Practical, robust and optimal quantum state preparation: bucket-brigade approach

Xiao-Ming Zhang[1, 2]

[1]*School of Physics, South China Normal University, Guangzhou 510006, China*
[2]*Center on Frontiers of Computing Studies, School of Computer Science, Peking University, Beijing 100871, China*

Quantum state preparation, as a general process of loading classical data to quantum device, is essential for end-to-end implementation of quantum algorithms. Yet, existing methods suffer from either high circuit depth or complicated hardware, limiting their practicality and robustness. The bucket-brigade architecture for quantum random access memory (QRAM) potentially offers a practical and robust solution, but it is currently limited to transformations in the form of $|j\rangle|0\rangle \to |j\rangle|D_j\rangle$. In this work, we overcome these limitations with bucket-brigade approaches for general quantum state preparation. The tree architectures of the approaches represent the simplest connectivity required for achieving sub-exponential circuit depth. Leveraging the bucket-brigade mechanism, our approaches exhibit exponential robustness compared to conventional methods, with infidelity scaling as $O(\varepsilon \text{polylog}(N))$ for gate error ε and data size N. Moreover, our algorithms is the first one that simultaneously achieve linear Clifford+T circuit depth, gate count number, and space-time allocation. These advancements offer the opportunity for processing big data in both near-term and fault-tolerant quantum devices.

An end-to-end application of quantum algorithm requires the loading of classical data to a quantum device. Quantum state preparation, as a general classical data loading process, lies at the foundation of quantum algorithms. Given an $N = 2^n$ dimensional normalized vector $[\psi_0, \psi_1, \cdots, \psi_{N-1}]$, quantum state preparation, from a trivial initial state $|0\rangle^n$, can be expressed as the following transformation

$$U_{\text{sp}}|0\rangle^n \otimes |\text{anc}\rangle = |\psi\rangle \otimes |\text{anc}\rangle, \tag{1}$$

where $|\psi\rangle \equiv \sum_{j=0}^{N-1} \psi_j |j\rangle$ is the target state, and $|\text{anc}\rangle$ represents the state of an ancillary system. The study of the quantum state preparation also has its fundamental motivations, as it indicates the space-time resource required to transform an arbitrary pure quantum state to another.

Various protocols has been proposed in the literature to realized Eq. (1) [1–14]. The first nontrivial idea was proposed independently by Long, Sun[1] and Grover, Rudolph [2] based on multi-controlled-rotations. Subsequent works have improved the gate count to $O(N)$, which is optimal [3, 4]. Although large gate count is inevitable in general, it is possible to trade time (circuit depth) for space (ancillary qubit). Recently, low-depth quantum state preparation with $O(n)$ circuit depth that matches the lower bound has been achieved by several protocols [5–13], provided sufficient number of ancillary qubits. These results represent the ultimate speed limit of loading general classical to a quantum device.

Despite the remarkable progress that has been made in low-depth quantum state preparation, current protocols are far from practical. On one hand, the robustness of existing protocols [6–14] can not be guaranteed. The worst-case single- and two-qubit gate count of state preparation is $O(N)$, regardless of the space-time trade-off. A direct evaluation indicates that to achieve a constant preparation fidelity, one should suppress the error at each elementary gate to the level of $O(N^{-1})$. For applications on large data set, this requirement is too stringent to be practical, especially for near-term quantum devices. Even in the fault-tolerant setting, the gate error requirement of $O(N^{-1})$ is also challenging. Take the surface code [15] scheme as an example, the code distance of each logic qubit should be at least $d = \Omega(n)$, which means that substantial among of classical data processing and corrections gates are required. On the other hand, existing protocols, such as [6–9, 11–13] assumes fully connectivity, which is not friendly for current quantum devices. In superconducting circuit system, qubits are typically connected by couplers [16, 17], and only nearest neighbor interaction is available. There are other systems where better connectivity is available, such as trapped ion [18] and neutral-atom arrays [19]. However, simultaneous rearrangement of connectivities requires complicated shuttling, which is time-costly and may substantially affect the control accuracy. Although protocols in [10, 14] have sparse connectivity, the architecture is far from optimal.

Besides, quantum random access memory (QRAM) [20–25] enjoys both robustness and simple connectivity, which aims at performing the following transformation

$$|j\rangle|0\rangle \to |j\rangle|D_j\rangle \tag{2}$$

coherently for $0 \leqslant j \leqslant N - 1$, for some data D_j to be encoded. Among various implementation methods, the circuit-based, bucket-brigade QRAM stands out, because it is proven to be exponentially more noise resilience [24, 26]. Moreover, qubits in Bucket-brigade QRAM are connected as a binary tree. This architecture is friendly for practical implementation as it can be realized in 2-dimension, while each qubit only need to be connected to three of its nearest neighbor, which is optimal. Various schemes has been proposed to realize the Bucket-brigade QRAM in different systems, such as neutral atom [21, 22], superconducting circuit [23], spin-photon network [27], etc. Unfortunately, recent proposal can only realize Eq. (2), as oppose to the general data loading process in Eq. (1). This substantially restrict the applicability of QRAM. As a typical example, the block encoding of linear combination of unitaries [28, 29] requires Eq. (2) sandwiched by Eq. (1) and its inverse.

In this work, we overcome the practicality and robustness challenges of quantum state preparation by bucket-brigade ap-

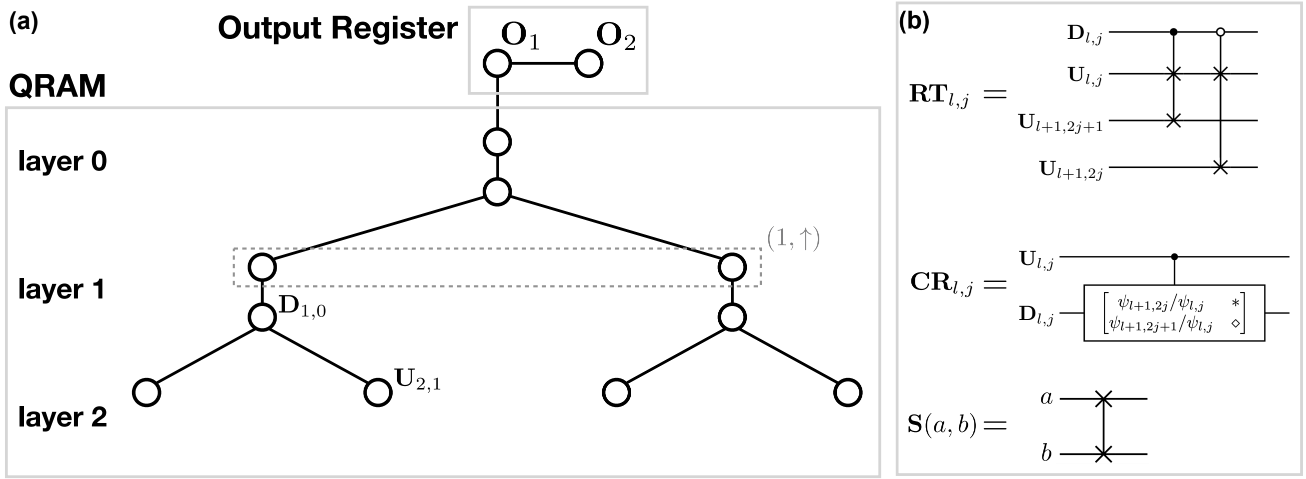


FIG. 1: (a) Hardware architecture of quantum state preparation protocols and corresponding notations in the main text. Each circle represent a qubit, and each line represents to the connection between a pair of qubits. (b) Definitions of routing ($\mathbf{RT}_{l,j}$), controlled-rotations ($\mathbf{CR}_{l,j}$), and swap $\mathbf{S}(a,b)$ operations. In the operation $\mathbf{CR}_{l,j}$, labels $*$ and $\diamond$ represent some values that make the matrix to be a unitary.

proaches. The hardware of our state preparation task is as simple as the binary tree architecture of bucket-brigade QRAM. In particular, each qubit connects to at most three of other qubits, which is optimal for achieving a subexponential circuit depth. Our algorithm is simple, and based on a novel fanin process that iteratively moves a pointer from the lth layer to the $(l+1)$th layer, and at the same time encodes the amplitudes $\{\psi_j\}$ through rotations controlled by the pointer. Then, the QRAM is uncomputed with a fanout process [24]. Thanks to the bucket-brigade mechanism, a dramatic advantage is that the infidelity scales polylogarithmically with data size. We present two variations of our method. The 2-qubit-per-node protocol is simpler while achieving optimal single- and two-qubit circuit complexity. Its state preparation infidelity scales as $1-F \leqslant A\varepsilon n^3$ for some constant A and elementary gate error ε. Yet, its Clifford+T circuit complexity is suboptimal. The 3-qubit-per-node protocol is more complex, but can simultaneously achieve the state-of-the-art Clifford+T gate count, circuit depth and space-time allocaltion. It also improve the infidelity scaling to $1-F \leqslant A\varepsilon n^2$. Remarkably, this infidelity scaling is even better than the best known result for two-level-based bucket-brigade QRAM, which aims at performing Eq. (2).

Hardware architecture. In the main text, we introduce the 2-qubit-per-node protocol, while the improved 3-qubit-per-node protocol is provided in Appendix. B.

As shown in Fig. 1, our protocol contains a bucket-brigade QRAM and an n-qubits output register. The bucket-brigade QRAM resembles an $(n+1)$ layer binary tree and each node of the tree corresponds to two qubits. To be specific, the lth $(0 \leqslant l \leqslant n)$ layer contains an upper and a lower sublayers, denoted as $(l,\uparrow)$ and $(l,\downarrow)$ respectively. An exception is that the leaf layer has only upper sublayers. Both sublayers contain totally 2^l qubits. We denote the jth qubit of $(l,\uparrow)$ and $(l,\uparrow)$ as $\mathbf{U}_{(l,j)}$ and $\mathbf{D}_{(l,j)}$. Qubits in QRAM are connected as a tree architecture, and each qubit connects only to their parent or children. $\mathbf{U}_{l,j}$ has one child $\mathbf{D}_{l,j}$, and $\mathbf{D}_{l,j}$ for $l \neq n$ has two children $\mathbf{U}_{l+1,2j}$ and $\mathbf{U}_{l+1,2j+1}$. The output register contains n qubits, each denoted as $\mathbf{O}_j$ (from $j=1$ to $j=n$). They are arranged as a line with nearest-neighbor coupling, and $\mathbf{O}_1$ also connect to the root of QRAM, i.e. $\mathbf{U}_{0,0}$. In this architecture, each qubit connects to at most 3 of the other qubits, which is optimal. This is a because graph with degree 2 can only form trivial lines or rings, qubit connections in such ways are insufficient for achieving subexponential circuit depth.

Fanin phase. In this phase, we only perform operations in QRAM. We first introduce some notations of quantum states. Suppose $\mathcal{S}$ is a set of qubits, we use the "activation" basis $|\mathcal{S}\rangle$ to represent that all qubits in S is activate (i.e. at state $|1\rangle$), while all other qubits are at state $|0\rangle$. Formally, we have $|\mathcal{S}\rangle \equiv \otimes_{v\in\mathcal{V}^{\mathrm{QRAM}}} |v \in \mathcal{S}\rangle_v$, where $|\cdot\rangle_v$ represents the state of qubit v, and the "True" or "False" result of $v \in \mathcal{S}$ correspond to the binary 1 or 0. $\mathcal{V}^{\mathrm{QRAM}}$ represents all qubits in QRAM.

