Load Balancing For High Performance Computing Using Quantum Annealing

QSE Meetup 2025

Omer Rathore,
Nicholas Chancellor, Alastair Basden, Halim Kusumaatmaja

Talk Outline

- Quantum annealing overview
- Load balancing
  - Definition
- Motivation
  - Why should we care and why bother with quantum annealing?
- Methods
  - Grid based vs particle based
- Results
  - Comparison with classical algorithms
  - Scalability

Quantum Annealing Theory

- A quantum system in it’s ground state, remains in the ground state if perturbations to the Hamiltoninan are slow “enough”…
- Interpolate: $$H(t) = A(t)H_A + B(t)H_B$$

Initial (H_A) Final (H_B)

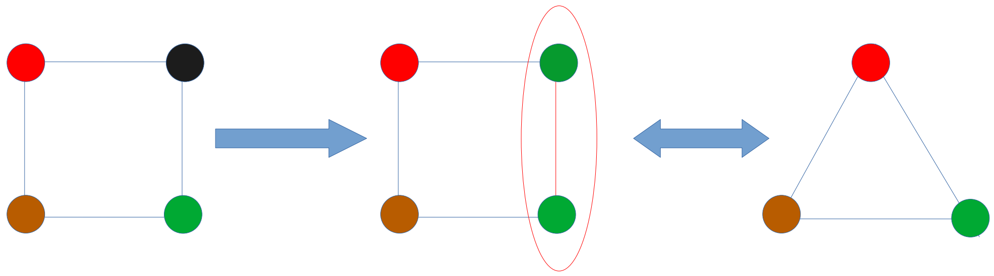
- Choose initial Hamintonian with easy to prepare ground state
- Encode problem of interest into final Hamiltonian
- Result: ground state of problem of interest
- QA is a <u>heuristic</u> algorithm for combinatorial optimisation

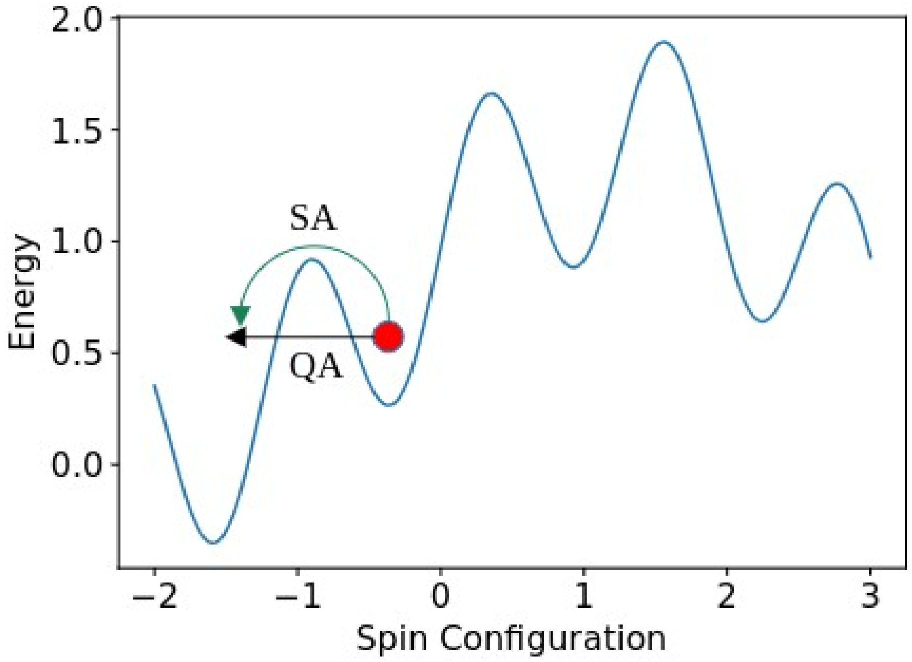
Quantum Annealing Implementation

- D-Wave accepts problem Ising Hamiltonians:

$$H_B = \sum_{i \in V} h_i \sigma_i^z + \sum_{(i,j) \in E} J_{ij} \sigma_i^z \sigma_j^z$$

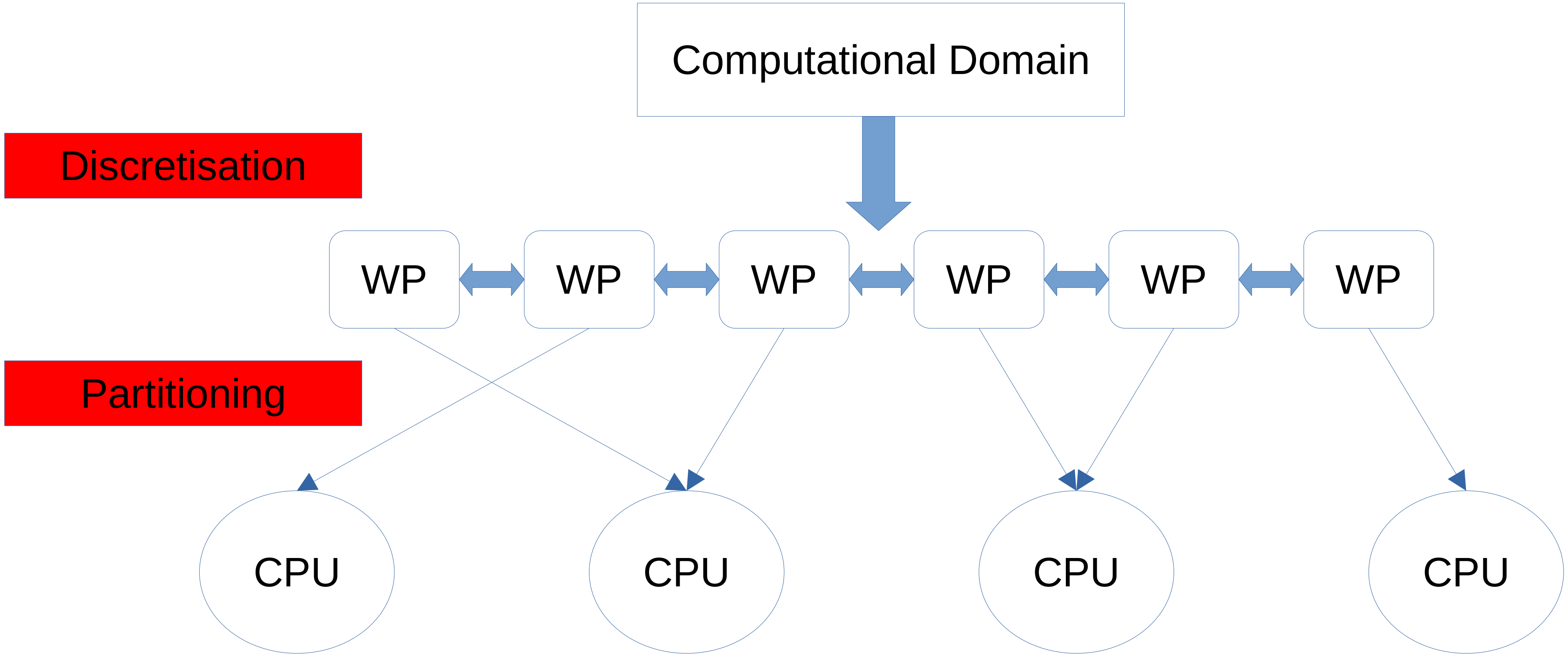
- Limited hardware connectivity
- Embedding uses chains of qubits to compensate





