What is the Computational Value of Finite-Range Tunneling?

Vasil S. Denchev,[1,*] Sergio Boixo,[1,†] Sergei V. Isakov,[1] Nan Ding,[1] Ryan Babbush,[1]
Vadim Smelyanskiy,[1] John Martinis,[2] and Hartmut Neven[1]
[1]*Google Inc., Venice, California 90291, USA*
[2]*Google Inc., Santa Barbara, California 93117, USA*
(Received 4 March 2016; revised manuscript received 22 June 2016; published 1 August 2016)

Quantum annealing (QA) has been proposed as a quantum enhanced optimization heuristic exploiting tunneling. Here, we demonstrate how finite-range tunneling can provide considerable computational advantage. For a crafted problem designed to have tall and narrow energy barriers separating local minima, the D-Wave 2X quantum annealer achieves significant runtime advantages relative to simulated annealing (SA). For instances with 945 variables, this results in a time-to-99%-success-probability that is $\sim 10^8$ times faster than SA running on a single processor core. We also compare physical QA with the quantum Monte Carlo algorithm, an algorithm that emulates quantum tunneling on classical processors. We observe a substantial constant overhead against physical QA: D-Wave 2X again runs up to $\sim 10^8$ times faster than an optimized implementation of the quantum Monte Carlo algorithm on a single core. We note that there exist heuristic classical algorithms that can solve most instances of Chimera structured problems in a time scale comparable to the D-Wave 2X. However, it is well known that such solvers will become ineffective for sufficiently dense connectivity graphs. To investigate whether finite-range tunneling will also confer an advantage for problems of practical interest, we conduct numerical studies on binary optimization problems that cannot yet be represented on quantum hardware. For random instances of the number partitioning problem, we find numerically that algorithms designed to simulate QA scale better than SA. We discuss the implications of these findings for the design of next-generation quantum annealers.

DOI: 10.1103/PhysRevX.6.031015

Subject Areas: Computational Physics, Quantum Physics, Quantum Information

I. INTRODUCTION

Simulated annealing (SA) [1] is perhaps the most widely used algorithm for global optimization of pseudo-Boolean functions with little known structure. The objective function for this general class of problems is

$$H_P^{\mathrm{cl}}(\mathbf{s}) = -\sum_{k=1}^{K} \sum_{j_1 \ldots j_k = 1}^{N} J_{j_1 \ldots j_k} s_{j_1} \ldots s_{j_k}, \tag{1}$$

where N is the problem size, $s_j = \pm 1$ are spin variables, and the couplings $J_{j_1 \ldots j_k}$ are real scalars. In the physics literature, $H_P^{\mathrm{cl}}(\mathbf{s})$ is known as the Hamiltonian of a K-spin model. SA is a Monte Carlo algorithm designed to mimic the thermalization dynamics of a system in contact with a slowly cooling reservoir. When the temperature is high, SA induces thermal excitations that can allow the system to escape from local minima. As the temperature decreases, SA drives the system towards nearby low-energy spin configurations.

Almost two decades ago, quantum annealing (QA) [2] was proposed as a heuristic technique for quantum enhanced optimization. Despite substantial academic and industrial interest [3–34], a unified understanding of the physics of quantum annealing and its potential as an optimization algorithm remains elusive. The appeal of QA relative to SA is due to the observation that quantum mechanics allows for an additional escape route from local minima. While SA must climb over energy barriers to escape false traps, QA can penetrate these barriers without any increase in energy. This effect is a hallmark of quantum mechanics, known as quantum tunneling. The standard time-dependent Hamiltonian used for QA is

$$H(t) = -A(t) \sum_{j=1}^{N} \sigma_j^x + B(t) H_P, \tag{2}$$

where H_P is written as in Eq. (1) but with the spin variables s_j replaced with σ_j^z Pauli matrices acting on qubit j, and the functions $A(t)$ and $B(t)$ define the annealing schedule parametrized in terms of time $t \in [0, T_{\mathrm{QA}}]$ (see Fig. 10). These annealing schedules can be defined in many different ways as long as the functions are smooth, $A(0) \gg B(0)$ and $A(T_{\mathrm{QA}}) \ll B(T_{\mathrm{QA}})$. At the beginning of the annealing, the

*Corresponding author.
denchev@google.com
†Corresponding author.
boixo@google.com

transverse field magnitude $A(t)$ is large, and the system dynamics are dominated by quantum fluctuations due to tunneling (as opposed to the thermal fluctuations used in SA).

The question of whether D-Wave processors realize computationally relevant quantum tunneling has been a subject of substantial debate. This debate has now been settled in the affirmative with a sequence of publications [6,8,9,12,13,18,21] demonstrating that quantum resources are present and functional in the processors. In particular, Refs. [35,36] studied the performance of the D-Wave device on problems where eight [37] qubit cotunneling events were employed in a functional manner to reach low-lying energy solutions.

In order to investigate the computational value of finite-range tunneling in QA, we study the scaling of the exponential dependence of annealing time with the size of the tunneling domain D,

$$T_{\mathrm{QA}} = B_{\mathrm{QA}} e^{\alpha D}, \tag{3}$$

where $\alpha = a_{\min}/\hbar$ and $a_{\min}$ is the rescaled instanton action [see Eqs. (24) and (27)]. In SA, the system escapes from a local minimum via thermal fluctuations over an energy barrier ΔE separating the minima. The time required for such events scales as

$$T_{\mathrm{SA}} = B_{\mathrm{SA}} e^{\Delta E/k_B T}, \tag{4}$$

which is exponentially long with respect to ΔE. However, for sufficiently tall and narrow barriers such that

$$\frac{\Delta E}{k_B T} > \alpha D, \tag{5}$$

QA can overcome barriers exponentially faster than SA. This situation was studied in Ref. [38] and it also occurs in the benchmark problems studied in this paper.

The path-integral quantum Monte Carlo (QMC) method is used for sampling the quantum Boltzmann distribution of a d-dimensional stoquastic Hamiltonian as a marginalization of a classical Boltzmann distribution of an associated $d+1$ dimensional Hamiltonian. For specific cases, it was recently shown that the exponent α for physical tunneling is identical to the corresponding quantity for the QMC algorithm [39]. However, in the present work, we find a very substantial computational overhead associated with the prefactor B_{QMC} in the expression for the runtime of the QMC algorithm, $T_{\mathrm{QMC}} = B_{\mathrm{QMC}} e^{\alpha D}$. In other words, B_{QMC} can exceed B_{QA} by many orders of magnitude. The role of this prefactor becomes essential in the situations where the number of cotunneling qubits D is finite, i.e., is independent of the problem size N (or depends on N very weakly).

Because tunneling is more advantageous when energy barriers are tall and narrow, we expect this resource to be most valuable in the upper part of the energy spectrum. For instance, a random initial state is likely to have an energy well above the ground-state energy for a difficult optimization problem such as the one in Eq. (1). However, the closest lower energy local minimum will often be less than a dozen spin flips away. Nevertheless, the energy barriers separating these minima may still be high. In such situations, if the transverse field is turned on to facilitate tunneling transitions, the transition rate to lower energy minima will often increase. By contrast, once the state reaches the low part of the energy spectrum, the closest lower local minimum is asymptotically N spin flips away [2,40–42]. There, finite-range tunneling may assist by effectively "chopping off" narrow energy ridges near the barrier top, but the transition probability is still largely given by the Boltzmann factor. This description illustrates that finite-range tunneling can be useful to quickly reach an approximate optimization, but will not necessarily significantly outperform SA when the task is to find the ground state (see Fig. 1).

The canonical QA protocol initializes the system in the symmetric superposition state $|+\rangle^{\otimes N}$, which is the ground state at $t = 0$. By a similar argument, we expect that finite-range tunneling will drive the system adiabatically across energy gaps associated with narrow barriers, preventing transitions to higher energies. However, in general, finite-range tunneling will not be able to prevent Landau-Zener diabatic transitions for very small gaps resulting from emerging minima in the energy landscape separated by a wide barrier. This will often include the gap separating the ground state from the first excited state [2,40–42].

This paper is organized as follows: In Sec. II, we present our main results consisting of benchmarking the D-Wave 2X processor against SA and the QMC algorithm on a crafted problem with a rugged energy landscape; Sec. III introduces the theory of instantons in multispin systems, discusses tunneling simulation in the QMC algorithm, and presents numerical results from theoretical modeling comparing QA and QMC calculations for the "weak-strong cluster pair" problem; Sec. IV presents numerical studies of generic problems with rugged energy landscapes that can potentially benefit from QA and discusses the challenges associated with designing future annealers of practical relevance; Sec. V concludes with an overview and discussion. Further technical details can be found in Appendixes A and B [43].

II. BENCHMARKING PHYSICAL QUANTUM ANNEALING ON A CRAFTED PROBLEM WITH A RUGGED ENERGY LANDSCAPE

Here, we consider a problem designed so that finite cotunneling transitions of multiple spins strongly impacts the success probability. The previous section outlined several reasons why QA has a chance to outperform SA for problems with a rugged energy landscape.

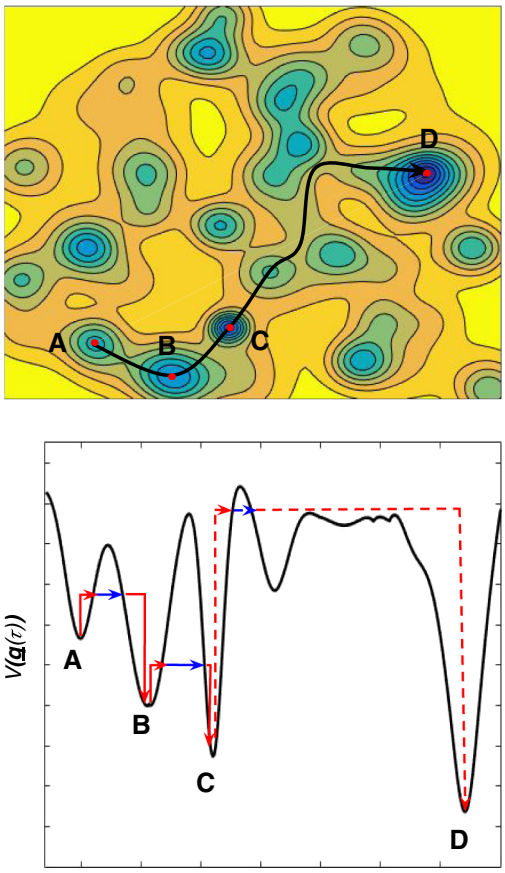

FIG. 1. Top: Quantum annealer dynamics are dominated by paths through the mean-field energy landscape that have the highest transition probabilities. Shown is a path that connects local minimum A to local minimum D. Bottom: Mean-field energy $V(q(\tau))$ along the path from A to D, as defined by Eqs. (12) and (16). Finite-range, thermally assisted tunneling can be thought of as a transition consisting of three steps: (i) the system picks up thermal energy from the bath (red arrow up), (ii) the system performs a tunneling transition between the entry and exit points (blue arrow), and (iii) the system relaxes to a local minimum by dissipating energy back into the thermal bath (red arrow down). In transitions A → B or B → C, finite-range tunneling considerably reduces the thermal activation energy needed to overcome the barrier. For long distance transitions in the lower part of the energy spectrum, such as C → D, the transition rate is still dominated by the thermal activation energy, and the increase in transition rate brought about by tunneling is negligible.

In previous work we proposed a problem consisting of a pair of strongly connected spins (called clusters) to study the presence of functional cotunneling in D-Wave processors [35,36]. Each cluster coincides with a unit cell of the native hardware graph, known as the Chimera graph. The problem Hamiltonian H_P in Eq. (2) is of Ising form:

$$H_P = H_P^1 + H_P^2 + H_P^{1,2}, \tag{6}$$

$$H_P^k = -J \sum_{\langle j,j' \rangle \in \text{intra}} \sigma_{k,j}^z \sigma_{k,j'}^z - \sum_{j=1}^{8} h_k \sigma_{k,j}^z, \tag{7}$$

$$H_P^{1,2} = -J \sum_{j \in \text{inter}} \sigma_{1,j}^z \sigma_{2,j}^z. \tag{8}$$

All the couplings are ferromagnetic with $J = 1$. The index $k \in \{1, 2\}$ indexes a unit cell of the Chimera graph, the first sum in Eq. (7) goes over the *intracell* couplings depicted in Fig. 2, the second sum goes over the *intercell* couplings corresponding to $j = 5, 6, 7, 8$ in Fig. 2, and h_k denotes the local fields within each cell.

The local fields $0 < h_1 < 0.44$ (weak cluster) and $h_2 = -1$ (strong cluster) are equal for all the spins within the cell. In this parameter regime, all spins of both clusters will point along the direction of the strong local fields in the ground state of the problem Hamiltonian H_P. However, in the initial phase of the annealing evolution, all spins in the weak cluster orient themselves by following the local field in the opposite direction. At a later stage of the annealing evolution, the pairwise coupling between clusters becomes dominant; however, in the mean-field picture, the state of the weak cluster is driven into a local minimum. Using a noise model with experimentally measured parameters for the D-Wave 2X processor, we numerically verify that the most likely mechanism by which all spins arrive at the energetically more favorable configuration is multiqubit cotunneling (see Ref. [35] and Appendix A [43]).

