A Survey Comparing Specialized Hardware And Evolution In TPUs For Neural Networks

Amna Shahid
Department of Computer Engineering,
NUST College of Electrical & Mechanical Engineering, Rawalpindi
Email: amnashahid901@gmail.com

Malaika Mushtaq
Department of Computer Engineering,
NUST College of Electrical & Mechanical Engineering, Rawalpindi
Email: malaikamushtaq00@gmail.com

2020 IEEE 23rd International Multitopic Conference (INMIC) | 978-1-7281-9893-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/INMIC50486.2020.9318136

***Abstract*—This survey paper is based on the evolution of TPUs from first generation TPUs to edge TPUs and their architectures. This paper compares CPUs, GPUs, FPGAs and TPUs, their hardware architectures, their similarities and differences will be discussed. Modern neural networks are immensely used these days but they require more time, computation and energy. Due to the greater demand and attractive options for architects to explore, companies are continuously working to reduce training and inference response time. Due to the demands and cost factors different kinds of ASICs (application specific integrated circuits) are developed and research is increased in this area. Many models of CPUs, GPUs and TPUs have been developed to support these networks and to improve training and inference phase. Intel developed CPUs for this purpose, NVIDIA developed GPUs and Google developed cloud TPUs. The hardware of CPUs and GPUs can be sold to businesses while Google offers TPU processing for everyone from the cloud. When the data is away from the computational source, it increases the overall cost and to reduce this cost companies implements memory management and caching techniques close to ALUs.**

***Keywords*—** GPU, TPU, Neural Networks, Deep Learning

1. Introduction

Artificial Intelligence (AI) make machines smart enough to automatically learn without human backing and perform actions accordingly. The major application of AI is machine learning (ML). The aim of ML is to make computers smart enough that they can access data and use to determine that what actions to be performed in any specific situation. For making a computers smart, AI is an umbrella term which consist of machine learning and deep learning (DL). ML is a subset of AI and DL is a subset of ML, all of these are nested within each other as shown in figure 1. Workhorses to achieve the goal of automation are Neural Networks (NN).

To turn input to a correct output using mathematical manipulation Deep Neural Networks (DNN) are used, which consist of multi layers between the input and output layers [1]. DNN helps in increasing the accuracy of image identification [2] and reduction of speech recognition error [3]. This accuracy is achieved at the cost of more time, more computation and more energy. These factors are to be handled properly at *'inference'*, Inference is a term used when a trained system is making predictions. Response time and throughput are also considered in inference [4] [5]. There are three major factors which increases the computation in neural networks. First, the neural networks gets deeper and involve many hidden layers. Second, the input size is increased mostly when images are being processed. Third, a large amount of data is used while training the model.

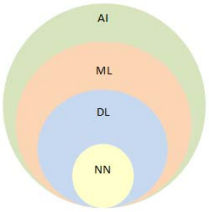

Figure 1. Relation of AI, ML, DL, NN.

Due to these costs and requirements, potential amount of researchers are working on specialized hardware and development of applications, which consist of specific integrated circuits. The main motive to design specialized hardware is to reduce the response time during inference.

Central Processing Units (CPUs) have been complemented by the computer systems with special purpose accelerators to accomplish different tasks, such CPUs are called co-processors. In late nineties for neural network accelerators, digital signal processors were used for optical character recognition [6]. In 1990s, workstations for parallel high-throughput systems were created that targets various applications including AI [7]. For training and inference accelerators based on Field Programmable Gate Arrays (FPGAs) were also designed [8], [9]. An accelerator neuronet Complementary Metal Oxide Semiconductor CMOS named ANNA was developed by Yann LeCun [10]. In heterogeneous computers more than one specialized processors are incorporated on a single chip, such as cell microprocessor [11]. AI accelerators and heterogeneous microprocessors have some overlapping features such as prioritization of throughput over latency, low precision arithmetic support and architecture of data flow. AI was one of the subsequent application of cell processors [12], [13], [14].

This survey focuses on the specialized hardware that are used for ML and DNN. For the domain of AI, understanding the relative advantages of these technologies are must and the constraints such as weight, size and power should be considered. The other factors include the memory bandwidth for the model parameters which are used for loading and updating the data.