Quantum advantage?

Load balancing



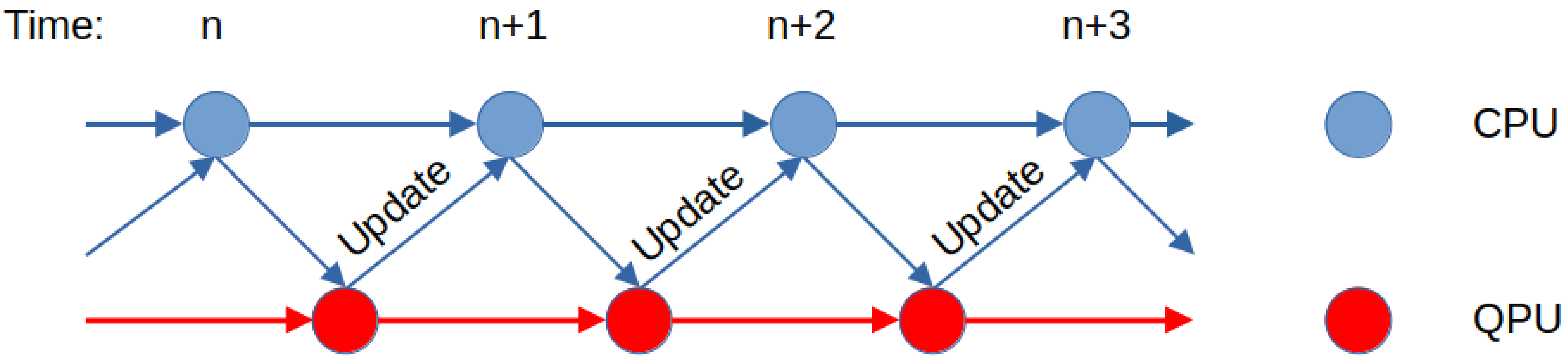
Choice of discretisation influences structure of WPs!

Why should we care?

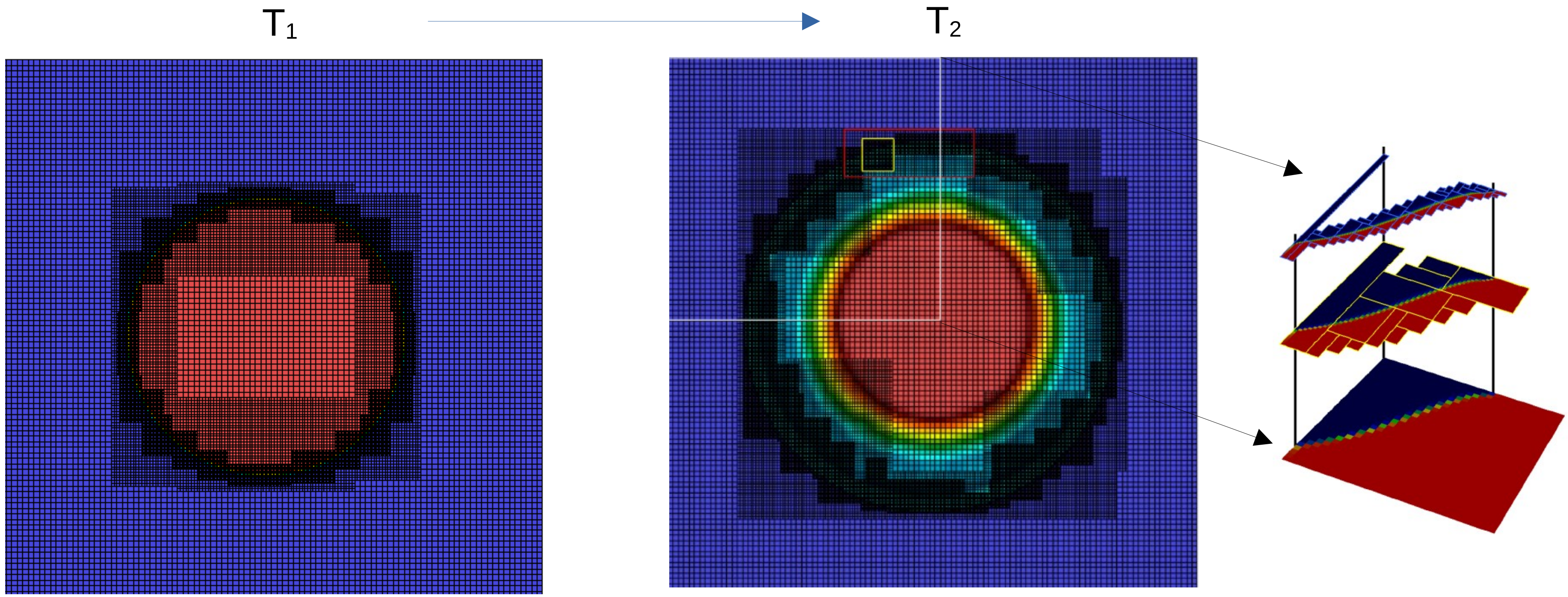
Crucial in leveraging modern HPC! Especially as we scale up to many cores.

