# Latency-constrained Fidelity Optimization for NISQ Execution via Quantum Circuit Knitting

M. Mendula*, D. Volpe†, G. A. Massa‡, F. Riente‡, G. Turvani‡, C. F. Chiasserini‡§
*CTTC, Barcelona, Spain; †Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy
‡ Politecnico di Torino, Italy; §CNIT, Parma, Italy
deborah.volpe@ingv.it, s359855@studenti.polito.it,
fabrizio.riente@polito.it, giovanna.turvani@polito.it, carla.chiasserini@polito.it

***Abstract*—Executing quantum circuits on NISQ devices poses challenges due to noise, decoherence, and limited qubit counts. Circuit Knitting (CK) allows solving these issues by breaking down large circuits into smaller sub-circuits. However, the subsequent reconstruction adds sampling and processing time. Quantum Error Mitigation (QEM) can enhance output quality, but it also raises execution costs. To address the above issues, we introduce MOSAIQ, a cost-aware framework optimizing both circuit partitioning and error mitigation for reliable NISQ execution. MOSAIQ uses calibration-driven noise estimation, mitigation-aware fidelity modeling, and cut-dependent reconstruction costs to choose partitioning strategies that balance end-to-end fidelity with practical feasibility. We define the problem as an optimization task and test the MOSAIQ on noisy simulators and actual quantum HW. Results indicate that MOSAIQ prevents overly aggressive partitioning and improves the balance between fidelity, sampling overhead, and execution time.**

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

## I. Introduction

Quantum computing is a promising computational paradigm [1], with the potential to accelerate the solution of challenging problems such as optimization, simulation, cryptography, and machine learning. However, current Noisy Intermediate-Scale Quantum (NISQ) devices are limited by a small number of reliable qubits, noisy operations, restricted connectivity, device-dependent calibration profiles, and non-negligible execution latency [2]. Thus, executing quantum workloads is not only a Hardware (HW)-level challenge, but also a software-defined management problem: quantum circuits must be adapted to the available backend resources while preserving the reliability of the final result.

Circuit Knitting (CK) [3], [4] mitigates NISQ resource limitations by decomposing large circuits into smaller sub-circuits that are executed independently and recombined classically. While this can make infeasible workloads executable on constrained hardware, each cut introduces sampling overhead, reconstruction cost, and possible fidelity loss, especially for two-qubit gate cuts. Quantum Error Mitigation (QEM) [5], [6] improves NISQ reliability by reducing the impact of HW noise on measured observables, but it also increases sampling and computational cost. Therefore, partitioned and mitigated execution must balance sub-circuit fidelity, mitigation gains, reconstruction overhead, and latency constraints. Fidelity-optimal partitions may become impractical due to cut reconstruction costs, while overly conservative ones may miss the benefits of smaller, less noisy sub-circuits.

Existing approaches often treat CK, and QEM as separate stages of the quantum execution pipeline. This may lead to sub-optimal execution strategies on real quantum hardware, where device constraints, noise levels, and execution costs vary over time. Cut placement should thus be treated as an optimization problem that jointly accounts for the expected fidelity gain and the operational cost of the selected cuts. This is critical for emerging quantum–classical infrastructures, where the ability to select feasible and cost-effective execution configurations is as important as improving fidelity.

In this work, we propose MOSAIQ, a framework that jointly optimizes circuit partitioning and error mitigation under HW, latency, and reconstruction-cost constraints (Fig. 1). MOSAIQ treats cut placement as a software-defined execution decision, selecting partitions that improve end-to-end quality while limiting the overhead of two-qubit gate reconstruction. We formulate the resulting cut placement problem as an optimization task and use it to balance fidelity gains and operational feasibility. Evaluation on noisy simulators and real quantum hardware shows that MOSAIQ avoids overly aggressive decompositions and improves the trade-off among fidelity, sampling overhead, and reconstruction cost.

Our main contributions are thus as follows:

- We introduce a cost-aware CK framework that treats quantum circuit partitioning as a software-defined execution problem over constrained NISQ resources.
- We address fidelity-aware cut placement by modeling the reconstruction cost of two-qubit gate cuts, together with sampling overhead, mitigation cost, and latency constraints.
- We formulate the cut placement optimization jointly capturing sub-circuit fidelity, error mitigation benefits, backend-specific HW constraints, and recombination overhead.
- We integrate calibration-driven noise information into the optimization flow, enabling backend-aware fidelity and cost estimation for realistic NISQ execution scenarios.
- We validate MOSAIQ on noisy simulators and real quantum HW, showing the trade-off between execution fidelity, reconstruction cost, and operational feasibility.