Let $|\psi_0\rangle = |\{\mathbf{U}_{0,0}\}\rangle$ be the initial state, we perform the following transformation

$$|\psi_l\rangle \longrightarrow |\psi_{l+1}\rangle, \quad |\psi_l\rangle \equiv \sum_{j=0}^{2^l-1} \psi_{l,j} |\mathscr{B}_{l,j}\rangle, \tag{3}$$

iteratively from $l=0$ to $l=n-1$, where $\psi_{l,j}$ will be defined later, and $\mathscr{B}_{l,j}$ is a set of qubits that will be clarified as follows. For qubit $\mathbf{D}_{(l,j)}$ at lower sublayers, we let $\mathcal{P}_\downarrow\left[\mathbf{D}_{(l,j)}\right] = \mathbf{D}_{(l-1,\lceil j/2 \rceil)}$ be its grandparent, which is also at the lower sublayers. Accordingly, we represent all ancestors of $\mathbf{D}_{(l,j)}$ in the lower sublayers as $\mathscr{A}_{l,\downarrow,j} = \left\{\mathcal{P}_\downarrow^{\odot m}[\mathbf{D}_{(l,j)}] \middle| 1 \leqslant m \leqslant l\right\}$, which contains totally l qubits. $|\mathscr{B}_{l,j}\rangle$ represents the following quantum state: at the subset of upper sublayers, only single qubit, $\mathbf{U}_{l,j}$, is activated and it serves as a *pointer*. At the subset of lower sublayers, $\mathscr{A}_{l,\downarrow,j}$ (which is also all ancestors of $\mathbf{U}_{l,j}$ at lower sublayers) is at computational basis $|j_1 j_2 \cdots j_l\rangle$, i.e. the first l bits of j. All other qubits are at state $|0\rangle$. More specifically, qubit set $\mathscr{B}_{l,j}$ has the following formal definition

Algorithm 1 Quantum state preparation. Initial state $|\mathrm{U}_{0,1}\rangle$, amplitudes $\{\psi_{l,j}\}$ as input

```
for $l = 0, \cdots, n-1$:
    implement $\mathbf{PRT}_l \mathbf{PCR}_l$
for $m = 0$ to $n$:
    start $\mathbf{Fanout}(n-m)$
    idle for 3 steps
```

Algorithm 2 Subroutine **Fanout**(l)

```
if $l \neq n$, implement $\mathbf{PS}_l$                    # takes 1 step
implement $\mathbf{PRT}_{l-1:0}$                        # takes $l$ steps
if $l \neq n$, implement $\mathbf{S}(\mathrm{U}_{0,1}, \mathbf{O}_{l+1})$   # takes $l$ step
if $l = n$, $\mathbf{NOT}(\mathrm{U}_{0,1})$                    # takes $l$ steps
```

$$\mathscr{B}'_{l,j} = \left\{\mathbf{D}_{l',j'} \in \mathscr{A}_{l,\downarrow,j} \middle| j'_{l'} = 1\right\} \tag{4a}$$

$$\mathscr{B}_{l,j} = \mathscr{B}'_{l,j} \cup \{\mathbf{U}_{l,j}\}. \tag{4b}$$

which clarifies Eq. (3).

We then define $\psi_{l,j}$. Firstly, each amplitude may be represented as $\psi_j = a_j \angle \phi_j$, where a_j and ϕ_j are absolute value and argument of ψ_j respectively. we set $\phi_0 = 0$ without loss of generality. Let $\psi_{n,j} \equiv \psi_j$, we recursively define $\psi_{l,j} = e^{i\phi_{l+1,2j}} \sqrt{a^2_{l+1,2j} + a^2_{l+1,2j+1}}$.

We then introduce the gates required for this pahse. Let $\mathbf{CR}_{l,j}$ be a controlled-rotation with $\mathbf{U}_{l,j}$ and $\mathbf{D}_{l,j}$ as controlled and target qubits, which satisfies (see also Fig. 1(b))

$$\mathbf{CR}_{l,j} |1\rangle \otimes (\psi_{l,j}|0\rangle) = |1\rangle \otimes (\psi_{l,2j}|0\rangle + \psi_{l,2j+1}|1\rangle) \tag{5a}$$

$$\mathbf{CR}_{l,j} |0\rangle \otimes |0\rangle = |0\rangle \otimes |0\rangle. \tag{5b}$$

Let $\mathbf{PCR}_l = \prod_{j=0}^{2^l-1} \mathbf{CR}_{l,j}$ be the parallel controlled-rotation, which can be realized with single layer of quantum circuit. This parallel rotation is crucial for data encoding.

Another operation that is critical for our protocol is *routing*. Let $\mathbf{S}(a,b)$ be the swap gate between qubit a and b, it is the following transformation

$$|0\rangle_{\mathrm{rt}}\langle 0| \otimes \mathbf{S}(\mathrm{in}, \mathrm{lo}) + |1\rangle_{\mathrm{ctrl}}\langle 1| \otimes \mathbf{S}(\mathrm{in}, \mathrm{ro}) \tag{6}$$

If the routing qubit (rt) is at state $|0\rangle$, we swap the states of incident qubit (in) and left output qubit (lo); if the routing qubit is at state $|1\rangle$, we swap the states of input qubit and right output qubit (ro). We denote $\mathbf{RT}_{l,j}$ as the routing operation defined in Eq. (6) with $\mathbf{U}_{l,j}$, $\mathbf{D}_{l,j}$, $\mathbf{U}_{l,2j}$ and $\mathbf{U}_{l+1,2j+1}$ as the input, control, left output and right output qubits respectively (see also Fig. 1(b)). Note that $\mathbf{RT}_{l,j}$ for different j can be implemented in parallel. Accordingly, we define let $\mathbf{PRT}_l = \prod_{j=0}^{2^l-1} \mathbf{RT}_{l,j}$, this parallel routing can be implemented with constant circuit depth.

With elementary gates being explained, we are ready for discussing the transformation in Eq. (3). We first apply parallel controlled rotation $\mathbf{PCR}_l$. Except for qubit connected to the pointer (currently at $\mathbf{U}_{l,j}$), other qubits at layer $(l,\downarrow)$ are not activated. So it can be verified that $\mathbf{PCR}_l(\psi_{l,j}|\mathscr{B}_{l,j}\rangle) = \psi_{l+1,2j}|\{\mathscr{B}_{l,j}\}\rangle + \psi_{l+1,2j+1}|\mathscr{B}_{l,j} \cup \{\mathbf{D}_{l,j}\}\rangle$. Then, we move the pointer from the lth to the $(l+1)$th layer using parallel routing operation $\mathbf{PRT}_l$. Recall that $\mathbf{D}_{l,j}$ are controlled qubits of our routing operations. According to the property defined by Eq. (6), if $\mathbf{D}_{l,j}$ is not activated , the pointer moves to $\mathbf{D}_{l+1,2j}$, otherwise the pointer moves to $\mathbf{D}_{l+1,2j+1}$. Following the definition of $\mathscr{B}_{l,j}$, we have

$$\mathbf{PRT}_l|\mathscr{B}_{l,j}\rangle = |\mathscr{B}_{l+1,2j}\rangle \tag{7a}$$

$$\mathbf{PRT}_l|\mathscr{B}_{l,j} \cup \{\mathbf{D}_{l,j}\}\rangle = |\mathscr{B}_{l+1,2j+1}\rangle \tag{7b}$$

Combining with the recursive definition of $\psi_{l,j}$, we have $\mathbf{PCR}_l\mathbf{PRT}_l|\psi_l\rangle = |\psi_{l+1}\rangle$. Therefore, at the lth step, it suffices to implement $\mathbf{PCR}_l\mathbf{PRT}_l$ to realize the transformation in Eq. (3).

Fanout stage. In this stage, our goal is to prepare the output register to the quantum state in Eq. (1), while uncompute the QRAM. In other words, we perform the basis transformation $|\mathscr{B}_{n,j}\rangle|0\cdots 0\rangle_{\mathrm{out}} \to |\varnothing\rangle|j\rangle_{\mathrm{out}}$, where $|\cdot\rangle_{\mathrm{out}}$ is the quantum state of output register in binary representation, while the state of QRAM is still in activation representation. This transformation has been introduced in [24] for binary data, and subsequently been generalized to continuous data by adding an extra pointer [10]. For completeness of our presentation, we briefly introduce the fannout stage in below.

We define the shorthand $\mathbf{PRT}_{a:b} \equiv \mathbf{PRT}_a \cdots \mathbf{PRT}_{b+1}\mathbf{PRT}_b$ for some $b > a$. We first perform operation $\mathbf{PRT}_{1:n}$. The pointer then moved to the root of the QRAM, i.e. $\mathbf{PRT}_{1:n}|\mathscr{B}_{n,j}\rangle = |\mathscr{A}_{n,\downarrow,j} \cup \{\mathbf{U}_{0,1}\}\rangle$ for arbitrary j. So we can then apply $\mathbf{NOT}(\mathbf{U}_{0,1})$ (i.e. NOT gate at qubit $\mathbf{U}_{0,1}$) to uncompute the pointer. After this step, the basis $|\mathscr{B}_{n,j}\rangle|0\cdots 0\rangle_{\mathrm{out}}$ has been transferred to $|\mathscr{B}'_{n,j}\rangle|0\cdots 0\rangle_{\mathrm{out}}$.

We then define

$$|\Psi_{l,j}\rangle = |\mathscr{B}'_{l,j_{1:l}}\rangle \otimes |0\cdots 0 j_{l+1}\cdots j_n\rangle_{\mathrm{out}}. \tag{8}$$

The current basis and target basis corresponds to $l = n$ and $l = 0$ respectively. We now show how to perform the basis transformation $|\Psi_{l+1,j}\rangle \longrightarrow |\Psi_{l,j}\rangle$ iteratively. We define $\mathbf{PS}_l = \prod_{j=1}^{2^l} \mathbf{S}(\mathbf{U}_{l,j}, \mathbf{D}_{l,j})$ as the parallel swap gate applied between sublayers $(l,\uparrow)$ and $(l,\downarrow)$. By applying $\mathbf{PS}_l$ to $|\psi'_{n-l+1}\rangle$, activation at sublayer $(l,\downarrow)$ are transferred to sublayers $(l,\uparrow)$. We then implement routing $\mathbf{PRT}_{l-1:0}$, after which the sublayer $(l,\uparrow)$ is uncomputed, while the root of QRAM is prepared at state $|j_{n-l}\rangle$. Therefore, by further performing swap gate $\mathbf{S}(\mathbf{U}_{0,0}, \mathbf{O}_{l+1})$, we complete the transformation. Note that $\mathbf{S}(\mathbf{U}_{0,0}, \mathbf{O}_{l+1})$ is a non-local operation, which should be decomposed into totally $(l+1)$ steps of local swap gates applied at pairs of connected qubits. To conclude, let $\mathbf{Fanout}(l) \equiv \mathbf{S}(\mathbf{U}_{0,1}, \mathbf{O}_{l+1})\mathbf{PRT}_{l-1:0}\mathbf{PS}_l$ (see algorithm. 2), we have

$$\mathbf{Fanout}(l)|\Psi_{l+1}\rangle = |\Psi_l\rangle \tag{9}$$

for $0 \leqslant l \leqslant n-1$. Transformation $\mathbf{Fanout}(l)$ has circuit depth $O(l)$. If we naively implement Eq. (9) for different l sequentially, the total circuit depth is $O(n^2)$. Fortunately, $\mathbf{Fanout}(l-1)$ dose not need to start after $\mathbf{Fanout}(l)$ finish. If fact, after the third step of $\mathbf{Fanout}(l)$ (one swap gate and two parallel routing), state $|j_l\rangle$ has already been routed to layer $(l-2, \uparrow)$ (or into the output register for small l). If we now start $\mathbf{Fanout}(l-1)$, the subsequent operations in $\mathbf{Fanout}(l)$ and $\mathbf{Fanout}(l-1)$ with not affect each other. Therefore, our fanout stage works as follows (see also Algorithm. 1). We start the uncomputation of pointer (denoted as $\mathbf{Fanout}(n)$). Then, from $l = n-1$ to $l = 0$, we idle for 3 steps and then start the $\mathbf{Fanout}(l)$. In this way, the fanout process contains totally $(4n+3)$ steps, so the circuit depth of the fanout process is $O(n)$.

We now discuss the circuit complexity under single- and two-qubit elementary gate set. The circuit depth is of our approach is $O(n)$, and the total single- and two-qubit gate count is $O(N)$, both of which match the lower bound [7, 8]. The total qubit number is $O(N)$, which is the minimum requirement for achieving optimal circuit depth [8]. One may also care about the space-time allocation [13], i.e. sum of the individual duration that each qubit is active. The lth layer of QRAM contains totally 2^l qubits, and each qubit is active for time $O(n-l)$. So the total space-time allocation (STA) for QRAM is $\sum_{l=1}^{n} 2^l O(n-l) = O(N)$. Besides, the output register has STA $O(n) \times O(n) = O(n^2)$, so the total STA of our approach is $O(N)$, which is comparable to the best known results [13]. Therefore, in aspects other than practicality and robustness, our approach is also highly competitive.

Robustness. One of the crucial advantages of bucket-brigade architecture is the noise resiliency. It has been shown that the error of bucket-brigade QRAM scales only polylogarithmically with n for Eq. (2) with binary $|D_j\rangle$. Because we are using a similar hardware architecture, it is expected that our state preparation protocol also enjoys the similar the noise resiliency.

