ISSN: 2320 – 8791 (Impact Factor: 2.317)
www.ijreat.org

VLSI Technology in Design of Future Hardware for Machine Learning: A Review

[1]Modalavalasa Hari Krishna, Dept. ECE, JNTUH, Hyderabad, India, mhkrishna9s@gmail.com
[2]Dr. Makkena Madhavi Latha, Dept. ECE, JNTUH, Hyderabad, India, mmadhavilatha@jntuh.ac.in

***Abstract*—Machine Learning (ML) and Very Large-Scale Integration (VLSI) technology are the two major and emerging research fields. Machine Learning consumes infinite amount of data, understands the hidden structures and provides automated regression or classification performance accurately. VLSI technology creates low power and high-speed Integrated Circuits (ICs) by properly placing millions of MOS transistors on a single chip. Many Machine Learning algorithms are developed to accelerate the VLSI design procedure by automating various design phases of VLSI chips. Now-a-days VLSI technology accelerating the Machine Learning algorithms' training as well as validation procedures. Previously Machine Learning algorithms are trained and validated on general processors but now-a-days advanced sensors easily generating huge training datasets with improved digital technology. Processors consume huge amount of development time for training and validation of algorithms with big datasets. Graphics Processing Units (GPUs) can reduce the training and validation times with few hundreds of cores but not enough to handle the training datasets in size of many Tera-bytes. VLSI technology can handle this problem with Programmable Logics devices (PLDs) and Application-Specific Integrated Circuits (ASICs). In this paper, we consider different existing hardware units for training and validation of ML algorithms with huge datasets and made a comparative analysis to exhibits the importance of VLSI technology in future hardware design for Machine Learning.**

I. Introduction

Artificial Intelligence (AI) change the technology to new era by providing decision making intelligence to machines. AI has its applications in all scientific fields. Machine Learning (ML) is the heart of AI, which deals with core decision making, classification or regression. ML algorithms easily consume huge amount of data and extract the hidden relations or structures inside data as features. ML algorithms create ML models by training it with extracted features [1]. These models will be used for future inference of unknown data. Deep Learning (DL) is the major category of ML algorithms with Neural Network structures. DL models has highly complex architectures with multiple hidden layers and need more computational resources for training and validation [1],[2].

The advanced digital technology developing new generation sensors, which are accurate, flexible, portable and cost-effective. For example, smart phone has many advanced sensors like high resolution cameras, Gyroscope, Proximity sensor, Ambient Light Sensor, Microphone etc. With the help of Advanced sensors people generating profusion of data every day and creating datasets of sizes from few Megabytes to few Terabytes. Generic computational resources are not capable to process this much huge data and need few years of time to train and validate ML or DL models [3]. To overcome this problem and reduce the training and validation times, many advanced hardware resources are developed with high degree of parallelism. Central Processing Units (CPUs) with multiple cores and special processing add-ons [4],[5],[6], Graphics Processing Units (GPUs) [4],[7],[8], Tensors Cores and AI engines [4], [9] are the few existing advanced resources for ML and DL applications. Even though these resources provide greater parallelism, still the model training and validation taking few weeks to few months for bigger datasets [10].

II. Procedure

In this paper, we have considered different available datasets for machine learning and deep learning applications to understand the future requirements of resources like memory, computational powers. Here different computational resources like CPUs, GPUs, FPGAs and ASICs are considered to perform comparative analysis if different aspects with respect to constraints related to ML algorithms and training and validation datasets. In each category, different industry level models are considered and compared to understand currently available and to estimate the future projections of requirements of memory and computational resources. Finally, we create comparative analysis on all analyzed resources

III. Comparative Analysis and Discussions

Huge storage memory and computational resources are the two major constraints of future Machine learning and deep learning algorithms. As the technology advances, people easily generating different datasets in sizes of few Gigabytes to Terabytes. Criteo data is the one of the biggest datasets with more than 4 billion rows of data and requires more than 1 Terabyte of storage. Open Images dataset has millions of images with size more than 500 Gigabytes. Million Song Dataset, Image-net, Free Music Archive (FMA) are few examples of available datasets with size more than 100 Gigabytes. Fig.1 shows the sizes of different available datasets. The increased dataset size also increases the accuracy of the machine learning model by reducing the overfit. From Fig.1 we can project the size of future datasets in terms of few Terabytes with advanced image and data acquisition sensors. Training and validation of ML models with these gigantic datasets require huge memory and computational resources. Microprocessors or CPUs are the

earliest computational resources and optimized for sequential programming with limited parallelism. With advanced VLSI technology, two or more cores are fabricated into single CPU to increase the computational performance and throughput. Intel, IBM, AMD and Snapdragon are the few real time

