

Quantum Tensor Representation via Circuit Partitioning and Reintegration

Ziqing Guo

Texas Tech University
Lubbock, TX, USA
ziqguo@ttu.edu

Kewen Xiao

Rochester Institute of Technology
Henrietta, NY, USA
kewenxiao@gmail.com

Jan Balewski

National Energy Research Scientific Computing Center,
Lawrence Berkeley National Laboratory
Berkeley, CA, USA
balewski@lbl.gov

Ziwen Pan

Texas Tech University
Lubbock, TX, USA
ziwen.pan@ttu.edu

Abstract

Quantum computing enables faster computations than classical algorithms through superposition and entanglement. Circuit cutting and knitting are effective techniques for ameliorating current noisy quantum processing unit (QPUs) errors via a divide-and-conquer approach that splits quantum circuits into subcircuits and recombines them using classical post-processing. The development of circuit partitioning and recomposing has focused on tailoring the simulation framework by replacing generic non-local gates with probabilistic local gates and measuring the classical communication complexity. Designing a protocol that supports algorithms and non-all-to-all qubit-connected physical hardware remains underdeveloped owing to the convoluted properties of cutting compact controlled unitary gates and hardware topology. In this study, we introduce **shardQ**, a method that leverages the **SparseCut** algorithm with matrix product state (MPS) compilation and a global knitting technique. This method elucidates the optimal trade-off between the computational time and error rate for quantum encoding with a theoretical proof, evidenced by an ablation analysis using an IBM Marrakesh superconducting-type QPU. This study also presents the results regarding application readiness.

1 Introduction

Scaling quantum computation to the utility regime, where quantum processors deliver consistent application-level advantages, remains a central challenge in the **NISQ** era. Despite progress in quantum algorithms and hardware, limited qubit connectivity, high two-qubit gate error rates, and the overhead of long-distance entanglement operations hinder the practical execution of large-scale quantum circuits on current superconducting quantum processing units (QPUs). These constraints are particularly problematic for high performance computing (HPC) integrated quantum workflows, in which large circuits must be partitioned, executed, and recombined efficiently.

Overcoming these limitations would enable the execution of resource-intensive quantum algorithms, such as Grover’s search [14], Shor’s factoring [31], and Quantum Fourier Transform (**QFT**) [10], on near-term devices, unlocking practical applications in cryptography, optimization, and scientific simulation. Importantly, circuit cutting and knitting techniques, which utilize quasi-probability decomposition (**QPD**) [7, 23], offer a promising route for reducing hardware requirements without sacrificing algorithmic universality. This capability is critical for hybrid HPC–quantum platforms, in which quantum resources must be tightly integrated with classical post-processing.

While **QPD** enables resource trade-offs through local operations and classical communication (LOCC) [8, 29], it introduces a sampling overhead and error amplification, particularly in algorithms requiring deep entanglement. Moreover, superconducting QPUs typically have finicky connectivity, forcing the use of additional SWAP operations or long-distance entanglement gates, which increase circuit depth and error probability. Existing quantum data encoding schemes, such as Vectorized Quantum Data Encoding (**QCrack**) [2], still require high-dimensional correlated superposition states that are costly to implement on real hardware.

State-of-the-art circuit decomposition methods, including space cut [22] and time cut [24], have demonstrated compatibility with universal two-qubit gate architectures [11], but they often neglect hardware-aware optimization for connectivity-limited QPUs. Alternative approaches, such as variational quantum clustering [4] and quantum embedding analysis [27], provide valuable insights into data representation but are not designed for direct classical data encoding and require iterative quantum-classical optimization. Consequently, these methods either fail to fully exploit **QPD** in practical **NISQ** systems or incur prohibitive resource overheads.

In this study, we introduce **shardQ**, an end-to-end partition-to-recomposition quantum tensor encoding model specifically optimized for superconducting quantum chips in the **NISQ** era. Our approach employs dynamic cut control based on the number of qubits to minimize two-qubit entanglement gates, integrates hardware-aware circuit knitting to address restricted qubit connectivity, and supports HPC-integrated execution, enabling the decomposition, distribution, and recombination of large-scale quantum algorithms with reduced error rates and circuit depths. Although the method is tailored to superconducting architectures, its principles can be adapted to other platforms with constrained connectivity.

2 Related Work

The concept underlying our approach traces back to the Clifford-gate group simulation protocol [6], which reduces the classical–quantum computation overhead and strengthens the hybrid paradigm [7]. This foundation has driven both empirical [24] and theoretical [21] advances in clustered simulations for molecular systems, which were later enhanced by maximum-likelihood fragment tomography [25]. Subsequent studies have examined the overhead of circuit cutting and knitting for entanglement gates [13, 19, 26, 35] and error mitigation in low-depth entanglement circuits [34], leading to practical frameworks such as CutQC [32] and Qiskit Circuit Knitting [30]. However, these frameworks lack algorithm-level, QPU-aware optimizations for hardware with limited connectivity. Recent innovations include heuristic randomized measurement cutting for the quantum approximate optimization algorithm (**QAOA**) [20] and high-fidelity cutting with gradient-based reconstruction [17]. The Qdislib framework [33] extends cutting to distributed settings by integrating HPC with QPUs. However, the scalability and efficiency of such cutting methods—particularly for quantum tensor encoding in the **NISQ** era—remain uncertain. Our proposed approach is motivated by the error mitigation benefits of tightly coupling QPUs with classical HPC resources [9].

3 Preliminaries

In this section, we present the essential background for the proposed **shardQ** protocol. We first introduce the tensor network simulation technique. Subsequently, we outline the approximate quantum data encoding approach. Finally, we provide the fundamental concepts underlying circuit cutting and the knitting process.

3.1 Quantum circuit simulation

MPS compilation. The quantum state can be written as

$$|\psi\rangle = \sum_{i=0}^{2^n-1} c_i |i\rangle, \quad (1)$$

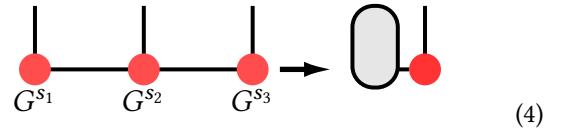
where classical hardware requires exponentially growing random access memory (RAM) to simulate the state vector, specifically 2^n complex amplitudes for n qubits. Matrix product states (MPS), based on tensor networks (TN), are widely used to mitigate RAM overhead. For example, the Greenberger–Horne–Zeilinger (GHZ) state can be represented as

$$\mathcal{G}_{s_1 s_2 s_3} = [A^{s_1} A^{s_2} A^{s_3}], \quad (2)$$

where $\mathcal{G}_{s_1 s_2 s_3}$ yields the computational basis amplitude for each $(s_1, s_2, s_3) \in \{0, 1\}^3$. The contraction of a tensor network representing a quantum circuit is given by

$$\mathcal{A} = \sum_{i_k} \prod_j T_{i_{j-1}, i_j}^{[j]}, \quad (3)$$

where $T^{[j]}$ is the local tensor at site j , and indices $\{i_k\}$ are contracted according to the network (see Diagram (4)). Here, the contraction of three tensors produces an effective tensor by summing the internal indices.



An important advantage of this method is that, for circuits exhibiting structured entanglement, the time complexity is $O(\text{poly}(N) \cdot 2^w)$ [3], where w represents the minimal width determined by the circuit connectivity.

