Efficient Approximation of Diagonal Unitaries over the Clifford+$T$ Basis

Jonathan Welch†, Alex Bocharov∗, and Krysta M. Svore∗

†Department of Chemistry and Chemical Biology,
Harvard University, Cambridge, MA (USA)
∗Quantum Architectures and Computations Group,
Microsoft Research, Redmond, WA (USA)

(Dated: December 23, 2014)

We present an algorithm for the approximate decomposition of diagonal operators, focusing specifically on decompositions over the Clifford+$T$ basis, that minimize the number of phase-rotation gates in the synthesized approximation circuit. The equivalent $T$-count of the synthesized circuit is bounded by $kC_0\log_2(1/\varepsilon) + E(n,k)$, where $k$ is the number of distinct phases in the diagonal $n$-qubit unitary, $\varepsilon$ is the desired precision, $C_0$ is a quality factor of the implementation method ($1 < C_0 < 4$), and $E(n,k)$ is the total entanglement cost (in $T$ gates). We determine an optimal decision boundary in $(k,n,\varepsilon)$-space where our decomposition algorithm achieves lower entanglement cost than previous state-of-the-art techniques. Our method outperforms state-of-the-art techniques for a practical range of $\varepsilon$ values and diagonal operators and can reduce the number of $T$ gates exponentially in $n$ when $k << 2^n$.

I. INTRODUCTION

Diagonal unitary (DU) operators are used in many quantum algorithms, for example, as simple analytical potential operators for quantum simulation and as complex oracles used to divine the answer in quantum searches. However, in order to implement a quantum algorithm on a given quantum device, each operator must be decomposed into a sequence of fault-tolerant, device-level instructions. Therefore, efficient low-level implementation of DU operators is essential to any quantum compiler framework. In this work we develop methods for the approximate decomposition of diagonal operators, focusing specifically on decompositions over the Clifford+$T$ basis. Since the $T$ gate requires substantial resource overhead as compared to Clifford gates [1, 2] we seek to minimize the number of $T$ gates in our approach. We analyze the tradeoffs between the $T$-cost of entangling operators and the $T$-cost of decomposing single-qubit rotations.

Methods for the exact decomposition of diagonal unitary operators [3, 4] focus on minimization of the number of one- and two-qubit gates, where the total number of single-qubit axial rotations $R_z$ and two-qubit CNOT operators is upperbounded by $O(2^n+1 - 3)$. In Ref. [3], an $n$-qubit diagonal unitary operator is treated as a discrete function \{f_k\}_{k=0}^{2^n-1}. The discrete function is then amenable to a fourier-like decomposition over the basis of Walsh functions. A series expansion over this basis affords the following relations

\[
\hat{U}(k) = \exp(i f_k) \\
= \exp \left( \sum_{j=0}^{2^n-1} a_j \hat{w}_{jk} \right) \\
= e^{i a_0 \hat{w}_0} e^{i a_1 \hat{w}_1} \ldots \\
= \prod_{j=0}^{2^n-1} e^{i a_j \hat{w}_{jk}},
\]

1 These are a binary periodic functions. See [3] for more details and c.f. Figure 1 of that reference for examples of the first eight such functions
where \( a_j = 1/(2^n) \sum_{k=0}^{2^n-1} f_k \hat{w}_{jk} \). Each of these basis functions has a one-to-one mapping with a corresponding quantum circuit which implements the basis function as a tensor product of \( Z \)-rotations on the input register. The splitting of the sum in the exponent into individual exponential terms in the product uses the commutative nature of the \( \hat{w}_{jk} \) operators. The \( \hat{w}_{jk} \), in addition to having a functional form, should more generally be thought of as a set of basis circuits.

Indeed, one does not require that a functional basis be explicitly used to perform this decomposition as evidenced in Ref. [4]. The authors of [4] perform a tensor product decomposition and show (c.f., Figure 3 of reference [4] and text therein) that the effective basis for their decomposition is those circuits corresponding to the basis circuits utilized in [3] as the Walsh basis operators. This correspondence explains why the two approaches converge on identical upper bounds and equivalent quantum circuits. The primary benefit associated with utilizing the map between the quantum circuits and the corresponding Walsh series expansion is the ability to utilize the tools of fourier analysis directly on the corresponding DU operator. That is to say, one may allow for a certain error tolerance \( \varepsilon \) such that 

\[
\left| \hat{U}_\varepsilon(k) - \hat{U}(k) \right| \leq \varepsilon \quad \text{for all } k,
\]

and utilize any approximation tools valid for the functional basis as a way of reducing the number of non-zero expansion coefficients that are required for reconstruction of the operator within the specified error tolerance. For DU operators whose discrete functional equivalent \( f_k \) has a rapidly converging Walsh series, the corresponding quantum circuit complexity can be reduced significantly as the number of single-qubit rotations (which corresponds exactly with the number of non-zero expansion coefficients in the Walsh series) can often be small and hence efficiently implemented.

Existing methods [3, 4] for the exact decomposition of DU operators exhibit the property that all entanglement occurs through the use of elementary CNOT gates, which have negligible fault-tolerant cost. This results in the entirety of the cost being placed on the fault-tolerant implementation of single-qubit rotations, of which in general an exponential number are required. When the phases of the DU are drawn from a finite collection of phases, exact methods in general produce an overly pessimistic number of single-qubit rotations because they delocalize the rotations required over the entire \( n \)-qubit operational space. In certain cases, delocalization benefits from highly non-local correlations in the phase values and leads to a small number of single-qubit rotations. In general, however, the rotation angles produced are not exactly implementable over the Clifford+\( T \) basis\(^2\) and hence approximation methods are required.