In Appendix. A, we show that this is the case. In particular, we consider the local depolarization model [30–32] that is standard in the noisy quantum circuit study: after each layer of the elementary single- and two-qubit gates, depolarization channel $(1-\varepsilon)\mathcal{I}+\varepsilon/3(\mathcal{X}+\mathcal{Y}+\mathcal{Z})$ is applied on *all* qubits with fixed ε, where $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ and $\mathcal{I}$ are single qubit Pauli X, Y, Z and I channels respectively. Under local Pauli noise, we show that the state preparation infidelity for Algorithm. 1 satisfies $1 - F \leqslant A\varepsilon n^3$ for some constant A. As a comparison, for a general quantum circuit with $O(2^n)$ elementary gates, the total infidelity scales exponentially with n. Our result also indicates that to achieve a constant state preparation fidelity, it suffices to suppress the gate error to $\varepsilon = O(n^{-3})$, which is polynomial.

The main idea of our proof about noise robustness is as follows. The noisy circuit can be decomposed into the linear combination of unitary evolutions, and each unitary evolution represents a specific space-time error configuration c. By a carful analysis on how error propagates between different branches of the QRAM, the final output state can be expressed as $|\tilde{\psi}(c)\rangle_{\text{out}} = \sum_{j\in g'(c)} \psi_j |f(c)\rangle_{\text{qram}} \otimes |j\rangle_{\text{out}} + |\text{garb}\rangle$. Where $g'(c)$ represents some *error-free* branches, that error will never propagate into it. $|\text{garb}\rangle$ is an unnormalized garbage state orthogonal to the first term. An important fact is that after tracing out QRAM part of $|\tilde{\psi}(c)\rangle_{\text{out}}$, the infidelity satisfies $1 - F(c) \leqslant \sum_{j\in g'(c)} |\psi_j|^2 \equiv \Lambda'(c)$. In sampling different error configuration c, we have $\mathbb{E}[\Lambda'(c)] \geqslant (1 - A\varepsilon n^3)$. The cubic infidelity scaling then follows from the concavity of fidelity.

3-qubit-per-node protocol. The circuit complexity of the method introduced above is optimal in terms of single- and two-qubit gates, its fault-tolerant implementation. However, the circuit complexity in terms of Clifford+T decomposition is not yet optimal. In Appendix. B, we show that the fault-tolerant performance can be further improved with our 3-qubit-per-node protocol. In this architecture, a middle sublayer is inserted between $(l, \uparrow)$ and $(l, \downarrow)$, while each qubit is still connect to at most 3 other qubits. The advantage of this revision that it enables us to use the *pre-rotation* [13] technique, i.e. rotations encoding amplitudes ψ_j are implemented prior to the routing operations. This allow us to simultaneously achieve the linear circuit depth $O(n + \log(1/\varepsilon))$, gate count number $O(N \log(1/\varepsilon))$ and STA $O(N \log(1/\varepsilon))$, in terms of Clifford+T decomposition. Moreover, the 3-qubit-per-node protocol can further improve the noise robustness. In this improved scheme, all routing operations should be controlled by extra pointer qubits in the middle sublayers. This revision can block the error propagation from bad branches to good branches. The infidelity scaling can be improved to

$$1 - F \leqslant A\varepsilon n^2. \tag{10}$$

This mechanism is also applicable for QRAM in Eq. (2). While the best-known infidelity scaling for two-level circuit-based QRAM is $1 - F \leqslant A\varepsilon n^3$ [24], we can improve the scaling from cubic to quadratic, which is of independent interest.

To conclude, we have introduced a bucket-brigade approach for general quantum state preparation. Our methods achieve exponential improvement on noise robustness. The simple tree architecture with minimal connectivity also ensures that it is easier to be implemented in practice compared to protocols based on the all-to-all connectivity. In terms of Clifford+T circuit complexity, to our best-knowledge, our approach is the first one that simultaneously achieve the linear circuit depth, gate count, and STA.

Acknowledgement. This work is supported by National Natural Science Foundation of China (No. 12405013).

[1] G.-L. Long and Y. Sun, Phys. Rev. A **64**, 014303 (2001).

[2] L. Grover and T. Rudolph, Preprint at https://arxiv.org/abs/quant-ph/0208112 (2002).

[3] M. Möttönen, J. J. Vartiainen, V. Bergholm, and M. M. Salomaa, Quantum. Inf. Comput. **5**, 467 (2005).

[4] M. Plesch and Č. Brukner, Phy. Rev. A **83**, 032302 (2011).

[5] G. H. Low, V. Kliuchnikov, and L. Schaeffer, Quantum **8**, 1375 (2024).

[6] Z. Zhang, Q. Wang, and M. Ying, Quantum **8**, 1228 (2024).

[7] X.-M. Zhang, M.-H. Yung, and X. Yuan, Phys. Rev. Res. **3**, 043200 (2021).

[8] X. Sun, G. Tian, S. Yang, P. Yuan, and S. Zhang, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (2023).

[9] G. Rosenthal, Preprint at https://arxiv.org/abs/2111.07992 (2021).

[10] X.-M. Zhang, T. Li, and X. Yuan, Phys. Rev. Lett. **129**, 230504 (2022).

[11] B. D. Clader, A. M. Dalzell, N. Stamatopoulos, G. Salton, M. Berta, and W. J. Zeng, IEEE Transactions on Quantum Engineering **3**, 1 (2022).

[12] P. Yuan and S. Zhang, Quantum **7**, 956 (2023).

[13] K. Gui, A. M. Dalzell, A. Achille, M. Suchara, and F. T. Chong, Quantum **8**, 1257 (2024).

[14] X.-M. Zhang and X. Yuan, npj Quantum Information **10**, 42 (2024).

[15] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, Phys. Rev. A **86**, 032324 (2012).

[16] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell, *et al.*, Nature **574**, 505 (2019).

[17] Y. Wu, W.-S. Bao, S. Cao, F. Chen, M.-C. Chen, X. Chen, T.-H. Chung, H. Deng, Y. Du, D. Fan, *et al.*, Physical review letters **127**, 180501 (2021).

[18] S. A. Moses, C. H. Baldwin, M. S. Allman, R. Ancona, L. Ascarrunz, C. Barnes, J. Bartolotta, B. Bjork, P. Blanchard, M. Bohn, *et al.*, Physical Review X **13**, 041052 (2023).

[19] D. Bluvstein, S. J. Evered, A. A. Geim, S. H. Li, H. Zhou, T. Manovitz, S. Ebadi, M. Cain, M. Kalinowski, D. Hangleiter, *et al.*, Nature **626**, 58 (2024).

[20] V. Giovannetti, S. Lloyd, and L. Maccone, Phys. Rev. Lett. **100**, 160501 (2008).

[21] V. Giovannetti, S. Lloyd, and L. Maccone, Phys. Rev. A **78**, 052310 (2008).

[22] F.-Y. Hong, Y. Xiang, Z.-Y. Zhu, L.-z. Jiang, and L.-n. Wu, Phys. Rev. A **86**, 010306 (2012).

[23] C. T. Hann, C.-L. Zou, Y. Zhang, Y. Chu, R. J. Schoelkopf, S. M. Girvin, and L. Jiang, Phys. Rev. Lett. **123**, 250501 (2019).

[24] C. T. Hann, G. Lee, S. Girvin, and L. Jiang, PRX Quantum **2**, 020311 (2021).

[25] S. Jaques and A. G. Rattew, Preprint at https://arxiv.org/abs/2305.10310 (2023).

[26] S. Arunachalam, V. Gheorghiu, T. Jochym-O?Connor, M. Mosca, and P. V. Srinivasan, New Journal of Physics **17**, 123010 (2015).

[27] K. C. Chen, W. Dai, C. Errando-Herranz, S. Lloyd, and D. Englund, PRX Quantum **2**, 030319 (2021).

[28] A. M. Childs and N. Wiebe, Preprint at https://arxiv.org/abs/1202.5822 (2012).

[29] G. H. Low and I. L. Chuang, Quantum **3**, 163 (2019).

[30] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, Nature Physics **14**, 595 (2018).

[31] S. Bravyi, D. Gosset, R. König, and M. Tomamichel, Nature Physics **16**, 1040 (2020).

[32] D. Aharonov, X. Gao, Z. Landau, Y. Liu, and U. Vazirani, in *Proceedings of the 55th Annual ACM Symposium on Theory of Computing* (2023) pp. 945–957.

[33] P. Selinger, Preprint at https://arxiv.org/abs/1212.6253 (2012).

Appendix A: Robustness analysis for 2-qubit-per-node protocol

1. noise model

As explained in the main text, state preparation protocol contains totally $O(n)$ layers of quantum circuit. We can abstractly expressed the quantum circuit as $\prod_{m=1}^{M} U_m|\psi_{\text{ini}}\rangle = |\psi_{\text{targ}}\rangle$, where U_m is the mth layer of single- and two-qubit gates. The specific form of U_m depends on how we decompose the operations (e.g. elementary routing and control rotation operations), but we typically have $M = O(n)$. In practice, we should deal with mixed state due to the existence of noise, so we also define the corresponding unitary channels as $\mathcal{U}_m[\cdot] = U_m[\cdot]U_m^{\dagger}$. Let

$$\mathcal{U} = \mathcal{U}_M \circ \cdots \circ \mathcal{U}_2 \circ \mathcal{U}_1, \tag{A1}$$

be the ideal evolution, we have $\mathcal{U}[\rho_{\text{ini}}] = \rho_{\text{end}}$, where $\rho_{\text{ini}} = |\psi_{\text{ini}}\rangle\langle\psi_{\text{ini}}|$, and $\rho_{\text{end}} = |\psi_{\text{end}}\rangle\langle\psi_{\text{end}}|$ are initial and ideal output state. Let $\rho_{\text{id}} = |\psi\rangle\langle\psi|$ be the target state of the quantum state preparation, we have $\rho_{\text{id}} = \text{Tr}_{\text{qram}}[\rho_{\text{end}}]$, where Tr_{qram} the partial trace over the QRAM.

We then introduce the local depolarization noise model. We define

$$\mathcal{E}_q = (1-\varepsilon)\mathcal{I} + \frac{1}{3}\varepsilon\left(\mathcal{X}_q + \mathcal{Y}_q + \mathcal{Z}_q\right) \tag{A2}$$

as the noisy quantum channel applied at qubit q, where $\varepsilon \in (0,1)$ is the error probability, $\mathcal{I}[\rho] = \rho$, $\mathcal{X}[\rho] = X_q\rho X_q$, $\mathcal{Y}[\rho] = Y_q\rho Y_q$, $\mathcal{Z}[\rho] = Z_q\rho Z_q$, are Pauli I, X, Y and Z channels applied at qubit q respectively. After the implementation of each layer of quantum circuit $\mathcal{U}_m$, $\mathcal{E}_q$ is applied at all qubits in the system. In other words, let $\mathcal{E} \equiv \prod_{q\in\mathscr{V}} \mathcal{E}_q$, where $\mathscr{V}$ is the set of all qubits in both QRAM and output register, the ideal channel $\mathcal{U}_m$ is replaced by the noisy channel $\tilde{\mathcal{U}}_m = \mathcal{E}\circ\mathcal{U}_m$. So the noisy quantum state preparation can be described by the following quantum channel

$$\tilde{\mathcal{U}} \equiv \tilde{\mathcal{U}}_M \circ \cdots \circ \tilde{\mathcal{U}}_2 \circ \tilde{\mathcal{U}}_1. \tag{A3}$$

2. Linear combination of unitary evolutions

We then show how to decompose $\tilde{\mathcal{U}}$ into the linear combination of unitary evolutions. We first rewrite $\mathcal{E}$ as the linear combination of all possible qubit distribution of error

$$\mathcal{E} \equiv \sum_{Q\in\text{Power}(\mathscr{V})} \mathcal{E}_Q \equiv \sum_{Q\in\text{Power}(\mathscr{V})} p_Q\mathcal{D}_Q, \tag{A4}$$

where Power($\mathscr{V}$) is the power of $\mathscr{V}$, i.e. all possible subset of all qubits. Moreover, $\mathcal{D}_q = \mathcal{X}_q + \mathcal{Y}_q + \mathcal{Z}_q$, represents the depolarization part of Eq. (A2), and $\mathcal{D}_Q = \prod_{q\in Q}\mathcal{D}_q$, $p_Q = (1-\varepsilon)^{|\mathscr{V}|-|Q|}\varepsilon^{|Q|}$. Here, $\mathcal{D}_Q$ represents that errors are applied at qubits in set Q while all qubits not in Q is free of errors. The probability distribution p_Q is normalized, and decreases with $|Q|$.