3.2 Gate-based quantum data encoding

Data encoding is essential for embedding and image encoding because it enables a compact Hilbert space representation of classical data. Quantum data encoding transforms classical information into quantum states using three methods. *Basis encoding* maps discrete values directly to computational basis states using binary representation; for example, $[01, 11]$ becomes $|x^1\rangle = |01\rangle$ and $|x^2\rangle = |11\rangle$. *Amplitude encoding* embeds normalized data \vec{x} into quantum amplitudes $\sum_i \alpha_i |i\rangle$, subject to normalization. *Angle encoding* maps features to qubit states using rotation gates, so each feature is encoded as $\cos(\theta_i/2) |0\rangle + \sin(\theta_i/2) |1\rangle$, with θ_i rescaled to $[0, \pi]$. The resulting quantum state is

$$|\psi\rangle = \bigotimes_{i=1}^n \left(\cos \frac{\theta_i}{2} |0\rangle + \sin \frac{\theta_i}{2} |1\rangle \right). \quad (5)$$

Following angle encoding, **QCrank** encodes classical data with the Unitary Controlled Rotation (UCR_y) gate

$$\text{UCR}_y(\theta) = \begin{pmatrix} R_y(\theta_0) & & \\ & \ddots & \\ & & R_y(\theta_{2^{n_a}-1}) \end{pmatrix}. \quad (6)$$

Each address qubit configuration (representing the position of each classical data input) $|i\rangle$ is prepared in a superposition using the Walsh–Hadamard Transform (WHT). For each address, the associated data values are encoded onto the data

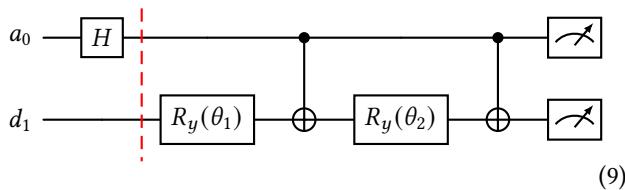
qubits by applying the single-qubit rotations. Specifically, for every address, each data qubit receives a rotation by an angle corresponding to the classical data value assigned to that address and the data qubit.

$$|c_{i,j}\rangle = \cos(\theta_{i,j}/2)|0\rangle + \sin(\theta_{i,j}/2)|1\rangle. \quad (7)$$

The UCR_y gate applies these rotations in a block-diagonal form, which is controlled by the address qubits. Thus, the **QCrank** encoder produces

$$|\psi_{\text{qcrank}}(\vec{\theta})\rangle = \frac{1}{\sqrt{2^{n_a}}} \sum_{i=0}^{2^{n_a}-1} |i\rangle \otimes |c_{i,0}\rangle \otimes \cdots \otimes |c_{i,n_d-1}\rangle. \quad (8)$$

with each $|c_{i,j}\rangle$ angle-encoded via UCR_y , n_a address qubits, and n_d data qubits. For example, encoding the 3D tensor $(1, 2, 1)$ yields the circuit



resulting in $|\psi_{\text{qcrank}}\rangle = \frac{1}{\sqrt{2}} (|0\rangle|0\rangle + |1\rangle [c|0\rangle + s|1\rangle])$, where $c = \cos\left(\frac{\theta_1+\theta_2}{2}\right)$, $s = \sin\left(\frac{\theta_1+\theta_2}{2}\right)$. To retrieve classical data, projective measurements estimate $|c|^2$ and $|s|^2$, and the encoded angle is reconstructed by

$$\theta_1 + \theta_2 = 2 \arctan\left(\sqrt{\frac{|s|^2}{|c|^2}}\right), \quad (10)$$

Assuming that the angle is rescaled to $[0, \pi]$ using $\arccos(0, 1)$ via EVEN encoding [1].

3.3 Circuit Cutting and Knitting

The circuit knit-and-cut technique enables the efficient simulation of large quantum circuits by dividing them into smaller, manageable subcircuits (SC), which are simulated separately and then recombined to construct the global observables. This approach exploits the fact that the expectation value of an operator on the entire circuit, such as $\mathcal{S}(U_1 \otimes U_2)$ [22], can be represented as a linear combination of the expectation values of each subcircuit, each weighted by classical coefficients, where \mathcal{S} represents the operator in Pauli groups. For two arbitrary unitaries U , the observable in the QPD framework is reconstructed as

$$\mathcal{S}(U_1 \otimes U_2) = \mathbf{c}^T \mathbf{S} + \frac{1}{4} \sin(2\theta) \sin(2\phi) \sum_{\alpha} \alpha_1 \alpha_2 \mathcal{R}(\alpha). \quad (11)$$

Here, \mathbf{S} contains the expectation values of the subcircuits with different operator insertions at the cut, and \mathbf{c} provides the corresponding weights of the expectation values. The sum over α captures the correlations introduced by the cut (see Appendix A in [21]). This technique relies on efficient

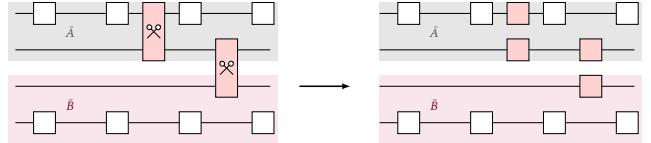


Figure 1. The generic example of cutting two qubit gates into separate one qubit gates. Note that, the right side only shows one of the subcircuits.

classical post-processing to recover the outcome of the original circuit (see Table 1 in [29]). Here, we provide the general QPD sampling overhead for a CX gate for our protocol

Definition 1 (QDP overhead for CX gate). *Let C be a quantum circuit; the QDP overhead is*

$$O_{\text{QDP}}(C) = 3^{2k}, \quad (12)$$

where k is the number of cut qubits, 2 symbolizes the decomposition of the control and target gates, and each factor of base 3 reflects the three nontrivial Pauli insertions (X, Y, Z) per cut qubit, as shown in Fig. 1.

4 Methods

Fig. 2 shows the end-to-end **shardQ** protocol that leverages circuit cutting algorithm, quantum approximate compilation, and global result reconstruction. Note that the proposed protocol is for gate-based quantum simulation. Specifically, we provide the general cutting strategy **SparseCut** for the quantum encoder, which further allows the best MPS compilation. In addition, we propose a generic global measured bit string reconstruction algorithm.

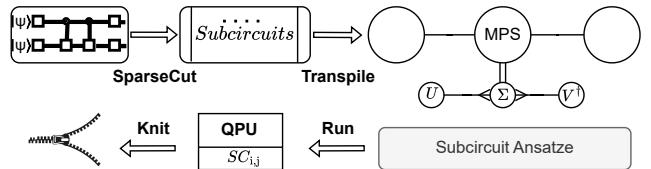


Figure 2. The **shardQ** protocol is outlined as follows: Initially, the original data encoder circuit employs the **SparseCut** algorithm to divide the circuit into subcircuits (SC). Subsequently, approximate quantum compilation is used to transpile these subcircuits into MPS. The ansatze are executed on the QPU, where the index i denotes the partition and j represents the decomposed gates. Ultimately, the results are globally reconstructed into classical tensor data using local saved intermediate results.

