Quantum Annealing: Theory, Applications & Latest Advances

Quantum Computing – SciFi or the Future of Finance?

Nicholas Chancellor

September, 2022

Joint Quantum Centre

Durham University

QCi Innovative Quantum Solutions Company

What this talk is about (+ collaborator acknowledgments)

Some background as well as some perspective on the theory

1. Background: quantum computing, and where quantum annealing fits within it
2. Moving theory beyond a limit which is often impractically slow for real devices (the adiabatic limit)
   - Work with Adam Callison, Max Festenstein, Jie Chen, Laurentiu Nita, and Viv Kendon

     Durham University
3. How does the encoding of optimisation problems affect dynamics?
   - Work with Jesse Berwald and Raouf Dridi

     QCi Innovative Quantum Solutions Company

Quantum computing

Big idea: harness the fundamental physics of discrete systems (quantum mechanics) to solve important problems

- We know it works in theory: quantum search of unstructured database with N entries in a time proportional to $\sqrt{N}$
- This is not possible without using quantum mechanics (only option without QM is random guess or exhaustive search)

(011) (010)

(111)

(000)

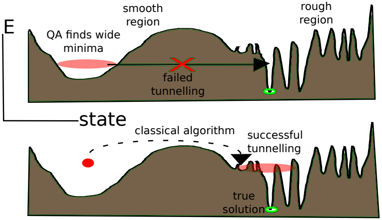
(001) (110)

...but how do we use real, imperfect, quantum machines to solve problems people care about?

Applied Quantum computing

How do we use real, imperfect, quantum machines to solve problems people care about?

1. Only use them for what they are good at do the rest classically hybrid quantum/classical algorithms*, build theory around what can be done physically, not the other way around



2. Find the right problems $\rightarrow$ need to be the right shape and size for near term the machines... and still be problems people care about important! but not the topic of this talk

But first... some background on continuous time QC and quantum annealing

*see: Callison and Chancellor Phys. Rev. A 106, 010101

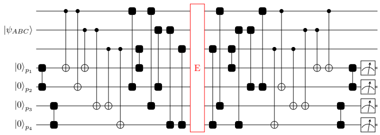
Two different approaches to quantum computing

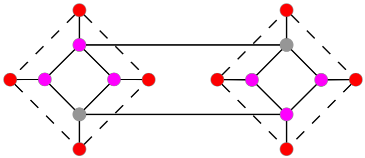
'Gate' based quantum computing

- Discrete quantum operations on qubits
- Construct 'circuits' out of these gates
- Detect and correct errors to reduce effect of noise

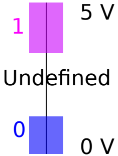
Continuous time

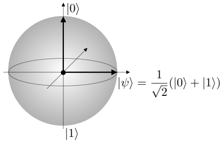
- Map problems directly to physical system
- Allow quantum physics to help search solution space
- Low temperature environment could help solve problems





Why continuous time?





Classical bits: fundamentally discrete $\rightarrow$ 0 or 1, nothing in between

Lends itself to a discrete *digital* description: bit flips either happen or they don't

Quantum bits: continuous rotations are possible

Breaking operations up into discrete chunks is not natural $\rightarrow$ an (exact) bit flip is just as hard as any other rotation

Bonus feature: applied gate based algorithms similar to continuous time operations $\rightarrow$ cont. time algorithms have implications for gate based

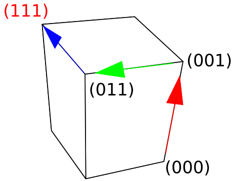
Getting physics to solve hard problems → transverse field Ising model

Physics Language, Hamiltonian:

$$H = -A(t) \sum_{i}^{n} X_i + B(t) \left(\sum_{i}^{n} h_i Z_i + \sum_{i,j}^{n} J_{ij} Z_i Z_j \right)$$

What this means in non-physics language:

$\sum_i^n X_i \rightarrow$ Bit flips, hops state through n dimensional hypercube



$\sum_i^n h_i Z_i + \sum_{i,j}^n J_{ij} Z_i Z_j \rightarrow$ Ising spin glass, defines interesting problem to be solved (as bitstring energies) more on next slides