This work was supported by the Centro meridionale per l'innovazione quantistica PN RIC 21-27, Q-Sud (D.D 307/2025, CUP B99H26000440007), and SNS JU under the EU's HE Grant Agreement No. 101192521.

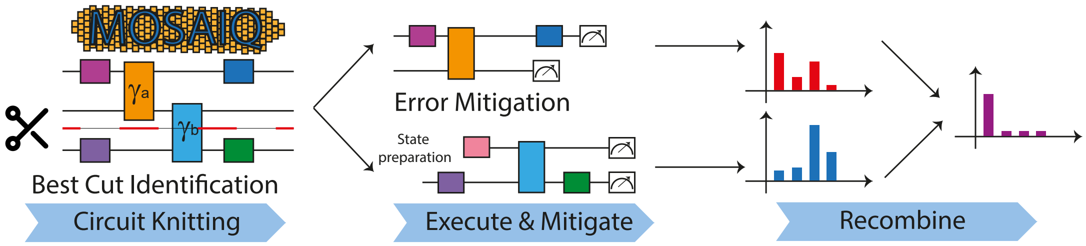


Fig. 1. MOSAIQ framework for fidelity-aware circuit partitioning and mitigation.

In the rest of the paper, Sec. II presents the motivation and an overview of MOSAIQ; Sec. III describes the reconstruction-aware fidelity and cost model, and the optimization formulation; Sec. IV presents the simulation and quantum HW setup, our benchmarks, and the experimental results; Sec. VI draws our conclusions and future research directions.

## II. Background and Problem Overview

**Quantum Circuit Execution on NISQ Devices.** Quantum algorithms are commonly represented as circuits, i.e., ordered sequences of gates applied to qubits. Their width $W$ and depth $D$ denote, respectively, the number of required qubits and sequential gate layers [1]. On NISQ devices, these quantities determine both feasibility and reliability [2]: circuits may exceed the available qubit count, while deeper executions are more exposed to gate errors, decoherence, and readout noise.

Execution quality further depends on backend-specific constraints, such as noisy gates, limited connectivity, and time-varying calibration data. Since quantum outputs are estimated from repeated shots, any increase in the number of circuit evaluations directly increases execution time and cost.

NISQ execution is not only a HW but also a software-defined resource management problem. A quantum workload must be adapted to constrained and heterogeneous backends, while accounting for qubit availability, circuit depth, noise, queueing and latency, and the classical post-processing required to obtain the final result. This makes quantum execution increasingly relevant for cloud-native and software-defined infrastructures, where orchestration decisions determine whether a workload is feasible, reliable, and cost-effective.

**CK and Reconstruction Overhead.** CK is designed to execute quantum circuits that are too large or noisy to be run directly on a target device [3], [4]. In general, two main cutting primitives can be considered. A wire cut interrupts the evolution of a qubit line and replaces the missing quantum channel with a set of measurement-and-preparation operations. A gate cut, instead, decomposes a non-local or two-qubit operation into a linear combination of local operations that can be executed in separate sub-circuits. In both cases, the final observable is reconstructed by combining the results of multiple sub-circuit evaluations. This reconstruction process increases the sampling cost. Notably, each cut may require the execution of several circuit variants over different measurement, preparation. The total number of required shots can thus grow rapidly with the number of cuts. This is especially relevant for two-qubit gate cuts, where the decomposition of an entangling operation introduces multiple reconstruction terms, each associated with a coefficient that contributes to the total estimator variance. Thus, a partitioning strategy that improves sub-circuit fidelity may still be impractical if it induces excessive reconstruction overhead.

**QEM and Execution Costs.** QEM improves NISQ computation reliability by reducing the noise impact at the level of measured expectation values. Examples include readout-error mitigation, symmetry verification, probabilistic error cancellation, and Zero-Noise Extrapolation (ZNE). These techniques however also introduce sampling and classical post-processing overhead. When CK and QEM are combined, the resulting execution flow is governed by a complex trade-off. On the one hand, partitioning a circuit into sub-circuits can reduce the accumulated operational error and make QEM more effective. On the other hand, each cut increases the reconstruction cost, while QEM multiplies the number of circuit evaluations required to estimate the final observable. Thus, the best execution strategy must jointly balance fidelity, mitigation benefit, reconstruction overhead, sampling cost, and latency. This observation is particularly important in realistic quantum-clouds where devices are shared, heterogeneous, and subject to time-varying calibration data. There, circuit partitioning becomes an orchestration decision: the execution framework must decide how to decompose the workload, which cuts are worth introducing, whether mitigation should be applied, and whether the resulting execution cost is compatible with the resource and latency budget.

