Benchmarking the Coherent Ising Machine for Large Time Horizon Vehicle Routing Problems with Time Windows

Tyler King*
McMahon Laboratory, Cornell University
(Dated: July 8, 2022)

The Ising model, a mathematical model of ferromagnetism, is an optimization problem generally found in statistical mechanics that consists of discrete variables that are either +1 or -1. Optimizing these models has benefits for commercial maritime problems like vehicle routing, as these problems reduce well into an Ising formulation. Due to the computational complexity of Ising models, optimization-based techniques are required, which includes physical solvers such as the coherent Ising machine (CIM). This report mainly focuses on digital implementations of CIMs that replicate the networks of coupled degenerate optical parametric oscillators to run theoretical simulations of physical Ising solvers, particularly as the time horizon (TH) of problem formulations grow up to 180 days.

I. ISING MODELS

The Ising model is a model between binary variables in an $N \times N$ matrix format that is popular across various subfields. It was first introduced as a theoretical model for alignment between positive and negative spins (+1 and -1, respectively) [1]. Optimizing this model tends to come in the form of finding

$$\arg\min_{\vec{\sigma}} H(\vec{\sigma}),$$

where $H(\vec{\sigma})$ represents the Hamiltonian energy of the Ising model for some configuration $\vec{\sigma}$ [2]. For the Ising problem, this Hamiltonian is computed by

$$H = -\sum_{1\leq i\leq j\leq N} J_{ij}\sigma_i\sigma_j - \sum_{1\leq i\leq N} h_i\sigma_i.$$

Note that J is an upper triangular $N \times N$ matrix representing spin couplings, $\vec{h}$ is an $N \times 1$ vector that represents external field terms, and $\vec{\sigma}$ is an $N \times 1$ vector that consists of only +1 and -1 values (i.e. $\vec{\sigma} \in \{-1, +1\}^N$) [2–4]. Since the Ising model is NP-complete, it is hard to solve this problem in reasonable time for larger configurations [5]. This gives rise to computationally hard optimization problem solvers such as coherent Ising machines, which are artificial optical systems of oscillators that typically have binary states of either amplitude (0 or α_0) or phase (0 or π) [6].

A. Ising Model Formulation

One of the largest appeals of Ising models is the ease with which various problems map well into Ising format (also known as Quadratic Unconstrained Binary Optimization). Some well known problems such as MAX-CUT and the Sherrington-Kirkpatrick spin glass problems reduce easily into Ising format, allowing these problems to be solved with the use of coherent Ising machines (CIMs) [7].

However, since these problems tend to be NP-complete (and thus NP-hard), it is extremely difficult to find "optimal" solutions except for small problem formulations. As a result, CIMs tend to compute high quality, but not necessarily optimal, solutions [8].

Another well-known problem with commercial usage that reduces well into the Ising problem is the Vehicle Routing Problems with Time Windows (VRPTWs).

II. VEHICLE ROUTING PROBLEMS WITH TIME WINDOWS

Routing problems are well-studied combinatorial optimization problem that appear frequently in logistics and operations research. They tend to involve a fleet of vehicles that try to satisfy some goal (such as maximal profit or minimal time). [9]

VRPTWs are a specific variation of this problem, involving choosing routes for a fixed fleet of vehicles that serve certain nodes within specific time windows [10]. They are a particular subset of the maritime inventory routing problem (MIRP), which are a more general class of combinatorial optimization problems which track inventory levels of products. These problems can be formulated mathematically in four different ways: route-based, arc-based, sequence-based, and sequence-based with continuous time [9]. For this semester, we focused on the first three formulations since they reduce well into Ising format as they consist primarily (or exclusively) of binary variables and linear/quadratic constraints [9].

A. Route-based Ising Model

Route-based formulation attempts to solve the traversal of certain routes. Routes consist of sequences of nodes

* ttk22@cornell.edu

that satisfy all constraints (such as the time window, number of available nodes that could be visited, etc). All routes begin and end at a particular node that is considered the depot. These routes are then returned by the route-based formulation.

