PERSPECTIVE

Test-driving 1000 qubits

To cite this article: Joshua Job and Daniel Lidar 2018 *Quantum Sci. Technol.* **3** 030501

IOP Publishing Quantum Sci. Technol. 3 (2018) 030501 https://doi.org/10.1088/2058-9565/aabd9b

Quantum Science and Technology

PERSPECTIVE

Test-driving 1000 qubits

PUBLISHED
19 June 2018

Joshua Job[1,2] **and Daniel Lidar**[1,2,3,4,5]

[1] Department of Physics and Astronomy, University of Southern California, Los Angeles, California 90089, United States of America
[2] Center for Quantum Information Science & Technology, University of Southern California, Los Angeles, California 90089, United States of America
[3] Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089, United States of America
[4] Department of Chemistry, University of Southern California, Los Angeles, California 90089, United States of America
[5] Author to whom any correspondence should be addressed.

E-mail: lidar@usc.edu

Keywords: quantum annealing, quantum computing, quantum algorithm validation, quantum computing benchmarking, quantum error correction

Abstract

Quantum computing is no longer a nascent field. Programmable quantum annealing (QA) devices with more than 1000 qubits are commercially available. How does one know that a putative QA device is indeed quantum? How should one go about benchmarking its performance and compare it to classical algorithms? How can its performance be improved by error correction? In this contribution to the focus collection on 'What would you do with 1000 qubits?', we review the work we and others have done in this area, since the first D-Wave quantum annealer with 108 qubits was made available to us. Some of the lessons we have learned will be useful when other quantum computing platforms reach a similar scale, and practitioners will attempt to demonstrate quantum speedup.

1. Introduction

As we look forward to the rapid development of new quantum computing devices with hundreds or a few thousand qubits, particularly commercial devices and non-gate-based devices such as quantum annealers, we are faced with a challenge. How does one ensure such devices really do what they claim, and are not effectively classical? How does one evaluate the performance of such a device, what methods should one use to estimate performance on a given metric, and what metrics should one use? How do we do maintenance on the quantum state and ensure we can prevent or correct breakdowns and errors? These questions have to be settled before we can decide where to take our device on a test drive, and what problems we should use our quantum computing devices to try to solve.

At this time, these new devices and plans for quantum annealing (QA) devices and various other quantum computing platforms are no longer the first of their kind. Several generations of programmable quantum annealers from D-Wave Systems have been made available to a small community of researchers, which has worked hard to answer the aforementioned questions. This community began largely groping in the dark, and has over the last six years answered many of the most basic questions, developing techniques to validate quantum annealers, methods to benchmark and estimate performance, and developing methods to suppress errors given the constraints of existing quantum annealers.

We have been fortunate to be members of the aforementioned community, which has given us an opportunity to work with the first several generations of quantum annealers, starting from the first commercially available such device, the 128-qubit D-Wave One (DW1) 'Rainier' processor, through two more generations of 512 and 1152 qubits, to the current 2048-qubit D-Wave 2000Q processor[6]. As such, rather than

[6] A brief history: the Rainier processor (108 operational qubits) was the first to be installed at the USC-Lockheed Martin Quantum Computing Center at the USC Information Sciences Institute in 2011. Upgrades to the 'Vesuvius' (504 operational qubits) and 'Washington' (1098 operational qubits) processors followed in 2013 and 2016, respectively. Google installed 'Vesuvius' (509 operational qubits) and 'Washington' (1097 operational qubits) processors in the same years at NASA Ames. Los Alamos National Lab installed a 'Washington' processor (1095 operational qubits) in 2016. The 2000Q processor was deployed at NASA Ames in 2017 (2027 operational qubits).

answering the question 'what would you do with 1000 qubits?', in this work we will answer the question 'what have we done with 1000 qubits?'. The discussion will draw mainly from the research we have done on quantum annealers, and we apologize in advance to the many others who have contributed to this enterprise for not doing their work justice. We expect that some of the lessons learned will inform studies of future classes of quantum computing devices with many qubits. Our presentation aims to remain at a fairly high level, without giving a detailed technical account, for which we refer the reader to the original literature cited.

2. Quantum validation testing

Perhaps the first question one might ask when offered a quantum computational device is whether or not it is, in fact, quantum. In the case of quantum computational devices based on the circuit model and/or gates for quantum computing, the task of validation can be reduced to a Clauser–Horne–Shimony–Holt test between two parts of the device that are treated as black boxes [1]. Alternatively, one may opt for quantum process tomography [2, 3] or quantum gate set tomography [4, 5], wherein one applies many small computations and measures the results, verifying that they match the predictions of quantum theory. These predictions are available because the quantum computations in question typically involve few qubits and are thus readily implementable [6, 7].

However, for other quantum computing paradigms, such as QA [8, 9] and the broader field inspired by adiabatic quantum computing (AQC) [10–13], quantum tomography is not currently available for validation. This is for a variety of reasons. The key difference is that gate-based computations are modular: they can be broken into discrete time-local and space-local operations, operating effectively on only one or two qubits at a time, with the others left essentially unaffected, so the only requirement to validate even a long chain of computations is to validate those one- and two-qubit operations on individual qubits and pairs of qubits. For AQC-like platforms, the quantum computation is composed of a continuously time-varying Hamiltonian with many computational operators acting on the system at the same time. They are non-modular in the sense that they cannot be easily broken down into discrete chunks which can be validated separately. Future versions of such platforms may be more flexible and allow for approaches such as quantum tomography, but will still be unable to validate arbitrarily large computations due to the aforementioned nonmodularity of the computation. Meanwhile, partial alternatives such as tunneling spectroscopy have already been explored [14]. Of course, in the absence of error correction and fault tolerance neither the gate model nor AQC are guaranteed successful validation.

Nevertheless, certain lessons can be ported over to non-gate-based approaches. One should, as in the circuit model, focus on small problems, with a small number of qubits, and one may hope that by studying such problems applied to many such overlapping sets that one can at least partially validate the operation of the device. From here, two paths for validation become available, depending on whether one can 'open the black box' and perform measurements during the anneal or use measurements beyond what may be considered 'native' to the device, or whether one is only able to use the device's output at the end of complete runs for testing.

2.1. Types of validation: proof of quantumness, quantum supremacy, speedup-inferred quantumness, and classical model rejection

As a case in point, for AQC-style algorithms a system is typically initialized in the ground state of a simple driver Hamiltonian, in most cases a transverse field 'driver Hamiltonian' $H_0 = -\sum_{i=1}^{N} \sigma_i^x$ for N qubits (the σ's are Pauli operators), and then the Hamiltonian is slowly modified into the 'problem Hamiltonian' H_1 via the transformation

$$H(s) = A(s)H_0 + B(s)H_1, \tag{1}$$

where $A(s)$ and $B(s)$ are monotonically decreasing and increasing functions of the dimensionless time $s = t/t_f$, respectively, where $t \in [0, t_f]$ and t_f denotes the annealing time. Here H_1 encodes the problem via programmable parameters $\{h_i\}$ ('local fields' or 'biases') and $\{J_{ij}\}$ ('couplers'):

$$H_1 = \sum_{i=1}^{N} h_i \sigma_i^z + \sum_{i<j}^{N} J_{ij} \sigma_i^z \sigma_j^z. \tag{2}$$

For the D-Wave quantum annealers the h_i and J_{ij} terms are programmable with values bounded between $[-2, 2]$ and $[-1, 1]$, respectively. The form in equation (1) is the general form for an Ising model quantum annealer, and as written every qubit in the final Hamiltonian is connected to every other qubit. In reality (for example the D-Wave architecture), full connectivity is difficult to achieve, and as such there may be additional restrictions on the J_{ij}, such that they can be nonzero if the nodes i and j are connected on the hardware graph of the device. An example of the 'Chimera' hardware connectivity graph of a D-Wave Two X (DW2X) processor is shown in figure 1, and its 'annealing schedule', the $A(s)$ and $B(s)$ curves, are shown in figure 2.

