FPGAs for Machine Learning Applications

1. **Abstract:**

[a] This research paper explains the potential that the field-programmable gate array chips (FPGAs) have for the development of deep learning applications in specific, and for advancements in the artificial intelligence (AI) and machine learning (ML) domains in general. The growing need for robust and fast hardware that can serve as AI accelerators corresponds to the currently-proliferating market of AI applications and the expanding amounts of data required in this field [1]. [b] Without specialized hardware, AI applications would require months or even years to analyze large quantities of data and make decisions using traditional computational power, which would make real-time AI applications impossible. [c] Using the flexibility and computational speed of current FPGAs, engineers can implement whole complex machine learning systems like convolutional neural networks along with supporting algorithms and circuits on one chip. Also, the flexibility of FPGAs allows users to modify the system during operation in case the conditions or the data sets of the application change. FPGAs provide a means to implement both parallel computing and hardware optimization on AI accelerators, increasing the system’s efficiency and speed during both the training and the inference phases. [d] Due to the current FPGAs’ computational power, FPGAs have gained a significant share in the AI market among other AI chips and have attracted top companies like Amazon and Microsoft for their AI applications.

[e] **Keywords** – AI, Accelerators, Machine Learning, Deep Learning, Neural Networks, FPGAs, GPUs, CPUs

1. **Executive Summary:**

[f] The FPGA device is one of the technology options that companies and individuals use when designing AI accelerators for artificial intelligence and machine learning applications. Machine learning is the part of artificial intelligence that involves analyzing patterns in streams of data for computer systems to learn, make decisions without the human’s help, and solve problems beyond the human’s ability [2]. Deep learning is a special branch of machine learning in which the system learns directly from raw data without rules placed by a human [g] For the machine learning system to be accurate, the system requires large amounts of data to analyze and learn from before creating a model for the application. Since this process is power and time consuming, it needs special hardware called “AI accelerators” that perform parallel computations to many processes at once. Using traditional computing power through general-purpose central processing units (CPUs) will require far more time to learn and give decisions which is not practical in real-time ML applications. [h] The FPGA is suitable for building AI accelerators due to its flexibility that allows developers to design parallel computing blocks onto the FPGA fabric. Also, this flexibility allows constantly evolving ML systems to re-shape as the learning process continues and allows developers to implement extra functions around the main ML core, such as cryptographic algorithms [3]. [i] For FPGAs to dominate the AI market, top FPGA vendors should focus on providing AI software libraries supported by FPGAs, similar to that currently available for GPUs, and advertise the benefits of FPGAs in the AI domain to AI companies and research groups at top universities.

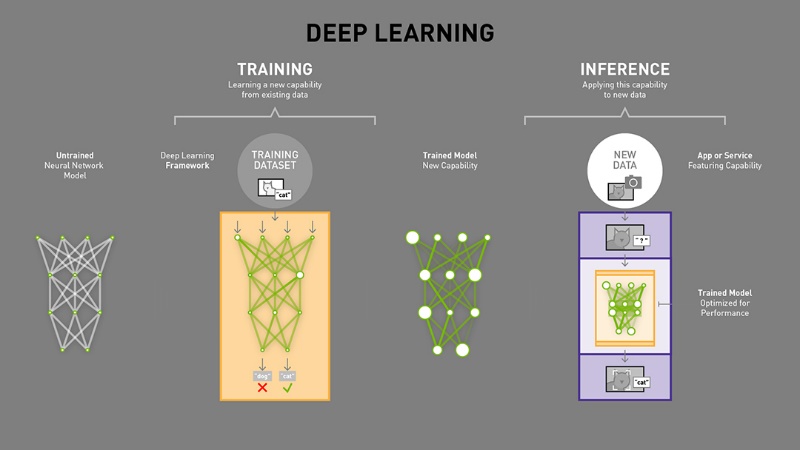
1. **Research Background:**
   1. *Machine Learning:*

While artificial intelligence is all about mimicking human behavior in computer systems, machine learning is the part of AI that involves teaching the machine or computer how to learn independently. The life cycle of the machine learning system begins by asking the question or identifying the problem. Then comes the data collection step to train the computer algorithm using that data. After training the machine, developers try out the algorithm in an application. The trained algorithm produces information that the system uses as feedback data to further train the algorithm. The importance of machine learning systems arises from the ability of the ML system to make decisions independently without the need for human intervention. This idea can revolutionize the status quo of various fields and industries like medicine, transportation, education, and more.

* 1. *Deep Neural Networks:*

In machine learning, deep learning is the approach of using artificial neural networks for computational analysis, similar to how biological neurons work inside the human brain to learn, analyze, and make decisions. The deep neural network (DNN) consists of several layers between an input layer and an output layer. Each layer consists of artificial neurons, which are non-linear math functions and communicate with other neurons to form the neural network. A simple artificial neuron does three main operations to the inputs; multiplication, addition, and passing them through an activation function. Each neuron first multiplies each input by a modifiable weight, then adds each term together, before passing the result through an activation function. With the activation function, the neuron passes a bounded data to the output instead of dealing with the unbounded input data [4]. The number of the hidden neural layers between the input layer and the output layer depends on the complexity of the application. The aim of the neural network is to receive raw data as inputs and learn by adapting its weights according to the training data [5].

