






Optimal joint cutting of two-qubit rotation gates

Christian Ufrecht ^{*}, Laura S. Herzog , Daniel D. Scherer , Maniraman Periyasamy, Sebastian Rietsch, Axel Plinge , and Christopher Mutschler 

Division Positioning and Networks, Fraunhofer IIS, [Fraunhofer Institute for Integrated Circuits IIS](https://www.fraunhofer-iis.de), Nuremberg, Germany



(Received 29 February 2024; accepted 6 May 2024; published 30 May 2024)

Circuit cutting, the partitioning of quantum circuits into smaller independent fragments, has become a promising avenue for scaling up current quantum-computing experiments. Here, we introduce a scheme for joint cutting of two-qubit rotation gates based on a virtual gate-teleportation protocol. With that, we significantly lower the previous upper bounds on the sampling overhead and prove the optimality of the scheme. Furthermore, we show that no classical communication between the circuit partitions is required. For parallel two-qubit rotation gates we derive an optimal ancilla-free decomposition, which includes CNOT gates as a special case.

DOI: [10.1103/PhysRevA.109.052440](https://doi.org/10.1103/PhysRevA.109.052440)

I. INTRODUCTION

Current quantum computing hardware faces serious limitations, such as low qubit numbers and high susceptibility to noise. As a result, quantum hardware will likely be unable to execute algorithms with provable speedup like Shor's [1] or Grover's [2] algorithm in the near future. On the other hand, experiments with heuristic quantum algorithms have been conducted on a small number of qubits, primarily applied to toy problems from the fields of, e.g., optimization [3,4], machine learning [5,6], and chemical simulations [7,8]. However, scaling up both problem sizes and the number of qubits is crucial for assessing the potential usefulness of quantum computers. Even though recent progress [9] suggests that quantum hardware based on superconducting qubits might be on the verge of entering a so-called *utility regime* where the first practically interesting problems might be approached, to achieve relevant hardware size, modular quantum computing has been proposed [10,11]. In this paradigm, multiple quantum computing platforms are interconnected through quantum links possibly enhanced by virtual entanglement distillation [12,13]. Until a higher technology readiness level is reached, circuit cutting could serve as a useful approach, replacing quantum links with classical links and a postprocessing step.

Given a unitary quantum channel \mathcal{V} , circuit cutting describes the method of decomposing the channel as

$$\mathcal{V} = \sum_i a_i \mathcal{F}_i, \quad (1)$$

with real coefficients a_i , to reduce the circuit size or the impact of noise.

Circuit cutting can be categorized into two methods. *Wire cutting* [14–19] effectively corresponds to the cutting of horizontal empty qubit wires which are modeled mathematically by the identity channel. Consequently, wire cutting is a decomposition of this channel into measure-and-prepare

channels. The second method, on which this work will focus, is known as *gate cutting* [20–24]. Here, \mathcal{F}_i are elements of $\text{LOCC}(A, B)$, that is local operations on two partitions A and B of the quantum circuit and classical communication between them. Since \mathcal{F}_i are local operations, no entanglement is created by the channel, but it is effectively simulated as described below. If all wires or gates that connect partitions A and B are cut, the subcircuits become independent and can be run on two quantum computers connected by only classical communication links. Most circuit-cutting research so far has focused on applications to algorithms in which the output is an expectation value of an observable. To evaluate Eq. (1) in this context, we first define $\kappa = \sum_i |a_i|$ and the probability distribution $p_i = |a_i|/\kappa$. Next, Eq. (1) is rewritten as

$$\mathcal{V} = \kappa \sum_i p_i \text{sgn}(a_i) \mathcal{F}_i \quad (2)$$

and is evaluated via Monte Carlo sampling. At each experimental shot we select the i th channel \mathcal{F}_i with probability p_i and evaluate the circuit with \mathcal{V} replaced by \mathcal{F}_i . The measurement outcome is weighted by $\kappa \text{sgn}(a_i)$, and the mean over many runs produces an unbiased estimate for the expectation value of the observable. This sampling procedure is referred to as quasiprobability sampling [25] because of the appearance of the sign of a_i in Eq. (2), which is referred to as a quasiprobability decomposition (QPD).

Quasiprobability sampling incurs a sampling overhead, the factor of more samples required to estimate the expectation value of an observable with the same accuracy, as with respect to the original uncut circuit. This sampling overhead has been shown to be κ^2 [25,26] and originates from the factor of κ in front of the sum in Eq. (2).

Note the similarity of the approach to probabilistic error cancellation, in which the (possibly unphysical) inverse of a noisy quantum channel is decomposed into physical operations [27].

The application of circuit cutting to sampling tasks was studied initially in Ref. [17] and was recently studied more

^{*}christian.ufrecht@iis.fraunhofer.de

extensively in Ref. [28], with the result that circuit cutting can also be meaningful in this situation.

When gates or wires are cut individually, the overall sampling overhead increases exponentially in the number of gates and wires cut. This strong increase of sampling overhead renders circuit cutting prohibitively expensive when the cutting location in a circuit and the cutting scheme are not carefully selected. The task is, therefore, to find decompositions of gates or collections of gates with minimal κ , which we will refer to as γ .

Reference [23] employed CNOT-gate teleportation [29–31] consuming one Bell state initially connecting the two partitions per CNOT gate. The optimal QPD with $\gamma = 3$ of the Bell-state density matrix then translates into the optimal decomposition of the gate. The crucial insight of Ref. [23] was that the joint QPD of n Bell states required for the teleportation of n CNOT gates can be constructed more efficiently than the individual decomposition of the gates. As a result, the γ parameter is reduced from $O(3^n)$ to $O(2^n)$, but at the cost of one ancilla qubit per partition and gate. In this cutting scheme, $\mathcal{F}_i \in \text{LOCC}(A, B)$, and two-way classical information is shared between the partitions. The protocol was extended to general Clifford gates and later also to wire cutting [15]. Recently, it was shown that optimal cutting of parallel wires is possible without ancilla qubits [16].