Exact methods face two challenges, which we aim to address in the present work:

1. The decompositions are performed to only the single-qubit \( R_z(\theta_i) \) rotation level and assume that the set of rotation angles \( \{\theta_i\} \) are exactly implementable on a given quantum architecture. In the Clifford+\( T \) basis, such rotations are seldom exactly representable [1, 2].

2. The methods do not take advantage of the existence of a phase context underlying the diagonal operator.

We develop the idea of a phase context and explicitly determine the circuit cost of each circuit element at the level of Clifford+\( T \) gates. Our approach is based around systematically distinguishing between the actual phase rotation and entanglement degrees of freedom within the diagonal operator. We rely on the fact that the amount of entanglement required in certain circuits depends on the number of qubits, and not on the desired precision \( \varepsilon \). In the case of a single multi-controlled rotation, this dependence has been shown, for example, in the circuit shown in Figure 8 of [5]. To the best of our knowledge, our application of this approach to the broader context is new.

We present an algorithm to decompose over the phase context of the operator, which offers flexibility in how the rotation angles are distributed among the elements of the circuit implementing

\(^2\) This is in the sense that the angles correspond to the real valued expansion coefficient of a Fourier type series.
the rotations. The phase context of the operator is the finite alphabet of phases found in the operator’s diagonal. As we will show, the rotation angles that need to be approximated will always be given as the ratio of two distinct phases from the context. This provides the ability to choose one of several possible phase context decompositions so that the required single-qubit rotations can be adjusted to have angles with a minimal cost of $\varepsilon$-approximation. Additionally, since the decomposition is over the phase context rather than the operator itself, and the context is assumed to have as most $k$ distinct phase values, there will never be more than $k$ single-qubit rotations required to implement the diagonal unitary operator.

Unlike the previously mentioned exact methods, the phase-context decomposition requires non-trivial entangling operations which, in general, are multi-controlled Toffoli gates. The generalized Toffoli gate may appear to have a potentially significant fault-tolerant cost, however the fault-tolerant cost is independent of the target accuracy for the single-qubit phase approximations. Therefore the asymptotic cost will in general be dominated by the number of single-qubit phase rotations. In the case of a phase-sparse matrix where the size of the phase context $k$ is much less than the length of the diagonal $2^n$, the low number of single-qubit rotations results in an overall lower fault-tolerant cost for the phase-context approach, despite incurring an additional entanglement cost which turns out to be asymptotically constant.

Our paper is organized as follows. In Section II we rigorously define phase contexts and cascaded entanglers, and their role in decomposition. In Section III we motivate our approach by example. Our phase-context decomposition algorithm is presented in Section IV. We present numerical results in Section V and finally conclude in Section VI.

II. PHASE CONTEXTS AND CASCADED ENTANGLERS

Consider a diagonal unitary operator on $n$ qubits with $k << 2^n$ distinct phases, such that each distinct phase $\phi_i$ appears in blocks along the diagonal as follows:

$$U = \text{diag}(\phi_1, \ldots, \phi_1, \phi_2, \ldots, \phi_2, \ldots, \phi_k, \ldots, \phi_k).$$

Throughout, we drop the index on the phase when it is clear from the surrounding text. We may represent $U$ as a product of a global phase (e.g., $\phi_1$) and a set of one-parameter diagonal operators of the form

$$V(\phi, \ell) = \text{diag}(1, \ldots, 1, \phi, \ldots, \phi).$$

(1)

A diagonal operator whose diagonal is sampled from the set of $k$ distinct phase values $\Phi = \{\phi_i\}_{i=1}^k$, called the phase context, requires at most $k - 1$ operators of the form $V(\phi, \ell)$ to implement.

Suppose the phase $\phi_j$ occurs $\ell_j$ times on the diagonal and suppose $L_m = \sum_{j=1}^{m} \ell_j$. Then, by direct computation

$$U = \phi_1 V(\phi_2/\phi_1, L_k - L_1) V(\phi_3/\phi_1, \phi_2, L_k - L_2), \ldots, V\left(\prod_{j=k \text{mod } 2} \frac{\phi_j}{\prod_{j \neq k \text{mod } 2} \phi_j}, \ell_k = L_k - L_{k-1}\right).$$

We present an algorithm to perform a phase context decomposition (PCD), where the task is to decompose $U$ into a product of $k - 1$ phase rotation gates of the form (1). Suppose $\phi = e^{i\theta}$, where $\theta \in \mathbb{R}$. We prove that the operator $V(\phi, \ell)$ can be realized, up to a global phase, using:

- One ancillary qubit initialized to $|0\rangle$;
We require several definitions before describing the algorithm. Let $J = \{|j_1\rangle, \ldots, |j_\ell\rangle\}$ be the set of basis vectors rotated by $V = V(\phi = e^{i\theta}, \ell)$, where

$$V |j\rangle = \begin{cases} \phi |j\rangle, & |j\rangle \in J \\ |j\rangle, & \text{otherwise} \end{cases}.$$  

Let $\Omega_\ell(j)$ be a logical activation function such that

$$\Omega_J(j) = \begin{cases} 1, & |j\rangle \in J \\ 0, & \text{otherwise} \end{cases}.$$  

Then $V(\phi, \ell)$ can be written as, where $J = \{|j\rangle : j \geq 2^n - \ell\}$,

$$V = V(\phi, \ell) = \sum_{j=0}^{2^n-1} \left(\phi^{\Omega_J(j)}\right) |j\rangle \langle j|.$$  \hspace{1cm} (2)

Each operator $V$ is associated with a so-called cascaded entangler, denoted $X^n(V)$, and formally defined on the $(n+1)$-qubit basis as

$$X^n(V) |j\rangle |b\rangle = |j\rangle |b \oplus \Omega_J(j)\rangle,$$

where $j \in \{0, \ldots, 2^n - 1\}$, $b \in \{0, 1\}$. Figure 1 shows the the circuit implementation for the case $\ell = 1$.

