvani@polito.it, carla.chiasserini@polito.it

***Abstract*—Executing quantum circuits on current quantum devices is fundamentally limited by noise, decoherence, and hardware constraints. Circuit Knitting (CK) techniques enable the execution of larger circuits by decomposing them into smaller sub-circuits, at the cost of reconstruction overhead. At the same time, Quantum Error Mitigation (QEM) methods improve fidelity but introduce additional complexity. In this work, we propose a fidelity-aware framework that jointly optimizes circuit partitioning and error mitigation to maximize end-to-end execution fidelity. We introduce a unified model that captures the trade-off between sub-circuit operational fidelity, reconstruction penalties induced by cuts, and the impact of error mitigation. Based on this model, we formulate the cut placement problem as an optimization task and solve it to identify fidelity-optimal partitioning strategies accounting for error mitigation and under realistic hardware constraints. Notably, our framework integrates calibration-driven noise models and supports practical fidelity estimation using backend-specific parameters. We validate the proposed method on both *noisy simulators* (IBM Fake Backends) and *real quantum hardware* (IQM Lagrange), proving consistent improvements in execution fidelity with gains of up to 40% over standard circuit execution.**

***Index Terms*—NISQ devices, circuit cutting, circuit knitting, error mitigation, fidelity optimization**

I. Introduction

Current quantum devices in the Noisy Intermediate-Scale Quantum (NISQ) era are severely limited by noise, decoherence, and hardware constraints, whose impact increases with circuit size and depth [1], [2]. Circuit Knitting (CK) partially addresses these limitations by decomposing large circuits into smaller sub-circuits that can be executed on NISQ devices, at the cost of reconstruction overhead, and fidelity degradation [3], [4]. Quantum Error Mitigation (QEM) techniques can be used for compensating for hardware noise, but further increasing execution cost and interacting non-trivially with circuit partitioning [5], [6]. Therefore, executing effectively on NISQ devices entails making choices not only regarding how to break up the circuits, but whether or not it is worthwhile doing so from the standpoint of improved fidelity.

Existing approaches typically address CK and QEM as separate stages of the execution pipeline. CK methods focus on feasible partitioning strategies under hardware constraints [3], [4], [7], while QEM approaches aim at reducing noise at the circuit or sub-circuit level [5], [6]. However, treating these steps independently can lead to sub-optimal execution plans, since aggressive partitioning may improve sub-circuit reliability but also amplify sampling and reconstruction costs, whereas limited partitioning preserves correlations but leaves deeper, noisier circuits.

Fig. 1. Fidelity-aware circuit partitioning and error mitigation workflow.

In this work, we propose MOSAIQ (Mixed-Integer Optimization for Sub-circuit Aware Integrated Quantum execution), a hardware-aware software framework for jointly reasoning about circuit knitting and error mitigation. MOSAIQ models the combined impact of circuit depth, number of cuts, backend-specific noise, reconstruction penalties, and mitigation strategies, and formulates cut placement as an optimization problem. This workflow, shown in Fig. 1, enables execution planning for fidelity-aware computation with respect to practical hardware limitations and latencies. In this first validation experiment, we consider only spatial cuts to highlight the CK vs QEM trade-off while avoiding the Measure-and-Prepare errors incurred through temporal cuts.

MOSAIQ exploits backend calibration data to estimate execution fidelity and adapt partitioning decisions to device-specific noise characteristics. We validate the framework on noisy IBM fake backends and on real quantum hardware, namely the IQM Lagrange processor at Politecnico di Torino. The results show that circuit cutting alone is not always beneficial, and that joint CK–QEM optimization is needed to improve end-to-end execution fidelity under realistic reconstruction and latency constraints.

The contributions of this work are:

- A calibration-driven noise model for realistic fidelity estimation;
- A joint optimization framework for circuit partitioning and error mitigation;
- Validation of the framework's spatial partitioning capabilities on both simulated and real quantum hardware.
- Experimental results highlighting the importance of jointly optimizing circuit partitioning and QEM for improved end-to-end fidelity on NISQ devices.