Using the weak-strong cluster pairs as building blocks, larger problems are formed by connecting the strong clusters to one another in a glassy fashion. That is, the four connections between two neighboring strong clusters

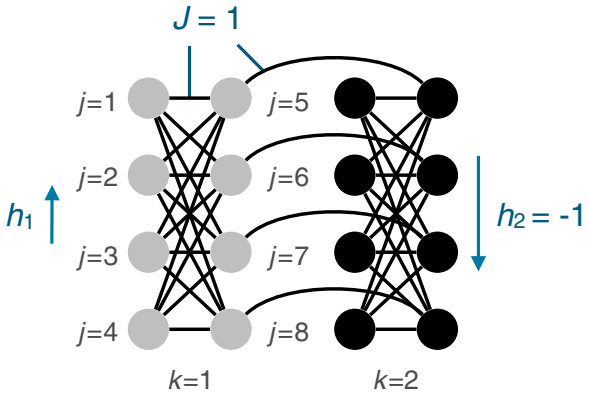

FIG. 2. A pair of weak-strong clusters, consisting of 16 qubits in two unit cells of the Chimera graph. All qubits are ferromagnetically coupled and evolve as part of two distinct qubit clusters. At the end of the annealing evolution, the right cluster is strongly pinned downwards due to strong local fields acting on all qubits in that cell. However, the local magnetic field h_1 in the left cluster is weaker and serves as a bifurcation parameter. For $h_1 < 1/2$, the left cluster will reverse its orientation during the annealing sweep and eventually align itself with the right cluster.

are all set to either $+1$ (ferromagnetic) or -1 (antiferromagnetic), at random. With this procedure, we define a large class of instances having any size that we refer to as the "weak-strong cluster networks" problem. We expect that for large sizes beyond the current D-Wave 2X processor the spin-glass backbone of this problem will dominate the computational hardness and finite-range tunneling will no longer be sufficient for reaching the ground state. This is also consistent with numerical studies of the energy landscape conducted in Ref. [44]. Even so, as suggested by Fig. 1, under those circumstances it would be interesting to study the role of finite-range tunneling in achieving good approximate solutions. Figure 3 shows several examples of the layout of instances that were used in these benchmark tests [45]. Because of the fact that not all of the qubits in the D-Wave 2X processor are properly calibrated, and hence available for computation, the instances become somewhat irregular.

A. D-Wave versus simulated annealing

We now compare the total annealing time of SA to the total annealing time of the D-Wave 2X processor. Figure 4 shows the time to reach the ground state with 99% success probability, as a function of problem size for different quantiles.

For D-Wave, we fix the annealing time at 20 μs, the shortest time available due to engineering compromises, and estimate the probability p_j of finding the ground state for a given instance j [46]. Shorter times are optimal in this benchmark, as explained in Appendix A and Ref. [35]. The total annealing time to achieve $p_j = 0.99$ is $20\ \mu\text{s}[\log(1\text{–}0.99)/\log(1-p_j)]$. See Appendix A [43] for details of the physical QA parameters used in D-Wave machines (see also Ref. [47]).

We measure the computational effort of SA in units of runtime on a single core, which is natural when using server centers and convenient in order to compare to other classical approaches. Of course, the total runtime can be shortened by parallelizing the overall computation. Part of that process is embarrassingly parallel since SA finds its best solutions not in a single run but by restarting from random bit configurations. Every restart could be executed on a different core. In fact, we use this strategy for our numerical benchmark. We find that median-case instances

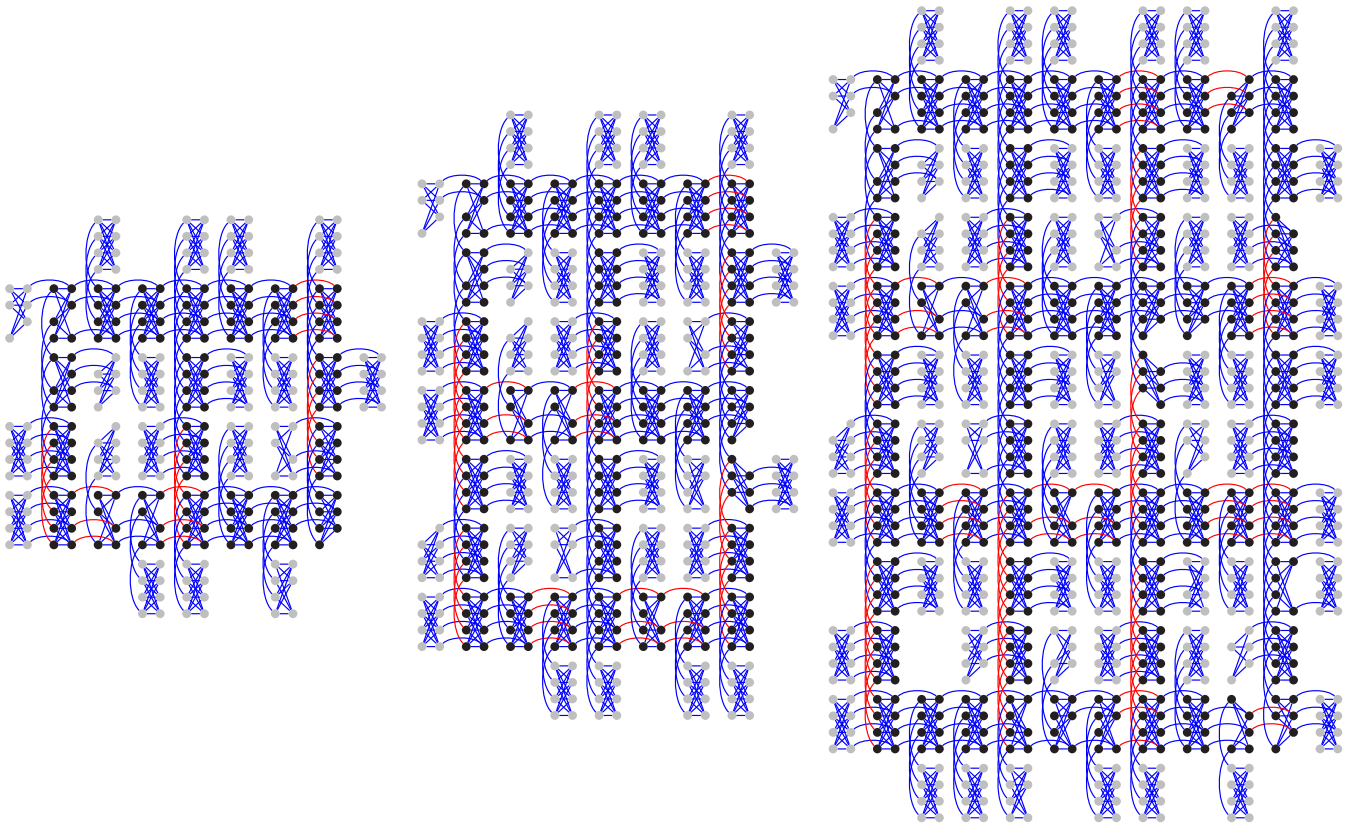

FIG. 3. Layout of several instances of the weak-strong cluster network problem on the D-Wave 2X processor. Shown are three different sizes with 296, 489, and 945 qubits. Each cluster consists of an eight-qubit unit cell of the Chimera graph. Black dots depict qubits subject to a strong local field $h_R = -1$ while the gray dots represent qubits with the weak field $h_L = 0.44$. Blue lines correspond to strong ferromagnetic couplings ($J = 1$) and red lines to strong antiferromagnetic couplings ($J = -1$). Note that the graphs are somewhat irregular due to the fact that not all 1152 qubits are operational.

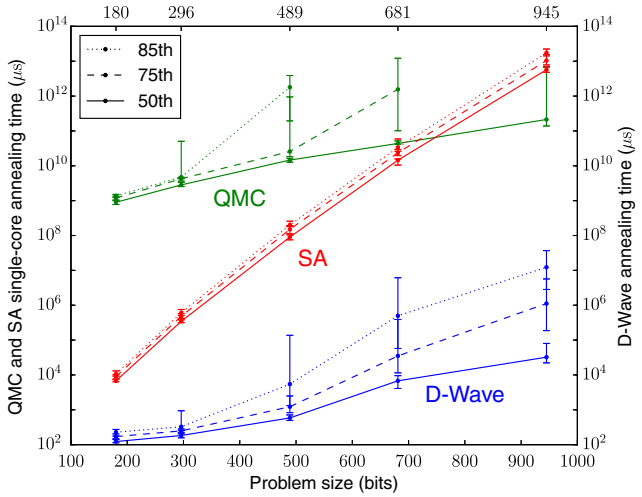

FIG. 4. Time to find the optimal solution with 99% probability for different problem sizes. We compare simulated annealing (SA), the quantum Monte Carlo (QMC) method, and the D-Wave 2X. To assign a runtime for the classical algorithms we take the number of spin updates (for SA) or worldline updates (for QMC method) that are required to reach a 99% success probability and multiply that with the time to perform one update on a single state-of-the-art core. Shown are the 50th, 75th, and 85th percentiles over a set of 100 instances. The error bars represent 95% confidence intervals from bootstrapping. This experiment occupied millions of processor cores for several days to tune and run the classical algorithms for these benchmarks. The runtimes for the higher quantiles for the larger problem sizes for the QMC algorithm were not computed because the computational cost was too high. For a similar comparison with the QMC method with different parameters, please see Fig. 12.

require 10^9 independent runs (with $945 \times 5 \times 10^4$ spin updates each) to find the optimal solution with 945 variables, on average.

We note that for instances with more qubits, more quantum hardware resources are brought to bear and therefore a fair comparison needs to take this into account [12,15]. As an extreme example, one could contemplate building special purpose classical hardware that would update as many spins in parallel as possible at state-of-the-art clock rates. The sets of spins that could be updated in parallel depends on the connectivity graph. Though such considerations are reasonable, we do not explore this possibility here as it is well known that graphs with sufficiently dense connectivity will severely limit the usefulness of such parallelism.

When estimating runtimes for SA, we follow the protocol laid out by Refs. [12,15,48] and tune SA for every problem size and quantile. Tuning means that the starting and end temperature, as well as the number of spin updates and the number of restarts, are optimized to achieve a short overall runtime. We first measure the computational effort in units of sweeps (one sweep attempts to update all the spins). These times are plotted as $n_{\text{sweeps}} N T_{\text{su}}$, where N is the number of spins. We use a spin update time $T_{\text{su}} = 1/5$ ns (see Ref. [12]).

Our key finding in this comparison is that SA performs very poorly on the weak-strong cluster networks. The D-Wave 2X processor is 1.8×10^8 faster at the largest size instances we investigate, which consists of 945 variables. This problem is specifically engineered to cause the failure of SA: as we explain above, the "weak-strong cluster networks" problem is intended to showcase the performance of annealers on a problem that benefits from finite-range cotunneling. By contrast, the random Ising instances studied in Refs. [12,15] have only low-energy barriers, as explained in Ref. [49].

B. D-Wave versus quantum Monte Carlo method

Next, we compare the performance of the path-integral quantum Monte Carlo method with that of D-Wave for the same benchmark. The QMC method samples the Boltzmann distribution of a classical Hamiltonian which approximates the transverse field Ising model. In the case of a 2-spin model, the discrete imaginary time QMC classical Hamiltonian is

$$H_{\text{cl}} = -\sum_{\tau=1}^{M}\left(\sum_{jk}\frac{J_{jk}}{M}\sigma_j(\tau)\sigma_k(\tau) + J^{\perp}(s)\sum_{j}\sigma_j(\tau)\sigma_j(\tau+1)\right), \tag{9}$$

where $\sigma_j(\tau) = \pm 1$ are classical spins, j and k are site indices, τ is a replica index, and M is the number of replicas. The coupling between replicas is given by

$$J^{\perp}(s) = -\frac{1}{2\beta}\ln\tanh\frac{A(s)\beta}{M}, \tag{10}$$

where β is the inverse temperature. The set of configurations for a given spin j across all replicas τ is called the worldline of spin j. Periodic boundary conditions are imposed between $\sigma_j(M)$ and $\sigma_j(1)$. We use the continuous path-integral QMC method, which corresponds to the limit $\Delta\tau \to 0$ [50], and, unlike the discrete path-integral QMC method, does not suffer from discretization errors of order $1/M$.

We numerically compute the number of sweeps n_{sweeps} required for QMC calculations to find the ground state with 99% probability at different quantiles. In our case, a sweep corresponds to two update attempts for each worldline. The computational effort is $n_{\text{sweeps}} N T_{\text{worldline}}$, where N is the number of qubits and $T_{\text{worldline}}$ is the time to update a worldline. We average $T_{\text{worldline}}$ over all the steps in the quantum annealing schedule; however, the value of $T_{\text{worldline}}$ depends on the particular schedule chosen. As explained above for SA, we report the total computational

effort of the QMC algorithm in standard units of time per single core. For the annealing schedule used in the current D-Wave 2X processor, we find

$$T_{\text{worldline}} = \beta(870\ \text{ns}), \tag{11}$$

using an Intel(R) Xeon(R) CPU E5-1650 @ 3.20 GHz. The reason why $T_{\text{worldline}}$ is approximately linear in β is because the number of spin flips in a worldline grows as β increases [see Eq. (10)], which affects the runtime of the continuous path-integral QMC method [12].