The decomposition of a cascaded entangler $X^n(V)$ results in a circuit whose cost depends only on the number of qubits $n$ and on the structure of the operator $V(\phi, \ell)$, and is independent of the desired precision $\varepsilon$. Using constructions in Refs. [6, 7], a cascaded entangler can be represented exactly by a Clifford+$T$ circuit where the number of $T$ gates, called the $T$-count, depends only on $n$ and $\ell$. Moreover, the cost of representing a pair of matching entanglers separated by an ancilla rotation (c.f., Figure 1) will be less than twice the cost of one entangler in most cases, since many of the gates in the decomposition of the first entangler will cancel with those of the matching entangler. We denote the minimal $T$-count of a Clifford+$T$ circuit that implements a pair of matching cascaded entanglers as $E[n, \ell]$.

We prove the following statements on cascaded entanglers:
1. A cascaded entangler $X^n(V)$ can be represented exactly as a composition of local NOT gates, CNOTs, and multi-controlled-NOT gates.

2. Any multi-controlled-NOT gate on $n$ qubits is represented exactly as a Clifford+$T$ circuit with a $T$-count proportional to the number of controls (using at most one additional ancillary qubit).

To avoid ambiguity, when given $m < n$ we reserve the notation $\Lambda^m(X)$ for the $n$-qubit gate defined as

$$\Lambda^m(X)|j\rangle = |j_1, \ldots, j_m, j_{m+1}, j_m \oplus c_m(j)\rangle,$$

where $j_1, \ldots, j_n$ is the bit string representation of $j$ and $c_m(j) = j_1 \land \ldots \land j_m$. As per [6, 7], a $\Lambda^m(X)$ can be implemented exactly by a Clifford+$T$ circuit with $T$-count in $\mathcal{O}(m)$ (assuming an additional ancillary qubit when $n = m + 1$).

A single-qubit axial rotation $P(\theta)$ can be approximated to precision $\varepsilon$ by a Clifford+$T$ circuit with an expected $T$-count of $C_0 \log_2(1/\varepsilon) + O(\log(\log(1/\varepsilon)))$, where $C_0$ is a constant depending on the decomposition scheme (c.f., [6, 8]).

Thus the $T$-count required to approximate the operator $V(\phi, \ell)$ is bounded by

$$C_0 \log_2(1/\varepsilon) + O(\log(\log(1/\varepsilon))) + E[n, \ell].$$

The target diagonal operator $U$ is represented as the product

$$U = \prod_{m=1}^{k-1} V(\phi_m, \ell_m),$$

and thus can be approximated to precision $\varepsilon$ by concatenating circuits that approximate the respective operators $V(\phi_m, \ell_m)$ to precision $\varepsilon/(k-1)$, i.e.,

$$U_\varepsilon = \prod_{m=1}^{k-1} V_{\frac{\varepsilon}{k-1}}(\phi_m, \ell_m).$$

Therefore the required $T$-count of the overall approximation circuit is bounded by

$$(k-1) C_0 \log_2(1/\varepsilon) + O(\log(\log(1/\varepsilon))) + E[n, k, \{\ell_1, \ldots, \ell_{k-1}\}],$$

where $E[n, k, \{\ell_1, \ldots, \ell_{k-1}\}]$ is an overall upper bound for the total cost of all cascaded entanglers generated in this decomposition.

### III. A MOTIVATING EXAMPLE

Let $W \in U(2)$ be an arbitrary single-qubit unitary. Now suppose we intend to implement an $(n+1)$-qubit controlled version $G = \Lambda^n(W)$. A traditional approach would be to decompose $G$ into a network of cascaded CNOTs and uncontrolled single-qubit unitaries. However, unless the resulting single-qubit unitaries can be performed exactly and fault-tolerantly, the major contributor to the asymptotic cost of the decomposition will be the cost of approximating the single-qubit unitaries.

$\varepsilon$ Also note that $[V(\phi, \ell), V(\theta, m)] = 0$ and hence the circuits can be implemented in any order.

$\sum_m E[n, \ell_m]$ can be used as a proxy for $E[n, k, \{\ell_1, \ldots, \ell_{k-1}\}]$, although the bound can be tightened in most cases.
unitaries. When the desired approximation precision $\varepsilon$ is very small, the cost of the single-qubit approximations dominates the total cost of the decomposition of $G$.

Therefore at a small precision level a more cost-efficient approach is to first consider an Euler-angle decomposition of $W$ in order to minimize the number of single-qubit unitaries in the final decomposition. Let $W = \delta \Lambda(\alpha) H \Lambda(\beta) H \Lambda(\gamma)$, where $\alpha, \beta, \gamma, \delta$ are phase factors and $H$ is the Hadamard gate. Then,

$$\Lambda^n(W) = \Lambda^n(\delta) \Lambda^{n+1}(\alpha) \Lambda^n(H) \Lambda^{n+1}(\beta) \Lambda^n(H) \Lambda^{n+1}(\gamma).$$

Recall that $\Lambda^n(H)$ is representable exactly in the Clifford+$T$ basis with a $T$-count of $O(n)$, which is independent of the desired precision [6, 9].