The learning journey of a deep learning system starts with the “training” stage, as shown in figure 1 [6]. In the “training” phase, the system receives unstructured data to learn from. When these inputs go through the neurons of the first layer, the neurons adjust their weights according to how close or how far the data is to a certain feature measured by this layer. Then, the data is passed to the next layer that measures another feature, thus adjusting its neurons’ weights differently. The data passes through several layers before reaching the output layer, which generates the decision based on the information coming from each intermediate layer. This deep learning system goes through this process over numerous times and cycles so that the neurons can learn and modify their weights until the model generated becomes accurate enough for the application. The system quantifies how good the trained model is by using a function called the “loss function” or the “cost function.” The aim of the ML system during the training stage is to minimize the loss parameter generated by the loss function. So, by receiving various and diverse amounts of right data and solving lots of straightforward problems, the ML system can keep re-shaping the neural network till the model is confident enough to make decisions in similar but harder problems.

  
Figure 1. The two stages of Deep learning [6]

After the training stage, the deep learning system goes through the “inference” stage, where it deploys the trained neural network model in the actual application. This is the stage where the neural network is actually useful for the application since the trained model now experiences new real data and makes decisions for problems with unknown results [7]. Each neural layer in the network analyzes a feature in the coming data and gives a decision based on the weights obtained from the training stage; then, the final output layer generates the decision accordingly. For example, in an image-recognition system, one layer may look for shapes, another may look for faces, and others may look for other particular features. After collecting these different pieces of information from the intermediate layers, the output neural layer can determine whether the picture is of a person or an animal, what the gender is, or how many objects there are, depending on what the model is trained to detect. So, for deep learning systems, the inference stage is the production phase, where the model generates new information independently for sophisticated applications.

1. **Analysis Criteria**

Since machine learning and deep learning deal with large amounts of data to build an accurate model, ML systems use dedicated hardware called “AI accelerators.” Unlike general-purpose CPUs, AI accelerators are dedicated hardware whose purpose is to perform AI-specific tasks efficiently [8]. There are different types of chips that people can use as AI accelerators, like neural network processing units (NNPUs), graphics processing units (GPUs), and field-programmable gate arrays (FPGAs). Each type of AI accelerator has advantages and disadvantages, which led certain applications to prefer one type over another or in some cases, a combination of two or more types [9].  For developers to use the AI accelerators, these hardware accelerators must support software platforms on which AI libraries can run [10]. The recent widespread popularity of AI and ML is only possible due to the advancement of the AI accelerators’ hardware technology and the progression in the software libraries that run on these accelerators.

Certain hardware-related capabilities determine whether a hardware chip is eligible as a candidate as an AI accelerator or not [11]. ML and deep learning applications deal with tremendous sets of data during both the training stage. So, using the traditional general-purpose CPUs will take months or years to go through all the data, which is not practical for industrial use. Also, during the inference stage, analyzing data in real-time applications will be challenging if the ML system is slow. Thus, the top characteristic in an AI accelerator is its ability to speed up the AI tasks, which is usually achievable through parallel computation. Also, the amount of unstructured data requires a lot of hardware resources along with high memory capacity per chip. In addition, AI accelerators must have high modifiability to accommodate any evolution to the used algorithms or the trained model. What differentiates the different AI accelerator architectures from each other is how much of each capability it can offer. There is not one AI accelerator that excels in all these characteristics, which resulted in AI companies using different chips for different applications.

For the software part of the system, the AI accelerators should support a software platform that is easy for developers to use and has an abundant amount of useful AI libraries [12]. AI frameworks like TensorFlow and Theano offer useful software libraries to developers to easily code within their machine learning applications using typical programming languages like Python and C++. But, to use high-level programming languages with AI libraries, AI accelerator chip vendors must provide a software platform for developers to use the hardware efficiently. For example, NVIDIA has created a parallel computing platform for its GPU chips called “CUDA” that can work with high-level programming languages. Xilinx has also created software platforms like “PYNQ” so that AI developers can program the FPGA using Python instead of hardware description languages like Verilog or VHDL. Thus, AI accelerator chips must satisfy both the hardware and the software requirements before machine learning developers can widely use them.

1. **Technology Description:**
   1. *FPGA Architecture and Design Flow:*

[j] The reason for the attractiveness of FPGAs in the AI field is that FPGAs are flexible chips that developers can re-configure their internal structure on the hardware level. Unlike application-specific chips (ASICs), the FPGA chips’ main fabric consists of configurable logic blocks (CLBs), and configurable interconnects. The CLBs consists of a few logic cells, RAM cells, and flipflops, while the configurable interconnects consist of wire connections and switch boxes at each intersection between each two wires. Using electronic design automation tools (EDAs), hardware engineers can use hardware description languages (HDLs) like Verilog and VHDL to control the configuration of the CLBs to create digital blocks and control the switches in the switch boxes to connect these digital blocks. In addition to the configurable fabric, recent FPGAs contain specialized blocks like AI engines, high-speed interfaces, and real-time processing units [13], as shown in figure 2. Having all these different units on the FPGA with flexible connectivity between them enables the customer to implement all performance functions and AI inference processes on one compact chip [14].

The FPGA design cycle consists of four main stages: design entry, synthesis, implementation, and device programming. Design entry involves adding the design file that contains the Verilog code containing the design and the constraints file that specifies the timing conditions, the power regions, and the pin selections. Logic synthesis is the stage where the Verilog abstract description of the system is converted into a netlist file describing the design in terms of logic and circuit elements. In the implementation stage, the EDA maps the elements in the netlist file to the available resources on the FPGA device, places them according to the specified constraints, and then routes the connections between the different blocks. The EDA then produces a bitstream file to be programmed on the actual chip. The benefit of using an FPGA is that whenever the user needs to modify the design, the user only needs to re-generate the bitstream file then re-program the chip without any additional costs.

