Five two-qubit gates are necessary for implementing the Toffoli gate

Nengkun Yu,* Runyao Duan, and Mingsheng Ying

State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China and Center for Quantum Computation and Intelligent Systems (QCIS), Faculty of Engineering and Information Technology, University of Technology, Sydney, New South Wales 2007, Australia

(Received 22 January 2013; published 30 July 2013)

In this Rapid Communication, we consider the open problem of the minimum cost of two-qubit gates for simulating the Toffoli gate and show that five two-qubit gates are necessary. Before our work, it was known that five two-qubit gates are sufficient to implement the Toffoli gate, and numerical evidence indicates that five two-qubit gates are also necessary. The idea introduced here can also be used to solve the problem of optimal simulation of Deutsch three-qubit gates.

DOI: 10.1103/PhysRevA.88.010304 PACS number(s): 03.67.Ac, 03.65.Ud, 89.70.Eg

Since quantum computation provides the possibility of solving certain problems much faster than any classical computer using the best currently known algorithms [1–5], a huge amount of effort has been devoted to building functional and scalable quantum computers over the last two decades. The quantum circuit model is one popular model of quantum computer hardware [6–18]. In order to be a general purpose computational device, a quantum computer must implement a small set of quantum logical gates [14], which are universal, that is, can serve as the basic building blocks of quantum circuits, in the same way as do classical logical gates for conventional digital circuits. It is quite natural to choose certain gates operating on a small number of qubits as the basic gates.

Theoretically, any two-qubit gate that can create entanglement, like the controlled-\text{NOT} (CNOT) gate, together with all single-qubit gates, is universal [18]. It has also been experimentally demonstrated that two-qubit gates can be realized with high fidelity using the current technology, for example, two-qubit gates with superconducting qubits have been presented with fidelities higher than 90% [19]. Finding more efficient ways to implement quantum gates may allow small-scale quantum computing tasks to be demonstrated on a shorter time scale. More precisely, it would be quite helpful for defeating quantum decoherence to realize multiqubit gates with the least number of possible basic gates. Thus an important problem is how to implement multiqubit gates using only two-qubit gates. Indeed, a study of the minimum cost of two-qubit gates for simulating a multiqubit gate is not only of theoretical importance, but also an experimental requirement: to accomplish a quantum algorithm, even at a small size, one has to implement a relatively high level of control over the multiqubit quantum system. A lot of experiments demonstrate multiqubit controlled-\text{NOT} gates in ion traps [20], linear optics [21], superconductors [22], and atoms [23].

Making controlled unitaries is an essential task for many algorithms in quantum computing [1]. Among all quantum controlled gates, those highly controlled unitaries (i.e., unitaries controlled on more than one other qubit) are useful in numerous quantum algorithms including the oracle in the binary welded tree algorithm [5] and quantum simulation [24]. The Toffoli gate is perhaps one of the most important highly controlled unitaries for three reasons: (1) The Toffoli gate is universal for classical reversible computation in the sense that all conventional Boolean circuits can be built upon it in a reversible way [25]; (2) it is also universal for quantum computation with little extra help in the sense that the one-qubit Hadamard gate is provided as a free resource [26]; (3) it is the simplest highly controlled unitary. Furthermore, a series of works showed that the Toffoli gate is an indispensable ingredient in realizing fault-tolerant quantum computation [23,27–30]. Recently, experimental implementation of the Toffoli gate has received considerable attention. The first experimental realization of the quantum Toffoli gate was presented in an ion-trap quantum computer, in 2009 at the University of Innsbruck, Austria [20]. Then, the Toffoli gate was realized in linear optics [21] and superconducting circuits [22,31,32].

Due to its significance in quantum computing, the theoretical pursuit of efficient implementation of the Toffoli gate using a sequence of single- and two-qubit gates has a quite long history [7,8,11,12,33–37]. It was explicitly stated as an open problem by Nielsen and Chuang in their influential textbook on quantum computation [14]: How many general two-qubit gates (or CNOT gates) are required to implement the Toffoli gate (see [14], p. 213, Problem 4.4)? What is known already is that the Toffoli gate can be decomposed as a circuit consisting of five two-qubit gates (or six CNOT gates), and numerical evidence has been gathered indicating that the five-gate implementation is optimal [8,11,12]. The cost of using a CNOT gate was first solved by Shende and Markov; they showed that six CNOT gates are optimal when single-qubit unitaries are free [36], a situation which was studied in [11,35,38]. However, the optimal simulation of the Toffoli gate by using general bipartite quantum logical gates remains unknown. This problem has attracted attention from many researchers in the last two decades [8,11,12,34–37]. In this Rapid Communication, we address this problem by showing that five two-qubit gates are necessary for implementing the Toffoli gate. The main difficulty of analyzing a circuit consisting of general two-qubit gates is that it contains more degrees of freedom. We would like to point out that the main

*nengkunyu@gmail.com
tool used in [36], the Cartan decomposition does not fit the implementation of using general two-qubit gates very well.