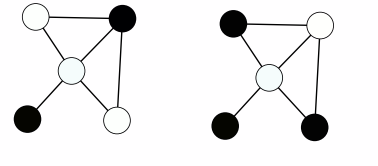
Example of Ising problem mapping *

Have:

- Binary variables $Z_i \in \{-1, 1\}$
- Minimisation over Hamiltonian made of single and pairwise terms $\mathrm{H}_{\mathrm{Ising}} = \sum_i h_i Z_i + \sum_{j>i} J_{i,j} Z_i Z_j$

Want:

- Maximum independent set: how many vertexes on a graph can we colour so none touch? $\rightarrow$ NP hard



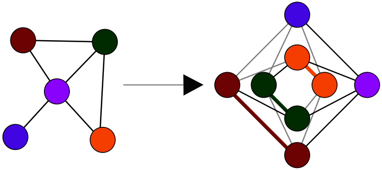
Method:

1. For an edge between vertex i and j add $Z_i + Z_j + Z_i Z_j \rightarrow$ penalizes colouring ($Z = 1$) adacent vertexes
2. Add $-\lambda Z_i$ to reward coloured vertexes $(0 < \lambda < 1)$

*Taken from the notes of a physics level 3 computing project I wrote, full notes at: http://nicholas-chancellor.me/QOpt_project_final.pdf

Minor embedding

- Strong ‘ferromagnetic’ $(-Z_i Z_j)$ coupling energetically penalizes variables disagreeing
- If strong enough than entire ‘chain’ acts as a single variable
- Mathematically corresponds to mapping one graph to graph minors of another

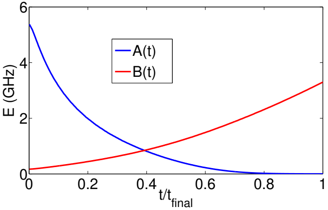


Can embed arbitrary graphs into quasi-planar hardware graph with polynomial (n^2 for fully connected) overhead $\rightarrow$ Ising model restricted to hardware graph is also NP-hard

In practice this leads to a large overhead $\rightarrow$ important consideration for solving real problems

Actually solving problems

Quantum Hamiltonians generalize classical Monte Carlo algorithms e.g. simulated annealing



$$H = -A(t) \sum_{i}^{n} X_i + B(t) \left(\sum_{i}^{n} h_i Z_i + \sum_{i,j}^{n} J_{ij} Z_i Z_j \right)$$

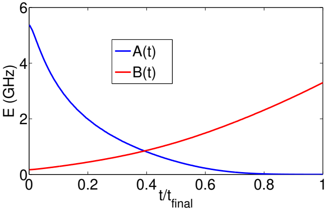
- Parameter sweeps (a.k.a. annealing) can be used to solve problems
- Low temperature dissipation can help too

This algorithm is called quantum annealing

Adiabatic quantum computing (a theorist's version of quantum annealing)

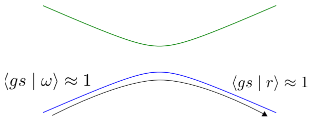
Traditional picture:

- Map an NP-hard optimization problem to a Hamiltonian, unknown ground state is solution
- *Slowly* change from a (driver) Hamiltonian with an easily prepared ground state to problem Hamiltonian
- Adiabatic theorem of quantum mechanics → success probability arbitrarily close to 100 % by running long enough



$$H(t) = A(t)H_{\mathrm{driver}} + B(t)H_{\mathrm{problem}}$$

Advantages and disadvantages of this picture



Theoretically satisfying

- Algorithm is effectively deterministic $\rightarrow$ "always" succeeds
- Intuitive picture involving only ground and first excited state

Let's assume P$\neq$NP *

- Algorithm succeeds roughly 100% of the time
- Total runtime needs to be exponential in size of problem $\rightarrow$ system needs to remain coherent for exponentially long time*

*For those unfamiliar with complexity theory, this is basically saying "let's assume that hard optimization problems exist", most experts believe P$\neq$NP

*For more sophisticated adiabatic theorem to faster quenches see: Crosson and Lidar, Nature Reviews Physics volume 3, pages 466-489 (2021)

What can be done?