In this work, we extend previous proposals to the joint cutting of non-Clifford two-qubit rotation gates. We significantly simplify existing methods and, as a result, enable implementation on current noisy intermediate-scale quantum hardware. Two-qubit rotation gates play a crucial role, for example, in the quantum approximate optimization algorithm [4] and in the simulation of spin systems. In Sec. II, we introduce a virtual teleportation scheme that effectively consumes less than one entanglement bit (ebit). After explaining the protocol for a single gate instance, we generalize the scheme to the joint cutting of n two-qubit rotation gates and demonstrate the optimality of the derived circuit-cutting method. With optimality, we refer to the circuit-cutting scheme based on quasiprobability sampling with the minimal possible sampling overhead. A two-qubit rotation gate with rotation angle θ is defined as

$$R_{zz}(\theta) = \cos(\theta/2)\mathbb{I} \otimes \mathbb{I} - i\sin(\theta/2)Z \otimes Z, \quad (3)$$

with \mathbb{I} being the single-qubit identity and Z being the Pauli Z matrix. The γ parameter for this gate is [21,23]

$$\gamma = 1 + 2|\sin(\theta)|. \quad (4)$$

As we will show, when n instances of this gate are jointly cut, we find the reduced effective γ parameter per gate

$$\gamma = 1 + |\sin(\theta)| \quad (5)$$

asymptotically as n approaches infinity, which matches the lower bound. Our scheme offers several improvements over previous work. First, it eliminates the need for classical communication between the partitions, and consequently, no real-time feedback is required. This is in contrast to the gate-teleportation protocol used in Ref. [23], where local correction operators conditioned on intermediate measurement results are necessary. Moreover, as explained in more detail in Sec. II B, our protocol can be executed sequentially on

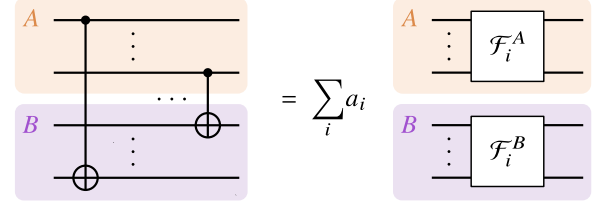


FIG. 1. Cutting of parallel CNOT gates. Consider two partitions, A and B , of a quantum circuit connected by n CNOT gates. In the case of gates that can be executed at the same time slice of the circuit (parallel gates), we present a joint optimal ancilla-free decomposition with coefficients a_i and unitary or measurement operations \mathcal{F}_i^A and \mathcal{F}_i^B . The sampling overhead is $O(\gamma^2)$ where $\gamma = O(2^n)$. This is a strong improvement compared to cutting n CNOT gates independently for which $\gamma = O(3^n)$. Note that the equality here has to be understood on the superoperator rather than on the gate level.

the same quantum computer, unlike schemes in which two-way classical communication necessitates the use of two separate quantum computers in parallel. Second, it was recently emphasized that the number of operations \mathcal{F}_i can be the limiting factor for the execution of circuit-cutting protocols [16]. We derive a quasiprobability decomposition of the entanglement resource states with an exponentially reduced number of elements. Third, in Sec. III, we show that parallel gates can be cut without the need for ancilla qubits. Parallel gates refer to gates that can be executed within the same time slice of a circuit, as shown in Fig. 1. Note that controlled rotation gates and CNOT gates (for $\theta = \pi/2$) are equivalent to two-qubit rotation gates up to local operations. Thus, all the results of our work also apply to these types of gates.

II. VIRTUAL GATE TELEPORTATION

In this section, we aim to find an optimal joint decomposition of two-qubit rotation gates based on gate teleportation. A teleportation scheme for a two-qubit rotation gate $R_{zz}(\theta)$ [31] is shown in Fig. 2. It consumes one initial Bell state (shown as a wiggly line) prepared on two ancilla qubits. These qubits are subsequently measured in the computational basis, and the outcomes $k, l \in \{0, 1\}$ are communicated between partitions A and B via classical communication to select the local Pauli correction operators X^k and Z^l . For $\theta = \pi/2$ the

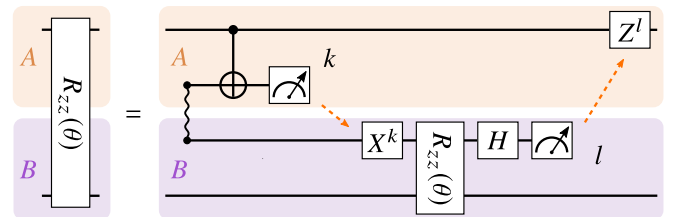


FIG. 2. Teleportation protocol for a two-qubit rotation gate $R_{zz}(\theta)$ [31]. The scheme consumes an initial Bell state (wiggly line) as an entanglement resource between partitions A and B and requires two-way classical communication visualized by the dashed arrows. The measurements are performed in the computational basis with the outcomes $l, k \in \{0, 1\}$. Here, H is the Hadamard gate, and Z and X are the Pauli operators.

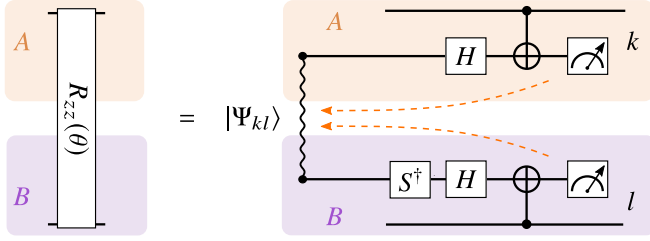


FIG. 3. Virtual gate teleportation. As explained in the text, for the state shown in Eq. (6), the correct gate is teleported only for the measurement results $k = l$, resulting in suboptimal sampling overhead when discarding other outcomes. Since we are interested in only a teleportation scheme together with a quasiprobability simulation of the initial state, we show that in this case, the result $k \neq l$ can be corrected in the final classical postprocessing step. This effectively corresponds to the preparation of the initial state $|\Psi_{kl}\rangle$ defined in Eq. (11), which depends on the outcome of the measurement performed later in time, hence the term *virtual teleportation*. Here, S is the phase gate.

two-qubit rotation gate is equivalent to a CNOT gate up to local operations. In this case, simulating the initial Bell state by a quasiprobability distribution with $\gamma = 3$ translates to an optimal cutting scheme for a CNOT gate (also $\gamma = 3$) [23]. When employing the teleportation scheme for n CNOT gates together with a joint quasiprobability decomposition of the n Bell states, the γ parameter behaves submultiplicatively, reducing from $O(3^n)$ to $O(2^n)$.

When attempting to achieve a similar reduction for joint cutting of two-qubit rotation gates, we encounter the following difficulty: A decomposition of a two-qubit rotation gate based on the teleportation scheme would be suboptimal since it always consumes one Bell state, resulting in $\gamma = 3$, while the γ parameter of a two-qubit rotation gate in Eq. (4) is smaller for most values of θ . On the other hand, a deterministic gate-teleportation protocol consuming less than one ebit does not exist [32,33], only probabilistic implementations were proposed [34–36], which would lead to suboptimal sampling overhead. As a result, a naive generalization of Ref. [23] is impossible.