4.1 Cutting Algorithm

Similar to the use of permuted controlled-unitary operations in the **QFT**, the **QCrank** encoder employs layers of gray-coded CX gates to facilitate data entanglement and connectivity in a high-dimensional Hilbert space. Therefore,

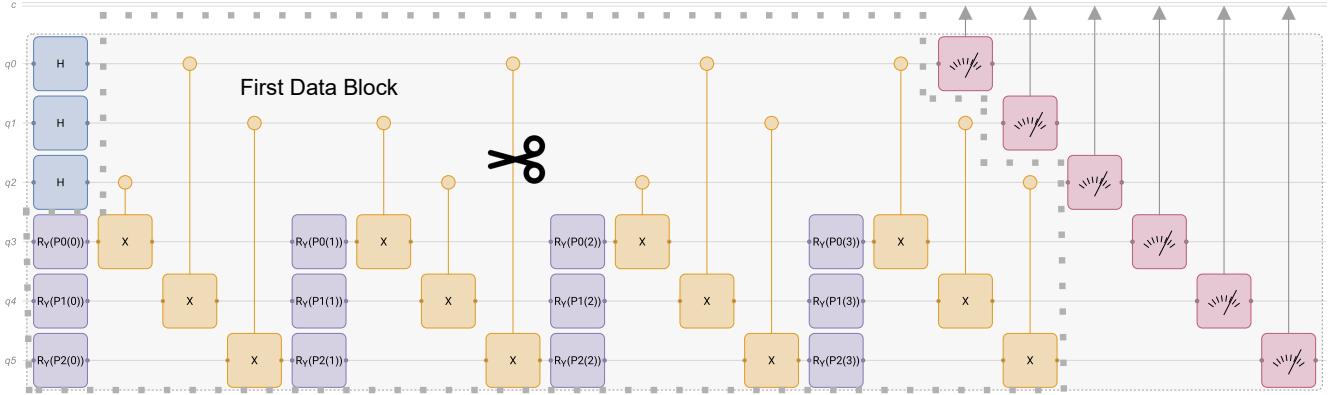


Figure 3. Example of three-by-three tensor encoder circuit cutting paradigm. The scissors are placed under the longest entanglement gate, corresponding to the physical qubit mapping, as shown in Fig. 4. We only show the first data block, as indicated by the gray dashed box. We refer to the rest of the encoding block in Fig. 1 (c) [2]. Note that, each data qubit (q_3, q_4, q_5) corresponding to first encoded dimension P with the address qubit encoded position as the rotation parameters noted by the indexes of P.

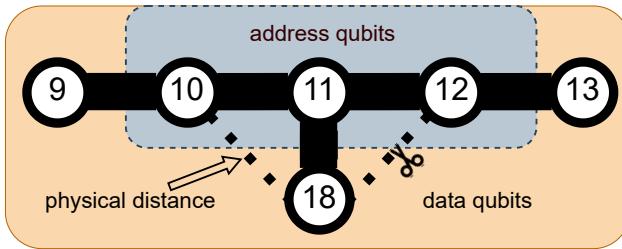


Figure 4. The three-by-three tensor encoder physical qubit mapping diagram. We denote the address qubits corresponding to the q_0, q_1, q_2 in Fig. 3 represented by the blue-shaded area. The yellow area presents the data qubits q_3, q_4, q_5 . The longest entanglement cut is q_0 (the first address qubit) and q_5 (third data qubit) shown in the dashed diagonal lines.

we denote that it is highly time-consuming to split the circuit into small partitions with smaller clusters simulated with fewer qubits because of the scaling overhead shown in Eq. (12). To tackle this, we provide the **SparseCut** algorithm illustrated in Alg. 1, where an example of a circuit cutting scheme with three address qubits and three data qubits is shown in Fig. 3. Here, we provide the definition of the cutting selection rule and its goal.

Definition 2. Given by the set of the cutting candidates

$$\mathcal{G}_{\text{cross}} = \left\{ \begin{array}{l} g = (n_a, n_d), \\ n_a \in A, n_d \in D, \\ d(n_a, n_d) \end{array} \right\}. \quad (13)$$

Here, n_a and n_d refer to the address and data qubits, respectively; d denotes distance; and the goal is to minimize the qubit map distance in the cutting pool \mathcal{G} . First, we recall that the address and data qubits encode Eq. (9) data = $\begin{bmatrix} \theta_{0,0,0} \\ \theta_{0,1,0} \end{bmatrix}$. Here, we denote that data[c][a][d], where the parameters

correspond to the circuit, address, and data indexes. The coupling map of Fig. 4 is retrieved from the state-of-the-art IBM 156 qubit Marrakesh QPU; the qubit indexes correspond to the latest layout [18]. By taking the address and data qubit

Algorithm 1 SparseCut Selection

Require: circuit C , observable qubit sets A, D , maximum cuts max_cuts

- 1: $\mathcal{G}_{\text{cross}} \leftarrow \emptyset$
- 2: **for** each two-qubit gate $g = (q_i, q_j)$ in C **do**
- 3: **if** $(q_i \in A \wedge q_j \in D) \vee (q_i \in D \wedge q_j \in A)$ **then**
- 4: $d \leftarrow |i - j|$
- 5: $\mathcal{G}_{\text{cross}} \leftarrow \mathcal{G}_{\text{cross}} \cup \{(gate_index, d)\}$
- 6: **sort** $\mathcal{G}_{\text{cross}}$ by d in **descending** order
- 7: **return** first $\min(|\mathcal{G}_{\text{cross}}|, \text{max_cuts})$ elements

indexes, Alg. 1 iteratively updates the current gate index (as shown in the 15th gate from left to right in our example in Fig. 3) based on the absolute distance between two sets A and D . The benefit of retrieving absolute differences becomes clearer when using different quantum platforms, as their resulting quantum bit strings may be in the reverse order. We also introduce a hyperparameter max_cuts that defines the upper bound of the cutting protocol; that is, in each recursion, the algorithm finds the longest entanglement distance based on the gray code law by looping the control and target qubit indexes and returning the minimum number of gate indexes in the group of the cutting candidate. The absolute virtual qubit distance is calculated by subtracting the target and control qubit indices from each other. We refer to the example of the QPD case in § A.

Algorithm 2 Reconstruct global counts

Require: job set $\mathcal{R} = \{r_1, \dots, r_m\}$; QPD coeffs. C
Ensure: global counter $\mathcal{G}: \Sigma^* \rightarrow \mathbb{R}$

- 1: $\mathcal{Y} \leftarrow \emptyset$ ▷ multiset of local counters.
- 2: **for all** $r \in \mathcal{R}$ **do**
- 3: $\mathbf{o} \leftarrow \text{LABELS}^1(r.\text{observables})$, $\mathbf{q} \leftarrow \text{LABELS}(r.\text{qpd})$
- 4: $\text{cnt} \leftarrow [\ s \mapsto 0 \]$
- 5: **for** $k = 1$ **to** $|\mathbf{o}|$ **do**
- 6: $s \leftarrow \mathbf{o}_k \parallel \mathbf{q}_k$; $\text{cnt}(s) += 1$
- 7: $\mathcal{Y} \cup= \{\text{cnt}\}$
- 8: $n_o \leftarrow n_q^{\text{tot}}$; $n_q \leftarrow |\text{dom}(\mathcal{Y}[1])| - n_o$
- 9: $\mathcal{G} \leftarrow [\ s \mapsto 0 \]$
- 10: **for all** $(\text{cnt}, c) \in \mathcal{Y} \times C$ **do**
- 11: **for all** $(s, n) \in \text{cnt}$ **do**
- 12: $\text{obs} \leftarrow s[1:n_o]$, $\text{qpd} \leftarrow s[n_o+1:]$
- 13: $\sigma \leftarrow \begin{cases} \text{PARITY}^2(\text{qpd}) & n_q > 0 \\ 1 & \text{otherwise} \end{cases}$
- 14: $\mathcal{G}(\text{reverse}(\text{obs})) += c \sigma n$
- 15: **for all** $s \in \text{dom}(\mathcal{G})$ **do**
- 16: $\mathcal{G}(s) \leftarrow \max(0, \lfloor \mathcal{G}(s) + 0.5 \rfloor)$
- 17: **return** \mathcal{G}