By counting the degrees of freedom (DOFs), in [7], it was shown that almost any \( n \)-qubit gate requires at least \( \left\lceil \frac{2^{n-1} - 1}{9} \right\rceil \) two-qubit gates to implement without ancilla. To see the validity of this statement, one needs to study quantum gates with unit determinant and notice that each two-qubit unitary operation \( U \) can be expressed in the form [39]

\[
U = (uA \otimes uB)U_d(vA \otimes vB),
\]

where \( uA, uB, vA, \) and \( vB \) are one-qubit unitary gates with unit determinants, with

\[
U_d = \exp[i(\alpha_x X \otimes X + \alpha_y Y \otimes Y + \alpha_z Z \otimes Z)]
\]

with Pauli matrices \( X, Y, Z \). Then the DOFs of the circuit with any structure are upper bounded by the summation of the DOFs of all \( U_d \)'s and one-qubit unitaries.

Let \( n = 3 \); then the simulation of a general three-qubit gate would require at least \( \left\lceil \frac{2^{3-1} - 1}{9} \right\rceil = 6 \) two-qubit gates. The following natural question then arose [12]: Are six two-qubit gates sufficient to implement a three-qubit gate? The general answer to this question remains open. In the following discussion, we focus only on the minimal cost of simulating \( V_{ABC} \) using two-qubit gates.

The Toffoli gate \( T_{ABC} \) can also be written as the following operator by considering the output over the computational basis:

\[
I - |110\rangle\langle110| - |111\rangle\langle111| - |110\rangle\langle111| - |111\rangle\langle110|.
\]

Let \( V_{ABC} = I_{ABC} - 2|111\rangle\langle111| \) with \( I_{ABC} \) being the identity operator on the Hilbert space \( \mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_C \). It is evident that

\[
V_{ABC} = (I_{AB} \otimes \mathcal{H}_C)T_{ABC}(I_{AB} \otimes \mathcal{H}_C),
\]

where \( H \) is the Hadamard gate given by

\[
H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.
\]

In other words, \( V_{ABC} \) and the Toffoli gate \( T_{ABC} \) are equivalent up to the local unitary \( \mathcal{H}_C \). By absorbing \( \mathcal{H}_C \) into any two-qubit gates acting on \( AC \) or \( BC \), we can easily conclude that \( V_{ABC} \) and \( T_{ABC} \) require the same number of two-qubit gates for realization. Thus in the following discussion, we focus only on the minimal cost of simulating \( V_{ABC} \) using two-qubit gates.

The gate \( V_{ABC} \) is a real Hermitian matrix that is invariant under any permutation of subsystems \( A, B, \) and \( C \). Thus it can be regarded as a controlled gate with control on each qubit. Note that any bipartite unitary \( U_{AB} \) acting on a qubit system \( A \) and a general system \( B \) is said to be a controlled gate with control on \( A \) if it can be decomposed into the form of \( U_{AB} = |0\rangle\langle0| \otimes U_0 + |1\rangle\langle1| \otimes U_1 \). This simple observation is helpful to reduce the number of cases we need to consider.

Since \( V_{ABC} \) is regarded as a three-qubit gate acting on \( ABC \), any two-qubit gate used to implement \( V_{ABC} \) can be simply classified into three types: \( K_{AB} \), the gate acting on the systems \( A \) and \( B \), and likewise \( K_{BC} \) and \( K_{AC} \). Clearly, it is impossible that all two-qubit gates used to simulate \( V_{ABC} \) belong to the same type. Furthermore, we can verify that two two-qubit gates are not sufficient for the simulation of \( V_{ABC} \).

Let us illustrate this by a concrete example. Suppose \( U_{AB} \) is a three-qubit unitary which can be regarded as a controlled gate between the bipartition \( A-BC \) with control system \( A \). Then there exist one-qubit gates \( v_{B1}, v_{B2} \) on \( \mathcal{H}_B \) and one-qubit gates \( w_{C1}, w_{C2} \) on \( \mathcal{H}_C \) such that \( U_{AB}U_{AC} = |0\rangle\langle0| \otimes v_{B1} \otimes w_{C1} + |1\rangle\langle1| \otimes v_{B2} \otimes w_{C2} \).