The operators $\Lambda^n(\delta), \Lambda^{n+1}(\alpha), \Lambda^{n+1}(\beta), \Lambda^{n+1}(\gamma)$ are each one-parameter diagonal unitaries. By allowing one ancillary qubit and using cascaded entanglers, we can approximate each of these unitaries with a circuit whose cost, up to the cost of cascaded entanglers, is dominated by the $T$-count of approximating a single axial rotation. Assuming a given method for approximating any given axial rotation has asymptotic cost of the form:

$$C_0 \log_2(1/\varepsilon) + O(\log(\log(1/\varepsilon))),$$

we conclude that the cost of implementing $\Lambda^n(W)$ is asymptotically dominated by $4 C_0 \log_2(1/\varepsilon) + O(\log(\log(1/\varepsilon)))$. The associated entanglement cost, which depends only on $n$, becomes increasingly less relevant as $\varepsilon$ tends to 0.

Of the standard decomposition methods, the technique described in Lemma 7.11 of [7] is suitable to this approximation context. The decomposition is ancilla-assisted and expresses $\Lambda^n(W)$ as a composition of two entanglers and $\Lambda^1(W)$. Unless the latter is reduced to diagonal unitaries, as proposed above, the decomposition recipes in [7] call for representing $\Lambda^1(W)$ as a circuit consisting of CNOTs and three single-qubit unitaries, where each of the latter may be represented with at most three axial rotations, resulting in a total of nine axial rotations and a corresponding $T$-count upper bounded by $9 C_0 \log_2(1/\varepsilon) + O(\log(\log(1/\varepsilon)))$. This example demonstrates the approach for the case of (at most) four very special diagonal unitaries, each containing only one non-trivial phase. In the next section we present the solution for the general case.

### IV. DIAGONAL OPERATOR DECOMPOSITION ALGORITHM

We present our algorithm to decompose diagonal unitaries using phase contexts, called *Decomposition with Cascaded Entanglers*, in Algorithm [1]. The algorithm takes as input a diagonal operator $U$, the precision $\varepsilon$, and a method for single-qubit unitary decomposition $A$, where $A$ could be based on techniques in Refs. [8, 10–12].

In Line [1] the input diagonal operator $U$ undergoes phase context decomposition, where $U$ is represented as the product of a global phase $\phi = e^{i\theta}$ and $k - 1$ operators of the form $V(\phi, \ell)$, where $k$ is the number of distinct phase values appearing on the diagonal. In Line [2] the global phase parameter is saved in $\text{ret}$. In Line [3] for each of the $k - 1$ remaining phases, we store local copies of the operator and the phase gate we wish to associate it with (Line [4]), then use the cascaded entangler algorithm to construct the pair of cascaded entanglers associated to the diagonal operator $V$ (Line [5]). Line [6] calls CED2 (cascaded entangler decomposition for a pair) to obtain an exact representation of a pair of matching entanglers over the Clifford+$T$ basis. The design of the cascaded entangler decomposition (CED) algorithm is described in Section IV.B given a gate of the form $X^n(V)$, it decomposes the gate into multi-controlled-NOT gates and Clifford+$T$ circuits [6, 7] for exact representation of each $\Lambda^n(X)$ at a $T$-count in $O(m)$. Line [7] calls the desired single
qubit unitary approximation algorithm to decompose the phase gate \( R = P(\phi_j) \). The result is then used as the \( \varepsilon \)-approximate implementation of \( R \) and Line 8 moves that result in place of \( R \) in the circuit for \( X^n(V)(I_n \otimes R)X^n(V) \). The latter is an implementation of the individual factor \( V(\phi_j, \ell_j) \). Finally in Line 9 this result is combined with the global phase calculated earlier and stored in \( \text{ret} \) as the return value for the algorithm.

Algorithm 1 Decomposition with Cascaded Entanglers

Require: \( n, U = \text{diag}(\phi_1, \ldots, \phi_{2^n}), \varepsilon > 0, \mathcal{A} \)

1: \( \text{factors} \leftarrow \phi \prod_{j=1}^{\ell} V(\phi_j, \ell_j) \) \{Phase-context decomposition\}
2: \( \text{ret} \leftarrow \{\phi\} \)
3: for \( j = 1, \ldots, k - 1 \) do
4: \( V \leftarrow V(\phi_j, \ell_j); R \leftarrow P(\phi_j) \)
5: \( X^n(V) \leftarrow \text{cascaded_entangler}(V) \)
6: \( \text{cv} \leftarrow \text{CED2}(X^n(V), I_n \otimes R, X^n(V)) \) \{Exact Clifford+T representation\}
7: \( \text{cv} \leftarrow \mathcal{A}(R, \varepsilon) \) \{Approximation circuit\}
8: \( \text{cv} \leftarrow \text{replace}(R \mapsto \text{cv} \text{ in } \text{cv}) \)
9: \( \text{ret} \leftarrow \text{cv} \text{ ret} \)
10: end for
11: return \( \text{ret} \)

A. Decomposition of Cascaded Entanglers

We demonstrate that for \( V = V(\phi, \ell) \) the cascaded entangler \( X^n(V) \) is completely defined by \( \ell \) and develop an algorithm for expanding \( X^n(V) \) into a network of multi-controlled-NOT gates based on the binary representation of \( \ell \). We then use the network for estimating an upper bound on the cost of \( X^n(V) \) over the Clifford+T basis.

