# A Coupled Oscillator Based Ising Chip with a Fully-connected Annealing Core, a Hamiltonian Engine, and a RISC-V Gradient Search for Solving 2,500+ Spin Combinatorial Optimization Problems

Chaohui Li*, Yonghwan Hong*, Alex Vanasse, Tahmida Islam,
Peter Kreye, Xunqi Li, Ruihong Yin, Hao Lo, William Moy, Chris Kim
*Department of Electrical and Computer Engineering University of Minnesota, Twin Cities*
*Minneapolis, 55455, USA (*Equal Contribution)*
Email: {li003135, hong0625, vanas073, islam112, krey0031, li001674, yin00473, lo000076, moyxx111, chriskim}@umn.edu

***Abstract*—A 0.9V, 28nm system-on-chip is presented for solving combinatorial optimization problems (COPs) using a network of all-to-all coupled oscillators. This work demonstrates that large COP benchmarks with 2,500 spin variables and 8+ bit resolution weights can be solved to 99.99% accuracy of the best known software solution using a decomposing hybrid compute framework powered by the proposed Ising solver chip, consuming 0.52% the energy of a classical CPU for solving sub-problems.**

***Index Terms*—Combinatorial optimization, coupled oscillator, Ising solver, annealing, Hamiltonian energy**

## I. Introduction

Physics-based annealing chips harness nature's intrinsic ability to relax to the ground state for solving combinatorial optimization problems (COPs). By leveraging the time-dynamical behavior of a network of coupled oscillators, these annealers offer exponentially higher energy efficiency compared to classical algorithms like simulated annealing running on classical CPUs. Unlike digital accelerators of software algorithms or traditional quantum computers relying on exotic devices with cryogenic cooling requirements, CMOS annealing chips based on bistable oscillators or latches have the advantage of emulating quantum properties like superposition and entanglement while being compatible with current foundry technologies. The phase of each oscillator is read out to determine the corresponding spin value. One of the major challenges in deploying coupled oscillator Ising solvers for real-world problems is finding high quality solutions for large-scale problems with many (e.g. $> 2,500$) spin variables and precise (e.g. 8 bit) weights. This paper demonstrates that by using a highly energy-efficient 45 spin Ising solver core, combined with modest amount of post-processing at the sub-problem and full-problem level, it is possible to solve quadratic unconstrained binary optimization (QUBO) problems with over 2,500 spins. The solution achieves 99.99% accuracy and comparable time-to-solution while consuming less than 1% of the energy of a classical CPU for solving sub-problems. This hybrid approach enables large COP problems with thousands of spins and high precision weights to be solved to near-optimal (e.g. 99.99%) levels, utilizing proven decomposer/recomposer approaches such as energy-impact based or breadth first search algorithms, along with a light-weight co-processor CPU for solution quality checking. In this paper, we describe a coupled oscillator based Ising solver chip, the main compute hardware, which includes a physics-based all-to-all connected annealer core, a Hamiltonian energy compute engine, and a RISC-V CPU.

## II. Chip Architecture

The block diagram of the chip is shown in Fig. 1. The annealer core is composed of 45 all-to-all coupled oscillators featuring programmable coupling weights. Each spin is formed by a pair of horizontal and vertical oscillators shorted together in the diagonal cell, ensuring consistent phase propagation throughout the array while coupling with other spins at each intersection [1]. The paired ring oscillators denoted as $s_1$ which corresponds to the local field spin are phase-locked and provide the timing reference for the entire array. The local field spin $s_1$ remains fixed, while the other spins are either +1 or -1, indicating whether they are in phase or out of phase with $s_1$. Each unit cell comprises one memory cell and fourteen coupling cells. The memory cell contains four static random access memroy (SRAM) memory cells, while each coupling cell incorporates the coupling logic as illustrated in Fig. 2. Weights are loaded via a direct memory access interface block, while the RISC-V CPU communicates with the annealing core through memory mapped registers. This configuration allows the RISC-V core to handle various tasks, including initiating the coupling operation, setting the sampling interval, performing gradient searches on the sampled solution, checking the Hamiltonian energy of the solution, and terminating the search after a predefined number of attempts.