This formulation is NP-hard, since for N nodes, there are $N!$ total possibilities. Furthermore, this route-based formulation tends to cover the largest total number of nodes, but this count can be reduced with preprocessing steps that influence the characteristics of formulation. As a result of route-based formulation being more general, it is used to develop both arc-based and sequence-based formulations.

B. Arc-based Ising Model

Arc-based formulation computes discrete routes based on the time when ships travel. When a ship travels from one node to the next, the time is incremented based on the travel time between the nodes. Various constraints exist to ensure that the routes satisfy certain constraints such as hitting all nodes and arriving before or during a particular time window.

Due to the large number of constraints, this formulation tends to form the largest number of variables (and thus J matrix size), leading to harder problems. These issues are only exacerbated when specific ships are tracked.

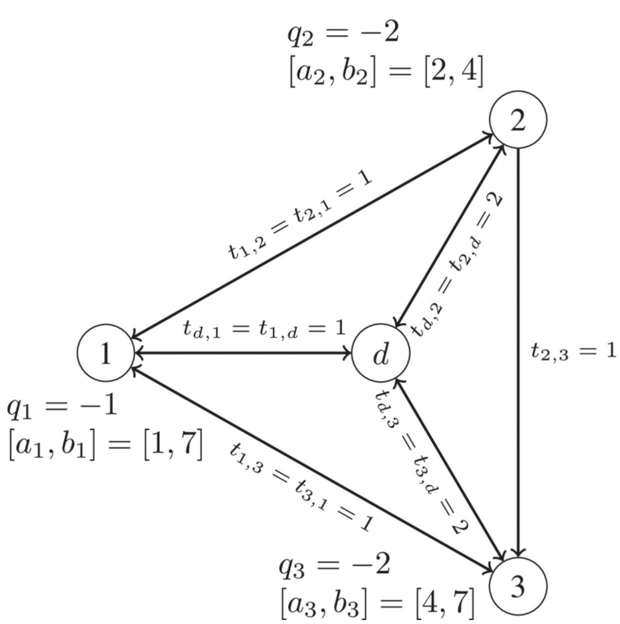

FIG. 1: Small example of VRPTW from ExxonMobil [9]. Visual model of an arc-based implementation.

C. Sequence-based Ising Model

Sequence-based formulation consists of putting an upper bound P on the number of stops that each vehicle can make (which is $P-2$ since the start and end nodes must be the depot). Once the vehicle returns to the depot, it cannot travel again. Additional constraints exist to check that all nodes are hit and each position is hit in the sequence only once, etc.

III. COHERENT ISING MACHINE

With Ising models being NP-complete and thus being computationally expensive to find the optimal solution, approximation solutions are necessary [2, 4]. One way to approach this is coherent Ising machines, which consist of a network of degenerate optical parametric oscillators (DOPOs) in a physical system [4]. Both numerical [11] and experimental results [2, 12] indicate that CIMs are an efficient heuristic solver for NP-hard optimization problems (including problems like MAX-CUT). To replicate the +1 (up-spin) and -1 (down-spin) states, phase amplitudes are manipulated to be positive or negative. These phases are then coupled together, allowing the Ising problem to be modeled. For large-scale tests with CIMs, virtual simulators are used to solve Ising problems that mirror the physical computations conducted by CIMs.

A. Simulated CIM Progress

Using code setup by Brian and others, it is possible to simulate the effects of a CIM. One of our first tasks was allowing this script to be interruptable: i.e. stop the script while it is computing results and record all pertinent information up to that point. The implementation of try-catch statements allowed for `InterruptException` to be thrown and caught, stopping the program and saving the relevant completed runs.

We also developed a script to test the number of violated constraints for the solution generated by the CIM simulator. This information was compiled and leveraged in the April report with ExxonMobil.

IV. ANALOG OPTIMIZATION SOLVERS

A large focus of our work was gaining experience with analog optimization solvers, which involved working with QUBO formulations of problems based on past work done by Hiro, where ODEs of the form

$$\dot{q_i} = f(q_i) + \sum_{ij} J_{ij} q_j$$

were created, such that q_i is a vector that consists of all spins (i.e. $\overrightarrow{\sigma}$) and J_{ij} is the QUBO problem matrix [13].