Given \( \ell \leq 2^n \), let us introduce the \((n+1)\)-qubit entanglement operator

\[
X^n(\ell)|k\rangle|b\rangle = \begin{cases} 
|k\rangle|b \oplus 1\rangle & k \geq 2^n - \ell, \\
|k\rangle|b\rangle & k < 2^n - \ell,
\end{cases} \tag{5}
\]

where \( |k\rangle \) is an \( n \)-qubit basis state and \( b \in \{0, 1\} \).

This operator can be alternatively defined in terms of activation functions as

\[
X^n(\ell) = |k\rangle|b \oplus \Omega_{2^n-\ell,2^n}(|k\rangle)\rangle |b\rangle |k\rangle. \tag{6}
\]

Observation 1. Using a single \((n+1)\)-st ancillary qubit, \( V(\phi, \ell) \) can be implemented as

\[
X^n(\ell) (I_n \otimes P(\phi)) X^n(\ell).
\]

Indeed, when \( X^n(\ell) \) is so defined, then the state \( |k\rangle|0\rangle \) picks up the phase factor \( \phi \) iff \( k \geq 2^n - \ell \), which is exactly how \( V(\phi, \ell) \otimes I \) acts on the state \( |k\rangle|0\rangle \).

A suboptimal way of implementing \( X^n(\ell) \) would be to factor it into a product \( \prod_{j=0}^{2^n-1} Y_n(j) \), where \( Y_n(j) |k\rangle|b\rangle = |k\rangle|b \oplus \delta_{kj}\rangle \), for \( k = 0, \ldots, 2^n - 1 \) and \( b \in \{0, 1\} \), is a \( \Lambda^n(X) \) gate. Under this factorization, the cost \( X^n(\ell) \) is dominated by \( \ell \) times the cost of \( \Lambda^n(X) \), which is a uniform worst-case bound. We show how this bound can be effectively tightened (unless, indeed, it is the definitive worst case).
B. The Cascaded Entangler Decomposition (CED) Algorithm

We propose a simple recursive algorithm for decomposing the $X^n(\ell)$ operator. We consider the cost of Pauli and Clifford gates to be negligible and count the number of $T$ gates. To this end we consider a slightly more general operator $X^n(p,q)$, where $p \leq q \leq 2^n$, and the operator is defined in terms of the activation function $\Omega_{[p,q]}$ as

$$X^n(p,q)|k\rangle|b\rangle = |k\rangle|b \oplus \Omega_{[p,q]}(k)\rangle.$$  

(7)

The gate $X^n(p,q)$ acts as an activation operator which inverts the register $|b\rangle$, $b \in \{0,1\}$, when $p \leq k < q$. The following identity immediately follows: $X^n(\ell) = X^n(2^n - \ell, 2^n)$.

**Observation 2.** Any $X^n(p,q)$ operator is effectively equivalent to $X^n(q-p)$ modulo the Pauli group.

Indeed $X^n(p,q)$ is an entanglement operator, where the $X$ on the ancillary qubit is activated by the primary basis states $|p\rangle, \ldots, |q-1\rangle$. Let $|b_{n-1} \ldots b_0\rangle$ be the bit representation of $q-1$. Apply an adjoining pair of $X$ gates on all qubit wires $k$ for which $b_k = 0$, that is, that are conditioned on 0. We obtain an entanglement operator such that the inversion of the ancillary qubit is activated by the primary basis states $|2^n - (q-p)\rangle, \ldots, |2^n - 1\rangle$ which is the $X^n(q-p)$ gate. In other words, $X^n(p,q)$ can be written as $(X^{b_{n-1}} \otimes \cdots \otimes X^{b_0})X^n(q-p)(X^{b_{n-1}} \otimes \cdots \otimes X^{b_0})$, where $\delta_b$ is the Kronecker delta.

**Proposition 3.** $X^n(p,p) = \hat{I}(n)$, where $\hat{I}(n)$ is the n-qubit identity operator for any $p \leq 2^n$.

**Proof.** This follows from Definition (5) and the fact that $|p,p\rangle$ is empty and thus $\Omega_{[p,p]}$ is constant zero.

**Proposition 4.** For $p < q_1 < q_2$ we have $X^n(p,q_1)X^n(q_1,q_2) = X^n(p,q_2)$

**Proof.** $[p,q_2]$ is the disjoint union of $[p,q_1]$ and $[q_1,q_2]$ and thus

$$\Omega_{[p,q_1]} \oplus \Omega_{[q_1,q_2]} = \Omega_{[p,q_2]}.$$  

(8)

Furthermore,

$$X^n(p,q_1)X^n(q_1,q_2)|k\rangle|b\rangle = |k\rangle|b \oplus \Omega_{[p,q_1]}(k) \oplus \Omega_{[q_1,q_2]}(k)\rangle$$  

(9)

$$= |k\rangle|b \oplus \Omega_{[p,q_2]}(k)\rangle.$$  

(10)

The above properties allow us to reduce $X^n(p,q)$ into cascaded entanglers.