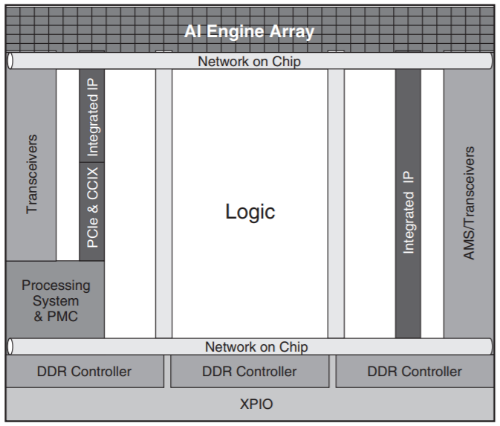


Figure 2. Xilinx’s Versal chip with AI Engine Architecture [15]

* 1. *AI Engines:*

Xilinx has added special logic blocks called “AI engines” in its latest family of FPGA chips to deliver more compute capacity, high-performance real-time DSP capabilities, improved development environment, and reduced power consumption for AI and ML applications. Each AI engine module contains a group of dedicated processors and memory blocks, as shown in figure 3, to support parallel execution of compute-intensive operations every clock cycle. Since the AI engines’ system works as a group of specialized processors, each responsible for a particular task, the FPGA offers a platform for heterogeneous computation, which results in low latency and high data throughput [15].

AI engines provide both data-level parallelism and instruction-level parallelism. AI engines are single instruction multiple data systems (SIMDs), which are a type of parallel computing systems in which multiple processor units execute the same instruction or operation on multiple data inputs at the same time. This process is called “data-level parallelism” [16]. AI engines are also very long instruction word machines (VLIWs), which are hardware architectures that rely on the compiler to combine multiple instructions into one long instruction word. This process is called “instruction-level parallelism” since the hardware understands the VLIW code that the compiler generates and executes the individual instructions in parallel. The advantage here is that the compiler (which is on the software level) can look at a larger window of instructions than the hardware and decide which instructions execute simultaneously. This reduces the need for more complicated hardware to discover the parallelism on its own. So, the new generation of FPGAs achieves software-level parallelism through SIMD and VLIW architectures, in addition to the inherent characteristic of FPGAs to achieve hardware-level parallelism or concurrency using the programmable logic around the AI engines.

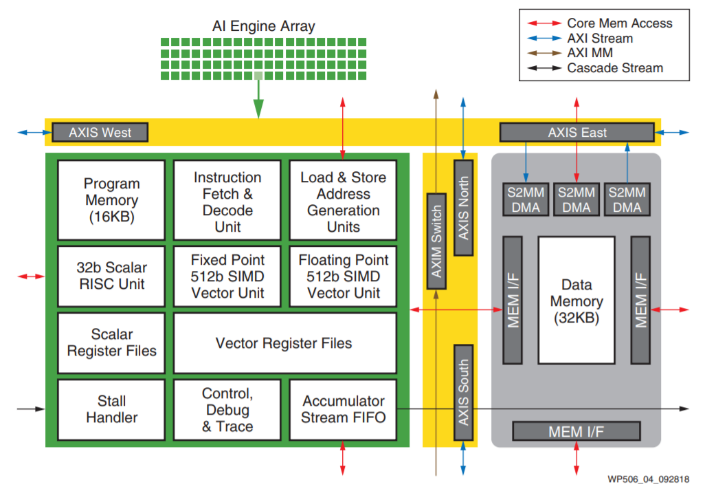


Figure 3. AI Engine Module [15]

1. **Technology Comparison:**

The tremendous increase in applying AI in recent applications has driven many of the hardware tech companies to try to deploy their technology in the AI market [17]. People usually compare CPUs, GPUs, and FPGAs when addressing applications that require heavy and complex computation power. Table 1 shows a comparison between the main chip technologies used in the AI industry [18]. In the AI market, GPUs and FPGAs are significantly preferable over CPUs, due to the reason that CPUs are general-purpose chips and are optimized for sequential operations, so their speeds to perform specific tasks are relatively low [19]. Another set of chip family that the industry considers for AI applications are AI accelerator ASICs like Google’s tensor processor units (TPUs) or Cerebras’ wafer-scale engines (WSEs). Despite these ASICs offering extremely high computation speeds, these chips are only useful for specific tasks and lack the flexibility to be used across different ML applications. For that reason, AI companies using AI ASICs still need GPUs or FPGAs along with the ASICs to build a complete AI or ML system.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CPU | FPGA | GPU | AI ASIC |
| Outline | General-purpose processor. | Configurable combination of logic and hard IP blocks. | Processor units originally designated for graphics applications. | Integrated circuit chips designed specifically for a certain application. |
| Processing power | Single and multi-core microprocessor units. | Contains hard and soft microprocessor cores. | Thousands of identical processor cores. | Number of processor cores is application-specific. |
| Programming | OSes and APIs run high-level languages, and Assembly language. | Normally HDL, but newer systems include C++ via OpenCL. | High-level programming via CUDA. | Application-specific.  (e.g. TensorFlow on Google’s TPU) |
| Peripherals | Lots of analog and digital peripherals. | Flexible transceiver blocks and I/O banks. | Limited peripherals. | Tailored to application. |
| Advantages | - Multitasking.  - Ease of programming. | - Hardware flexibility.  - Parallel computation.  - Broad choice of features. | - Massive processing power.  - Parallel computation.  - Ease of programming. | - Optimized design for application. |
| Disadvantages | - Optimized for sequential processing, so considered slow for AI applications. | - Difficult to program.  - Implementation complexity.  - Lack of AI libraries. | - High power consumption.  - Limited additional features. | - High development cost and time.  - Rigid, can’t change after production. |
| Top Vendors | - Intel  - Qualcomm  - IBM | - Xilinx  - Altera (Intel)  - Lattice Semiconductors | - Nvidia  - AMD | - Google  - Apple |

