Hardware Architectures for processing Artificial Intelligence and Machine Learning processes

Prof. Asadinia

By
Isaiah Martinez
Summer Shin

December 16 2024

1 Traditional Architecture Artificial Intelligence Chip Technology

This publication is authored by Qinze Jiang and Jiajun Zhan who presented their work at the 2nd International Conference on Computer Science and Management Technology (ICCSMT) held in 2021.

The main goal is this paper is to try to compare different hardware architectures for handling AI and ML tasks and looking at various ways to improve performance for each. There are a total of 3 different types of architectures discussed:

- GPUs
- FPGAs
- ASIC Chips

These different technologies all have different purposes for being used. GPUs are the classical solution to handling ML tasks. They are capable of handling the complex tensor computations due to their many simplistic designed cells (processing units). This simplistic design is contrasted by that of the FPGA, which is more complex and allows them to be programmed via logical blocks (gates) and can be configured before usage. This enables them to be more efficient than the classical GPU which is more general purpose. An alternative to both would be the ASIC chips which are specifically designed to the purpose instead of being programmed for it. By instead designing the hardware for the particular task for AI/ML processing, this can further improve performance.

To improve each of these technologies, the authors propose a series of solutions. For GPUs, they suggest combining their operation via more sophisticated software such as Kubernetes and Docker. For FPGAs, they suggest to use FPGA Accelerators to improve performance for a particular type of ML model. For ASIC chips, they suggest a design using Word Lines and particular charges on each side of the line for maintaining a highly optimized form of calculating Matrix vector multiplication.

The main weaknesses of this paper are that it contains some false information, as Multiple GPUs don't need to be connected via Docker/Kubernetes, and that it focuses primarily on only Neural Networks. Perhaps other ML algorithms may see some improvement so these technologies would be used more regularly. This paper does excel at providing several viable implementations of these various technologies to improve performance.

2 Paper 2

paper 2 stuff...

3 Paper 3

paper 3 stuff...

4 Paper 4

paper 4 stuff...

References

[1] Baker, N. 1966, in Stellar Evolution, ed. R. F. Stein & A. G. W. Cameron (Plenum, New York) 333

[2] Balluch, M. 1988, A&A, 200, 58

[3] Cox, J. P. 1980, Theory of Stellar Pulsation (Princeton University Press, Princeton) 165

[4] Cox, A. N.,& Stewart, J. N. 1969, Academia Nauk, Scientific Information 15, 1

[5] Mizuno H. 1980, Prog. Theor. Phys., 64, 544

[6] Tscharnuter W. M. 1987, A&A, 188, 55

[7] Terlevich, R. 1992, in ASP Conf. Ser. 31, Relationships between Active Galactic Nuclei and Starburst Galaxies, ed. A. V. Filippenko, 13

[8] Yorke, H. W. 1980a, A&A, 86, 286

[9] Zheng, W., Davidsen, A. F., Tytler, D. & Kriss, G. A. 1997, preprint