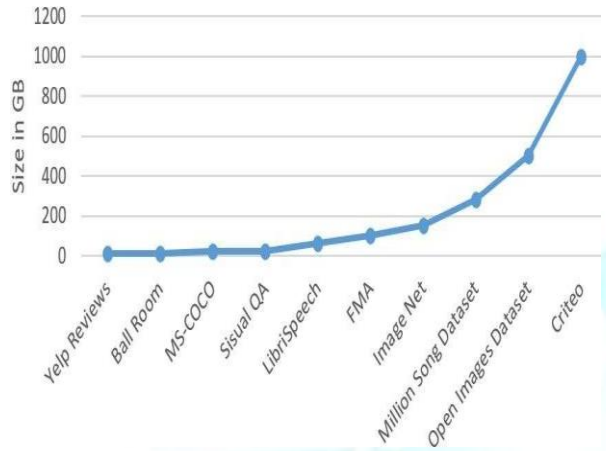

Fig.1. Increase in size of ML or DL Training datasets.

industry level CPUs available in the market.

Fig.2 show the comparison of different versions of the advanced Intel processors like core-i3, core-i5, core-i7 and core-i9. Core-i3 have 4 cores and supports 4 treads with a maximum frequency of 3.6 Giga Hertz. The advanced Intel core-i9 processors have up to 8 cores and support 16 threads with a maximum frequency of 5 Giga Hertz. CPUs have many advantages like Versatility, multitasking, ease of programming but not suitable to train models with large datasets due to limited cores and threads. Another disadvantage is huge overhead added by OS capability [4],[11],[12].

GPUs are the optimized for massively parallel processing with huge number of cores. These are Configurable for specific application and can be changed after installation. GPUs are highly suitable for machine learning algorithms

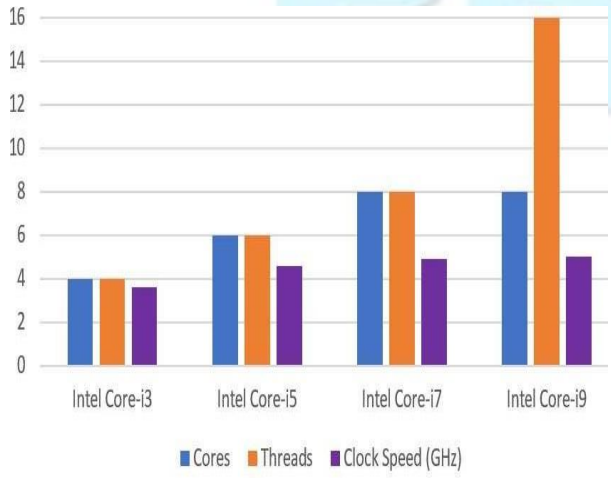

Fig.2. Comparison of Intel CPUs in terms of Cores, Threads and Clock Frequency

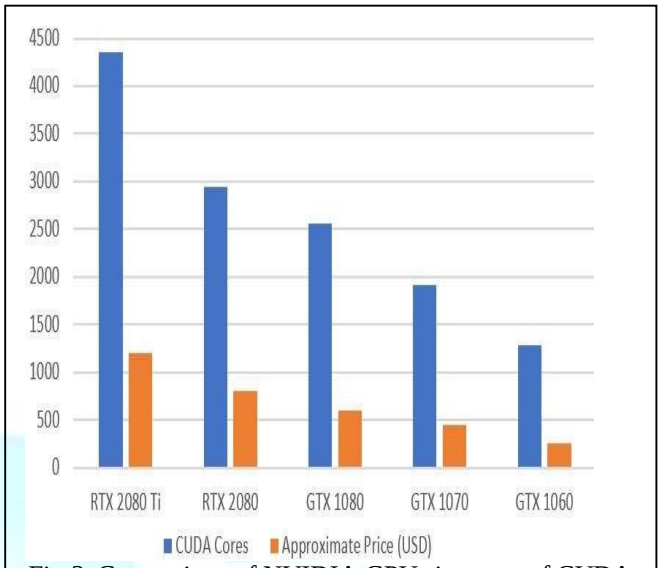

Fig.3. Comparison of NVIDIA GPUs in terms of CUDA cores and Price (Approximately on Jan-2020)

