KLIMA: K−Local In-Memory Accelerator for Combinatorial Optimization

Dmitri Strukov

University of California at Santa Barbara, USA
E-mail address: strukov@ece.ucsb.edu

Ising machines, along with related models such as Hopfield neural networks and Boltzmann machines, are promising computational paradigms for solving complex optimization problems. Many real-world problems are naturally described by higher-degree polynomial objective functions [1], necessitating the development of efficient high-order Ising machines. However, the hardware implementations of such machines remain limited, largely due to the difficulty of efficiently computing partial derivatives of high-degree functions in a scalable and reconfigurable manner. A common workaround is to convert high-order problems into equivalent quadratic formulations and solve them using conventional second-order Ising machines. However, this transformation incurs significant complexity and latency overheads, including an increased number of variables and slower convergence due to spurious local minima [1,2].

This talk focuses on KLIMA (K−Local In-Memory Accelerator), a comprehensive framework developed by my research group with our DARPA QuICC program collaborators [1-9] to address the above challenges. A central innovation is an approach for the massively parallel gradient computation for high-degree polynomials, which is well-suited to efficient mixed-signal in-memory computing [3,9]. This method enables circuit implementations whose area complexity scales linearly with the number of variables and polynomial terms, independent of the polynomial's degree. We have built experimental KLIMA prototypes using ReRAM [3,8] and SRAM-based [6,7] in-memory computing circuits, and evaluated them on standard satisfiability problem (SAT) benchmarks. These prototypes, along with simulations on larger-scale problems, demonstrate orders-of-magnitude improvements in speed and energy efficiency over state-of-the-art solutions. Finally, I will discuss a higher-level KLIMA architectures [4,5] inspired by island-style field-programmable gate arrays. Such an architecture leverages the high sparsity of coupling matrices found in many practical (hard) problems, enabling solving efficiently SATs with tens of thousands of variables on a single chip.

References

1. M. Hizani et al., *ISCAS'24* (**2024**)
2. D. Dobrynin *et al. Physics Review B* (**2024**)
3. T. Bhattacharya *et al. Nature Communications* 15, 8211 (**2024**)
4. G. Hutchinson *et al. ISPLED'24* (**2024**)
5. T. Bhattacharya *et al. ISVLSI'24* (**2024**)
6. T. Bhattacharya *et al. VLSI Symp'25* (**2025**)
7. T. Bhattacharya *et al. HotChips'25* (**2025**)
8. G. Pedretti *et al. Nature Unconventional Computing* 2, 7 (**2025**)
9. T. Bhattacharya *et al. submitted* (**2025**).