Then, let $\boldsymbol{Q} \equiv [Q_1,\cdots,Q_M]$ be a vector of qubit set for some $Q_m \in \text{Power}(\mathscr{V})$. $\boldsymbol{Q}$ describes a specific space-time configuration of the depolarization error. More specifically, we define

$$\tilde{\mathcal{U}}(\boldsymbol{Q}) = \mathcal{D}_{Q_M} \circ \mathcal{U}_M \circ \cdots \circ \mathcal{D}_{Q_2} \circ \mathcal{U}_2 \circ \mathcal{D}_{Q_1} \circ \mathcal{U}_1, \tag{A5}$$

and $p_{\boldsymbol{Q}} = \prod_{m=1}^{M} p_{Q_m}$. Let $\mathbb{Q} = \left\{[Q_1,\cdots,Q_M]\middle|Q_m \in \text{Power}(V) \text{ for all } 1 \leqslant m \leqslant M\right\}$ be all possible space-time configuration, we can rewrite $\tilde{\mathcal{U}}$ in Eq. (A3) as

$$\tilde{\mathcal{U}} = \sum_{\boldsymbol{Q}\in\mathbb{Q}} p_{\boldsymbol{Q}}\tilde{\mathcal{U}}(\boldsymbol{Q}). \tag{A6}$$

We further decompose each $\tilde{\mathcal{U}}(\boldsymbol{Q})$ into the linear combination of unitary evolutions. Recall that in Eq. (A25), each depolarization $\mathcal{D}_{Q_m}$ is the linear combination of three unitary channels. Let $\mathscr{P}_{Q_m} = \left\{\prod_{q\in Q_m}\mathcal{P}_q\middle|\mathcal{P}_q \in \{\mathcal{X}_q, \mathcal{Y}_q, \mathcal{Z}_q,\}\right\}$, we have

$$\mathcal{D}_{Q_m} = \frac{1}{\left|\mathscr{P}_{Q_m}\right|}\sum_{\mathcal{P}\in\mathscr{P}_{Q_m}} \mathcal{P}, \tag{A7}$$

where $\left|\mathscr{P}_{Q_m}\right| = 3^{|Q_m|}$. Let $[\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_M]$ be the polarization configuration of errors, we define all possible $[\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_M]$ under a space-time configuration $\boldsymbol{Q}$ as $\mathscr{P}_{\boldsymbol{Q}} \equiv \left\{[\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_M] \middle| \mathcal{P}_m \in \mathscr{P}_{Q_m}\right\}$. $\tilde{\mathcal{U}}(\boldsymbol{Q})$ can therefore be decomposed as

$$\tilde{\mathcal{U}}(\boldsymbol{Q}) = \frac{1}{\left|\mathscr{P}_{\boldsymbol{Q}}\right|} \sum_{c \in \mathscr{P}_{\boldsymbol{Q}}} \tilde{\mathcal{U}}(c), \tag{A8}$$

where c represents a specific space-time-polarization configuration of error, and

$$\tilde{\mathcal{U}}([\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_M]) \equiv \mathcal{P}_M \mathcal{U}_M \cdots \mathcal{P}_2 \mathcal{U}_2 \mathcal{P}_1 \mathcal{U}_1. \tag{A9}$$

Because each $\mathcal{P}_m$ is a unitary channel, Eq. (A9) is also a unitary channel. The total noisy evolution can then be decomposed as the linear combination of unitary evolutions as

$$\tilde{\mathcal{U}} = \sum_{\boldsymbol{Q} \in \mathscr{Q}} \sum_{c \in \mathscr{P}_{\boldsymbol{Q}}} p_c \tilde{\mathcal{U}}(c) \tag{A10}$$

for some $p_c = p_{\boldsymbol{Q}} / |\mathscr{P}_{\boldsymbol{Q}}|$. Let

$$\tilde{\rho}_{\text{out}}(c) = \text{Tr}_{\text{qram}}\left[\tilde{\mathcal{U}}(c)[\rho_{\text{ini}}]\right], \tag{A11}$$

be the final noisy output state is therefore $\tilde{\rho}_{\text{out}} = \sum_c p_c \tilde{\rho}_{\text{out}}(c)$. We denote $\text{Fid}(A, B)$ as the fidelity between two density matrix A and B, the total state preparation fidelity is just $F \equiv \text{Fid}(\rho_{\text{id}}, \tilde{\rho}_{\text{out}})$. Due to the concavity of fidelity, we have

$$\begin{aligned} F &\geqslant \sum_{\boldsymbol{Q} \in \mathscr{Q}} \sum_{c \in \mathscr{P}_{\boldsymbol{Q}}} p_c \text{Fid}[\rho_{\text{id}}, \tilde{\rho}_{\text{out}}(c)] \\ &= \mathbb{E}\left[\text{Fid}(\rho_{\text{id}}, \tilde{\rho}_{\text{out}}(c))\right]. \end{aligned} \tag{A12}$$

where $\mathbb{E}[\cdot]$ represents the expectation value with c sampled according to p_c. The remaining of this section is to study the unitary evolution under space-time-depolarization configuration c, and estimate Eq. (A12).

3. Definition of good branch

Before discussing the infidelity of $\tilde{\rho}_{\text{out}}(c)$, we give the definition of *good* branch and related terminologies that are useful. To begin with, we define the parent of each node as

$$\mathcal{P}[X] = \begin{cases} \mathbf{O}_{l+1} & X = \mathbf{O}_l \text{ for } 1 \leqslant l \leqslant n-1 \\ \mathbf{O}_0 & X = \mathbf{U}_{0,1} \\ \mathbf{D}_{(l-1, \lceil j/2 \rceil)} & X = \mathbf{U}_{(l,j)} \text{ for some } l \neq 0 \\ \mathbf{U}_{(l,j)} & X = \mathbf{D}_{(l,j)} \text{ for some } l \neq 0 \end{cases} \tag{A13}$$

We the define $\mathscr{A}_{l,j}$ as the all ancestors of qubit $\mathbf{U}_{l,j}$ as

$$\mathscr{A}_{l,j} = \{\mathcal{P}^{\circ t}[\mathbf{U}_{l,j}] \big| 1 \leqslant t \leqslant 3n\}. \tag{A14}$$

Let $\mathscr{A}_{l,j}^{(\text{neighbor})} \equiv \left\{b \notin \mathscr{A}_{l,j} \middle| \{a, b\} \in E \text{ and } a \in \mathscr{A}_{l,j}\right\}$ be the nearest-neighbor of $\mathscr{A}_{l,j}$, and

$$\hat{\mathscr{A}}_{l,j} = \mathscr{A}_{l,j} \cup \mathscr{A}_{l,j}^{(\text{neighbor})}. \tag{A15}$$

As will be demonstrated later, if $\boldsymbol{Q} \cap \hat{\mathscr{A}}_{n,j} = \varnothing$, the basis of the final output state with respect to label j is some how free of errors.

We consider a specific space-time-polarization configuration of error $c \in \mathscr{P}_{\boldsymbol{Q}}$ for some $\boldsymbol{Q} = [Q_1, Q_2, \cdots, Q_M]$. We define the set of survived qubits with respect to c as

$$\mathcal{S}_{\text{surv}}(c) \equiv \{q \in V \big| q \notin Q_m \text{ for all } 1 \leqslant m \leqslant M\}. \tag{A16}$$

If a qubit is in $\mathcal{S}_{\text{surv}}(c)$, it means that no error has been applied at it during the algorithm. We then introduce the set of all *good* branch at the lth spatial layer of QRAM as

$$g_l(c) = \left\{ j \middle| \mathcal{S}_{\mathrm{surv}}(c) \cup \hat{\mathscr{A}}_{l,j} = \varnothing \right\}. \tag{A17}$$

For a lighter notation, we also define

$$\hat{\mathscr{A}}_j \equiv \hat{\mathscr{A}}_{n,j}, \tag{A18}$$

and

$$g(c) \equiv g_n(c). \tag{A19}$$

It turns out that the infidelity is closely related to $g(c)$. In below, we discuss the evolution during fanin phase and fanout phase separately.

4. Fanin phase

We assume that before the lth step, the quantum state is in the form of

$$|\tilde{\psi}_{l-1}\rangle = E_{l-1} \sum_{j \in g_{l-1}(c)} \psi_{l-1,j} |\mathscr{B}_{l-1,j}\rangle + |\mathrm{garb}_{l-1}\rangle \tag{A20}$$

for some unitary E_{l-1} acting trivially in the good branch, and $|\mathrm{garb}_{l-1}\rangle$ orthogonal to the first term. For a lighter notation, in Eq. (A20), we have neglected the dependency on c, and set $|\tilde{\psi}_{l-1}\rangle \equiv |\tilde{\psi}_{l-1}(c)\rangle$, $E_{l-1} \equiv E_{l-1}(c)$ and $|\mathrm{garb}_{l-1}\rangle = |\mathrm{garb}_{l-1}(c)\rangle$.

For $l = 0$, Eq. (A20) holds because the initial state $|\mathbf{U}_{0,0}\rangle = |\mathscr{B}_{0,0}\rangle$ is assumed to be error-free. At the lth step, we denote the ideal evolution as $\prod_{j=0}^{2^{l-1}-1} U_{l-1,j}$, with $U_{l-1,j} = \mathbf{RT}_{l-1,j}\mathbf{CR}_{l-1,j}$. Note that $U_{l-1,j}$ for different j acts on different qubits and do not have overlap, so they commute with each other. Moreover, errors act trivially on qubits in good branches. So we can express the unitary at the lth step as

$$\tilde{U}_l = \prod_{j \in g_{l-1}(c)} U_{l,j} \prod_{j \notin g_{l-1}(c)} \tilde{U}_{l,j}. \tag{A21}$$

For $j \notin g_{l-1}(c)$, the unitary $\tilde{U}_{l,j}$ is the noisy implementation of $U_{l,j}$, which acts trivially at good branches. So the quantum state at the lth step satisfies

$$\begin{aligned}
|\tilde{\psi}_l\rangle =& \tilde{U}_l |\tilde{\psi}_{l-1}\rangle && \text{(A22)}\\
=& \tilde{U}_l E_{l-1} \sum_{j \in g_{l-1}} \psi_{l-1,j} |\mathscr{B}_{l-1,j}\rangle + \tilde{U}_l |\mathrm{garb}_{l-1}\rangle && \text{(A23)}\\
=& \left(\prod_{j \notin g_{l-1}(c)} \tilde{U}_{l,j} \right) E_{l-1} \sum_{j \in g_{l-1}(c)} \tilde{U}_{l,j} \psi_{l-1,j} |\mathscr{B}_{l-1,j}\rangle + \tilde{U}_l |\mathrm{garb}_{l-1}\rangle && \text{(A24)}\\
=& E_l \sum_{\lfloor j/2 \rfloor \in g_{l-1}(c)} \psi_{l,j} |\mathscr{B}_{l,j}\rangle + \tilde{U}_l |\mathrm{garb}_{l-1}\rangle && \text{(A25)}\\
=& E_l \sum_{j \in g_l(c)} \psi_{l,j} |\mathscr{B}_{l,j}\rangle + |\mathrm{garb}_l\rangle. && \text{(A26)}
\end{aligned}$$

In Eq. (A25), we have defined $E_l \equiv (\prod_{j \notin g_{l-1}(c)} \tilde{U}_{l,j}) E_{l-1}$; in Eq. (A26), we have defined

$$|\mathrm{garb}_l\rangle = E_l \sum_{j \in \{j' | \lfloor j'/2 \rfloor \in g_{l-1}(c) \& j' \notin g_l(c)\}} \psi_{l,j} |\mathscr{B}_{l,j}\rangle + U_l^{\mathrm{enc}} |\mathrm{garb}_{l-1}\rangle.$$