Observations (i) and (ii) are obvious. To see (iii), we can assume \( U_{AC}|0\rangle\langle\gamma| = |0\rangle\langle0| \otimes \gamma|c\rangle \) by moving the local unitary to the left of \( U_{AB} \). Then \( U_{AB}U_{AC}|0\rangle\langle\gamma| = U_{AB}|0\rangle\langle\gamma|B \otimes \gamma|c\rangle \). Note that the state of \( A \)'s part of \( U_{AB}|0\rangle\langle\gamma|B \) is always \( |0\rangle \), which means that \( U_{AC} \) is a controlled gate with control on \( A \). Similarly, \( U_{AB} \) is also a controlled gate with control on \( A \). Hence the result follows.

Now we show that three two-qubit gates are not sufficient to implement \( V_{ABC} \). We achieve this goal by analyzing all possible circuits consisting of three two-qubit gates. Due to the highly symmetric properties of \( V_{ABC} \), we need to consider only the following two cases:

**Case I.** These three gates belong to just two types. Without loss of generality (WLOG), we can assume that two gates are of the type \( K_{AB} \) and the third one is of the type \( K_{BC} \), and the circuit is (note that the time goes from left to right)

\[
\begin{aligned}
A & \quad \xrightarrow{V_{AB}} \quad B \\
\quad & \quad \xrightarrow{U_{BC}} \quad C
\end{aligned}
\]

We only need to show that there is no solution of the equation

\[
U_{AB}U_{BC}V_{AB} = V_{ABC},
\]

where \( U_{AB} \) and \( V_{AB} \) are of type \( K_{AB} \), and \( U_{BC} \) of type \( K_{BC} \). Then \( U_{BC} \) must be a controlled gate with control on \( C \) by noticing that \( U_{BC} = U_{AB}V_{ABC}V_{AB} \), where \( \dagger \) stands for the Hermitian conjugate. We write \( U_{BC} = |0\rangle\langle0| \otimes I_B + |0\rangle\langle0| \otimes w_B \). A direct calculation leads us to the conclusion that \( I_A \otimes w_B \) and \( I - 2|11\rangle\langle11| \) share the same set of eigenvalues (counting multiplicity). That is impossible.
Case 2. Three gates belong to different types. WLOG, we assume the circuit is
\[ U_{AB}U_{BC}U_{AC} = V_{ABC}. \]
Then \( U_{BC}U_{AC} \) is a controlled gate with control bit \( C \). As discussed in observation (iii), we can obtain that \( U_{BC} \) is a controlled gate with control system \( C \), and so is \( U_{AC} \). Consequently, we can assert that \( I - 2|11\rangle\langle 11| \) is a local unitary by figuring out directly the form of the control unitary. That is again impossible.

We can generalize this technique to show that \( V_{ABC} \) cannot be implemented by any circuit consisting of four nonlocal two-qubit gates. We do not count the number of one-qubit gates as they can be easily absorbed into relevant two-qubit gates. Again the symmetric property of \( V_{ABC} \) enables us to consider only the following two cases:

Case 1. Four gates belong to only two types, say \( K_{AB} \) and \( K_{BC} \). Due to the symmetry of \( V_{ABC} \), we need to show only that the following circuit cannot simulate \( V_{ABC} \):

\[ \begin{array}{c}
A \\
B \\
C \\
\end{array} \quad \begin{array}{cc}
\text{V}_{ABC} \\
\text{U}_{BC} \\
\text{U}_{AB} \\
\end{array} \]

This is necessary to show that the following equation has no solution:
\[ U_{AB}U_{BC}V_{AB}V_{BC} = V_{ABC}. \]

For readability, we postpone the detailed proof of this conclusion to the Appendix.