Note that $f(q)$ is some function that has zeros at $q = \pm a$ and has negative slopes at those points (i.e. they are stable fixed points).

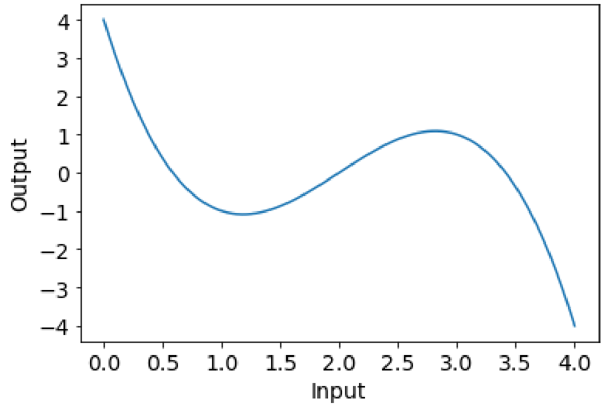

FIG. 2: Example $f(q) = 2q - q^2$ function that is shifted to the positive x-axis. [14]

While there are various functions that satisfy the aforementioned constraints (including piecewise functions), one of the most common examples in CIM literature includes $f(q) = 2q - q^2$. $f(q)$ can be shifted to the positive x-axis (i.e. $q_i := q_i + b$ where $b > a$) to allow for the implementation of an external field.

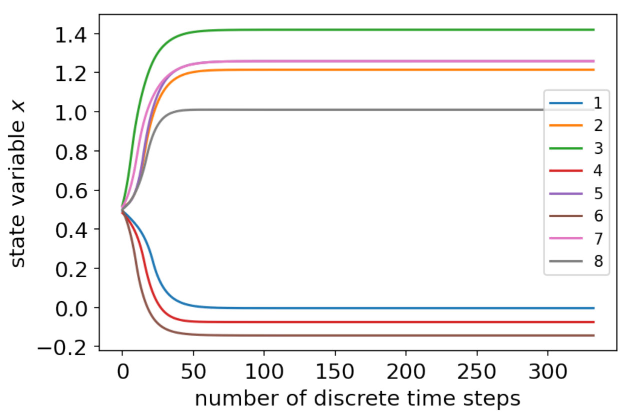

FIG. 3: Bifurcation of initial spins (state variables) over ~350 time steps given an 8-spin system. [14]

This naive approach was utilized as a baseline for which amplitude heterogeneity correction was built on.

A. Amplitude Heterogeneity Correction

While the current implementation of simulated amplitude heterogeneity correction builds off the QUBO formulation, we hope to modify it to interface with Ising problems in the near future and compete with the current iteration CIM simulator (based on a MAX CUT approach).

After implementing a simple QUBO optimizer, we tackle Timothee's amplitude heterogeneity correction implementation that incorporated auxiliary degrees of freedom (i.e. error variables) into the phase space that corrected the aforementioned amplitude heterogeneity that was observed in the prior subsection (ex. figure 3) [7]. The error variables play a key role in guaranteeing positive entropy production in the system, allowing the system to continuous evolve towards a local optimum.

While Brian has done past work with the MAX-CUT implementation of these algorithms, we mainly focused on leveraging the algorithm for Sherrington-Kirkpatrick spin glass problems and expanding these problems to different VRPTW formulations.

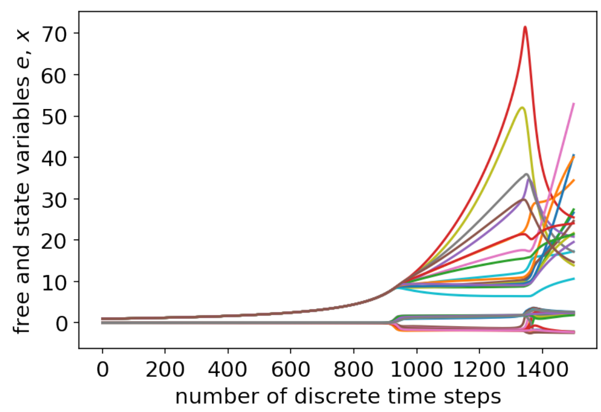

FIG. 4: Bifurcation of initial spins (state variables) over ~1500 discrete time steps and the related free variables given an 18-spin system. [14]