A. Single gate instance

In the following, we introduce the alternative scheme depicted in Fig. 3 for which we prepare the initial state $|\Psi_{kl}\rangle$, to be specified below, on the ancilla qubits. As a first step, we choose the initial state

$$|\Psi\rangle = \cos(\theta/2)|00\rangle + \sin(\theta/2)|11\rangle. \quad (6)$$

As we detail in Appendix A, an optimal QPD of a pure-state density matrix can be readily calculated based on the robustness of entanglement measure [23,37]. The γ parameter is calculated as

$$\gamma = 2\left(\sum_j c_j\right)^2 - 1, \quad (7)$$

where c_j are the Schmidt coefficients of the state. Consequently, for the state shown in Eq. (6), we obtain $\gamma = 2[|\cos(\theta/2)| + |\sin(\theta/2)|]^2 - 1$, matching Eq. (4). However,

it is straightforward to convince ourselves that the correct gate is teleported only for $k = l$ and its Hermitian conjugate is teleported for $k \neq l$ [35]. On the one hand, local operators on partitions A and B conditioned on the measurement results to achieve a deterministic protocol do not exist [32,33]. On the other hand, a *repeat-until-success* version that disregards all measurement outcomes with $k \neq l$ would lead to the suboptimal sampling overhead $4\gamma^2$. Note, however, that we are not interested in deriving an actual deterministic gate-teleportation protocol but in a QPD to simulate the two-qubit rotation gate with optimal sampling overhead.

In Appendix A, we derive an alternative optimal QPD for a pure-state density matrix compared to the one presented in Ref. [37]. Applied to Eq. (6), we find

$$|\Psi\rangle\langle\Psi| = \cos^2(\theta/2)|00\rangle\langle 00| + \sin^2(\theta/2)|11\rangle\langle 11| \quad (8)$$

$$+ \sin(\theta)(\sigma^+ - \sigma^-). \quad (9)$$

The explicit form of the separable states σ^+ and σ^- is specified in Appendix A. To see that the decomposition is optimal, we sum over the absolute of the prefactors, finding again Eq. (4). The states σ^+ and σ^- are independent of θ , which is the crucial property to perform our protocol as follows: As mentioned above, for $k \neq l$, the Hermitian conjugate of the two-qubit rotation gate is teleported, which corresponds to $\theta \rightarrow -\theta$ in Eqs. (8) and (9). Because $\sin(-\theta) = -\sin(\theta)$, we could equally choose a minus sign in front of Eq. (9), keep results with $k \neq l$, and disregard results with $k = l$. To make use of both situations, we define the QPD

$$|\Psi_{kl}\rangle\langle\Psi_{kl}| = \cos^2(\theta/2)|00\rangle\langle 00| + \sin^2(\theta/2)|11\rangle\langle 11| \\ + (-1)^{k+l}\sin(\theta)(\sigma^+ - \sigma^-), \quad (10)$$

in which the signs of the quasiprobabilities depend on the measurement outcomes in the form of the factor $(-1)^{k+l}$. To execute the protocol, we proceed as follows: To estimate an expectation value experimentally, one typically defines a postprocessing function $f(s) \in [-1, 1]$ [14] as a function of the bit string s observed in each experimental run and then takes the sample mean over many runs. As explained in the Introduction, we perform Monte Carlo sampling as, for example, explained in Ref. [24]. For each run, we prepare one of the states in Eq. (10) sampled according to the probability given by the absolute of the prefactor divided by γ . In case we initialize the ancilla qubits with one of the states constituting σ^+ or σ^- , we accumulate $\pm(-1)^{k+l}\gamma f(s)$ in the sample mean, where the sign depends on the outcome of the measurements on the ancilla qubits. We will refer to our protocol as *virtual gate teleportation* since an actual entanglement resource state is not physically prepared but is only simulated by a quasiprobability decomposition. In this respect, the protocol effectively makes use of the state

$$|\Psi_{kl}\rangle = \cos(\theta/2)|00\rangle + (-1)^{k+l}\sin(\theta/2)|11\rangle, \quad (11)$$

which depends on the outcomes of the measurement performed later in time.

B. Joint cutting of multiple gates

In this section, we generalize the virtual protocol to the joint cutting of n two-qubit rotation gates at arbitrary positions

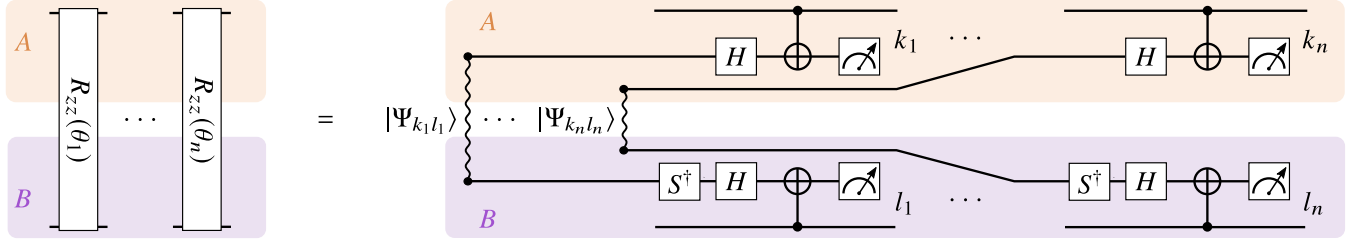


FIG. 4. Generalization of the virtual teleportation protocol from the single gate instance shown in Fig. 3 to n two-qubit rotation gates. Each gate requires one ancilla qubit per partition on which we effectively prepare the states $|\Psi_{k_s l_s}\rangle$ for $s = 1, \dots, n$ using a quasiprobability decomposition. The result of the measurements on the ancilla qubits determines the sign of the quasiprobabilities in the classical postprocessing step.

in the circuit. When n gate instances are cut independently, the overall γ parameter defined here as $\gamma_{\text{ind.}}^{(n)}$ behaves multiplicatively, that is,

$$\gamma_{\text{ind.}}^{(n)} = \prod_{s=1}^n [1 + 2|\sin(\theta_s)|], \quad (12)$$

where we used Eq. (4). In order to investigate whether joint decomposition can be done more efficiently, the single-instance virtual teleportation protocol is generalized in the following to n gates, as shown in Fig. 4. In Fig. 4 we view the n two-qubit states $|\Psi_{k_s l_s}\rangle$ for $s = 1, \dots, n$ on the ancilla qubits as a $2n$ -qubit state $|\Psi_{k_1 l_1 \dots k_n l_n}\rangle$. Again, a specific choice of k_s, l_s for $s = 1, \dots, n$ defining the resource state translates into only the signs of the quasiprobabilities of its QPD. In turn, we can choose $k_s = l_s = 0$ for all s of the resource state, but correct unwanted measurement results in the postprocessing step.