Case 2. Each of three types contains at least one of the four two-qubit gates. Again due to the symmetry of \( V_{ABC} \), we need to deal only with the following two subcases:

Case 2.1. The circuit is represented by \( U_{AC}U_{AB}U_{BC}V_{AB} = V_{ABC} \). We can reduce this circuit to the circuit considered in Case 1 by observing that \( S_{AB}V_{ABC}V_{AB} = V_{ABC} \) and
\[ (S_{AB}U_{AC}S_{AB})(S_{AB}U_{AB})U_{BC}(V_{AB}S_{AB}) = V_{ABC}, \]
where \( S_{AB} \) is the swap gate on system \( \mathcal{H}_A \otimes \mathcal{H}_B \) given by \( S|x\rangle\langle y| = |y\rangle\langle x| \) for any two states \( |x\rangle \) and \( |y\rangle \). Here we have employed the fact that \( S_{AB}U_{AC}S_{AB} \) is a two-qubit gate acting on \( BC \), and \( S_{AB}U_{AB} \) and \( V_{AB}S_{AB} \) are two-qubit gates acting on \( AB \).

Case 2.2. The circuit is represented by \( U_{AB}U_{BC}U_{AC}V_{AB} = V_{ABC} \). We know that \( U_{BC}U_{AC} = \text{a controlled gate with control on system } C \). Directly, we observe that \( U_{BC} \) and \( U_{AC} \) are controlled gates with control on \( C \). This leads us to the conclusion that \( I - 2|11\rangle\langle 11| \) shares eigenvalues, counting multiplicity with a local unitary, which means that the product of two eigenvalues of \( I - 2|11\rangle\langle 11| \) equals the product of the other two. This is impossible.

By summarizing the above arguments, we have shown that four two-qubit gates are not sufficient for simulating the Toffoli gate. This further implies that any circuit consisting of fewer than five two-qubit gates has a positive distance to the Toffoli gate since the set of three-qubit gates that can be implemented by using up to four two-qubit gates forms a compact set; in other words, the Toffoli gate cannot be well approximated by such circuits.

This proof technique above can also be used to show that the following three-qubit controlled phase gate (three-qubit quantum gates with two control systems and one target qubit) introduced by Deutsch [6] cannot be implemented by four two-qubit gates:
\[ V_{\theta} = I - (1 - e^{i\theta})|111\rangle\langle 111|, \]
where \( 0 < \theta < 2\pi \). Note that \( V_{ABC} \) is the special case of \( \theta = \pi \). Together with the result in [8], we conclude that five two-qubit gates are optimal for simulating the the two-qubit controlled phase gate.

More generally, by employing the proof technique here, we completely solve the optimal simulation for general three-qubit controlled gates [7,8] and Fredkin gates [10]:

Theorem 1. The three-qubit controlled unitary \( C^2(U) \) is implementable by four two-qubit gates if and only if \( \det(U) \) is odd; otherwise, five is optimal. Five two-qubit gates are optimal for implementing the Fredkin gate.

A detailed proof of this theorem is presented in [40].

In this Rapid Communication, we study the problem of implementing a multiqubit gate using two-qubit unitaries. In particular, we demonstrate that four two-qubit unitaries are not sufficient for constructing the three-qubit Toffoli gate; thus, the implementation with five two-qubit gates is optimal.

More generally, our idea can be used to characterize the two-qubit gate cost of implementing a three-qubit controlled gate. We hope this work will be helpful for further study of the minimal cost of implementing larger quantum logical gates, e.g., the multiqubit controlled gate, and for studying optimization of quantum logical circuits, a crucial issue in the design and implementation of quantum computer hardware and architecture.

We are grateful to Professor D. P. DiVincenzo for pointing out useful references. This work was partly supported by the Australian Research Council (ARC) (Grants No. DP110103473 and No. DP120103776) and the Overseas Team Program of the Academy of Mathematics and Systems Science, Chinese Academy of Sciences. R. Duan was also supported in part by the National Natural Science Foundation of China under Grant 61179030 and an ARC Future Fellowship under Grant FT120100449.

APPENDIX

In this Appendix, we show that there are no unitaries \( U_{AB} \), \( V_{AB} \) and \( U_{BC} \), \( V_{BC} \) such that
\[ U_{AB}U_{BC}V_{AB}V_{BC} = V_{ABC}. \]

Notice that \( U_{AB}U_{BC}V_{AB} \) is a controlled gate on the bipartition \( A-BC \) with control on \( A \). Moreover, part A’s state of the output state \( U_{AB}U_{BC}V_{AB}|i\rangle_A|\psi\rangle_{BC} \) is still \( |i\rangle_A \) for any input state \( |i\rangle_A|\psi\rangle_{BC} \) with \( i = 0, 1 \). Since \( V_{AB} \) maps some state \( |0\rangle_A|\xi\rangle_B \) to a product state, we can assume that \( V_{AB}|0\rangle_A|0\rangle_B = |0\rangle_A|0\rangle_B \) by absorbing one-qubit gates into \( U_{BC} \) and \( U_{AB} \).