Table 1. A comparison between the main chip technologies in the AI industry [18]

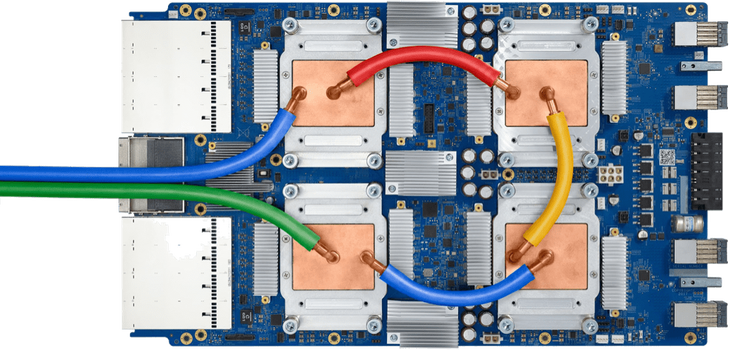
* 1. *CPUs:*

Even though the speeds of CPUs are not as high as that of GPUs and FPGAs, CPUs still play a role in AI and ML systems [20]. AI system designers may not use CPUs for AI acceleration and tasks that deal with extensive data, but the designers may pair up the CPU with a bunch of GPUs, ASICs, or FPGAs to act as the central unit that controls the data flow within the AI system and across the different components. In most computer systems, the CPU acts as the brain or the core of the system that perceives the instructions coming from the software level and allocates the different tasks to the suitable hardware components to execute. The basic operation of the CPU consists of three stages: fetching the instructions from the program memory, decoding each instruction into control signals, and executing the instruction by sending the control signals to other parts of the CPU or to other hardware components. Since CPUs are general-purpose chips and their operations are sequential, CPUs are not efficient for specific tasks that depend on parallelization like AI acceleration, but they are still important to serve as multi-purpose processors for most electronic systems like computers and mobile phones.

  
Figure 4. Intel’s 5GHz-capable Core i9-9900KS (Source: engadget.com)

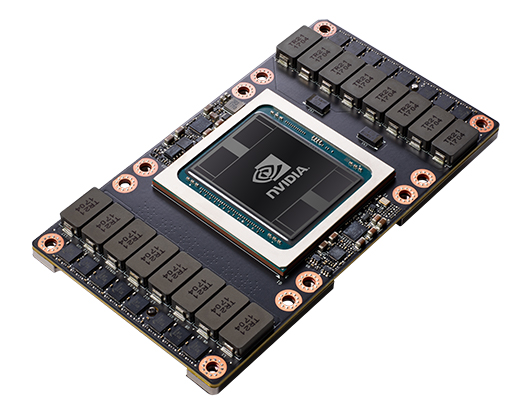
* 1. *TPUs:*

The TPU is one of the AI accelerator ASICs developed by Google as part of Google’s strategy as an “AI-first” company. In 2017, Google’s CEO Sundar Pichai announced that the company’s strategy moving forward would be to apply AI and ML across all Google’s platforms, products, and services [21]. For that reason, Google is developing its own optimized AI chips to be used in all Google’s AI systems around the world. The TPU chip specifically serves neural network machine learning applications and Google’s TensorFlow framework, which is a free, open-source software library widely used for AI and ML applications. The version of the TPU defines the TPU’s hardware architecture, the amount of high-bandwidth memory, the type of interconnects between the cores, and the internal networking interfaces. For data center applications, clusters of TPUs can work in a configuration called a “TPU pod” in which the TPU devices connect to each other through dedicated high-speed network connections. For the intended Google applications, TPUs show higher performance and efficiency than GPUs and CPUs [22]. Also, for Google, TPUs are cheaper than other chips since Google produces them in bulk to serve all of Google’s platforms and products. But despite the high performance of TPUs, they lack the flexibility to move to applications other than what they are designed for. Thus, Google still uses CPUs, GPUs, and FPGAs, along with the TPUs in many of its AI applications.

  
Figure 5. Google’s Cloud TPU V3 (Source: cloud.google.com)

* 1. *GPUs:*

NVIDIA has been dominating the AI market with its GPUs in recent years since GPUs offer high-speed computation using parallel programming architectures [23]. The initial use of GPUs was in dealing with graphics-related data in personal computers and in gaming consoles. The architecture of GPUs deals with huge chunks of data by breaking them up into millions of separate tasks, then processing them in parallel on hundreds of processor cores. This concept was ideal for high-definition graphics that require the processing of display components like textures, lighting, and shapes to be super-fast [24]. In addition to the hardware architecture, NVIDIA provided a software platform called “CUDA” that NVIDIA continuously updates for software developers to make use of the parallel computing architecture of GPUs using high-level programming languages like C, C++, and Fortran. For these reasons, 97% of the AI accelerators of the top cloud companies are based on NVIDIA’s GPUs alone as of May 2019, while FPGAs of Xilinx and Intel and AMD’s GPU make up less than 3% of the cloud AI market [25].