Using Eq. (11) and the qubit arrangement shown in Fig. 4, we find

$$|\Psi_{k_1 l_1 \dots k_n l_n}\rangle = \sum_{j \in \{0,1\}^n} c_j |j_1 \dots j_n\rangle |j_n \dots j_1\rangle, \quad (13)$$

with $c_j = \prod_{s=1}^n c_{j_s}$ and

$$c_{j_s} = \begin{cases} \cos(\theta_s/2) & \text{for } j_s = 0, \\ (-1)^{k_s + l_s} \sin(\theta_s/2) & \text{for } j_s = 1. \end{cases} \quad (14)$$

In Eq. (13), the first (second) ket describes the state on the ancilla qubits of partition A (B). If we redefine, e.g., the first ket on the right-hand side of Eq. (13) by absorbing the sign of c_j , this equation is the Schmidt decomposition of $|\Psi_{k_1 l_1 \dots k_n l_n}\rangle$, allowing us to calculate $\gamma_{\text{joint}}^{(n)}$ for the QPD as

$$\gamma_{\text{joint}}^{(n)} = 2 \left(\sum_{j \in \{0,1\}^n} |c_j|^2 \right) - 1 \quad (15)$$

$$= 2 \prod_{s=1}^n [1 + |\sin(\theta_s)|] - 1 \quad (16)$$

$$< \gamma_{\text{ind.}}^{(n)}. \quad (17)$$

In Appendix A, we explicitly state an optimal quasiprobability decomposition for any pure-state density matrix. Applied to Eq. (13), the measurement-dependent signs in Eq. (14) appear only as the signs of the quasiprobabilities, again allowing correction in the classical postprocessing step. The γ parameter in Eq. (16) is, in general, significantly smaller than the γ parameter for independent cuts. Therefore, we have shown that a decomposition for multiple instances of two-qubit

rotation gates with submultiplicative behavior of the γ parameter exists. In Appendix B, we prove that $\gamma_{\text{joint}}^{(n)}$ also is a lower bound for the optimal decomposition of n two-qubit rotation gates. Consequently, joint optimal virtual gate teleportation also translates into optimality of the gate decomposition.

Note that no classical communication is required between the partitions [38]. Joint gate cutting without the need for classical communication significantly alleviates implementation on current hardware for two reasons. First, no real-time feedback for local operations conditioned on measurement results is needed, distinct from conventional gate-teleportation protocols such as that shown in Fig. 2. Second, if all gates connecting the two partitions are cut, the severed subcircuits can be executed sequentially on the same hardware. Indeed, given a decomposition like that in Eq. (1), we first determine the number of shots N required to achieve a certain accuracy of the expectation value estimate with high probability. Next, we determine how many times the circuit with channel \mathcal{F}_i needs to be executed by sampling N times from the probability distribution with $p_i = |a_i|/\kappa$ and counting the number of times i is realized. If the decomposition does not require classical communication, $\mathcal{F}_i = \mathcal{F}_i^A \otimes \mathcal{F}_i^B$, and if \mathcal{F}_i on one partition involves measurements, the i th channel on the other partition is independent of the measurement result. If all gates connecting the two partitions are cut, the circuits can be evaluated sequentially and independently of each other since \mathcal{F}_i^A and \mathcal{F}_i^B are completely independent. After all circuits have been executed, the bit strings that were measured on both partitions for the i th channel are concatenated and passed to the postprocessing function. This procedure stands in stark contrast to protocols employing two-way classical communications in which two quantum devices connected by classical links must operate in parallel. This results from the structure of \mathcal{F}_i when allowing classical communication. Here, we allow operations that may depend on the outcome of a measurement on the other partition. Since measurement results are probabilistic, the number of times each channel has to be applied cannot be calculated prior to the experiment.

Last, we mention again that our protocol contains the joint cutting of n CNOT gates as a special case with optimal $\gamma_{\text{joint}}^{(n)} = 2^{n+1} - 1$ [23].

C. Multiqubit rotation gates

Finally, we consider multiqubit Z rotation gates. A multiqubit rotation gate can be written as a two-qubit rotation gate sandwiched by ladders of CNOT gates [39], as shown

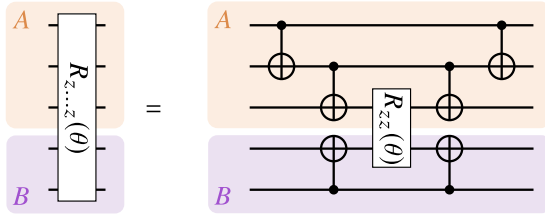


FIG. 5. A multiqubit rotation gate is equivalent to a two-qubit rotation gate up to local operations (with respect to the two partitions A and B). Consequently, the γ parameter of this gate equals that of the two-qubit rotation gate.

in Fig. 5. The task of cutting a multiqubit Z rotation gate then reduces to the cutting of a two-qubit rotation gate. The resulting decomposition is optimal. This can be seen by noting that if the γ parameter of the multiqubit rotation gate were smaller than that of a two-qubit rotation gate, we could use the former to construct a decomposition of the latter with smaller γ , leading to a contradiction.

III. CUTTING PARALLEL GATES WITHOUT ANCILLA QUBITS

Interestingly, as we will show in the following, there is an ancilla-free joint decomposition of parallel gates. By “parallel gates,” we refer to gates that can be executed at the same time slice of the circuit, as shown in Fig. 1 for the case of CNOT gates, for example. In Appendix C, we start from the protocol in Fig. 4 and subsequently show that it can be reduced to an ancilla-free decomposition in terms of local operations on partitions A and B . Remarkably, again, no classical communication between the partitions is required, and optimality is preserved. The explicit formula for the decomposition is shown in Eq. (C16). It contains single-qubit Z gates, as well as multiqubit operations on all subsets of the n qubits on each of partitions A and B . The operations consist of two different types. The first type is multiqubit rotation gates with angles $\pm\pi/2$. They can be implemented by single-qubit rotation gates sandwiched by ladders of CNOT gates [39]. The second is nonunitary operations, in which the single-qubit rotation gates within the CNOT ladders are replaced by computational-basis measurements; a detailed discussion is given in Appendix C. As a consequence, each term in the decomposition can be implemented with $O(n)$ CNOT gates. Note that an n -qubit rotation gate about $\pm\pi/2$ can be implemented with exactly n CNOT gates. Since the majority of terms in the decomposition of an n -qubit rotation gate shown in Eq. (C16) contain operations on fewer than n qubits, we expect strong noise reduction by using our scheme. Noise mitigation due to the reduction of costly two-qubit gates in circuit-cutting schemes was recently observed experimentally in Ref. [24].

