Large-scale FPGA Simulator for Coherent Ising Machines

Farad Khoyratee[1,2], Timothee Leleu[3,4], Satoshi Kako[5], Franco Nori[1,2] and Yoshihisa Yamamoto[5]

[1]Theoretical Quantum Physics Laboratory, RIKEN Cluster for Pioneering Research, Wako-shi, Saitama 351-0198, Japan
[2]Physics Department, University of Michigan, Ann Arbor, Michigan 48109-1040, USA
[3]Institute of industrial Science, University of Tokyo, Japan
[4]International Research Center for Neurointelligence, University of Tokyo, Japan
[5]Physics & Informatics Laboratories, NTT Research Inc., 1950 University Ave., #600, East Palo Alto, CA94303, USA

1. Introduction

Combinational optimization searches the **best configurations** of a set of variables in order to achieve a certain goal. It can be used in various applications, such as: **financial portfolios**, **cluster analysis**, **interacting proteins, drugs discovery,** or **hardware routing**.

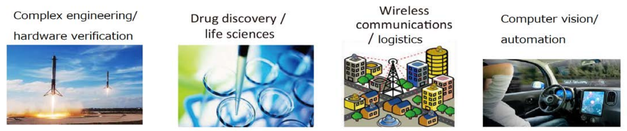

Figure 1: A few examples of applications that could benefit from combinational optimization

Combinational optimization problems require a **computation time that grows exponentially** with the number of variables, when computed on a classical computer. Conventional computers then become inefficient, because of the **Von Neumann bottleneck,** leading to the need of **unconventional hardware**. These problems can be solved by using hardware based on the Ising model, and called **Ising machines**. Photonic implementations of the Ising machine show promising results to solve larger-size combinatorial optimization problems.

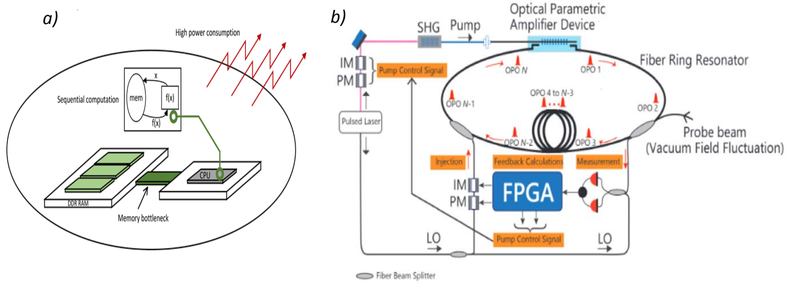

Figure 2: a) Von Neuman bottleneck showing memory-transfer limitation, sequential instructions and high-power dissipation. b) A proposed schematic diagram of a laser-based machine using error detection and correction feedback (Kako et al., 2020)

Alternatively, hardware such as the **Field Programmable Gate Array (FPGA)** exhibit good results for implementing such Ising model. Here we propose to use a **Field Programmable Gate Array (FPGA)** using **brain-inspired dynamics** and **architecture** as a **fast**, **low-power** consumption, **large-scale**, and **flexible Coherent Ising Machine (CIM)** simulator. FPGAs are devices that can be reprogrammed to modify its internal interconnection between logic blocks for computation offering high flexibility.

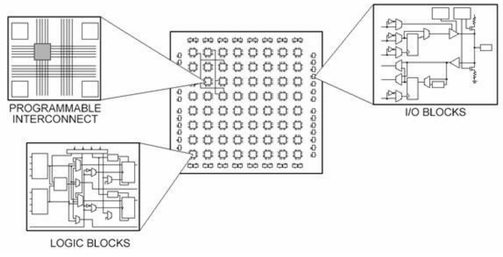

Figure 3: Principle of FPGA from the National Instrument website (https://www.ni.com/)

2. Brain-inspired proposed scheme

Recent Ising machines use dynamics that can be expressed as the descent of a potential function. Rate-based neural networks, as well as CIM classical dynamics, can be described by the following dynamical system:

$$\frac{dx_i}{dt} = f_{i(x_i)} + \beta(t)\sum_j w_{ij} g_{j(x_j)} + \sigma\eta_i$$

Which can be rewritten as the gradient descent of a potential function described as:

$$V = \beta \mathrm{H}(y) + Vb(y)$$

Where,

$$V_b = -\sum_i \int_0^{y_i} f_i\left(g_i^{-1}(y)\right) dy$$

$$H = -\sum_i w_{ij}\, y_i y_j$$

The well-known problem of this approach is that this potential function is non-convex. One strategy is to gradually deform this landscape; but there is no guarantee that the system converges to the global minima of the Ising Hamiltonian. Leleu et al. (2019) proposed the introduction of a micro-structure to the system which induces a chaotic search rather than an annealing process searching the ground-state of the Ising Hamiltonian.