II. The MOSAIQ Framework

MOSAIQ models a quantum circuit as a sequence of layers $\mathcal{L} = \{0, \dots, L\}$, where each layer ℓ acts on $0 < Q_\ell \leq Q$ active qubits. The goal is to select a set of cuts $\mathbf{P}$ that

partitions the circuit into sub-circuits $\mathcal{S} = \{S_1, \ldots, S_N\}$, where each sub-circuit S_i has width q_i and depth Δ_i. Although the formulation can support both spatial and temporal cuts, this work focuses on spatial cuts to isolate the interplay between circuit depth, reconstruction overhead, and QEM without introducing intermediate Measure-and-Prepare errors.

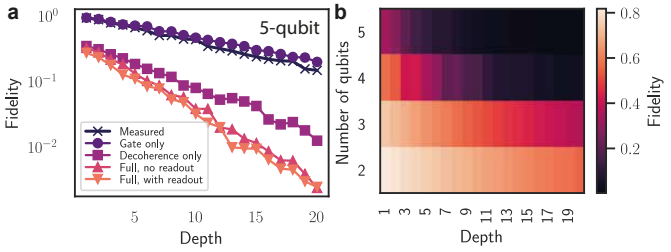

Fig. 2. Fidelity vs. circuit depth. **a**: Experimental results compared to individual contributions of the considered noise model, for a 5-qubit system. **b**: Fidelity heatmap vs. circuit depth and number of qubits (2–5) under a full noise model without readout.

Fidelity-aware model. We estimate execution fidelity using a backend-calibrated multiplicative noise model, where gate errors, decoherence, and readout errors are obtained from device calibration data [1], [6], [8]–[10] (Fig. 2). For a sub-circuit S_i, the operational fidelity is approximated as an exponential decay with width and depth,

$$F_{\text{ops}}(S_i) = e^{-\epsilon q_i \Delta_i}, \tag{1}$$

where ϵ denotes an effective backend-dependent error rate, computed by Qiskit calibration data combined as in [1], [6], [8]–[10]. To capture the effect of QEM, we introduce an error-suppression factor $\mu \in (0, 1]$, yielding

$$F^{\text{mit}}_{\text{ops}}(S_i) = e^{-(\mu\epsilon) q_i \Delta_i}. \tag{2}$$

Each cut $c \in \mathbf{P}$ introduces a reconstruction penalty $\eta_c \in (0, 1)$, accounting for sampling overhead and recombination errors. The end-to-end fidelity of a partitioned and mitigated execution is therefore

$$F_{\text{total}}(\mathbf{P}) = \prod_{S_i \in \mathcal{S}} F^{\text{mit}}_{\text{ops}}(S_i) \prod_{c \in \mathbf{P}} \eta_c. \tag{3}$$

This formulation captures the central MOSAIQ trade-off: cuts reduce the size and depth of the executed sub-circuits, but introduce reconstruction penalties and additional sampling cost.

Latency constraint. For each cut, reconstruction requires measurements over multiple Pauli bases, increasing the number of shots by a factor γ. QEM introduces an additional multiplier M_{QEM}. Hence,

$$N_{\text{total}} = N_{\text{base}} \gamma^{|\mathbf{P}|} M_{\text{QEM}}. \tag{4}$$

Assuming a constant time per shot t_{shot}, feasible executions must satisfy

$$t_{\text{shot}} N_{\text{base}} \gamma^{|\mathbf{P}|} M_{\text{QEM}} \leq T_{\max}, \tag{5}$$

where $T_{\max}$ is the maximum allowed execution latency. Taking logarithms yields the linear cut-budget constraint

$$\sum_{c \in \mathbf{P}} \log(\gamma) \leq \log\left(\frac{T_{\max}}{t_{\text{shot}} N_{\text{base}} M_{\text{QEM}}}\right). \tag{6}$$