Restore coherence somehow

- Error correction, difficult to do in continuous time, but progress being made
- Low temperature dissipation can restore coherence $\rightarrow$ would have to be very low temperature

- Have to mitigate *enough* errors for a *very* long time
- Not the subject of this talk

Succeed with low probability

- *Total* runtime is still exponential in problem size
- Each run is short $\rightarrow$ exponentially many needed to hit right answer

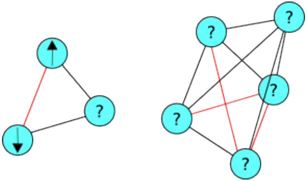
- Exponentially low success each run is *conceptually* unsatisfying...
- ... but much less demanding for coherence

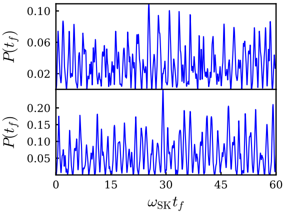
Lottery

image public domain from wikimedia commons

A simpler algorithm: continuous time quantum walk on spin glass

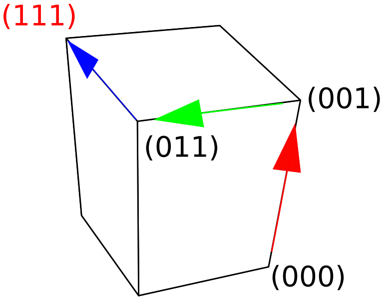
- Start with an equal positive superposition of all solutions, $|\omega\rangle = \frac{1}{\sqrt{N}} \sum_i |i\rangle$
  - Evolve with a fixed Hamiltonian $H_{\mathrm{walk}} = \gamma H_{\mathrm{hop}} + H_{\mathrm{problem}}$
- $H_{\mathrm{hop}} = -\sum_i X_i \to$ superposition is ground state
- $H_{\mathrm{problem}} = \sum_{i,j} J_{i,j} Z_i Z_j + \sum_i h_i Z_i$ where h_i and $J_{i,j}$ drawn from the same Gaussian distribution
- Measure after random short period of time, repeat many times

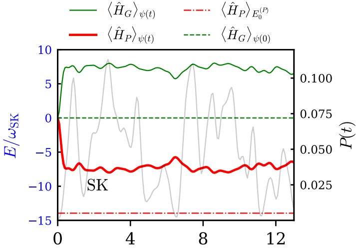




See Adam Callison et al 2019 New J. Phys. 21 123022 for details, work with Adam Callison, Viv Kendon, and Florian Mintert

How is this a ‘walk’? How does it find solutions?





- H_{hop} effectively forms a hypercube with a bitstring at each vertex, probability amplitude ‘walks’ between different states
- H_{problem} contributes phases which guide the walk

Energy is conserved $\langle H_{\text{walk}} \rangle_{t=0} = \langle H_{\text{walk}} \rangle_{t>0}$ since the system starts in the ground state of H_{hop}:
$\langle H_{\text{problem}} \rangle_{t>0} - \langle H_{\text{problem}} \rangle_{t=0} = \langle H_{\text{hop}} \rangle_{t=0} - \langle H_{\text{hop}} \rangle_{t>0} \leq 0$

Walk seeks out ‘good’ solutions!

Rapid quenches?