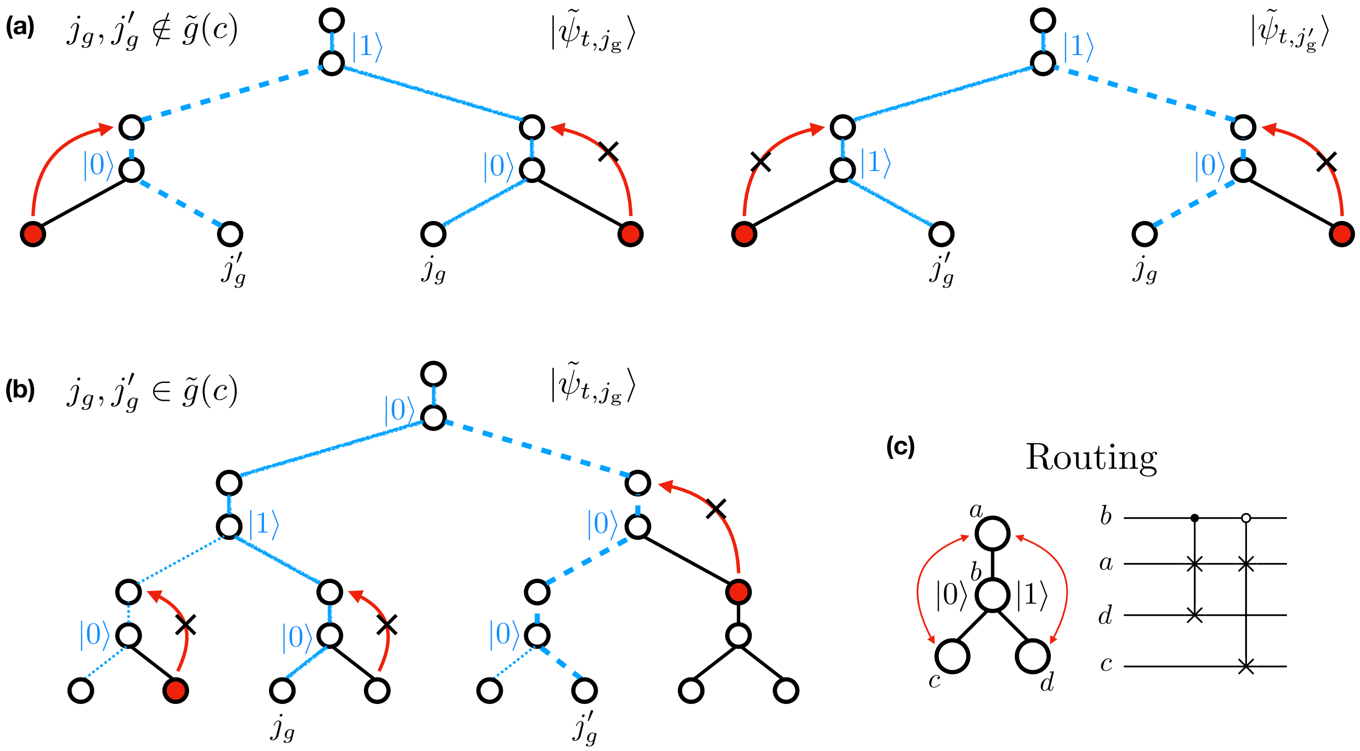


FIG. 2: (a) For $j_g, j'_g \in g(c)$ but $j_g, j'_g \notin \tilde{g}(c)$, errors may propagate upward into the good branch from the left hand side. In particular, at the left figure for basis $|\psi_{t,j_g}\rangle$, error propagate into the branch j'_b. (b) For basis $|\psi_{t,j_g}\rangle$, routing qubits at all good branches $j'_g \neq j_g$ are always at state $|0\rangle$. With this property, error will never propagate into branches in $\tilde{g}(c)$. (c) Sketch of the routing operations.

Accordingly, the final state of the encoding phase is in the form of

$$|\tilde{\psi}_n\rangle = \sum_{j\in g(c)} \psi_j E|\mathscr{B}_j\rangle + |\mathrm{garb}\rangle, \tag{A27}$$

where $E \equiv E_n$, $\mathscr{B}_j = \mathscr{B}_{n,j}$, and $|\mathrm{garb}\rangle \equiv |\mathrm{garb}_n\rangle$. Note that unitary E acts trivially at qubits in good branches, and $|\mathrm{garb}\rangle$ is orthogonal to the first term.

5. Fanout phase

We then study the fanout phase. Our discussion mainly follows the ideal in Sec.V of [24]. In the fanout phase, all operations (under a specific error configuration c) only transfer a computational basis to another computational basis, up to a phase. So we can always express the quantum state before the tth step as

$$|\tilde{\psi}'_t\rangle = \sum_{j\in g(c)} \psi_j |\psi'_{t,j}\rangle + |\mathrm{garb}'_t\rangle, \tag{A28}$$

where $|\psi'_{t,j}\rangle$ is some computational basis up to a phase, and $|\mathrm{garb}'_t\rangle$ is orthogonal to the first term. Similar to the encoding phase, the expression of states neglect the dependency on c. For $t = 0$, Eq. (A28) corresponds to $|\psi'_{0,j}\rangle = E|\mathscr{B}_j\rangle$ and $|\mathrm{garb}'_0\rangle = |\mathrm{garb}\rangle$.

At each step, we suppose a routing operation $\mathbf{RT}_{l',j'}$ acts nontrivially at a good branch $j \in g(c)$ (this also indicates that it is error-free). It can be verified that $\mathbf{RT}_{l',j'}$ only swaps $\mathbf{U}_{l,j}$ and one of the children qubit of $\mathbf{D}_{l,j}$ that is within the branch j, while another child qubit of $\mathbf{D}_{l,j}$ remains unchanged (see Fig. 2(a)). Moreover, swap gates in the good branch is also error-free. Applying this argument on each elementary routing and swapping operations, the fanout phase performs the basis transformation $E|\mathscr{B}_j\rangle \to |f_j\rangle_{\mathrm{qram}} \otimes |j\rangle_{\mathrm{out}}$ for all $j \in g(c)$. Here, $|f_j\rangle_{\mathrm{qram}} \equiv |f_j(c)\rangle_{\mathrm{qram}}$ is some quantum state of the QRAM. Accordingly, the final state of the fanout phase, $|\tilde{\psi}'\rangle \equiv |\tilde{\psi}'_{t_{\rm end}}\rangle$ with $t_{\rm end}$ the last step, is in the form of

$$|\tilde{\psi}'\rangle = \sum_{j\in g(c)} \psi_j |f_j\rangle_{\rm qram} \otimes |j\rangle_{\rm out} + |{\rm garb}'\rangle, \tag{A29}$$

for some $|{\rm garb}'\rangle$ orthogonal to the first term.

However, Eq. (B22) is still not sufficient for us to estimate the infidelity. For $j, j' \in g(c)$, we in general have $|f_j\rangle_{\rm qram} \neq |f_{j'}\rangle_{\rm qram}$ when $j \neq j'$. After tracing out the QRAM part of Eq. (B22), the coherence between basis in $g(c)$ may be destroyed. To understand why $|f_j\rangle_{\rm qram} \neq |f_{j'}\rangle_{\rm qram}$ (see also Sec.V of [24]), we should analyze how error terms propagate from different branches. We first consider basis $|\psi_{t,j_{\rm g}}\rangle$ with $j_{\rm g} \in g(\boldsymbol{Q})$. As shown in left subfigure of Fig. 2 (a), suppose an error occurs at the bad branch $j_{\rm b} \notin g(\boldsymbol{Q})$, it may propagate into another good branch $j'_{\rm g} \in g(\boldsymbol{Q})$ ($j'_{\rm g} \neq j_{\rm g}$) through a sequence of routing operations (Fig. 2(c)). On the other hand, if we consider the basis $|\psi_{t,j'_{\rm g}}\rangle$ instead, errors will never propagate into $j'_{\rm g}$ (see also right subfigure of Fig. 2). So in general the final state of the QRAM is different for different basis in $g(c)$.

Fortunately, we can identify a large portion of basis in Eq. (B22), such that errors will still *not* propagate from bad branches to any of the good branches. For these j, the final states of QRAM, $|f_j\rangle_{\rm qram}$, is independent of j. To begin with, we notices that in every good branches, error only propagate into it from the right hand side (instead of left hand side). The reason is as follows. Let us consider a basis $|\psi_{t,j_g}\rangle$ with $j_g \in g(c)$. For branch j_g, errors will not propagate into it as mentioned previously. For another good branch $j'_g \neq j_g$, all routing qubits in it (those in lower sublayers) is at the default state $|0\rangle$. Therefore, swap is only performed between its parent and its right child.

With the argument above, we suppose $k_1k_2\cdots k_l0\cdots0$ is a good branch, then no errors will ever propagate upward through $\mathbf{RT}_{l-1,k}$ (with $k = k_1k_2\cdots k_l$). Therefore, for index j, if we have $j_1j_2\cdots j_l0\cdots0 \in g(c)$ for all $0 \leqslant l \leqslant n-1$, no error will propagate into the branch j from any site. Accordingly, we can define the set of all *error-free* branches

$$g'(c) \equiv \{j \in g(c) | \tilde{j}_l \in g(c) \text{ for } 0 \leqslant l \leqslant n-1\} \tag{A30}$$

where

$$\tilde{j}_l \equiv j_1j_2\cdots j_l\underbrace{0\cdots0}_{n-l}. \tag{A31}$$

For the basis $|f_j\rangle_{\rm qram} \otimes |j\rangle$ of the final state, if $j \in g'(c)$, errors are only applied at good branches. So for all $j \in g'(c)$, their QRAM part is identical, i.e. $|f_j\rangle = |f\rangle$ for some quantum state $|f\rangle$. Then, Eq. (B22) can be rewritten as

$$|\tilde{\psi}'\rangle = \sum_{j\in g'(c)} \psi_j |f\rangle_{\rm qram} \otimes |j\rangle_{\rm out} + |\widetilde{{\rm garb}}'\rangle. \tag{A32}$$

Note that $|f\rangle$ is independent of j, but still depends on c.

6. State preparation infidelity

In Sec. A 5, the final output state is $|\tilde{\psi}'\rangle$. Comparing Eq. (A32) to Eq. (A11), we have

$$\tilde{\rho}_{\rm out}(c) = {\rm Tr}_{\rm qram}\left[|\tilde{\psi}'\rangle\langle\tilde{\psi}'|\right]. \tag{A33}$$

We define $|\psi'\rangle \equiv \sum_{j=1}^{N-1} \psi_j |f\rangle_{\rm qram} \otimes |j\rangle_{\rm out}$. The fidelity between $|\psi'\rangle$ and Eq. (A32) is

$${\rm Fid}\left(|\psi'\rangle\langle\psi'|, |\tilde{\psi}'\rangle\langle\tilde{\psi}'|\right) = \sum_{j\in g'(c)} |\psi_j|^2 \equiv \Lambda'(c). \tag{A34}$$

Here, $\Lambda'(c)$ highlights that it is depends on c. Because fidelity is non-decreasing under partial trace, we have

$${\rm Fid}\left({\rm Tr}_{\rm qram}\left[|\psi'\rangle\langle\psi'|\right], \tilde{\rho}(c)\right) \geqslant \Lambda'(c). \tag{A35}$$

Moreover, it can be verified that ${\rm Tr}_{\rm qram}\left[|\psi'\rangle\langle\psi'|\right] = \rho_{\rm id}$. So

$${\rm Fid}\left(\rho_{\rm id}, \tilde{\rho}(c)\right) \geqslant \Lambda'(c). \tag{A36}$$

Combining with Eq. (A12), the total state preparation infidelity satisifes

$$F \geqslant \mathbb{E}[\Lambda'(c)], \tag{A37}$$

where $\mathbb{E}[\Lambda'(c)]$ represents the expectation value of $\Lambda(c)$ when sampling c according to p_c.