Given some $p < q$, let $\underline{m} = \lceil \log_2(q-p) \rceil$ and $\overline{m} = \underline{m} + 1$. Then there are two ways to recursively split the $X^n(\ell)$ operator:

1. $X^n(\ell) = X^n(2^n - \ell, 2^n - 2^\overline{m})X^n(2^\overline{m}),$

2. $X^n(\ell) = X^n(2^\underline{m})X^n(2^n - 2^\overline{m}, 2^n - \ell).$

Both follow directly from Proposition [4]. As per Observation [2], the first split can be written in the form

$$X^n(\ell) = P_1 X^n(\ell - 2^\overline{m}) P_1 X^n(2^\overline{m})$$

and the second in the form

$$X^n(\ell) = X^n(2^\underline{m}) P_2 X^n(2^\overline{m} - \ell) P_2,$$

where $P_1, P_2$ are algorithmically computable Pauli involutions.
Observation 5. Given $0 \leq m \leq n$, $X^n(2^m)$ is equivalent to a tensor product of $\Lambda^{n-m}(X)$ and identity operator on $m$ qubits.

Indeed, by direct inspection of the activation states it is easy to see that $X^n(2^m)$ is the inversion of the last $(n+1)$-st qubit controlled by the first $(n-m)$ qubits, while the qubits with numbers $n-m+1, \ldots, n$ are not affected by the operator and have no control on it.

Decomposition of an $X^n(\ell)$ into a Clifford+$T$ circuit with near-optimal $T$-count is based on the result of Ref. [6]: as $k$ is growing, $\Lambda^k(X)$ can be effectively implemented by a Clifford+$T$ network with a $T$-count in $O(k)$. Should this $T$-count have been perfectly proportional to $k$, the algorithm proposed below would achieve absolute optimality. However we can only establish that the $T$-count required for implementing $\Lambda^k(X)$ is less than or equal to some $\kappa k$. (Based on Lemma 7.2 in [7] the coefficient $\kappa$ may be upper-bounded by 8 assuming available ancillae qubits for decomposition of multi-controlled-NOTs.)

Algorithm 2 recursively splits $X^n(\ell)$ into a sequence of gates of the form $X^n(2^m)$ with the possible interleaving local $X$ gates. Since at every level there is a choice of whether to round up or round down $\ell$ to the closest power of 2, the algorithm explores, exhaustively, both choices.

---

**Algorithm 2 CED1-full: Full-recursive decomposition of $X^n(\ell)$**

**Require:** $n, \ell$

1: if $\ell = 0$ then
2: return $\{I_{n+1}, 0\}$
3: end if
4: $m \leftarrow \lfloor \log_2(\ell) \rfloor$
5: if $\ell = 2^m$ then
6: return $\{\text{Library circuit for } \Lambda^{n-m}(X) \otimes I(m), \text{Library } T\text{-count estimate} \}$
7: end if
8: $\{c_1, t_1\} \leftarrow \text{CED1f}(n, \ell - 2^m)$
9: $\{c_2, t_2\} \leftarrow \text{CED1f}(n, 2^m)$
10: $\{c_3, t_3\} \leftarrow \text{CED1f}(n, 2^{m+1} - \ell)$
11: $\{c_4, t_4\} \leftarrow \text{CED1f}(n, 2^{m+1})$
12: if $t_1 + t_2 < t_3 + t_4$ then
13: return $\{c_1, c_2, t_1 + t_2\}$
14: end if
15: return $\{c_3, c_4, t_3 + t_4\}$

---

Lines 5, 6 address the case when $\ell$ is an exact power of 2. This is a termination leaf for the recursion, since library circuits for this case are available from [6, 7]. In Lines 8, 9 the operator is split recursively by round down to the closest power of 2, and in Lines 10, 11 this is done by rounding up. The outcomes of the two choices are then compared and the best outcome is returned. The runtime complexity of Algorithm 2 is exponential in the bit size of $\ell$ and therefore is in $O(\ell)$.

This complexity could be problematic for larger values of $\ell$, therefore we also develop a tail-recursive version, given in Algorithm 3 with runtime complexity in $O(\log(\ell))$. Here the decision between round up and round down is based on the empirical threshold as shown in Line 8, without exhaustive analysis of the post-split outcomes. Although we have not yet proven the equivalence of Algorithms 2 and 3 our numerical experiments indicate that the ultimate $T$-count of the circuits generated by these two algorithms is been identical. The threshold in Line 8 of Algorithm 3 is designed such that both of the tail recursion arguments are smaller than $\ell/2$. Therefore the depth of the tail recursion never exceeds $\log_2(\ell)$.
Algorithm 3 CED1-tail: Tail-recursive decomposition of $X^n(\ell)$

**Require:** $n, \ell$

1. if $\ell = 0$ then
2. \hspace{1em} return \{$I_{n+1}, 0$\}
3. end if
4. $m \leftarrow \lfloor \log_2(\ell) \rfloor$
5. if $\ell = 2^m$ then
6. \hspace{1em} return \{$\text{Library circuit for } \Lambda^{n-m}(X) \otimes I_m, \text{Library T-count estimate}$\}
7. end if
8. if $\ell < \frac{3}{2} 2^m$ then
9. \hspace{1em} \{$c_1, t_1$} \leftarrow \text{CED1t}(n, \ell - 2^m)
10. \hspace{1em} \{$c_2, t_2$} \leftarrow \text{CED1t}(n, 2^m)
11. \hspace{1em} return \{$c_1 c_2, t_1 + t_2$\}
12. end if
13. \{$c_1, t_1$} \leftarrow \text{CED1t}(n, 2^{m+1} - \ell)
14. \{$c_2, t_2$} \leftarrow \text{CED1t}(n, 2^{m+1})
15. return \{$c_1 c_2, t_1 + t_2$\}

---

**Example 6.** Full recursive expansion of $X_6(23)$ using Algorithm 3 yields

$$X_6(23) = X_6(32) P_1 X_6(8) P_1 P_2 X_6(1) P_2,$$

where $P_1, P_2$ are appropriate tensor products of $X$ gates. The high-level representation of the circuit is shown in Fig. 5.