It has been emphasized that the sampling overhead is not the only relevant metric measuring the quality of a circuit-cutting scheme [16]; the number L of different channels to be applied is also relevant. In principle, the number of experimental shots required to evaluate an expectation value to additive error ϵ scales as $O(\gamma^2/\epsilon^2)$, which is independent of L . In practice, however, the latency introduced by the additional compilation time needed for each of the L subcircuits

corresponding to the elements \mathcal{F}_i for $i = 1, \dots, L$ on topologically constrained hardware can be the dominant factor in the runtime of the algorithm [16]. As discussed in Appendix A, our decomposition guarantees $L = O(4^n)$ for the case of parallel as well as nonparallel gates, which is an exponential improvement over previous results. It should be noted that we could trade compilation time for additional circuit depth. Indeed, a multiqubit rotation gate can be trivially routed to topologically constrained hardware by SWAP insertion. This decreases compilation latency to almost zero but, in turn, might drastically increase the CNOT-gate count and the additional circuit depth d , resulting in a more severe impact of noise. Note that a multiqubit rotation gate can be implemented on fully connected hardware with depth $d = O(\lceil \log(n) \rceil)$ when replacing each CNOT gate ladder by a balanced tree [39]. For hardware with constrained connections, efficient compilation methods exist for multiqubit rotation gates based on minimum spanning trees [40] and CNOT-gate resynthesis [41]. In what sense the low depth of the circuits can be retained will be the focus of further investigations.

IV. CONCLUSION

In this work, we introduced joint cutting of non-Clifford two-qubit rotation gates based on virtual teleportation and proved optimality. In the case of parallel gates, we further derived an ancilla-free optimal decomposition. We conclude by stressing the following aspects: As mentioned before, controlled rotation gates are equivalent to two-qubit rotation gates up to local unitary operations. In addition, CNOT gates are special instances of this gate class. Consequently, the joint virtual teleportation protocol proposed in this work can also be used for optimal joint cutting of CNOT gates with several improvements over previous work [23]. (1) In our virtual teleportation protocol, no classical communication needs to be shared. As a result, no real-time feedback during the quantum computation in the form of local correction operations conditioned on the measurement outcomes on the ancilla qubits is required. (2) As a consequence, the cut circuits can be run in sequence on the same hardware as opposed to parallel execution, which was necessary for previous methods. (3) With our alternative quasiprobability decomposition of pure-state density matrices, we reach an exponential reduction in the number of terms required compared to Refs. [23,37], directly translating into an exponential reduction in the number of channels in the cutting scheme. (4) Our decomposition method allows us to derive ancilla-free optimal joint cutting schemes for CNOT gates, two-qubit rotation gates, and controlled rotation gates, which has not yet been considered in the literature.

Note added. Recently, Refs. [42,43] appeared as preprints and generalize some of our ideas.

ACKNOWLEDGMENTS

C.U. would like to thank M. Bechtold, D. Sutter, and F. Wagner for helpful discussions. The research was funded by the project QuaST, supported by the Federal Ministry for Economic Affairs and Climate Action (Grant No. 01MQ22004A) on the basis of a decision by the German Bundestag.

APPENDIX A: OPTIMAL QPD FOR A PURE-STATE DENSITY MATRIX

In this Appendix, we derive an alternative optimal QPD for arbitrary pure-state density matrices. Vidal and Tarrach [37] considered the following problem: Given two partitions and a bipartite state ρ , what is the minimal mixing of the state with a separable state such that the mixed state becomes separable as well? This problem can be rephrased as follows [23]: What is the minimal R , called the robustness of entanglement, for which

$$\rho = (1 + R)\eta_1 - R\eta_2, \quad (\text{A1})$$

where η_1 and η_2 are separable. While such decomposition is hard to find in general, for a pure state $\rho = |\psi\rangle\langle\psi|$ the robustness can be calculated via the m Schmidt coefficients c_j , with $j = 0, \dots, m-1$, of the $|\psi\rangle$ state as

$$R = \left(\sum_{j=0}^{m-1} c_j \right)^2 - 1. \quad (\text{A2})$$

Consequently,

$$\gamma = 1 + 2R. \quad (\text{A3})$$

In Ref. [37], the authors also provided an explicit representation for η_1 and η_2 in which, however, the separable states themselves are functions of the Schmidt coefficients. In contrast, the alternative QPD derived in the following contains only the Schmidt coefficients in the quasiprobabilities. As discussed in the main text, this guarantees that, when our alternative QPD is applied, to the protocol in Figs. 3 and 4, we can correct the unwanted measurement results in the classical postprocessing step. A similar representation was derived in Refs. [44,45]. In the following, we first derive the general form of the decomposition and subsequently show optimality.

Consider a bipartite state $|\psi\rangle$ on partitions A and B and an expansion of the form

$$|\psi\rangle = \sum_{j=0}^{m-1} c_j |\varphi_j\rangle \otimes |\varphi'_j\rangle, \quad (\text{A4})$$

with c_j being real but not necessarily positive. Then

$$\begin{aligned} |\psi\rangle\langle\psi| &= \sum_{j=0}^{m-1} c_j^2 |\varphi_j\rangle\langle\varphi_j| \otimes |\varphi'_j\rangle\langle\varphi'_j| \\ &\quad + 2 \sum_{i>j} c_i c_j (\sigma_{ij}^+ - \sigma_{ij}^-) \end{aligned} \quad (\text{A5})$$

holds. In this expression

$$\sigma_{ij}^\pm = \frac{1}{\alpha} \sum_{r=1}^{\alpha} |\xi_{rij}^\pm\rangle\langle\xi_{rij}^\pm| \otimes |\tau_{rij}\rangle\langle\tau_{rij}| \quad (\text{A6})$$

are separable density matrices with respect to partitions A and B . The states take the forms

$$|\xi_{rij}^\pm\rangle = \frac{1}{\sqrt{2}} (|\varphi_i\rangle \pm e^{i\phi_r} |\varphi_j\rangle) \quad (\text{A7})$$

and

$$|\tau_{rij}\rangle = \frac{1}{\sqrt{2}} (|\varphi'_i\rangle + e^{-i\phi_r} |\varphi'_j\rangle). \quad (\text{A8})$$

Finally, the phases ϕ_r can be chosen such that

$$\sum_{r=1}^{\alpha} e^{i\phi_r} = \sum_{r=1}^{\alpha} e^{i2\phi_r} = 0. \quad (\text{A9})$$

Furthermore, if Eq. (A4) is the Schmidt decomposition of $|\psi\rangle$ (in this case $c_j \geq 0$), then Eq. (A5) is optimal in that it achieves the minimal possible γ .