. The rest of this paper is organized as follows. In Section II, we have discussed the specified hardware which are used for AI machines, data is gathered from publicly available materials including technical trade press, research papers and company benchmarks etc. We have compared different processors and accelerators on the basis of throughput and performance metrics along with the power consumption in section III. Section IV presents the architecture and performance metrics of TPUs and explains how TPUs are the best processors for Machine Learning and Deep Neural Networks. The conclusion of this paper is given in section V.

2. Specialized Hardware

The advent of Neural Networks has increased the input parameters which ultimately effects the computational cost of Machine Learning. With this advent of technology, the need of more specialized hardware increases and to address this, a lot of work has been done on the design of processors and integrated circuits. The hardware used for ML ranges from a single core chip to multi-core neural processing systems. All specialized processors are different from each other on the basis of hardware architecture and on-chip parallel processing. Public clouds like Microsoft azure, Amazon AWS etc. are used to run ML algorithms. Nowadays, five different processors are used for Machine Learning.

2.1. Central Processing Unit

The integral part of a computer system that performs computations is known as Central Processing Unit. Arithmetic Logic Unit (ALU), Control Unit (CU) and Memory Unit (MU) are the majors components of a CPU. All arithmetic and logical operations in a computer system are performed by the ALU. The control functions related to data bus, control bus etc. are controlled by CU. MU consist of registers, caches, hard drives, RAMs, ROMs and other storage devices. When an ALU request for some values loading, retrieving and all such operations are performed by MU, it took necessary steps to serve such requests in the CPU [15]. Figure 2 shows the architecture of a CPU. CPUs works on the principle of fetch and execute. CPUs revolutionized from transistor CPUs to small-scale integrated CPU, large-scale integrated CPU and microprocessors.

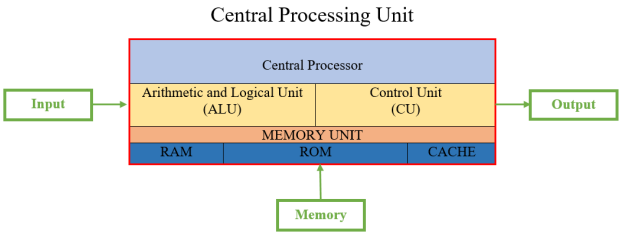

Figure 2. Central Processing Unit.

2.2. Graphic Processing Unit

A graphic Processing Unit is a specially designed circuit, intended for quality output on the display devices. These are the most efficient processors to manipulate the image processing and computer graphics. GPUs have the ability to perform parallel processing, this feature makes them more efficient from the generally used CPUs. GPUs can be present on two locations in a die, it can be embedded on the motherboard or it can be present on the video card of the computer [16]. For the first time in 1994, Toshiba designed GPUs for Sony play stations named as Sony GPU [17]. In 1999, GPUs got hype when NVIDIA developed the first single-chip processor [18], [19]. ATI technologies coin the term GPU as Virtual Processing Unit (VPU) in 2002 with the development of Radeon9700 [20]. The emergence of Deep Learning has increases the importance of GPUs because they can work 250 times faster than the CPUs and work much better while training a neural network. Figure 3 shows the architecture of a GPU.

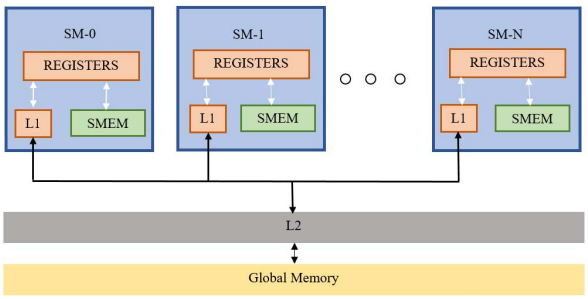

Figure 3. Graphic Processing Unit.