For completeness, we note that for an implementation of $X_6(1) = \Lambda^6(X)$ it is advisable to use an additional ancilla qubit, and thus to implement it as an 8-qubit circuit. As per [7], Lemma 7.4, the implementation requires $8 \times (8 - 5) = 24$ three-qubit Toffoli gates and so the required $T$-count is bounded by $24 \times 7 = 168$. With simple cancellation of some Pauli gates, a near-optimal circuit for implementing $V_6(\varphi, 23)$ would be similar to the one presented in Fig. 6.

---

C. The CED2($x, r, x$) procedure.

The CED2($x, r, x$) block referenced in Algorithm 1 can be implemented by concatenating a circuit $c$ for the entangler $x$ in its first argument obtained via Algorithm 3 with a Clifford+$T$ circuit for the rotation $r$ and then with the mirror image of the circuit $c$, as we have just done for the $V_6(\varphi, 23)$ in the above example. However, in most cases this is not going to be optimal and the eventual synthesis method for $V(\phi, \ell)$ would benefit from further optimization of its entangling part.

Thus, it is well-known that a pair of matching identical multi-controlled-NOTs (such as those in Figure 1) can be implemented at less than twice the cost of a single multi-controlled-NOT with the same number of controls since instead of implementing each of the multi-controlled-NOTs faithfully we can implement them up to mutually-canceling local phases of $-1$ at a lower cost. As evidence, Figure 2 shows the exact Clifford+$T$ decomposition of the Toffoli gate using 7 $T$ gates. Note that we have used the identity $HZH = X$ to rewrite the Toffoli as a double-controlled-$Z$ gate, which is diagonal and hence has an exact decomposition in the Clifford+$T$ basis. Figure 3 shows an exact decomposition of two paired Toffoli gates with a phase gate on the target qubit. We use the Clifford+$T$ decomposition of the Toffoli shown in Figure 2. The first two qubits in this example have a sequence of gates which are adjoints of one another, thus if we carry through all commutations to move the gates on the top two qubits towards the center, we find that not only
Figure 2: An exact decomposition of the Toffoli gate on the Clifford+T basis. We use the fact that $HZH = X$, where $H$ is the Hadamard gate and $Z$, $X$ are the standard Pauli gates.

Figure 3: An exact decomposition of two paired Toffoli gates with a phase gate on the target qubit.

do they mutually cancel one another, but all commutations in between different CNOT operators will also mutually cancel. Independently they would cost 7 $T$ gates each for a total of 14 per pair, however in this paired configuration a pair of matched Toffoli entanglers costs only 8 $T$ gates.

Another type of simplification can occur as part of the multi-controlled entangler decompositions proposed in Ref. [7] that can be used to calculate decompositions of the entanglers like those in Figure 5. A typical pattern that appears in these decompositions is shown in Figure 4. In this case, there is no phase gate separating the target qubits of a pair of matched entanglers, and as a result a significant simplification occurs when we pair two toffoli’s and examine them under the Clifford+T decomposition of Figure 2. As it turns out the two paired entanglers in this configuration require only 2 $T$ gates each, resulting in a total of 4 $T$ gates for the pair. Several other types of simplifications may occur in these types of paired entangler scenarios. Overall, finding the true optimum for entanglement cost is beyond the scope of this paper and merits future work.

V. NUMERICAL RESULTS

We apply Algorithm 1 to a series of randomly generated phase-sparse, $n$-qubit diagonal unitary operators, for $n$ in $\{4, 5, 6, 7, 8\}$. For each qubit count $n$, we sweep the phase context size $k$ from 1 to $2^n$. Each DU was generated by first generating a phase context, that is, a random set of $k$ distinct phases between 0 and $2\pi$. Then a set of $2^n$ phases are generated by sampling at random from the generated phase context. A phase context decomposition is performed over the generated diagonal operator and the resulting cascaded entangler cost is calculated. Finally the asymptotic cost is calculated using Eqn (4) for a range of $\varepsilon$ between $10^{-3}$ and $10^{-30}$ with $C_0 = 1$. This is the smallest $C_0$ known that can be achieved at the cost of one additional ancilla qubit by using either probabilistic circuits with fallback as described in [8] or repeat-until-success circuits as described in [12].

We compare our decomposition algorithm to the following decomposition scheme:
Figure 4: Similar to the case in Figure 3, we pair two Toffoli gates with a third Toffoli in between. This type of pattern appears often in the multi-controlled entangler decomposition proposed in [7].

Figure 5: A circuit for $X_{6}(23)$.


2. Perform independent $\varepsilon$-approximation of all the single-qubit rotations generated in Step 1.

Each single-qubit unitary is then assumed to be further decomposed into a chosen basis by a given algorithm $A$. Algorithms given in Refs. [8, 10–12], for example, may be considered.

We find that a Walsh-based exact decomposition rarely takes advantage of context sparseness when given a general context-sparse diagonal unitary and tends to generate $2^n$ distinct single-qubit rotations (except in a small random fraction of cases). In contrast, our cascaded entangler method always takes full advantage of context sparseness. As a result, the log scatter plot of the circuit

Figure 6: A circuit for $V_6(\varphi, 23)$. 
The joint scatter plot for \( n = 7 \) and \( \varepsilon = 10^{-20} \) is shown in Fig. 7. The blue curve represents the cost of the Walsh-based circuits, while the green curve represents the cost of circuits with cascaded entanglers. For very sparse contexts (\( k \) in single digits), the circuits with cascaded entanglers tend to be orders of magnitude better compared to their exact decomposition counterparts. It is also apparent from the plot that the former circuits become more expensive than the latter after \( k \) crosses \( 2^6 = 64 \). Overall, our numerical experiments suggest that there exists a significant range of \( \varepsilon \) and \( k \) for which the decomposition of the diagonal unitary benefits from the use of cascaded entanglers as compared to exact methods, and for very sparse contexts the reduction in the \( T \)-count is exponential in \( n \).