Why Quantum Annealing?

- Large and complex solution space
- Small increase in solution quality becomes important when scaled
- Asynchronous implementation can leverage QPU/CPU synergy



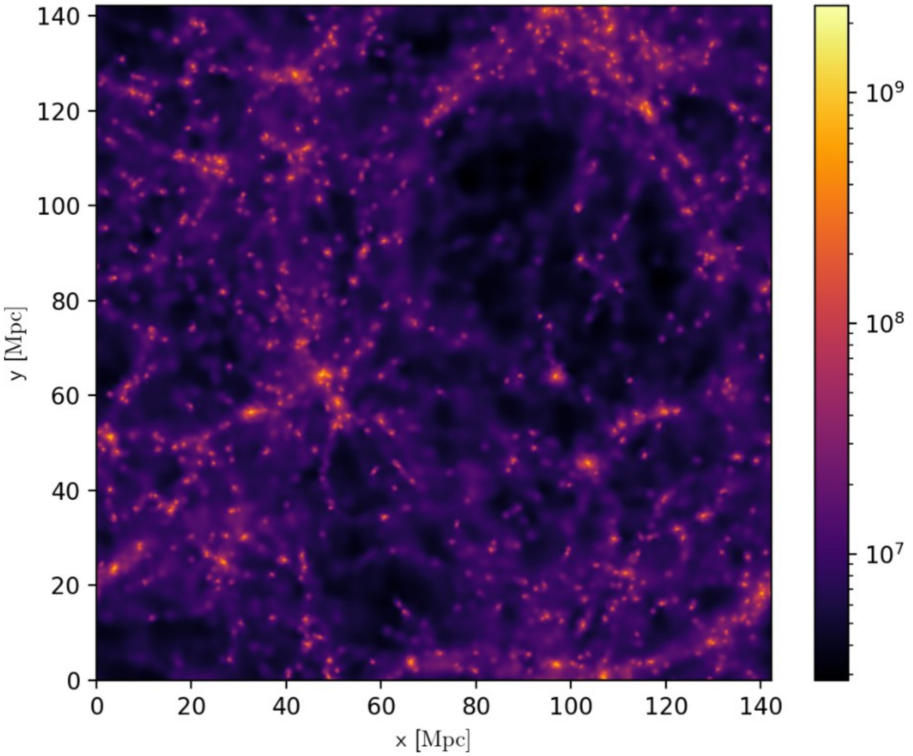
Methods: Grid Based

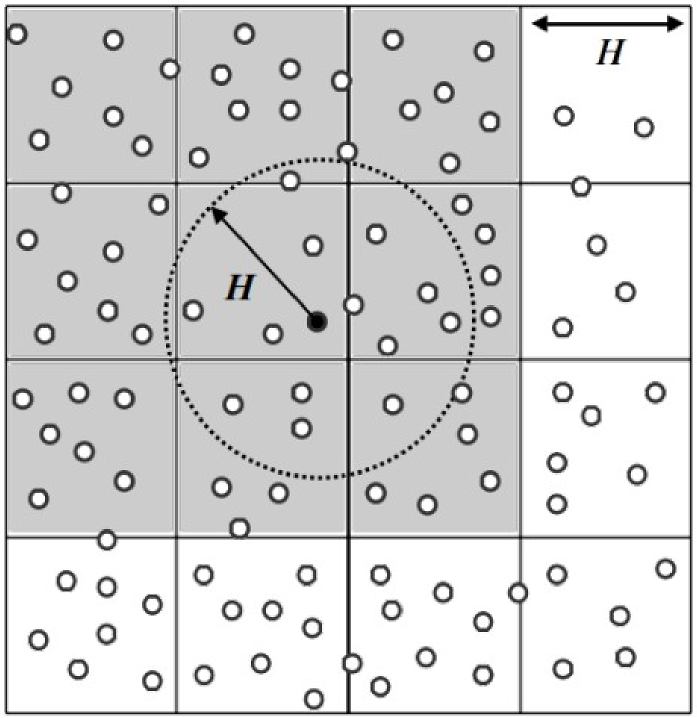


- Nested hierarchy of grids
- High intra-connectivity and low inter-connectivity

Bell, J., et al.
github.com/BoxLib-Codes/BoxLib (2012)

Methods: Particle Based





Schaller, Matthieu, et al. Monthly Notices of the Royal Astronomical Society 530.2 (2024).