2.3. Field Programmable Gate Array

Field Programmable Gate Array (FPGA) are integrated circuits, which are used to be configured after manufacturing according to the customers requirements and needs. The language used to program the FPGA hardware is Hardware Define Language (HDL). FPGAs performance in accordance to computation is less than the GPUs but their performance related to energy consumption is better than the GPUs. Fundamental blocks of FPGAs are build with RAM blocks and ample amount of logic gates [21]. Analog features are provided by the FPGAs in processing due to which low latency than CPU is achieved. Software are used to design the instruction based architecture and hardware circuits are used to configure the FPGAs. Figure 4 shows the architecture of a FPGA.

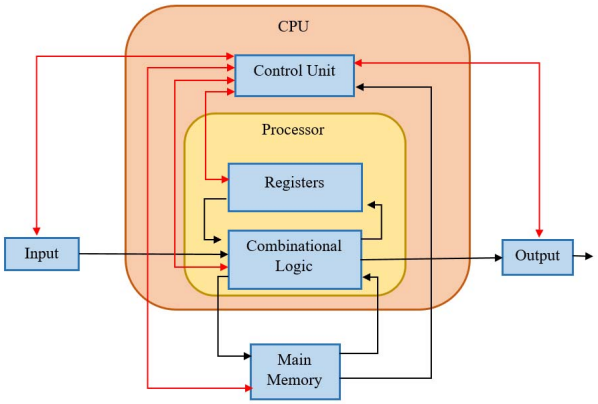

Figure 4. Field Programmable Gate Array.

2.4. Application Specific Integrated Circuit

Application Specific Integrated Circuit (ASIC) is basically an integrated circuit (IC) which can be customized for any specific use, it is not intended for any general purpose. For example ASICs are designed for machine learning, deep learning or data mining etc [22]. CPUs, GPUs and FPGAs have much more flexibility than ASICs. There are three types of ASICs: Full-customed ASIC, Semi-customed ASIC and Platform ASIC. For designing an application for a specific purpose full-custom ASICs are used. Semi-customs ASICs allows to do a few customisation. Designs in which the design duration is to be decreased and the cost utilization factor is to be increased, in such cases Platform ASICs are used. Figure 5 shows the architecture of an ASIC.

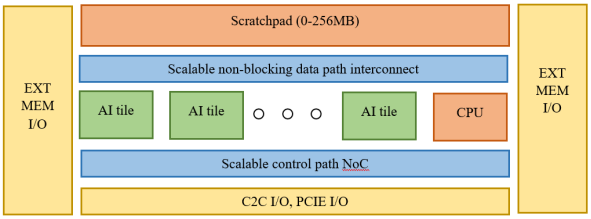

Figure 5. Graphic Processing Unit.

2.5. Tensor Processing Unit

Tensor Processing Unit (TPU) is the recent development in ASICs by google, mainly designed for deep learning inference and training. These are the processors which are used in the data centers of Google Cloud to handle enormous amount of data [23]. These specifically designed chips use a math library named as tensorflow framework, which is mainly used for deep learning and neural networks. Only some models of Google TPUs are available commercially because google TPUs are proprietary. These chips are designed to work on more input/output operations per joules with a large volume of low precision computing. Figure 6 shows the architecture of an TPU.

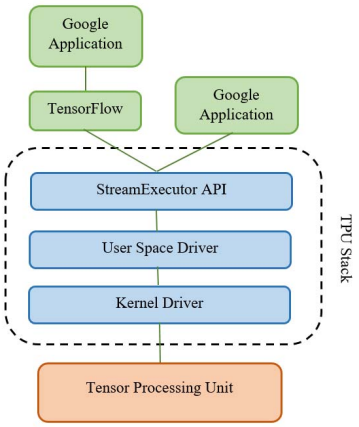

Figure 6. Tensor Processing Unit.

3. Comparison of Processors

In the previous section we have discussed different specialized hardware used for Machine Learning and Neural Networks. We have seen the hardware implementation of each processor chip. In this section we will consider that which processor is most efficient to work on Neural Networks and which has the ability to process huge amount of data. During the survey, it is surprisingly seen that due to the software toolbox CPU approach works very well and can optimize other frameworks, which are commonly used. Each Processor chip has its own specifications. These specialized hardware has decreased the cost of training the Deep Learning models. Table 1 shows the specifications of Tensor Processing Unit(GPU) (TPU), Graphic Processing Unit and Central Processing Unit (CPU). Table 2 and 3 shows the similarities and differences in the processors chips according to their specifications. For each model, the requirements decide which hardware is ultimately best.