  
Figure 6. Nvidia’s Tesla V100 GPU (Source: nvidia.com)

* 1. *FPGAs:*

Recently, the top FPGA companies have evolved the FPGAs to compete with GPUs in AI acceleration applications. Xilinx re-branded its FPGA into adaptive compute acceleration platforms (ACAPs), which are basically the traditional FPGA with AI and DSP engines added to it and which is fully software-programmable [26]. Intel also decided to expand its FPGA business into AI applications by acquiring companies like Altera and Omnitek [27]. As seen in Table 1, both FPGAs and GPUs offer higher speeds and higher data throughput than CPUs and a higher degree of flexibility than ASICs. GPUs have an advantage over FPGAs where their software platform is well-developed since GPUs have been in the AI and ML industry for over a decade, unlike FPGAs, who have just started to enter the AI race. But, Xilinx and Intel are currently developing software platforms and AI libraries for their newer generation FPGAs with the help of AI companies like Baidu and Microsoft, so that software engineers can leverage the FPGA’s capabilities using high-level programming languages. Additionally, FPGAs offer higher hardware flexibility and additional hardware features along with the parallel computation capabilities, which can result in integrating the AI-related tasks within a sophisticated system in one compact chip. GPUs still need to communicate with CPUs and other chips to perform non-AI-related tasks that can affect the overall speed of the system. So, with the increasing usage of FPGAs in AI acceleration and with further developments in the FPGAs’ software platform and AI libraries, FPGAs can add additional value to AI applications that current GPUs don’t offer.

  
Figure 7. Xilinx’s Versal Chip (Source: xilinx.com)

1. **Results and Implications:**

Due to the rising potential of FPGAs in the AI market, top AI and ML companies are starting to shift from using only GPUs to giving part of the market to FPGAs in their AI applications. GPUs still dominate over 95% of the AI accelerator hardware market since developers and companies have been using GPUs for AI acceleration for over a decade, which resulted in giving Nvidia, the top GPU vendor, a head start in the AI and ML race. So, Nvidia has been optimizing the hardware design of GPUs and their software computing platform “CUDA” for AI acceleration and ML applications over the years. But despite GPUs’ current superiority in terms of market share, FPGAs can offer more than just AI acceleration on one chip, unlike GPUs. FPGAs can offer extra computation related to the specific application in addition to AI acceleration, removing any latency or slowdown due to moving data across different chips. FPGAs can also help reduce the production and the testing costs, since the designers can easily correct any mistakes in the hardware design by re-programming the FPGA and they can build test logic within the FPGA that monitors the system’s control and output signals. As a result, AI and ML companies and research groups are encouraging more developers to develop software libraries for FPGA-based platforms and to utilize the full potential of FPGAs in their AI applications.

* 1. *Microsoft’s Data Centers:*

Microsoft started to deploy Intel FPGAs for data center applications and is currently working with Xilinx to use their FPGAs as well [28]. Microsoft deploys AI and ML algorithms in its data center infrastructure to optimize how these data centers operate. For example, ML algorithms can use predictive analytics to distribute the workload efficiently across servers. Machine learning can also discover the best ways to cool down the servers, which reduces the amount and cost of energy needed for cooling activities. Additionally, ML systems can autonomously perform quick IT tasks to massive amounts of servers like security patching, file backups, and handling every user request or incident flag without the help of an IT personnel [29]. So, with FPGAs, Microsoft can use AI acceleration along with the additional hardware features in its data centers to improve security, conserve energy, reduce downtime, implement server optimization, and monitor its equipment.

Furthermore, Microsoft started offering machine learning models accelerated by FPGAs and powered by Project Brainwave to Microsoft’s Azure customers. Azure is Microsoft’s cloud computing service that covers a wide range of software, platform, and infrastructure services such as virtual machines, storage, data management, blockchain, internet of things, and machine learning. Regarding the machine learning services, Project Brainwave is a deep learning platform that provides Azure with real-time AI inference capabilities using Intel’s Stratix 10 FPGA devices. Project Brainwave loads the deep neural network model onto the FPGA after the software application level creates the model using a Python compiler and Azure’s machine learning SDK. Even though Project Brainwave currently only uses FPGAs for AI inference applications, Microsoft is showing an interest in reducing its dependence on GPUs and moving to FPGAs for AI training as well [30].

* 1. *Amazon’s Cloud Services:*

Amazon partnered with Xilinx to provide FPGA-accelerated software platforms for its AWS cloud computing services [31]. Amazon is offering to customers FPGA-based compute instances called “EC2 F1” that can utilize up to eight dedicated 16nm Xilinx Ultrascale+ FPGAs per cloud instance. The cloud instance is a virtual machine that offers a certain compute capability, memory, and storage capacities depending on the type of hardware instance the user chooses. The cloud instance implements an equivalent of a single hardware device across multiple computers, which results in a dynamic cloud computing, preventing major slowdowns if many users are working at the same time. Each FPGA in the EC2 F1 instance offer 2.5 million logic elements, 6800 digital signal processing (DSP) engines, and 64 GiB DDR4 memory with dedicated PCIe x16 connection. Currently, Amazon is offering the FPGA cloud instances for the experienced FPGA developers and is working on developing high-level tools like OpenCL for developers who are less experienced with FPGA design. Amazon has also set up a dedicated Github account for the AWS FPGA development kit with documentation, examples, and software guides to encourage developers to leverage the benefits of the FPGAs on the cloud. The current use cases of EC2 F1 instances are in genomics research, financial analytics, real-time video processing, large-throughput image processing, big data analytics, and network security.