This study is designed to explore the utility of the QMC algorithm as a classical optimization routine. Accordingly, we optimize the QMC algorithm by running at a low temperature, 4.8 mK. We also observe that the QMC algorithm with open boundary conditions (OBC) performs better than the standard QMC algorithm with periodic boundary conditions in this case [39]; therefore, OBC is used in this comparison. We further optimize the number of sweeps per run which, for a given quantile, results in the lowest total computational effort. We find that the optimal number of sweeps for the 50th percentile at the largest problem size is 10^6. This enhances the ability of the QMC algorithm to simulate quantum tunneling, and gives a very high probability of success per run in the median case, $p_{\text{success}} = 0.16$.

All the qubits in a cluster have approximately the same orientation in each local minimum of the effective mean-field potential. Neighboring local minima typically correspond to different orientations of a single cluster. Here, the tunneling time is dominated by a single purely imaginary instanton and is described by Eq. (27) below. It was recently demonstrated that, in this situation, the exponent $a_{\text{min}}/\hbar$ for physical tunneling is identical to that of the QMC method [39]. As seen in Fig. 4, we do not find a substantial difference in the scaling of the QMC algorithm and D-Wave. However, we find a very substantial computational overhead associated with the prefactor B in the expression $T = Be^{Da_{\text{min}}/\hbar}$ for the runtime. In other words, B_{QMC} can exceed B_{QA} by many orders of magnitude. The role of the prefactor becomes essential in situations where the number of cotunneling qubits D is finite, i.e., is independent of the problem size N (or depends on N very weakly). Between some quantiles and system sizes we observe a prefactor advantage as high as 10^8.

We note that the QMC and D-Wave error bars are consistently larger than SA's error bars. This is due to the sensitivity of the confidence interval bootstrapping procedure to the fact that QMC and D-Wave's mean performance results are significantly more spread out across quantiles than SA's. The latter phenomenon is explained by the presence of instances whose glassy backbones necessitate long-range tunneling. The success probabilities of the QMC method and D-Wave are significantly lowered on such instances, while the success probabilities of SA are already severely diminished by the individual weak-strong cluster pairs. Similarly to observations made in Ref. [30], this suggests that, for the tails of the complexity distributions, the scalings might be substantially different. As illustrated by Fig. 1 and suggested in Ref. [51], in situations where long-range tunneling is required for reaching the global minimum, quantum annealing may still be advantageous for reaching good approximate solutions.

C. D-Wave versus other classical solvers

Based on the results we present here, one cannot claim a quantum speedup for D-Wave 2X, as this would require that the quantum processor in question outperforms the best known classical algorithm (see also Ref. [44], which follows up on the study of this paper). This is not the case for the weak-strong cluster networks. This is because a variety of heuristic classical algorithms can solve most instances of Chimera structured problems much faster than SA, the QMC algorithm, and the D-Wave 2X [52–54] (for a possible exception, see Refs. [25,51]) [55]. For instance, the Hamze–de Freitas–Selby algorithm [52,53] performs a greedy sequence of random large neighborhood optimizations. Each large neighborhood is defined by first replacing each 4-qubit column in a cluster with a large spin, and then expanding a tree of large spins which covers ∼77% of all spins (and more than half of all 8-qubit clusters). In the particular case of a weak-strong cluster pair, this algorithm avoids the formation of the energy barrier.

However, it is well known that such solvers will become ineffective for sufficiently dense connectivity graphs. In such dense graphs, the largest tree-structured subgraphs become too small to be useful for large neighborhood search [52], and cluster moves become too costly or break down altogether due to lowered percolation thresholds [56]. Unsurprisingly, these two failure modes affect the tailored and nontailored algorithms studied in Ref. [44]. On the contrary, multiqubit cotunneling is a general phenomenon in spin glasses that is not limited to sparse graphs. Nevertheless, the situation for advanced connectivity graphs that can be engineered in practice remains an open question and is presently subject to active research.

We have also learned from the Janus team, working with special purpose FPGAs to thermalize Ising models on a 32^3 cube, that they found cluster finding not to be helpful [57].

1. A Remark on scaling

Certain quantum algorithms have asymptotically better scaling than the best-known classical algorithms. While such asymptotic behavior can be of tremendous value, this is not always the case. Moreover, there can be substantial value in quantum algorithms that do not show asymptotically better scaling than classical approaches. The first reason for this is that current quantum hardware is restricted to rather modest problem sizes of less than order 1000

qubits. At the same time, when performing numerical simulations of quantum dynamics, in particular when doing open system calculations, we are often limited to problem sizes smaller than 100 qubits. Extrapolating from such finite size findings can be misleading, as it is often difficult to determine whether a computational approach has reached its asymptotic behavior.

When forecasting the future promise of a given hardware design, there is a tendency to focus on the qubit dimension. However, this perspective is not necessarily helpful. For instance, when looking at runtime as a function of qubit dimension, one may conclude that the measurements we report here indicate that QMC calculations and physical annealing have a comparable slope, scale similarly, and that, therefore, the up side of physical annealing is bounded. However, the large and practically important prefactor depends on a number of factors such as temperature. Furthermore, we expect future hardware to have substantially richer connectivity graphs and dramatically improved T_1 and T_2 times. With such changes, next-generation annealers may drastically increase the constant separation between algorithms, leading to very different performance from generation to generation. To illustrate how dramatic this effect can be, when we ran smaller instances of the weak-strong cluster networks on the older D-Wave Vesuvius chips we predicted that at 1000 variables D-Wave would be 10^4 times faster than SA. In fact, we observe a speedup of more than a factor of 10^8. This is because certain noise parameters are improved and the new dilution refrigerator operates at a lower temperature. Similarly, we suspect that a number of previous attempts to extrapolate the D-Wave runtimes for 1000 qubits will turn out to be of limited use in forecasting the performance of future devices. For this reason, the current study focuses on runtime ratios that are actually measured on the largest instances solvable using the current device, rather than on extrapolations of asymptotic behavior which may not be relevant once we have devices that can attempt larger problems.

III. SPIN COTUNNELING IN QA AND THE QMC ALGORITHM

A. Instantons in systems with multiple spins

Cotunneling consists of system state transitions in which a group of spins simultaneously change their orientation with energy well below the energy of the (mean-field) potential barriers. Tunneling is a quintessential quantum phenomenon; real-time dynamics of classical trajectories cannot describe barrier penetration when the system wave function extends to classically forbidden regions. In such situations, the exponential decay of the wave function under the barrier is often captured through the path-integral formalism by computing the minimum action of the trajectories in imaginary time [58,59]. This approach was also extended to treat the tunneling of large magnetic moments with conserved total spin [60].

Tunneling in mean-field spin models can be described using the path integral over spin-coherent states in imaginary time [61]. The tunneling path connects the minima of the mean-field potential,

$$V(\underline{\mathbf{m}}, t) = \langle \Psi_{\underline{\mathbf{m}}} | H(t) | \Psi_{\underline{\mathbf{m}}} \rangle, \tag{12}$$

where $H(t)$ is the time-dependent QA Hamiltonian from Eq. (2) and $|\Psi_{\underline{\mathbf{m}}}\rangle$ is a product state

$$|\Psi_{\underline{\mathbf{m}}}\rangle = \bigotimes_j \left[\cos\frac{\theta_j}{2}|0\rangle + e^{-i\phi_j}\sin\frac{\theta_j}{2}|1\rangle \right]. \tag{13}$$

The coherent state of the jth spin is defined by a vector on the Bloch sphere,

$$\mathbf{n}_j = (\sin\theta_j\cos\phi_j, \sin\theta_j\sin\phi_j, \cos\theta_j). \tag{14}$$

and the corresponding state of the N-spin system is defined by the vector $\underline{\mathbf{m}} = (\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_N)$.

Towards the beginning of a QA evolution, the system remains near $\underline{\mathbf{m}}_0(t)$, the global minimum of the time-dependent potential $V(\underline{\mathbf{m}}, t)$, which connects to the global minimum at the initial time. Later on, $V(\underline{\mathbf{m}}, t)$ undergoes a bifurcation, which may cause the initial minimum to become metastable. At that point, the system may be able to tunnel to the new global minimum $\underline{\mathbf{m}}_1(t)$ when $V(\underline{\mathbf{m}}_0) \simeq V(\underline{\mathbf{m}}_1)$. Here, we omit the argument t, whose value corresponds approximately to the moment when the minima exchange order. Such tunneling events are sometimes accompanied by thermal activation if QA is performed at finite temperatures (i.e., thermally assisted tunneling) [38]. The sequence of bifurcations and associated tunneling events can continue multiple times before the global minimum of H_P is reached.

Typically, the quasiequilibrium Gibbs distribution associated with a given local minimum is formed on a time scale much shorter than the time scale of the tunneling decay of the metastable state. Therefore, the spin tunneling rate $W = W(t)$ can be described in terms of the imaginary part of the partition function [38,62]:

$$W_{\mathrm{QA}} = -\frac{2}{\beta}\frac{\mathrm{Im}(Z)}{\mathrm{Re}(Z)}, \qquad Z = \mathrm{Tr}\, e^{-\beta H}. \tag{15}$$

The partition function can be represented via path integral over the spin-coherent states Eqs. (13) and (14). The individual paths,

$$\underline{\mathbf{q}}(\tau) = [\mathbf{n}_1(\tau), \mathbf{n}_2(\tau), \ldots, \mathbf{n}_N(\tau)], \qquad \tau \in (0, \beta), \tag{16}$$

are periodic in imaginary time, $\underline{\mathbf{q}}(\tau) = \underline{\mathbf{q}}(\tau + \beta)$, where $\beta = \hbar/k_B T$ and T is the system temperature. The

multiqubit tunneling transition is a "rare event," corresponding to a a large group of spins $\{\mathbf{n}_j\}$ performing a concerted motion that connects the domains of the different minima of the effective potential V. Therefore, the tunneling can be described by a dominant path in the path integral that determines $\mathrm{Im}(Z)$ similarly to how it is done in the tunneling problems in continuous space [63]. The details of this analysis will be provided elsewhere [64]. Here, we simply outline the main argument.

From the spin path integral, the action along the instanton trajectory in imaginary time is

$$A = \frac{i\hbar}{2}\sum_{i=1}^{N}\omega\left[\mathbf{n}_j(\tau)\right] + \int_0^\beta d\tau V[\mathbf{n}_1(\tau),\ldots,\mathbf{n}_N(\tau)\Big], \quad (17)$$

where the first term corresponds to the sum over the Berry phases of individual spins:

$$\omega[\mathbf{n}(\tau)] = \int_0^\infty d\tau[1-\cos\theta(\tau)]\dot{\phi}(\tau).$$

The instanton trajectory corresponds to the extremum of the action $\delta A/\delta\mathbf{n}_j(\tau) = 0$. From here, the system of equations for the instanton path components has the form

$$\frac{\hbar}{2}\frac{d\mathbf{n}_j(\tau)}{d\tau} = \mathbf{n}_j(\tau)\frac{\partial V}{\partial\mathbf{n}_j(\tau)}, \qquad \mathbf{n}_j(\tau) = \mathbf{n}_j(\tau+\beta). \quad (18)$$

We note that the first (Berry phase) term in Eq. (17) contains an additional factor i compared to the second term. Therefore, at the instanton trajectory the vectors $\mathbf{n}_j(\tau)$ are complex [cf. Eqs. (67) and (68) in the Supplemental Material of Ref. [39] and also Ref. [62]]. They can be written in the form

$$\mathbf{n}_j = (\sin\theta_j\cosh\varphi_j, -i\sin\theta_j\sinh\varphi_j, \cos\theta_j), \quad (19)$$

corresponding to a purely imaginary azimuthal angle $\phi_j(\tau) = -i\varphi_j(\tau)$. This substitution makes the Berry phase terms in Eq. (17) purely real along the instanton path. The terms involving V are also real due to the fact that H is Hermitian. One can easily see this for the Hamiltonian H given in Eqs. (1) and (2) because in this case V depends only on azimuthal angles via $\cos\phi_j(\tau) = \cosh\varphi_j(\tau)$. Therefore, despite the presence of the imaginary Berry phase in Eq. (17), after the substitution Eq. (19) the instanton trajectory equations (18) involve only purely real quantities similar to the instantons for the particle in the potential. Similarly to this case, the initial point of the instanton $\underline{\mathbf{q}}(0)$ corresponds to the initial minimum of V where the instanton starts. The midpoint of the trajectory $\underline{\mathbf{q}}(\beta/2)$ typically corresponds to the exit point of the potential barrier in the vicinity of the final minima:

$$\underline{\mathbf{q}}(0) = \underline{\mathbf{q}}(\beta) \simeq \underline{\mathbf{m}}_0, \qquad \underline{\mathbf{q}}(\beta/2) \simeq \underline{\mathbf{m}}_1. \quad (20)$$

Finally, the transition rate is determined with logarithmic equivalence as

$$W_{\mathrm{QA}} B\exp(-A[q(\tau)]), \quad (21)$$

where $\mathbf{q}(\tau)$ is an instanton trajectory and the prefactor B can be, in principle, obtained in terms of the functional determinant of the kernel $\delta^2 A[\mathbf{q}(\tau)]$.