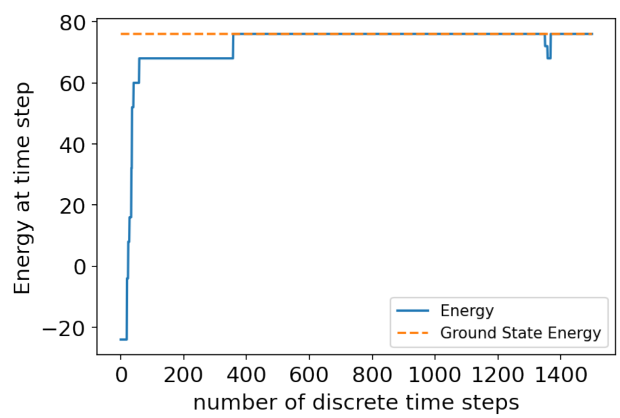

FIG. 5: Evolution of the energy of the 18-spin system over ~1500 discrete time steps. [14]

The full derivation for the algorithms used can be found in the appendix and come from the supplemental material of Timothee's amplitude heterogeneity correction paper [7]. While the algorithm is meant for Ising formulations, it still maintains some level of functionality on low-spin QUBO formulations.

V. GENERATING INSTANCES

`generate_test_set.py` allows for the formulation of large Ising models given a particular time horizon (TH) and number of routes [15]. These larger problems are correlated with significantly higher spin counts in Ising problems, leading to difficulty optimizing formulations. For every TH specified when generating test sets, 6 distinct files are created. One for each formulation (path-based, arc-based, sequence-based) which are either a feasibility or optimality instance. Feasibility tests whether the global minimum can be reached, while optimality deals with finding the lowest minimum.

These instances are then tested with Brian's implementation of the CIM simulator (`MIRP_TestSet4_vrp.jl`).

VI. SIMULATED RESULTS

After working through the Amplitude Heterogeneity Correction model, Hiro tasked us with replicating Brian's results from the early April ExxonMobil meeting. We thus transitioned from QUBO back to Ising formulations, utilizing past code in the `CIMSimulator` repository to obtain results. This would segue our progress into generating large-TH results for the May ExxonMobil meeting.

While optimizing larger instances is a key goal, we've had greater success working with time horizons problems ≤ 180 (particularly for sequence-based and path-based). Although our focus has centered around feasibility instances of problem formulations, we have generated a technique that allowed for large objective instance testing by tapering the energy (i.e. repeatedly scaling down by 10) if it overflowed the maximum float value. While untested for sufficiently large problems, this modification should allow for the script to mitigate overflow errors that we experienced earlier on in the semester.

TABLE I: A table with various `MAX_TIME` (number of discrete time steps) implementations for single-instance solving of `test_sb_15785_f.rudy` (TH=180 instance).

Ground State Energy	Lowest Energy Achieved	% Error	`MAX_TIME`[a]
-1282422	-1282213	0.0163%	2500
-1282422	-1282223	0.0155%	10000
-1282422	-1282214	0.0162%	50000

[a] These results were generated before `@elapsed` was implemented and will be updated before the ExxonMobil meeting.

Although it was a single run per trial, the results still have low deviation. Viewing the energy of the 50,000 time-step system indicates that the minimum energy is reached much earlier, and after ∼3,000 time-steps the energy tends to oscillates at higher energy values.

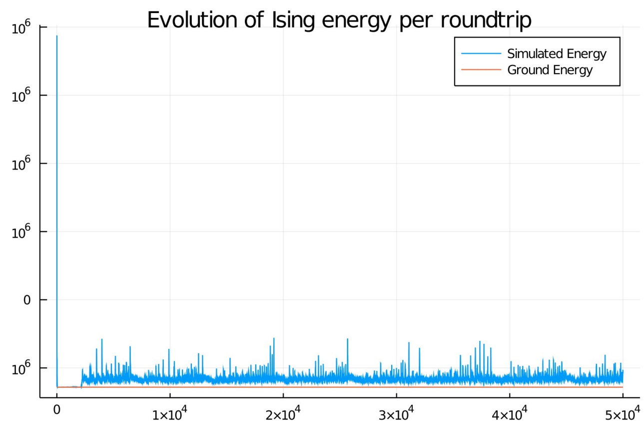

FIG. 6: Evolution of Ising energy for `test_sb_15785_f.rudy` over 50,000 discrete time steps.