4.2 MPS Compilation

In alignment with the principles of **SparseCut**, the algorithm is distinctively characterized by its adherence to universal optimality, as the selection of the shortest path for severing the longest non-direct connecting edge aligns with Dijkstra's algorithm [16]. Consequently, the approximate quantum compilation (**AQC**) technique [28] employed in our protocol facilitates further reduction in gate depth post-cut Alg. 1, thereby compressing the tensor encoder in Eq. (8). The complete encoder circuit C is decomposed into a prefix C_1 and a suffix C_2 such that $C = C_2 C_1$, with only C_1 being compiled. All two-qubit operations that couple the address register A to the data register D are ranked according to their Manhattan distance d ; the $k = \text{max_cuts}$ gates with the largest d form the cutting set $\mathcal{G}_{\text{cross}}^*$. The removal of these gates results in the truncated circuit C_1^{trunc} , whose output state can be simulated as an MPS with bond dimension $\chi \ll 2^{|A|+|D|}$. It is important to note that the MPS provides an explicit representation of the tensor encoder target state $|\psi_{\text{tar}}\rangle \approx |\psi_{\text{qcrank}}\rangle$, as demonstrated in Eq. (8).

The ansatz \tilde{C}_θ , which is hardware-native and incorporates only nearest-neighbor couplings, is derived from C_1^{trunc} through a KAK-based [36] block factorization. This ensures that the initial parameter vector θ_0 precisely reconstructs C_1^{trunc} , except for the global phase. Subsequently, optimization is conducted by minimizing the infidelity cost function

$$\mathcal{L}(\theta) = 1 - |\langle \psi_{\text{tar}} | \psi(\theta) \rangle|^2, \quad |\psi(\theta)\rangle = \tilde{C}_\theta |0\rangle^{\otimes L}, \quad (14)$$

¹**LABELS** converts a bit-array (rows of 0/1 or non-negative integers) into a list of binary strings.

²**PARITY** returns $(-1)^{\#\text{1's}}$ of its input.

where the gradients are obtained by automatic differentiation through the tensor-network contraction. Because the long-range gates in $\mathcal{G}_{\text{cross}}^*$ are absent from the simulation, their entangling effect is reproduced variationally by the ansatz parameters, allowing χ to remain small while still capturing the dominant short-range correlations inside A and D . After convergence the compiled circuit is reconstructed as

$$C_{\text{AQC}} = C_2 \mathcal{G}_{\text{cross}}^* \tilde{C}_{\theta^*}, \quad (15)$$

which approximates the original encoder with fidelity exceeding $1 - \varepsilon$ while containing k fewer long-range CX layers than the original encoder.

4.3 Global Reconstruction

We provide the pseudocode in Alg. 2 for global bit string reconstruction. Note that **dom** represents the domain of the Hilbert Space. To recover the original circuit statistics, we start from the basic **QPD** recursion. For a single Pauli cut, the quasi-probability expansion of an observable \mathcal{M} acting on circuit C is

$$T(C, \mathcal{M}) = \sum_{i=1}^8 c_i T(C', \mathcal{M}_i), \quad (16)$$

where $T(C, \mathcal{M}) = \text{Tr}[\mathcal{M} C(\rho)]$ and C' denotes the circuit after inserting one Pauli completion. By iterating the rule over M cuts gives

$$T(C, \mathcal{M}) = \sum_{\alpha \in \{1, \dots, 8\}^M} \left(\prod_{m=1}^M c_{\alpha_m} \right) T(C', \mathcal{M}_\alpha), \quad (17)$$

with $\alpha = (\alpha_1, \dots, \alpha_M)$ enumerating the 8^M completion patterns (see Eq. (29)). Removing every cut edge splits C' into K independent fragments,

$$C' = C^{(1)} \sqcup C^{(2)} \sqcup \dots \sqcup C^{(K)}. \quad (18)$$

Because of the independent expectation values, we provide

$$T(C', \mathcal{M}_\alpha) = \prod_{k=1}^K T\left(C^{(k)}, \mathcal{M}_{\alpha_{\text{cuts}(k)}}^{(k)}\right), \quad (19)$$

from Eq. (17) and Eq. (18) where $\alpha_{\text{cuts}(k)} \subset \alpha$ holds only the completions that act on the fragment k . We note that $\text{SC}_{i,j}$, $i = 1, \dots, K$, $j = 1, \dots, 8^{m_i}$, in Fig. 2 represents the subcircuits that runs fragment $C^{(i)}$ with the j -th Pauli completion of its m_i local cuts. Executing all $\text{SC}_{i,j}$ on the QPU provides the conditional probabilities $P^{(i)}(b_i | \alpha_{\text{cuts}(i)})$. Inserting these probabilities into the squared-modulus of (19) yields the knitted distribution

$$P(\mathbf{b}) = \sum_{\alpha} \left(\prod_{m=1}^M c_{\alpha_m} \right) \prod_{i=1}^K P^{(i)}(b_i | \alpha_{\text{cuts}(i)}). \quad (20)$$

Each non-Clifford gate within C can be expressed through the QPD expansion $U(\theta) = \sum_j c_j(\theta) G_j$, with an associated overhead $\Gamma(\theta) = \sum_j |c_j(\theta)|$, such Pauli measurement pairs with a cut set S increases the Monte Carlo variance

by at most $\prod_{g \in S} \Gamma(\theta_g) \leq 9^{|S|}$. Given that each Pauli completion G_j is a Clifford gate, every $SC_{i,j}$ can be efficiently simulated classically in polynomial time, as per the Gottesman-Knill theorem. Consequently, the total post-processing effort scales as $O(9^{|S|} \text{poly}(n))$. Notice that the final global result can be represented by the trigonometric function $\mathcal{G}[s] \leftarrow \mathcal{G}[s] + c_i \cdot \text{sign}(q) \cdot \mathcal{P}$, where sign is given by the PARITY function and \mathcal{P} is the probability of the expected results of the sub-experiments calculated by measured bit strings based on the shots. This is because achieving the optimal local operation classical communication (LOCC) [8] overhead requires internal communication in each subcircuit. In the realistic noisy quantum simulation scenario, we emphasize that the advantage of using the cutting pool \mathcal{G} defined in Def. 2 allows the reconstruction algorithm to process the quantum bit string results with one pass to substitute the sequential processing.

5 Result

5.1 Ablation study

Q^{*}: Does the **shardQ** protocol facilitate state-of-the-art QPD results in the context of three-dimensional tensor encoding?