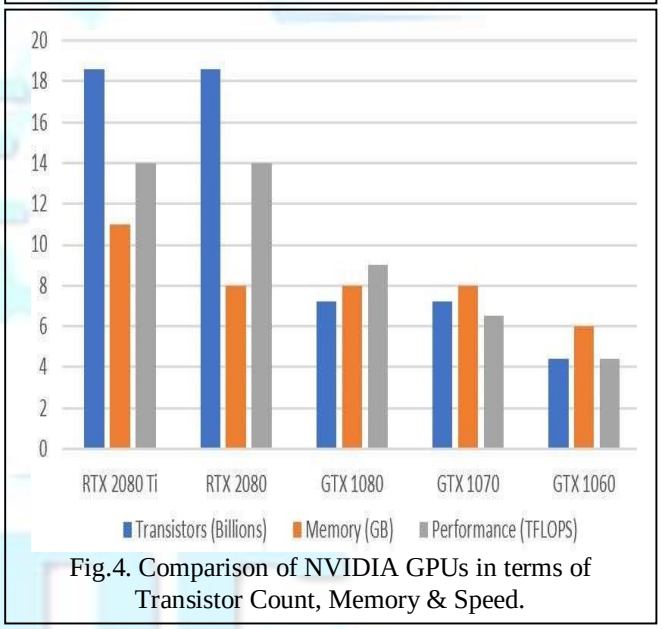

Fig.4. Comparison of NVIDIA GPUs in terms of Transistor Count, Memory & Speed.

which require massive parallelism and also provides lower power consumption compared to CPUs. Many advanced GPUs are developed with few hundreds of cores. NVIDIA GEFORCE and AMD RADEON GPUs are some of best GPUS available in market. Fig.3 shows number of cores and price of some advanced NVIDIA GPUs. GTX-1060 has 1280 cores where as RTX-2080 has 4352 cores but the price also increased from 200 USD to 1200 USD [4],[7].

With VLSI technology, the density of the GPU cores increasing constantly. Fig.4 shows the number of transistors integrated into each GPU. From GTX-1060 to RTX-2080 Ti , the number of transistors increased from 4.4 billion to 18.6 billion and also the computational power also increased from 4.4 TFLOPS to 14 FLOPS. GTX series has up to 8 GB of DDR5 RAM where as RTX supports up to 11 GB of DDR6 RAM. Both base clock and boost clock are maintaining gradual improvement as shown in Fig.5 to maintain the stability of the GPU as the number of cores increasing rapidly.

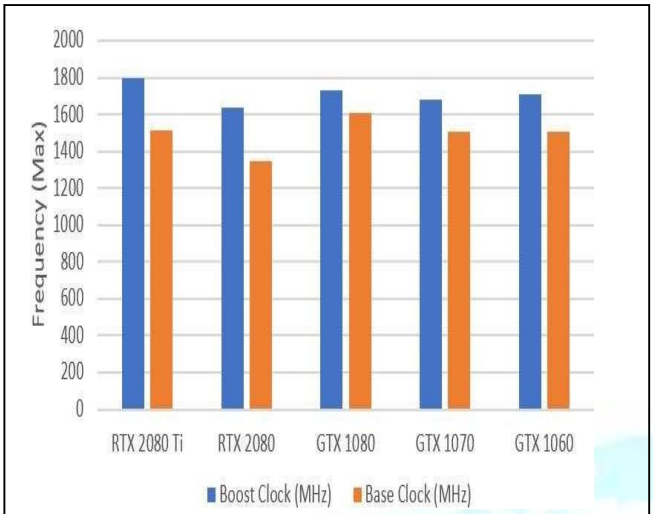

Fig.5. Comparison of NVIDIA GPUs in terms of Clock speed or Frequency.