Proof. The statement is most easily proven by inserting Eqs. (A6)–(A9) into Eq. (A5). The optimality in the case when Eq. (A4) is the Schmidt decomposition of $|\psi\rangle$ remains to be shown. To that end, we sum over the absolute of the coefficients for all terms with the result stated in Eqs. (A2) and (A3). Therefore, the decomposition is optimal.

We choose $\phi_r = 2\pi r/\alpha$ for integer $\alpha \geq 3$. The choice $\alpha = 3$ involves the least number of terms, but we observe in Appendix C that $\alpha = 4$ allows removing ancilla qubits for parallel gates. Equation (A5) involves $O(m^2)$ terms. This is an exponential improvement of the decomposition used in Ref. [37] with $O(2^m)$ terms.

APPENDIX B: LOWER BOUNDS

In this Appendix we prove a lower bound on $\gamma_{\text{joint}}^{(n)}$, the γ parameter for jointly cutting n two-qubit rotation gates. This bound will agree with the upper bound established in Eq. (16) via the joint virtual teleportation scheme, thereby proving optimality. A lower bound on the γ parameter of a quantum gate can be obtained from the QPD of the Choi state of the gate's unitary [23]. The argument is as follows: Since the Choi state of a gate's unitary can be formed by local (with respect to partitions A and B) operations, the γ parameter of the gate cannot possibly be smaller than that of the Choi state. If it were, construction of the Choi state with smaller γ would be possible, leading to a contradiction. Thus, Eq. (7) provides

$$\gamma \geq 2 \left(\sum_j c_j \right)^2 - 1, \quad (\text{B1})$$

where c_j are the Schmidt coefficients of the Choi state of the gate. Further note that the Choi state of a gate whose matrix representation is diagonal in the computational basis has the same Schmidt coefficients as the normalized vector containing the diagonal elements. In our case, this vector is just $|\Psi_{k_1 l_1 \dots k_n l_n}\rangle$ for $k_s = l_s = 0$ for all s up to local operations. Consequently, the lower bound matches the upper one, proving the optimality of the joint virtual teleportation approach for joint cutting of two-qubit rotation gates.

Reference [23] proved optimality for the large class of gate decompositions derived in Ref. [22]. Note, however, that the lower bound obtained in this way can be relatively loose for some gates. For example, in the case of a K -qubit version of a Toffoli gate, it can be shown that the right-hand side of Eq. (B1) for a cut after one qubit tends to 1 as $K \rightarrow \infty$. Since, however, this gate can be used to create a Bell state between two partitions, $\gamma \geq 3$ must be true.

APPENDIX C: PARALLEL GATES

In this Appendix, we derive the ancilla-free decomposition of n parallel two-qubit rotation gates; we start from the protocol in Fig. 4. First, we define the map

$$\Lambda(|\psi\rangle\langle\psi|)[\cdot] = 2^n \sum_{i,j} \langle i|\psi\rangle\langle\psi|j\rangle P_i \cdot P_j, \quad (\text{C1})$$

where $P_i = |i\rangle\langle i|$ and $\{|i\rangle\}_{i=0}^{2^n-1}$ is the computational basis. This map corresponds to the transformation $U \cdot U^\dagger$, where U contains the computational basis elements of $|\psi\rangle$ multiplied by $\sqrt{2^n}$ on its diagonal. Thus, U is unitary if $|\langle i|\psi\rangle| = 1/\sqrt{2^n}$ for all i . Further, note that Λ is linear in its first argument, and if $|\psi\rangle$ factorizes over two partitions, then the map does as well.

Let us define as $\mathcal{R}_{zz}^{(n)}$ the $2n$ -qubit unitary channel corresponding to the n two-qubit rotation gates we want to cut and a state $|\psi\rangle$ such that

$$\Lambda(|\psi\rangle\langle\psi|)[\cdot] = \mathcal{R}_{zz}^{(n)}. \quad (\text{C2})$$

With the Hadamard gate H and the phase gate S this state is

$$|\psi\rangle = H^{\otimes n} \otimes (HS^\dagger)^{\otimes n} |\Psi\rangle, \quad (\text{C3})$$

with $|\Psi\rangle$ as in Eq. (13) but with $k_s = l_s = 0$ for all s , that is,

$$|\Psi\rangle = \sum_{j \in \{0,1\}^n} c_j |j_1 \dots j_n\rangle |j_n \dots j_1\rangle, \quad (\text{C4})$$

with $c_j = \Pi_{s=1}^n c_{j_s}$ and

$$c_{j_s} = \begin{cases} \cos(\theta_s/2) & \text{for } j_s = 0, \\ \sin(\theta_s/2) & \text{for } j_s = 1. \end{cases} \quad (\text{C5})$$

The following steps remain: inserting a QPD of $|\psi\rangle\langle\psi|$ calculated with the help of Appendix A into Eq. (C1) and then using the linearity and factorization properties of Λ . Comparison with Eq. (A4) shows $m = 2^n$, $|\varphi_j\rangle = H^{\otimes n} |j_1, \dots, j_n\rangle$, and $|\varphi'_j\rangle = (HS^\dagger)^{\otimes n} |j_n, \dots, j_1\rangle$. To calculate the map Λ , we first make use of linearity and calculate the different terms individually. When we define the unitary channel

$$\mathcal{Z}_j[\cdot] = Z^{j_1} \otimes \dots \otimes Z^{j_n} \cdot Z^{j_1} \otimes \dots \otimes Z^{j_n} \quad (\text{C6})$$

corresponding to the application of single-qubit Z gates, we find

$$\Lambda(|\varphi_j\rangle\langle\varphi_j|)[\cdot] = \mathcal{Z}_j[\cdot], \quad (\text{C7})$$

where we used

$$\sqrt{2} \sum_{k_s \in \{0,1\}} \langle k_s | H | j_s \rangle P_{k_s} = \sum_{k_s \in \{0,1\}} (-1)^{j_s k_s} P_{k_s} = Z^{j_s}. \quad (\text{C8})$$

Equally, we find the same result for $|\varphi'_j\rangle$ for j replaced by $\tilde{j} := (j_n, \dots, j_1)$. The evaluation of states $|\xi_{rij}^\pm\rangle$ and $|\tau_{rij}\rangle$ requires a little bit more work. First, we insert $|\xi_{rij}^\pm\rangle$, defined in Eq. (A7),

with the result

$$\sqrt{2}^n \sum_k \langle k | \xi_{rij}^\pm \rangle P_k = Z^{i_1} \otimes \dots \otimes Z^{i_n} [(\mathbb{I}^{\otimes n} \pm e^{i\phi_r} Z^{j_1-i_1} \otimes \dots \otimes Z^{j_n-i_n}) / \sqrt{2}]. \quad (\text{C9})$$