To address this inquiry, we demonstrate that the protocol effectively reduces crosstalk errors in the current noisy QPU, where the goal is to compare the **shardQ** protocol with and without the original encoder.

ShardQ Analysis. We demonstrate that the protocol provides a lower error rate for quantum circuit simulation because the protocol physically cuts the longest entanglement gate into local unitary operations, as shown in Fig. 5. The benefits arise from two reasons. First, the idle qubit performs single-qubit Pauli operations after the cut, which allows a longer coherence time because of the probe of the pulse in the conductor of the superconducting circuit hardware. Second, our MPS-enabled compilation further reduces the transpiled circuit depth, which limits the entanglement gates, allowing the results to have better locality using shallower subcircuits because of the tensor approximate contraction. We emphasize that the root mean square error (RMSE) of the quantum reconstructed data and true data trend reveal that the cut simulation constantly outperforms the original uncut simulation.

Overhead Evaluation. However, we note that the classical simulation overhead is unavoidable because of the QPD technique. In the ablation test, we show that beyond two cuts, diminishing performance returns coupled with exponential time growth demonstrate computational intractability, validating the practical selection of the two-cut as the optimum configuration for real-world deployment, as shown in Fig. 6. Additionally, we indicate that the optimum settings of the

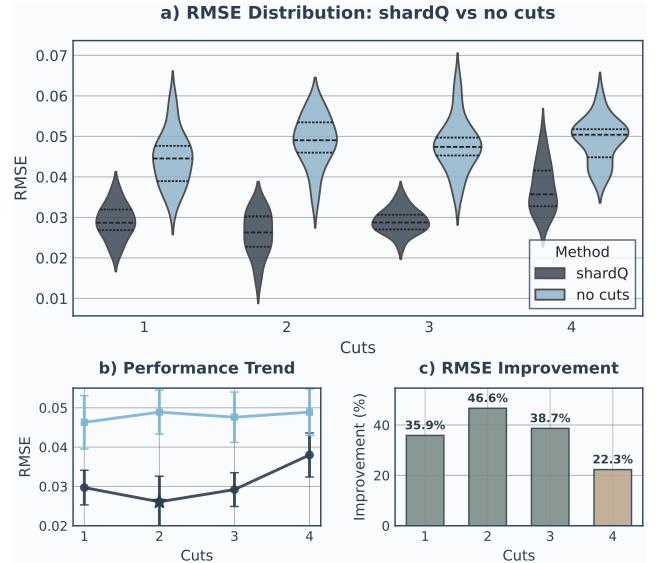


Figure 5. Ablation study evaluating **shardQ** performance across cuts. (a) the middle dashed line symbolize the median with two dashed lines above and below indicate the 25% and 75% percentile. (b) RMSE performance trending line with two cuts at optimum. (c) Relative RMSE improvement quantification with color-coded bars indicating improvement magnitude (beige: 15-25%, sage: >25%). Error bars represent standard deviation across independent trials.

two cuts also have the best trade-off concerning the classical overhead and error rates.

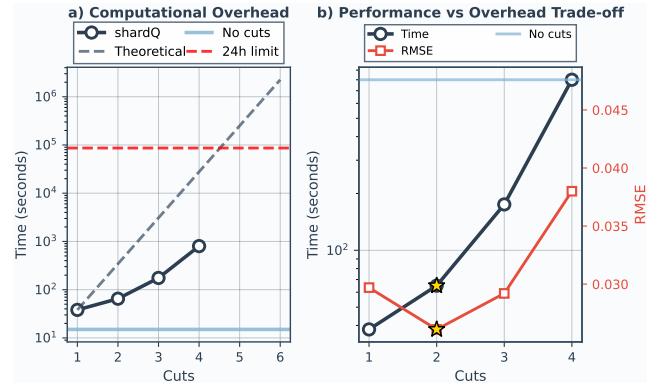


Figure 6. Computational overhead analysis and performance trade-off evaluation for **shardQ** method. (a) Exponential computational scaling showing measured execution times (circles) closely following Eq. (12) with 24-hour practical limit (red dashed line) exceeded beyond 4 cuts. Baseline no-cuts method maintains constant ~ 15 s execution time. (b) The gold star indicates the optimum trade-off corresponding with Fig. 5.

5.2 Application

Furthermore, we demonstrate that the optimal two-cut enabled quantum tensor encoding simulation on the IBM ideal

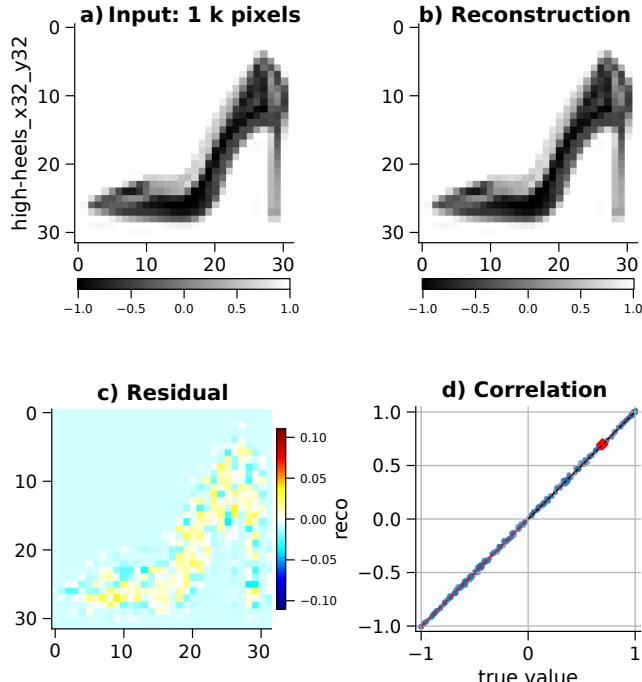


Figure 7. The correlation between the reconstructed values and the ground truth is tight.

simulator facilitates the near-perfect reconstruction of a grayscale image, as shown in Fig. 7. The total encoded tensor length is determined by $2^{n_a} * n_d$; thus, for an image comprising 1,000 pixels, we select nine address qubits and two data qubits for encoding the image. We allocate 3,000 shots per classical data-encoded position (2^9), aligning with the methodology of employing ideal GPU quantum image simulation [15]; consequently, the total number of shots per subcircuit amounted to 1.5 million. The results indicate that the **shardQ** protocol yields a quantum-encoded image with an error rate of less than 1% and a negative four orders of magnitude standard deviation.

6 Discussion

The **shardQ** protocol presents an **NISQ**-friendly framework for quantum tensor encoding circuits, particularly within gate-based quantum platforms such as IBM, and could be extended to trapped-ion-based platforms such as IonQ. By utilizing the **SparseCut** and global bit string reconstruction techniques, our approach addresses a significant challenge: extending quantum circuit simulation to fault-tolerant quantum computing (**FTQC**). This is crucial because future HPC-integrated quantum platforms will necessitate the division of quantum circuit simulations or approximations across different hardware. Our experimental findings validate the feasibility of the optimal-cut strategy and low-error-rate quantum image encoding. We anticipate that our protocol

will facilitate the future development of quantum computers capable of executing deeper and more intricately structured entangled circuits, thereby producing reliable results. Our artifact is available at: <https://anonymous.com>

Acknowledgments

This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award NERSC DDR-ERCAP0034486.