* 1. *Baidu’s AI Platform:*

Baidu is also cooperating with Xilinx and Intel to expand its data centers and cloud’s AI hardware capabilities. Being one of the largest AI and internet services providers in the world, Baidu needs robust AI acceleration hardware to apply machine learning and security features to its data centers. Baidu uses these data centers for numerous machine learning applications like autonomous vehicles and voice and image recognition. Besides using the data centers for internal use, Baidu also offers workload acceleration services through its data centers to developers for their AI-related projects. So, recently, Baidu started deploying Xilinx and Intel FPGAs to its data centers and also launched its new cloud service, the “Baidu FPGA Cloud Server,” that support AI acceleration development activities using FPGAs [32] [33].

Moreover, Baidu is using FPGAs for its AI-based hardware products. For its smart camera products used in retail applications, Baidu is using several Intel chips like FPGAs and visual processing units (VPUs) to apply machine learning algorithms to detect objects and people, creating a personalized shopping experience to the customers in the retail business. Furthermore, Baidu released a development board for applications that use edge computing, which is a type of computing used in distributed systems that brings the resources and the data close to where they are used to decrease latency, improve response time, and enhance security measures. Baidu’s board is called the “EdgeBoard,” and it mainly depends on a Xilinx Zynq UltraScale+ multi-processor system-on-chip (MPSoC) which is programmable on both the software level and the hardware level. In order to promote this powerful edge computing tool, Baidu offers extensive documentation and examples to get the users started on the EdgeBoard through its open platform, the “Brain AI Hardware Platform.” So, with the increasing number of edge computing users and AI developers that are using Baidu’s vast platform, Baidu and the FPGA companies are expecting further expansion in the collaboration between the two sides to deliver high-quality AI and ML services [34].



Figure 8. Baidu’s EdgeBoard based on Xilinx’s Zynq UltraScale+ MPSoC device [34]

* 1. *Face-Detection Research Work:*

Several research groups in the academic community are also looking into using the hardware flexibility to optimize machine learning algorithms for applications like face detection. Researchers at Effat University in Jeddah, Saudi, used an Intel Cyclone VI FPGA to synthesize a Viola-Jones detection algorithm applicable for real-time face detection applications like in surveillance systems and for children supervision [35]. The Viola-Jones algorithm goes through four main stages: detects the Haar-like features, creates an integral image, passes through the learning algorithm, and finally cascading the classifiers. In the first stage, the system looks into a detection window of a certain size that moves across the picture. It then compares adjacent rectangular regions within this window and calculates the Haar-like feature for each subsection. In the second stage, the system uses a summed-area table algorithm to create an integral image based on the numbers obtained from the first stage. This integral image is useful for fast calculation evaluations. In the third stage, the algorithm uses a machine-learning algorithm called “AdaBoost” to create stronger feature classifiers. Then, in the fourth stage, the classifiers are cascaded (placed in a sequence). Each sub-region of the image goes through a classifier; if it passes, it goes through to the next classifier, and if it fails, that sub-region is filtered-out. So, the sub-regions of the image that are positive for all the classifiers and reach the output layer are the faces detected.

The FPGA hardware configuration consists of two main sub-systems: the main data path and the face detection algorithm block. The Cyclone VI FPGA board connects to an OV7670 camera from one end and connects to a VGA monitor on the other end. The main data path consists of a camera control logic, a RAM memory frame buffer to store a single image frame, and a VGA screen driver to continuously display the captured images on a VGA monitor along with the face detection indicators. The face detection block comprises four sub-blocks: an integral-image generation unit, some buffers, a sub-window kernel parallelization unit that represents the cascaded classifiers, and a face-box logic to draw a box around the face on the VGA screen. This system is able to successfully detect multiple faces per frame in real time with a maximum detection distance of 6 meters.

* 1. *Handwriting Recognition Research Work:*

Another research group at Ton Duc Thang University in Vietnam proposed an architecture that uses both an Intel Cyclone FPGA and an Nvidia GTX 1080GPU for handwriting recognition [36]. The project comprises two main parts: the training system using a GPU, and the inference system using an FPGA. The GPU trains the system’s neural network using the MNIST database, which includes a large database of different handwritten numbers and is commonly used to train machine learning systems that deal with handwriting recognition. After the training phase, the researchers download the parameters of the trained neural network onto the DRAM memory on the Intel Cyclone FPGA development kit. The AI acceleration algorithm built on the FPGA performs the recognition inference on the input written on a smartphone and displays the predicted digit on a 7-segment LED display. The inference runtime of the FPGA-based system is 10-4 seconds, with a recognition rate of 98.15%. Besides the performance, the benefit of using an FPGA was that the researchers were able to build the receiver, the DRAM, the 7-segment controller, and the inference algorithm block on one compact chip. So, as seen in the previous examples, FPGAs are obviously gaining more shares in the AI and the machine learning markets, as well as in academia, and some experts are expecting FPGAs to surpass GPUs in terms of AI market share soon [37].

1. **Conclusion:**

[l] This research paper shows the usefulness of FPGAs in AI and machine learning applications. FPGAs have the computational power, speed, and flexibility that allow users to build sophisticated and energy-efficient machine learning systems. The re-programmability of FPGAs also allows the system to easily re-model according to new data and application changes. With the extra logic inside FPGAs, design engineers can also apply software and hardware cryptography algorithms to prevent any hacking to the machine learning model from outsiders. The re-programmability feature of FPGAs can also minimize system design risks by allowing designers to add extra logic for testing and to load the improved design on the FPGA for any number of times. With the multi-level parallelism, speed, and flexibility offered by the new FPGAs like Xilinx’s Versal & Intel’s Agilex chip families, FPGAs can present a more powerful solution than GPUs for machine learning and AI applications [38]. For that reason, top AI companies are investing in further collaboration with FPGA vendors and are promoting the idea of leveraging the power of FPGAs within the AI market. So, the concept of FPGA deployment in AI applications is a promising idea since it allows companies and developers to build deep learning systems integrated within the application’s top-level system in one compact module, that is also adaptable to changes and improvements that the system might need in the future as machine learning models develop.