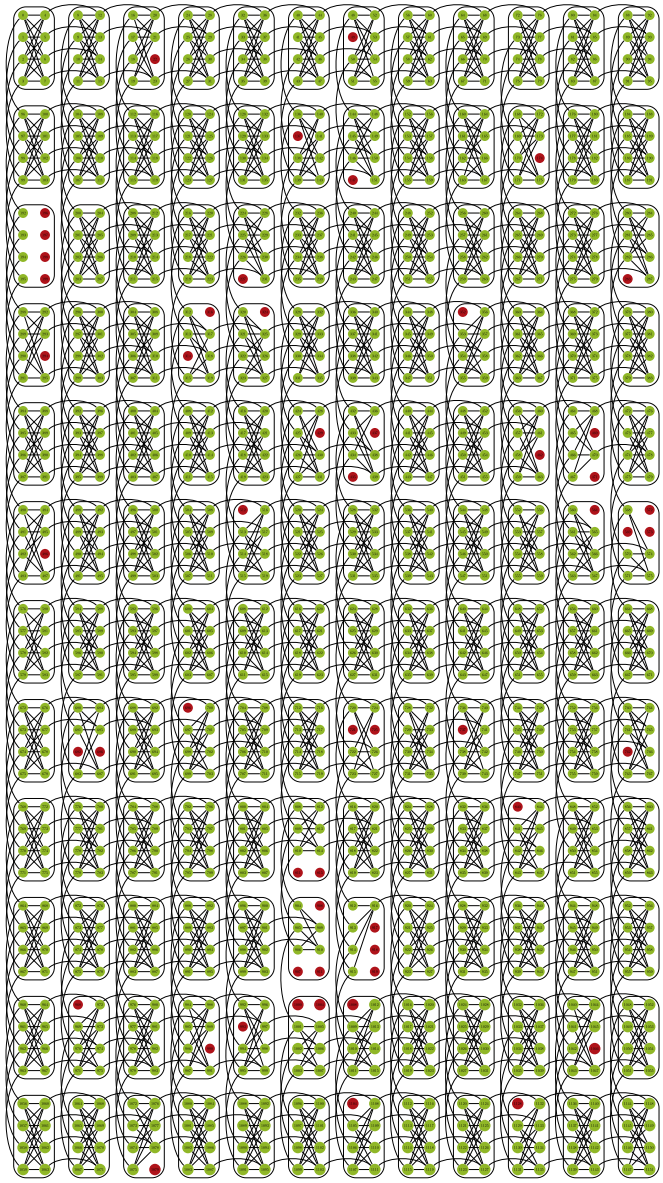


Figure 1. An 1152 qubit Chimera graph describing the D-Wave Two X processor at the University of Southern California's Information Sciences Institute. Inactive qubits are marked in red, active qubits (1098) are marked in green. Black lines denote active couplings (where J_{ij} is programmable to be in the range $[-1, 1]$) between qubits.

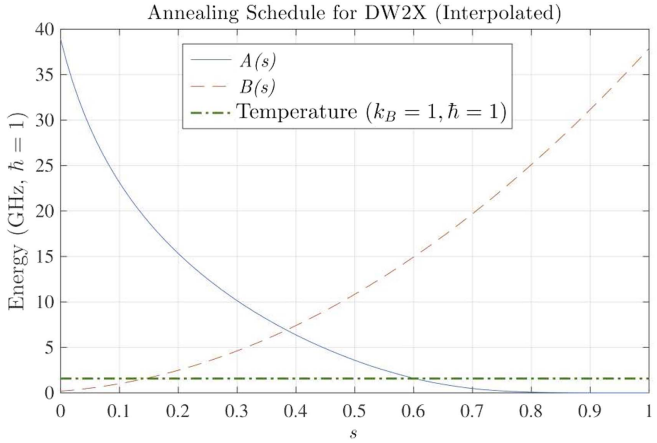


Figure 2. Annealing schedules for the D-Wave Two X processor described in figure 1.

In validating quantum annealers, one seeks to create an assignment to the h's and J's such that one can take some measurements which will conclusively demonstrate, for instance, quantum entanglement, in what might be called an experimental '*proof of quantumness*'.

A somewhat weaker and indirect type of validation is provided by '*quantum supremacy*' experiments [15], since they have the potential for complexity theoretic guarantees[7]. More specifically, quantum supremacy is a scenario where (part of) the polynomial hierarchy of complexity theory collapses if the quantum result could be replicated classically without slowdown [17–22]. While weaker than a direct proof of quantumness, a demonstration of quantum supremacy would be considered strong evidence for quantum computational power of a device, which may be considered inherently more interesting than a direct demonstration of, e.g., entanglement.

'*Speedup-inferred quantumness*' is a related type of indirect validation based on a demonstration of quantum speedup [23] over the best classical solvers known for a task, which is often considered the holy grail of quantum information processing. Unlike quantum supremacy tests, speedup-inferred quantumness tests do not have complexity theoretic guarantees (an example in the circuit model would be Shor's algorithm [24]). It appears that an unqualified quantum speedup would necessarily have to invoke quantum properties, and this might happen even if these properties remain poorly understood or characterized. Thus a certificate of quantumness might be assigned even in the absence of a direct demonstration of quantum properties such as entanglement. It should be recognized that this carries a certain element of risk. For example, suppose a new *classical* optimization is discovered that outperforms all other classical and quantum optimization algorithms known to date (this is in fact what happened recently in a tug-of-war between quantum and classical optimization for the Max E3LIN2 problem [25]). This algorithm could be deceptively marketed as a quantum algorithm providing speedup-inferred quantumness by a shrewd company claiming to own quantum computers, that provides black box access only to run the new optimization algorithm. Thus any claim of speedup-inferred quantumness should always be treated with a healthy degree of skepticism as related to its quantum underpinnings, until actual evidence of quantum effects driving the algorithm is presented.

If none of prior three types (proof of quantumness, quantum supremacy, speedup-inferred quantumness) of validation are attainable, one may alternatively seek to show that on sufficiently small scale problems the results are only readily reproducible using a truly quantum model of the device, and cannot be replicated qualitatively using any existing classical model, in what might be called '*classical model rejection*'. This type of validation experiment does not provide a certificate of quantumness, since one can always invent a new and better classical model. Instead, one can only hope to exclude all 'physically reasonable' classical models for the device. Moreover, classical model rejection can only be performed as long as it is feasible to carry out quantum model simulations, which limits system sizes to about 20 qubits for master equation type models, using the quantum trajectories method [26]. Extrapolations to larger sizes are, as always, risky in the absence of fault tolerance guarantees.

One caveat regarding 'proof of quantumness' experiments is noteworthy. While demonstrations of entanglement can be considered 'proof of quantumness', they often require additional physical resources and measurement possibilities beyond those that may natively be embedded in a (commercial) quantum computational device or that are strictly required to implement the core algorithm, and thus may be impossible on certain platforms. Additionally, in practice, certain assumptions may be made in a 'proof of quantumness' experiment which, when relaxed, render it effectively a 'classical model rejection' experiment; we shall shortly see an example of this with the D-Wave quantum annealers.

2.2. Experimental implementations of quantum validation tests

The primary 'proof of quantumness' experiment for quantum annealers was performed in [27], using an entanglement test on the D-Wave Two (DW2) generation of processors. Briefly, the work used quantum tunneling spectroscopy [14] to estimate the populations of the first and second excited states of a combined probe-system Hamiltonian. They also measured the energy spectrum and found it to be consistent with the Hamiltonian the device was designed to implement, which provided a justification for the assumption that the measured populations were those of the energy eigenstates of the Hamiltonian. This allowed for a reconstruction of the density matrix under the assumption that it is diagonal in the energy eigenbasis, enabling a computation of the negativity [28] for all possible bipartitions of the system, the geometric mean of which was taken as a measure of the entanglement of the system. As it was found to be nonzero, the system is entangled. Further, by exploiting the theory of entanglement witnesses [29], [27] was able to show that even if the diagonality assumption is relaxed, the entanglement remains. This was used to conclude that the DW2 system tested displays entanglement at least on the scale of a single 8-qubit unit cell.

[7] The term 'supremacy' has generated considerable controversy [16]. While we would prefer the adoption of an alternative such as 'hegemony' or 'supremeness', we recognize that 'supremacy' is likely here to stay due to its current widespread usage.

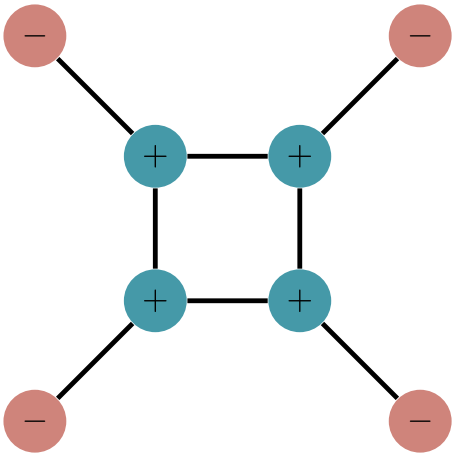


Figure 3. The eight-spin Ising quantum signature Hamiltonian introduced in [32]. The inner 'core' spins (green circles) have local fields $h_i = +1$ while the outer spins (red circles) have $h_i = -1$. All couplings are ferromagnetic: $J_{ij} = 1$ (black lines).

It was noted in [30] that these tests depended on the assumption that the device was well-described by equation (1) for an appropriate (programmed) choice of local fields and couplers for which the ground state is entangled, and that this assumption is not directly demonstrable by the experiments in [27]. Without that assumption, one must revert to a 'classical model rejection' experiment in which one compares results of direct quantum simulations of the device and available classical alternatives to demonstrate that only the quantum model is consistent with the experimental observations. Reference [30] provides a detailed description of the experiments, but for our purposes the key takeaway is that only the quantum adiabatic master equation [31] can reproduce the output distribution from experiments, validating the approach in [27].

Another branch of validation experiments of the classical model rejection type are the so-called 'quantum signature' Hamiltonians and the consistency tests derived therefrom, introduced in [32], critiqued in [33], and further explored in [34–36]. Unlike the aforementioned entanglement tests, these experiments do not require access to the system during the annealing process, and are appropriate for cases in which the quantum device is a 'black box' in which one can only control the inputs and measure the outputs. An example of a quantum signature Hamiltonian is shown in figure 3. These Hamiltonians take the form of a ring of tightly bound qubits each connected to a single outer qubit. The resulting Hamiltonian has the property that there is a large ($2^{N/2}$-dimensional) degenerate subspace of ground state configurations corresponding to arbitrary assignments to the outer qubits where all qubits in the inner ring are in the state 0 (forming a 'cluster' connected by single-spin flips applied to the outer ring). There is one additional ground state corresponding to flipping all the inner qubits to the state 1, dubbed the 'isolated' state, since for a signature Hamiltonian with $2N$ qubits it is at least N spin flips away from all other ground states. Thermal algorithms such as classical simulated annealing (SA) [37] will be weighted toward the isolated state, such that it will have the highest probability of occurrence of any ground state configuration, whereas an adiabatic quantum evolution will find the isolated state to be suppressed relative to the cluster states. Extensive simulations and experiments on a 108-qubit DW1 'Rainier' processor matched qualitatively with the adiabatic master equation across all the parameters and statistics of the output distributions tested, though noise due to cross-talk made it very difficult to find quantitative agreement; at the same time all existing classical models failed to qualitatively match the DW1 in at least one of the tests [35].