We implemented a time counter using `@elapsed` to confirm reasonable runtime for generated results. This is particularly relevant as the TH passes 180, as results for a `MAX_TIME` of 50,000 can take days to generate.

VII. FUTURE WORK

With the ExxonMobil meeting on 5/27, there is a short-term goal to generate numerous results for time horizon problems up to 180 (particularly for sequence-based and path-based instances). In particular, we are interested in tracking the time it takes to run `MIRP_TestSet4_vrp.jl` with various `MAX_TIME` setups and the accuracy it achieves with particular configurations. We also plan to develop thorough plots that characterize the relationship between the time horizon of instances and performance/computational time.

Over a longer time-frame, we plan to finalize flag-planting for the Sherrington-Kirkpatrick spin glass problem along with tuning its functionality towards Ising problem formulations instead of QUBO formulations.

ACKNOWLEDGMENTS

I would like to thank Brian for his patience and support along with Professor Peter McMahon and Dr. Tatsuhiro Onodera for the opportunity to work in this lab and their guidance throughout this semester.

Appendix A: Derivation of Sherrington-Kirkpatrick Spin Glass Problems

This problem can be mathematically modeled by an Euler approximation of ODEs where:

$$x_i(t+\Delta t) = x_i(t) + \Delta x_i(t)\Delta t$$

$$e_i(t+\Delta t) = e_i(t) + \Delta e_i(t)\Delta t$$

if the quantity $(\langle e_i(t)+\Delta e_i(t)\Delta t\rangle) < \Gamma$ where $\langle X_i\rangle$ is the ensemble mean, Γ is the max mean error, x_i are the spin terms, and e_i are the error terms. If the earlier statement is false, then

$$e_i(t+\Delta t) = \frac{e_i(t)+\Delta e_i(t)\Delta t}{\langle e_i(t)+\Delta e_i(t)\Delta t\rangle}.$$

Furthermore, $\Delta x_i(t)$ and $\Delta e_i(t)$ are defined as

$$\Delta x_i(t) = (-1+p(t))x_i(t) - x_i(t)^3 + \epsilon e_i(t)\sum_j J_{ij}x_j(t)$$
$$\Delta e_i(t) = -\beta(x_i(t)^2 - a(t))e_i(t).$$

If configuration $\sigma(t)$, which is defined by $\sigma_i(t) = \frac{x_i(t)}{|x_i(t)|}$, $\forall i$, is not a stable local minimum, which means that for some index j, $\sigma_j(t)h_j(t) < 0$ (where $h(t) = J\sigma(t)$), then $a(t)$ and $p(t)$ (amplitude and linear gain, respectively) are defined as:

$$a(t) = \alpha + \epsilon\langle e_i(t_c)h_i(t)\sigma(t)\rangle$$
$$p(t) = \pi$$

where π is a constant and $t_c < t$ is the last time $x_i(t)$ changed sign for any i (i.e. a change in configuration σ).

If $\sigma(t)$ is instead at a local minimum, i.e. $\sigma_i(t)h_i(t) > 0$ for all i and $\theta(t) > \eta\left(\mu_0\left(\sigma(t)\right)\right)$ where $\theta(t) = \frac{a(t)}{1-p(t)}$, $0 < \eta < 1$, and $\mu_0(\sigma(t)) < 2$, then $a(t)$ and $p(t)$ are instead defined as:

$$a(t) = \eta F\left(\mu_0\left(\sigma(t)\right)\right)(1-p(t))$$
$$p(t) = \pi + a(t) - a(t-1)$$

where $F(X) = -\frac{X}{X-2}$ and $\mu_0\left(\sigma(t)\right)$ is the dominant eigenvalue of $D\left[(\sigma(t)\cdot h(t))^{-1}\right]J - I$.