The energy conservation argument given previously can be extended to any monotonic (closed system) quench

$$H(t) = A(t)\,H_{\mathrm{drive}} + B(t)\,H_{\mathrm{problem}} \quad \frac{A(t)}{B(t)} \geq \frac{A(t+\delta t)}{B(t+\delta t)} \forall_t$$

Sketch of proof:

1. Trotterize time evolution: $A(t) \rightarrow A(t+\delta t)$ and $B(t) \rightarrow B(t+\delta t)$ and apply $|\psi(t+\delta t)\rangle = \exp(-iH(t)\delta t)|\psi(t)\rangle$ in separate steps
2. Rescale time so that Hamiltonian always resembles quantum walk $H_{eff}(\Gamma(t)) = \Gamma(t)\,H_{\mathrm{drive}} + H_{\mathrm{problem}}$
3. In rescaled version $\Gamma(t) \geq \Gamma(t+\delta t)$ $\therefore$ $\langle H_{eff}(\gamma(t))\rangle_{\psi(t)} - \gamma(t)\,n \geq \langle H_{eff}(\Gamma(t+\delta t))\rangle_{\psi(t)} - \Gamma(t+\delta t)\,n$
4. Because $\langle H_{eff}(\Gamma(t))\rangle_{\psi(t)} \geq -\Gamma(t)\,n\ \forall_t$, $\langle H_{\mathrm{problem}}\rangle_{\psi(t)} \leq 0\ \forall_t$

Details can be found in Callison et. al. PRX Quantum 2, 010338

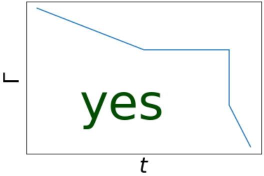
A very general result!

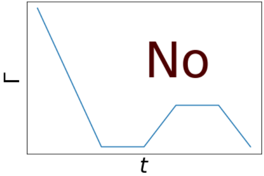
What is needed for result to hold:

1. Monotonic $\Gamma(t) \geq \Gamma(t + \delta t)$ where $\Gamma(t) = \frac{A(t)}{B(t)}$
2. Start in ground state of H_{drive}
3. Driver not gapless $\rightarrow$ not a concern for real problems

What is allowed:

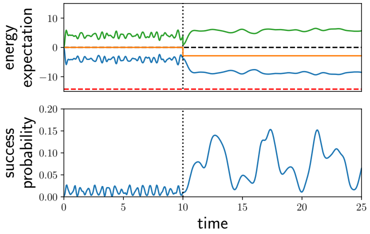
1. No limit on how fast algorithm runs
2. Discontinuities in $\Gamma(t)$ are ok
3. H_{drive} does not need to be diagonal in an orthogonal basis to H_{problem} $\rightarrow$ starting state can be biased

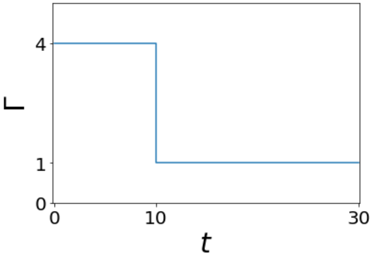




Intuitive example: two stage quantum walk

Perform a quantum walk at γ_1, and than use result as an input state for a second walk at $\gamma_2 < \gamma_1$





- Energy expectations: Green$= \gamma_{1,2}\langle H_{\mathrm{drive}}\rangle$; Blue$= \langle H_{\mathrm{problem}}\rangle$; Gold$= \gamma_{1,2}\langle H_d\rangle + \langle H_{\mathrm{problem}}\rangle$
- Total energy conserved except for at dashed line where γ decreases
- Non-instantaneous quench effectively infinite stage quantum walk

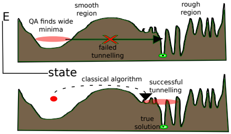
Why is the rapid quench result important?

General, but rather weak:

Any monotonic quench at least as good as measuring the initial state

1. Design protocols to maximize dynamics $\rightarrow$ don't need to worry about dynamics being counter-productive not space to discuss here, but this allows us to design better protocols
2. A **biased** search can already start from a very good guess next slide
3. Mechanism to understand dynamics very far from adiabatic limit

Hybrid protocols using this mechanism?