- Particles grouped into cells
- Operations span at most 1 neighbouring cell

Problem Formulation

- Adaptive Mesh Refinement (grid based)

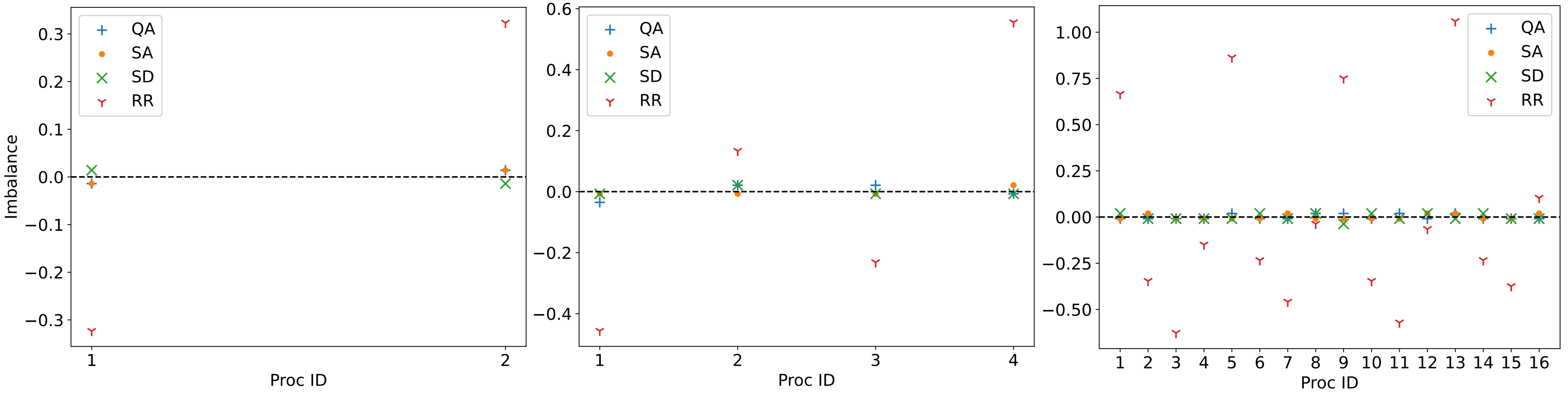
$$H = A\left(\sum_{i=1}^{N} n_i s_i\right)^2$$

- Smoothed Particle Hydrodynamics (particle based)

$$H_1 = \left(\sum_{n=1^N} w_i s_i\right)^2, \qquad H_2 = \sum_{(uv)\in E} e_i \frac{1 - s_u s_v}{2}$$

AMR (Grid Based)