A different approach to classical model rejection was taken in [38], which used random $J_{ij} = \pm 1$ instances of the Ising Hamiltonian in equation (2) to test the hypothesis that QA correlates well with two classical models: SA and classical spin dynamics [33] (also known as the Landau–Lifshitz–Gilbert model). The hypothesis was tested using the same DW1 processor. This work showed that these two classical models failed to correlate with the results for the distribution of ground state probabilities generated by the DW1 device, while the DW1 correlated very well with simulated quantum annealing (SQA), implemented using quantum Monte Carlo [39]. This was taken as evidence for QA on the scale of more than 100 qubits, thus generalizing the conclusion of the earlier result [32] based on the 8-qubit 'gadget' shown in figure 3. Shortly thereafter a new semiclassical spin-vector Monte Carlo (SVMC) model was introduced, also known as SSSV, the author initials of [40] [8]. In this model spins are treated as $O(2)$ rotors (effectively as single qubits), evolved according to the annealing schedule given in equation (1), with Monte Carlo angle updates. The SVMC model correlated well with both the DW1 and SQA

[8] Both the spin dynamics and SVMC models can be derived in a strong coupling limit from the anisotropic Langevin equation, starting from Keldysh field theory [41].

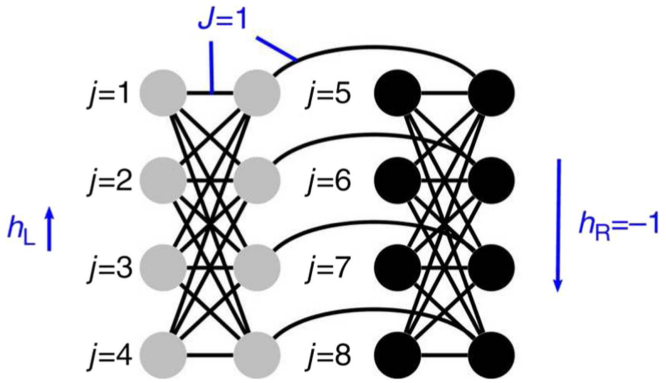


Figure 4. The 16-spin Ising Hamiltonian composed of two $K_{4,4}$ unit cells introduced in [46]. All couplings are set to $J = 1$, all qubits in the left unit cell have a local field $0 < h_L < 0.5$ applied to them while all spins in the left unit cell have $h_R = 1$ applied to them. Two local minima form, one with the cells internally aligned but in opposite states from each other (a local minimum) and the other with all states aligned with h_R (the global minimum). By tightly binding each unit cell, they effectively act like single large spins.

data, suggesting that although the DW1 device's performance is consistent with QA, it operated in a temperature regime where, for most random Ising spin glass instances, a quantum annealer may have an effective semiclassical description. This conclusion was challenged in [42], which considered the excited state distribution rather than just the ground state distribution over random $J_{ij} = \pm 1$ Ising instances, as well as the ground state degeneracy. This work presented evidence that for these new measures neither SQA nor SVMC, which are classically efficient algorithms, correlated well with the DW1 experiments. The close correlation between SQA and the SVMC model was explained by showing that the SVMC model represents a semiclassical limit of the spin-coherent states path integral, which forms the foundation for the derivation of the SQA algorithm.

The intense debate that arose around the original classical model rejection tests presented in [32, 38], in particular the critique presented in [33, 40], illustrates the risks associated with such tests—risks that materialize whenever a sufficiently clever new classical model is found that agrees with (some of) the data—as well as the fruitfulness of the classical model rejection approach, which can lead to a healthy updating and sharpening of models and assumptions.

Black box classical model rejection tests such as the quantum signature Hamiltonians provide the basis for the testing of new putative quantum devices for which available controls and potential measurements are limited, and ultimately even the best experiments that seek to prove entanglement will depend on a series of such experiments to demonstrate that only quantum models can reproduce the experimental data from the device. Quantum supremacy tests are a type of limiting case of this, in which one can prove that should any classical device be able to produce a particular output distribution in polynomial time then the computational complexity hierarchy will at least partially collapse. Since this is not expected to occur, building a device which can produce said distribution efficiently will then immediately rule out all classical models for the device [17].

Another kind of black box classical model rejection test is based on the phenomenon of quantum tunneling, whereby a quantum state has sizable probability on either side of an energy barrier which the system could not move through classically, or at least will only be able to do so with reasonable probability at high temperature. The first QA experiments involving tunable tunneling were carried out using the disordered ferromagnet $LiHo_xY_{1-x}F_4$ in a transverse magnetic field [43, 44], and served as inspiration for the design of programmable superconducting flux-qubit based quantum annealers. These experiments indicated that QA hastens reduction of the residual energy (i.e., the energy above the ground state) via tunneling, compared to simple thermal hopping models. The first programmable quantum annealer experiment was reported in [45], in which it was demonstrated that an 8-qubit QA device was able to reproduce the domain wall tunneling predictions of quantum theory for a chain of superconducting flux qubits by modifying the time during the annealing process at which a local field is abruptly applied to the qubits. This contradicted the temperature dependence predictions of a classical thermal hopping model, thus serving as a classical model rejection experiment.

More recently, [46] reported on a specially designed tunneling probe Hamiltonian for QA, illustrated in figure 4. The probe uses two unit cells of the D-Wave Chimera graph, binding each one together tightly so they each act like a single effective spin, or cluster. Opposite magnetic fields are applied to each unit cell, one weak and one strong, so that the spins in the 'strong' cluster align before the spins in the 'weak' cluster. Initially, there is only a single minimum. A second minimum develops over the course of the anneal, and eventually becomes the global minimum of the final Ising Hamiltonian. The only way to reach the global minimum is to overcome an energy barrier whose strength increases as the anneal progresses, a good example of tunneling. Using the non-interacting blip approximation (NIBA) it was shown in [46] that the system effectively acts like a two-level

system even in the open-system setting with a strongly coupled bath. NIBA-based predictions without free parameters for tests at different values of h_L and different temperatures demonstrated very good agreement with experiments involving a DW2 device, and were not reproducible using classical models for the device such as SVMC [40]. A variant of this experiment was reported on in [47], which introduced a new class of problem instances which couples the weak-strong clusters of the tunneling probe as sub-blocks of the Hamiltonian. This work can be interpreted as an attempt to go from classical model rejection to speedup-inferred quantumness, as it claimed a large tunneling-induced constant-factor speedup over classical SA and SQA for a DW2X device. However, this claim was critiqued in [48] on the basis of a comparison to classical algorithms with better performance. Moreover, as we discuss below, speedup-inferred quantumness requires a demonstration of an optimal annealing time [23], which was absent in the results reported in [47].

Validating non-gate-based quantum devices will continue to be a challenge as new such systems come online, but applying combinations of the techniques discussed above, from the construction of quantum signature Hamiltonians and tunneling probes to (in)direct proofs of entanglement via entanglement witnesses and direct computation of entanglement, should allow one to boost confidence that the system obeys the predictions of quantum theory *over small scales*. The challenge remains to extend these techniques so that they are able to demonstrate conclusively that a device with hundreds or thousands of qubits displays coherence and long-range entanglement. Due to decoherence this presents a challenge for gate-based quantum devices as well even at a smaller scale [49, 50], and speedup-inferred quantumness tests may prove to be simpler to execute than direct quantumness tests even in the gate model setting.

Once one has validated that the device works in approximately the manner one would expect from the instruction manual, one can then turn to the question: 'Toward what practical purpose may this device be put?'. The choice of appropriate problems in this domain is a complex issue that we cannot address here. Indeed, before we can answer that question, we must first focus on an operational question: 'How does one go about comparing performance in a specified problem domain between a verified quantum computing device and existing classical strategies?'. We discuss this next.

3. Benchmarking

Assume we have at our disposal a device verified to be quantum, at least provisionally on the small scales covered by classical model rejection, and we would like to compare its performance to competing classical solvers. This is the task we refer to here as benchmarking, which belongs more generally to the field of experimental algorithmics [51]. Specifically, consider the problem of estimating the value of some function of merit (or 'reward') R from the output of a given solver (e.g., our quantum device or some classical algorithm) for a given problem family $\mathcal{P} = \{P\}$. Each problem instance P is parametrized by some parameters θ. In the case of quantum annealers, particularly studies of the D-Wave devices thus far, the goal has generally been to find the ground state of Ising Hamiltonians as defined in equation (2). In that context, typically the reward is taken to be the negative of the time to solution (TTS), defined as $\mathrm{TTS} = t_f \log(1-p_d)/\log(1-p)$ for a probability p of finding the ground state at least once with desired probability p_d (typically 0.99), and annealing time t_f[9]. In the language above, $R = -\mathrm{TTS}$ (one would like to minimize the TTS), and the problem is parametrized by $\theta = \{h_i, J_{ij}\}$. Many similar metrics have been proposed, such as time-to-epsilon and time-to-target (TTT) [52], which amount to mild generalizations of TTS. A more elaborate notion of cost, based on optimal stopping theory, has also been considered and shown to recover the previous metrics as special cases [53]. We shall return to this below.