Known techniques*:

Dissipative reverse annealing NC 2017 New J. Phys. 19 023024 as implemented on D-Wave devices

Relies on dissipation, not suitable for coherent algorithms

Coherent reverse annealing Perdomo-Ortiz et. al. Quantum Inf Process (2011) 10: 33. doi:10.1007/s11128-010-0168-z

Involves three separate Hamiltonians, not compatible with rapid sweep proof in Callison et. al. PRX Quantum 2, 010338

Biased driver Hamiltonian Chinese Physics Letters, 30 **1** 010302 and Tobias Graß Phys. Rev. Lett. 123, 120501 (2019)

Compatible with proof in Callison et. al. PRX Quantum 2, 010338, can apply the mechanisms discussed here

*for details of each, see Callison and Chancellor Phys. Rev. A 106, 010101

Effect of problem structure and encoding*

Consider higher-than-binary discrete problems; appear often in real world optimisation, for example:

- A truck can go down any of three roads...
- A tasks can be scheduled at any of five times...
- A component can be placed any of seven places on a chip...

- Define two index objects:

$$x_{i,\alpha} = \begin{cases} 1 & \text{variable } i \text{ takes value } \alpha \\ 0 & \text{otherwise} \end{cases}$$

- Discrete Quadratic models, (DQM), made from pairwise interactions of x terms:

$$H_{\mathrm{DQM}} = \sum_{i,j} \sum_{\alpha,\beta} D_{(i,j,\alpha,\beta)} x_{i,\alpha} x_{j,\beta}$$

*Details in arχiv:2108.12004, accepted in Royal Society Philosophical Transactions A

Discrete variables into binary, three ways

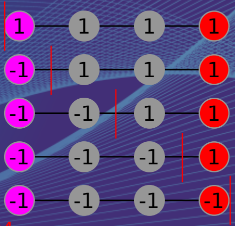
Innovative Quantum Solutions Company

Variable size$=m$

performance metric	binary	one-hot	domain wall*
# binary variables	$\lceil \log_2(m) \rceil$	m	$m-1$
# couplers for encoding	0 if $m=2^n$ $n \in \mathbb{Z}$ complicated otherwise	$m(m-1)$	$m-2$
intra-variable connectivity	N/A or complicated	complete	linear
maximum order needed for two variable interactions	$2\lceil \log_2(m) \rceil$	2	2

Binary= assign bitstrings to configurations
One hot= constrain variables so exactly one can be 1
Domain wall= new encoding w/ better performance: Chen et. al. IEEE Transactions on Quantum Engineering 3102714 (2021)

encoded value	qubit configuration
0	1111
1	-1111
2	-1-111
3	-1-1-11
4	-1-1-1-1

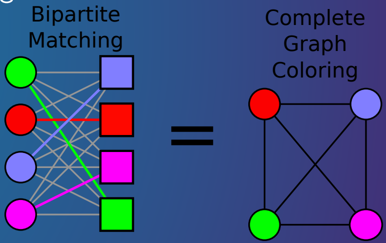


*Chancellor, Quantum Sci. Technol. 4 045004

Quadratic Assignment Problem (QAP)

Innovative Quantum Solutions Company

Assign m facilities to m locations such that a single facility is only assigned to one location and vice-versa

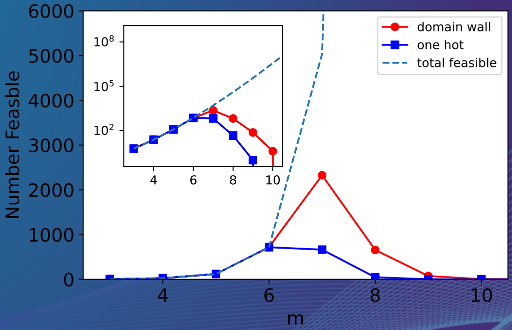


- General (hard) version → pairs of assignments are weighted, we use unweighted → **not** hard, but symmetry and large degeneracy useful for analysis
- Can be thought of as a colouring problem on an m-node fully connected graph
- $m!$-fold degenerate ground state

Experimental tests

Innovative Quantum Solutions Company

Run on D-Wave *Advantage* annealer 10 embeddings at each size with $10,000$ reads for total of $100,000$ reads at each size (default settings otherwise)