## III. The MOSAIQ Framework

MOSAIQ addresses the above problem by treating CK as a cost-aware and backend-aware optimization task, and evaluating the expected impact of candidate cuts on both execution fidelity and reconstruction cost. It combines calibration-driven noise estimation, mitigation-aware fidelity modeling, HW constraints, and reconstruction-cost penalties in a unified optimization flow. The key idea is to select cuts only when the fidelity improvement enabled by smaller and less noisy sub-circuits outweighs the cost introduced by reconstruction and mitigation. Notably, MOSAIQ accounts for the overhead associated with two-qubit gate cuts. This prevents overly aggressive partitioning that may appear beneficial for fidelity but becomes inefficient or infeasible due to sampling and reconstruction costs. In this sense, MOSAIQ provides a software-defined control layer for reliable NISQ execution. Importantly, differently from topology-driven cutting approaches, MOSAIQ does not treat all cuts as equivalent. Instead, it associates each candidate cut with a cut-dependent reconstruction cost.

This allows the optimizer to distinguish between coarse wire cuts and finer-grained two-qubit gate cuts, selecting the best configuration.

**Circuit Partitioning Model.** We model a quantum circuit as a layered computation acting on $W$ logical qubits and with depth $D$. A partitioning configuration $P$ is a set of selected cuts; a cut $c{\in}P$ can be either a wire cut or a two-qubit gate cut. The selected cuts decompose the original circuit into a set of sub-circuits $\mathcal{S}(P){=}\{S_1, \ldots, S_N\}$. Each sub-circuit $S_i$ is characterized by its width $w_i$, i.e., the number of qubits required for its execution, and its depth $d_i$, and it is feasible only if every generated sub-circuit can be executed on the target backend. Thus, we impose:

$$w_i \leq W_{\max}, \forall S_i \in \mathcal{S}(P), \tag{1}$$

where $W_{\max}$ is the number of physical qubits on the selected device. This constraint captures the HW-size limitation that motivates circuit knitting: cuts are introduced to make the generated fragments compatible with constrained NISQ backends while reducing their exposure to accumulated noise.

**Compact Fidelity Model.** MOSAIQ relies on a lightweight fidelity proxy to compare alternative partitioning strategies without simulating the full noisy evolution of each candidate circuit. We adopt a compact exponential decay model where the operational fidelity of a sub-circuit $S_i$ depends on its width $w_i$, depth $d_i$, and an effective backend-dependent error rate $\epsilon$: $F_{\text{ops}}(S_i) = \exp(-\epsilon w_i d_i)$. The parameter $\epsilon$ summarizes the dominant contribution of gate errors, decoherence, and readout noise, and is estimated from backend calibration data combined as in [2], [6]–[9]. When QEM is enabled, its average effect is modeled via a mitigation factor $\mu \in (0, 1]$, which reduces the effective error rate: $F^{\text{mit}}_{\text{ops}}(S_i) = \exp(-\mu\epsilon w_i d_i)$. This model is intentionally simple: the goal is not to reproduce all device-level noise mechanisms, but to provide a backend-aware ranking criterion for candidate partitions. Notably, it captures the expected qualitative trend that shallower and narrower sub-circuits are less affected by noise accumulation. The model consistency with calibration-driven fidelity trends is shown in Fig. 2.

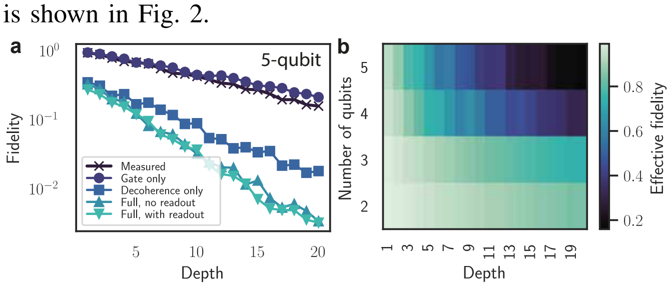


Fig. 2. Fidelity vs. circuit depth. **a)**: Experimental results vs. individual noise model contributions, for a 5-qubit system. **b)**: Fidelity heatmap vs. circuit depth and no. of qubits (2–5) under a full noise model without readout.