Optimization problem. MOSAIQ selects the cut configuration maximizing the logarithm of Eq. (3) subject to hardware and latency constraints:

$$\max_{\mathbf{P}} \quad -\sum_{S_i \in \mathcal{S}} \mu\epsilon q_i \Delta_i + \sum_{c \in \mathbf{P}} \log(\eta_c) \tag{7}$$

$$\text{s.t.} \quad p_{s,t} \leq Q_\ell, \quad \forall (s,t),\ \ell \in \text{sub-circuit}(s,t), \tag{8}$$

$$\sum_{s,t} p_{s,t} \log(\gamma) \leq \log\left(\frac{T_{\max}}{t_{\text{shot}} N_{\text{base}} M_{\text{QEM}}}\right), \tag{9}$$

$$p_{s,t} \in \{0, 1\}.$$

The binary variable $p_{s,t}$ indicates whether a cut is placed at candidate position (s, t). Constraint (8) enforces the available hardware width, while Constraint (9) limits the number of cuts according to the latency budget. The resulting Mixed-Integer Linear Programming (MILP) problem can be solved with standard optimization solvers such as Gurobi[1].

III. Evaluation

Experimental protocol. We evaluate MOSAIQ on parameterized 3-qubit circuits targeting the $\langle ZZZ \rangle$ observable. The circuits are composed of RX/RY rotations and CZ gates, and include a spatial cut on the middle wire to study the trade-off between partitioning, mitigation, and reconstruction overhead. We consider three depths: *Shallow*, *Original*, and *Deep*, obtained by progressively increasing the number of rotation and entangling layers.

We compare three strategies: (i) full execution without partitioning; (ii) baseline CK, using a deterministic spatial bisection on the middle wire; and (iii) MOSAIQ, combining optimized cut placement with QEM, including ZNE and SPAM correction. Experiments are performed on IBM noisy fake backends (*FakeLimaV2*, *FakeManilaV2*, *FakeJakartaV2*) and on the IQM Lagrange 5-qubit superconducting processor at Politecnico di Torino [11], [12]. All measured expectation values are compared against the corresponding noiseless reference.

Fidelity across depths and backends. Fig. 3 shows that, as circuit depth increases, full execution degrades because of noise accumulation, while CK alone provides limited or even negative gains due to reconstruction errors. On the other hand, CK combined with QEM systematically moves the measured $\langle ZZZ \rangle$ closer to the ideal reference, with the largest improvements on deeper circuits. The trend is consistent across *FakeLimaV2*, *FakeManilaV2*, and *FakeJakartaV2*, indicating robustness to different noise models.

Real-hardware validation. Fig. 4 compares the Original circuit on IQM Lagrange and the IBM fake backends. The results confirm the simulation trend: full execution deviates