Then the state of A’s part of \( U_{AB}U_{BC}V_{AB}|0\rangle_A|0\rangle_B|\psi\rangle_C = U_{AB}U_{BC}|0\rangle_A|0\rangle_B|\psi\rangle_C \) is still \( |0\rangle_A \). We now need to consider three cases according to different forms of the state \( U_{BC}|0\rangle_B|\psi\rangle_C \):
Case 1. There is some $|z\rangle_C$ such that $U_{BC}|0\rangle_B|z\rangle_C$ is entangled. Assume that there is $0 < \lambda < 1$ such that

$$U_{BC}|0\rangle_B|z\rangle_C = \sqrt{\lambda}|0\rangle_B|\psi^+_C\rangle + \sqrt{1 - \lambda}|1\rangle_B|\psi^-_C\rangle,$$

where we have absorbed a local unitary acting on $B$ into $U_{AB}$. Let $(\Phi) = U_{AB}|00\rangle$ and $|\Psi\rangle = U_{AB}|01\rangle$. Then

$$|\lambda\rangle_{ABC} = U_{AB}U_{BC}|0\rangle_A|0\rangle_B|z\rangle_C = \sqrt{\lambda}(\Phi)_{AB}|\alpha\rangle_C + \sqrt{1 - \lambda}|\Psi\rangle_{AB}|\alpha^-\rangle_C.$$

Thus, we obtain

$$\lambda_A = |0\rangle\langle 0| = \chi \Phi_A + (1 - \lambda) \Psi_A,$$

$$\Rightarrow \Phi_A = \Psi_A = |0\rangle\langle 0|.$$

Consequently, $U_{AB}$ is a controlled gate with control on system $A$, and one knows that $V_{AB} = U_{BC} U_{AB} V_{ABC} V_{AB}^\dagger$ is a controlled gate with control $A$. Assume that $U_{AB} = |0\rangle\langle 0| \otimes I_B + |1\rangle\langle 1| \otimes u_B$ and $V_{AB} = |0\rangle\langle 0| \otimes I_B + |1\rangle\langle 1| \otimes v_B$. We conclude that $U_{ABC} V_{AB} U_{BC}^\dagger = u_B |0\rangle\langle 0| + \phi_B Z_B |1\rangle\langle 1|$, where $Z_B$ is the Pauli matrix given by $Z|0\rangle = |0\rangle$ and $Z|1\rangle = -|1\rangle$.

The set of eigenvalues of $U_{BC} v_B U_{BC}^\dagger$ counting multiplicity is $\{e^{i\theta}, e^{i\theta}, e^{-i\theta}, e^{-i\theta}\}$, which is also the set of eigenvalues counting the multiplicity of the right-hand side of the above equality. Note that $u_B$ should not equal the identity up to a global phase. Then $u_B$ and $u_B Z_B$ have the same set of eigenvalues. Thus their determinants are equal, say,

$$\det(u_B) = \det(u_B Z_B) = \det(u_B) \det(Z_B) = -\det(u_B),$$

and $\det(u_B^\dagger) = 0$. This contradicts the fact that $u_B$ is unitary. Thus $U_{BC}|0\rangle_B|z\rangle_C$ is always a product for any $|z\rangle_C$. This leads us to consider the following two cases.

Case 2. There is a $|\gamma\rangle_C$ and a local unitary $w_B$ on system $B$ such that $U_{BC}|0\rangle_B|\gamma\rangle_C = w_B|\gamma\rangle_C$. Then $U_{ABC}$ maps $|0\rangle_A \otimes H_B$ to itself; hence $U_{AB}$ is a controlled gate with control system $A$. Similarly, $V_{AB}$ is also a controlled gate with the same control bit. The rest of the proof is the same as in Case 1.

Case 3. There is a state on system $B$, say $|0\rangle_B$, and a local unitary $w_C$ on system $C$ such that $U_{BC}|0\rangle_B|z\rangle_C = |0\rangle_B w_C|z\rangle_C$.