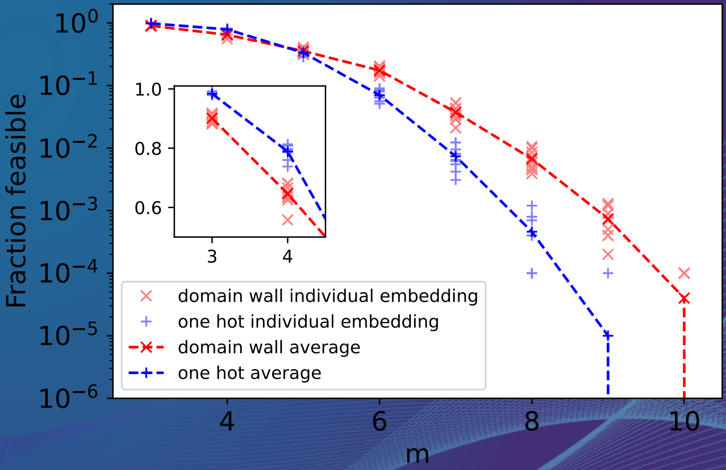


Able to find all feasibles up until about size 6, then both struggle, but domain-wall encoding performs much better.

Rate of feasible solutions

Innovative Quantum Solutions Company

Stars represent fractions of returned solutions which are feasible

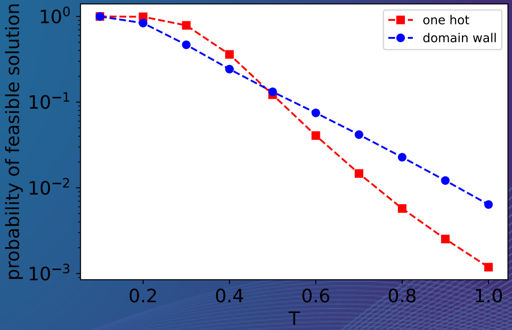


At largest size ($n = 10$) domain-wall encoding finds solutions while one-hot finds none.

One explanation: thermal excitations

Innovative Quantum Solutions Company

Symmetry of problem means Metropolis algorithm converges quickly, efficient thermal sampling

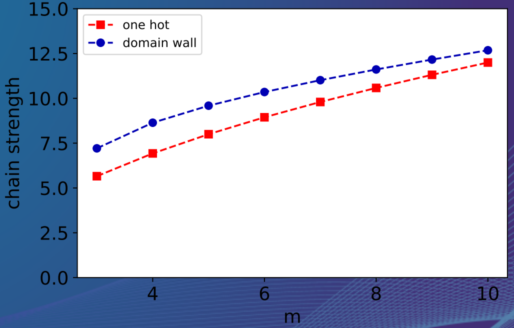


Probability of feasible solution is better at higher temperature with domain-wall encoding, makes sense one fewer qubit $\rightarrow$ smaller solution space

Dynamic range squeezing

Innovative Quantum Solutions Company

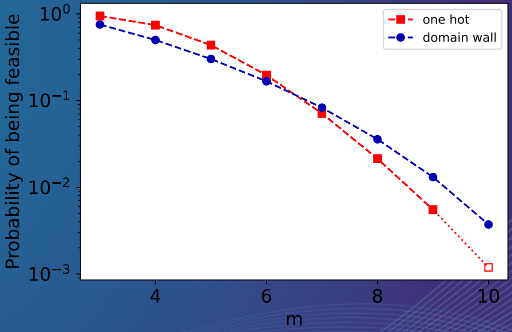
Minor embedding chains need to be stronger for larger problems → less range left for problem, effectively higher temperature ⋆