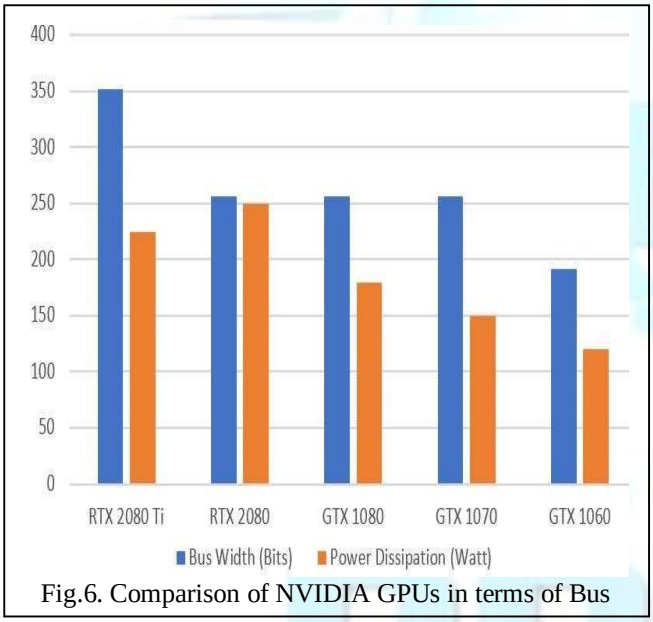

Fig.6. Comparison of NVIDIA GPUs in terms of Bus

One more advantage of the GPUs is relatively a greater number of inputs and outputs (IOs)compared to CPUs. RTX-2080 Ti support up to 350 IOs and consume around 225W of power on full load on 4352 cores. Even though the 225W is huge power but at 14 TFLOPS throughput, the power dissipation per computation is very less.

Even though GPUs provides low power high speed parallel computations they are not meeting the computational requirements of advanced Machine Learning and Deep Learning algorithms with huge datasets. Limited IOs and bandwidths are also limiting the applications of GPUs. Few other drawbacks of GPUs are relatively difficult to programmability, poor floating-point operational speed, relatively longer development time and poor performance and high-power consumption for sequential operations.

CPUs and GPUs have fixed hardware architecture and perform wide range of generic applications but they are not suitable for high end scientific applications like video processing, image processing and signal processing, which require dedicated computational resources.

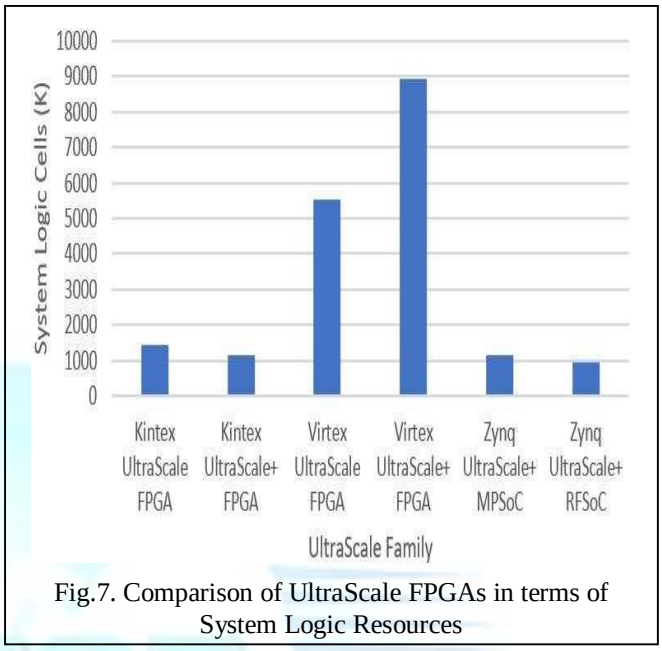

Fig.7. Comparison of UltraScale FPGAs in terms of System Logic Resources

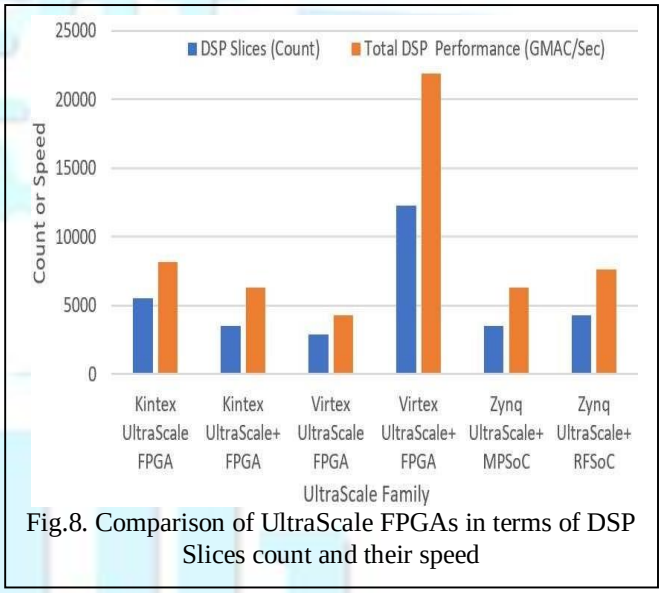

Fig.8. Comparison of UltraScale FPGAs in terms of DSP Slices count and their speed