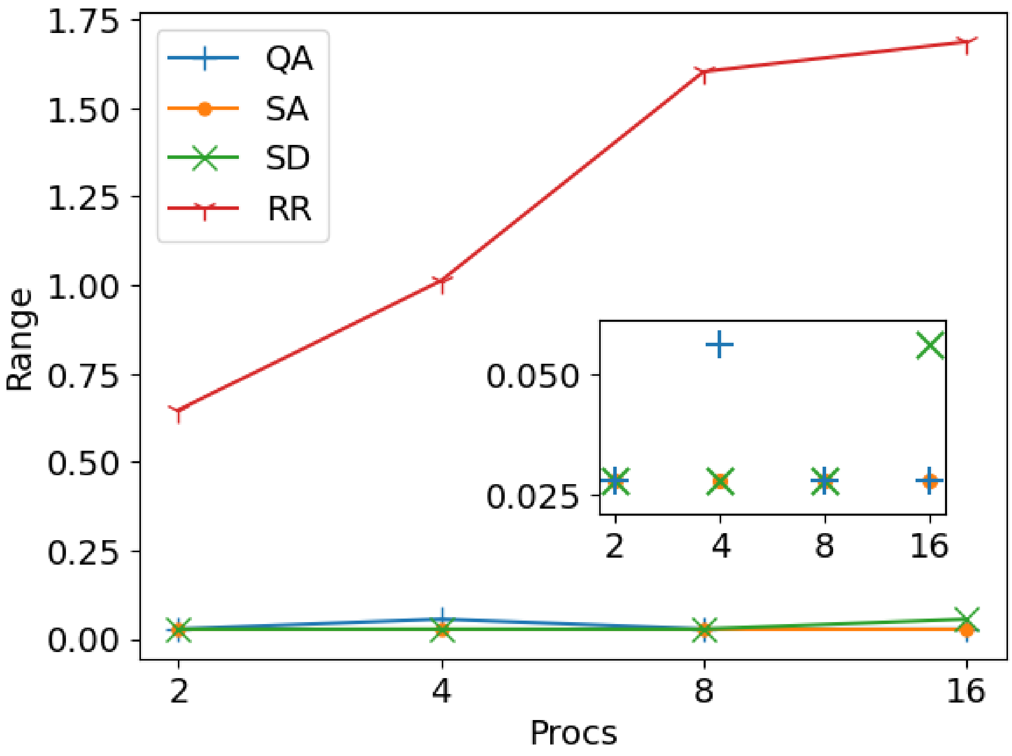
Example Load Balancing Partitions



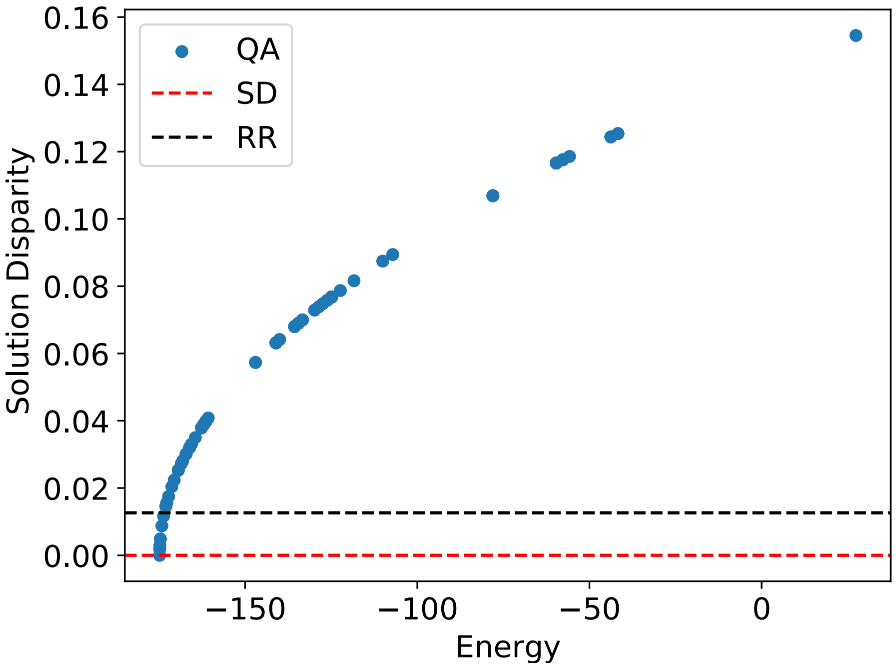
- Partitioning carried out recursively

Overall Performance

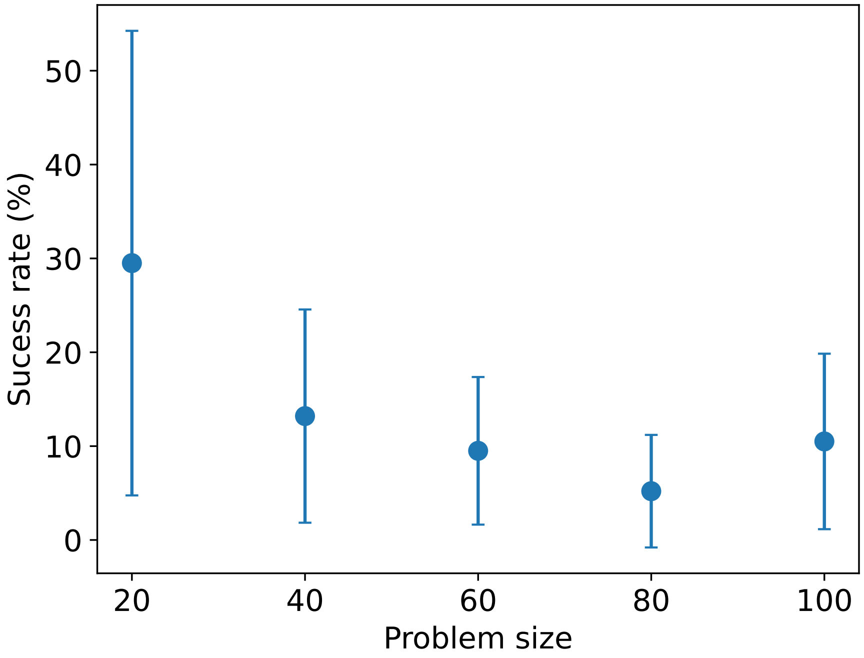
- What about the maximum work disparity?
- Good performance at small problem size
- Clear advantage over RR
- Close agreement with SA/SD in general
- Parameter tuning? Obstacles to scalability?