**Reconstruction and Sampling Cost.** The classical reconstruction overhead induced by circuits cuts depends on the cut type [10]. A wire cut interrupts the evolution of a qubit line and requires reconstructing the missing quantum channel through multiple measurement-and-preparation configurations. A two-qubit gate cut, instead, decomposes an entangling operation into a linear combination of local operations.

To capture this diversity, we associate each candidate cut $c$ with a reconstruction factor $\gamma_c{>}1$, and distinguish wire cuts and and two-qubit gate cuts with factor $\gamma_{\text{wire}}$ and $\gamma_{\text{gate}}$ (resp.). The total reconstruction overhead associated with a partitioning configuration $P$ is then

$$\Gamma(P) = \prod_{c\in P} \gamma_c^2, \tag{2}$$

where the squared factor accounts for the sampling overhead of quasi-probability reconstruction. (2) allows comparing heterogeneous cutting choices under a common cost model.

In addition to sampling overhead, each cut may also degrade the quality of the reconstructed observable. We thus introduce a cut-dependent reconstruction penalty $\eta_c{\in}(0, 1]$, and write the total quality of a partitioned execution as the product of the mitigated operational fidelities of the sub-circuits and the reconstruction penalties associated with the selected cuts:

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

(3) captures CK's fundamental trade-off: cuts can improve sub-circuit fidelity by reducing width or depth, but it also increases the reconstruction cost and may degrade the final estimator.

**Latency Constraint.** The reconstruction overhead directly affects the number of circuit executions required to estimate the final observable. Let $N_{\text{base}}$ be the number of shots used for the reference execution and $M_{\text{QEM}}$ the multiplicative overhead induced by the selected QEM procedure. The total number of required shots can be written as $N_{\text{tot}}(P) = N_{\text{base}} M_{\text{QEM}} \Gamma(P)$. Assuming an average execution time per shot $t_{\text{shot}}$, the corresponding Quantum Processing Unit (QPU) execution time is:

$$T_{\text{exec}}(P) = t_{\text{shot}} N_{\text{base}} M_{\text{QEM}} \Gamma(P). \tag{4}$$

To ensure operational feasibility, MOSAIQ enforces a maximum latency budget $T_{\max}$: $T_{\text{exec}}(P) \leq T_{\max}$. By applying a logarithmic transformation, the multiplicative overhead constraint becomes additive over the selected cuts:

$$\sum_{c\in P} 2\log(\gamma_c) \leq \log\left(\frac{T_{\max}}{t_{\text{shot}} N_{\text{base}} M_{\text{QEM}}}\right). \tag{5}$$

This form is useful for optimization, since it preserves the reconstruction overhead's multiplicative nature while enabling additive reasoning over candidate cuts. It also prevents MOSAIQ from selecting partitions that are valid from a circuit perspective but infeasible under the available execution budget.

**Optimization Problem.** MOSAIQ's aim is to select the partitioning configuration $P$ that maximizes the expected fidelity while satisfying HW and latency constraints. Taking the logarithm of $F_{\text{tot}}(P)$, we formulate the problem as

$$\max_{P} \quad -\sum_{S_i\in\mathcal{S}(P)} \mu\epsilon w_i d_i + \sum_{c\in P} \log(\eta_c) \tag{6}$$

$$\text{s.t. (1) and (5) hold} \tag{7}$$

The resulting formulation can be solved via standard optimization solvers such as Gurobi (https://www.gurobi.com). In practice, the optimizer selects cuts only when the reduction in operational error enabled by sub-circuits compensates for the reconstruction penalty and remains compatible with the available latency and shot budget.

## IV. Experimental Setup

**Benchmarks.** We evaluate MOSAIQ on two complementary classes of quantum circuits. The first benchmark consists of parameterized 3-qubit circuits targeting the expectation value of the $ZZZ$ observable. The circuits include single-qubit rotations and $CZ$ gates, and are partitioned through a spatial cut on the middle qubit. This benchmark evaluates how CK and QEM affect the reconstructed expectation value as circuit depth and noise accumulation increase. The second benchmark uses Quantum Approximate Optimization Algorithm (QAOA) circuits [11] with depths from $p = 1$ to $p = 5$, dominated by $RZZ$ interactions. This benchmark tests whether MOSAIQ can choose between low-cost gate cuts and higher-cost wire cuts as the cumulative reconstruction overhead grows.