4. Case Study- TPU Architectures

Tensor processing unit is google designed ASIC optimized to run Machine learning. It is proved to be more efficient as compared to GPUs, CPUs and FPGAs. Till 2006, GPUs and FPGAs were believed to be sufficient to fulfill demands. But huge computing hardware was required after discovery of neural networks. Deep Neural Networks(DNNs) would twofold our data centres, calculation demands which would be very costly utilizing conventional CPUs [24]. Thus high priority project were started and goal was to make it 10 times better in cost and performance over GPU. And in 15 months TPUs were designed. In order to reduce delay, TPU was intended to be a co-processor on the PCIe I/O bus. In order to simplify design and debugging, TPU does not fetch instruction to execute directly but host server sends instruction to TPU. Below figure show TPU block diagram, whose objective was to run whole inference model in order to reduce interaction with host CPU and to be adaptable enough to meet the NNs needs of 2015 [32].

The instructions are sent into instruction buffer over PCIe Gen13x16 bus from host. There is 256-byte wide path to which internal blocks are connected together. TPU block diagram shows Matrix Multiply unit in yellow, which is heart of TPU where main computation takes place. Blue weight FIFOs are its input and have blue unified buffer, blue accumulators are its output. Non linear functions are performed on accumulators by yellow activation unit, which later go to UB(unified buffer) [34]. Figure 8 follows same shading as in figure 7. Data path in light blue is 67%, I/O green is 10% and controls in red are only 2%.

Figure 9 shows a printed circuit board, which are used by Google Cloud in their data centers. TPUs uses tensors to provide high performance Machine Learning and DL operations. TPUs consists of four generations i.e First Generation TPUS, Second Generation TPUs, Third Generation TPUs and EDGE TPUs.

4.1. First Generation TPUs

First generation TPUs, include CISC instruction set along with PCIe 3.0 bus and is 8 bit multiplication engine. It works on integer operation and has limited bandwidth and performs well during inference phase in neural networks however not during training phase. It moves data to and from host, apply activates functions and perform convolutions and matrix operations [35]. TPUs are manufactured on 28nm process and 331 mm2 die size. They have on chip memory of 28Mib and 32 bit accumulators of 4Mib, that take results of 256x256 systolic array of 8 bit multipliers.TPUs consist of 8MiB of dual channel 2133 MHz DDR3 SDRAM that offers 34 GB/s bandwidth [25].

TABLE 1. SPECIFICATIONS OF SPECIALIZED PROCESSORS

Metrics	TPUs	GPUs	CPUs
Cores	Multiple cores are connected together to form a supercomputer called "TPU Pods" [27], with which an ML system can be trained overnight. Under a full TPU Pod training of some models is done in 30 minutes rather than a whole day .	Nividia tensor cores in GPU provides ultra-fast performance and large-matrix operations. These cores have built in instruction set language [28].	The scalable processors of Intel XEON can support upto 28 physical cores at the frequency of 25GHz at each socket, this can be extended to 3.80GHz in turbo mode .
Softwares	TensorFlow are used to optimize google's harware. TensorFlow is a open-source library which is used to build models for ML and NNs.	CUDA software technology is used for NVIDIA GPUs. CUDA language excels in large scale astronomical calculations for deep learning and parallel computing [29] .	Many softwares are used by INTEL for optimization but all of them use a kernel math library MKL-DNN for DL and NN. The performance of INTEL increases by 100 times with this library [30].
Performance	A TPU has 3.5 times more on-chip memory as compared to GPU and 25 times more multiplier accumulators. Google claims that a TPU is 15-30 times faster than a GPU.	GPU Tesla V100 can work in a single clock cycle on 64 floating point operations. Tesla V100 can perform at peak throughput with 5120 CUDA cores and 640 tensor cores.	Xeon processors can process 3.57-5.18 TFLOPS per socket. 28 cores can be connected per socket.
Hardware	TPU architecture uses 4 stage-pipeline. ALUs have their own pipeline for matrix computation. The available pipelines varies while the 1000 of clock cycles are used in any NN instruction. Detrimental to peak performance is reduced by using more memory or L2/L3 caches.	Tesla V100 has improved the shared memory and L1 cache system. Bigger and Faster L1 cache helps in achieving peak performance and reduce the usage for memory for programmers.	There are 6 memory channels in a CPU, which can support upto 1.5TB of memory. There are three different levels in each core. L2 is !MB private cache and L3 cache is a shared cache of 38.5MB.
Power	In a TPU the read and writes are performed on buffer and memory due to which power optimization is achieved. The idle time of a TPU is more then GPU and CPU.	The maximum power required by a Tesla V100 is 250W. The speed and additional energy is achieved from the memory units built into the GCU.	A standard CPU uses the power between 65-86 watts. The power of Quad core processors ranges from 95-140 watts.