Likelihood of good solution

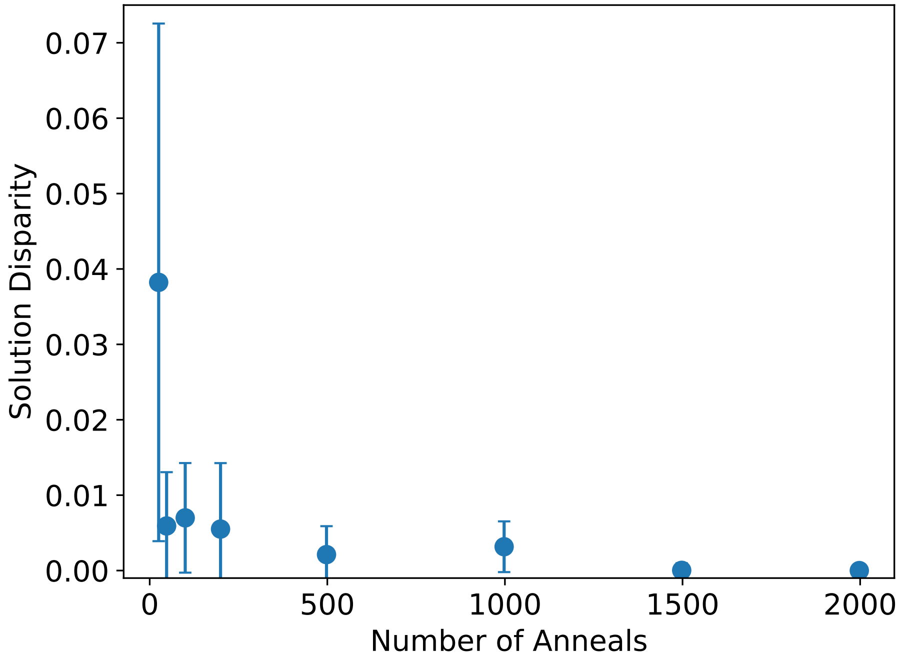


Effectively guaranteed improvement over RR

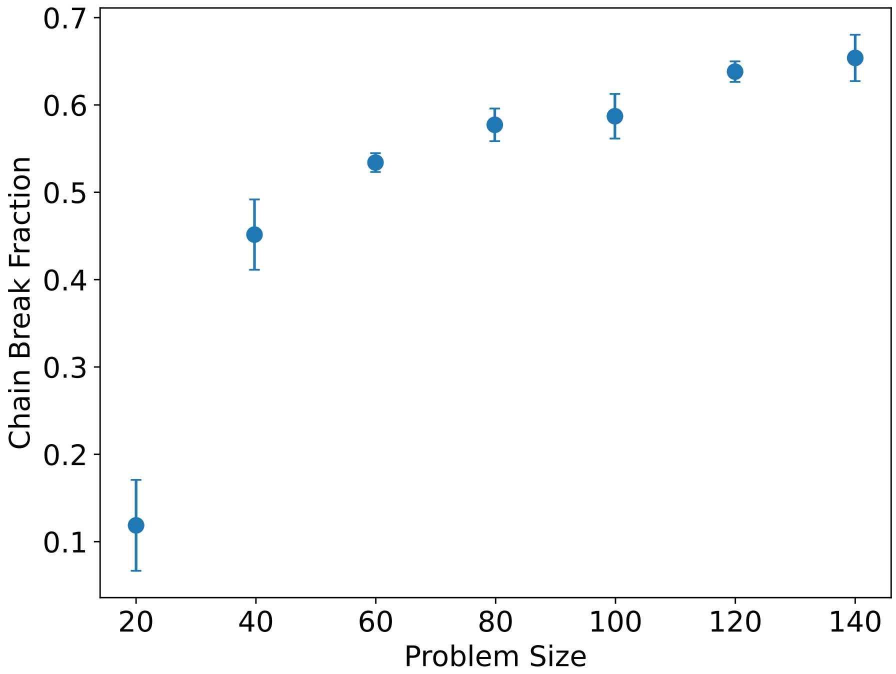


Degradation with problem size

Roadblocks to Scalability

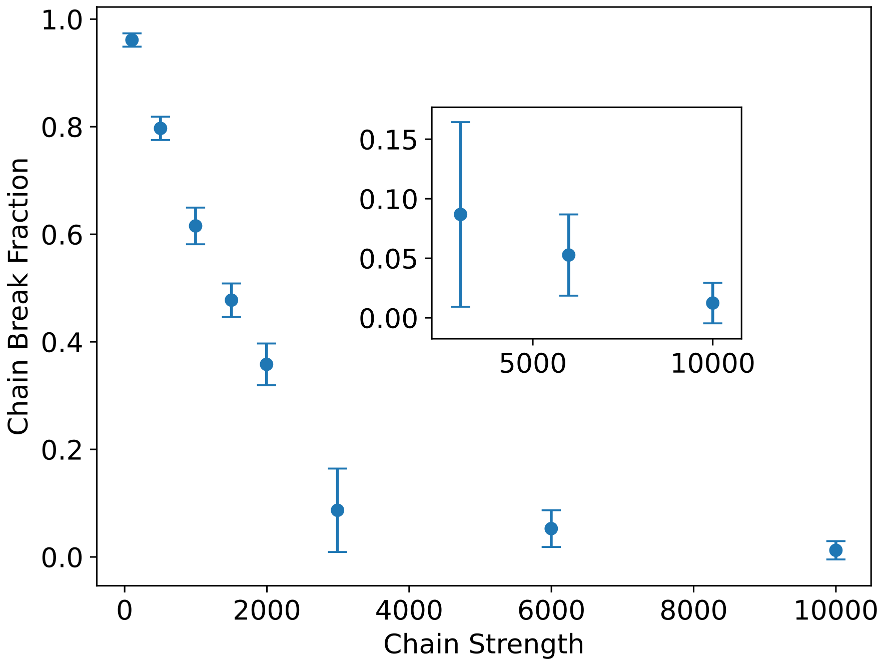


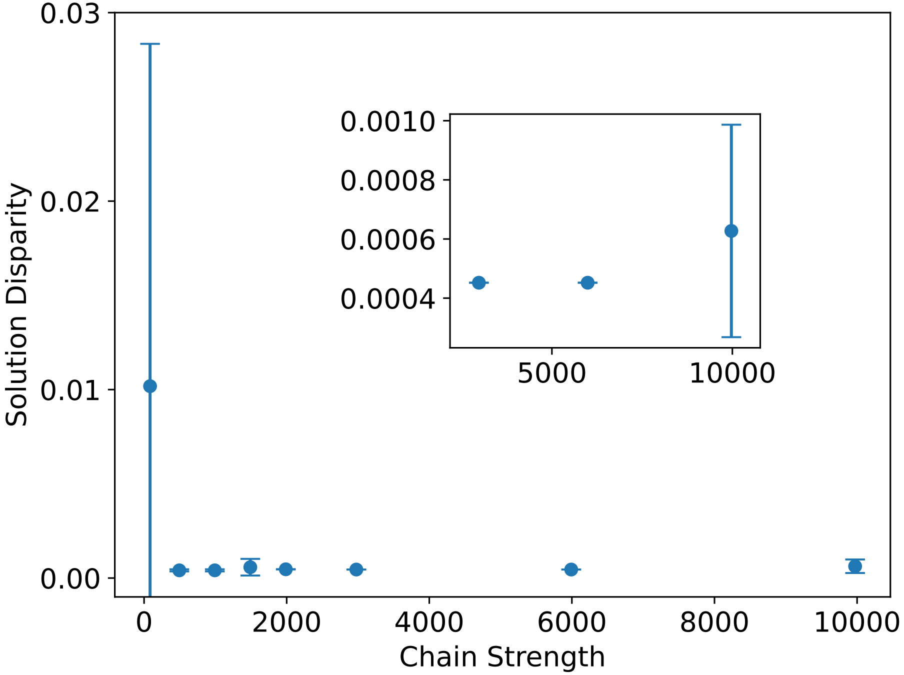
Improvement with number of anneals quickly saturates



Embedding becomes a problem as expected...

Parameter Tuning?