3.1. Solvers for comparison

The choice of classical solvers against which to compare the quantum device involves a few considerations. It is important to perform an apples-to-apples comparison, in that if the device is probabilistic, it would be misleading to measure its performance against a deterministic algorithm [38, 54, 55]. For a quantum annealer, a performance comparison to known heuristic algorithms for sampling low-energy states from Ising models is natural, such as SA [37, 56], parallel tempering [57–59], and the Hamze–Freitas–Selby (HFS) algorithm (which searches all states on nodes that make up induces trees or small-treewidth subgraphs of the Ising model's connectivity graph) [60, 61]. One might also compare to approximations of QA itself, in particular SQA [39, 62, 63], or the SVMC algorithm [40]. All of these can be said to be 'solvers' for the Ising problem on QA. But, to determine if the quantum device is truly useful in practice, it must also be compared to the *best* algorithm for solving the *original* (typically non-Ising problem) task. For example, when solving the graph isomorphism

[9] The probability of not finding the ground state even once after k independent runs of duration t_f each is $(1-p)^k$, so the probability of finding it at least once is $1-(1-p)^k$, which we set equal to p_d. Solving for k and substituting into $\mathrm{TTS} = t_f k$ gives the TTS formula. See, e.g., [23] for a more detailed derivation.

problem [64], job-shop scheduling [65], operational planning [66], or portfolio optimization [67], the original problem must first be mapped into an Ising problem [68] and then embedded using the existing hardware connectivity graph [69–71]; the performance of the quantum device must be compared to the best algorithm for solving the original problem, and the mapping plus embedding steps can severely reduce performance. Note also that determining what the truly optimal classical algorithm is can be a daunting, or even impossible, challenge. In many cases one settles for an educated guess: the standard and/or currently best known algorithm(s). Finally, it is important to remember that any tests run on a quantum device that does not enjoy a fault tolerance guarantee cannot be reliably extrapolated to arbitrarily large sizes. I.e., in the absence of such a guarantee, a finite-size device provides evidence of what can be expected at larger sizes only provided that quantities such as the device temperature, coupling to the environment, and calibration and accuracy errors, can be appropriately scaled down. With this in mind, let us turn to a discussion of much of the benchmarking work done so far and some of the considerations that go into using large, noisy quantum devices.

3.2. The state of benchmarking

The first comprehensive study benchmarking QA devices was [38], using a 108-qubit DW1 processor. This article introduced many of the concepts used in later studies in the field, including the above definition of the TTS. It focused on the performance on the set of random Ising problems with binary ± 1 local fields and couplings, and introduced the use of SA and SQA as important comparison algorithms. It also noted the importance of comparing against parallelized versions of classical algorithms, as quantum annealers such as the D-Wave device consume linearly more computational hardware with increasing problem size, and in many cases SA and SQA can be effectively parallelized in much the same way.

Another significant contribution of [38] was the use of 'gauge averaging' in benchmarking, a technique that was introduced in [32] (where it was called 'spin inversions') and which has become so universal that it is now included natively in the D-Wave API for their processors, and which points toward a more general consideration for noisy quantum devices in the absence of quantum error correction. The need for gauge averaging arises from the observation that in QA, one may have per-qubit or per-edge random and systematic biases from stray fields or interactions. In such cases, performance may be dramatically impacted by the choice of mapping from a logical Hamiltonian as defined in equation (2) to a physically implemented computation. In essence, a gauge transformation corresponds to swapping which physical spin state corresponds to a computational 0 or 1. In an ideal annealer, this transformation commutes with (i.e., is a symmetry of) the total Hamiltonian and so has no dynamical effect. However in the presence of noise, this symmetry is broken and the choice of gauge does make a difference, and indeed was found to have a significant effect on the performance of the DW1 quantum annealer, to such a degree that the device did not even correlate with itself if one compares one gauge to another, or even one gauge with itself when run later (most likely the result of slow drift $1/f$ noise resulting in the effect that each time the annealer is programmed, a small random error term is added to the Hamiltonian). However, when results for the same Hamiltonian were averaged across many gauges, the DW1 processor correlated quite well with itself [38]. Since then, applying many gauge transformations to the same Hamiltonian and averaging the results has become a standard practice in the QA community, and the idea behind it has been steadily generalized since then to include sampling over every known potentially broken symmetry of the Hamiltonian.

For example, if one is solving a fully connected Ising problem, the Hamiltonian has a permutation symmetry. Since every logical spin has an interaction with every other, one can relabel which spin is which without changing anything about the logical problem. However, when one goes to implement such a problem on an actual quantum annealer with limited connectivity, such as the DW2, one has to perform a minor embedding (ME) in which each logical spin is mapped to a chain of spins on the physical device [69, 70]. Those physical spins may have local field biases which vary from chain to chain, and thus the distribution over logical states will depend, in part, on the assignment of the logical spin variables to the physical chains, as shown in [72]. This work was the first case study of both ME of fully connected problems as well as permutation embeddings for such problems, and demonstrated the importance of optimizing the strength of the coupling in ME applications, a topic which is discussed in more detail in section 4.

Reference [55] demonstrated evidence for the easy-hard-easy phase transition for Max 2-SAT problems (wherein one wishes to find the maximal number of simultaneously satisfiable two-variable Boolean clauses over a set of variables from some ensemble of clauses) near a clause density of one, on the 108-qubit DW1 processor. It performed a rudimentary benchmarking comparison between the DW1 and an exact Max 2-SAT solver (akmaxsat) (see also [54]), and noted that there was no correlation between the two solvers over randomly selected instances of Max 2-SAT. This work also introduced the important idea of bootstrapping into the QA community, variants of which (such as the Bayesian bootstrap [73]) formed the backbone of error analyses for later studies, as a nonparametric method for approximating the distribution over the problem space and over the aforementioned broken computational symmetries.

A decisive step forward was taken in [23], which introduced the notion of different quantum speedup categories. Of particular interest in the benchmarking context are *potential quantum speedup*, defined as a speedup compared to a specific classical algorithm or a set of classical algorithms (e.g., simulation of the time evolution of a quantum system implemented on a quantum computer as compared to directly solving the Schrödinger equation on a classical computer), a *limited quantum speedup*, defined as a speedup against algorithms that may be said to be analogous to the quantum solver (e.g., SA or SQA compared with a quantum annealer), and an unqualified *quantum speedup*, defined as a speedup against the best available algorithms for solving the problem (e.g., Shor's algorithm for factoring). A crucial observation made in [23] was that unless an optimal annealing time can be explicitly demonstrated (i.e., an observed minimum in the TTS as a function of the annealing time), a scaling analysis performed over a finite range of problem sizes cannot be trusted to reveal any type of quantum speedup. The reason is that an annealing time t_f that is too large (suboptimal) can artificially inflate the TTS at small problem sizes, thus leading to artificially shallow scaling, and potentially to a false conclusion that (some type of) quantum speedup is present. These notions were then applied to random Ising instances with a wide range of integer couplings (up to $|J_{ij}| = 7$, which is renormalized to $[-1, 1]$ when submitted to the processor), and tested on a DW2 device with up to 503 qubits. The analysis of the scaling of TTS with system size mostly demonstrated a disadvantage for the DW2 over SA, but an advantage for the DW2 over SA on lower hardness percentiles of the problem distribution (i.e., the easier problems). However, due to the aforementioned issue with suboptimal annealing times, this was not taken as evidence of any type of quantum speedup. Very recently evidence of a scaling advantage over SA with optimal annealing times was reported [74], as we discuss below.

An interesting critique of the scaling results presented in [23] was made in [75], which argued that random Ising instances restricted to the Chimera graph are 'too easy', essentially since their phase space exhibits only a zero-temperature transition. This would imply that classical thermal algorithms such as SA see a simple energy landscape with a single global minimum throughout the entire anneal (except perhaps at the very end as the simulation temperature is lowered to near zero), instead of the usual glassy landscape with many local traps associated with hard optimization problems. This work highlighted the importance of a careful design of benchmark problems, to ensure that classical solvers would not find them trivial. Of course, it should be stressed that quantum speedup is always relative, and it can be observed even when efficient (polynomial time) classical algorithms exist, as in, e.g., the solution of linear systems of equations [76]. In light of this one may interpret the message of [75] to mean that a quantum speedup might not be *detectable* over a finite range of problem sizes if the problem is classically easy, since the difference between the quantum and classical scaling is too subtle to be statistically significant.