References

- [1] Jan Balewski, Mercy G Amankwah, E Bethel, Talita Perciano, and Roel Van Beeumen. 2025. EHands: Quantum Protocol for Polynomial Computation on Real-Valued Encoded States. *arXiv preprint arXiv:2502.15928* (2025).
- [2] Jan Balewski, Mercy G Amankwah, Roel Van Beeumen, E Wes Bethel, Talita Perciano, and Daan Camps. 2024. Quantum-parallel vectorized data encodings and computations on trapped-ion and transmon QPUs. *Scientific Reports* 14, 1 (2024), 3435.
- [3] Aleksandr Berezutskii, Minzhao Liu, Atithi Acharya, Roman Ellerbrock, Johnnie Gray, Reza Haghshenas, Zichang He, Abid Khan, Viacheslav Kuzmin, Dmitry Lyakh, et al. 2025. Tensor networks for quantum computing. *Nature Reviews Physics* (2025), 1–13.
- [4] Pablo Bermejo and Román Orús. 2023. Variational quantum and quantum-inspired clustering. *Scientific Reports* 13, 1 (2023), 13284.
- [5] Abraham Bookstein, Vladimir A Kulyukin, and Timo Raita. 2002. Generalized hamming distance. *Information Retrieval* 5 (2002), 353–375.
- [6] Sergey Bravyi and Alexei Kitaev. 2005. Universal quantum computation with ideal Clifford gates and noisy ancillas. *Physical Review A—Atomic, Molecular, and Optical Physics* 71, 2 (2005), 022316.
- [7] Sergey Bravyi, Graeme Smith, and John A Smolin. 2016. Trading classical and quantum computational resources. *Physical Review X* 6, 2 (2016), 021043.
- [8] Lukas Brenner, Christophe Piveteau, and David Sutter. 2023. Optimal wire cutting with classical communication. *arXiv preprint arXiv:2302.03366* (2023).
- [9] Almudena Carrera Vazquez, Caroline Tornow, Diego Riste, Stefan Woerner, Maika Takita, and Daniel J Egger. 2024. Combining quantum processors with real-time classical communication. *Nature* 636, 8041 (2024), 75–79.
- [10] Don Coppersmith. 2002. An approximate Fourier transform useful in quantum factoring. *arXiv preprint quant-ph/0201067* (2002).
- [11] David P DiVincenzo. 1995. Two-bit gates are universal for quantum computation. *Physical Review A* 51, 2 (1995), 1015.
- [12] Simone Faro, Arianna Pavone, Caterina Viola, et al. 2024. Families of Constant-Depth Quantum Circuits for Rotations and Permutations. In *Proceedings of the 25nd Italian Conference on Theoretical Computer Science, Torino, Italy*.
- [13] Gian Gentilletta, Friederike Metz, and Giuseppe Carleo. 2024. Overhead-constrained circuit knitting for variational quantum dynamics. *Quantum* 8 (2024), 1296.
- [14] Lov K Grover. 1996. A fast quantum mechanical algorithm for database search. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing* (Philadelphia, Pennsylvania, USA) (STOC '96). Association for Computing Machinery, New York, NY, USA, 212–219. doi:10.1145/237814.237866
- [15] Ziqing Guo, Ziwen Pan, and Jan Balewski. 2025. Q-GEAR: Improving quantum simulation framework. *arXiv preprint arXiv:2504.03967* (2025).

- [16] Bernhard Haeupler, Richard Hladík, Václav Rozhoň, Robert E Tarjan, and Jakub Tetěk. 2024. Universal optimality of dijkstra via beyond-worst-case heaps. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2099–2130.
- [17] Michael Hart and John McAllister. 2024. Reconstructing Cut Quantum Circuits Maximising Fidelity between Quantum States. In *Proceedings of the 21st ACM International Conference on Computing Frontiers*. 224–231.
- [18] IBM Quantum. 2024. IBM Quantum System: Marrakesh. https://quantum.cloud.ibm.com/computers?system=ibm_marrakesh. Accessed: November 10, 2025.
- [19] Mingrui Jing, Chengkai Zhu, and Xin Wang. 2025. Circuit knitting facing exponential sampling-overhead scaling bounded by entanglement cost. *Physical Review A* 111, 1 (2025), 012433.
- [20] Angus Lowe, Matija Medvidović, Anthony Hayes, Lee J O’Riordan, Thomas R Bromley, Juan Miguel Arrazola, and Nathan Killoran. 2023. Fast quantum circuit cutting with randomized measurements. *Quantum* 7 (2023), 934.
- [21] Kosuke Mitarai and Keisuke Fujii. 2019. Methodology for replacing indirect measurements with direct measurements. *Physical Review Research* 1, 1 (2019), 013006.
- [22] Kosuke Mitarai and Keisuke Fujii. 2021. Constructing a virtual two-qubit gate by sampling single-qubit operations. *New Journal of Physics* 23, 2 (2021), 023021.
- [23] Hakop Pashayan, Joel J Wallman, and Stephen D Bartlett. 2015. Estimating outcome probabilities of quantum circuits using quasiprobabilities. *Physical review letters* 115, 7 (2015), 070501.
- [24] Tianyi Peng, Aram W Harrow, Maris Ozols, and Xiaodi Wu. 2020. Simulating large quantum circuits on a small quantum computer. *Physical review letters* 125, 15 (2020), 150504.
- [25] Michael A Perlin, Zain H Saleem, Martin Suchara, and James C Osborn. 2021. Quantum circuit cutting with maximum-likelihood tomography. *njp Quantum Information* 7, 1 (2021), 64.
- [26] Christophe Piveteau and David Sutter. 2023. Circuit knitting with classical communication. *IEEE Transactions on Information Theory* 70, 4 (2023), 2734–2745.
- [27] Minati Rath and Hema Date. 2024. Quantum data encoding: A comparative analysis of classical-to-quantum mapping techniques and their impact on machine learning accuracy. *EPJ Quantum Technology* 11, 1 (2024), 72.
- [28] Niall Robertson, Albert Akhriev, Jiri Vala, and Sergiy Zhuk. 2025. Approximate quantum compiling for quantum simulation: A tensor network based approach. *ACM Transactions on Quantum Computing* 6, 3 (2025), 1–15.
- [29] Lukas Schmitt, Christophe Piveteau, and David Sutter. 2025. Cutting circuits with multiple two-qubit unitaries. *Quantum* 9 (2025), 1634.
- [30] Ibrahim Shehzad, Edwin Pednault, James R Garrison, Caleb Johnson, Bryce Fuller, and Jennifer R Glick. 2024. Automated cut finding and circuit knitting on large quantum circuits. In *2024 IEEE International Conference on Quantum Computing and Engineering (QCE)*, Vol. 2. IEEE, 406–407.
- [31] Peter W Shor. 1999. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM review* 41, 2 (1999), 303–332.
- [32] Wei Tang and Margaret Mardonosi. 2022. Cutting quantum circuits to run on quantum and classical platforms. *arXiv preprint arXiv:2205.05836* (2022).
- [33] Mar Tejedor, Berta Casas, Javier Conejero, Alba Cervera-Lierta, and Rosa M Badia. 2025. Distributed Quantum Circuit Cutting for Hybrid Quantum-Classical High-Performance Computing. *arXiv preprint arXiv:2505.01184* (2025).
- [34] Kristan Temme, Sergey Bravyi, and Jay M Gambetta. 2017. Error mitigation for short-depth quantum circuits. *Physical review letters* 119, 18 (2017), 180509.
- [35] Songqinghao Yang and Prakash Murali. 2024. Understanding the Scalability of Circuit Cutting Techniques for Practical Quantum Applications. *arXiv preprint arXiv:2411.17756* (2024).
- [36] Jun Zhang, Jiri Vala, Shankar Sastry, and K Birgitta Whaley. 2003. Geometric theory of nonlocal two-qubit operations. *Physical Review A* 67, 4 (2003), 042313.