To illustrate this concept, we consider the simplified situation where the total spin of the tunneling domain is conserved and all spins in the domain move identically through the instanton trajectory, $\mathbf{n}_j(\tau) \equiv \mathbf{n}(\tau)$ for all $j \in [1, D]$, where D is the number of cotunelling spins. Then, the mean-field potential for the instanton can be rescaled as

$$V(\underline{\mathbf{q}}(\tau)) = D\upsilon(\mathbf{n}(\tau)). \quad (22)$$

In general, the environment can lead to thermally assisted effects in multispin tunneling. Their role in the analysis of spin instantons is discussed in Ref. [62]. For this effect to become important the energy scale $k_B T$ corresponding to the environmental temperature needs to be comparable with the energy separation ΔE_0 between the levels of the initial potential well near its minimum. However, in many situations typical for QA the tunneling is of a Landau-Zener type where only the two lowest levels are involved in the tunneling transition while the rest of the levels are separated by the large energy gap that is much bigger than the temperature. For example, in the case of the weak-strong cluster network discussed in Sec. II, the typical gap separating the tunneling doublet from the rest of the levels is $\Delta E_0/h \gtrsim 3$ GHz. This is about 10 times greater then the thermal energy scale at the device temperature of 15 mK.

In the limit of low temperatures $\beta \to \infty$, the instanton action takes the form $S[\mathbf{n}(\tau)] = Da[\mathbf{n}(\tau)]$, with

$$a[\mathbf{n}(\tau)] = \frac{\hbar}{2}\omega[\mathbf{n}(\tau)] + \int_0^\infty d\tau\upsilon[\mathbf{n}(\tau)], \quad (23)$$

where ω describes a "Berry phase" type contribution. The exponential dependence of the tunneling rate,

$$W_{\mathrm{QA}} \sim e^{-Da_{\min}/\hbar}, \qquad a_{\min} = \min_{\mathbf{n}(\tau)} a[\mathbf{n}(\tau)], \quad (24)$$

is given by the minimum of the rescaled action. We consider the relevant case where the tunneling of the spin domain is enabled by the transverse field and the Hamiltonian is of the type in Eq. (2),

$$H = -A\sum_{j=1}^{N}\sigma_j^x - BH_P^{\rm cl}(\sigma_1^z,\ldots,\sigma_D^z), \tag{25}$$

where $H_P^{\rm cl}(s_1,\ldots,s_D)$ is a classical cost function of binary variables $s_k = \pm 1$. Assuming that $H_P^{\rm cl} = H_P^{\rm cl}(\sum_{j=1}^{D}\sigma_j^z)$, in the zero-temperature limit the system is in a state of maximum total spin and all spins will tunnel together. Thus,

$$v(\mathbf{n}(\tau)) = -A\sin\theta(\tau)\cosh\varphi(\tau) - Bg[\cos\theta(\tau)], \tag{26}$$

where the function $g(x) = D^{-1}H_P^{\rm cl}(Dx)$ is the rescaled mean-field potential energy of the spin system. Solving the set of Eqs. (18) under the ansatz Eq. (22), one gets

$$\frac{a_{\rm min}}{\hbar} = \int_{\theta_0}^{\theta_1}\operatorname{arcsinh}\left(\frac{v(\theta)}{A\sin\theta}\right)\sin\theta d\theta, \tag{27}$$

where $v(\theta) = \sqrt{B^2[g(\cos\theta) - g(\cos\theta_0)]^2 - A^2\sin^2\theta}$ is a linear velocity along the instanton path and the angles θ_i correspond to the minima of the potential $v(\mathbf{n})$ (the values of $\sinh\varphi_j = 0$ at the minima). This expression corresponds to the familiar zero-temperature result that can be obtained by other methods (cf. Supplemental Material in Ref. [39] and also Ref. [65]). We note that, in the general case, the action A in Eq. (21) also grows linearly with the number of cotunneling spins D as in the simplified case of Eq. (24) because the individual spin contributions to the action are highly correlated at the instanton trajectory. The general condition for the validity of the instanton calculus is a large number of cotunnelling spins, $D \gg 1$.

Even in the limit of low temperatures, dissipative effects of the environment can substantially affect both the prefactor and the exponent in the expression for the tunneling rate Eq. (21). Instantonic calculus allows us to take these effects into account consistently in the calculation of the transition rates (as has been done, e.g., in the case of tunneling with dissipation in SQUIDs [66]). However, in the cases relevant to the present study and discussed in Sec. II, the dissipative corrections in the exponent of the transition rate will be relatively small. This happens because the typical separation between the intrawell energy levels ($\Delta E_0/h \gtrsim 3$ GHz, see above) is much greater than the characteristic level broadening due to low frequency noise (~0.5 GHz, see Appendix A [43]) and than the decay rate due to Ohmic noise ($\simeq\eta\Delta E_0 \sim 0.3$ GHz with $\eta \simeq 0.12$, Appendix A [43]).

B. Tunneling simulation in the QMC algorithm

In the path-integral QMC algorithm, one introduces an extra dimension associated with the imaginary time axis in order to simulate the multispin tunneling phenomenon on a classical computer, as seen above. This is done by using a Suzuki-Trotter decomposition and representing the partition function of the system Z in terms of the path integral over the spin trajectories $\underline{\sigma}(\tau) = \{\sigma_j(\tau)\}_{j=1}^{N}$; see Eq. (9). These trajectories are periodic along the imaginary time axis $\sigma_j(0) = \sigma_j(\beta)$. For each spin j, the set of values $\sigma_j(\tau) = \pm 1$ form a path component referred to as a worldline. The time step along the worldline is $\Delta\tau = \beta/M$, where M is the number of Trotter slices (i.e., the number of spin replicas in the worldline). Sampling the system states along this extra dimension introduces an additional overhead in classical computation that does not exist for the corresponding quantum dynamics.

The runtime $T_{\rm QMC}$ of the QMC algorithm can be thought of as a product of three factors:

$$T_{\rm QMC} = Nn_{\rm sweeps}T_{\rm worldline}, \tag{28}$$

where N is the problem size that is equal to the number of worldlines. The number of sweeps $n_{\rm sweeps}$, in general, depends exponentially on the typical size D of the cotunneling domain, $n_{\rm sweeps} \propto e^{\alpha D}$, where $\alpha = \alpha(\beta)$ also depends on the inverse temperature β. In cases where $D = \mathcal{O}(N)$, the growth of this factor with N reflects a major computational bottleneck of QMC and QA. As was shown recently by some of us [39], the exponent in QMC calculations and QA is the same for a broad class of problems.

According to the findings we report in this paper, the prefactor in $n_{\rm sweeps}$ along with the factors N and $T_{\rm worldline}$ in $T_{\rm QMC}$ are significantly different for QMC calculations and QA. The value of $T_{\rm worldline}$ we find in our simulations is given in Eq. (B1). We also find that

$$n_{\rm sweeps} \gg 1{\rm ns}/T_{\rm QA}, \tag{29}$$

where $T_{\rm QA}$ is the duration of QA and we use a normalization factor of 1 ns, corresponding to the typical time scale of superconducting QA devices. We expect that the above relation will remain true even if the scaling of both quantities with $Da_{\rm min}/\hbar$ is the same.

C. Comparison of QA and QMC results for the "weak-strong cluster pair" problem

We use the modeling considerations described above to theoretically compare QA and QMC results in the system corresponding to the weak-strong cluster pair problem [see discussion in Sec. II, Fig. 2, and Eqs. (6)–(8)]. Tunneling in this system corresponds to an avoided crossing between the two lowest energy levels of the Hamiltonian, shown in Fig. 5. All other levels lie high above the first two and are never excited. During tunneling, the total spin of the left cluster reverses orientation. The number of cotunneling spins is $D = 8$ in this case, while the total number of spins is $N = 16$.

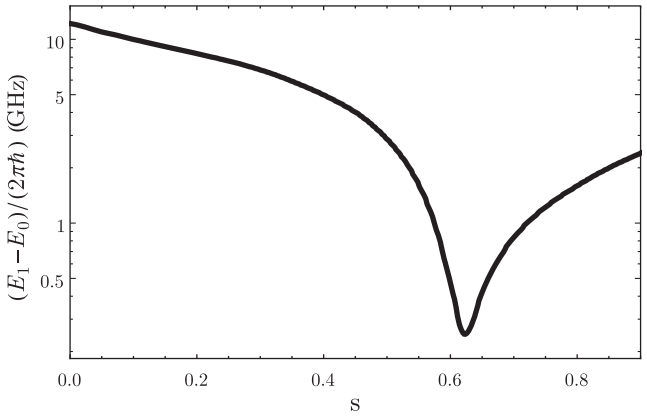

FIG. 5. Gap of the quantum Hamiltonian for $h_1 = 0.44$ as a function of the annealing parameter. The solid line is the energy difference between the ground state and the first excited state. The avoided crossing at $t/T_{\mathrm{QA}} = 0.62$ corresponds to a minimum gap of 248 MHz. The next excited state is separated by a gap in excess of 2 GHz.

Proper analysis of the tunneling probabilities and related QA success rates should also account for coupling to the environment. We study the success probabilities in QA for the 2-cluster problem using the theoretical model developed in Ref. [35]. The results are summarized in Fig. 6. Decreasing the temperature of QA compared to the temperature of the D-Wave device suppresses steeply the transition rates between the states because of the increase in the reconfiguration energy [35,67,68] (see Fig. 7). This, together with the suppression from the Boltzman factor in the transition rates, leads to an increase of the final success probability p_0 to find the ground state. Once the temperature reaches 5 mK, the success probability stays above 90%, even at QA schedules that are faster than $T_{\mathrm{QA}} =$ 300 ns (cf. Fig. 6).

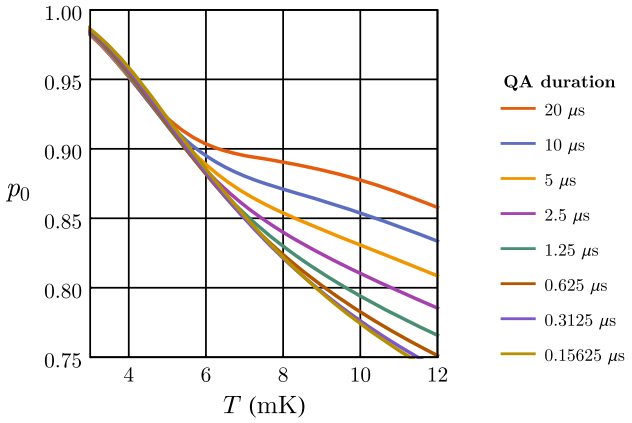

FIG. 6. Success probability of QA p_0 versus T, given by theoretical modeling [Eq. (A2)]. Different colors correspond to different durations of the QA process T_{QA}. Plots correspond to $h_1 = 0.44$.

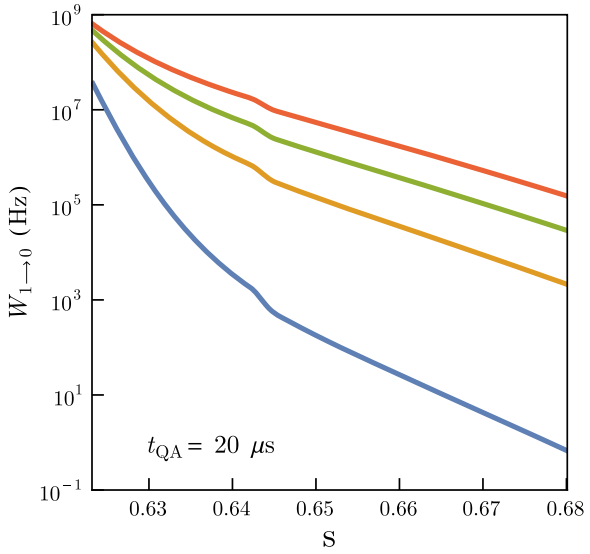

FIG. 7. Logarithmic plots of the theoretically modeled transition rate $W_{10}(s)$ versus s after the avoided crossing for different temperatures. Red, green, mustard, and blue correspond to $T = 12$, 8, 6, and 4 mK, respectively. All plots correspond to $h_1 = 0.44$ and $T_{\mathrm{QA}} = 20$ μs.

On the other hand, adiabatic transitions near the avoided crossing are suppressed even at $T_{\mathrm{QA}} \simeq 71$ ns, as can be seen from solutions to the time-dependent Schrödinger equation, shown in Fig. 8.