**References:**

1. L. Columbus, (2019), *“Roundup Of Machine Learning Forecasts And Market Estimates For 2019”* [Online], Available: https://www.forbes.com/sites/louiscolumbus/2019/03/27/roundup-of-machine-learning-forecasts-and-market-estimates-2019/#5275610f7695, Accessed: October 27, 2019.
2. M. Elliott, (2018), “Machine Learning, Deep Learning and Prescriptive Analytics: What’s the Difference?” [Online], Available: https://blog.netapp.com/machine-learning-deep-learning-and-prescriptive-analytics-whats-the-difference/, Accessed: October 27, 2019.
3. M. Yatsenko, (2018), “FPGAs for Artificial Intelligence: Possibilities, Pros, and Cons” [Online], Available: https://www.apriorit.com/dev-blog/586-fpgas-for-ai, Accessed: October 27, 2019.
4. V. Zhou, (2019), “Machine Learning for Beginners: An Introduction to Neural Networks” [Online], Available: https://towardsdatascience.com/machine-learning-for-beginners-an-introduction-to-neural-networks-d49f22d238f9, Accessed: November 9, 2019.
5. C. Thomas, (2019), “The basics of Deep Neural Networks” [Online], Available: https://towardsdatascience.com/the-basics-of-deep-neural-networks-4dc39bff2c96, Accessed: October 27, 2019.
6. M. Copeland, (2016), “What’s the Difference Between Deep Learning Training and Inference?” [Online], Available: https://blogs.nvidia.com/blog/2016/08/22/difference-deep-learning-training-inference-ai/, Accessed: October 14, 2019.
7. S. Sharma, (2019), “Four Surprising Ways Inference Is Putting AI into Action” [Online], Available: https://blogs.nvidia.com/blog/2019/06/14/ai-inferencing/, Accessed: October 29, 2019.
8. M. Rouse, (2018), “AI Accelerator” [Online], Available: https://searchenterpriseai.techtarget.com/definition/AI-accelerator, Accessed: October 29, 2019.
9. M. Rojek, (2018), “Which AI accelerator shall I pick for my next device?” [Online], Available: https://becominghuman.ai/which-ai-accelerator-shall-i-pick-for-my-next-device-e267a64892bb, Accessed: October 29, 2019.
10. A. Thakur, (2019), “10 AI Platforms to Build Your Modern Application” [Online], Available: https://geekflare.com/ai-platforms/, Accessed: October 14, 2019.
11. *White Paper on AI Chip Technologies*, Beijing Innovation Center for Future Chips (BICFC), Tsinghua University, Beijing, China 2018.
12. J. MSV, (2018), “How AI Accelerators Are Changing The Face Of Edge Computing” [Online], Available: https://www.forbes.com/sites/janakirammsv/2019/07/15/how-ai-accelerators-are-changing-the-face-of-edge-computing/#5fa602c2674f, Accessed: October 14, 2019.
13. *Exploring Zynq® MPSoC, With PYNQ and Machine Learning Applications*, 1st ed., Xilinx Inc., San Jose, CA, 2019.
14. M. Feldman, (2019), “FPGAs Open Gates in Machine Learning” [Online], Available: https://www.nextplatform.com/2019/05/01/fpgas-open-gates-in-machine-learning/, Accessed: October 15, 2019.
15. *Xilinx AI Engines and Their Applications White Paper*, WP506 (v1.0.2), Xilinx Inc., San Jose, CA, 2018.
16. B. Barney, (2019), “Introduction to Parallel Computing” [Online], Available: https://computing.llnl.gov/tutorials/parallel\_comp/, Accessed: November 17, 2019.
17. C. Wheelock, (2019), “Deep Learning Chipset Shipments to Increase to 2.9 Billion Units Annually by 2025, According to Tractica” [Online], Available: https://www.businesswire.com/news/home/20190506005260/en/Deep-Learning-Chipset-Shipments-Increase-2.9-Billion, Accessed: October 14, 2019.
18. “FPGA vs CPU vs GPU vs Microcontroller: How Do They Fit into the Processing Jigsaw Puzzle?” [Online], 2019, Available: https://www.arrow.com/en/research-and-events/articles/fpga-vs-cpu-vs-gpu-vs-microcontroller, Accessed: November 10, 2019.
19. J. Dorsch, (2016), “CPU, GPU, Or FPGA?” [Online], Available: https://semiengineering.com/cpu-gpu-or-fpga/, Accessed: October 14, 2019.
20. T. Morgan, (2019), “The CPU Still Matters In The AI Stack” [Online], Available: https://www.nextplatform.com/2019/05/07/the-cpu-still-matters-in-the-ai-stack/, Accessed: November 24, 2019.
21. S. Condon, (2019), “Google I/O: From 'AI first' to AI working for everyone” [Online], Available: https://www.zdnet.com/article/google-io-from-ai-first-to-ai-working-for-everyone/, Accessed: November 24, 2019.
22. K. Sato, C. Young, and D. Patterson, (2017), “An in-depth look at Google’s first Tensor Processing Unit (TPU)” [Online], Available: https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu, Accessed: November 24, 2019.
23. P. Teich, (2019), “NVIDIA Dominates The Market For Cloud AI Accelerators More Than You Think” [Online], Available: https://www.forbes.com/sites/paulteich/2019/06/17/nvidia-dominates-the-market-for-cloud-ai-accelerators-more-than-you-think/#5ceae20d5edb, Accessed: October 14, 2019.