An attempt to address the critique that random Ising instances are too easy was made in [77], which introduced a new class of 'planted solution' instances (see also the follow-up study [78], though neither study demonstrated a nonzero critical temperature). The problem Hamiltonian H_1 (equation (2)) is constructed out of a sum of frustrated small-loop Hamiltonians, each one designed so its ground state is a chosen bit-string. In so doing, the total Hamiltonian is guaranteed to have as one of its ground states the chosen bit-string, dubbed a 'planted solution'. Knowing a solution in advance is an important advantage of this problem class over the classes tested before, for which solutions could only be found either heuristically or by directly solving the Ising problem at a cost which is generally exponential with the system size (in particular, scaling like 2^{4L} for an $L \times L$ unit cell problem on the Chimera graph), which is prohibitive for systems much larger than those tested in previous studies. By knowing a solution in advance, the ground state energy can be computed instantly, and any further global optima can be recognized immediately, providing a sound basis for TTS comparisons for problems that may turn out to be too large for brute force search. This study was also one of the first (following [79]) to compare against the HFS algorithm [60, 61], which has been considered ever since to be the 'algorithm to beat' thanks to its superior scaling and direct exploitation of the large treewidth induced subgraphs possible in the Chimera graph. It was found in [77] that the DW2 had flatter scaling than all algorithms that had been tested up to that point (SA, SQA, SSV) over virtually the entire range of frustrated loop-to-spin density. In the comparison to the HFS algorithm it was found that the latter was able to achieve superior scaling compared to the DW2 for all but the easiest and largest loop densities. These results invited the possibility of a limited quantum speedup, but due to the lack of an optimal annealing time, this could again not be demonstrated. Moreover, [77] provided a proof (under the assumption that the TTS increases monotonically with the annealing time) that without an optimal annealing time, one could only demonstrate a slowdown, but never conclusively demonstrate even a limited quantum speedup.

Before we turn to a discussion of the evidence for a scaling advantage over SA, we first briefly discuss alternatives to the TTS as a performance measure. One such alternative is the TTT, i.e., the total time required by a solver to reach the target energy at least once with a desired probability, assuming each run takes a fixed time [52]. This reduces to the TTS if the target is the ground state. A unified approach that includes a variety of other measures was presented in [53], drawing upon optimal stopping theory, specifically the so-called 'house-selling' problem [80]. Within this framework one answers the question of how long, *given a particular cost for each sample drawn from a solver*, one should sample in order to maximize one's reward, analogously to the decision problem about when to sell one's house given that bids accrue over time but that waiting longer carries a higher monetary cost. This allows the TTS and TTT, among other measures, to be shown to be specific choices of the cost and reward functions. The optimal

stopping framework also paves the way for a more detailed comparison between quantum and classical approaches and the trade-offs of each, as by altering the cost per sample one can see the impact of the distribution over states (rather than just the ground state) for the various solvers. Optimal stopping is appropriate for applications where finding the minimum energy is not strictly the most important consideration for the application, such as many machine learning contexts and even various business-origin optimization problems. In those cases, there is a trade-off between the cost to perform the computation and the benefit from receiving a result. Tests were performed demonstrating the optimal stopping approach with a DW2X device (with 1098 qubits) on frustrated loop problems much like those in [77], demonstrating identical scaling (modulo concerns about the lack of an optimal annealing time) to the HFS algorithm at multiple values of the cost to draw a sample, an improvement over the DW2. However, these results could still not qualify as a limited quantum speedup due to the problem of suboptimal annealing times.

This problem was finally overcome in [74], which for the first time demonstrates an optimal annealing time, and can thus make positive claims about limited quantum speedup as originally defined in [23]; however, since it was unclear to what extent the origin of the speedup demonstrated in [74] was quantum, the authors preferred to use the more careful terminology of 'scaling advantage.' Previous studies could not find an optimal annealing time since a class of problem instances had not been identified for which the shortest available annealing time (20 μs in the DW2, 5 μs in the DW1 and DW2X, 1μs in the DW2000Q) was sufficiently short to observe on optimum given the largest problem size that could be tested. Using the D-Wave 2000Q (DW2KQ) device (with 2027 qubits) [74] demonstrated a simple one-unit cell gadget Hamiltonian which, when added randomly to a constant fraction of the unit cells on top of similar frustrated loop problems as in [81], resulted in the observation of an optimal annealing time for frustrated loops defined on the hardware graph (also when using the DW2X device), as well as for frustrated loops defined on the logical graph of unit cells (each unit cell then being bound together tightly as a pseudo-spin in the physical problem, modulo the gadget Hamiltonian). For the latter, logical-planted instances, the DW2KQ exhibited a statistically significant scaling advantage over single-spin-flip SA. These results amount to the first observation of a lscaling advantage of a quantum annealer over a general purpose classical optimation algorithm, since the existence of an optimal annealing time was certified. However, this did not amount to an unqualified quantum speedup since the DW2KQ's scaling was worse than SVMC, the HFS algorithm, unit cell cluster-flip SA, and SQA, which was found to have the best scaling among the algorithms tested. Nevertheless, this result paves the way towards future demonstrations of problems with optimal annealing times and hence certifiable scaling, a necessary requirement for any type of scaling speedup claim. However, even this may not be sufficient since other quantities remain that must eventually be optimized, such as the annealing schedule, which is known to play a crucial role in provable quantum speedups (specifically the Grover search problem [82, 83]), and can conversely be used to potentially overturn (limited) quantum speedup claims.

3.3. Lessons

What lessons may be gleaned from this story for future benchmarks of QA devices with limited or no error correction and hundreds or thousands of qubits?

1. It is vitally important to carefully account for resource use, lest one be led astray with a fake speedup. In particular, QA requires a demonstration of an optimal annealing time before any definitive conclusion can be drawn about a quantum speedup. More generally, optimizing all known free parameters is almost certainly necessary to demonstrate a quantum speedup which will hold up to scrutiny.
2. One must distinguish between different types of quantum speedup. Comparisons between a quantum computational device and a single other solver are inherently limited to a demonstration of a 'potential quantum speedup'. To go further, one must be sure to compare performance against a suite of algorithms, in particular those that mimic the device to some degree (such as SA or SQA). A speedup against such solvers would be considered a 'limited quantum speedup'. If there is a consensus about the solvers that are the best at the original task, then a speedup against such solvers would be considered an unqualified 'quantum speedup'. This would be a game-changing result.
3. Users of such quantum computational devices should perform something akin to gauge averaging in order to effectively estimate performance, by averaging over many different mappings from the logical problem to physical states, at least so long as the devices are not fully error corrected. Given that there is no good distribution for problem hardness as a function of this ensemble of mappings, nonparameteric techniques are appropriate[10].

[10] We suggest using a variant of the bootstrap, the Bayesian bootstrap, first introduced in [73], which can be shown to be the limit in the case of negligible prior information or large amounts of data of a Dirichlet process. Thus, it is arguably the only bootstrap procedure which is well-founded on Bayesian grounds. It involves reweighting the observed data, much like every bootstrap, but rather than sampling from a Multinomial(N, [1/N, …, 1/N]) distribution as in the frequentist bootstrap, one instead samples from the related Dirichlet(1, …,1) distribution. The primary advantage of the Bayesian bootstrap is that, unlike the frequentist bootstrap, the weight assigned any element in the dataset is always positive, i.e., there are no reweighted data vectors which leave out a data element, whereas the frequentist bootstrap has probability 1/e of dropping any given data element in a reweighted sample.

4. Choice of benchmark problem is key, and should be made with an eye toward the day when classical machines are vastly outpaced by quantum devices. For example, the transition from random Ising problems to frustrated loop/planted solutions problems was forced by the need to have reasonable benchmarks for devices so large that classical systems cannot solve them in a human lifetime.

So far we have only addressed the question of benchmarking without any attempt at error correction. Since it is impossible to achieve a scalable quantum speedup without some means of correcting for errors, while benchmarking native problems may give some indication of the abilities of QA by looking at relatively small problem sizes, work on error correction lays the foundation for potential lasting quantum advantages over classical computing.

4. Error correction

The discussion of error correction in gate-based quantum computing is usually dominated by questions of fault tolerance and the error thresholds on one and two-qubit gates necessary to satisfy the fault tolerance theorems [89–91]. The state for QA and AQC is quite different, as there is currently no known mechanism for achieving fault tolerance in such devices. Without the benefit of fault tolerance theorems, the best techniques we have available for managing errors in AQC and QA are energy gap protection [92], dynamical decoupling [93], and the Zeno effect [94], three intimately related techniques [95]. Work on error correction in physically realized quantum annealers has focused on energy gap protection, as techniques for dynamical decoupling are unavailable in current generations of annealers due to the associated high bandwidth requirements. Both of these techniques are really more error suppression than correction, as rather then correct errors they can lower the probability that errors will occur and mitigate their consequences should they happen.

The first demonstration of error suppression via energy energy gap protection in quantum annealers came with the introduction of the technique of quantum annealing correction (QAC) [84], which as the name suggests, also includes an active error correction component. In addition, [84] introduced a method for energy boosting by encoding the final Hamiltonian of a QA algorithm via multiple copies of the logical Hamiltonian operating on separate sets of physical qubits. This is in effect a simple classical repetition code. The copies are bound together by a penalty qubit whose action is to increase the energy of states of the physical system in which the copies are not in alignment with each other. The energy penalty for disagreement between the states effectively suppresses excitation to error states. Figure 5 shows the structure of the encoding and the nature of the encoded problem graph on a DW2 device. The logical Hamiltonian is boosted to have an effective strength three times that achievable in the hardware by directly programming the Hamiltonian. A restriction of this approach is that only the final Hamiltonian is encoded, not the driver Hamiltonian $-\sum\sigma_i^x$, which means that while the final Hamiltonian's gap is significantly larger than in the unencoded case, it is difficult to verify that the minimum ground to first excited state gap of the quantum Hamiltonian is also enhanced. However, a mean-field analysis shows that QAC softens the gap closure dependence on system size N, in the sense that for models exhibiting a first order quantum phase transition with the gap Δ scaling as C^N, the coefficient $0 < C \leqslant 1$ grows monotonically with the QAC penalty strength γ, and saturates at $C = 1$ for sufficiently large γ [96, 97]. The same work also showed that after QAC, the free energy barrier between the global minimum and the local minimum the system is initially trapped in is reduced in both width and height for a variety of transverse field Ising spin models, including models with disorder such as the Hopfield model.