VLSI technology solving all above discussed problems with advanced FPGA and ASIC architectures. FPGAs have huge number of computational resources like DSPs, Logic cells and IOs. Xilinx and Altera FPGAs are very popular in market and have around 90% market share. With wide range of FPGAs from low and Spartan to high end Virtex FPGAs, Xilinx have 52% of market share whereas Altera have 34% of market share from low end Cyclone to high end Stratix FPGA families. Fig.7 show the availability of logic cells in different high-end Xilinx FPGAs. Virtex UltraScale+ FPGA has around 9000K system logic cells which represents its capability to provide highest parallel processing pool. It also has the highest number of DSP slices (around 12300) as shown in the fig.8. UltraScale+ family not only best in DSP resource size but in performance also. It can perform up to 21897 Giga MAC operations per second [4],[10],[13],[14].

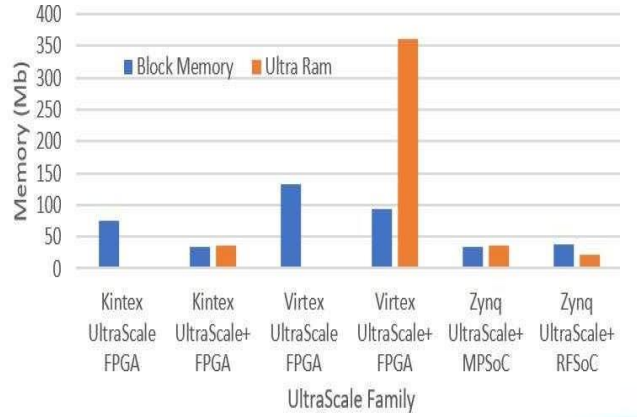

Fig.9. Comparison of UltraScale FPGAs in terms of Memory (Block & Ultra RAM)

Fig.9 shows the richness of Virtex UltraScale+ family in terms of Block memory and ultra-RAM. It has 95 MB of Block memory and 360 MB of ultra-RAM. It also provides greatest serial connectivity with wide bandwidth up to 8384 GB per second and also best memory interface in range with speed up to 2666 MB per second [15],[16].

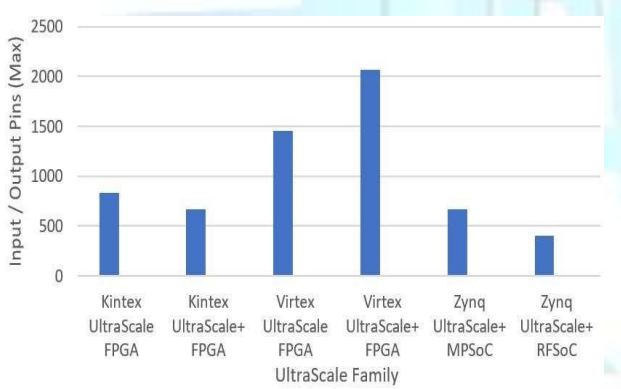

Fig.10. Comparison of UltraScale FPGAs in terms of Input Output Pins (IOs)

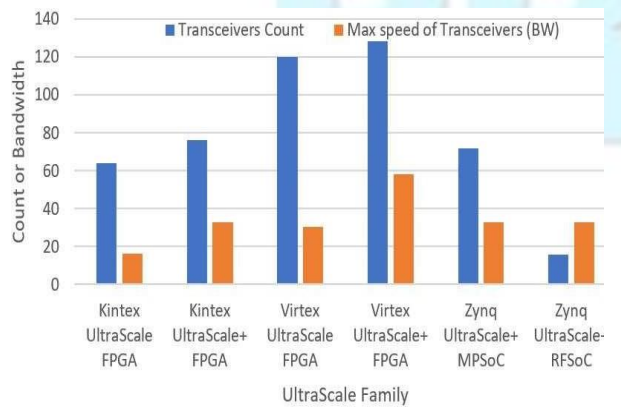

Fig.11. Comparison of UltraScale FPGAs in terms of Transceivers count and their speed