Fig. 2 shows a graph representation of the Ising core, with 45 all-to-all connected oscillators and 45x44/2 = 990 couplers. It worth noting that the natively all-to-all connected architecture requires fewer couplers than previous locally

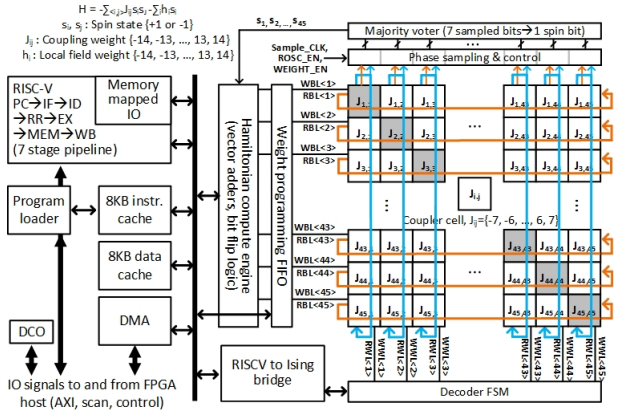


Fig. 1. Chip organization; 45 spin all-to-all connected coupled oscillator Ising core; RISC-V CPU, cache memory, Hamiltonian compute engine, and direct memory access. A 45x45 adjacency matrix of a specific COP is programmed to the Ising core. Oscillator phases automatically settle to a high-quality ground state through coupling dynamics.

connected architectures such as lattice graph or King's graph [2,3,4,5,6]. The corresponding Hamiltonian energy function that the core minimizes is listed below the graph image. The oscillator/coupler circuit depicted in Fig. 2 becomes the unit tile of the 45x45 oscillator array. Each coupler can be programmed to integer values ranging from -7 to 7. As $J_{ij}$ weights can be programmed in both (i,j) and (j,i) location in the 2D array, the overall weight range is from -14 to 14 (=29 weight levels). We introduce a logic-based coupler that eliminates static current consumption of previous resistive or capacitive couplers and can be realized using digital logic gates [7]. Each stage of the coupler cell comprises a baseline inverter in parallel with a tri-state inverter. The enable signal of the tri-state inverter is controlled by the coupling signal from the other path, achieving the same delay shifting functionality as the previous resistive coupler [1]. Multi-bit coupling can be achieved by cascading coupling stages and configuring the individual control logic (Fig. 2, upper right). Each coupler cell also contains 4 memory bits to store the -7 to 7 integer coupling weights. Fig. 2 demonstrates both positive and negative couplings. Taking the S and W couplers as an example, positive coupling begins with the couplers in the out-of-phase. They gradually couple until they are in phase, resulting in identical final spins values (either +1 or -1). For negative coupling, the couplers start from the in-phase and end with out-of-phase, with the opposite final spins values, vice versa. The phase sampling circuit (Fig. 1) is located at the upper edge of the array to compare the phase difference of each vertical oscillator with a reference horizontal oscillator through an XOR gate. The XOR gate's output is sampled seven times over a half-cycle period to determine whether a specific oscillator's phase is in-phase or out-of-phase compared to a reference oscillator also embedded in the array.