TABLE 2. SIMILARITIES IN PROCESSORS [31]

Metrics	Description
Specialized Softwares	Each company uses a specific software for its processors that works best for thet hardware integration. TPU, GPU and CPU all processors work in a way to keep its main processing unit as busy as possible.
Cache and Memory	For an NN application data caching is very important. Each company works on the improvement of data moving between ALUs, memory and cache. All processors include memory and caching ability.
Horizontal Scaling	All processors support horizontal scaling, parallelism and distributed processing. An NN model can be splitted into parts and run onto different machines to reduce the training time.
Instructions	For the basic computation of NN each company uses combined add-multiply instrucions. It is basic vector/matrix multiplication.
Power to Processor	Advanced memory management and caching techniques are used by all companies to keep as much as data near the ALUs. This saves power because moving data increases the power cost
ResNet-50	All of these companies reduce time of training from a whole day to about 30 minutes by using ResNet-50 by using its hardware and software stack.

TABLE 3. DIFFERENCES IN PROCESSORS [31]

Metrics	TPUs	GPUs	CPUs
Purchasing	Google is the only company that offers TPUs that is why its actual hardware could not be bought.	NVIDIA sell their hardware in the market so any business can use it..	Like NVIDIA, Intel also sell its hardware in market.
Processing Location	PCIe busses are used in google TPUs.	NVIDIA hardware also uses PCIe busses.	Intel do not require stacks because its data transfer is very high.
Software Differences	TensorFlow can be used with NVIDIA and CUDA softwares for parallelization optimization	NVIDIA and CUDA softwares can be used for parallelization optimization	Intel library MKL-DNN was designed to support different frameworks for DNN.

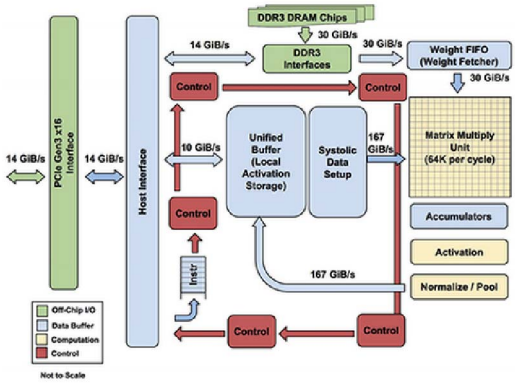

Figure 7. TPU Block Diagram [33]

.

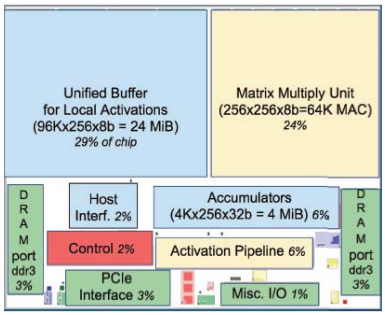

Figure 8. TPU Die FloorPlan [33]

.

