2021 NSF-FET workshop on Post-Von Neumann Ising Machines

Steering Committee: Hideo Mabuchi (Stanford), Alex Gaeta (Columbia), Surya Ganguli (Stanford), Helmut Katzgraber (Amazon), Zoltan Torockai (Notre Dame), Yoshihisa Yamamoto (NTT)

As we consider future contributions that high-performance computing could make to our society and economy, few if any technical challenges loom larger than that of radically improving our ability to find – with lower latency and energy consumption – better solutions to non-convex ("NP-hard") optimization problems. Optimization problems lie at the core of broad application areas spanning operations research, public policy, engineering design of structures and molecules, machine learning and artificial intelligence. In these contexts, optimization tasks are integrated deeply within complex heterogeneous workflows that incorporate nontraditional factors such as substantial energy costs, price-performance tradeoffs, and cloud computing. As a result, conventional research on algorithms and code development will be insufficient on their own to enable the kind of progress required for revolutionary impact. At the same time, the hardware substrate of computing is evolving rapidly with the onset of formidable obstacles to Dennard/Moore/Koomey scaling, and the introduction of novel materials and integrated photonics to minimize energy dissipation and improve data movement. High-profile industrial efforts to demonstrate quantum computers have likewise reignited broader interest in special-purpose information processing architectures, including optimizers such as the D-Wave quantum annealers that do not fit within the framework of conventional scaling analysis.

Recasting the foundations of computational optimization to leverage emerging physics, mathematics, and management science will thus require an expansive technology co-design approach that facilitates simultaneous and coordinated rethinking of devices, architectures, algorithms and performance metrics. While there is of course an important role for high-level theory to elucidate key issues related to mapping among distinct problem classes (QUBO, knapsack, MAX-CUT, etc.), our aim in this NSF workshop proposal will be to develop a concrete vision for optimization co-design grounded in recent work on post-von Neumann Ising Machines (PVNIMs). By focusing our outlook in this way, we hope to spark the coalescence of vertically-integrated research communities around existing experimental hardware platforms and build consensus regarding overarching goals for improving both our basic understanding of non-convex optimization and the practical impact of our frontier efforts.

PVNIMs generate solutions to the Ising ground-state problem – a paradigmatic non-convex optimization problem – by embedding discrete optimization in the continuous dynamics of analog physical or cyber-physical systems. We may include in this class, for example, Coherent Ising Machines [1], quantum annealers [2], Ising machines based on spatial light modulation [3,4] or laser networks [5,6], coupled CMOS ring oscillator networks [7] and autonomous p-computers [8]. While implementation and benchmarking efforts for this class of physical optimization engines have proliferated over the past few years, very little is yet understood about their scaling or advantages in real-world application contexts. Two central questions have begun to emerge as organizing principles for fundamental research: first, how do Ising machines fit into the heuristic model of "performance complementarity" for computationally hard problems [9], and second, what should we make of the fact that even numerical simulations of novel Ising machine "architectures" running on conventional computers [10,11] may constitute highly competitive new heuristic algorithms for combinatorial optimization? As a corollary to the latter, what advantages can native hardware implementations of Ising machine-type architectures offer in terms of lowered latencies, capital cost, or energy consumption?

The issue of performance complementarity – the empirical observation that different algorithms work best for solving different instances of a computationally hard problem – of course represents a broad challenge for contemporary computer science, but with specific attention to PVNIMs we may ask what makes a given optimization instance easy or hard for our architectures. Can modifications or generalizations of an architecture mitigate key bottlenecks to deliver leading performance on broader classes of instances? To make a real impact on applied optimization, research on the latter questions must be coupled with statistical characterization of instances arising in real-world practice, as well as mathematical analysis of cross-cutting features – What kinds of instances actually arise in practice? Only by combining deeper perspectives on these two aspects of performance complementarity can we develop a sharper view of the role that Ising machines may play in future optimization technologies.

The apparent power of conventional numerical simulations of PVNIM architectures – viewed as optimization algorithms in their own right – gives rise to a view of an underlying scheme for embedding a discrete optimization problem into a continuous physical model (specified at the level of differential equations) as a type of *meta-architecture*. By lifting the notion of "algorithm" into something more like a policy for controlling salient topological properties of a physical phase portrait, we introduce new complexities (such as the potential for chaotic behavior) but gain opportunities to apply methods of dynamical systems theory, singularity theory, and the mathematics of systems of polynomial equations to the study of combinatorial optimization. Can modeling and simulation studies of the empirical performance of prototype hardware actually teach us something fundamentally new in the theory of algorithms? Conversely, can an analysis of the essential dynamical features of the Ising machine meta-architecture lead us to discover hardware implementation strategies with dramatically improved resource efficiencies? What does it mean that the coupled differential equations we use to model prototype Coherent Ising Machines (networks of coupled optical parametric oscillators) are essentially identical to those used for studying the statistical mechanics of large neural networks?

Our aims in this workshop will be to elaborate grand challenges and near-term research priorities for Ising machines and related emerging technologies, and to identify needs and opportunities for novel interdisciplinary collaboration. The workshop will take place online over two days, most likely in the October-November 2021 timeframe. While the final agenda will be subject to speaker/panelist availability and further discussion within the steering committee and NSF program managers, we propose the following session themes and list of tentative presenters/panelists.