The major advantage of the FPGAs is a greater number of IOs. Virtex UltraScale+ family has more than 2000 IOs. These many number of inputs allow parallel loading of huge datasets as small chunks. The feature increases the applications of FPGAs in ML and Deep Learning areas. Fig.10 shows the number of available IOs in high-end Xilinx FPGAs. UltraScale+ family also provides largest number of transceivers with highest speed up to 58 GB per second. With these high-end features, FPGAs are providing many optimized computational solutions to train and validate advanced Machine Learning and Depp Learning applications with bigger datasets. The main problems with FPGAs are their higher power consumption and also, they need re-formulation of design to utilize the great extent of parallelism provided by these FPGAs.

ASICs exhibits the real fruitfulness of VLSI technology with their low power-area architectures with smallest delays to provide highest performance. The major problems with ASICs are their largest design cycle and development cost. Bur for long-term real time applications like dedicated computational servers, cloud services these ASICs provide infinite flexibility with best feature set. NVIDIA Tensor core, ML cores, Snapdragon AI Engines (AIE) are the few existing examples of ASICs [9],[17].

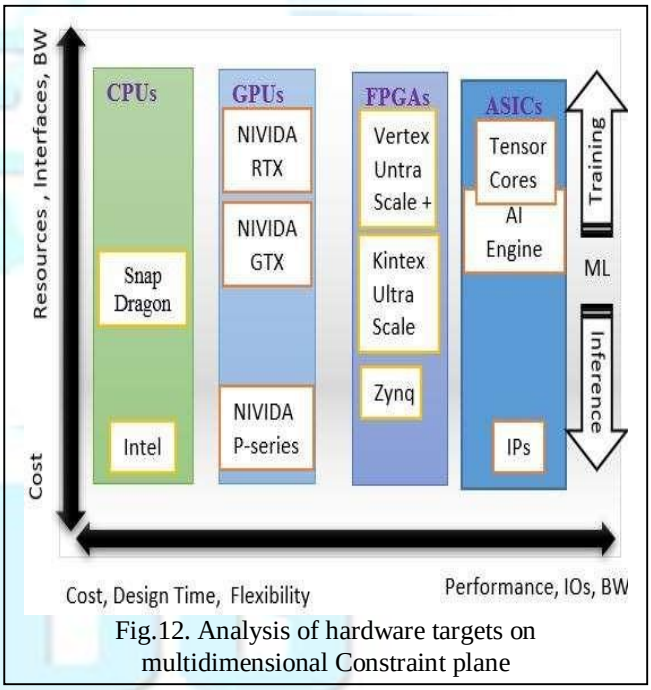

Fig.12. Analysis of hardware targets on multidimensional Constraint plane

All the computational resources discussed here are the VLSI products but CPUs and GPUs have fixed hardware architectures and provides zero flexibility for optimization but have lower design time. FPGAs provides semi-custom designs with high speed and bandwidth (BW) but power consumption is high. Fig.12 shows the placement and fitness of various hardware resources on multi-dimensional constraint plane. One should properly select target hardware depending on their application, dataset and other design constraints. In generic, FPGAs are the optimized solutions for research and development of ML or DL models and ASICs are the future hardware for low power, low area and high-performance Machine Learning and Deep Learning applications.

IV. CONCLUSION

Training and Validation of Machine Learning and Deep Learning applications with bigger datasets is very complex and time-consuming procedure. CPUs has limited cores, parallelism and suitable for sequential applications. GPUs have few hundreds of cores and provides high degree of parallelism but has higher development cycle. GPUs well suited to train ML models with parallel datasets. The sequential and floating-point performance of GPUs are poor. FPGAs and ASICs are the two future hardware solutions with advanced VLSI technology. FPGAs are optimum targets for ML and DL applications during research and Development phase with ocean of semi-fabricated resources like DSP slices. ASICs are the Area-Power efficient and highest-performance resources for real time high-end applications like Computing Clusters and Cloud services.

REFERENCES

[1] S. Ray, "A Quick Review of Machine Learning Algorithms," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 35-39, doi: 10.1109/COMITCon.2019.8862451.