For $\phi_r = 2\pi r/\alpha$, with $\alpha = 4$, we first consider $\exp(i\phi_1) = i$ and $\exp(i\phi_3) = -i$. In this case, the expression in the brackets is unitary. Indeed,

$$R_{ij}(\pm\pi/2) = \frac{\mathbb{I}^{\otimes n} \mp i Z^{j_1-i_1} \otimes \dots \otimes Z^{j_n-i_n}}{\sqrt{2}} \quad (\text{C10})$$

is a multiqubit rotation gate on the qubits with $|i_s - j_s| = 1$. Again, we define the corresponding channel $\mathcal{R}_{ij}(\pm\pi/2)$ using calligraphic notation. Furthermore, we introduce the abbreviation

$$\mathcal{R}_{ij} = \frac{\mathcal{R}_{ij}(\pi/2) - \mathcal{R}_{ij}(-\pi/2)}{2} \quad (\text{C11})$$

for the unitary channels. We now turn to Eq. (C9) with $\exp(i\phi_2) = -1$ and $\exp(i\phi_4) = 1$. In this case, we define

$$P_{ij}^k = \frac{\mathbb{I}^{\otimes n} + (-1)^k Z^{j_1-i_1} \otimes \dots \otimes Z^{j_n-i_n}}{2} \quad (\text{C12})$$

for $k = 0, 1$, which is nonunitary. The channel \mathcal{P}_{ij}^k can be evaluated on a quantum computer using a ladder of CNOT gates [46] and a projector on $|0\rangle$ or $|1\rangle$, e.g.,

$$\frac{\mathbb{I}^{\otimes t} + (-1)^k Z^{\otimes t}}{2} = \text{Circuit diagram showing } t \text{ qubits with CNOT gates and a projector on } |k\rangle\langle k|. \quad (\text{C13})$$

for integer $t > 0$. We then define

$$\mathcal{P}_{ij} = \mathcal{P}_{ij}^0 - \mathcal{P}_{ij}^1, \quad (\text{C14})$$

which is neither trace preserving nor positive. However, \mathcal{P}_{ij} is the difference of two completely positive trace nonincreasing maps, whose sum is completely positive and trace preserving. Such maps can be simulated on a quantum device without extra sampling overhead [21–24]. Thus, both \mathcal{R}_{ij} and \mathcal{P}_{ij} can be simulated with $\gamma = 1$. The evaluation of Eq. (C1) for $|\tau_{rij}\rangle$ proceeds along the same lines. We only need to replace $i, j \rightarrow \tilde{i}, \tilde{j}$, choose the plus sign in Eq. (C9), and replace $\phi_r \rightarrow -\phi_r$. Due to the extra S^\dagger gate in the definition of $|\varphi'\rangle$, we additionally have to replace $Z \rightarrow -iZ$, introducing a factor $(-i)^{v_{ij}}$ with

$$v_{ij} = \sum_{s=1}^n (j_s - i_s). \quad (\text{C15})$$

Collecting all the terms, we finally arrive at

$$\mathcal{R}_{zz}^{(n)} = \sum_{j \in \{0,1\}^n} c_j^2 \mathcal{Z}_j \otimes \mathcal{Z}_{\tilde{j}} + 2 \sum_{i>j} c_i c_j [\mathcal{Z}_i \otimes \mathcal{Z}_{\tilde{i}}] \circ \begin{cases} (-1)^{v_{ij}/2} [\mathcal{P}_{ij} \otimes \mathcal{P}_{\tilde{i}\tilde{j}} - \mathcal{R}_{ij} \otimes \mathcal{R}_{\tilde{i}\tilde{j}}] & \text{for even } v_{ij}, \\ (-1)^{(v_{ij}-1)/2} [\mathcal{R}_{ij} \otimes \mathcal{P}_{\tilde{i}\tilde{j}} + \mathcal{P}_{ij} \otimes \mathcal{R}_{\tilde{i}\tilde{j}}] & \text{for odd } v_{ij}. \end{cases} \quad (\text{C16})$$

This decomposition is optimal with

$$\gamma = 2 \Pi_{s=1}^n (1 + |\sin(\theta_s)|) - 1. \quad (\text{C17})$$

It consists of single-qubit Z gates, multiqubit Z rotations by the angles $\pm\pi/2$, and measurements within CNOT-gate ladders. Due to the double sum, the number of terms in Eq. (C16)

grows as $O(4^n)$. Interestingly, the decomposition achieves the optimal γ parameter even though no exchange of classical information is required between the partitions.