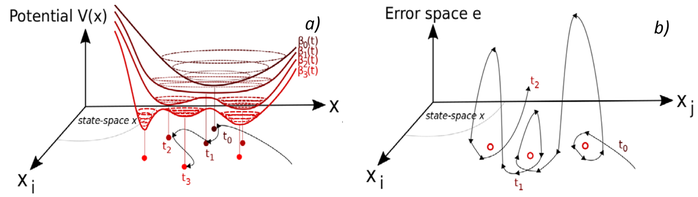

Figure 4: a) Gradual deformation of the landscape (annealing); b) Proposed scheme with the introduction of a micro-structure inducing a chaotic search.

3. Neuromorphic architecture

The brain has inspired many research on various fields. When compared to current technology, the biology remains more efficient in term of speed execution of highly parallelized tasks and low power consumption at low frequency (20W). Thus, neuromorphic hardware that can be described as technologies that implements neural models, has been developed these last decade.

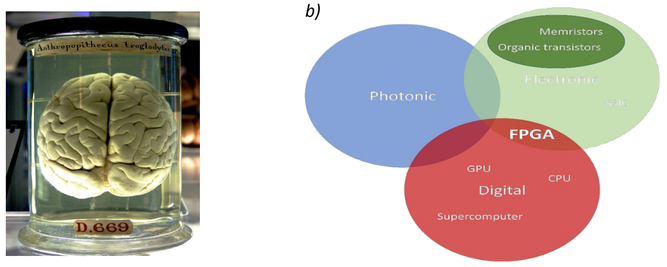

Figure 5: a) The brain can process complex functions at low frequency and low power consumption (~20 W); b) Various neuromorphic technologies that have been developed to provide efficient computation. FPGA is a good compromise between an Application Specified Integrated Circuit (ASIC) and digital computing.

The proposed approach, named Chaotic Amplitude Control (CAC), has been implemented into a XCVU13P FPGA provided by Xilinx and integrated into a board provided by Bittware. The current implementation works using different frequencies to compute the terms shown in figure 6.a). Here, we proposed a neuromorphic architecture to solve Ising problems with several features:

- The use of several frequencies and time steps to reduce power consumption;
- Highly parallelized and pipelined neurons (for a total of 40K neurons);
- Hierarchical organization to optimize communication between neurons;
- All-to-all connection.

a) *b)*

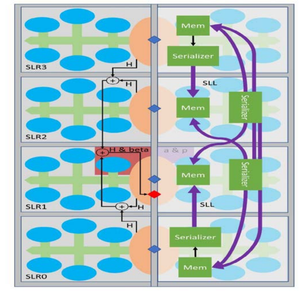

Figure 6: Chaotic amplitude equations computed at different frequencies where f_g=200 MHz, f_x=600 MHz and f_e=400 MHz; b) Proposed FPGA architecture implementation based on the nervous system organization.

4. Results

Here we show the results of the current implementation (XUPVPP design A) and an estimation of different scenarios of the implementation, showing that increasing the computation capacity of one dot product leads to better results.

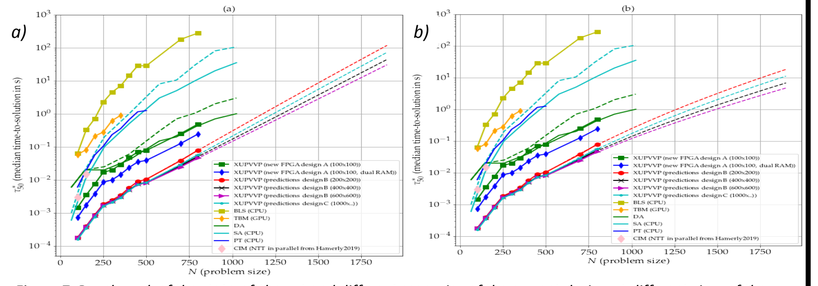

Figure 7. Benchmark of the state-of-the-art and different scenarios of the current design at different sizes of the matrix computation per clock cycle. Two hypothesis are shown here for the scaling of the Time-To-Solution (TTS) according to the problem size N; a) The TTS scales as e^N; b) the TTS scales as $e^{\sqrt{N}}$.

5. Perspectives

One of the advantages of such architecture is its adaptability. We plan to extend this implementation with the addition of a truncated Wigner, quantum Gaussian model, real weight value, and Zeeman terms. Also, the presented neuromorphic architecture will be implemented into a rack of 8 FPGAs to achieve higher speed and larger problem size (N=10^5 to 10^6).

Finally, most Ising machines do not scale when N > 2000, and CPU heuristics then become unreliable. Having the fastest machine would allow us to estimate the scaling for larger system sizes according to the Replica Symmetry Breaking (RSB) theory of Parisi (Boettcher, 2020).