**Platforms and noise models.** We validate MOSAIQ on both noisy simulators and real quantum HW. Simulations use IBM Fake Backends, namely, *FakeLimaV2*, *FakeManilaV2*, and *FakeJakartaV2*, with *FakeJakartaV2* used for the QAOA budget experiments. Real-HW experiments are run on *IQM Lagrange*, a 5-qubit superconducting processor [12], [13], enabling validation under hardware noise, calibration imperfections, and runtime variability. The optimizer does not simulate the full noisy evolution of every candidate partition. Instead, it relies on the compact fidelity and reconstruction-cost model introduced in Sec. III, while noisy simulation and HW execution are used only for validation.

**Execution strategies.** We compare different execution strategies depending on the benchmark. For the 3-qubit circuits, we consider full execution (Full), CK with a deterministic spatial cut followed by classical reconstruction (Cut), and CK combined with SPAM mitigation (Cut+SPAM). The reconstructed expectation values are compared against the ideal noiseless outcome. For the QAOA circuits, we compare MOSAIQ to a topology-driven cutting baseline inspired by automatic cutting strategies, considering raw CK as well as CK combined with ZNE and SPAM mitigation. The baseline selects a wire cut with fixed $16\times$ reconstruction overhead, whereas MOSAIQ compares candidate $RZZ$ gate cuts against the wire-cut cost and selects the partition that best satisfies the fidelity-cost-latency trade-off defined in Sec. III.

**Shot and latency budgets.** The QAOA experiments are designed to explicitly test the operational feasibility of the selected partitioning strategies under fixed resource budgets. We impose a global shot budget of $150{,}000$ shots and a QPU execution-time budget of 2 hours. Under this setting, the topology-driven baseline always selects a wire cut with $16\times$ overhead, which results in 9,375 shots per fragment for all QAOA depths. MOSAIQ uses cut-dependent overheads: for QAOA depths $p$=1, $p$=2, and $p = 3$, it selects low-cost $RZZ$ gate cuts with cumulative overheads of $2.4\times$, $5.8\times$, and $13.9\times$, respectively. This corresponds to $62{,}500$, $25{,}862$, and $10{,}791$ shots per fragment. For $p$=4 and $p$=5, the cumulative gate-cut overhead becomes too high, and MOSAIQ falls back to a wire cut with $16\times$ overhead and $9{,}375$ shots per fragment.

The latency experiment uses $N_{\text{base}} = 81{,}920$ shots, a QEM multiplier $M_{\text{QEM}} = 3$, and an average shot time $t_{\text{shot}} = 0.002\,\text{s}$. This setup directly evaluates whether a partitioning strategy is not only structurally valid, but also operationally feasible under realistic execution constraints.

**Evaluation metrics.** We evaluate MOSAIQ's reconstructed expectation-value accuracy, execution latency under the imposed QPU budget, and statistical stability across repeated trials. The last metric is crucial for QEM, since ZNE and SPAM correction rely on finite-shot estimates: if reconstruction overhead consumes most of the shot budget, mitigation may become unstable. We therefore track both shots per fragment and estimator variance to assess whether MOSAIQ preserves enough statistical quality for reliable mitigation.

## V. Experimental Results

**Real HW Validation.** We validate MOSAIQ by comparing the IQM Lagrange quantum processor performance to noisy IBM Fake Backend simulation results. Fig. 3 shows $\langle ZZZ \rangle$ for the Original circuit under full execution, CK, and CK with SPAM mitigation. For real HW and simulators, full execution is degraded by noise, CK alone provides limited gains due to reconstruction overhead, while CK+SPAM moves the result closer to the ideal reference. Although the values vary with backend-specific noise and calibration, the trend is consistent, confirming that MOSAIQ effectively trades off sub-circuit reliability with reconstruction overhead, and mitigation benefit.

**Impact of Fidelity-aware Routing on Mitigation.** We finally evaluate how reconstruction-aware routing affects the practical effectiveness of QEM. While the previous experiments assess the fidelity improvement obtained by combining circuit knitting and QEM, this analysis focuses on a complementary question: whether the selected partitioning strategy preserves enough execution budget and statistical accuracy for mitigation to operate reliably.