⋆we use default "uniform torque compensation" method

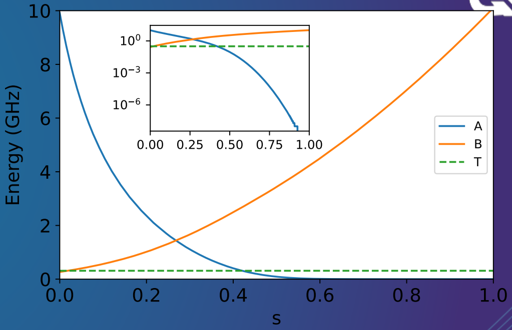
Thermal equilibrium model

Assuming an energy scale of ≈ 5 GHz at the freezing point we find feasible probability for a purely thermal model



- Shows same crossover as real data
- Not in the same location, but...
  1. Estimate of energy scale is rough
  2. Not all sizes will freeze at the same time each will have different scales

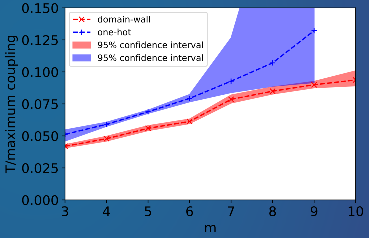
Estimate energy scale

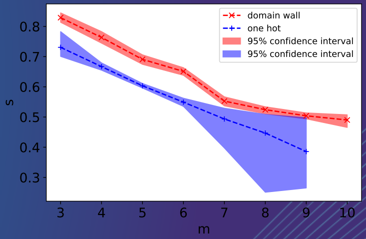


1. Assume "frozen in" thermal distribution $\rightarrow$ Kibble-Zurek style approximation
2. Known physical temperature and experimental success probabilities
3. Back calculate energy scale (B) and therefore freeze point (s)
4. Verify that quantum fluctuations (A) can be safely ignored at freeze point

Effective temperature and freeze point

QCi Innovative Quantum Solutions Company



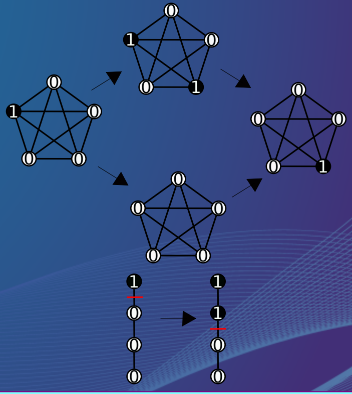


- Already taken into account embedding strength
- Domain-wall version effectively sampled at lower temperature $\leftrightarrow$ later freezing

Encoding has a strong effect on the dynamics of how the problem is solved

Why might this be true?

- One-hot value cannot be changed by flipping a single binary variable
- Domain-wall can therefore easier for transverse field to update



Need to consider underlying physics with encoding

Key points

The continuous-time setting (including quantum annealing) is a promising setting for understanding how to solve problems

Important practical considerations exist for near-term quantum computing

Most convenient theoretical setting often does not match what is practical, we make steps toward advancing this kind of theory

We need a better understanding of the interface between problem encoding and physical dynamics

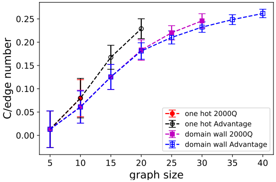
Encoding can have a fundamental and dramatic effect on the physics

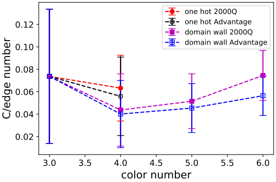
Supplemental slides

Comparing domain-wall and one-hot on hard colouring problems*

For both k and three colouring problems the domain-wall encoding performs better on both Advantage and 2000Q

three colouring (left), k-colouring (right)





C=number of places same colour touches

Even looks like domain-wall on 2000Q out-performs one-hot on Advantage!