-
- [1] P. Shor, Algorithms for quantum computation: Discrete logarithms and factoring, in *Proceedings 35th Annual Symposium on Foundations of Computer Science* (IEEE Computer Society Press, Santa Fe, NM, USA, 1994), p. 124.
 - [2] L. K. Grover, A fast quantum mechanical algorithm for database search, in *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, 1996), p. 212.
 - [3] M. P. Harrigan *et al.*, Quantum approximate optimization of non-planar graph problems on a planar superconducting processor, *Nat. Phys.* **17**, 332 (2021).
 - [4] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, [arXiv:1411.4028](https://arxiv.org/abs/1411.4028).
 - [5] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Supervised learning with quantum-enhanced feature spaces, *Nature (London)* **567**, 209 (2019).
 - [6] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, *Nature (London)* **549**, 195 (2017).
 - [7] B. P. Lanyon, J. D. Whitfield, G. G. Gillett, M. E. Goggin, M. P. Almeida, I. Kassal, J. D. Biamonte, M. Mohseni, B. J. Powell, M. Barbieri, A. Aspuru-Guzik, and A. G. White, Towards quantum chemistry on a quantum computer, *Nat. Chem.* **2**, 106 (2010).
 - [8] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets, *Nature (London)* **549**, 242 (2017).
 - [9] Y. Kim, A. Eddins, S. Anand, K. X. Wei, E. van den Berg, S. Rosenblatt, H. Nayfeh, Y. Wu, M. Zaletel, K. Temme, and A. Kandala, Evidence for the utility of quantum computing before fault tolerance, *Nature (London)* **618**, 500 (2023).
 - [10] M. H. Devoret and R. J. Schoelkopf, Superconducting circuits for quantum information: An outlook, *Science* **339**, 1169 (2013).
 - [11] C. Monroe and J. Kim, Scaling the ion trap quantum processor, *Science* **339**, 1164 (2013).
 - [12] X. Yuan, B. Regula, R. Takagi, and M. Gu, Virtual quantum resource distillation, *Phys. Rev. Lett.* **132**, 050203 (2024).
 - [13] M. Bechtold, J. Barzen, F. Leymann, and A. Mandl, Circuit cutting with non-maximally entangled states, [arXiv:2306.12084](https://arxiv.org/abs/2306.12084).
 - [14] T. Peng, A. W. Harrow, M. Ozols, and X. Wu, Simulating large quantum circuits on a small quantum computer, *Phys. Rev. Lett.* **125**, 150504 (2020).
 - [15] L. Brenner, C. Piveteau, and D. Sutter, Optimal wire cutting with classical communication, [arXiv:2302.03366](https://arxiv.org/abs/2302.03366).
 - [16] H. Harada, K. Wada, and N. Yamamoto, Doubly optimal parallel wire cutting without ancilla qubits, [arXiv:2303.07340](https://arxiv.org/abs/2303.07340).
 - [17] A. Lowe, M. Medvidović, A. Hayes, L. J. O’Riordan, T. R. Bromley, J. M. Arrazola, and N. Killoran, Fast quantum circuit cutting with randomized measurements, *Quantum* **7**, 934 (2023).
 - [18] G. Uchihara, T. M. Aamodt, and O. D. Matteo, Rotation-inspired circuit cut optimization, in *2022 IEEE/ACM Third International Workshop on Quantum Computing Software (QCS)* (IEEE Computer Society, Los Alamitos, CA, 2022), p. 50.
 - [19] P. Pednault, An alternative approach to optimal wire cutting without ancilla qubits, [arXiv:2303.08287](https://arxiv.org/abs/2303.08287).
 - [20] H. F. Hofmann, How to simulate a universal quantum computer using negative probabilities, *J. Phys. A* **42**, 275304 (2009).
 - [21] K. Mitarai and K. Fujii, Constructing a virtual two-qubit gate by sampling single-qubit operations, *New J. Phys.* **23**, 023021 (2021).
 - [22] K. Mitarai and K. Fujii, Overhead for simulating a non-local channel with local channels by quasiprobability sampling, *Quantum* **5**, 388 (2021).
 - [23] C. Piveteau and D. Sutter, Circuit knitting with classical communication, *IEEE Trans. Inf. Theory* **70**, 2734 (2023).
 - [24] C. Ufrecht, M. Periyasamy, S. Rietsch, D. D. Scherer, A. Plinge, and C. Mutschler, Cutting multi-control quantum gates with ZX calculus, *Quantum* **7**, 1147 (2023).
 - [25] H. Pashayan, J. J. Wallman, and S. D. Bartlett, Estimating outcome probabilities of quantum circuits using quasiprobabilities, *Phys. Rev. Lett.* **115**, 070501 (2015).
 - [26] C. Piveteau, D. Sutter, and S. Woerner, Quasiprobability decompositions with reduced sampling overhead, *npj Quantum Inf.* **8**, 12 (2022).
 - [27] K. Temme, S. Bravyi, and J. M. Gambetta, Error mitigation for short-depth quantum circuits, *Phys. Rev. Lett.* **119**, 180509 (2017).
 - [28] L. S. Herzog, F. Wagner, C. Ufrecht, L. Palackal, A. Plinge, C. Mutschler, and D. D. Scherer, Improving quantum and classical decomposition methods for vehicle routing, [arXiv:2404.05551](https://arxiv.org/abs/2404.05551).
 - [29] D. Collins, N. Linden, and S. Popescu, Nonlocal content of quantum operations, *Phys. Rev. A* **64**, 032302 (2001).
 - [30] D. Gottesman and I. L. Chuang, Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations, *Nature (London)* **402**, 390 (1999).
 - [31] J. Eisert, K. Jacobs, P. Papadopoulos, and M. B. Plenio, Optimal local implementation of nonlocal quantum gates, *Phys. Rev. A* **62**, 052317 (2000).
 - [32] D. Stahlke and R. B. Griffiths, Entanglement requirements for implementing bipartite unitary operations, *Phys. Rev. A* **84**, 032316 (2011).
 - [33] A. Soeda, P. S. Turner, and M. Murao, Entanglement cost of implementing controlled-unitary operations, *Phys. Rev. Lett.* **107**, 180501 (2011).
 - [34] J. I. Cirac, W. Dür, B. Kraus, and M. Lewenstein, Entangling operations and their implementation using a small amount of entanglement, *Phys. Rev. Lett.* **86**, 544 (2001).

- [35] W. Dür and J. I. Cirac, Nonlocal operations: Purification, storage, compression, tomography, and probabilistic implementation, *Phys. Rev. A* **64**, 012317 (2001).
- [36] B. Groisman and B. Reznik, Implementing nonlocal gates with nonmaximally entangled states, *Phys. Rev. A* **71**, 032322 (2005).
- [37] G. Vidal and R. Tarrach, Robustness of entanglement, *Phys. Rev. A* **59**, 141 (1999).
- [38] It can be shown that all operations in the virtual teleportation protocol are elements of the extended definition of local operations used in Refs. [15,23].
- [39] A. Cowtan, S. Dilkes, R. Duncan, W. Simmons, and S. Sivarajah, Phase gadget synthesis for shallow circuits, *Electron. Proc. Theor. Comput. Sci.* **318**, 213 (2020).
- [40] S. Gogioso and R. Yeung, Annealing optimisation of mixed ZX phase circuits, *Electron. Proc. Theor. Comput. Sci.* **394**, 415 (2023).
- [41] B. Nash, V. Gheorghiu, and M. Mosca, Quantum circuit optimizations for NISQ architectures, *Quantum Sci. Technol.* **5**, 025010 (2020).
- [42] L. Schmitt, C. Piveteau, and D. Sutter, Cutting circuits with multiple two-qubit unitaries, [arXiv:2312.11638](https://arxiv.org/abs/2312.11638).
- [43] A. W. Harrow and A. Lowe, Optimal quantum circuit cuts with application to clustered Hamiltonian simulation, [arXiv:2403.01018](https://arxiv.org/abs/2403.01018).
- [44] J. Sperling and W. Vogel, Representation of entanglement by negative quasiprobabilities, *Phys. Rev. A* **79**, 042337 (2009).
- [45] J. Sperling and W. Vogel, Erratum: Representation of entanglement by negative quasiprobabilities [Phys. Rev. A **79**, 042337 (2009)], *Phys. Rev. A* **80**, 029905(E) (2009).
- [46] K. Mitarai and K. Fujii, Methodology for replacing indirect measurements with direct measurements, *Phys. Rev. Res.* **1**, 013006 (2019).