4.2. Second Generation TPUs

In 2017, Google announced second generation TPUs, which have increased bandwidth of 600 GB/s and performance have also increased to 45 tera FLOPS. To increase performance to 180 tera FLOPS ,four chips of TPUs were arranged. For parallelism Google introduced TPU pod, which has 64 TPUs arranged to chip. First

Figure 9. Google TPU on Printed Circuit Board [23].

generation has limitation that it has only integer operations and perform well during inference phase and not during training phase but second generation TPUs also work on floating point operation and perform well during both phases i.e inference and training [26].

4.3. Third Generation TPUs

In May 2018, Google announced third generation TPUs which were twice powerful and have 8 times better throughput than second generation TPUs as they have 4 time more chips [37], [38]. They have increased bandwidth upto 128GB, more Tera FLOPS upto 420. With increased bandwidth memory of 320GB, TPU third generation pod can deliver 100 Peta FLOPs.

4.4. EDGE TPUs

In July 2018, Google announced edge TPUs designed for neural networks inference and training on edge computing [39]. They give high performance under small physical and power limitation. Using tensor-flow lite, they can be used to perform inference in field of IoT [40].

5. Conclusion

TPU will advance Moore's law prediction into future by seven years. Due to the increased demand of deep neutral networks, machine learning specialized hardware are of main focus. They decrease the overall training cost in deep neural networks. They

are specifically designed for high parallelism and to execute SIMD instructions with multi-threading. Neural networks require to process a lot of complex computations due to which manufacturers are focusing to improve performance by introducing parallelism in ALU operations to cater the need of matrix multiplication. All specialized hardware approaches i.e CPU, GPU and TPUs shows massive, similar upgrades to neural networks training and inference. It truly depends upon the model, how it is defined. Due to the new innovations happening every year, the choice of specialized hardware changes for neural networks according to the requirements.

References

[1] Sze, Vivienne, et al. "Efficient processing of deep neural networks: A tutorial and survey." Proceedings of the IEEE 105.12 (2017): 2295-2329.

[2] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[3] Zhang, Zewang, et al. "Deep recurrent convolutional neural network: Improving performance for speech recognition." arXiv preprint arXiv:1611.07174 (2016).

[4] Friedman, Daniel. "SC: Hardware approaches to machine learning and inference." 2018 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2018.

[5] Avati, Anand, et al. "Improving palliative care with deep learning." BMC medical informatics and decision making 18.4 (2018): 122.

[6] Wei, Tan Chiang, U. U. Sheikh, and Ab Al-Hadi Ab Rahman. "Improved optical character recognition with deep neural network." 2018 IEEE 14th International Colloquium on Signal Processing Its Applications (CSPA). IEEE, 2018.

[7] ASANOVIĆ, KRSTE, et al. "Designing a connectionist network supercomputer." International journal of neural systems 4.04 (1993): 317-326.

[8] Gschwind, Michael, Valentina Salapura, and Oliver Maischberger. "Space efficient neural net implementation." Proc. of the Second International ACM/SIGDA workshop on Field-Programmable Gate Arrays, Berkeley, CA. 1994.

[9] Gschwind, Michael, Valentina Salapura, and Oliver Maischberger. "A generic building block for hopfield neural networks with on-chip learning." IEEE International Symposium on Circuits and Systems, Atlanta, GA. 1996.

[10] Säckinger, Eduard, et al. "Application of the ANNA neural network chip to high-speed character recognition." IEEE Transactions on Neural Networks 3.3 (1992): 498-505.

[11] Gschwind, Michael, et al. "Synergistic processing in cell's multicore architecture." IEEE micro 26.2 (2006): 10-24.

[12] Ansari, Daniel, et al. "Artificial neural networks predict survival from pancreatic cancer after radical surgery." The American Journal of Surgery 205.1 (2013): 1-7.

[13] Kwon, Bomjun, et al. "Parallelization of the scale-invariant keypoint detection algorithm for cell broadband engine architecture." 2008 5th IEEE Consumer Communications and Networking Conference. IEEE, 2008.

[14] Duan, Rubing, and Alfred Strey. "Data mining algorithms on the cell broadband engine." European Conference on Parallel Processing. Springer, Berlin, Heidelberg, 2008.