[1]https://www.gurobi.com

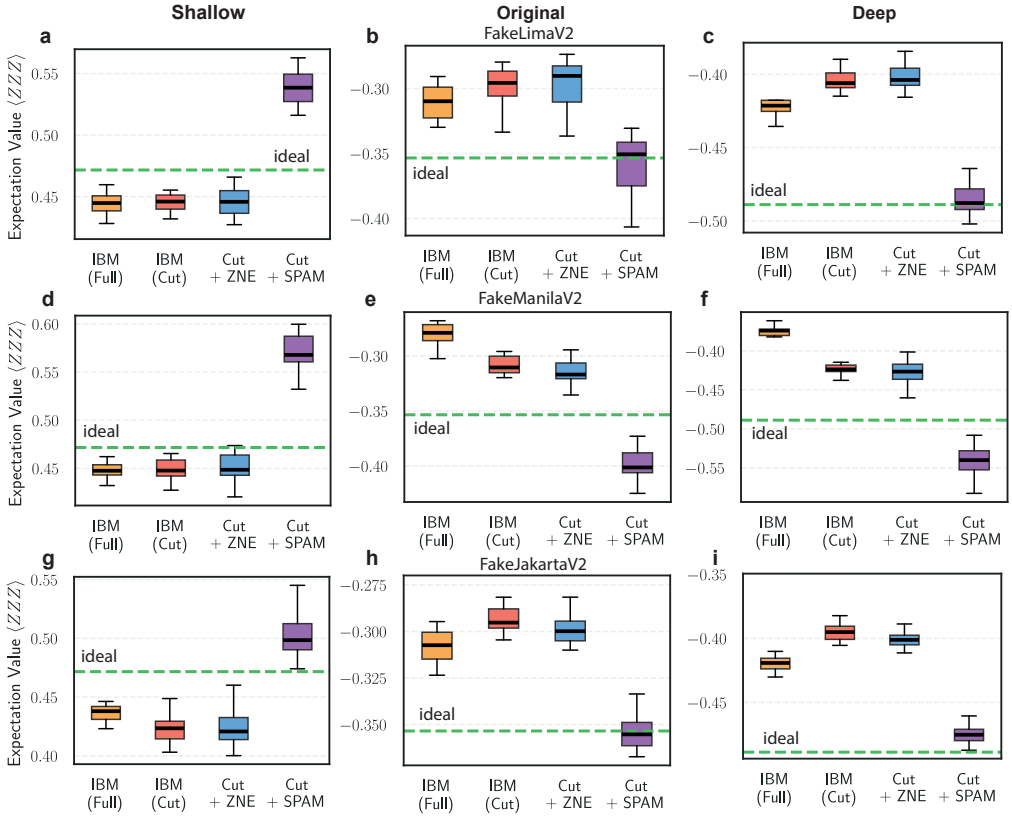

Fig. 3. Expectation values across circuit depths and IBM fake backends. Boxplots compare full execution, CK, and CK with QEM over $N = 8{,}192$ shots; the dashed line marks the ideal value.

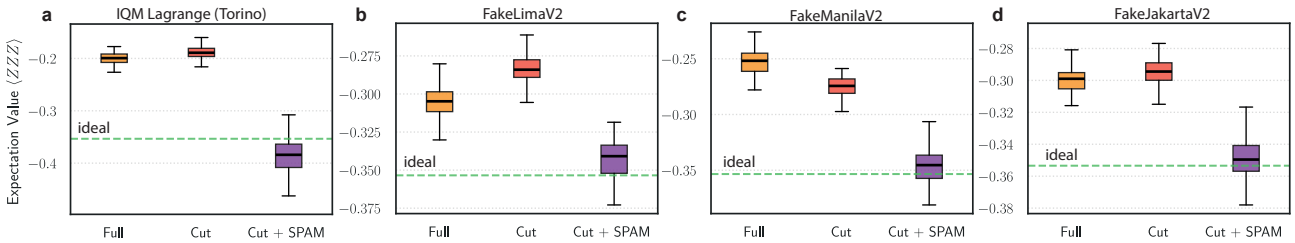

Fig. 4. Expectation values on IQM Lagrange and noisy simulators for the Original circuit. Boxplots compare full execution, CK, and CK with SPAM mitigation; the dashed line marks the ideal value.

from the ideal value, CK alone is inconsistent, and CK with QEM yields values closer to the ideal reference. Although absolute values vary across platforms because of hardware-specific noise and runtime variability, the relative behavior remains consistent.

Cut placement under hardware constraints. We then evaluate MOSAIQ on a representative 30-qubit, 60-layer target circuit by varying the available hardware width $Q_{\max}$ and the latency budget. Fig. 5 shows the resulting fidelity–latency trade-off: limited $Q_{\max}$ requires more aggressive partitioning, which reduces sub-circuit size but increases sampling overhead and may violate latency constraints.