We now estimate $\mathbb{E}[\Lambda'(c)]$. Let $\mathcal{J} = \{0, 1, \cdots, N-1\}$ be all indexes, and Power$(\mathcal{J})$ be the set of all subset of $\mathcal{J}$. By definition, we have

$$\mathbb{E}[\Lambda'(c)] = \sum_{J \in \text{Power}(\mathcal{J})} \sum_{j\in J} |\psi_j|^2 \times \Pr[J_1 \in g'(c)]\Pr[J_2 \in g'(c)|J_1 \in g'(c)]\Pr[J_3 \in g'(c)|J_1, \mathcal{J}_2 \in g'(c)] \cdots . \tag{A38}$$

In Eq. (A38), $J_1, J_2, \cdots$ are elements of J arranged in arbitrary order. Note that different branches may have overlap, and we always have

$$\Pr[J_2 \in g'(c)|J_1 \in g'(c)] \geqslant \Pr[J_2 \in g'(c)], \tag{A39}$$

$$\Pr[J_3 \in g'(c)|J_1, \mathcal{J}_2 \in g'(c)] \geqslant \Pr[J_3 \in g'(c)], \tag{A40}$$

and so on. Therefore, we have

$$\mathbb{E}[\Lambda'(c)] \geqslant \sum_{J \in \text{Power}(\mathcal{J})} \sum_{j\in J} |\psi_j|^2 \times \Pr[J_1 \in g'(c)][J_2 \in g'(c)][J_3 \in g'(c)] \cdots \tag{A41}$$

$$= \sum_{j=0}^{N-1} |\psi_j|^2 \Pr[j \in g'(c)] \tag{A42}$$

$$= \Pr[j \in g'(c)] \tag{A43}$$

Eq. (A42) is because the right hand side of Eq. (A41) corresponds to a summation of multiple variables sampled independently. Eq. (A43) is because of the normalization of ψ_j, and the probability is independent of j. By definition in Eq. (A30), j is an error-free branch in $g'(c)$, if and only if all qubits in $\tilde{j}_l$ (for all $0 \leqslant l \leqslant n-1$) are free of error at all time. There are at most $O(n^2)$ of these qubits. For each individual qubit, the probability that it is error free at all time is $(1-\varepsilon)^{O(n)}$, because the algorithm has totally $O(n)$ steps. Therefore, with probability $((1-\varepsilon)^{O(n)})^{O(n^2)} = (1-\varepsilon)^{O(n^3)}$, j is a good branch. By Bernoulli inequality, we have $\Pr[j \in g'(c)] \geqslant 1 - A\varepsilon n^3$ for some constant A. So we have

$$\mathbb{E}[\Lambda'(c)] \geqslant (1 - A\varepsilon n^3). \tag{A44}$$

Combining with Eq. (A37), we have

$$1 - F \leqslant A\varepsilon n^3. \tag{A45}$$

Appendix B: 3-qubit-per-node protocol

1. Hardware architecture and basic operations

In our 3-qubit-per-node protocol, each layer contains 3 sublayers. The upper, middle, and lower sublayers of the lth layer are denoted as $(l, \uparrow)$, $(l, \bullet)$, and $(l, \downarrow)$ respectively. Each sublayer contain 2^l qubits, each denoted as $\mathbf{U}_{l,j}$, $\mathbf{M}_{l,j}$, $\mathbf{U}_{l,j}$ respectively with $0 \leqslant j \leqslant 2^l - 1$. The children of $\mathbf{U}_{l,j}$, $\mathbf{M}_{l,j}$ and $\mathbf{D}_{l,j}$ are $\{\mathbf{M}_{l,j}\}$, $\{\mathbf{D}_{l,j}\}$, and $\{\mathbf{U}_{l+1,2j}, \mathbf{U}_{l+1,2j+1}\}$ respectively. Moreover, the output register is identical to the 2-qubit-per-node protocol. The hardware architecture contains more qubits (totally $6N-3$ qubits), but each qubit is still connected to at most 3 other qubits.

We then introduce some basic operations. Let $r_{l,j} \equiv \begin{pmatrix} \psi_{l+1,2j}/\psi_{l,j} & * \\ \psi_{l+1,2j+1}/\psi_{l,j} & \diamond \end{pmatrix}$, where $*$ and $\diamond$ are some complex values that make $r_{l,j}$ be a unitary. We have the following basic operations.

- $\mathbf{R}_{l,j}$: rotation $r_{l,j}$ applied at qubit $\mathbf{D}_{l,j}$
- $\overline{\mathbf{CR}}_{l,j}$ controlled rotation $|0\rangle\langle 0| \otimes r_{l,j}^{\dagger} + |1\rangle\langle 1| \otimes \mathbb{I}$ with $\mathbf{M}_{l,j}$ and $\mathbf{D}_{l,j}$ as controlled and target qubits
- $\mathbf{CNOT}_{l,j}$: CNOT gate with $\mathbf{U}_{l,j}$ and $\mathbf{M}_{l,j}$ the control and target qubits

- $\mathbf{CRT}_{l,j}$: Five-qubit-gate

$$|0\rangle_{\mathbf{M}_{l,j}}\langle 0| \otimes \mathbb{I} + |1\rangle_{\mathbf{M}_{l,j}}\langle 1| \otimes |0\rangle_{\mathbf{D}_{l,j}}\langle 0| \otimes \mathbf{S}(\mathbf{U}_{l,j}, \mathbf{U}_{l+1,2j}) + |1\rangle_{\mathbf{M}_{l,j}}\langle 1| \otimes |1\rangle\langle 1| \otimes \mathbf{S}(\mathbf{U}_{l,j}, \mathbf{U}_{l+1,2j+1})$$

- $\mathbf{S}_{l,j}^{(\uparrow,\bullet)}$ and $\mathbf{S}_{l,j}^{(\uparrow,\downarrow)}$: swap gate between $\mathbf{U}_{l,j}, \mathbf{M}_{l,j}$, and swap gate between $\mathbf{U}_{l,j}, \mathbf{D}_{l,j}$

Accordingly, we define the following parallel operations

$$\mathbf{PR} \equiv \sum_{l=0}^{n-1}\sum_{j=0}^{2^l-1} \mathbf{R}_{l,j}, \quad \mathbf{P\overline{CR}} \equiv \sum_{l=0}^{n-1}\sum_{j=0}^{2^l-1} \mathbf{R}_{l,j} \tag{B1}$$

that applies at all l, j, and

$$\mathbf{PCNOT}_l \equiv \sum_{j=0}^{2^l-1} \mathbf{R}_{l,j}, \quad \mathbf{PCRT}_l \equiv \sum_{j=0}^{2^l-1} \mathbf{R}_{l,j}, \quad \mathbf{PS}_l^{(\uparrow,\bullet)} \equiv \sum_{j=0}^{2^l-1} \mathbf{S}_{l,j}^{(\uparrow,\bullet)}, \quad \mathbf{PS}_l^{(\uparrow,\downarrow)} \equiv \sum_{j=0}^{2^l-1} \mathbf{S}_{l,j}^{(\uparrow,\downarrow)} \tag{B2}$$

that act on a specific layer $0 \leqslant l \leqslant n$. All parallel operations above can be implemented with $O(1)$ layer of single- and two-qubit gates.

2. Fanin phase

The pseudo code of our algorithm is illustrated in Algorithm. 3 and Algorithm. 4. Our method is inspired by the pre-rotation technique in [13], which encodes angles $\{\psi_{l,j}\}$ before the controlled routing. The advantage is that pre-rotation can push the Clifford+T depth to a linear scaling. For clarity, we discuss the single- and two-qubit decomposition in this section, while the Clifford+T decomposition will be introduced in Sec. B 5.

Let $\mathscr{D} = \{\mathbf{D}_{l,j} | 0 \leqslant l \leqslant n-1, 0 \leqslant j \leqslant 2^l - 1\}$ be the set of all qubits in the lower sublayers. We first implement parallel rotation $\mathbf{PR}$, and $\mathscr{D}$ is prepared as

$$|\theta\rangle_{\mathscr{D}} \equiv \bigotimes_{\mathbf{D}_{l,j} \in \mathscr{D}} \left(\psi_{l+1,2j}/\psi_{l,j}|0\rangle_{\mathbf{D}_{l,j}} + \psi_{l+1,2j+1}/\psi_{l,j}|1\rangle_{\mathbf{D}_{l,j}}\right). \tag{B3}$$

Qubits in $\mathscr{D}$ serves as the routing qubits of our subsequent controlled-routing operations.

Let $\mathscr{A}_{l,\downarrow,j}$be all ancestors of $\mathbf{D}_{l,j}$ at lower sublayers, and we further define $\mathscr{D}_{l,j} = \mathscr{D} - \mathscr{A}_{l,\downarrow,j}$. $\mathscr{D}_{l,j}$ represents all routing qubits in $\mathscr{D}$ that are irrelevant to the operation $\mathbf{CRT}_{l,j}$ during our fanin process. We also define

$$|\theta\rangle_{\mathscr{D}_{l,j}} \equiv \bigotimes_{\mathbf{D}_{l',j'} \in \mathscr{D}_{l,j}} \left(\psi_{l'+1,2j'}/\psi_{l',j'}|0\rangle_{\mathbf{D}_{l',j'}} + \psi_{l'+1,2j'+1}/\psi_{l',j'}|1\rangle_{\mathbf{D}_{l',j'}}\right) \tag{B4}$$

as the quantum state of the subsystem $|\theta\rangle_{\mathscr{D}_{l,j}}$ for Eq. (B3). Let $\mathscr{M}_{l,j} \equiv \{\mathbf{M}_{0,0}\} \cup \left\{\mathbf{M}_{l',j_{1:l'}} \middle| 1 \leqslant l' \leqslant l-1\right\}$, and $\mathscr{C}_{l,j} \equiv \mathscr{M}_{l,j} \cup \mathscr{B}_{l,j}$ (with $\mathscr{B}_{l,j}$ defined in Eq. (4)), we will then iteratively perform the transformation

$$|\psi_l\rangle \rightarrow |\psi_{l+1}\rangle, \tag{B5}$$

where

$$|\psi_l\rangle = \sum_{j=0}^{2^l-1} \psi_{l,j} |\theta\rangle_{\mathscr{D}_{l,j}} \otimes |\mathscr{C}_{l,j}\rangle_{\mathscr{V}-\mathscr{D}_{l,j}}. \tag{B6}$$

Note that $|\psi_0\rangle = |\theta\rangle_{\mathscr{D}} \otimes |\{\mathbf{D}_{0,0}\}\rangle$. In Eq. (B6), quantum state of qubit set $\mathscr{D}_{l,j}$ and $\mathscr{V} - \mathscr{D}_{l,j}$ (all qubits not in $\mathscr{D}_{l,j}$) are expressed in the form of computational basis representation and activation representation, respectively.

By implementing parallel CNOT gates, we have $\mathbf{PCNOT}_l|\mathscr{C}_{l,j}\rangle_{\mathscr{V}-\mathscr{D}_{l,j}} = |\mathscr{C}_{l,j} \cup \{\mathbf{M}_{l,j}\}\rangle_{\mathscr{V}-\mathscr{D}_{l,j}}$, and hence

$$\mathbf{PCNOT}_l|\psi_l\rangle = \sum_{j=0}^{2^l-1} \psi_{l,j}|\theta\rangle_{\mathscr{D}_{l,j}} \otimes |\mathscr{C}_{l,j} \cup \{\mathbf{M}_{l,j}\}\rangle_{\mathscr{V}-\mathscr{D}_{l,j}}. \tag{B7}$$

Then, if we apply $\mathbf{CRT}_{l,j}$ on Eq. (B7), only basis with label j will be changed. This is because the routing is controlled on $\mathbf{M}_{l,j}$. We notice that the basis with label j can be rewritten as

$$\psi_{l,j}\left(|\theta\rangle_{\mathscr{D}_{l,j}} \otimes |\mathscr{C}_{l,j} \cup \{\mathbf{M}_{l,j}\}\rangle_{\mathscr{V}-\mathscr{D}_{l,j}}\right) \tag{B8}$$

$$=\psi_{l,j}|\theta\rangle_{\mathscr{D}_{l,j}-\{\mathbf{D}_{l,j}\}} \otimes \left(\psi_{l+1,2j}/\psi_{l,j}|\mathscr{C}_{l,j} \cup \{\mathbf{M}_{l,j}\}\rangle_{\mathscr{V}-\mathscr{D}_{l,j}-\{\mathbf{D}_{l,j}\}} + \psi_{l+1,2j+1}/\psi_{l,j}|\mathscr{C}_{l,j} \cup \{\mathbf{M}_{l,j}, \mathbf{D}_{l,j}\}\rangle_{\mathscr{V}-\mathscr{D}_{l,j}-\{\mathbf{D}_{l,j}\}}\right) \tag{B9}$$

$$=\psi_{l+1,2j}|\theta\rangle_{\mathscr{D}_{l+1,2j}} \otimes |\mathscr{C}_{l,j} \cup \{\mathbf{M}_{l,j}\}\rangle_{\mathscr{V}-\mathscr{D}_{l+1,2j}} + \psi_{l+1,2j+1}|\varphi\rangle_{\mathscr{D}_{l+1,2j+1}} \otimes |\mathscr{C}_{l,j} \cup \{\mathbf{M}_{l,j}, \mathbf{D}_{l,j}\}\rangle_{\mathscr{V}-\mathscr{D}_{l+1,2j+1}}. \tag{B10}$$

It can be verified that

$$\mathbf{CRT}_{l,j}|\mathscr{C}_{l,j} \cup \{\mathbf{M}_{l,j}\}\rangle_{\mathscr{V}-\mathscr{D}_{l+1,2j}} = |\mathscr{C}_{l+1,2j}\rangle_{\mathscr{V}-\mathscr{D}_{l+1,2j}} \tag{B11}$$

$$\mathbf{CRT}_{l,j}|\mathscr{C}_{l,j} \cup \{\mathbf{M}_{l,j}, \mathbf{U}_{l,j}\}\rangle_{\mathscr{V}-\mathscr{D}_{l+1,2j+1}} = |\mathscr{C}_{l+1,2j}\rangle_{\mathscr{V}-\mathscr{D}_{l+1,2j+1}}. \tag{B12}$$