24. B. Caulfield, (2009), “What’s the Difference Between a CPU and a GPU?” [Online], Available: https://blogs.nvidia.com/blog/2009/12/16/whats-the-difference-between-a-cpu-and-a-gpu/, Accessed: November 10, 2019.
25. P. Teich, (2019), “NVIDIA Dominates The Market For Cloud AI Accelerators More Than You Think” [Online], Available: https://www.forbes.com/sites/paulteich/2019/06/17/nvidia-dominates-the-market-for-cloud-ai-accelerators-more-than-you-think/#5bff819a5edb, Accessed: November 10, 2019.
26. R. Shrout, (2018), “Opinion: Xilinx shifts directions to compete in artificial intelligence with Nvidia and Intel” [Online], Available: https://www.marketwatch.com/story/xilinx-shifts-directions-to-compete-in-artificial-intelligence-with-nvidia-and-intel-2018-10-02, Accessed: October 14, 2019.
27. I. Lunden, (2019), “Intel acquires UK’s Omnitek to double down on FPGA solutions for video and AI applications” [Online], Available: https://techcrunch.com/2019/04/16/intel-acquires-uks-omnitek-to-double-down-on-fpga-solutions-for-video-and-ai-applications/, Accessed: October 14, 2019.
28. K. Freund, (2018), “Is Microsoft Adding Xilinx To Its Datacenters?” [Online], Available: https://www.forbes.com/sites/moorinsights/2018/11/02/is-microsoft-adding-xilinx-to-its-datacenters/#3bc8ade4cf45, Accessed: October 15, 2019.
29. A. Ahdoot, (2019), “Three Ways Artificial Intelligence Will Revolutionize Data Centers” [Online], Available: https://www.datacenterknowledge.com/industry-perspectives/three-ways-artificial-intelligence-will-revolutionize-data-centers, Accessed: November 10, 2019.
30. M. Feldman, (2018), “Microsoft Launches FPGA-Powered Machine Learning for Azure Customers” [Online], Available: https://www.top500.org/news/microsoft-launches-fpga-powered-machine-learning-for-azure-customers/, Accessed: November 26, 2019.
31. K. Freund, (2017), “Amazon And Xilinx Deliver New FPGA Solutions” [Online], Available: https://www.forbes.com/sites/moorinsights/2017/09/27/amazon-and-xilinx-deliver-new-fpga-solutions/#3c9b81362370, Accessed: October 15, 2019.
32. Xilinx Inc., (2017), “Baidu Deploys Xilinx FPGAs in New Public Cloud Acceleration Services” [Online], Available: https://www.xilinx.com/news/press/2017/baidu-deploys-xilinx-fpgas-in-new-public-cloud-acceleration-services.html, Accessed: November 23, 2019.
33. Intel Inc., (2018), “Intel AI at Baidu Create: AI Camera, FPGA-based Acceleration and Xeon Scalable Optimizations for Deep Learning” [Online], Available: https://newsroom.intel.com/news/intel-ai-baidu-create-ai-camera-fpga-based-acceleration-xeon-scalable-optimizations-deep-learning/#gs.hz9moc, Accessed: November 23, 2019.
34. C. Atwell, (2019), “Baidu Announces Xilinx-Based EdgeBoard for AI Applications” [Online], Available: https://www.hackster.io/news/baidu-announces-xilinx-based-edgeboard-for-ai-applications-31f32d4456cb, Accessed: November 23, 2019.
35. A. Alahmadi, and S. Qaisar, "Robust Real-time Embedded Face Detection Using Field Programmable Gate Arrays (FPGA)," 2019 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, UAE, 2019.
36. V.Phat, P. Tho, H. Dat, and C. Chou, "Deep Learning Accelerator on FPGA Using Handwritten Digit Recognition for Example," 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), Taichung, Taiwan, 2018.
37. L. Barney, (2017), “Can FPGAs Beat GPUs in Accelerating Next-Generation Deep Learning?” [Online], Available: https://www.forbes.com/sites/moorinsights/2017/09/27/amazon-and-xilinx-deliver-new-fpga-solutions/#3c9b81362370, Accessed: October 15 2019.
38. A. VanDer Ploeg, (2018), “Why use an FPGA instead of a CPU or GPU?” [Online], Available: https://blog.esciencecenter.nl/why-use-an-fpga-instead-of-a-cpu-or-gpu-b234cd4f309c, Accessed: October 15, 2019.

**Abstract & Executive Summary Checklist**

|  |  |
| --- | --- |
| **Item** | **CHECKLIST ITEM** |
| [a] | Abstract: Motivation Sentence |
| [b] | Abstract: Problem Statement |
| [c] | Abstract: Results Sentence |
| [d] | Abstract: Conclusions Sentence |
| [e] | Abstract: Keywords |
| [f] | Executive Summary: Research Background |
| [g] | Executive Summary: Problem Statement |
| [h] | Executive Summary: Value statement determined by analysis criteria |
| [i] | Executive Summary: Recommended action |
| [j] | Product Description: Product Description |
| [k] | Product Description:  Product Differentiators |
| [l] | Conclusion:   Recommendation Statement |

**EE295 Grammar Checklist**

|  |  |  |
| --- | --- | --- |
| **Item** | **Description** | **Compliance** |
| 1 | No personal words | Checked |
| 2 | No redundant phrases like “nothing but” or “very very” | Checked |
| 3 | No passive voice | Checked |
| 4 | All paragraphs with at least 3 sentences | Checked |