In the previous studies involving D-Wave devices (see Ref. [35] for references), it was inferred from the data that the low-frequency noise components of the spectral density, providing a leading contribution to the qubit linewidth W [Eqs. (A1) and (A2)], have effective frequency cutoffs

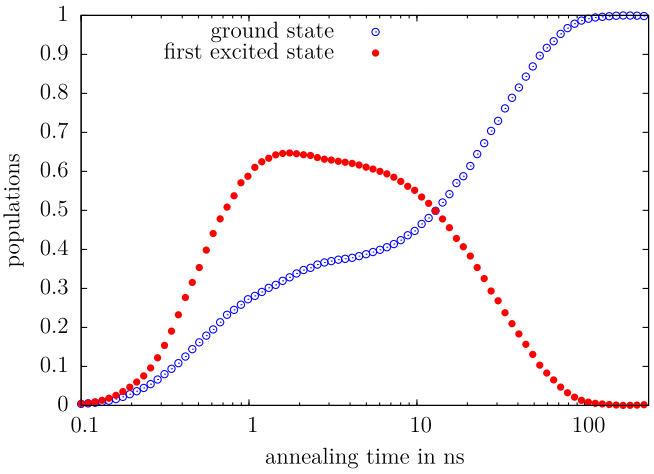

FIG. 8. Plots show the results of the solution of the time-dependent Schrödinger equation for the closed system QA. Probabilities of occupation of ground state and first excited state as a function of QA time are plotted with blue and red points, respectively. The QA time to reach probability of success 0.95 equals 70.9 ns. All plots correspond to $h_1 = 0.44$.

much below 314 MHz (15 mK). In the current QA schedule of 20 μs, the system spends only a small fraction of this time in the vicinity of the avoided crossing where thermal excitations from the ground state are possible. For a QA schedule duration of ∼100 ns, we expect that the effective noise strength will be weaker than at the current schedule. This would lead to the suppression of the thermal excitations from the ground state.

To compare simulations of QA at a temperature of 5 mK with the QMC performance, we choose a duration of QA such that the probability to reach the ground state at the end of QA equals 0.95. As we discuss above, this can be achieved at

$$T_{\mathrm{QA}} = 71 \text{ ns}. \tag{30}$$

In setting up the QMC simulations our objective is to select the two parameters, number of sweeps per qubit n_{sweeps} and β, to minimize T_{QMC} for a given probability of success p_0 to find the system at the end of the QA in the ground state where all spins point down. Essentially, we need to minimize the product of $\beta n_{\mathrm{sweeps}}$ keeping p_0 fixed.

In Fig. 9, we plot the success probability of QMC $p_0 = p_0(\beta, n_{\mathrm{sweeps}})$ as a function of β (inverse temperature) for different numbers of sweeps. We see that increasing β increases p_0; however, the success probability saturates at some value

$$p_0(\beta, n_{\mathrm{sweeps}}) \le p_0^{\mathrm{sat}}(n_{\mathrm{sweeps}}), \tag{31}$$

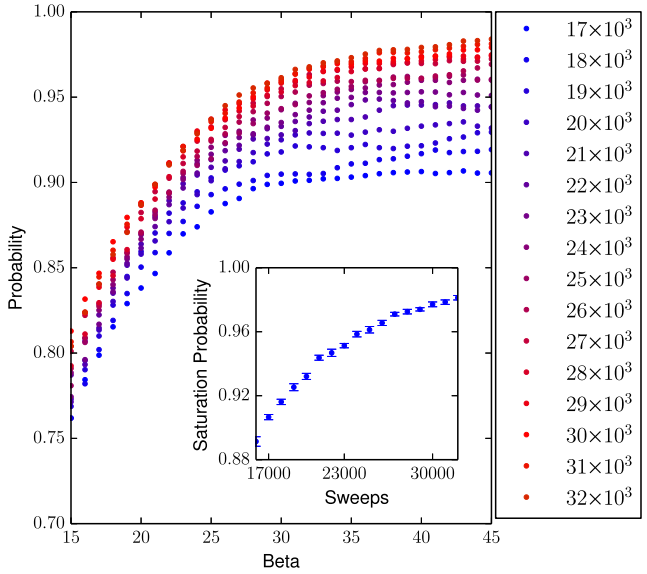

FIG. 9. Probability of success versus β for QMC with the D-Wave 2X schedule. Different colors correspond to different number of sweeps (see legend). The embedded plot shows the probability of success at saturation for different number of sweeps. We use periodic boundary conditions, which performed better than open boundary conditions in this case.

which itself depends on the number of sweeps. The saturation probability $p_0^{\mathrm{sat}}(n_{\mathrm{sweeps}})$ is plotted in the inset of Fig. 9. By fixing the success probability $p_0 = 0.95$, we select the optimal number of sweeps. Then, by looking on the main plot we determine the value of β_{sat} where saturation occurs. The optimal values are

$$n_{\mathrm{sweeps}} = 23000, \qquad \beta = 32.5. \tag{32}$$

The total time to update one worldline with the D-Wave 2X schedule is [see Eq. (11)]

$$T_{\mathrm{worldline}} = 28.3\ \mu\text{s}, \tag{33}$$

and the total runtime of the QMC calculations per qubit, according to Eq. (28), is

$$\frac{T_{\mathrm{QMC}}}{N} = 0.65 \text{ s}. \tag{34}$$

By comparing this with Eq. (30), we estimate that

$$\frac{T_{\mathrm{QMC}}/N}{T_{\mathrm{QA}}} \sim 10^7 \quad (T_{\mathrm{QA}} = 71 \text{ ns}, T = 5 \text{ mK}). \tag{35}$$

This ratio will need to be multiplied by the number of qubits to obtain the overall speed-up factor (e.g., $\sim 10^{10}$ for 1000 qubits).

Implementing fast QA schedules or operating flux qubits at 5 mK will require improvements in the control electronics and other elements of the design. Furthermore, readout will need to be made much faster than in the current D-Wave devices. However, the estimates we present above serve to emphasize the significant promise of QA, as compared to QMC results when the system adiabatic evolution "under the gap" becomes coherent and thermal excitations are suppressed.

IV. NUMERICAL STUDIES OF QUANTUM ANNEALING FOR GENERIC PROBLEMS WITH RUGGED ENERGY LANDSCAPES

Runtime advantages for the quantum processor we describe above are only valuable if they extend to problems of practical interest. While rather obvious, it may we worth delineating criteria for problems that are suitable for treatment with a quantum annealer:

(1) Solutions to the problem are valuable or interesting.
(2) The problem is representable on hardware that can be built in the near future.
(3) Quantum annealing offers a runtime advantage.

A. Number partitioning

A valuable and interesting practical problem is the number partitioning problem (NPP). The NPP is defined as follows: Given a set of N positive numbers $(a_1, \ldots, a_N)$

find a partition $\mathcal{P}$ of this set into two groups that minimizes the partition residue $E = |\sum_{j\in\mathcal{P}} a_j - \sum_{j\notin\mathcal{P}} a_j|$. A partition $\mathcal{P}$ can be encoded by Ising spin variables $s_j = \pm 1$: $s_j = +1$ if $j \in \mathcal{P}$ and $s_j = -1$ otherwise. Thus, the NPP cost function is

$$E(\mathbf{s}) = |\Omega_{\mathbf{s}}|, \qquad \Omega_{\mathbf{s}} = \sum_{j=1}^{N} a_j s_j, \tag{36}$$

where $\mathbf{s} = (s_1, \ldots, s_N)$ is a spin configuration and $\Omega_{\mathbf{s}}$ is a *signed* partition residue. Number partitioning is also one of Garey and Johnson's six basic NP-hard problems that lie at the heart of the theory of NP completeness [69]. In studies of the average-case computational complexity of NPP, one usually assumes that $\{a_1, \ldots, a_N\}$ are independent, uniformly distributed random numbers in the interval [0, 1). The average-case complexity of NPP is exponential in N when the number of bits b used to represent the numbers a_j satisfies the condition $b/N \geq \kappa_c \simeq 1 - (1/2N)\log_2 N$ [70]. Our focus is on hard instances of NPP, and we will be studying the random NPP ensemble with $b = N$. NPP has many practical applications including multiprocessor scheduling and the minimization of VLSI circuit size and delay, public key cryptography, and others (see references in Ref. [71]).

Unfortunately, practically interesting NPP instances cannot be represented on the current D-Wave 2X machine due to two obstacles: (i) the Ising representation of NPP is fully connected and (ii) the available bit precision of couplers on D-Wave 2X is insufficient for representing any nontrivial NPP instances. While in principle there are techniques, such as graph embedding [20], to address these problems, performance is likely to be significantly impacted by their usage. Nevertheless, NPP is attractive for numerical studies because, for $b/N > \kappa_c$, the typical runtime of all known algorithms for NPP scales exponentially with large coefficients in the exponent and often the asymptotic behavior can already be seen at sizes as low as 20 variables. For our purposes, NPP is a useful problem to study in the context of quantum annealing because it possesses extremely rugged energy landscapes where a single bit flip can result in dramatic energy changes. Its low-energy band resembles that of the random energy model as there is almost no correlation between the state and its energy [72]. The 2^N signed partition residues Ω can be thought of as drawn from the Gaussian distribution

$$p(\Omega) = \frac{2}{\sqrt{2\pi N\langle a^2\rangle}} \exp\left(-\frac{\Omega^2}{2N\langle a^2\rangle}\right), \tag{37}$$

where $\langle a^2\rangle = (1/N)\sum_{j=1}^{N} a_j^2$. The distribution of the cost function values $E = |\Omega|$ is given by $2p(E)$. By picking a bit string at random, one gets an average value of the cost function $\langle E\rangle = \sqrt{2\langle a^2\rangle N/\pi}$. The *minimum* value of this cost function is exponentially small, with median value $E_{\min} \sim \langle E\rangle 2^{-N}$ [71].

An obvious heuristic algorithm for NPP starts by placing the largest number in one of the two subsets. The next largest number is then placed in the set whose elements sum to the smallest value and this continues until all numbers are assigned. The idea behind this greedy heuristic is to keep the discrepancy small with every decision. This gives the scaling of the resulting partition residue as $\mathcal{O}(1/N)$. The differencing method of Karmarker and Karp [71], also called the KK heuristic, is a polynomial time approximation algorithm. The key idea of this algorithm is to reduce the size of the numbers by replacing the two largest numbers by the absolute value of their difference. It has been proven [73] that the differencing method gives a minimum residue $E_{\min}^{\mathrm{KK}}$ such that

$$E_{\min}^{\mathrm{KK}} \sim N^{-\alpha \log N}, \qquad \alpha = 0.72. \tag{38}$$

The time complexity of both greedy and KK heuristics is $N \log N$ [71]. The residual energies reached by both methods are much smaller than the average partition residue $\langle E\rangle$, but far greater than the minimum residue $E_{\min}$. The absence of efficient heuristics for these hard cases is a particular feature of NPP. It is attributed to the extremely rugged energy landscape in the low part of the energy spectrum [71]. The statistics of the NPP energy landscape was studied analytically in Ref. [74] and numerically in Ref. [75].

This type of landscape leads to the exponential complexity of QA for NPP that was obtained in Ref. [74] via direct solution of the time-dependent Schrödinger equation. We observe that the particularly challenging instances of NPP violate the second condition of our "suitability" criteria as the numbers $a_j \in (0, \ldots, 2^N - 1)$ should be drawn from a set whose cardinality grows exponentially with N. This translates to a requirement that the bit precision for the coupling coefficients J_{jk} grows as $2N$ if one were to express NPP as a quadratic binary optimization problem with objective function $\sum_{jk=1}^{N} a_j a_k s_j s_k$ corresponding to the form Eq. (1) with $K = 2$. However, for numerical studies this is not a concern.

Table I shows the runtime behavior of annealing as well as some more efficient algorithms that achieve asymptotically better performance by exploiting the existing problem structure. The relative performance of annealing algorithms applied to NPP is similar to their relative performance on the weak-strong cluster networks problem: QA, simulated using the Schrödinger equation [74], scales better than SA. This is the case because both problems are characterized by rugged energy landscapes for which tunneling transitions are a more useful way to reach low-energy states than thermal transitions.

To achieve such scaling behavior it will be necessary for the size of the domains of cotunneling qubits to grow with the problem size. However it is interesting to explore the problem from a different perspective and ask the question,

TABLE I. Runtime scaling exponent for different methods to solve the number partitioning problem. The scaling of simulated annealing is very poor and is not shown here. This is because for simulated annealing to work at all it has to be run at very high temperatures to overcome the enormous energy barriers present in this problem. However, at these high temperatures, SA behaves almost like random sampling and, hence, its scaling is almost that of exhaustive search. The value $\alpha = 0.8$ for the solution of the time-dependent Schrödinger equation was initially obtained in Ref. [74] and is significantly better than the nearly exhaustive-search behavior of SA. This mirrors the situation we encounter for the weak-strong cluster networks. We also give references for the state-of-the-art classical and quantum algorithms.

Method	α
Quantum adiabatic algorithm (time-dependent Schrödinger equation [74])	0.80
Moduli + representations [76]	0.337
Moduli + representations + overlap [77]	0.291
Quantum walk + moduli + representations [78]	0.241

How much can the residual energy be lowered in QA until the system reaches such a state where lowering the energy further would require cotunneling of spin domains with sizes greater than κ?