Extensive full-chip end-to-end RTL verification was performed to ensure correct functionality. This included a behavioral model of the annealing core extracted from the circuit schematic, which captures the coupling induced oscillator phase shifts. The hybrid analog-digital ASIC design flow facilitates the mapping of real COP problem weights and helps verify the quality of the solution and its improvement with various light-weight post-processing methods. Specifically in this design, we use the raw binary phase (i.e. +1 or -1) data from the annealing core as the seed, and sequentially flip the spin values to check for improvements in the Hamiltonian energy. If no further improvements are observed after flipping all the spin values, the greedy gradient search concludes, and the final solution is sent to the host. To expedite the self-checking routine, we designed a Hamiltonian accelerator engine that utilizes a throughput-optimized vector merge adder. Coupling weights stored in the data cache, along with the sampled spins are fed into the engine while, the RISC-V program orchestrates the entire compute workflow. This includes sampling the oscillator phases at specific intervals, running the post-processing algorithm, comparing the new solution with the previous best, terminating the search after a predetermined number of attempts, and transferring the final solution back to the host computer. Fig. 3 shows the gradient descent search algorithm running on the RISC-V CPU as well as the overall architecture and control signals between the various blocks within the Ising solver system-on-chip.

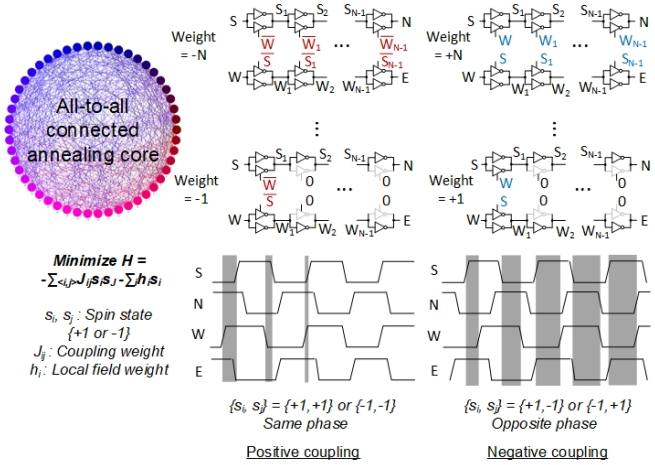


Fig. 2. 45 spin all-to-all connected core and corresponding Hamiltonian energy objective function. Proposed logic-based oscillator/coupler combo tile with cascaded delay stages achieving zero static current, uniform delay shifts, and compact layout.

## III. Chip Testing Results

Fig. 4 illustrates the complete chip testing flow from host API programming to the chip. The host PC leverages the host API to interface with Xilinx Kintex-7 FPGA for initializing the hardware, transmitting RISC-V instructions, and QUBO problems through PCIe 4.0, which achieves approximately six times faster data transmission speed than those of USB 3.0 on Raspberry Pi 4 Model B[1]. The instructions and data are temporarily stored in the FPGA before forwarded via AXI-Stream to the Ising solver chip (i.e., COBIRISCV Chip) for processing. The RISC-V CPU organizes the behavior of the

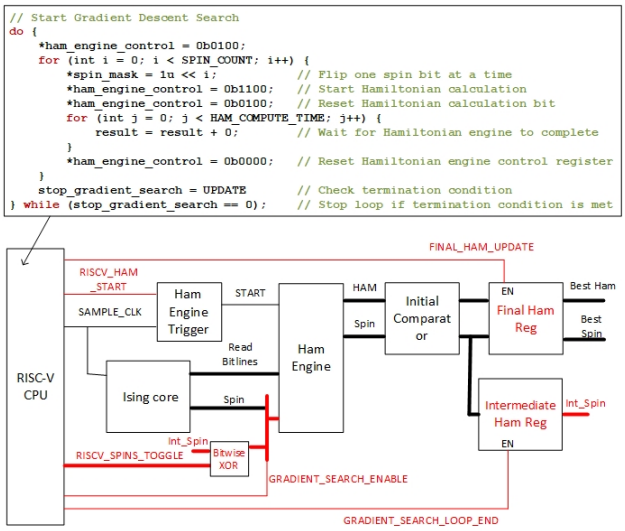


Fig. 3. RISC-V code and Hamiltonian energy compute engine architecture for gradient descent post-processing algorithm.