[2] Nan Zheng; Pinaki Mazumder, "Fundamentals and Learning of Artificial Neural Networks," in Learning in Energy-Efficient Neuromorphic Computing: Algorithm and Architecture Co-Design , IEEE, 2020, pp.11-60, doi: 10.1002/9781119507369.ch2.

[3] M. K. K. Leung, A. Delong, B. Alipanahi and B. J. Frey, "Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets," in Proceedings of the IEEE, vol. 104, no. 1, pp. 176-197, Jan. 2016, doi: 10.1109/JPROC.2015.2494198.

[4] Pooja Jawandhiya, "HARDWARE DESIGN FOR MACHINE LEARNING", in International Journal of Artificial Intelligence and Applications (IJAIA), Vol.9, No.1, January 2018 DOI : 10.5121/ijaia.2018.9105 63.

[5] S. M. Tam et al., "SkyLake-SP: A 14nm 28-Core xeon® processor," 2018 IEEE International Solid - State Circuits Conference - (ISSCC), San Francisco, CA, 2018, pp. 34-36, doi: 10.1109/ISSCC.2018.8310170.

[6] O. Lempel, "2nd Generation Intel® Core Processor Family: Intel® Core i7, i5 and i3," 2011 IEEE Hot Chips 23 Symposium (HCS), Stanford, CA, 2011, pp. 1-48, doi: 10.1109/HOTCHIPS.2011.7477509.

[7] S. Jean-Paul, T. Elseify, I. Obeid and J. Picone, "Issues in the Reproducibility of Deep Learning Results," 2019 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), Philadelphia, PA, USA, 2019, pp. 1-4, doi: 10.1109/SPMB47826.2019.9037840.

[8] M. A. Raihan, N. Goli and T. M. Aamodt, "Modeling Deep Learning Accelerator Enabled GPUs," 2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Madison, WI, USA, 2019, pp. 79-92, doi: 10.1109/ISPASS.2019.00016.

[9] van Lent, Michael, and John Laird. "Developing an artificial intelligence engine." In Proceedings of the game developers Conference, pp. 16-18. 1999.

[10] Talib, Manar Abu et al. "A systematic literature review on hardware implementation of artificial intelligence algorithms." The Journal of Supercomputing,2020,Pp. 1 - 42.

[11] Lee, Taek-Soo, Jingyan Xu and Benjamin M. W. Tsui. "Development of transfer learning datasets using realistic simulation of myocardial perfusion SPECT images for a deep learning model." The Journal of Nuclear Medicine 60 -2019.

[12] T. Lehinevych and H. Andii, "Analysis of Deep Metric Learning Approaches," 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT), Kyiv, Ukraine, 2019, pp. 369-372, doi: 10.1109/ATIT49449.2019.9030440.

[13] Sulaiman, Nasri & Obaid, Zeyad & Marhaban, Mohammad Hamiruce & Hamidon, Mohd Niza, 'Design and Implementation of FPGA-Based Systems -A Review', in. Australian Journal of Basic and Applied Sciences. 3.

[14] H. Wang and C. Choy, "Hardware acceleration of support vector machine based on high level synthesis," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, 2018, pp. 956-959, doi: 10.1109/IEMCON.2018.8614917.

[15] I. Ganusov, H. Fraisse, A. Ng, R. T. Possignolo and S. Das, "Automated extra pipeline analysis of applications mapped to Xilinx UltraScale+ FPGAs," 2016 26th International Conference on Field Programmable Logic and Applications (FPL), Lausanne, 2016, pp. 1-10, doi: 10.1109/FPL.2016.7577344.

[16] V. Boppana, S. Ahmad, I. Ganusov, V. Kathail, V. Rajagopalan and R. Wittig, "UltraScale+ MPSoC and FPGA families," 2015 IEEE Hot Chips 27 Symposium (HCS), Cupertino, CA, 2015, pp. 1-37, doi: 10.1109/HOTCHIPS.2015.7477457.

[17] M. K. Bhatti, B. Nawaz and A. M. Soomro, "Design & Analysis of Asynchronous (Clockless) Circuits and Implementation using Mentor Graphics ASIC Design Tools," 2019 2nd International Conference on Communication, Computing and Digital systems (C-CODE), Islamabad, Pakistan, 2019, pp. 243-246, doi: 10.1109/C-CODE.2019.8680988.