To answer this question, one can use an algorithm in which one starts at a random initial state and, at each step, (i) looks at all groups of bits of size κ and (ii) flips the bits of the group that results in the largest reduction of residual energy and then one iterates. We call this procedure "algorithmic tunneling" (AT). In other communities this would be referred to as κ-opt or large neighborhood search. We emphasize that AT does not provide any information about the actual system dynamics during QA, nor the runtime of QA. We investigate AT as an upper bound on the typical performance of QA. AT does not consider the entropic component of tunneling events arising from the statistics of different mechanisms for arriving in the same minima. Likewise, AT does not consider how the height of energy barriers affects tunneling events. However, AT allows us to develop our intuition about the value of an optimal tunneling procedure that always finds the lower-energy solution within a finite Hamming distance.

To investigate the lowest residual energies that AT can reach, we focus on the conditional distribution of the signed partition residues Ω' [Eq. (36)] over all possible spin configurations $\{\mathbf{s}'\}$ generated from a given (ancestor) configuration $\mathbf{s}$ by simultaneously flipping a fixed number of spins κ. This conditional distribution was studied in Ref. [74] and has the form

$$P_\kappa(\Omega|\Omega') = \frac{1}{\binom{N}{\kappa} P(\Omega)} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{dx dx'}{8\pi^2} \zeta(x+x') \times e^{i(\Omega-\Omega')x/2} e^{i(\Omega+\Omega')x'/2} \times \sum_{\mathbf{J}} \prod_{j\in\mathbf{J}} \cos(a_j x) \prod_{j\notin\mathbf{J}} \cos(a_j x'), \tag{39}$$

where $P(\Omega)$ is given in Eq. (37) and $\zeta(s) = \sin(\Delta\Omega s/4)/(\Delta\Omega s/4)$. In the distribution $P_\kappa(\Omega,\Omega')$, we average over the residue of the ancestor configuration within the interval $\Omega_{\mathbf{s}} \in [\Omega - \Delta\Omega/2, \Omega + \Delta\Omega/2]$, where $\langle E\rangle/\binom{n}{\kappa} \ll \Delta\Omega \ll \Omega$.

Near its maximum, the distribution $P_\kappa(\Omega|\Omega')$ has the form (cf. [74])

$$P_\kappa(\Omega'|\Omega) = \frac{1}{\sqrt{2\pi N\sigma^2(q)}} \exp\left[-\frac{(\Omega' - q\Omega)^2}{2N\sigma^2(q)}\right], \tag{40}$$

where

$$\sigma(q) = \langle a^2\rangle (1-q^2)^{1/2}, \qquad q = 1 - \frac{2\kappa}{N}. \tag{41}$$

For small $\kappa \ll N$, the width of the distribution is approximately $\sqrt{\kappa\langle a^2\rangle}$. After a single step of the AT algorithm, the average partition residue is reduced by a factor of $1 - 2\kappa/N$. Therefore, once the number of steps of AT far exceeds N/κ, the algorithm reaches the residues $|\Omega| \ll \sqrt{\kappa\langle a^2\rangle}$. For those residues the distribution becomes

$$P_\kappa(\Omega'|\Omega) \simeq P_\kappa(0|0) = \frac{1}{\sqrt{8\pi\kappa\langle a^2\rangle}}. \tag{42}$$

We now apply the results obtained in Ref. [74] to compare the minimum cost values reached in AT and the KK heuristic. In total, we have $\binom{N}{\kappa}$ samples of the distribution [Eq. (42)] (spin configurations located on a Hamming distance κ from the ancestor configuration). Therefore, the minimum value is given by extreme order statistics [79]. That is, within the set of bit configurations $\{\mathbf{s}'\}$ generated from a given (ancestor) configuration $\mathbf{s}$ by simultaneous flipping of a fixed number of spins κ, the minimum partition energy E is a random variable drawn from the exponential distribution

$$p_0(E) = \frac{1}{E_\kappa} e^{-E/E_\kappa}, \qquad E_\kappa = \left[\binom{N}{\kappa} P_\kappa(0|0)\right]^{-1}. \tag{43}$$

Therefore, average (and median) values of the minimum partition residue achieved in the AT algorithm are $E^{\mathrm{AT}}_{\mathrm{min}}(\kappa) = E_\kappa$. For $1 \ll \kappa \ll N$, we obtain

$$E^{\mathrm{AT}}_{\mathrm{min}}(\kappa) = 4\pi\kappa\langle a^2\rangle \left(\frac{\kappa}{N}\right)^\kappa e^{-\kappa}. \tag{44}$$

It is instructive to compare the minimum cost values reached in AT and KK heuristics. Using Eqs. (38) and (44), we get

$$\frac{E^{\mathrm{AT}}_{\mathrm{min}}}{E^{\mathrm{KK}}_{\mathrm{min}}} \sim N^{\alpha \log N - \kappa} \kappa^\kappa e^{-\kappa}, \qquad \alpha = 0.72. \tag{45}$$

One can see from here that tunneling over barriers of length $\kappa > \alpha \log N$ allows AT to reach cost values lower than that of the KK heuristics as N increases.

Consider AT with values of κ that do not scale with N. We observe that in the asymptotic limit,

$$N \gg N_\kappa = \frac{e}{\kappa}\exp(\kappa/\alpha), \tag{46}$$

the KK heuristic produces smaller residues than AT. If we consider tunneling with $\kappa = 8$, corresponding to the case of the weak-strong clusters we study in this paper, then $N_\kappa \simeq 22735$. We observe that for the high-precision ($B \geq N$) instances, NPP becomes intractable already for $N > 100$. Thus, these novel calculations show that for a broad range of problem sizes AT reaches cost values much smaller than conventional heuristics. This finding supports recent claims [51,80] that QA might be advantageous in achieving superior approximate solutions rather than global minima.

Motivated by the above observation, we conclude that asymptotic scaling behavior is not essential for this analysis. For example, one can choose AT with the barrier sizes $\kappa = \alpha_0 \log N$, where $\alpha_0 - \alpha \gg 1/\log N$. In this case, the ratio $E^{\mathrm{AT}}_{\mathrm{min}}/E^{\mathrm{KK}}_{\mathrm{min}}$ approaches 0 in the asymptotic limit $N \to \infty$. However, the length of the barriers remains relatively small in a very broad range of N. For example, consider $\alpha_0 = 1.16$. Then, for $N = 1000$, the barrier length is $\kappa = 10$ while $E^{\mathrm{AT}}_{\mathrm{min}}/E^{\mathrm{KK}}_{\mathrm{min}} \sim 10^{-9}$.

In summary, a greedy search procedure with flipping at most κ bits at a time (referred to above as algorithmic tunneling) allows us to find cost values in NPP that are much below those given by the KK heuristic for $\kappa \sim 10$ at all realistic values of N. However, while the KK heuristic terminates in just $\mathcal{O}(N \log N)$ time [71], the time complexity of AT is exponential in κ. Nevertheless, it would be interesting to compare the minimum residue obtained by AT with the minimum residues obtained by the KK heuristic and the algorithms in Table I when all of them are constrained to terminate in polynomial time.

B. Designing future annealers of practical relevance

Our current best candidate for a problem class that fulfills all three criteria consists of Kth-order binary optimization problems with $K > 2$. Kth-order binary optimization is NP hard and occurs naturally in many engineering disciplines and many computational tasks. In unpublished work under way, we seek to establish that for many K-local problems, QA indeed offers a runtime advantage over SA. Currently we are focusing on $K \in \{4, 5, 6\}$. As energy landscapes get more rugged with higher K, our conjecture is that we will see larger subsets of instances for which QA runs faster as K increases. However, representing K-body terms in hardware becomes more challenging as K grows.

Should numerical studies confirm that QA offers a substantial runtime advantage, there is still one more hurdle to overcome. We need to ensure that K-local problems can be economically represented in hardware. We would like to be able to tell a user "If you have a binary optimization problem with N variables and L terms, and the many-body order of the highest term is K, then you can send this problem to the quantum annealing coprocessor." However, annealers built to date support only pairwise qubit couplings, i.e., $K = 2$. Two routes have been proposed to increase the locality.

One route is to build physical K-body couplers. However, it may prove difficult to lay out K-local couplers on a two-dimensional chip or even in layered architectures. Of course, the general case in which one aims to implement all possible $\sum_{k=1}^{K} \binom{N}{k}$ couplings will be infeasible. While many applications will necessitate only $\mathcal{L} = O(N)$ coupling terms, this could still prove challenging. Furthermore, economically embedding problems in a fixed graph with only a limited number of specific K-local terms may prove difficult.

Another possibility is that we could use logical embeddings that map K-local problems to 2-local problems. A new proposal on how to accomplish such embeddings has been put forth [81], which has reinvigorated interest in this direction. Our main worry regarding any reduction to quadratic problems is that this will involve ancillary qubits. As we argue, it is crucial that the problem features tall and narrow barriers for tunneling transitions to contribute, and the introduction of additional variables may cause these barriers to become wider. This makes purely thermal annealing more competitive and may negate gains seen in the numerics prior to embedding.

V. SUMMARY

It is often quipped that simulated annealing is only for the "ignorant or desperate." Yet, in our experience we find that lean stochastic local search techniques such as SA are often very competitive and they continue to be one of the most widely used optimization schemes. This is because for sufficiently complex optimization tasks with little structure to exploit (such as instances of Kth-order binary optimization) it often takes considerable expert effort to devise a faster method. Therefore, we regard SA as the generic classical competition that quantum annealing needs to beat.

Here, we show that, for carefully crafted proof-of-principle problems with rugged energy landscapes that are dominated by large and tall barriers, QA can have a significant runtime advantage over SA. We find that for problem sizes involving nearly 1000 binary variables, quantum annealing is more than 10^8 times faster than SA running on a single core. We also compare the hardware to the QMC method. While the scaling of runtimes with size between these two methods is comparable, they are again separated by a large factor sometimes as high as 10^8.

For higher-order optimization problems, rugged energy landscapes will become typical. As we see in our experiments with the D-Wave 2X, problems with such landscapes stand to benefit from quantum optimization because quantum tunneling makes it easier to traverse tall and narrow energy barriers. Therefore, we expect that quantum annealing might also deliver runtime advantages for problems of practical interest such as Kth-order binary optimization with larger K.

More work is needed to turn quantum enhanced optimization into a practical technology. The design of next-generation annealers must facilitate the embedding of problems of practical relevance. For instance, we would like to increase the density and control precision of the connections between the qubits as well as their coherence. Another enhancement we wish to engineer is to support the representation not only of quadratic optimization but of higher-order optimization as well. This necessitates that not only pairs of qubits can interact directly but also larger sets of qubits. Such improvements will also make it easier for end users to input hard optimization problems.

The work we present here focuses on the computational resource that is experimentally most accessible for quantum annealers: finite-range tunneling. However, this analysis is far from complete. A coherent annealer could accelerate the transition through saddle points, an issue slowing down the training of deep neural networks, for reasons similar to those that make a quantum walk spread faster than a classical random walker [82–84]. It could also dramatically accelerate sampling from a probability distribution via the mechanism of many-body delocalization [85]. The computational value of such physics still needs to be properly understood and modeled.

ACKNOWLEDGMENTS

We would like to thank Matthias Troyer for discussions and Edward Farhi and Masoud Mohseni for helping to review the manuscript. We would also like to thank Daniel Lidar, Damian Steiger, Alex Selby, and Dvir Kafri for comments.

APPENDIX A: QUANTUM ANNEALING RESULTS FOR WEAK-STRONG CLUSTER PROBLEM WITH 16 QUBITS

We developed a detailed modeling of the quantum annealing process and incoherent multiqubit cotunneling for the weak-strong cluster problem in Ref. [35]. Using a noise model with experimentally measured parameters for the D-Wave 2X processor, we numerically verified that the spins arrive at the energetically more favorable configuration via multiqubit tunneling. In the following, we refer to the modeling in Ref. [35] for the details.