The encoding restriction on QAC is not intrinsic to the technique, but rather is the result of the lack of any higher order (more than single-body) σ^x terms in the system Hamiltonian, which renders it impossible to form an effective logical σ^x term. Decoding a QAC encoded Hamiltonian is as simple as applying a majority vote over the problem qubits. The method was tested in [84] on antiferromagnetic chains of various lengths, demonstrating a significant improvement in success probability of finding the ground state of the chains compared not only to the unencoded case but also the case of a four copy repetition code (since QAC uses four times the hardware resources, one could simply run four copies of the problem at once and pick the lowest energy solution from any copy). As implemented in [84], the technique is not scalable (in the sense that both the energy boost and the gap against errors is constant), but it provided the first hope for systematically overcoming errors in the experimental QA. Another innovation was the use of many embeddings of the same logical Hamiltonian into compatible subgraphs, an idea which has found its way into the benchmarking context.

Chains have a trivial (classical) ground state, so a natural next test of QAC was to apply the method to random Ising problem instances [85]. This provided a demonstration that QAC could improve performance also on NP-hard problems defined on the QAC logical graph. Moreover, not only was the absolute performance on random Ising problems improved over both the unencoded and the classical repetition cases, but the scaling of the TTS for those problems improved under QAC, with the caveat that no optimal annealing time was

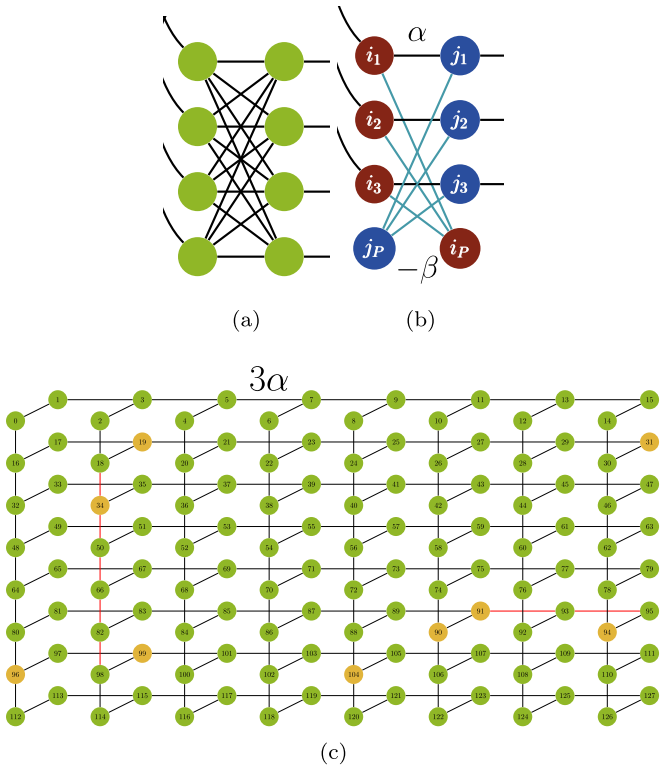
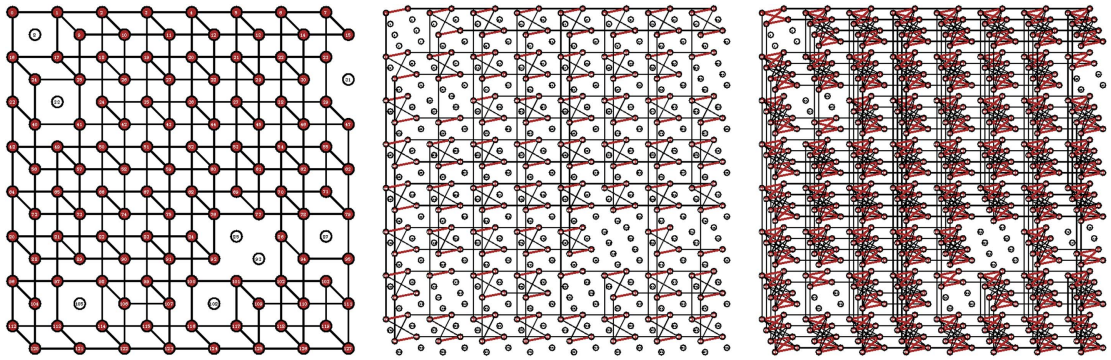


Figure 5. The QAC unit cell and encoded graph introduced in [84, 85]. (a) Schematic of one of the 64 unit cells of the DW2 processor. Unit cells are arranged in an 8 × 8 array forming a 'Chimera' graph between qubits. Each circle represents a physical qubit, and each line a programmable Ising coupling $\sigma_i^z \sigma_j^z$. Lines on the right (left) couple to the corresponding qubits in the neighboring unit cell to the right (above). (b) Two 'logical qubits' (*i*, red and *j*, blue) embedded within a single unit cell. Qubits labeled 1–3 are the 'problem qubits', the opposing qubit of the same color labeled *P* is the 'penalty qubit'. Problem qubits couple via the black lines with tunable strength α both inter- and intra-unit cell. Light blue lines of magnitude β are ferromagnetic couplings between the problem qubits and their penalty qubit. (c) Encoded processor graph obtained from the Chimera graph by replacing each logical qubit by a circle. This is a non-planar graph with couplings of strength 3α. Green circles represent complete logical qubits. Orange circles represent logical qubits lacking their penalty qubit. Red lines are groups of couplers that cannot all be simultaneously activated.

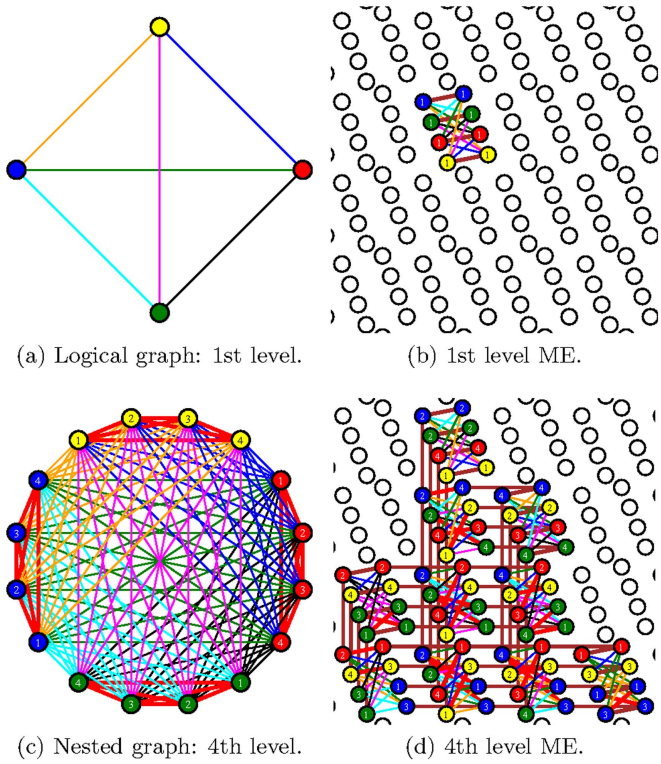
identified. In addition, QAC improved the robustness of the annealer to problem misspecification and increased the effective accuracy of the implemented problem Hamiltonian, as was shown by systematically reducing the energy scale of the final Hamiltonian. Since the hardware graph of the DW2 had missing qubits, encoding Ising problems required that some logical qubits went without a penalty qubit. Additional robustness was thus demonstrated by artificially increasing the number of penalty qubits lost: with up to 60% of the logical qubits going without penalty qubits, QAC continued to work with negligible performance loss.

The next step in testing QAC was to apply it to minor embedded problems [86], dubbed QAC-ME, thus going beyond natively embeddable problems such as chains and random Ising instances; see figure 6. A key innovation introduced in [86] is the introduction of non-uniform weights for both the QAC penalty terms as well as the strength of the chain in the ME, making them both proportional to the mean coupling strength in their respective logical Hamiltonians. This was in part informed by previous ME experiments such as [72], in which the optimal strength of the chain was found to be related to the emergence of the spin glass phase of the Hamiltonian. Since there is only a single strength σ^x term applied to every qubit while the strength of the σ^z terms depends on the choice of h_i and J_{ij}, one can easily find that with a uniform penalty strength some qubits will 'freeze' (i.e., no longer be effectively flipped by the driver Hamiltonian) before others, which can negatively impact solution quality. By locally fitting penalties to the strength of the logical problem Hamiltonian for each qubit, this process can be mitigated. Addressing decoding, [86] also proposed to use energy minimization, which involves directly minimizing the state of broken logical qubits (logical qubits whose physical qubits are not in alignment) given their neighboring qubits, and demonstrated that this can be done efficiently so long as the per-qubit probability of error is below the percolation threshold of the problem graph. And, going beyond the original QAC code of [84], a new, scalable QAC 'square code' whose logical graph forms a two-level-grid was proposed in [86] (see figure 6). The square code has the attractive feature that it can be concatenated. To benchmark QAC-ME, the same kind of frustrated loop problems with planted solutions that were first introduced in [77] were used. The results demonstrated a significant improvement in performance for non-uniform penalties over uniform penalty strengths, and that energy minimization was strongly preferable for



(a) 2LG logical graph connectivity (b) Minor embedding into Chimera graph (c) QAC-ME

Figure 6. Illustration of mapping from a logical problem to a QAC-ME embedding, introduced in [86]. (a) The two-level-grid (2LG) graph, for which the Ising spin glass problem with couplings in $\{-1, 0, 1\}$ is NP-hard [87]. Disconnected vertices indicate spins with couplings set to zero, as well as unused logical qubits in the logical DW2 Chimera graph. (b) Minor embedding of the 2LG graph into the physical DW2 Chimera graph. White circles correspond to unused or unusable qubits. (c) QAC-ME embedding of the 2LG problem using the 'square' code. In (b) and (c) penalties are represented by red (thick) couplings between groups of two (ME) and four (QAC-ME) physical qubits. Black (thin) links implement logical couplings.