A Theoretical Analysis of Quasi-Probability Decomposition

In this section, we present a detailed proof of the decomposition of the CX gate within our protocol, which involves two address qubits and one data-qubit tensor data encoder. Notably, the permutation of UCR_y gates represents the current state-of-the-art approach for differential data encoding blocks, which can be realized through the constant-depth quantum circuits demonstrated in [12]. Specifically, the single rotation gate and controlled Y rotation gate are

$$R_y(\theta) = \cos\left(\frac{\theta}{2}\right)I - i \sin\left(\frac{\theta}{2}\right)Y, \quad (21)$$

$$CRY(\theta) = |0\rangle\langle 0| \otimes I + |1\rangle\langle 1| \otimes R_y(\theta).$$

The commutative principle is expressed in Eq. (18-21) of [2], where the CRY can be decomposed into a R_y with a CX gate. Such a decomposition enables the construction of the encoder circuit with two address qubits and one data qubit, denoted as C_{21} shown in Fig. 8. Note that the permutation of CX is encoded with the gray code [5] that optimizes the Hamming distance between the neighbors with the maximum value of one. The advantages become more apparent when we have more address qubits because the traversal of control configurations allows more efficient updating of CX control patterns; only one control bit changes between consecutive operations, which minimizes the number of CX gates required to reconfigure the multi-controlled rotations. We recall that the number of the address qubits is n_a , hence, the R_y gates are parameterized by $P_0\{a_{00}\}, \dots, P_0\{binary(2^{n_a})\}$. Specifically, the data-to-angle encoding is given by

$$\theta^{\text{final}} = \text{Gray}(\text{FWHT}(\arccos(\mathbf{d}))) \quad (22)$$

where $\mathbf{d} \in [-1, 1]^N$ is the input data vector, $\arccos(\cdot)$ maps data to angles, FWHT is the scaled fast Walsh-Hadamard transform, and the gray code permutation is used to optimize the control pattern for efficient quantum circuit simulation.

Given that the simplest maximum cut number set is one using Alg. 1, we obtain the cutting gate index as the sixth from circuit C_{21} . Then, considering that the uncut circuit is decomposed into six subcircuits because of the Pauli group, each subcircuit is attained through the probabilistic Clifford gate representation. Here, the KAK decomposition for C_{21} is given by implementing a uniformly controlled rotation with address qubits q_0, q_1 and data qubit q_2 denoted by

$$|\psi\rangle = \frac{1}{2} \sum_{i=0}^3 |i\rangle \otimes \left(\cos \frac{\alpha_i}{2} |0\rangle + \sin \frac{\alpha_i}{2} |1\rangle \right). \quad (23)$$

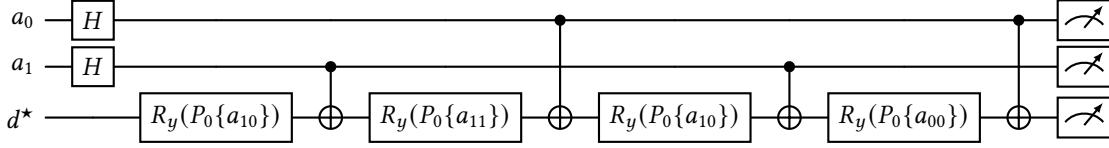


Figure 8. The example of two address and one data qubits **QCrank**.

Then, by recalling that § 4.3, each controlled rotation gate can be decomposed to

$$\text{UCR}_y(\vec{\alpha}) = \sum_{a_1, a_2 \in \{\pm 1\}} c_{a_1, a_2} \cdot P_{a_1}^{(0)} \otimes P_{a_2}^{(1)} \otimes R_y(\theta_{a_1, a_2}). \quad (24)$$

where P is the Pauli measurement projector. We note that the number of cut indices results in the $\mathcal{O}(n)$ linearly increasing of the QPD measurement stored as the temporary results with the coefficient shown in Eq. (29) because **SparseCut** allows cuts in the data block level, as shown in Fig. 2. To prove the CX cut, we first recall the CZ gate decomposition shown in Fig. 6 of [24]. Additionally, the CX can be represented by CZ with

$$CX = (I \otimes H) CZ (I \otimes H). \quad (25)$$

However, only three terms (Pauli X, Pauli Z, and Hadamard gate) are required for the **QCrank**. Here, CZ can be summed by R_z gate because of the two virtual qubit gate decomposition principle [21]

$$CZ = \sum_{a_1, a_2 \in \{\pm 1\}^2} a_1 a_2 \left\{ RZ \left(\frac{a_1 \pi}{2} \right) \otimes RZ \left(\frac{a_2 \pi}{2} \right) \right\}. \quad (26)$$

With Eq. (26), it applies the conjugation to Eq. (25)

$$CX = (I \otimes H) \cdot \left[\sum_{a_1, a_2 \in \{\pm 1\}^2} a_1 a_2 \left\{ RZ \left(\frac{a_1 \pi}{2} \right) \otimes RZ \left(\frac{a_2 \pi}{2} \right) \right\} \right] \cdot (I \otimes H) \quad (27)$$

Because $H \cdot RZ(\theta) \cdot H = RX(\theta)$ (referred to Clifford decomposition table Table 1), this gives us separate terms conditioned based on the coefficients

$$H \cdot RZ \left(\frac{a_2 \pi}{2} \right) \cdot H = RX \left(\frac{a_2 \pi}{2} \right). \quad (28)$$

The general decomposition for the two virtual qubit gates can be written in the format of super operators with a single qubit operation sandwiching the observable density matrix. Let us expand the observables into three Pauli matrices and identity term, therefore, the coefficients are can be defined

by

$$\begin{aligned} O_1 &= I, & \rho_1 &= |+\rangle\langle+|, & c_1 &= +\frac{1}{2}, \\ O_2 &= I, & \rho_2 &= |- \rangle\langle-|, & c_2 &= +\frac{1}{2}, \\ O_3 &= X, & \rho_3 &= |0\rangle\langle 0|, & c_3 &= +\frac{1}{2}, \\ O_4 &= X, & \rho_4 &= |1\rangle\langle 1|, & c_4 &= -\frac{1}{2}, \\ O_5 &= Y, & \rho_5 &= |-i\rangle\langle-i|, & c_5 &= +\frac{1}{2}, \\ O_6 &= Y, & \rho_6 &= |+i\rangle\langle+i|, & c_6 &= -\frac{1}{2}, \\ O_7 &= Z, & \rho_7 &= |+\rangle\langle+|, & c_7 &= +\frac{1}{2}, \\ O_8 &= Z, & \rho_8 &= |- \rangle\langle-|, & c_8 &= -\frac{1}{2}. \end{aligned} \quad (29)$$