Figures 9(a) and 9(c) present surface plots of the \( T \)-count given \( k \) and \( \varepsilon \) for the two methods in the case of 6 qubits and 8 qubits, respectively. The blue surfaces represent the \( T \)-count of circuits obtained by using the exact method of [3], and the green surface represents the \( T \)-count of the circuits obtained with the cascaded entangler algorithm. There exists a crossover line at which the cascaded entangler approach begins to accumulate a large \( T \)-cost in the entanglers. The cascaded entangler approach is seen to provide a lower \( T \)-count for \( k \ll 2^n \) at coarse (larger) \( \varepsilon \) and \( k \) less than about \( 2^{n-1} \) for finer (smaller) \( \varepsilon \). For the case of \( n = 1 \) there is no need for entanglement and hence the implementation is trivial. For \( n = 2, 3 \) the comparison between the two methods is less dramatic, but still present. The plots in Figures 8(a) and 8(b) demonstrate that the cascaded entangler approach (green) can still have a significant cost benefit when the rotation angles are in general position and the number of qubits is small.

The crossover line can more clearly be seen in Figures 9(b) and 9(d), which are top-down projections of the surface plots on the left. Green indicates where the circuits utilizing cascaded entanglers have a lower \( T \)-count and blue when the exact method has a lower \( T \)-count. While we do not have a closed form equation for the crossover line at this time, as a first approximation it appears to be closely fit by a log-hyperbolic curve of the form

\[
k = 2^n (\lambda n + \log_{10}(1/\varepsilon))/(\kappa n + \log_{10}(1/\varepsilon)),
\]

where \( \kappa \) is the leading cost coefficient on the implementation \( T \)-count for the multi-controlled Toffoli gate and \( \lambda \) is some empirical fitting parameter likely in [1, 2]. One might alternatively describe the
Figure 8: Scatter plots comparing T-count at $\varepsilon = 10^{-20}$ between the exact and cascaded entangler methods on a set of randomly generated (a) 2-qubit and (b) 3-qubit diagonal unitaries.

crossover in terms of a phase-context sparseness threshold $t = k/2^n$ which then should be fit as

$$t = (\lambda n + \log_{10}(1/\varepsilon)) / (\kappa n + \log_{10}(1/\varepsilon)).$$

We conclude from the numerical simulations that there exists a region for implementing phase-sparse diagonal unitary operators with a significantly lower T-count than exact methods. The position of the boundary is not fixed, but depends on the number of qubits in the system. As the number of qubits increases, the number of phase sparse diagonal unitaries that are more efficiently implemented using cascaded entanglers grows as well.

VI. CONCLUSIONS AND FUTURE WORK

We have analysed the cost of approximation of diagonal unitaries over the Clifford+$T$ basis. We find that by distinguishing between a phase sparse and phase dense diagonal unitary, the T-count and cost of approximation can be significantly reduced. Namely, if the target diagonal unitary is phase sparse, then the cost is dominated by the cost of approximating single-qubit rotations. We have designed a circuit decomposition framework and have shown that, given a diagonal unitary where phase factors on the diagonal belong to an alphabet with $k$ distinct values, the asymptotic implementation cost over the Clifford+$T$ basis is dominated by the $(k - 1) C_0 \log_2(1/\varepsilon)$ term when the desired approximation precision $\varepsilon$ tends to zero (and where $C_0$ is the known cost factor in the approximation cost of a general single qubit rotation). Intuitively our results suggest that when a target unitary can be described by a relatively small number of parameters, then that unitary can be approximated at a cost asymptotically proportional to the number of parameters and $\log_2(1/\varepsilon)$.

In the extreme case of an $n$-qubit diagonal with $k = 2$ distinct phase factors in general position, previously known decomposition methods tend to generate circuits with $\Omega(2^n)$ different single-qubit rotations. Our method generates a circuit with only one rotation and some cascaded entanglers, which allows us to claim asymptotic improvement by a factor of $\Omega(2^n/n)$.

While we have focused on asymptotic optimality of the decomposition of circuits, an important open topic for further research is establishing exact bounds for the complexity of cascaded entanglers. In the broader context of quantum circuit decompositions, the distinction between phase rotation and entanglement should be applied to the synthesis of circuits for other classes of content-sparse gates, beyond the class of diagonal unitaries. Block-diagonal unitaries would be a likely candidate for the next stage of this research.
Figure 9: Crossover plots for a 6-qubit (a,b) and 8-qubit (c,d) diagonal unitary decomposed using exact Walsh decomposition (blue) and cascaded entangler decomposition (green): (ac) surface plot; (bd) density plot in the $k, \varepsilon$ plane. The density plots are top-down projections of the surface plots; the cascaded entangler method achieves lower overall cost in the green region, and higher cost in the blue region. The $T$-count is calculated using Eqn (4) with $C_0 = 1$.

ACKNOWLEDGMENTS

The Authors wish to thank Nathan Wiebe for valuable tips and references, we also thank Martin Roetteler and Dave Wecker for useful discussions.