DOI: 10.14529/jsfi210101

# When Sally Met Harry or When AI Met HPC

*Ulises Cortés*[1,3], *Ulises Moya*[2], *Mateo Valero*[3,1]

## Introduction

The Artificial Intelligence (AI) explosion which we are witnessing today can also be, at least in part, credited to the current advances in computing power, in particular to High-Performance Computing. This is not a brand new relation as it can be traced from the very beginning of the hardware and AI developments. Maybe the first encounter between AI and hardware dates back to 1958. The perceptron – a more general computational model than McCullochPitts units – was intended to be a programable machine rather than a software program. While its first implementation was in software for the IBM 704, perceptron subfsequently implemented it in custom-built hardware as the *Mark 1 perceptron* [1, 6]. The perceptron was designed for image recognition: it was an array of 400 photocells, randomly connected to units called *neurons*. Weights were encoded in potentiometers, and electric engines performed weight updates during the learning phase. A seminal interaction between AI and the hardware design was the use of a perceptron for efficient branch prediction to boost instruction-level parallelism [4]. After the years, the evolution of those *neurons* brought the inception of the so-called Neural Networks (NN) as software that gave birth to Deep Learning (DL). In turn, to accelerate DL, nowadays the use of GPU's and more specialized hardware architectures has become the norm (*e.g.* Cerebras CS-2, SambaNova, INTEL's Habana, *etc.*).

In the 80's Thomas Knight, an AI researcher at the Massachusetts Institute of Technology (MIT) said: "*The bigger we make our programs, the 'smarter' they get and the slower they get ... We're in the embarrassing position that we give a program more information, and it gets worse* [9]." The explosion of cheap sensors, the wide use of the Internet and the smart telephones augmented the amount of available data, AI-based technologies request more and more computational power.

The convergence between AI & HPC is seriously and consistently pursued throughout the HPC ecosystem. This includes, as said before, Deep Learning (DL) as the main engine, DL is a greedy approach. It is clear that this AL and HPC union is an excellent opportunity to obtain better and faster scientific results and translate them into industrial applications.

These accomplishments all share a single common thread. Namely, the algorithms developed to accelerate Deep Learning models' training on HPC platforms have a strong experimental component (see [8]). To the date, there is no accurate framework to narrow down the ideal set of hyper-parameters to guarantee rapid convergence and optimal performance levels of AI models as the number of processor or GPU nodes increases to speed up the training stage. Furthermore, it is common, in this community, to compare distributed training algorithms on HPC platforms using idealised neural network models and data sets, *e.g.*, training a ResNet model [10] using the ImageNet dataset [2].

On the other hand, many researchers use AI-based technologies, mostly DL and Reinforcement Learning (RL), to transform how computer systems and chips are designed and opti-

[1] Universitat Politècnica de Catalunya, Catalunya, Spain
[2] Gobierno del Estado de Jalisco, Jalisco, México
[3] Barcelona Supercomputing Center, Catalunya, Spain

mised. Many core problems in systems and hardware design are combinatorial optimisation tasks [3, 5, 7]. AI development has been tightly interlinked with progress in chip design. This cooperation may speed up the achievements in both fields and, it is evident, in terms of scientific development, in general. Today, AI methods are available to solve various complex problems in the design and development of the HPC systems, such as predicting running times, resource utilisation, optimisation of load balancing, job scheduling, resource discovery, and process migration. Recent accomplishments of this program brought the idea of having a selection of those achievements under the official name *Advances on Parallel and High-Performance Computing for Artificial Intelligence.*

## Exploring the Selection

The present selection is meant to provide a snapshot of some of the latest work done in the confluence of AI and HPC. The collection we have made covers a small but illustrative choice of papers that clearly shows the importance of HPC in developing faster and powerful applications of Al. The papers in this special issue address the following topics:

- **Benchmarking**. Recent years witness a trend of applying large-scale distributed deep learning algorithms (HPC & AI) in both industry and scientific computing areas, which aim to speed up the training time to achieve a state-of-the-art quality. The HPC & AI benchmarks accelerate the process. Benchmarking HPC AI systems at scale raise serious challenges. The paper by Parés-Pont et al. touches on this topic.
- **Deep Learning Applications**. HPC & AI give the power to deal with ever growing complex models. More accurate predictions need to be fed by large amounts of data and they require large computing power, in return those systems create better applications. The paper by Manero and Béjar brings an application that may impact on the production of green energy.
- **Training Deep Neural Networks**. The paper by Torres et al. touches on the performed characterisation using a convolutional neural network implemented in TensorFlow and Pytorch. Likewise, the behaviour of the component interactions is discussed by varying the batch size for two sets of synthetic data.
- **An Exascale platform and AI**. The exascale platforms offer unique challenges to enabling functional hardware and software toolchains to manage vast volumes of data. In this regard, in their contribution, Fell et al. introduce the vision of MareNostrum Experimental Exascale Platform (MEEP), an open-source platform enabling HPC experimentation. One of the most exciting topics of this paper is the Accelerated Compute and Memory Engine (ACME) proposal. This tool facilitates the exploration of separating the computation from the memory operations and optimising the accelerator for dense (compute-bound) and sparse (memory-bandwidth bound) workloads that are very useful to AI workflows.

## Conclusions

HPC and, in general, Parallel and Distributed Computing has become a pervasive utility, from supercomputer facilities and server farms containing multicore CPUs, GPUs and TensorFlow, to individual PCs, laptops, and new powerful mobile devices. AI research and industrial application heavily depend on parallel processing.

Today, as AI & HPC continue to transform an ever-increasing number of scientific disciplines and successful industrial applications at an expeditious pace, we can only imagine what the future holds once AI is powered with a rigorous mathematical framework. This is a promising innovation space for developers and hardware engineers to build the application, compiler, libraries and the necessary hardware to solve future challenging endeavours in the HPC, AI, ML, and DL domains.

The fast confluence of AI & HPC gives the means to address science, engineering, and industry challenges. It enables the creation of disruptive approaches for data-driven discovery and innovation. Realising these goals demands a combined effort between AI practitioners, HPC and specific domain experts.

## Acknowledgements

The editors of this special issue would like to acknowledge and thank the external reviewers. They carried out a very professional, scientific and altruistic task in very short times and with very short deadlines.
Prof. U. Cortés (SNI-III) and Dr. U. Moya (SNI-I) are members of the Sistema Nacional de Investigadores, CONACyT (México).

## References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)

2. Deng, J., Dong, W., Socher, R., et al.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, 20-25 June 2009, Miami, FL, USA. pp. 248–255. IEEE (2009), DOI: 10.1109/cvprw.2009.5206848

3. Goldie, A., Mirhoseini, A.: Reinforcement Learning for Placement Optimization. In: Proceedings of the 2021 International Symposium on Physical Design, 22-24 March 2021, Virtual Event, USA. pp. 5–5 (2021), DOI: 10.1145/3439706.3446883

4. Jimenez, D., Lin, C.: Dynamic branch prediction with perceptrons. In: Proceedings HPCA Seventh International Symposium on High-Performance Computer Architecture, 19-24 Jan. 2001, Monterrey, Mexico. pp. 197–206. IEEE (2001), DOI: 10.1109/HPCA.2001.903263

5. Khailany, B., Ren, H., Dai, S., et al.: Accelerating chip design with machine learning. IEEE Micro 40(6), 23–32 (2020), DOI: 10.1109/mm.2020.3026231

6. Minsky, M., Papert, S.A.: Perceptrons: An introduction to computational geometry. MIT press (2017), DOI: 10.7551/mitpress/11301.001.0001

7. Nemirovsky, D., Arkose, T., Markovic, N., et al.: A general guide to applying machine learning to computer architecture. Supercomput. Front. Innov. 5(1), 95–115 (2018), DOI: 10.14529/jsfi180106

8. Sejnowski, T.J.: The unreasonable effectiveness of Deep Learning in Artificial Intelligence. In: Proceedings of the National Academy of Sciences. vol. 117, pp. 30033–30038. National Acad. Sciences (2020), DOI: 10.1073/pnas.1907373117

9. Waldrop, M.M.: Artificial intelligence in parallel. Science 225(4662), 608–610 (1984), DOI: 10.1126/science.225.4662.608

10. Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: Revisiting the ResNet model for visual recognition. Pattern Recognition 90, 119–133 (2019), DOI: 10.1016/j.patcog.2019.01.006