Table I presents the corresponding MILP-selected cut configurations and fidelity gains relative to an unmitigated baseline. Configurations with too many cuts are rejected as infeasible when their sampling cost exceeds the latency budget. Moreover, the benefit of partitioning depends strongly on the reconstruction penalty η_c: high penalties can make aggressive cutting detrimental, whereas optimistic penalties allow MOSAIQ to identify feasible configurations with positive fidelity gains.

Overall, the evaluation confirms that optimal partitioning depends jointly on circuit structure, backend constraints, latency budget, and reconstruction penalties. CK is advantageous only

TABLE I

OPTIMAL CUT CONFIGURATIONS AND RELATIVE FIDELITY IMPROVEMENT OVER UNMITIGATED BASELINE EXECUTION

Time Budget	$Q_{\max}$	No. Cuts	Width Cuts (Qubits)	Latency (h)	Fidelity Gain (%)		
					η_c=0.80	η_c=0.90	η_c=0.99
4h	4, 6			Infeasible			
	8	4	8, 16, 22	3.50	-41.3	-6.0	+37.7
	10	3	10, 20	0.87	-26.6	+4.5	+39.1
	16	2	14	0.22	-8.3	+16.1	+40.5
36h	4			Infeasible			
	6	5	6, 12, 18, 24	13.98	-53.0	-15.4	+36.3
	8	4	8, 16, 22	3.50	-41.3	-6.0	+37.7
	10	3	10, 20	0.87	-26.6	+4.5	+39.1
	16	2	14	0.22	-8.3	+16.1	+40.5

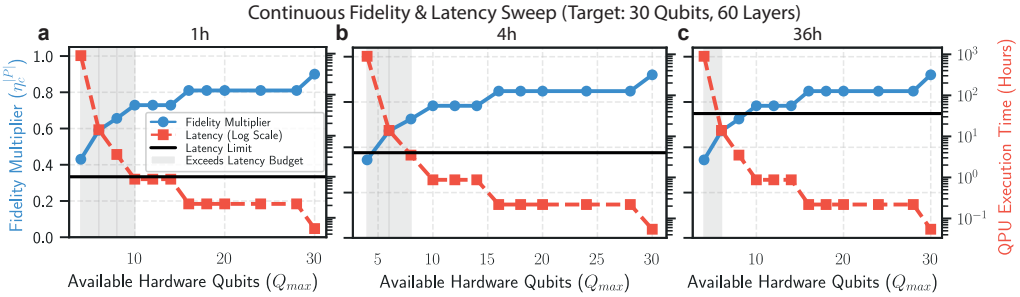

Fig. 5. Fidelity–latency trade-off for a 30-qubit, 60-layer target circuit under different latency budgets. Shaded regions indicate infeasible configurations.

when the fidelity improvement enabled by smaller mitigated sub-circuits outweighs the cumulative cost of recombination and sampling.

IV. Discussion

The results show that circuit cutting provides advantages only when the fidelity gain obtained from executing smaller mitigated sub-circuits outweighs reconstruction penalties and sampling overhead. Thus, cut placement cannot be treated as an isolated partitioning problem: it must be jointly optimized with QEM, backend noise, and latency constraints. As shown in Fig. 5, aggressive partitioning can improve sub-circuit reliability but may exceed feasible execution time, while limited partitioning reduces recombination cost at the expense of deeper, noisier circuits.

MOSAIQ currently focuses on spatial cuts and uses a backend-calibrated but independent Markovian noise model. This excludes correlated errors, crosstalk, leakage, and the additional SPAM errors introduced by temporal Measure-and-Prepare cuts. Moreover, while the MILP formulation is tractable for the evaluated circuits, larger instances may require heuristic or decomposed solvers. Future work will extend MOSAIQ with entanglement-aware and SPAM-aware penalties, larger-circuit validation, and tighter integration with quantum compilation and scheduling workflows.