Then $U_{BC}$ is a controlled gate with control system $B$. By moving this $w_C$ into $V_{BC}$, we can assume that $U_{BC} = |0\rangle\langle 0| \otimes I_C + |1\rangle\langle 1| \otimes u_C$. Note that for any $|z\rangle_C$, part C’s state of the output state $|\lambda\rangle_{ABC} = U_{AB} U_{BC} V_{ABC} |0\rangle_A |0\rangle_B |z\rangle_C = U_{AB} |0\rangle_A |0\rangle_B |z\rangle_C$ is still $|z\rangle_C$. By recalling that $|\lambda\rangle_{ABC} = V_{ABC} (U_{BC}^\dagger |0\rangle_B |z\rangle_C) = |0\rangle_A (V_{BC}^\dagger |1\rangle_B |z\rangle_C)$. We know that part C’s state of $V_{BC}^\dagger |1\rangle_B |z\rangle_C$ is $|z\rangle_C$ for all $|z\rangle_C \in \mathcal{H}_C$, which means that there is $|\beta\rangle_B$ such that $V_{BC} |\beta\rangle_B |z\rangle_C = |0\rangle_B |z\rangle_C$. Therefore, one can find a unitary $v_C$ such that $V_{BC} = |0\rangle \langle \beta| \otimes I_C + |1\rangle \langle \beta^\dagger| \otimes v_C$. In order to simplify the structure of the two-qubit gates, we observe that $V_{BC}^\dagger V_{BC}^\dagger U_{BC} V_{BC}^\dagger U_{BC} = U_{ABC}$. This also provides a simulation of $V_{ABC}$. Consider the state

$$V_{BC}^\dagger V_{BC}^\dagger U_{ABC}^\dagger U_{AB}^\dagger |0\rangle_A |0\rangle_B |x_A\rangle = V_{BC}^\dagger V_{BC}^\dagger U_{ABC}^\dagger U_{AB}^\dagger |0\rangle_A |0\rangle_B |x_A\rangle$$

for any $|x\rangle_A$. The argument of Cases 1 and 2 excludes the following possibilities: (i) there is some $|x\rangle_A$ such that $V_{ABC} |0\rangle_B |x\rangle_A$ is entangled, or (ii) there are a $|\beta\rangle_B$ and a local unitary $w_B$ on system $B$ such that $V_{ABC}^\dagger |0\rangle_A |0\rangle_B |x\rangle_A = w_B |x\rangle_B |\delta\rangle_A$. So the only possibility is that there is a state $|\phi\rangle_B$ on system $B$ and a local unitary $w_B$ on system $A$ such that $V_{ABC}^\dagger |0\rangle_A |0\rangle_B |x\rangle_A = |\phi\rangle_B w_B |x\rangle_A$. According to $V_{ABC}^\dagger |0\rangle_A |0\rangle_B |x\rangle_A = |0\rangle_A |0\rangle_B |x\rangle_A$, we can choose $|\phi\rangle = |0\rangle$. Thus $V_{ABC}^\dagger$ is a controlled gate with control system $B$, i.e.,

$$V_{ABC} = |0\rangle\langle 0| \otimes w_A + |1\rangle\langle 1| \otimes v_A.$$

By studying part C’s state of $U_{ABC} U_{BC} V_{ABC} |0\rangle_A |0\rangle_B |z\rangle_C = |0\rangle_A |0\rangle_B |z\rangle_C$, we see that $|\beta\rangle_B$ defined in $V_{BC}$ equals $|0\rangle_B$ or $|1\rangle_B$, up to some global phase. Otherwise, assume that $|0\rangle_B = a |\beta\rangle_B + b |\beta^\dagger\rangle_B$ for $ab \neq 0$. Then the state of part C becomes a mixed state for general input $|0\rangle_B |z\rangle_C$ since $u_C$ is not the identity up to some global phase and $U_{BC}$ is nonlocal. For the case $|\beta\rangle_B = |0\rangle_B$, we know that all the four two-qubit gates are controlled gates with control system $B$, which implies that $I - 2|11\rangle|11\rangle$ is a local unitary, a contradiction. For the case $|\beta\rangle_B = |1\rangle_B$, let $X_B$ be the NOT (flip) gate such that $X|0\rangle = |1\rangle$ and $X|1\rangle = |0\rangle$, one can verify that

$$(U_{AB} X_B)(X_B U_{BC} X_B)(X_B V_{AB} X_B)(X_B V_{BC}) = V_{ABC}.$$

$U_{AB} X_B$, $X_B U_{BC} X_B$, $X_B V_{AB} X_B$, and $X_B V_{BC}$ are controlled gates with control system $B$. This also leads us to a contradiction that $I - 2|11\rangle|11\rangle$ is local.

[38] N. Margolus (unpublished).