(a) Logical graph: 1st level. (b) 1st level ME.

(c) Nested graph: 4th level. (d) 4th level ME.

Figure 7. Illustration of the nested QAC scheme introduced in [88]. In the left column, a C-level nested graph is constructed by embedding a K_N into a $K_{C\times N}$, with $N = 4$ and $C = 1$ (top) and $C = 4$ (bottom). Red, thick couplers are energy penalties defined on the nested graph between the (i, c) nested copies of each logical qubit i. The right column shows the nested graphs after minor embedding (ME) on the DW2X Chimera graph. Brown, thick couplers correspond to the ferromagnetic chains introduced in the process.

decoding QAC-ME compared to standard majority vote decoding. The square code was compared with the original QAC code on chains in [98].

Both the original QAC scheme and QAC-ME induce a graph of lower degree than that of the initial Hamiltonian. To overcome this and deal from the start with arbitrary Ising model Hamiltonians, a 'nested QAC' (NQAC) method was introduced in [88]. NQAC starts from a fully connected K_N graph for the underlying problem and then maps this N qubit problem into C coupled copies of itself in a larger $K_{C\times N}$ graph. When run on a hardware graph of lower degree, this larger $K_{C\times N}$ graph is then minor-embedded, with the coupled copies doing the work of suppressing errors and helping to limit the formation of domain walls in the ME chains. In this way, the number of physical qubits required to implement level C NQAC is approximately $C^2N^2/4$. An

illustration of the embedding of the K_N Hamiltonian into the larger $K_{C\times N}$ Hamiltonian as well as a sample of the minor embedded graph are given in figure 7.

The key finding of [88] is that NQAC effectively rescales the temperature of the system down by a factor of μ_C for C nesting levels, with the theoretical expectation for a fully thermalized state being that $\mu_C \propto C^2$, based on a mean-field analysis. In practice, the scaling is not quite that fast, instead $\mu_C \approx C^{1.4}$ once the energy penalty tying the C copies of the problem Hamiltonian together and the chain strength are optimized. This result is important since it means that one can trade qubits for an effective temperature reduction, that is controllable via the nesting level C. This suggests, at least in principle, that the effective temperature can be kept below the gap. For the DW2 device, for $C \geqslant 3$ NQAC was no longer able to improve the success probability more than classical repetition of the $C = 1$ case, though this is probably because the base problem (a random antiferromagnetic Ising K_8 with $J_{ij} \in \{0.1, 0.2, \ldots, 1\}$) was too easy. The NQAC method was shown to continue to scale favorably on larger problems embedded on the DW2KQ processor [99]. Tests on future generations of more advanced processors will reveal whether techniques like NQAC hold at least part of the key to scalable quantum error suppression and correction on QA platforms, even if it and similar pure error suppression techniques will never be able to achieve true fault tolerance.

The story of error suppression and correction in experimental QA algorithms is one of a sequence of developments, building off both the results and lessons of native benchmarking work and the initial insight of energy gap protection as a fruitful and feasible path on current systems for error suppression, and generalizing to more and more useful encoded graphs, reaching NQAC with its fully general encoded Ising Hamiltonian and potential for arbitrarily large and strong error suppression. Future paths for investigation of experimental QA correction center on expanded benchmarking of NQAC for larger systems and for application-domain problems, as well as continued theoretical development of error suppression and correction techniques. For example, using subsystem codes it is possible to construct error suppression schemes appropriate for AQC that use only two-body interactions for certain problems such as the transverse field Ising model on a ring, so that by adding $\sigma_i^x \sigma_j^x$ terms to the driver Hamiltonian one could significantly improve over the current state of the art of QAC [100, 101].

5. Conclusion

As the field of quantum computing, and QA in particular, expands rapidly and the number of available platforms rises, methods to validate the fidelity of the platform to its stated physical model, verifying entanglement and tunneling, strong benchmarking methods, and error correction/suppression techniques will be vital in discerning which platforms truly can offer advantages over classical computation. Several years have been spent developing methods to meet each of these challenges, particularly targeting existing quantum annealers, but many of these methods can be readily adapted to other systems that implement programmable Ising model Hamiltonians [102]. Insights from each of these areas are informing development in the others. For example, insight into and experience with small gadgets from quantum validation informed recent work demonstrating scaling advantage over SA on D-Wave quantum annealers using a small gadget to generate an optimal annealing time on existing machines, while insights from benchmarking regularly inform developments in error suppression and correction. Work on benchmarking has provided guidelines and methods of analysis which can be used by anyone seeking to characterize the performance of a putative quantum computing device, while error suppression work has laid the foundation for more extensive experiments and solving more difficult problems on future, larger QA devices. Thus, one answer to the question of what one may want to use a 1000-qubit quantum computer for, is in our view the type of bootstrapping we have reviewed in this article, where a productive interplay among quantum validation testing, benchmarking, and error correction has led to a sequence of advances that will inform even larger quantum computation experiments, until one day a test drive with a brand new quantum computer will take us to the ultimate destination of undisputed quantum supremacy and unqualifed quantum speedup.

Acknowledgments

We are grateful to the many colleagues with whom we have collaborated over the past several years of test-driving quantum annealers, and in particular to Tameem Albash for helpful discussions and comments on this article, and to Siddharth Muthu Krishnan for prompting us to write this review. This work was supported under ARO grant number W911NF-12-1-0523, ARO MURI grant No. W911NF-15-1-0582, and NSF grant number INSPIRE-1551064. This material is based upon work supported by the Intelligence Advanced Research Projects Activity (IARPA) through the Army Research Office (ARO) under Contract No. W911NF-17-C-0050. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Intelligence Advanced Research Projects Activity (IARPA) and the Army Research Office (ARO).