We consider the QAOA benchmark introduced in Sec. IV and compare MOSAIQ against a topology-driven cutting baseline. The experiment is performed on the FakeJakartaV2 noise model under a global budget of $150{,}000$ shots and a QPU execution-time budget of 2 hours. The baseline always selects a wire cut, which induces a fixed reconstruction overhead of $16\times$. Conversely, MOSAIQ compares the cost of multiple $RZZ$ gate cuts against the cost of a wire cut and selects the configuration with the lowest reconstruction overhead compatible with the execution constraints.

TABLE I
MOSAIQ vs. Topology-driven cutting on QAOA circuits under a $150{,}000$-shot budget

| Circuit Depth | Baseline (Topological AutoCut) | | MOSAIQ (Fidelity-Aware) | | |
|---|---|---|---|---|---|
| | Cut | Overhead | Cut | Overhead | Shots / Fragment |
| QAOA $p = 1$ | 1x Wire | 16.0x | 2x Gate (RZZ) | 2.4x | 62,500 |
| QAOA $p = 2$ | 1x Wire | 16.0x | 4x Gate (RZZ) | 5.8x | 25,862 |
| QAOA $p = 3$ | 1x Wire | 16.0x | 6x Gate (RZZ) | 13.9x | 10,791 |
| QAOA $p = 4$ | 1x Wire | 16.0x | 1x Wire (Safety) | 16.0x | 9,375 |
| QAOA $p = 5$ | 1x Wire | 16.0x | 1x Wire (Safety) | 16.0x | 9,375 |

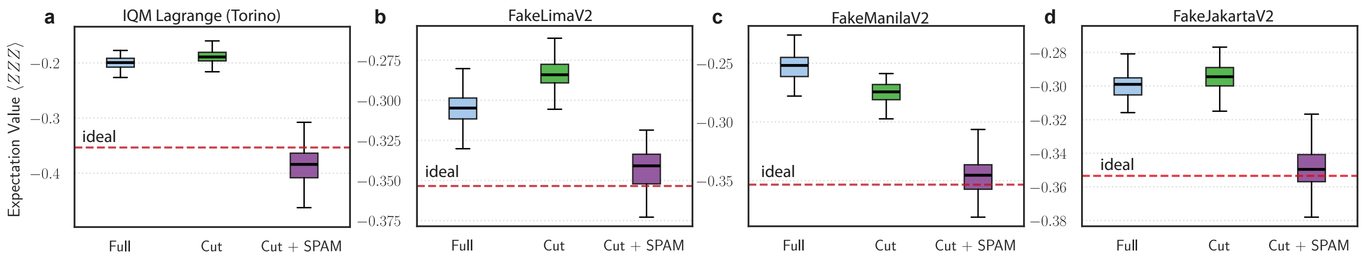


Fig. 3. Expectation values ($\langle ZZZ \rangle$) on real HW (IQM Lagrange) and noisy simulators for the Original circuit. Boxplots compare full execution (Full), circuit knitting (Cut), and circuit knitting with SPAM (Cut+SPAM) mitigation. The dashed line indicates the ideal value.

The detailed cut configurations are presented in Table I. For $p \leq 3$, MOSAIQ selects low-cost $RZZ$ gate cuts instead of the wire cut selected by the baseline. Specifically, it selects 2, 4, and 6 gate cuts for $p = 1$, $p = 2$, and $p = 3$, resulting in cumulative overheads of $2.4\times$, $5.8\times$, and $13.9\times$, respectively. These overheads remain below the $16\times$ wire-cut cost, allowing MOSAIQ to allocate a larger number of shots per fragment.

Fig. 4 shows the resulting execution latency as a function of QAOA depth. The topology-driven baseline exceeds the 2-hour budget already at $p = 1$, reaching 2.18 hours due to the $16\times$ overhead of the selected wire cut. This makes the corresponding execution infeasible under the imposed HW budget. MOSAIQ avoids this failure mode by selecting lower-cost $RZZ$ gate cuts for shallow and intermediate circuits, keeping the latency below the 2-hour constraint for $p \leq 3$.