In the present study, we apply this detailed model for the new schedule functions $A(s)$, $B(s)$ (see Fig. 10) and for the new values of the noise parameters, linewidth W, Ohmic coefficient η, and temperature of the device T for the D-Wave 2X processor. The noise parameters are measured near the end of the quantum annealing schedule, $s = 1$. The values of the noise parameters at a point during the annealing can be related to the measured ones (see Appendix A 5 in Ref. [35]):

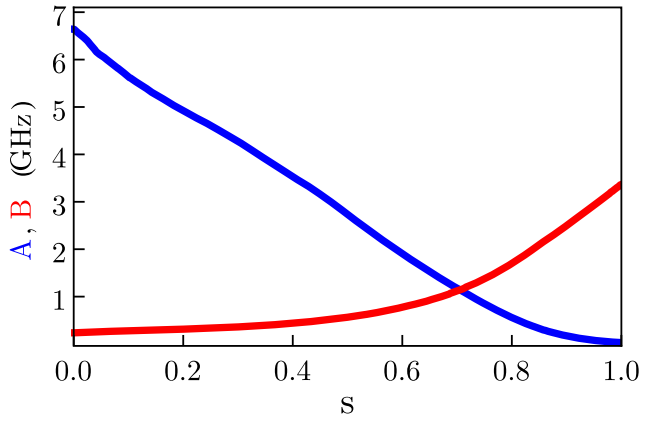

FIG. 10. Schedule functions for D-Wave 2X chip. The annealing parameter is $s = t/T_{\mathrm{QA}}$ for time t and total annealing time T_{QA}.

$$[W(s)/W_{\mathrm{MRT}}]^2 = \eta/\eta_{\mathrm{mRT}} = B(s)/B(1),$$
$$W_{\mathrm{MRT}} \simeq 661\ \mathrm{MHz}, \qquad \eta_{\mathrm{MRT}} \simeq 0.12, \qquad T \simeq 12\ \mathrm{mK}. \tag{A1}$$

The population $p_0(t)$ of the ground state during the QA process obeys the equations

$$\frac{dp_0}{dt} = -[W_{01}(s) + W_{10}(s)]p_0(s) + W_{10}(s),$$
$$\frac{W_{01}(s)}{W_{10}(s)} = e^{-\Delta_{10}(s)/k_B T}, \qquad \Delta_{10}(s) = E_1(s) - E_0(s), \tag{A2}$$

where $W_{jk}(s)$ is a transition rate from the state j to the state k whose explicit form is given in Ref. [35]. The transition rate $W_{10}(s)$ decays very fast after the avoided crossing (see Fig. 7) because the weak cluster (left cluster in Fig. 2) becomes progressively more polarized along the z axis and the effective size of the tunneling domain $D = D(s)$ grows. This gives rise to the multiqubit "freezing phenomenon," where the system population gets partially trapped in the excited state after a certain value of s in later stages of QA [35]. Figure 11 shows the ground-state population given by the solution of Eq. (A2). The success probability to be at the ground state at the end of the annealing is $p_0 = 0.85$, which is close to the experimentally observed mean value of 0.9. It is seen that the equilibrium population of the ground state

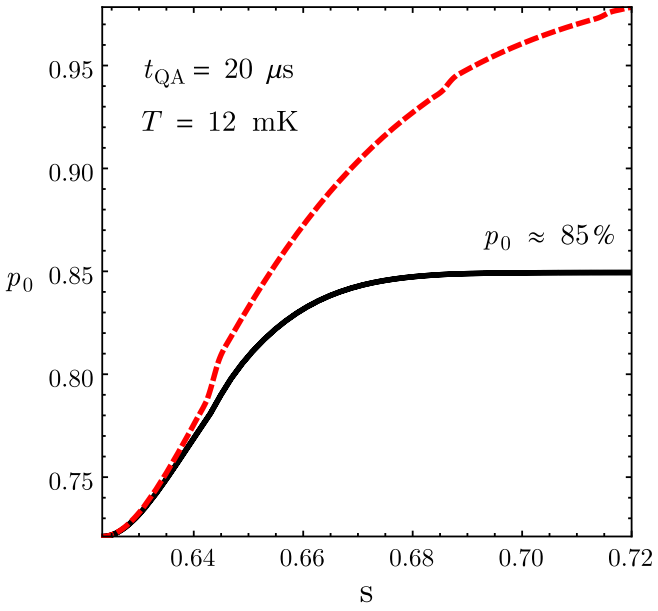

FIG. 11. Simulation of the probability of occupation of the ground state during the theoretically modeled QA process is shown with a solid black line. The red dashed line gives the equilibrium population of the ground state. Results are given at $T = 12$ mK and $h_1 = 0.44$. The dashed red line gives the equilibrium population of the ground state. The simulated quantum annealing time is $t_{QA} = 20$ μs.

exceeds the actual population for $s \gtrsim 0.64$, corresponding to the onset of freezing of the transition rates.

We note that if the effective temperature of the qubit environment can be lowered, then the QA success rates can be made sufficiently high using much faster annealing durations, as shown in Fig. 6.

APPENDIX B: D-WAVE VERSUS QUANTUM MONTE CARLO METHOD WITH LINEAR SCHEDULE

There is ongoing work directed to optimize the QMC parameters further. In preliminary results, we compare the QMC method with a linear schedule against D-Wave [86]. The transverse field in this case is lower, resulting in a faster time $T_{\text{worldline}}$ to update a worldline ($T_{\text{worldline}}$ scales linearly with the transverse field). We measure this time to be

$$T_{\text{worldline}} = \beta(115\ \text{ns}). \tag{B1}$$

This time is consistent with the one reported in Ref. [12].

We also take a different approach when optimizing β and the number of sweeps per run to minimize the total computational effort. In the case reported in Sec. II B, we optimize the number of sweeps for each quantile at fixed $\beta = 10$. In the case of a linear schedule, we use our knowledge of the structure of the weak-strong cluster networks problem to optimize β and the number of sweeps n_{sweeps} concurrently. We first measure the probability of success $p(n_{\text{sweeps}}, \beta)$ for a single weak-strong cluster pair. Then we estimate the performance for the cluster network problems taking into account the number of cluster pairs c for each size. The estimate is

$$\text{total time} \propto n_{\text{sweeps}}\beta\left\lceil \frac{\log(1\text{–}0.99)}{\log[1 - p(n_{\text{sweeps}}, \beta)^c]} \right\rceil.$$

Here, we also run with OBC in imaginary time. We estimate an optimal $\beta = 130$ for all sizes. We then optimize the number of sweeps for each quantile and size running the actual benchmark. Finally, we modifiy the path-integral QMC code to search for the minimum energy configuration along all replicas at the end of the annealing.

The results, following the same methodology as in Sec. II B, are plotted in Fig. 12. We obtain a prefactor $\sim 10^6$ for the median and up to $\sim 10^8$ for the 85th quantile.

When optimizing the QMC code to the extent performed here, a methodological concern arises. Since the QMC code has many parameters and modes of execution (e.g., temperature, number of sweeps, annealing schedule, open versus closed boundary conditions in imaginary time, discrete or continuous time Monte Carlo method), overlearning can become an issue when working with just 100 instances. Moreover, optimizations over many parameters will

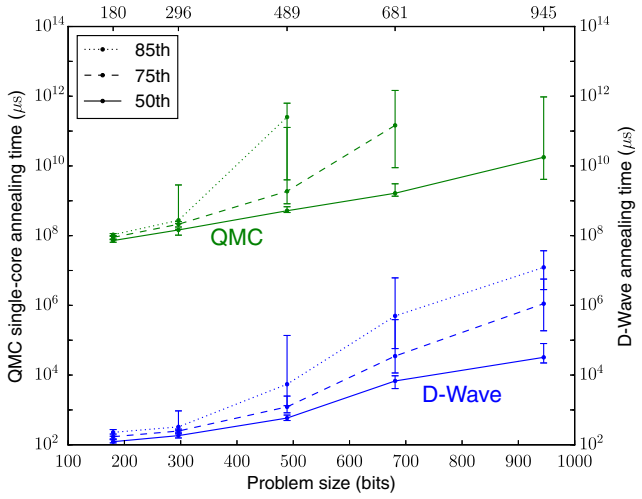

FIG. 12. Time to find the optimal solution with 99% probability for different problem sizes. We compare the QMC method with a linear schedule and the D-Wave 2X. To assign a runtime for the QMC code, we take the number of worldline updates that are required to reach a 99% success probability and multiply that with the time to perform one update on a single state-of-the-art core. Shown are the 50th, 75th, and 85th percentiles over a set of 100 instances. The error bars represent 95% confidence intervals from bootstrapping. The runtimes for the higher quantiles for the larger problem sizes for the QMC method were not computed because the computational cost was too high.

become computationally prohibitive as the problem size increases. By contrast, the current quantum hardware has only a single parameter that can be tuned: the number of annealing sweeps.

[1] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, *Optimization by Simulated Annealing*, Science **220**, 671 (1983).

[2] T. Kadowaki and H. Nishimori, *Quantum Annealing in the Transverse Ising Model*, Phys. Rev. E **58**, 5355 (1998); J. Brooke, D. Bitko, T. F. Rosenbaum, and G. Aeppli, *Quantum Annealing of a Disordered Magnet*, Science **284**, 779 (1999); Y. H. Lee and B. J. Berne, *Global Optimization: Quantum Thermal Annealing with Path Integral Monte Carlo*, J. Phys. Chem. A **104**, 86 (2000); E. Farhi, J. Goldstone, S. Gutmann, J. Lapan, A. Lundgren, and D. Preda, *A Quantum Adiabatic Evolution Algorithm Applied to Random Instances of an NP-Complete Problem*, Science **292**, 472 (2001); G. E. Santoro, R. Martonak, E. Tosatti, and R. Car, *Theory of Quantum Annealing of an Ising Spin Glass*, Science **295**, 2427 (2002).

[3] A. Das and B. K. Chakrabarti, *Quantum Annealing and Related Optimization Methods* (Springer Science and Business Media, Berlin, 2005), Vol. 679.

[4] A. Das and B. K. Chakrabarti, *Colloquium: Quantum Annealing and Analog Quantum Computation*, Rev. Mod. Phys. **80**, 1061 (2008).

[5] R. Harris *et al.*, *Experimental Investigation of an Eight-Qubit Unit Cell in a Superconducting Optimization Processor*, Phys. Rev. B **82**, 024511 (2010).

[6] M. W. Johnson, M. H. S. Amin, S. Gildert, T. Lanting, F. Hamze, N. Dickson, R. Harris, A. J. Berkley, J. Johansson, P Bunyk *et al.*, *Quantum Annealing with Manufactured Spins*, Nature (London) **473**, 194 (2011).

[7] V. Bapst, L. Foini, F. Krzakala, G. Semerjian, and F. Zamponi, *The Quantum Adiabatic Algorithm Applied to Random Optimization Problems: The Quantum Spin Glass Perspective*, Phys. Rep. **523**, 127 (2013).

[8] S. Boixo, T. Albash, F. M. Spedalieri, N. Chancellor, and D. A. Lidar, *Experimental Signature of Programmable Quantum Annealing*, Nat. Commun. **4**, 2067 (2013).

[9] N. G. Dickson, M. W. Johnson, M. H. Amin, R. Harris, F. Altomare, A. J. Berkley, P. Bunyk, J. Cai, E. M. Chapple, P. Chavez *et al.*, *Thermally Assisted Quantum Annealing of a 16-Qubit Problem*, Nat. Commun. **4**, 1903 (2013).

[10] C. C. McGeoch and C. Wang, in *Proceedings of the ACM International Conference on Computing Frontiers* (Association for Computing Machinery, New York, 2013), p. 23.

[11] S. Dash, *A Note on QUBO Instances Defined on Chimera Graphs*, arXiv:1306.1202.

[12] S. Boixo, T. F. Rønnow, S. V. Isakov, Z. Wang, D. Wecker, D. A. Lidar, J. M. Martinis, and M. Troyer, *Evidence for Quantum Annealing with More Than One Hundred Qubits*, Nat. Phys. **10**, 218 (2014).

[13] T. Lanting, A. J. Przybysz, A. Yu Smirnov, F. M. Spedalieri, M. H. Amin, A. J. Berkley, R. Harris, F. Altomare, S. Boixo, P. Bunyk *et al.*, *Entanglement in a Quantum Annealing Processor*, Phys. Rev. X **4**, 021041 (2014).

[14] S. Santra, G. Quiroz, G. V. Steeg, and D. A. Lidar, *MAX 2-SAT with Up to 108 Qubits*, New J. Phys. **16**, 045006 (2014).

[15] T. F. Rønnow, Z. Wang, J. Job, S. Boixo, S. V. Isakov, D. Wecker, J. M. Martinis, D. A. Lidar, and M. Troyer, *Defining and Detecting Quantum Speedup*, Science **345**, 420 (2014).

[16] W. Vinci, K. Markström, S. Boixo, A. Roy, F. M. Spedalieri, P. A. Warburton, and S. Severini, *Hearing the Shape of the Ising Model with a Programmable Superconducting-Flux Annealer*, Sci. Rep. **4**, 5703 (2014).

[17] S. W. Shin, G. Smith, J. A. Smolin, and U. Vazirani, *How "Quantum" Is the D-Wave Machine?*, arXiv:1401.7087.

[18] T. Albash, W. Vinci, A. Mishra, P. A. Warburton, and D. A. Lidar, *Consistency Tests of Classical and Quantum Models for a Quantum Annealer*, Phys. Rev. A **91**, 042314 (2015).

[19] C. C. McGeoch, *Adiabatic Quantum Computation and Quantum Annealing: Theory and Practice*, Synth. Lect. Quantum Comput. **5**, 1 (2014).

[20] D. Venturelli, S. Mandrà, S. Knysh, B. O'Gorman, R. Biswas, and V. Smelyanskiy, *Quantum Optimization of Fully Connected Spin Glasses*, Phys. Rev. X **5**, 031040 (2015).

[21] T. Albash, T. F. Rønnow, M. Troyer, and D. A. Lidar, *Reexamining Classical and Quantum Models for the D-Wave One Processor*, Eur. Phys. J. Spec. Top. **224**, 111 (2015).