V. Conclusions

In this work, we presented a fidelity-aware framework for CK that jointly optimizes circuit partitioning and error mitigation. By introducing a unified model that captures the combined effects of operational fidelity, reconstruction penalties, and mitigation strategies, we formulated the cut placement problem as a MILP optimization task. The proposed approach enables the identification of partitioning strategies that maximize end-to-end execution fidelity under realistic hardware constraints. Experimental validation on both noisy simulators and real quantum hardware (IQM Lagrange) proves consistent improvements with gains of up to 40% over standard circuit execution, highlighting the effectiveness of jointly optimizing CK and QEM. Overall, our results show that the co-design of circuit decomposition and error mitigation is essential to fully exploit the potential of circuit knitting in NISQ systems. By explicitly accounting for hardware limitations, noise characteristics, and execution constraints, our proposed methodology provides a practical approach towards more reliable quantum computations on NISQ devices.

References

[1] J. Preskill, "Quantum computing in the nisq era and beyond," *Quantum*, vol. 2, p. 79, 2018.

[2] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*, 10th ed. Cambridge ; New York: Cambridge University Press, 2010.

[3] M. Bechtold, J. Barzen, M. Beisel, F. Leymann, and B. Weder, "Patterns for quantum circuit cutting," in *Proceedings of the 30th Conference on Pattern Languages of Programs*, 2023, pp. 1–12.

[4] J. Nivala, "Quantum circuit knitting an overview of existing quantum circuit knitting methods," 2024.

[5] A. He, B. Nachman, W. A. de Jong, and C. W. Bauer, "Zero-noise extrapolation for quantum-gate error mitigation with identity insertions," *Physical Review A*, vol. 102, no. 1, p. 012426, 2020.

[6] Z. Cai, R. Babbush, S. C. Benjamin, S. Endo, W. J. Huggins, Y. Li, J. R. McClean, and T. E. O'Brien, "Quantum error mitigation," *Reviews of Modern Physics*, vol. 95, no. 4, p. 045005, 2023.

[7] S. Brandhofer, I. Polian, and K. Krsulich, "Optimal partitioning of quantum circuits using gate cuts and wire cuts," *IEEE Transactions on Quantum Engineering*, vol. 5, pp. 1–10, 2023.

[8] A. Morvan, B. Villalonga, X. Mi, S. Mandrà, A. Bengtsson, P. Klimov, Z. Chen, S. Hong, C. Erickson, I. Drozdov *et al.*, "Phase transitions in random circuit sampling," *Nature*, vol. 634, no. 8033, pp. 328–333, 2024.

[9] Y. Zhou, E. M. Stoudenmire, and X. Waintal, "What limits the simulation of quantum computers?" *Physical Review X*, vol. 10, no. 4, p. 041038, 2020.

[10] N. Samos, R. Bistroń, M. Rudziński, R. M. C. Pereira, K. Życzkowski, and P. Ribeiro, "Fidelity decay and error accumulation in random quantum circuits," *SciPost Physics*, vol. 19, no. 1, p. 013, 2025.

[11] P. Viviani, F. Bertone, G. Vitali, E. Dri, F. Stirano, G. Caragnano, F. Lubrano, A. Nespola, O. Terzo, M. Cocuzza, B. Montrucchio, G. Turvani, G. Bertaina, M. Coisson, D. Calonico, F. Pirri, and P. Asinari, "Lagrange: Operating italy's first publicly-accessible quantum computer for research and education," 2026. [Online]. Available: https://arxiv.org/abs/2604.21695

[12] J. Rönkkö, O. Ahonen, V. Bergholm, A. Calzona, A. Geresdi, H. Heimonen, J. Heinsoo, V. Milchakov, S. Pogorzalek, M. Sarsby *et al.*, "On-premises superconducting quantum computer for education and research," *EPJ Quantum technology*, vol. 11, no. 1, p. 32, 2024.