Here, O_i is the observable, ρ_i is the eigenstate, c_i is the

Gate	Decomposition
X	$H \cdot Z \cdot H$
Y	$H \cdot Z \cdot H \cdot Z$
Z	S^2
R_X	$H \cdot S^\dagger \cdot H$
R_Y	$S \cdot H \cdot S^\dagger \cdot H \cdot S^\dagger$
R_Z	S^\dagger
R_{YZ}	$H \cdot S^\dagger \cdot H \cdot Z$
R_{ZX}	$S^\dagger \cdot H \cdot S^\dagger \cdot H \cdot S^\dagger$
R_{XY}	$H \cdot Z \cdot H \cdot S^\dagger$
Π_X	$S \cdot H \cdot S \cdot H \cdot P_0 \cdot H \cdot S^\dagger \cdot H \cdot S^\dagger$
Π_Y	$H \cdot S^\dagger \cdot H \cdot P_0 \cdot H \cdot S \cdot H$
Π_Z	P_0
Π_{YZ}	$S \cdot H \cdot S \cdot H \cdot P_0 \cdot H \cdot S \cdot H \cdot S^\dagger$
Π_{ZX}	$H \cdot S^\dagger \cdot H \cdot P_0 \cdot H \cdot S \cdot H \cdot Z$
Π_{XY}	$P_0 \cdot H \cdot Z \cdot H$

Table 1. Decomposition of gates in terms of H Hadamard gate, S Phase gate, S^\dagger Reverse phase gate, Z Pauli Z gate, and P_0 Z-based measurement gate used in controlled-rotation circuits. We note that the decomposition for the Pauli groups is defined in the Clifford gate group [6] except for the Z measurement gate. In local operator classic communication, **QPD** stores the mid-circuit measurement result for the global measurement reconstruction.

coefficient. Note that the cut gate has one prepared state and a measured observable. Specifically, on the preparation side, we apply a 1-qubit density matrix $\rho_\lambda = |\lambda\rangle\langle\lambda|$ that is the eigenstate of the Pauli as appearing in Eq. (29). On the measurement side, $s = \pm 1$ serves as the measured and

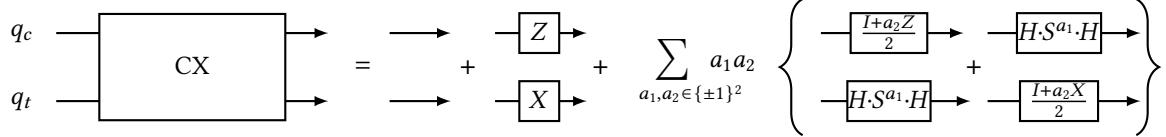


Figure 9. CX-decomposition written exclusively with H , S , S^\dagger , Z , X and the projector $P_0 = \frac{1}{2}(I + Z)$. Every rotation $R_Z(\pm\pi/2)$ has been replaced by S or S^\dagger ; every $R_X(\pm\pi/2)$ is implemented as the Clifford $H S^{(\dagger)} H$.

recorded eigenvalues. In the experiment, we used a gate-based quantum circuit simulation. After shot-averaging,

$$\langle s \rangle = \text{Tr}[P \rho_\lambda] = \lambda, \quad (30)$$

so the product of eigenvalue and shot average reproduces the Pauli and we denote the QPD Pauli representation of CX gate as

$$\text{CX}_{c \rightarrow t} = \frac{1}{2} \left(I_c \otimes I_t + Z_c \otimes I_t + I_c \otimes X_t - Z_c \otimes X_t \right), \quad (31)$$

where c is the conditioned (control) qubit and t is the target qubit. Therefore, to generalize the eight Pauli observable and re-prepare for the global states, we categorize the six terms denoted in Table 2. Here, Alg. 1 selects the Clifford

Table 2. Post-measurement bases associated with the six transformed Pauli-Stinespring operators Φ'_i . S : unitary on both sides, M_A : measure & reprepare qubit 1, M_B : mid-measurement on qubit 2.

Φ'_i from (E1 [22])	Applied gates	Computation basis
$S(I \otimes I)$	no mid measurement	identity term
$S(A \otimes B)$	no mid measurement	$Z \otimes X$ term
$M_A \otimes S(e^{i\pi B/4})$	$S^\dagger H + Z$ -meas.	S^\dagger basis ($ \pm i \rangle$)
$M_A \otimes S(e^{-i\pi B/4})$	$SH + Z$ -meas.	S basis ($ \mp i \rangle$)
$S(e^{i\pi A/4}) \otimes M_B$	$I + Z$ -meas.	computational ($ 0\rangle, 1\rangle$)
$S(e^{-i\pi A/4}) \otimes M_B$	$H + Z$ -meas.	Hadamard ($ \pm \rangle$)

bases for efficient quantum hardware simulation because of the efficiency of the Clifford group overhead simulation. We note that rows 1 and 2 share the same hardware setting; the difference is an S vs. S^\dagger gate; likewise 3 and 4 share Hadamard and no Hadamard settings. Each basis yields two possible classical outcomes, which correspond to the “ \pm ” states listed in the table. Because distinct quantum circuits (measurement settings) are required, the eight table rows are recovered by classical post-processing of the outcomes from the six circuits. Therefore, to produce the single cut as shown in the Fig. 8, we combine Eq. (31) and Eq. (29). Hereby we proved Fig. 9. Note that, in the transpiled version of quantum circuit mapping to the physical qubits, we do not consider the UCR_y gate in the **SparseCut** algorithm. To complete the proof for generalize UCR_y, therefore, by combining Eq. (24)

and E1 [22], the complete UCR_y gate decomposes as

$$\text{UCR}_y(\vec{\alpha}) = \frac{1}{4} \sum_{j=1}^6 w_j \cdot \mathcal{M}_j \otimes \mathcal{R}_j \quad (32)$$

$$= \frac{1}{4} \left[\begin{array}{l} w_1 \langle +i | \otimes R_y(\theta_1) + w_2 \langle -i | \otimes R_y(\theta_2) \\ + w_3 \langle +i | \otimes I \cdot R_y(\theta_3) + w_4 \langle -i | \otimes Z \cdot R_y(\theta_4) \\ + w_5 \langle 0 | \otimes R_y(\theta_5) + w_6 \langle + | \otimes R_y(\theta_6) \end{array} \right] \quad (33)$$

For each subcircuit produces measurement outcomes with probabilities

$$P(m_j = \pm 1) = \frac{1 \pm \langle \psi | \sigma_j^{(0)} \otimes I^{(1)} \otimes I^{(2)} | \psi \rangle}{2} \quad (34)$$

where $|\phi\rangle$ is Eq. (23). Finally, we produce the global measurement reconstruction using Alg. 2. The original expectation values are recovered through

$$\langle R_y(\alpha_i) \rangle = \sum_{j=1}^6 c_j \cdot \text{Corr}(m_j^{(0)}, m_j^{(2)}) \quad (35)$$

where $\text{Corr}(m_j^{(0)}, m_j^{(2)})$ represents the correlation between the address and data qubit measurements in subcircuit j , thereby completing the end-to-end quantum tensor encoder circuit partitioning and recomposition.