[22] A. D. King and C. C. McGeoch, *Algorithm Engineering for a Quantum Annealing Platform*, arXiv:1410.2628.

[23] P. J. D. Crowley, T. Duric, W. Vinci, P. A. Warburton, and A. G. Green, *Quantum and Classical Dynamics in Adiabatic Computation*, Phys. Rev. A **90**, 042317 (2014).

[24] W. Vinci, T. Albash, G. Paz-Silva, I. Hen, and D. A. Lidar, *Quantum Annealing Correction with Minor Embedding*, Phys. Rev. A **92**, 042310 (2015).

[25] I. Hen, J. Job, T. Albash, T. F. Rønnow, M. Troyer, and D. Lidar, *Probing for Quantum Speedup in Spin-Glass Problems with Planted Solutions*, Phys. Rev. A **92**, 042325 (2015).

[26] D. S. Steiger, T. F. Rønnow, and M. Troyer, *Heavy Tails in the Distribution of Time to Solution for Classical and Quantum Annealing*, Phys. Rev. Lett. **115**, 230501 (2015).

[27] D. Venturelli, D. J. J. Marchand, and G. Rojo, *Quantum Annealing Implementation of Job-Shop Scheduling*, arXiv:1506.08479.

[28] B. Bauer, L. Wang, I. Pižorn, and M. Troyer, *Entanglement as a Resource in Adiabatic Quantum Optimization*, arXiv:1501.06914.

[29] T. Albash, I. Hen, F. M. Spedalieri, and D. A. Lidar, *Reexamination of the Evidence for Entanglement in the D-Wave Processor*, Phys. Rev. A **92**, 062328 (2015).

[30] H. G. Katzgraber, F. Hamze, Z. Zhu, A. J. Ochoa, and H. Munoz-Bauza, *Seeking Quantum Speedup through Spin Glasses: The Good, the Bad, and the Ugly*, Phys. Rev. X **5**, 031026 (2015).

[31] N. Chancellor, S. Szoke, W. Vinci, G. Aeppli, and P. A. Warburton, *Maximum-Entropy Inference with a Programmable Annealer*, Sci. Rep. **6**, 22318 (2016).

[32] A. Perdomo-Ortiz, J. Fluegemann, R. Biswas, and V. N Smelyanskiy, *A Performance Estimator for Quantum Annealers: Gauge Selection and Parameter Setting*, arXiv:1503.01083.

[33] A. Perdomo-Ortiz, B. O'Gorman, J. Fluegemann, R. Biswas, and V. N. Smelyanskiy, *Determination and Correction of Persistent Biases in Quantum Annealers*, arXiv:1503.05679.

[34] W. Vinci, T. Albash, and D. A. Lidar, *Nested Quantum Annealing Correction*, arXiv:1511.07084.

[35] S. Boixo, V. N. Smelyanskiy, A. Shabani, S. V. Isakov, M. Dykman, V. S. Denchev, M. Amin, A. Smirnov, M. Mohseni, and H. Neven, *Computational Role of Collective Tunneling in a Quantum Annealer*, arXiv:1411.4036.

[36] S. Boixo, V. N. Smelyanskiy, A. Shabani, S. V. Isakov, M. Dykman, V. S. Denchev, M. H. Amin, A. Y. Smirnov, M. Mohseni, and H. Neven, *Computational Multiqubit Tunnelling in Programmable Quantum Annealers*, Nat. Commun. **7**, 10327 (2016).

[37] In still unpublished experiments we saw evidence of tunneling events involving up to 12 qubits. However, it is difficult to exclude higher-order tunneling processes that may break up the cotunneling group.

[38] K. Kechedzhi and V. N. Smelyanskiy, *Open-System Quantum Annealing in Mean-Field Models with Exponential Degeneracy*, Phys. Rev. X **6**, 021028 (2016).

[39] S. V. Isakov, G. Mazzola, V. N. Smelyanskiy, Z. Jiang, S. Boixo, H. Neven, and M. Troyer, *Understanding Quantum Tunneling through Quantum Monte Carlo Simulations*, arXiv:1510.08057.

[40] B. Altshuler, H. Krovi, and J. Roland, *Anderson Localization Makes Adiabatic Quantum Optimization Fail*, Proc. Natl. Acad. Sci. U.S.A. **107**, 12446 (2010).

[41] E. Farhi, D. Gosset, I. Hen, A. W. Sandvik, P. Shor, A. P. Young, and F. Zamponi, *Performance of the Quantum Adiabatic Algorithm on Random Instances of Two Optimization Problems on Regular Hypergraphs*, Phys. Rev. A **86**, 052334 (2012).

[42] S. Knysh, *Computational Bottlenecks of Quantum Annealing*, arXiv:1506.08608.

[43] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.6.031015 for further technical details.

[44] S. Mandra, Z. Zhu, W. Wang, A. Perdomo-Ortiz, and H. G. Katzgraber, *Strengths and Weaknesses of Weak-Strong Cluster Problems: A Detailed Overview of State-of-the-Art Classical Heuristics vs Quantum Approaches*, arXiv:1604.01746.

[45] The complete set of instances is available at http://goo.gl/yYKcb1.

[46] We take the mean over 16 gauges.

[47] R. Harris, J. Johansson, A. J. Berkley, M. W. Johnson, T. Lanting, S. Han, P. Bunyk, E. Ladizinsky, T. Oh, I. Perminov, E. Tolkacheva, S. Uchaikin, E. M. Chapple, C. Enderud, C. Rich, M. Thom, J. Wang, B. Wilson, and G. Rose, *Experimental Demonstration of a Robust and Scalable Flux Qubit*, Phys. Rev. B **81**, 134510 (2010).

[48] S. V. Isakov, I. N. Zintchenko, T. F. Rønnow, and M. Troyer, *Optimised Simulated Annealing for Ising Spin Glasses*, Comput. Phys. Commun. **192**, 265 (2015).

[49] H. G. Katzgraber, F. Hamze, and R. S. Andrist, *Glassy Chimeras Could Be Blind to Quantum Speedup: Designing Better Benchmarks for Quantum Annealing Machines*, Phys. Rev. X **4**, 021008 (2014).

[50] H. Rieger and N. Kawashima, *Application of a Continuous Time Cluster Algorithm to the Two-Dimensional Random Quantum Ising Ferromagnet*, Eur. Phys. J. B **9**, 233 (1999).

[51] J. King, S. Yarkoni, M. M. Nevisi, J. P. Hilton, and C. C. McGeoch, *Benchmarking a Quantum Annealing Processor with the Time-to-Target Metric*, arXiv:1508.05087.

[52] F. Hamze and N. de Freitas, in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, 2004 (UAI '04)* (AUAI Press Arlington, Virginia, 2004), pp. 243–250.

[53] A. Selby, *Efficient Subgraph-Based Sampling of Ising-Type Models with Frustration*, arXiv:1409.3934.

[54] I. Zintchenko, M. B. Hastings, and M. Troyer, *From Local to Global Ground States in Ising Spin Glasses*, Phys. Rev. B **91**, 024201 (2015).

[55] Nevertheless, it is interesting to note that Ref. [15] defines a limited quantum speedup as "a speedup obtained when comparing specifically with classical algorithms that correspond to the quantum algorithm in the sense that they implement the same algorithmic approach, but on classical hardware … a natural example is quantum annealing implemented on a candidate physical quantum information processor versus either classical simulated annealing, classical spin dynamics, or simulated quantum annealing." In the weak-strong cluster networks benchmark we show that D-Wave 2X outperforms these three algorithms. We find a substantial advantage against QMC in the prefactor, not the scaling. For the comparison with classical spin dynamics, see Ref. [35].

[56] Z. Zhu, A. J. Ochoa, and H. G. Katzgraber, *Efficient Cluster Algorithm for Spin Glasses in Any Space Dimension*, Phys. Rev. Lett. **115**, 077201 (2015).

[57] V. Martin-Mayor (private communication).

[58] S. Coleman, *Fate of the False Vacuum: Semiclassical Theory*, Phys. Rev. D **15**, 2929 (1977).

[59] A. I. Vainshtein, V. I. Zakharov, V. A. Novikov, and M. A. Shifman, *ABC of Instantons*, Sov. Phys. Usp. **25**, 195 (1982).

[60] E. M. Chudnovsky and J. Tejada, *Macroscopic Quantum Tunneling of the Magnetic Moment* (Cambridge University Press, Cambridge, England, 1998).

[61] N. Nagaosa, *Quantum Field Theory in Condensed Matter Physics* (Springer Science and Business Media, New York, 2013).

[62] Z. Jiang, V. N. Smelyanskiy, S. V. Isakov, S. Boixo, G. Mazzola, M. Troyer, and H. Neven, *Scaling Analysis and Instantons for Thermally-Assisted Tunneling and Quantum Monte Carlo Simulations*, arXiv:1603.01293.

[63] J. S. Langer, *Statistical Theory of the Decay of Metastable States*, Ann. Phys. (N.Y.) **54**, 258 (1969).

[64] V. N. Smelyanskiy, S. Isakov, S. Boixo, and H. Neven, "*Tunneling in Spin Systems with Disorder*" (to be published).

[65] A. Garg, *Application of the Discrete Wentzel-Kramers-Brillouin Method*, J. Math. Phys. (N.Y.) **39**, 5166 (1998).

[66] A. O. Caldeira and A. J. Leggett, *Quantum Tunnelling in a Dissipative System*, Ann. Phys. (N.Y.) **149**, 374 (1983).

[67] M. H. S. Amin and D. V. Averin, *Macroscopic Resonant Tunneling in the Presence of Low Frequency Noise*, Phys. Rev. Lett. **100**, 197001 (2008).

[68] R. Harris, M. W. Johnson, S. Han, A. J. Berkley, J. Johansson, P. Bunyk, E. Ladizinsky, S. Govorkov, M. C. Thom, S. Uchaikin, B. Bumble, A. Fung, A. Kaul, A. Kleinsasser, M. H. S. Amin, and D. V. Averin, *Probing Noise in Flux Qubits via Macroscopic Resonant Tunneling*, Phys. Rev. Lett. **101**, 117003 (2008).

[69] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, San Francisco, 1979).

[70] S. Mertens, *Phase Transition in the Number Partitioning Problem*, Phys. Rev. Lett. **81**, 4281 (1998).

[71] S. Mertens, *A Physicist's Approach to Number Partitioning*, Theor. Comput. Sci. **265**, 79 (2001).

[72] S. Mertens, *Random Costs in Combinatorial Optimization*, Phys. Rev. Lett. **84**, 1347 (2000).

[73] B. Yakir, *The Differencing Algorithm LDM for Partitioning: A Proof of a Conjecture of Karmarkar and Karp*, Math. Oper. Res. **21**, 85 (1996).

[74] V. N. Smelyanskiy, U. v. Toussaint, and D. A. Timucin, *Dynamics of Quantum Adiabatic Evolution Algorithm for Number Partitioning*, arXiv:quant-ph/0202155.

[75] P. F. Stadler, W. Hordijk, and J. F. Fontanari, *Phase Transition and Landscape Statistics of the Number Partitioning Problem*, Phys. Rev. E **67**, 056701 (2003).

[76] N. Howgrave-Graham and A. Joux, in *Advances in Cryptology—EUROCRYPT 2010* (Springer, Berlin, 2010), pp. 235–256.

[77] A. Becker, J.-S. Coron, and A. Joux, in *Advances in Cryptology—EUROCRYPT 2011* (Springer, Berlin, 2011), pp. 364–385.

[78] D. J. Bernstein, S. Jeffery, T. Lange, and A. Meurer, in *Post-Quantum Cryptography* (Springer, Berlin, 2013), pp. 16–33.

[79] J. E. Angus, *The Asymptotic Theory of Extreme Order Statistics*, Technometrics **32**, 110 (1990).

[80] V. N. Smelyanskiy, D. Venturelli, A. Perdomo-Ortiz, S. Knysh, and M. I. Dykman, *Quantum Annealing via Environment-Mediated Quantum Diffusion*, arXiv:1511.02581.

[81] W. Lechner, P. Hauke, and P. Zoller, *A Quantum Annealing Architecture with All-to-All Connectivity from Local Interactions*, Sci. Adv. **1**, e1500838 (2015).

[82] Andris Ambainis, *Quantum Walks and Their Algorithmic Applications*, Int. J. Quantum. Inform. **01**, 507 (2003).

[83] A. M. Childs, R. Cleve, E. Deotto, E. Farhi, S. Gutmann, and D. A. Spielman, in *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, 2003), pp. 59–68.

[84] J. Kempe, *Quantum Random Walks: An Introductory Overview*, Contemp. Phys. **44**, 307 (2003).

[85] C. L. Baldwin, C. R. Laumann, A. Pal, and A. Scardicchio, *The Many-Body Localized Phase of the Quantum Random Energy Model*, Phys. Rev. B **93**, 024202 (2016).

[86] The annealing function $A(s)$ decreases linearly from 1 to 0, while $B(s)$ increases linearly from 0 to 1.