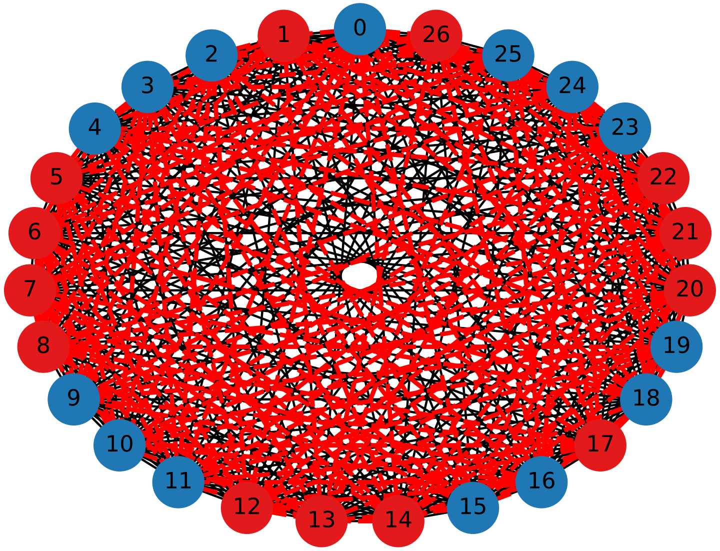


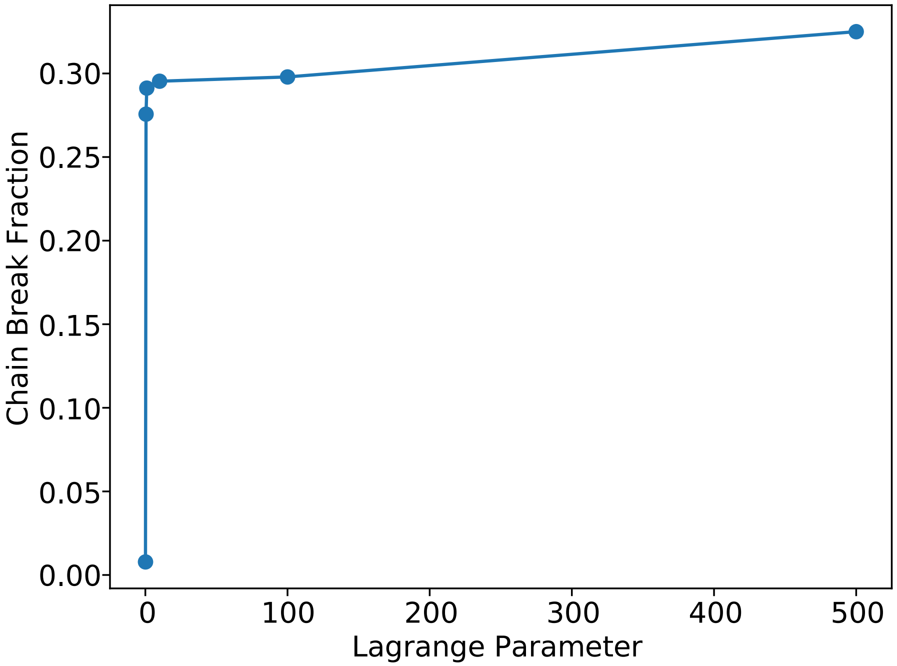


Although possible to obtain good solutions, the problem fundamentally remains fully connected!

SPH (Particle Based)