Use hypothesis testing to verify that this is a statistically significant result, test 100 instances on each and see how much each processor/encoding combination wins for all 6 combinations

* Chen et. al. IEEE Transactions on Quantum Engineering 3102714 (2021)

Hypothesis testing, three colour

Green=statistically significant result (95% confidence)

	Adv. dw/oh		2000Q dw/oh		dw Adv./2000Q		oh Adv./2000Q		(dw, Adv.)/(oh, 2000Q)		(dw, 2000Q)/(oh, Adv.)	
5 node (b,w)	0	0	0	0	0	0	0	0	0	0	0	0
5 node p												
10 node (b,w)	42	0	37	0	2	0	19	21	39	0	40	0
10 node p	2.27×10^{-13}		7.28×10^{-12}		2.50×10^{-1}		6.82×10^{-1}		1.82×10^{-12}		9.09×10^{-13}	
15 node (b,w)	85	2	95	3	32	34	70	22	94	1	91	2
15 node p	2.47×10^{-23}		4.95×10^{-25}		6.44×10^{-1}		2.67×10^{-7}		2.42×10^{-27}		4.41×10^{-25}	
20 node (b,w)	99	0	100	0	43	41	94	3	100	0	93	2
20 node p	1.58×10^{-30}		7.89×10^{-31}		4.57×10^{-1}		9.60×10^{-25}		7.89×10^{-31}		1.15×10^{-25}	
25 node (b,w)	100	0		FAIL	66	20		FAIL		FAIL	98	2
25 node p	7.89×10^{-31}				3.33×10^{-7}						3.98×10^{-27}	
30 node (b,w)	100	0		FAIL	72	20		FAIL		FAIL	97	2
30 node p	7.89×10^{-31}				2.30×10^{-8}						7.81×10^{-27}	
35 node (b,w)	100	0	FAIL	FAIL		FAIL		FAIL		FAIL	FAIL	
35 node p	7.89×10^{-31}											
40 node(b,w)	100	0	FAIL	FAIL		FAIL		FAIL		FAIL	FAIL	
40 node p	7.89×10^{-31}											

- Domain wall 2000Q beats one hot Advantage (in a statistically significant way)
- Trend continue up to size where no longer possible to embed in 2000Q (FAIL)
- Otherwise results are expected $\rightarrow$ 2000Q worse than Advantage, one hot worse than domain wall

Hypothesis testing, k colour

Green/red=statistically significant result (95% confidence)

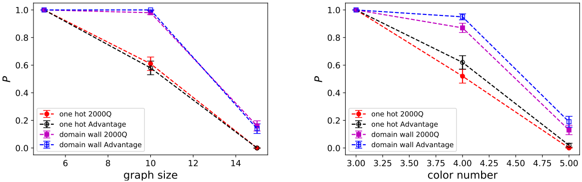
	Adv. dw/oh		2000Q dw/oh		dw Adv./2000Q		oh Adv./2000Q		(dw, Adv.)/(oh, 2000Q)		(dw, 2000Q)/(oh, Adv.)	
3 color (b,w)	0	0	0	0	0	0	0	0	0	0	0	0
3 color p												
4 color (b,w)	34	1	37	2	11	3	26	16	44	1	33	7
4 color p	1.05×10^{-9}		1.42×10^{-9}		2.87×10^{-2}		8.21×10^{-2}		1.31×10^{-12}		2.11×10^{-5}	
5 color (b,w)	91	1	78	1	34	18	23	59	88	1	91	1
5 color p	1.88×10^{-26}		1.32×10^{-22}		1.82×10^{-2}		≈ 1		1.45×10^{-25}		1.88×10^{-26}	
6 color(b,w)	99	0		FAIL	59	15		FAIL		FAIL	99	0
6 color p	1.58×10^{-30}				1.28×10^{-7}						1.58×10^{-30}	
7 color(b,w)	92	0	FAIL	FAIL		FAIL		FAIL		FAIL	FAIL	
7 color p	2.02×10^{-28}											

- Domain wall 2000Q beats one hot Advantage (in a statistically significant way)
- Trend continue up to size where no longer possible to embed in 2000Q (FAIL)
- One case where 2000Q beats advantage for the same decoding (one hot)*

*This goes away when the decoding strategy for broken chains is changed so probably an artefact of majority vote decoding

Same pattern holds for probability to find optimal

three colouring (left), k-colouring (right)



Note that each run was only performed with 100 reads, better results could be attained with more reads

All QPU-encoding combinations found optimal solution at smallest size → explains no "winners" in hypothesis testing