1. [1/3 day] Overarching questions
    a. What new possibilities for theoretical analysis are opened by embedding combinatorial optimization – as we do in the CIM model – within dynamical systems theory, statistical physics and the theory of polynomials?
    b. How fair is it to regard the Ising ground-state problem as a canonical non-convex optimization problem? What do we know about instance mapping from alternative problems and concomitant issues of approximation ratios and constraint feasibility?
    c. What have we learned so far from PVNIMs about hybrid opto-electronic computing? What is our strongest case for special-purpose optimization architectures that diverge from the CMOS roadmap?
    d. Do specific use cases and application domains, such as drug design versus deep learning, give rise to distinct end-to-end performance metrics for non-convex optimization? What does management science have to say about computational infrastructure for optimization?
    e. How do robustness and uncertainty factor into the practice of non-convex optimization, for example if some parameters in the specification of an instance are unknown, imprecisely measured, or time-varying? Is there any fundamental connection with physical noise in analog optimization architectures?
2. [2/3 day] Impact assessment beyond application-agnostic scaling of worst/average-case runtimes
    a. How can we construct real-world testbeds for optimizers, which include benchmark instances across scales that capture the statistics of instances that arise in practical optimization? How different are these statistics in different application domains? Is the proprietary nature of industrial data an insurmountable obstacle?
    b. If we can collect such a testbed, how could we develop mathematical abstractions of real-world instance statistics, including constraints?
    c. Taking performance complementarity for granted, can we utilize concepts from statistical physics or dynamical systems theory to develop good *instance-specific* predictors of the time/energy required by PVNIMs to find solutions within a given quality threshold? Or even just binary (good candidate/bad candidate) instance classifiers? Conversely, can we use machine learning and tailored testbeds to discover new "principles" of PVNIM performance?
    d. What are the specific requirements and metrics for non-convex optimization for machine learning and artificial intelligence?
3. [1 day] Existing/proposed PVNIM (including ASIC-CMOS) hardware architectures; next steps and desiderata
    a. Identifying and addressing speed/accuracy bottlenecks and energy inefficiencies
    b. Pareto optimality of solution latency and energy consumption?
    c. Expansion to higher-order and integer optimization
    d. Improved computational nonlinearities (including higher-than-quadratic couplings)
    e. Hardware implementation of solution constraints
    f. Coherent phase-space transport, including local transient quantum effects

References

[1] Y. Yamamoto, T. Leleu, S. Ganguli and H. Mabuchi, “Coherent Ising machines—Quantum optics and neural network Perspectives,” Applied Physics Letters **117**, 160501 (2020).

[2] A. Pearson, A. Mishra, I. Hen and D. A. Lidar, “Analog errors in quantum annealing: doom and hope,” NPJ Quantum Information **5**, 107 (2019).

[3] D. Pierangeli, G. Marcucci and C. Conti, “Large-Scale Photonic Ising Machine by Spatial Light Modulation,” Physical Review Letters **122**, 213902 (2019).

[4] S. Kumar, H. Zhang and Y.-P. Huang, “Large-scale Ising emulation with four body interaction and all-to-all connections,” Communications Physics **3**, 108 (2020).

[5] I. Gershenzon, G. Arwas, S. Gadasi, C. Tradonsky, A. Friesem, O. Raz and N. Davidson, “Exact mapping between a laser network loss rate and the classical XY Hamiltonian by laser loss,” Nanophotonics **9**, 4117 (2020).

[6] M. Babaeian, D. T. Nguyen, V. Demir, M. Akbulut, P.-A. Blanche, Y.Kaneda, S. Guha, M. A. Neifeld and N. Peyghambarian, “A single shot coherent Ising machine based on a network of injection-locked multicore fiber lasers,” Nature Communications **10**, 3516 (2019).

[7] I. Ahmed, P.-W. Chiu, W. Moy and C. H. Kimb, “A Probabilistic Compute Fabric Based on Coupled Ring Oscillators for Solving Combinatorial Optimization Problems,” IEEE Journal of Solid-State Circuits, doi: 10.1109/JSSC.2021.3062821.

[8] B. Sutton, R. Faria, L. A. Ghantasala, R. Jaiswal, K. Y. Camsari and S. Datta, “Autonomous Probabilistic Coprocessing With Petaflips per Second,” IEEE Access **8**, 157238 (2020).

[9] P. Kerschke, H. H. Hoos, F. Neumann and H. Trautmann, “Automated Algorithm Selection: Survey and Perspectives,” Evolutionary Computation **27**, 3 (2018).

[10] E. S. Tiunov, A. E. Ulanov, and A. I. Lvovsky, “Annealing by simulating the coherent Ising machine,” Optics Express **27**, 10288 (2019).

[11] T. Leleu, F. Khoyratee, T. Levi, R. Hamerly, T. Kohno and K. Aihara, “Chaotic Amplitude Control for Neuromorphic Ising Machine in Silico,” arXiv:2009.04084 (2020).