Weighted Graph Partitioning

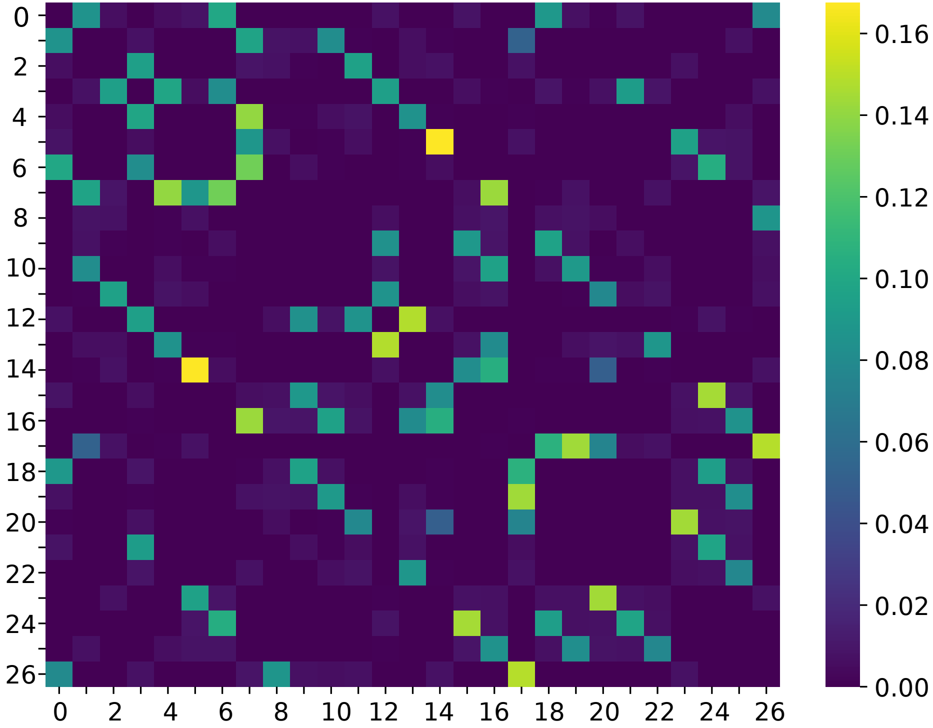




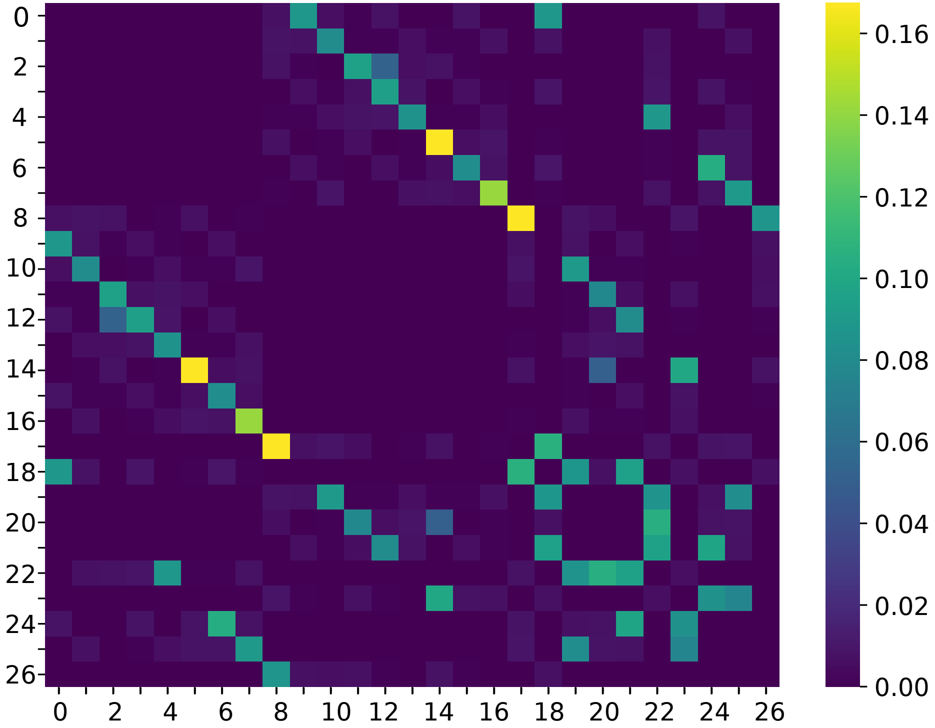
Fully connected problem! Or is it?

More resilient?

Cut Edge Weights



METIS



QA

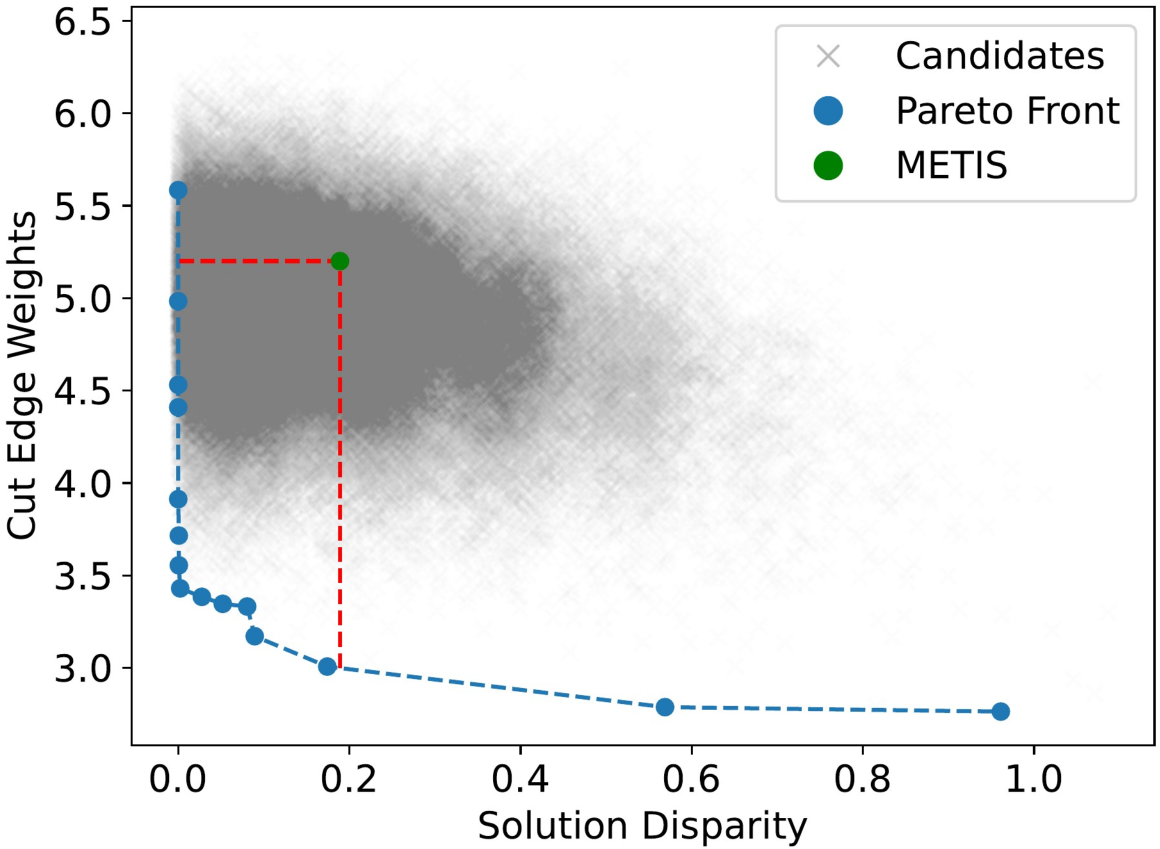
Overall Performance

	Solution Disparity	Cut Edge Weights
Quantum Annealing	0.057	3.69
METIS	0.189	5.20
Performance Ratio	3.32	1.41

- Can simultaneously improve both objectives
- Problem will not remain fully connected at larger problem sizes

Approximate Pareto Front

- Can match partition to individual architectures using Lagrange parameter
- User can determine whether intra or inter processor communication is the priority
- 41% of QA solutions are Pareto dominant compared to METIS
- Approach can be extended to simultaneous (instead of recursive) higher order partitions



Summary

- Motivation for using QA to address load balancing in HPC
- Grid based methods :
  - Possible to obtain as good a solution as optimised classical
  - Problem remains fully connected
- Particle based methods :
  - QA solutions are Pareto dominant over state of the art
  - Expected to scale better for larger problems

Thank you for your attention. Any questions are welcome.

https://doi.org/10.1103/PhysRevResearch.7.013067