Title: Parallel Optimization of Spiking Neural Networks

Supervisors: Prof. E-G. Talbi & Prof. P. Boulet
Laboratory: CRISTAL UMR CNRS, University of Lille & INRIA

Abstract: The optimization of neuromorphic networks is a complex problem characterized by expensive objective functions and high-dimensional search space. Hence, the development of parallel optimization algorithms is necessary to solve the problem. Moreover, an implementation on GPU-powered clusters of the simulation and surrogate-based parameter optimization of these architectures will be investigated.

Context

This PhD proposal focuses on new computer architectures having the potential to take over architectures based on transistors to further improve the performance of computers after the end of Moore's Law. The investigated model is based on a spiking neural networks. The exploration of the parameter space of these new architectures is performed using the combination of simulation and optimization (e.g. surrogated-based optimization). Large GPU-accelerated computing is highly useful to support the computational burden. The topic addressed in this proposal, a step on the road to the neuromorphic computing era.

Motivations and objectives

Surrogate-assisted optimization consists in replacing the exact computationally intensive simulations by cheaper data-driven approximations so-called surrogates or metamodels [1].

To further support the computational burden of the simulation and surrogate-based optimization process, GPU-accelerated cluster computing will be investigated. Indeed, typically, a modern supercomputer supplies two major coarse levels of parallelism: intra-node and inter-node (or cluster-level). The intra-node level is tending to be heterogeneous (typically CPU+GPU) providing a hierarchy of CPU-side and GPU-side levels of parallelism. Recently, intra-node heterogeneous parallelism for optimization has been the focus of many research works [4,5,7,8]. However, most of these later do not exploit all the supplied levels of parallelism. In addition, to the best of our knowledge the combination of the intra-and inter-node levels is never addressed within the context of optimization. Therefore, although existing parallel hybrid metaheuristics allow to significantly reduce the size of the search space and speed up its exploration they often fail when it comes to tackle very large complex optimization problems as it is the case in our setting. Finally, the proposed approaches will be extensively experimented. The experiments will be conducted on supercomputers.

Program and planning

The PhD thesis will include four major steps summarized in the following:

- State of the art on spiking neural networks, surrogate-assisted optimization and high-performance computing for neuromorphic computing.

- High-level description of the parameterization of these architectures, evaluation of their learning capabilities to allow the optimization of these architectures when the cost function is particularly heavy. The goal is to enable the most automatic “tuning” of spiking neural networks.
- Parallel design and implementation on GPU-powered clusters surrogate-based parameter optimization of these architectures. The experiments will be performed on GPU-powered clusters.

Scientific and economic impact

The results of the PhD thesis will be published in high-level international journals such as JPDC, Parallel Computing, CCPE, IEEE TPDS, etc. and high-ranked international conferences including IEEE IPDPS, ACM GECCO, NICE, etc. From an economic point of view, the thesis will contribute particularly with the parallel surrogate-assisted optimization for simulation-based neuromorphic computing which will be highly useful in many industrial domains.

References

[1] **E-G. Talbi**. “Automated design of deep neural networks”. ACM Computing Surveys, 2020.
[2] L'apprentissage profond : une révolution en intelligence artificielle : https://www.college-de-france.fr/site/yann-lecun/inaugural-lecture-2016-02-04-18h00.htm
[3] Akka : https://akka.io/
[4] Grid'5000 : https://www.cristal.univ-lille.fr/?page=plateforme&id=1
[5] H. Paugam-Moisy, S. Bohte. «Computing with Spiking Neuron Networks». In Handbook of Natural Computing, édité par Grzegorz Rozenberg, Thomas Bäck, et Joost N. Kok, 335-76. Springer Berlin Heidelberg, 2012. http://link.springer.com/referenceworkentry/10.1007/978-3-540-92910-9_10.
[6] P. Falez, P. Devienne, Pierre Tirilly, M. Bilasco, C. Loyez, I. Sourikopoulos, **P. Boulet**. «Flexible Simulation for Neuromorphic Circuit Design: Motion Detection Case Study». In Conférence d’informatique en Parallélisme, Architecture et Système (ComPAS). Sophia Antipolis, France, 2017. https://hal.archives-ouvertes.fr/hal-01538449.
[7] Shahsavari, M., **P. Boulet**. «Parameter Exploration to Improve Performance of Memristor-Based Neuromorphic Architectures». IEEE Transactions on Multi-Scale Computing Systems PP, n°99 (2017): 1-1. https://doi.org/10.1109/TMSCS.2017.2761231.
[8] T. Bartz-Beielstein, B. Filipic, P. Korosec, **E-G. Talbi**, « High-performance simulation-based optimization, Springer, 2020.
[9] A. Hebbal, L. Brevault, M. Balesdent, **E-G. Talbi**, N. Melab, “Bayesian optimization using deep Gaussian processes”, *Optimization and Engineering Journal, 2020.*
[10] J. Pelamatti, L. Brevault, M. Balesdent, **E-G. Talbi**, Y. Guerin, “Efficient global optimization of constrained mixed variable problems”, Journal of Global Optimization JOGO, 2019.