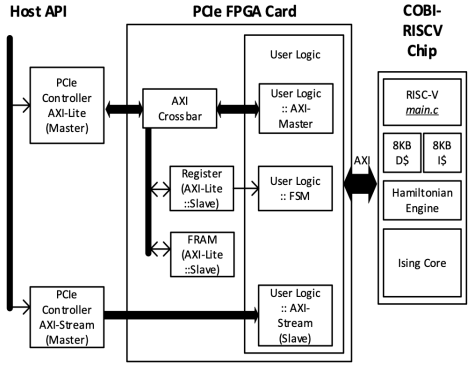


Fig. 4. Overall control software stack, from host application programming interface (API) to FPGA and test chip.

Ising solver core and hamiltonian accelerator, then transmits the computed results back to the FPGA via AXI-Stream. The final results are stored in either FRAMs or Registers via AXI-Lite, waiting for being retrieved by the host PC.

The Ising solver chip is fabricated in 28nm CMOS and has a core area of 0.68mm$^2$. The total chip power consumption of this Ising solver chip under different clock frequencies and operating voltages is shown in Fig. 5 (upper left). The maximum operating frequency was 630 MHz at the nominal 0.9V and room temperature test condition. The breakdown of the power consumption in Fig. 5 (upper right) highlights several key trends; (1) the power consumption of the Ising solver core remains constant as it is determined by the self-oscillating frequency of the coupled oscillators; (2) the CPU and cache power increases linearly with system clock frequency; (3) power consumption is distributed evenly between the analog Ising core and the digital block; (4) the Hamiltonian energy compute engine consumes less than 5% of the total chip power. To address large COP problems that cannot fit into a single core, we developed a decomposing quadratic unconstrained binary optimization (QUBO) software-hardware framework where large (e.g. 2,500+ spins) problems are divided into sub-problems based on either each spin's impact on the overall Hamiltonian energy or the connectivity between spins. These sub-problems are then processed by the solver chip. The best solution obtained after a fixed number of attempts is returned to the host computer and used to find the global solution of the original problem. We repeat this process within a given timeout constraint to find the best possible global solution. Using the decomposing solver framework, we measured the total computation time, sub-problem computation time, and the solution accuracy, normalized to the best known software solution, as a function of the amount of global search, for a 2,500 spin BQP benchmark problem [8] (Fig. 5, lower). Results show that increasing the global search improves the overall solution quality but at the expense of CPU computation time for performing the global search. Data from the chip suggests that reducing the global search until solution accuracy begins to degrade below 100% represents an optimal operating point, though specific requirements will vary depending on the user application.

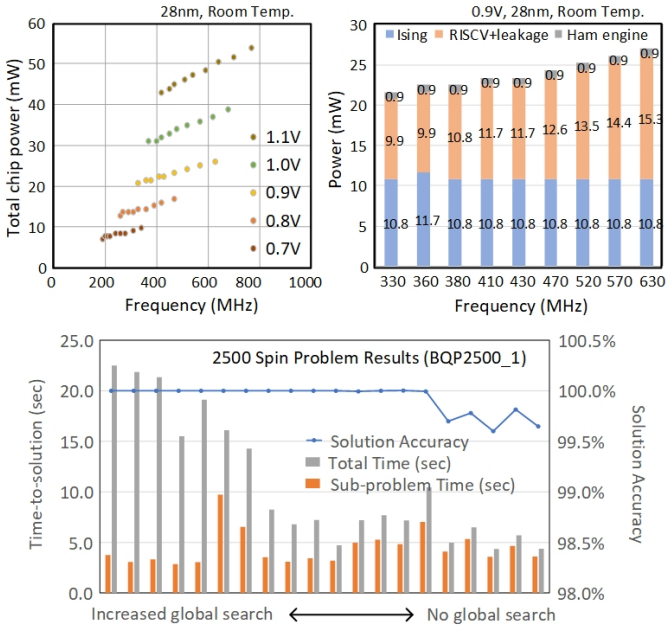


Fig. 5. Power, time-to-solution (2,500 spin problem) and accuracy measurements.