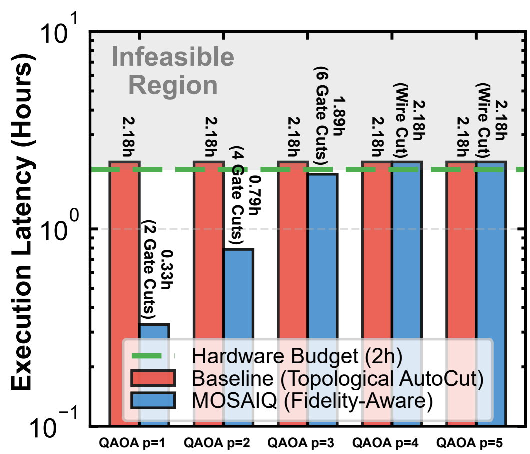


Fig. 4. Latency of topology-driven cutting and MOSAIQ on QAOA circuits under 2-hour QPU. The baseline selects a wire cut with $16\times$ overhead, exceeding the budget already at $p$=1. MOSAIQ selects lower-cost $RZZ$ gate cuts for $p$≤3, keeping the execution feasible, and falls back to a wire cut for deeper circuits when the cumulative gate-cut overhead is too large.

For deeper QAOA circuits, however, the cumulative cost of gate cutting becomes too large. At $p$=4 and $p$=5, MOSAIQ dynamically falls back to the wire-cut configuration, with $16\times$ overhead and 9,375 shots per fragment, as reported in Table I. This behavior shows that MOSAIQ does not blindly prefer fine-grained gate cuts, but uses them only when their reconstruction cost is compatible with the available latency and shot budget. The wire cut thus acts as a safety configuration when the gate-cut overhead approaches the feasibility limit.

The impact of this decision on the statistical quality of the reconstructed signal is shown in Fig. 5. Under a fixed global budget of 150,000 shots, the number of shots available per fragment is inversely proportional to the reconstruction overhead. As a consequence, the topology-driven baseline allocates only 9,375 shots per fragment for all QAOA depths. In contrast, MOSAIQ allocates 62,500 shots per fragment at $p = 1$, 25,862 at $p = 2$, and 10,791 at $p = 3$. This larger per-fragment sampling budget substantially reduces the statistical variance of the reconstructed expectation values, producing more compact distributions in Fig. 5 and more stable inputs for ZNE and SPAM mitigation.

These results highlight an important interaction between CK and error mitigation. QEM techniques operate on statistically estimated expectation values; hence, if the partitioning strategy consumes most of the available shot budget through excessive reconstruction overhead, the mitigated result may become unstable or dominated by estimator variance. By selecting lower-cost cuts when possible, MOSAIQ preserves the statistical quality required by mitigation and improves the reliability of the reconstructed observable.

For deeper QAOA circuits, the benefit progressively decreases as HW noise and decoherence become dominant. Notably, for $p$=4 and $p$=5, MOSAIQ and the baseline converge to the same wire-cut strategy, as shown in Table I. The residual deviation from the ideal value in Fig. 5 is mainly determined by the backend physical noise rather than the routing policy alone. This confirms that MOSAIQ does not mask the intrinsic limitations of current NISQ HW, but rather identifies the most feasible execution strategy within those limits.

**Summary.** The experimental results highlight that CK is effective only when the fidelity benefit obtained by executing smaller sub-circuits is balanced against the reconstruction overhead introduced by the selected cuts. Notably, the QAOA experiments show that the choice of the cutting primitive has a direct impact on both execution latency and the statistical quality of the reconstructed observable.

Interestingly, the topology-based baseline always selects a wire cut, which induces a fixed $16\times$ overhead. This choice is structurally simple, but it rapidly exhausts the execution budget and becomes infeasible already for shallow QAOA instances under the considered 2-hour QPU budget. In contrast, MOSAIQ selects lower-cost $RZZ$ gate cuts when their cumulative overhead is smaller than that of a wire cut, keeping the execution feasible for $p \leq 3$. This confirms that minimizing the number of cuts is not necessarily the right strategy: several fine-grained gate cuts may be preferable to a single coarse wire