Therefore, we have

$$\begin{aligned}\mathbf{PCRT}_l\mathbf{PCNOT}_l|\psi_l\rangle &= \sum_{j=0}^{2^l-1} \psi_{l+1,2j}|\theta\rangle_{\mathscr{D}_{l+1,2j}} \otimes |\mathscr{C}_{l+1,2j}\rangle_{\mathscr{V}-\mathscr{D}_{l+1,2j}} + \psi_{l+1,2j+1}|\theta\rangle_{\mathscr{D}_{l+1,2j+1}} \otimes |\mathscr{C}_{l+1,2j+1}\rangle_{\mathscr{V}-\mathscr{D}_{l+1,2j+1}}\\ &= \sum_{j=0}^{2^{l+1}-1} \psi_{l,j}|\theta\rangle_{\mathscr{D}_{l,j}} \otimes |\mathscr{C}_{l,j}\rangle_{\mathscr{V}-\mathscr{D}_{l,j}}\\ &=|\psi_{l+1}\rangle.\end{aligned} \tag{B13}$$

Applying $\mathbf{PCRT}_l\mathbf{PCNOT}_l$ iteratively from $l = 0$ to $l = n - 1$, we obtain

$$|\psi_n\rangle = \sum_{j=0}^{N-1} \psi_j|\theta\rangle_{\mathscr{D}_{n,j}} \otimes |\mathscr{C}_{n,j}\rangle_{\mathscr{V}-\mathscr{D}_{n,j}}. \tag{B14}$$

In the last step, we perform $\mathbf{P\overline{CR}}$, which performs of $r^\dagger_{l,j}$ on $\mathbf{D}_{l,j}$ conditioned on $\mathbf{M}_{l,j}$ not activated. For basis with label j, at the middle sublayers, only qubits $\mathbf{M}_{n-1,j}$, $\mathbf{M}_{n-2,j_{1:n-1}}$, $\cdots$, $\mathbf{M}_{0,j_1}$ are activated. These qubits are not in $\mathscr{D}_{n,j}$, so $|\theta\rangle_{\mathscr{D}_{l,j}}$ are uncomputed, and the final state is

$$|\psi\rangle = \mathbf{P\overline{CR}}|\psi_n\rangle = \sum_{j=0}^{N-1} \psi_j|\mathscr{C}_{n,j}\rangle_{\mathscr{V}}. \tag{B15}$$

Eq. (B15) is similar to the one for 2-qubit-per-node protocol. The only difference is that for basis j, all ancestors of $\mathbf{U}_{n,j}$ in sublayers $(l, \bullet)$ are activated. In the next section, with a mild modification of the fanout phase, we can uncompute the QRAM while obtain the target state in the output register.

3. Fanout phase

Let

$$\mathscr{C}'_{l,j} = \mathscr{M}_{l,j} \cup \mathscr{B}'_{l,j}, \tag{B16}$$

with $\mathscr{B}'_{l,j}$ defined in Eq. (4), it can be verified that

$$\mathbf{NOT}(\mathbf{U}_{0,0})\mathbf{PCRT}_1\mathbf{PCRT}_2\cdots\mathbf{PCRT}_{n-1}|\mathscr{C}_{n,j}\rangle = |\mathscr{C}'_{n,j}\rangle. \tag{B17}$$

In other words, performing parallel controlled routing from $l = n-1$ to $l = 0$ transfers the excitation at layer $(n, \uparrow)$ to $\mathbf{U}_{0,0}$, which can be uncomputed by an extra not gate. Our strategy is to perform the following transformation

$$|\mathscr{C}'_{l,j_{1:l}}\rangle \otimes |0\cdots 0 j_{l+1}\cdots j_n\rangle_{\text{out}} \longrightarrow |\mathscr{C}'_{l,j_{1:l-1}}\rangle \otimes |0\cdots 0 j_l\cdots j_n\rangle_{\text{out}}. \tag{B18}$$

iteratively. For basis $|\mathscr{C}'_{l,j_{1:l}}\rangle$, we can also deterministically route the activation at layer $(l, \bullet)$ to $\mathbf{U}_{0,0}$, and uncompute it with a NOT gate, i.e.

$$\mathbf{NOT}(\mathbf{U}_{0,0})\mathbf{PCRT}_1\mathbf{PCRT}_2\cdots\mathbf{PCRT}_{l-1}\mathbf{PS}_l^{(\uparrow,\bullet)}|\mathscr{C}'_{l,j_{1:l}}\rangle = |\mathscr{C}'_{l,j_{1:l}} - \{\mathbf{M}_{l,j_{1:l}}\}\rangle. \tag{B19}$$

Moreover, in analogy to the 2-qubit-per-node protocol, we can route the state $|j_l\rangle$ from layer $(l, \downarrow)$ to qubit $\mathbf{O}_l$ in the output register.

$$\begin{aligned}&\mathbf{S}(\mathbf{U}_{0,0}, \mathbf{O}_l)\mathbf{PCRT}_1\mathbf{PCRT}_2\cdots\mathbf{PCRT}_{l-1}\mathbf{PS}_l^{(\uparrow,\downarrow)}|\mathscr{C}'_{l,j_{1:l}} - \{\mathbf{M}_{l,j_{1:l}}\}\rangle \otimes |0\cdots 0 j_{l+1}\cdots j_n\rangle_{\text{out}}\\ =&|\mathscr{C}'_{l,j_{1:l-1}}\rangle \otimes |0\cdots 0 j_l\cdots j_n\rangle_{\text{out}}.\end{aligned} \tag{B20}$$

We can start the operation in Eq. (B20) after the operation in Eq. (B19) has finished the $\mathbf{PCRT}_{l-2}$, and two operations will not affect each other. With an abuse of notation, we also define this process as $\mathbf{Fanout}(l-1)$ (for $1 \leqslant l \leqslant n$), which performs the transformation claimed in Eq. (B18). We also define $\mathbf{Fanout}(n)$ as the process corresponding to Eq. (B17). By implementing $\mathbf{Fanout}(n), \mathbf{Fanout}(n-1), \cdots, \mathbf{Fanout}(0)$ iteratively, we can uncompute the QRAM, while prepare the target state at output register. Similar to the 2-site-per-node protocol, while implementing $\mathbf{Fanout}(l)$ sequentially is time costly, we can start another before one finished. More specifically, we can start $\mathbf{Fanout}(l)$, idle for 5 steps, and then start $\mathbf{Fanout}(l-1)$. In this way, operations $\mathbf{Fanout}(l)$ and $\mathbf{Fanout}(l-1)$ will not affect each other, and the total runtimes is $O(n)$.

Algorithm 3 Quantum state preparation. Initial state $|\mathbf{U}_{0,1}\rangle$, amplitudes $\{\psi_{l,j}\}$ as input

```
implement $\mathbf{PR}$
for $l = 0, \cdots, n-1$:
    implement $\mathbf{PCNOT}_l$
    implement $\mathbf{PCRT}_l$
implement $\mathbf{P}\overline{\mathbf{CR}}$
for $m = 0$ to $n$:
    start $\mathbf{Fanout}(n-m)$
    idle for 6 steps
```

Algorithm 4 Subroutine $\mathbf{Fanout}(l)$ for 3-qubit-per-node scheme

```
if $l \neq n$:
    for $l' = 1$ to $l' = l$
        implement $\mathbf{PRT}_{l-l'}$                        # takes 1 steps
else if $l \neq n$:
    implement $\mathbf{PS}_l^{(\uparrow,\bullet)}$                 # takes 1 step
    implement $\mathbf{PRT}_{l-2}\mathbf{PRT}_{l-1}$              # takes 2 steps
    implement $\mathbf{PS}_l^{(\uparrow,\downarrow)}$                 # takes 2 step
    for $l' = 1$ to $l' = l-2$
        implement $\mathbf{PRT}_{l-l'}\mathbf{PRT}_{l-l'-2}$          # takes 1 steps
    $\mathbf{NOT}(\mathbf{U}_{0,1})$                          # takes 1 steps
    implement $\mathbf{PRT}_0\mathbf{PRT}_1$                  # takes 2 steps
    $\mathbf{S}(\mathbf{U}_{0,1}, \mathbf{O}_l)$                        # takes $l$ steps
```

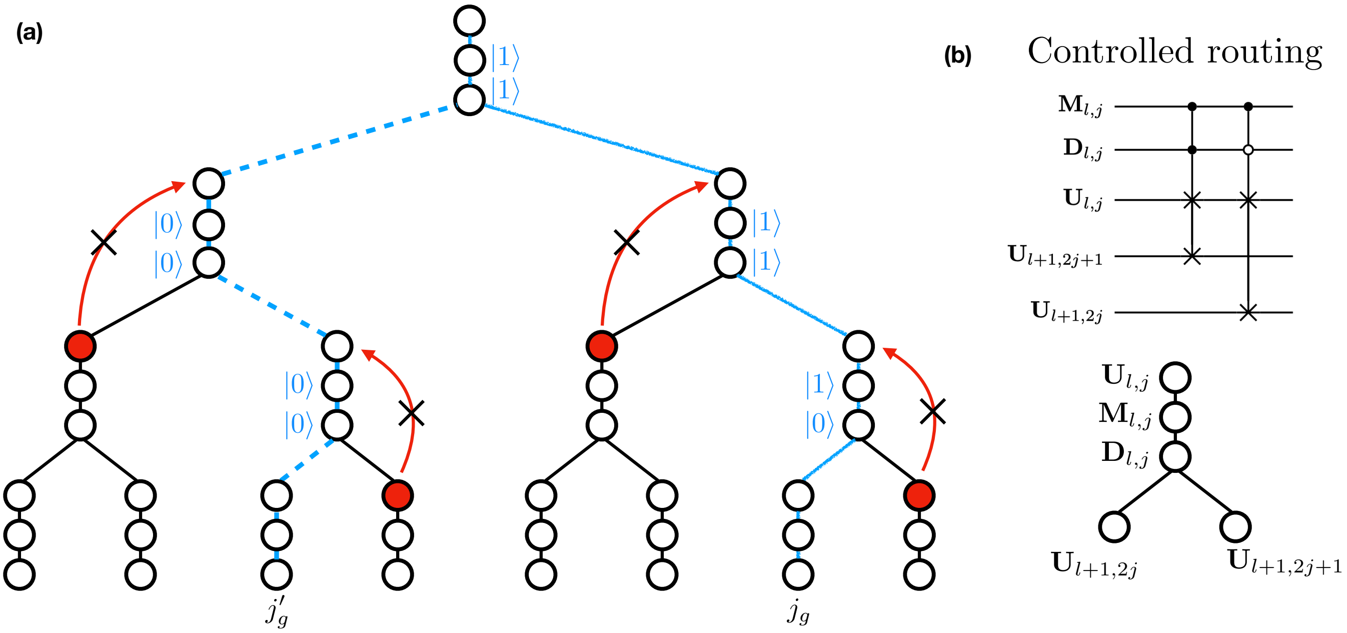


FIG. 3: (a) For both $j_g \in g(c)$ and $j'_g \in g(c)$, errors never propagate into the good branch, because all controlled qubits $\mathbf{M}_{l,j}$ are error free. (c) Sketch of the controlled routing operations.

4. Robustness analysis

With an abuse of notation, we define *good* index and relevant terminologies here in a similar way to the 2-qubit-per-node protocol. Let $\mathscr{A}_j$ be all ancestors of $\mathbf{U}_{n,j}$ in both QRAM and output register. We also define $\hat{\mathscr{A}}_j$ as the intersection of $\mathscr{A}_j$ and its nearest neighbour. Similar to the 2-qubit-per-node protocol, for a specific space-time-polarization configuration of error c, we define $g(c)$ as set of all *good* index j, such that all qubits in $\hat{\mathscr{A}}_j$ are free of errors at all time.

With the same argument to the 2-qubit-per-node protocol, the final output state of can be expressed as

$$|\tilde{\psi}'\rangle = \sum_{j\in g(c)} \psi_j |f_j\rangle_{\text{qram}} \otimes |j\rangle_{\text{out}} + |\text{garb}'\rangle. \tag{B21}$$

for some garbage state that is orthogonal to the first term. Yet, the main difference is that in the 3-qubit-per-node protocol here, errors will never propagate into the good branches $j \in g(c)$ (as oppose to $j \in g'(c)$ in the 2-qubit-per-node protocol). The reason is as follows (see also Fig. 3). During the fanout process, we suppose the quantum state at a certain step t is

$$|\tilde{\psi}'_t\rangle = \sum_{j\in g(c)} \psi_j |\psi'_{t,j}\rangle_{\text{out}} + |\text{garb}'_t\rangle. \tag{B22}$$

We now consider basis $|\psi'_{t,j_b}\rangle$ for some $j_g \in g(c)$. During controlled routing operations, errors will not propagate into the branch j_g, because the controlled and routing qubits are at correct state. We now consider other good branch $j'_g \in g(c)$ that $j'_g \neq j_g$. All of their control qubits in the middle sublayers are free of errors, and hence at state $|0\rangle$. Therefore, all corresponding routing operations does not perform any swapping, and errors will not propagate from bad branch to the branch j'_g.