Fig. 6 presents the measured solution accuracy, energy-to-solution, and time-to-solution of a 2,500 spin problem for different solver combinations; chip versus software (using only the tabu search algorithm), and with and without global search. The solution accuracy using the chip alone without any global search was 99.56% of the best solution found by the software with an extended timeout. Introducing a modest amount of global search improves the solution to 99.99% of the best known solution, while the energy consumption remains only 0.52% of that of a a classical CPU, excluding the the energy overhead of the decomposer. We expect the decomposer

overhead to become negligible once this computation step is offloaded to the RISC-V or FPGA controller (not implemented in this work).

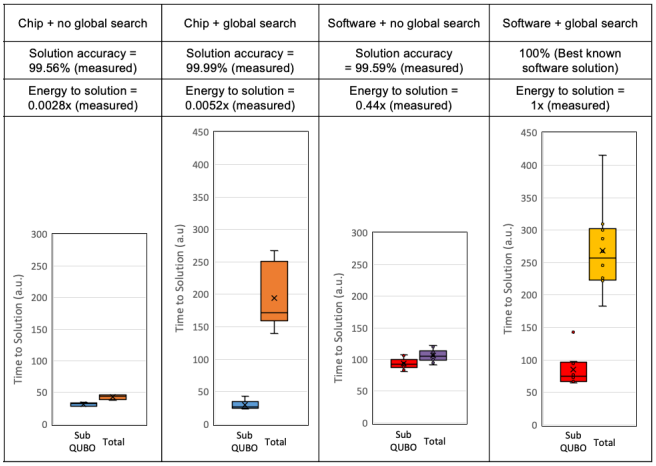


Fig. 6. 2,500 spin BQP results; chip vs. software, with and without global search. Chip results with modest global search achieve 99.99% of the best known software solution while consuming 0.52% the energy of a classical CPU for solving sub-problems.

A comparison with recent quantum [9,10], digital [3], oscillator-based [1,5], and latch-based [2] Ising hardware in Fig. 8 highlights the unique aspects of this work, including a natively all-to-all connected chip architecture with a connectivity of 45, a built-in RISC-V CPU for orchestrating the probabilistic compute flow, integration with a decomposing software stack for solving large COPs, and a hybrid computing approach achieving a 99.99% solution quality while consuming less than 1% of the energy of classical CPUs for solving sub-problems. Compared to prior work, our chip demonstrates superior area efficiency, higher connectivity, and reduced energy consumption. Finally, Fig. 7 displays the die photo, chip floorplan, and the overall testing setup, which includes an FPGA PCIe card. The FPGA host boad based on Xilinx Kintex-7 is connected to the host computer Dell Optiplex 7000. The customized FPGA mezzanine card is designed to support two parallel Ising chips and features 400 FMC connector pins to interface with the PCIe FPGA host board. The FPGA board is connected to the Host PC via PCIe Gen 4.0 slots.

## Acknowledgment

This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) Quantum-Inspired Classical Computing (QuICC) program, National Science Foundation (NSF), and Intel's Transformative HardWare for AI (THWAI) center. We thank Prof. J. Gu, Y. Ju, and X. Chen at Northwestern University for designing and verifying the RISC-V core.

## References

[1] H. Lo, et al., "An Ising solver chip based on coupled ring oscillators with a 48-node all-to-all connected array architecture", Nature Electronics, pp. 771–778, 2023.