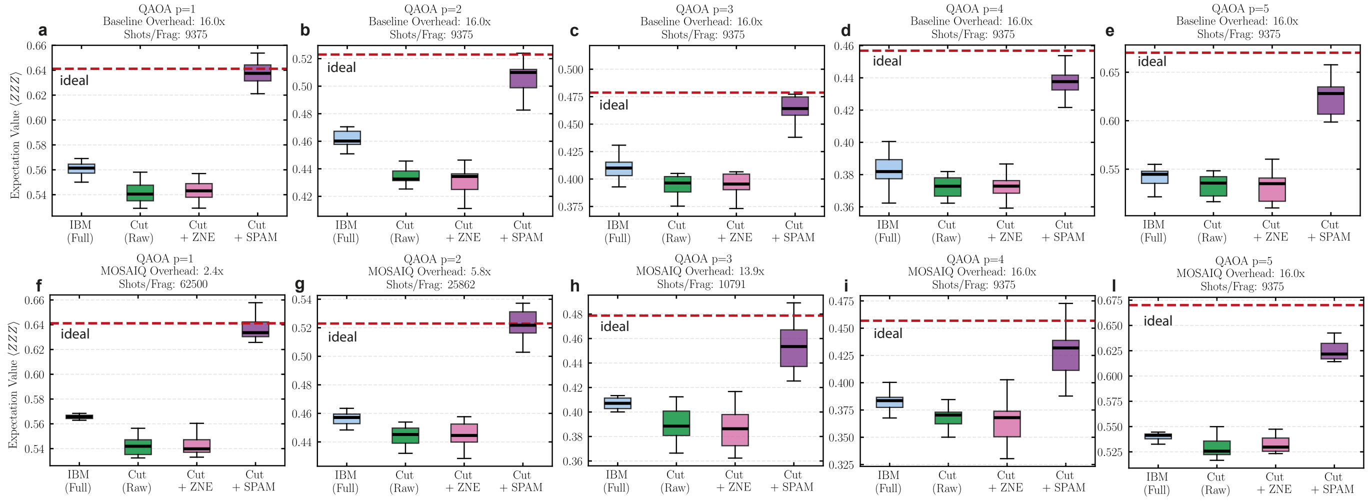


Fig. 5. Impact of partitioning strategy on the quality of mitigated QAOA expectation values under a fixed budget of 150,000 shots on FakeJakartaV2. By reducing reconstruction overhead for shallow and intermediate depths, MOSAIQ allocates more shots/fragment, reduces estimator variance, and provides more stable inputs to ZNE and SPAM mitigation. For deeper circuits, HW noise dominates and both strategies converge to the same safety wire-cut configuration.

cut if their cumulative reconstruction cost remains limited.

A second important finding is the interaction between circuit partitioning and QEM. Error mitigation relies on statistically estimated expectation values, i.e., its effectiveness depends not only on the backend noise but also on the number of shots available for each reconstructed fragment. If the selected partitioning consumes most of the global shot budget through excessive reconstruction overhead, the mitigated result may become dominated by estimator variance. MOSAIQ mitigates this issue by preserving a larger per-fragment shot budget whenever low-overhead gate cuts are available, producing more stable inputs for the mitigation pipeline.

These findings support the view of NISQ execution as a software-defined orchestration problem. Quantum workloads must be adapted to noisy and time-varying HW resources while accounting for target latency and QoS. MOSAIQ fits this vision by treating CK as an execution-planning decision rather than as a purely circuit-level transformation.

## VI. Conclusions

We presented MOSAIQ, a reconstruction-aware framework for latency-constrained NISQ execution via quantum circuit knitting. MOSAIQ jointly accounts for sub-circuit fidelity, error mitigation, reconstruction overhead, and execution latency, enabling the selection of partitioning strategies that are not only beneficial from a fidelity perspective, but also feasible under realistic shot and time budgets. Experimental results show that topology-driven cutting can lead to infeasible executions or high estimator variance when reconstruction overhead is ignored. In contrast, MOSAIQ selects low-cost two-qubit gate cuts when advantageous and dynamically falls back to wire cuts when their cumulative cost becomes too high. This preserves the statistical quality required by QEM techniques, improving the reliability of the reconstructed expectation values. Future work will extend MOSAIQ toward larger workloads, richer backend-aware cost models, and multi-backend quantum-cloud orchestration.

## References


[1] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*. Cambridge university press Cambridge, 2001, vol. 2.

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

[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] 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.

[8] 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.

[9] 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.

[10] 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. 023021, feb 2021. [Online]. Available: https://doi.org/10.1088/1367-2630/abd7bc

[11] K. Blekos, D. Brand, A. Ceschini, C.-H. Chou, R.-H. Li, K. Pandya, and A. Summer, "A review on quantum approximate optimization algorithm and its variants," *Physics Reports*, vol. 1068, p. 1–66, 2024. [Online]. Available: http://dx.doi.org/10.1016/j.physrep.2024.03.002

[12] 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

[13] 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.