[15] webopedia.org. Retrieved October 31, 2020 from https://www.webopedia.com/TERM/C/CPU.html

[16] computer.howstuffworks.com. Retrieved October 31, 2020 from https://computer.howstuffworks.com/graphics-card.htm

[17] computer.org. Retrieved October 31, 2020 from https://www.computer.org/publications/tech-news/chasing-pixels/is-it-time-to-rename-the-gpu.

[18] Buck, Ian. "Gpu computing with nvidia cuda." ACM SIGGRAPH 2007 courses. 2007. 6-es.

[19] Uralsky, Yury Y. "Isosurface extraction utilizing a graphics processing unit." U.S. Patent No. 7,965,291. 21 Jun. 2011.

[20] Pabst, Thomas, and Lars Weinand. "ATI Takes Over 3D Technology Leadership With Radeon 9700." Tom's Hardware. Retrieved 29 (2016).

[21] Vranesic, Zvonko G. "The FPGA challenge." Proceedings. 1998 28th IEEE International Symposium on Multiple-Valued Logic (Cat. No. 98CB36138). IEEE, 1998.

[22] Barr, Keith. ASIC design in the silicon sandbox: a complete guide to building mixed-signal integrated circuits. McGraw-Hill, 2007.

[23] Osborne, Joe. "Google's tensor processing unit explained: this is what the future of computing looks like." TechRadar. Available via http://www. techradar. com/. Accessed 6 (2017).

[24] Jouppi, Norman P., et al. "In-datacenter performance analysis of a tensor processing unit." 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA). IEEE, 2017.

[25] Wang, Yu Emma, Gu-Yeon Wei, and David Brooks. "Benchmarking TPU, GPU, and CPU platforms for deep learning." arXiv preprint arXiv:1907.10701 (2019).

[26] Jangamreddy, Nikhil. "A Survey on Specialised Hardware for Machine Learning." (2019).

[27] tensorflow.org. Retrieved October 31, 2020, from https://www.tensorflow.org/

[28] nvidia.com. Retrieved October 31, 2020 from https://www.nvidia.com/en-us/data-center/tesla-v100/

[29] nvidianews.nvidia.com. Retrieved October 31, 2020 from https://nvidianews.nvidia.com/news/nvidia-titan-v-transforms-the-pc-into-ai-supercomputer

[30] software.intel.com. Retrieved October 31, 2020 from https://software.intel.com/en-us/articles/intel-processors-for-deep-learning-training

[31] Rush, Allen, Ashish Sirasao, and Mike Ignatowski. "Unified deep learning with cpu gpu and fpga technologies." Advanced Micro Devices, Tech. Rep.. 2017.

[32] Jouppi, N., Young, C., Patil, N., Patterson, D. (2018). Motivation for and evaluation of the first tensor processing unit. IEEE Micro, 38(3), 10-19.

[33] Jouppi, Norman P., et al. "In-datacenter performance analysis of a tensor processing unit." Proceedings of the 44th Annual International Symposium on Computer Architecture. 2017.

[34] Pandey, Pramesh, et al. "GreenTPU: Improving Timing Error Resilience of a Near-Threshold Tensor Processing Unit." Proceedings of the 56th Annual Design Automation Conference 2019. ACM, 2019.

[35] Jouppi, Norman, et al. "Motivation for and evaluation of the first tensor processing unit." IEEE Micro 38.3 (2018): 10-19

[36] Civit-Masot, Javier, et al. "TPU Cloud-Based Generalized U-Net for Eye Fundus Image Segmentation." IEEE Access 7 (2019): 142379-142387.

[37] Fischer, Keno, and Elliot Saba. "Automatic full compilation of julia programs and ML models to cloud TPUs." arXiv preprint arXiv:1810.09868 (2018).

[38] Huot, Fantine, et al. "High-resolution imaging on TPUs." arXiv preprint arXiv:1912.08063 (2019).

[39] You, Yang, et al. "Fast Deep Neural Network Training on Distributed Systems and Cloud TPUs." IEEE Transactions on Parallel and Distributed Systems (2019).

[40] devopedia.org. Retrieved October 31, 2020 from https://devopedia.org/tensor-processing-unit