References

[1] Reichardt B W, Unger F and Vazirani U 2013 *Nature* **496** 456
[2] Chuang I L and Nielsen M A 1997 *J. Mod. Opt.* **44** 2455
[3] Mohseni M, Rezakhani A T and Lidar D A 2008 *Phys. Rev.* A **77** 032322
[4] Blume-Kohout R, Gamble J K, Nielsen E, Mizrahi J, Sterk J D and Maunz P 2013 arXiv:1310.4492
[5] Greenbaum D 2015 arXiv:1509.02921
[6] Childs A M, Chuang I L and Leung D W 2001 *Phys. Rev.* A **64** 012314
[7] Blume-Kohout R, Gamble J K, Nielsen E, Rudinger K, Mizrahi J, Fortier K and Maunz P 2017 *Nat. Commun.* **8** 14485
[8] Kadowaki T and Nishimori H 1998 *Phys. Rev.* E **58** 5355
[9] Das A and Chakrabarti B K 2008 *Rev. Mod. Phys.* **80** 1061
[10] Farhi E, Goldstone J, Gutmann S, Lapan J, Lundgren A and Preda D 2001 *Science* **292** 472
[11] Kaminsky W M and Lloyd S 2004 *Quantum Computing and Quantum Bits in Mesoscopic Systems* ed A Leggett *et al* (New York: Springer Science & Business Media)
[12] Kaminsky W M, Lloyd S and Orlando T P 2004 arXiv:quant-ph/0403090
[13] Albash T and Lidar D A 2018 *Rev. Mod. Phys.* **90** 015002
[14] Berkley A J *et al* 2013 *Phys. Rev.* B **87** 020502
[15] Preskill J 2012 arXiv:1203.5813
[16] Flammia S 2017 Quantum Advantage http://dabacon.org/pontiff/?p=11863
[17] Aaronson S and Chen L 2016 arXiv:1612.05903
[18] Bremner M J, Montanaro A and Shepherd D J 2016 *Phys. Rev. Lett.* **117** 080501
[19] Farhi E and Harrow A W 2016 arXiv:1602.07674
[20] Gao X, Wang S-T and Duan L M 2017 *Phys. Rev. Lett.* **118** 040502
[21] Boixo S, Isakov S V, Smelyanskiy V N, Babbush R, Ding N, Jiang Z, Martinis J M and Neven H 2018 *Nat. Phys.* 10.1038/s41567-018-0124-x
[22] Fefferman B, Foss-Feig M and Gorshkov A V 2017 *Phys. Rev.* A **96** 032324
[23] Rønnow T F, Wang Z, Job J, Boixo S, Isakov S V, Wecker D, Martinis J M, Lidar D A and Troyer M 2014 *Science* **345** 420
[24] Shor P W 1997 *SIAM J. Comput.* **26** 1484
[25] Farhi E, Goldstone J and Gutmann S 2014 arXiv:1412.6062
[26] Yip K-W, Albash T and Lidar D 2017 *Phys. Rev.* A **97** 022116
[27] Lanting T *et al* 2014 *Phys. Rev.* X **4** 021041
[28] Vidal G and Werner R F 2002 *Phys. Rev.* A **65** 032314
[29] Spedalieri F M 2012 *Phys. Rev.* A **86** 062311
[30] Albash T, Hen I, Spedalieri F M and Lidar D A 2015 *Phys. Rev.* A **92** 062328
[31] Albash T, Boixo S, Lidar D A and Zanardi P 2012 *New J. Phys.* **14** 123016
[32] Boixo S, Albash T, Spedalieri F M, Chancellor N and Lidar D A 2013 *Nat. Commun.* **4** 2067
[33] Smolin J A and Smith G 2014 *Frontiers Phys.* **2** 52
[34] Wang L, Rønnow T F, Boixo S, Isakov S V, Wang Z, Wecker D, Lidar D A, Martinis J M and Troyer M 2013 arXiv:1305.5837
[35] Albash T, Vinci W, Mishra A, Warburton P A and Lidar D A 2015 *Phys. Rev.* A **91** 042314
[36] Shin S W, Smith G, Smolin J A and Vazirani U 2014 arXiv:1404.6499
[37] Kirkpatrick S, Gelatt C D and Vecchi M P 1983 *Science* **220** 671
[38] Boixo S, Ronnow T F, Isakov S V, Wang Z, Wecker D, Lidar D A, Martinis J M and Troyer M 2014 *Nat. Phys.* **10** 218
[39] Martoňák R, Santoro G E and Tosatti E 2002 *Phys. Rev.* B **66** 094203
[40] Shin S W, Smith G, Smolin J A and Vazirani U 2014 arXiv:1401.7087
[41] Crowley P J D and Green A G 2016 *Phys. Rev.* A **94** 062106
[42] Albash T, Rønnow T F, Troyer M and Lidar D A 2015 *Eur. Phys. J. Spec. Top.* **224** 111
[43] Brooke J, Bitko D, Rosenbaum T F and Aeppli G 1999 *Science* **284** 779
[44] Brooke J, Rosenbaum T F and Aeppli G 2001 *Nature* **413** 610
[45] Johnson M W *et al* 2011 *Nature* **473** 194
[46] Boixo S, Smelyanskiy V N, Shabani A, Isakov S V, Dykman M, Denchev V S, Amin M H, Smirnov A Y, Mohseni M and Neven H 2016 *Nat. Commun.* **7** 10327
[47] Denchev V S, Boixo S, Isakov S V, Ding N, Babbush R, Smelyanskiy V, Martinis J and Neven H 2016 *Phys. Rev.* X **6** 031015
[48] Mandrà S, Zhu Z, Wang W, Perdomo-Ortiz A and Katzgraber H G 2016 *Phys. Rev.* A **94** 022337
[49] Monz T, Schindler P, Barreiro J T, Chwalla M, Nigg D, Coish W A, Harlander M, Hänsel W, Hennrich M and Blatt R 2011 *Phys. Rev. Lett.* **106** 130506
[50] Bohnet J G, Sawyer B C, Britton J W, Wall M L, Rey A M, Foss-Feig M and Bollinger J J 2016 *Science* **352** 1297
[51] McGeoch C C 2012 *A Guide to Experimental Algorithmics* (Cambridge: Cambridge University Press)
[52] King J, Yarkoni S, Nevisi M M, Hilton J P and McGeoch C C 2015 arXiv:1508.05087
[53] Vinci W and Lidar D A 2016 *Phys. Rev. Appl.* **6** 054016
[54] McGeoch C C and Wang C 2013 *Proc. 2013 ACM Conf. on Computing Frontiers*
[55] Santra S, Quiroz G, Steeg G V and Lidar D A 2014 *New J. Phys.* **16** 045006
[56] Isakov S V, Zintchenko I N, Rønnow T F and Troyer M 2015 *Comput. Phys. Commun.* **192** 265
[57] Geyer C J 1991 *Computing Science and Statistics Proc. 23rd Symp. on the Interface* ed E M Keramidas (New York: American Statistical Association) p 156
[58] Earl D J and Deem M W 2005 *Phys. Chem. Chem. Phys.* **7** 3910
[59] Katzgraber H G, Trebst S, Huse D A and Troyer M 2006 *J. Stat. Mech.* P03018
[60] Hamze F and de Freitas N 2004 *UAI* ed D M Chickering and J Y Halpern (Arlington, Virginia: AUAI Press) pp 243–50
[61] Selby A 2014 arXiv:1409.3934
[62] Heim B, Rønnow T F, Isakov S V and Troyer M 2015 *Science* **348** 215
[63] Crosson E and Harrow A W 2016 *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)* (Piscataway, NJ: IEEE) 714–23
[64] Zick K M, Shehab O and French M 2015 *Sci. Rep.* **5** 11168
[65] Venturelli D, Marchand D J J and Rojo G 2015 arXiv:1506.08479

[66] Rieffel E G, Venturelli D, O'Gorman B, Do M B, Prystay E M and Smelyanskiy V N 2015 *Quantum Inf. Process.* **14** 1
[67] Rosenberg G, Haghnegahdar P, Goddard P, Carr P, Wu K and Prado M L D 2016 *IEEE J. Sel. Top. Signal Processing.* **10** 1053–60
[68] Lucas A 2014 *Front. Phys.* **2** 5
[69] Choi V 2008 *Quantum Inf. Process.* **7** 193
[70] Choi V 2011 *Quantum Inf. Process.* **10** 343
[71] Klymko C, Sullivan B D and Humble T S 2014 *Quantum Inf. Process.* **13** 709
[72] Venturelli D, Mandrà S, Knysh S, O'Gorman B, Biswas R and Smelyanskiy V 2015 *Phys. Rev.* X **5** 031040
[73] Rubin D B *et al* 1981 *Ann. Stat.* **9** 130
[74] Albash T and Lidar D A 2017 arXiv:1705.07452
[75] Katzgraber H G, Hamze F and Andrist R S 2014 *Phys. Rev.* X **4** 021008
[76] Harrow A W, Hassidim A and Lloyd S 2009 *Phys. Rev. Lett.* **103** 150502
[77] Hen I, Job J, Albash T, Rønnow T F, Troyer M and Lidar D A 2015 *Phys. Rev.* A **92** 042325
[78] King A D, Lanting T and Harris R 2015 arXiv:1502.02098
[79] Selby A 2013 D-wave: Comment on comparison with classical computers www.archduke.org/stuff/d-wavecomment-on-comparison-with-classical-computers/
[80] Ferguson T S 2008 *Optimal Stopping and Applications* (Los Angeles, CA: UCLA) www.math.ucla.edu/~tom/Stopping/Contents.html
[81] King J, Yarkoni S, Raymond J, Ozfidan I, King A D, Nevisi M M, Hilton J P and McGeoch C C 2017 arXiv:1701.04579
[82] Roland J and Cerf N J 2002 *Phys. Rev.* A **65** 042308
[83] Rezakhani A T, Pimachev A K and Lidar D A 2010 *Phys. Rev.* A **82** 052305
[84] Pudenz K L, Albash T and Lidar D A 2014 *Nat. Commun.* **5** 3243
[85] Pudenz K L, Albash T and Lidar D A 2015 *Phys. Rev.* A **91** 042302
[86] Vinci W, Albash T, Paz-Silva G, Hen I and Lidar D A 2015 *Phys. Rev.* A **92** 042310
[87] Barahona F 1982 *J. Phys. A: Math. Gen.* **15** 3241
[88] Vinci W, Albash T and Lidar D A 2016 *Nat. Quantum Inf.* **2** 16017
[89] Aliferis P, Gottesman D and Preskill J 2006 *Quantum Inf. Comput.* **6** 97
[90] Raussendorf R 2012 *Phil. Trans. R. Soc.* A **370** 4541
[91] Lidar D and Brun T (ed) 2013 *Quantum Error Correction* (Cambridge: Cambridge University Press)
[92] Jordan S P, Farhi E and Shor P W 2006 *Phys. Rev.* A **74** 052322
[93] Lidar D A 2008 *Phys. Rev. Lett.* **100** 160506
[94] Paz-Silva G A, Rezakhani A T, Dominy J M and Lidar D A 2012 *Phys. Rev. Lett.* **108** 080501
[95] Young K C, Sarovar M and Blume-Kohout R 2013 *Phys. Rev.* X **3** 041013
[96] Matsuura S, Nishimori H, Albash T and Lidar D A 2016 *Phys. Rev. Lett.* **116** 220501
[97] Matsuura S, Nishimori H, Vinci W, Albash T and Lidar D A 2017 *Phys. Rev.* A **95** 022308
[98] Mishra A, Albash T and Lidar D A 2015 *Quantum Inf. Process.* **15** 609
[99] Vinci W and Lidar D A 2018 *Phys. Rev.* A **97** 022308
[100] Marvian M and Lidar D A 2017 *Phys. Rev. Lett.* **118** 030504
[101] Jiang Z and Rieffel E G 2017 *Quantum Inf. Process.* **16** 89
[102] Inagaki T *et al* 2016 *Science* **354** 603