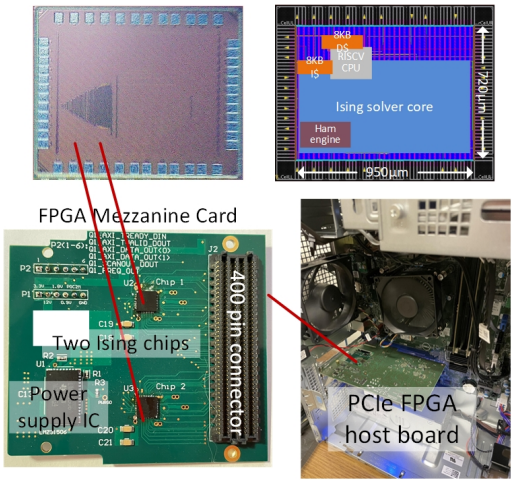


Fig. 7. Die photo, chip floorplan, high-speed FPGA board, and PCIe card setup.

| Technology (computing) | Superconductin g (quantum) | 65nm CMOS (oscillator) | 65nm CMOS (oscillator) | 40nm CMOS (latch) | 65nm CMOS (digital) | 28nm CMOS (oscillator) |
|---|---|---|---|---|---|---|
| Die photo or layout | | | | | | |
| Architecture | | | | | | |
| Publications | D-Wave | I. Ahmed, VLSI 20, JSSC 21 | H. Lo, Nature Electronics 23 | J. Bae, ISSCC 23 | Y. Su, ISSCC 22, JSSC 24 | This Work |
| Connectivity | 15 (Pegasus) | 6 (Hexagonal) | 47 (all-to-all) | 4 (local) | 4 (local) | 45 (all-to-all) |
| Weight resolution | 4-5 bits | -1 to +1 | -14 to +14 | -1 to +1 | -127 to +127 | -14 to +14 |
| Power | ~25,000W | 23mW (Ising core) | 110mW (Ising core) | N/A | 1.167mW @ 1.2V, 96MHz | 10.8mW (Ising), 9.9mW @ 0.9V 330MHz (RISC-V) |
| # of qubits / couplers | >5,000 / >35,000 | 560 / 3,360 | 46 / 1,081 | 1440 / N/A | 1024 / N/A | 45 / 990 |
| Equiv. complete graph | ~70 | 23 | 46 | ~18 | ~16 (or 32) | 45 |
| Minor embedding? | Yes | Yes | No | Yes | Yes | No |
| Decomposer? | Yes | No | No | No | No | Yes |
| Solution Accuracy | N/A | 80-95% (random) | 95-100% (randoms) | 100% (w/o contention) | 88-100% (w/, w/o contention) | 99-100% (random graphs) |

Fig. 8. Comparison with recent quantum, digital, oscillator-, and latch-based annealers.

[2] J. Bae, et al., "CTLE-Ising: A 1440-spin continuous-time latch-based Isling machine with one-shot fully-parallel spin updates featuring equalization of spin states," ISSCC, pp. 142–144, Feb. 2023.
[3] Y. Su, et al., "FlexSpin: A CMOS Ising Machine With 256 Flexible Spin Processing Elements With 8-b Coefficients for Solving Combinatorial Optimization Problems," JSSC, vol. 59, no. 8, pp. 2659-2670, Aug. 2024.
[4] M. Yamaoka, et al, "A 20k-spin Ising chip to solve combinatorial optimization problems with CMOS annealing," IEEE JSSC, vol. 51, no. 1, pp. 303–309, Jan. 2016.
[5] I. Ahmed, et al., "A Probabilistic Compute Fabric Based on Coupled Ring Oscillators for Solving Combinatorial Optimization Problems," JSSC, vol. 56, no. 9, pp. 2870-2880, Sept. 2021.
[6] S. Ebadi et al., "Quantum optimization of maximum independent set using Rydberg atom arrays," Science 376, pp. 1209-1215, 2022.
[7] U. Karpuzcu, et al., "COBI: A Coupled Oscillator Based Ising Chip for Combinatorial Optimization", April 2024, preprint.
[8] J.E. Beasley, "Heuristic algorithms for the unconstrained binary quadratic programming problem," Technical Report, Management School, Imperial College, UK, 1998.
[9] N. Dattani, et al., "Pegasus: The second connectivity graph for large-scale quantum annealing hardware", arXiv.
[10] C. McGeoch, et al., "Advantage Processor Overview".