As a result, errors perform trivially at all good branch, and for all $j \in g(c)$, we have $|f_j\rangle_{\text{qram}} = |f\rangle_{\text{qram}}$ for some computational basis f independent of j. Let $\Lambda = \sum_{j\in g(c)} |\psi_j|^2$, With the same argument for obtaining Eq. (A37) in Appendix. A 6, we have $F \geqslant \mathbb{E}[\Lambda]$. With the same argument for obtaining Eq. (A43), we also have

$$\mathbb{E}[\Lambda] \geqslant \Pr[j \in g(c)]. \tag{B23}$$

Because $\mathscr{A}_j = O(n)$ and the algorithm has runtime $O(n)$, we have $\Pr[j \in g(c)] \geqslant (1-\varepsilon)^{O(n)\times O(n)} \geqslant 1 - A\varepsilon n^2$ for some constant A. Therefore, the total infidelity satisfies

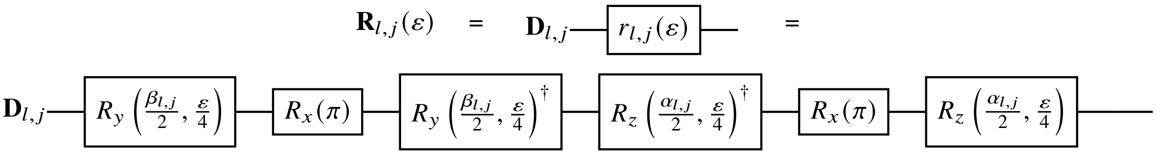
$$1 - F \leqslant A\varepsilon n^2. \tag{B24}$$

5. Clifford+T decomposition

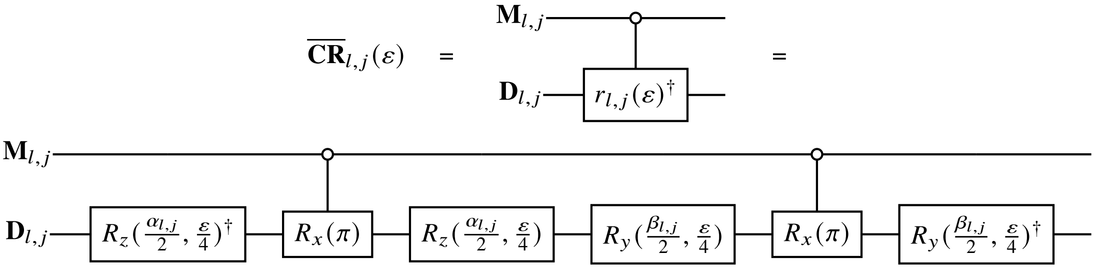
a. Decomposition protocol and error analysis

Among all elementary single- and two-qubit gates, only rotations $\mathbf{R}_{l,j}$ and controlled rotations $\overline{\mathbf{CR}}_{l,j}$ has decomposition error, while all other elementary gates can be ideally constructed with constant number of Clifford and T gates.

According to [33], given an arbitrary z-rotation $R_z(\alpha)$ and accuracy $\varepsilon > 0$, we can always construct a single qubit rotation $R_z(\alpha, \varepsilon)$ with $O(\log(1/\varepsilon))$ depth of H and T gates, such that $\|R_z(\alpha, \varepsilon) - R_z(\alpha)\| \leqslant \varepsilon$. For y-rotation $R_y(\beta)$, the result is similar. Moreover, we can always decompose each $r_{l,j}$ into the concatenation of a y-rotation and a z-rotation $r_{l,j} = R_z(\alpha_{l,j})R_y(\beta_{l,j})$. We approximate $\mathbf{R}_{l,j}$ with the following quantum circuit


$\mathbf{R}_{l,j}(\varepsilon)$
=
$\mathbf{D}_{l,j}$
$r_{l,j}(\varepsilon)$
=
$\mathbf{D}_{l,j}$
$R_y\left(\frac{\beta_{l,j}}{2}, \frac{\varepsilon}{4}\right)$
$R_x(\pi)$
$R_y\left(\frac{\beta_{l,j}}{2}, \frac{\varepsilon}{4}\right)^\dagger$
$R_z\left(\frac{\alpha_{l,j}}{2}, \frac{\varepsilon}{4}\right)^\dagger$
$R_x(\pi)$
$R_z\left(\frac{\alpha_{l,j}}{2}, \frac{\varepsilon}{4}\right)$

Note that $R_z\left(\frac{\alpha_{l,j}}{2}, \frac{\varepsilon}{4}\right)^\dagger$ can be constructed by inverse the H, T gate sequence of $R_z\left(\frac{\alpha_{l,j}}{2}, \frac{\varepsilon}{4}\right)$, then replace T and H by $T^\dagger$ and $H^\dagger$. $R_x(\pi)$ takes $O(1)$ gate count, and $\|\mathbf{R}_{l,j} - \mathbf{R}_{l,j}(\varepsilon)\| \leqslant \varepsilon$. The reason of using this decomposition is that together with the controlled rotation introduced below, qubits in $\mathscr{D}_{l,j}$ can be fully uncomputed after implementing $\overline{\mathbf{CR}}_{l,j}$. To be specific, $\overline{\mathbf{CR}}_{l,j}$ is approximated by


$\mathbf{M}_{l,j}$
$\overline{\mathbf{CR}}_{l,j}(\varepsilon)$
=
$\mathbf{D}_{l,j}$
$r_{l,j}(\varepsilon)^\dagger$
=
$\mathbf{M}_{l,j}$
$\mathbf{D}_{l,j}$
$R_z(\frac{\alpha_{l,j}}{2}, \frac{\varepsilon}{4})^\dagger$
$R_x(\pi)$
$R_z(\frac{\alpha_{l,j}}{2}, \frac{\varepsilon}{4})$
$R_y(\frac{\beta_{l,j}}{2}, \frac{\varepsilon}{4})$
$R_x(\pi)$
$R_y(\frac{\beta_{l,j}}{2}, \frac{\varepsilon}{4})^\dagger$

In our Cliffor+T circuit implementation, we just perform the following replacement in the fanin phase

$$\mathbf{R}_{l,j} \to \mathbf{R}_{l,j}(\varepsilon_l), \quad \overline{\mathbf{CR}}_{l,j} \to \overline{\mathbf{CR}}_{l,j}(\varepsilon_l). \tag{B25}$$

To analyse the decomposition accuracy, we define

$$U_l = \begin{cases} r_{0,0} \otimes \mathbb{I}_{n-1}, & l = 0 \\ \sum_{j=0}^{2^l-1} |j\rangle\langle j| \otimes r_{l,j} \otimes \mathbb{I}_{n-l-1} & 1 \leqslant l \leqslant n-1 \end{cases} \tag{B26}$$

and

$$U_l(\varepsilon_l) = \begin{cases} r_{0,0}(\varepsilon_0) \otimes \mathbb{I}_{n-1}, & l = 0 \\ \sum_{j=0}^{2^l-1} |j\rangle\langle j| \otimes r_{l,j}(\varepsilon_l) \otimes \mathbb{I}_{n-l-1} & 1 \leqslant l \leqslant n-1 \end{cases} \tag{B27}$$

where $\mathbb{I}_m$ is the m-qubit identity matrix. It can be verified that for ideal and Clifford+T implementations, the final state of the output register can be expressed as

$$|\psi\rangle = U_{n-1}\cdots U_1 U_0|0\rangle^{\otimes n} \tag{B28}$$

$$|\psi^{(\mathrm{CT})}\rangle = U_{n-1}(\varepsilon_{n-1})\cdots U_1(\varepsilon_1)U_0(\varepsilon_0)|0\rangle^{\otimes n} \tag{B29}$$

respectively, while the QRAM has been uncomputed for both cases. Moreover, we have $\|U_l - U_l(\varepsilon_l)\| \leqslant \varepsilon_l$. According to the triangular inequality, we have

$$\left\| |\psi\rangle - |\psi^{(\mathrm{CT})}\rangle \right\| \leqslant \sum_{l=0}^{n-1} \|U_l - U_l(\varepsilon_l)\| \leqslant \sum_{l=0}^{n-1} \varepsilon_l. \tag{B30}$$

Based on Eq. (B30), to achieve a given accuracy $\left\| |\psi\rangle - |\psi^{(\mathrm{CT})}\rangle \right\| \leqslant \varepsilon$, it suffices to set

$$\varepsilon_l = \varepsilon/2^{n-l}. \tag{B31}$$

b. Circuit complexity

Below, we analysis the Clifford+T circuit complexity based on the decomposition protocol above.

Clifford+T gate count. Each rotation $\mathbf{R}_{l,j}(\varepsilon_l)$ or controlled-rotation $\mathbf{CR}_{l,j}(\varepsilon_l)$ accounts for $O(\log 1/\varepsilon_l) = O(\log(2^{n-l}/\varepsilon))$ gate count. So **PR** and $\mathbf{P\overline{CR}}$ accounts for totally $\sum_{l=1}^{n} 2^l \times O(\log(2^{n-l}/\varepsilon)) = O(N\log(1/\varepsilon))$ gate count. Other operations during the implementation can be realized without decomposition error, and accounts for $O(N)$ gate count. Therefore, the total Clifford+T gate count is $O(N\log(1/\varepsilon))$.

Clifford+T depth. The decomposed parallel rotation and parallel controlled-rotation accounts for $\max_l O(\log(2^{n-l}/\varepsilon)) = O(\log(2^n/\varepsilon)) = O(n + \log(1/\varepsilon))$ circuit depth. Other operations during the implementation accounts for totally $O(n)$ depth. So the total Clifford+T depth is $O(n + \log(1/\varepsilon))$.

Clifford+T space-time allocation. We first consider the total STA of the lth spatial layer. Each qubit is activated for time $O(n-l) + O(\log(2^{n-l}/\varepsilon)) = O(n - l + \log(1/\varepsilon))$. There are $O(n)$ qubits at the lth spatial layer, so the total STA is $A_l = O(2^l(n - l + \log(1/\varepsilon)))$. Moreover, the STA of output register is $A_{\mathrm{out}} = O(n) \times O(n) = O(n^2)$. Therefore, the total STA of the algorithm is

$$\begin{aligned} A &= A_{\mathrm{out}} + \sum_{l=1}^{n} A_l \\ &= O(n^2) + \sum_{l=0}^{n} O(2^l(n - l + \log(1/\varepsilon))) \\ &= O(n^2) + O(N) + O(N\log(1/\varepsilon)) \\ &= O(N\log(1/\varepsilon)). \end{aligned} \tag{B32}$$

In Table. I below, we compare our circuit complexity to existing results.

TABLE I: Clifford+T complexities some typical low-depth state preparation protocols.

Protocols	Count	Depth	STA	connectivity	Robustness
Ref [8]	$O(N\log(N/\varepsilon))$	$O(n\log(N/\varepsilon))$	$O(Nn\log(N/\varepsilon))$	all-to-all	NA
Ref [13]	$O\ (N\log(n/\varepsilon))$	$\boldsymbol{O(n + \log(1/\varepsilon))}$	$\boldsymbol{O\ (N\log(1/\varepsilon))}$	all-to-all	NA
Ref [14]	$\boldsymbol{O\ (N\log(1/\varepsilon))}$	$O\ (n\log(n/\varepsilon))$	$O(N\log(n/\varepsilon))$	degree 4	NA
2-qubit-per-node	$\boldsymbol{O\ (N\log(1/\varepsilon))}$	$O\ (n\log(n/\varepsilon))$	$O\ (N\log(n/\varepsilon))$	**degree** 3	$1 - F \leqslant A\varepsilon n^3$
3-qubit-per-node	$\boldsymbol{O\ (N\log(1/\varepsilon))}$	$\boldsymbol{O\ (n + \log(1/\varepsilon))}$	$\boldsymbol{O\ (N\log(1/\varepsilon))}$	**degree** 3	$\boldsymbol{1 - F \leqslant A\varepsilon n^2}$