[1] J. M. B. Haslbeck, S. Epskamp, M. Marsman, and L. J. Waldorp, enInterpreting the ising model: The input matters, Multivariate Behavioral Research **56**, 303 (2021).

[2] P. L. McMahon, A. Marandi, Y. Haribara, R. Hamerly, C. Langrock, S. Tamate, T. Inagaki, H. Takesue, S. Utsunomiya, K. Aihara, R. L. Byer, M. M. Fejer, H. Mabuchi, and Y. Yamamoto, enA fully programmable 100-spin coherent Ising machine with all-to-all connections, Science **354**, 614 (2016).

[3] R. Hamerly, T. Inagaki, P. L. McMahon, D. Venturelli, A. Marandi, T. Onodera, E. Ng, C. Langrock, K. Inaba, T. Honjo, K. Enbutsu, T. Umeki, R. Kasahara, S. Utsunomiya, S. Kako, K. ichi Kawarabayashi, R. L. Byer, M. M. Fejer, H. Mabuchi, D. Englund, E. Rieffel, H. Takesue, and Y. Yamamoto, Experimental investigation of performance differences between coherent ising machines and a quantum annealer, Science Advances **5**, eaau0823 (2019), https://www.science.org/doi/pdf/10.1126/sciadv.aau0823.

[4] Y. Yamamoto, K. Aihara, T. Leleu, K.-i. Kawarabayashi, S. Kako, M. Fejer, K. Inoue, and H. Takesue, enPublisher Correction: Coherent Ising machines—optical neural networks operating at the quantum limit, npj Quantum Information **4**, 9 (2018).

[5] A. Lucas, Ising formulations of many NP problems, Frontiers in Physics **2**, 10.3389/fphy.2014.00005 (2014).

[6] F. Böhm, G. Verschaffelt, and G. Van der Sande, enA poor man's coherent Ising machine based on optoelectronic feedback systems for solving optimization problems, Nature Communications **10**, 3538 (2019).

[7] T. Leleu, Y. Yamamoto, P. L. McMahon, and K. Aihara, enDestabilization of Local Minima in Analog Spin Systems by Correction of Amplitude Heterogeneity, Physical Review Letters **122**, 040607 (2019).

[8] F. Glover, G. Kochenberger, and Y. Du, A Tutorial on Formulating and Using QUBO Models, arXiv:1811.11538 [quant-ph] (2019), arXiv: 1811.11538.

[9] S. Harwood, C. Gambella, D. Trenev, A. Simonetto, D. Bernal Neira, and D. Greenberg, Formulating and Solving Routing Problems on Quantum Computers, IEEE Transactions on Quantum Engineering **2**, 1 (2021).

[10] X. Gan, Y. Wang, S. Li, and B. Niu, enVehicle Routing Problem with Time Windows and Simultaneous Delivery and Pick-Up Service Based on MCPSO, Mathematical Problems in Engineering **2012**, 1 (2012).

[11] Z. Wang, A. Marandi, K. Wen, R. L. Byer, and Y. Yamamoto, enCoherent Ising machine based on degenerate optical parametric oscillators, Physical Review A **88**, 063853 (2013).

[12] A. Marandi, Z. Wang, K. Takata, R. L. Byer, and Y. Yamamoto, enNetwork of time-multiplexed optical parametric oscillators as a coherent Ising machine, Nature Photonics **8**, 937 (2014).

[13] T. Onodera, Simulations of native qubo solver `https://github.com/mcmahon-lab/IncoherentOpticalOptimizationMachines` (2021).

[14] T. King, Amplitude heterogeneity correction for cim `https://github.com/mcmahon-lab/CIMSimulator/blob/master/notebooks/CIM_Tim_Implementation_No_Flag.ipynb` (2022).

[15] S. Harwood, Test set generation for vrptws `https://github.com/mcmahon-lab/CIMSimulator/blob/master/instances/ExxonMobil/2021-10-08_TestSet4/generate_test_set.py` (2019).