Name: Shehbaz Patel

Institution: Universita degli studi di milano

Date : 16-10-2018

Hardware acceleration for computer intelligence

Table of contents

List of figures....1

Introduction....2

The evolution of Artificial intelligence accelerator chips....3

Artificial intelligence tiers....3

Artificial intelligence tasks....4

Artificial intelligence tolerance....5

State of the art....5

Methodologies/Best practices....7

Solving the memory challenges for AI hardware acceleration....7

Solving the computing challenges for AI applications....8

Experiments, Use cases, Examples....9

GPU....10

FPGA....12

ASICs....12

Heterogeneous computing....13

Virtual machines and environments for NN acceleration..... 14

Nvidia Volta/Tesla application for NN acceleration..... 15

Conclusion..... 15

Works cited..... 16

List of figures

Figure 1 Routes to key market for AI chips 3

Figure 2 Model flow through the deep learning deployment toolkit. source 5

Figure 3 AI inference for vision solution using Intel FPGA supported DL frameworks 7

Figure 4 AI hardware acceleration infrastructure 8

Figure 5 How GPU acceleration works 11

Figure 6 The ASIC advantage source 13

Figure 7 How Heterogeneous computing work source 14

Introduction

Over the past decade, the established and the emerging chip vendors have come up with a more robust generation of new hardware architectures whose optimization is dedicated to machine learning, natural language processing, deep learning, and other artificial intelligence workloads. According to (11), Such emerging artificial intelligence chipset architecture is field programmable gate arrays (FPGAs), neural network processing units (GPUs), application specific integrated circuits (ASICs) and other related approaches referred to as neurosynaptic architecture that adds to the already common generations of GPUs (3). As described from a periodical review by Yoshida, The modern artificial intelligence has no hardware monoculture

equivalent to the traditional X86 central processing unit (CPU) which has dominated the chipset market in the architecture of desktop computing space over the last several years. This new trend is influenced by the applications and adaptation of the new artificial intelligence architecture into the emerging specified roles in the competitive cloud-to-edge based ecosystems such as computer vision and learning.

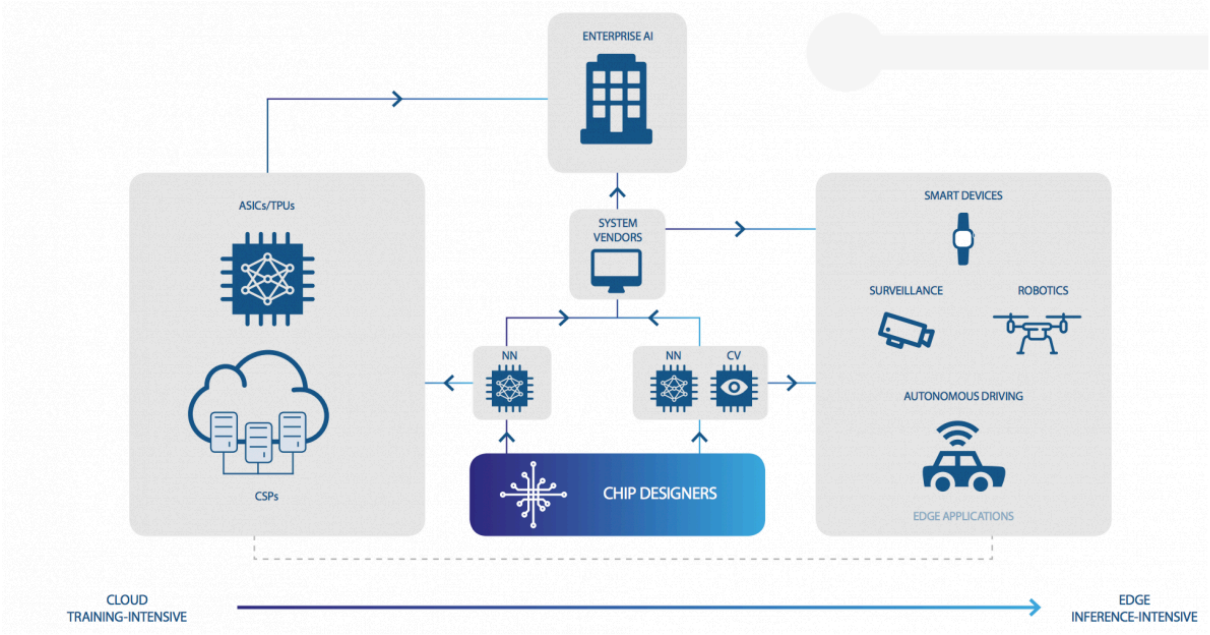


Figure 1 Routes to key market for AI chips

The evolution of Artificial intelligence accelerator chips

There has been a rapid evolution of the Artificial intelligence accelerator chips. This evolution can be better understood by focusing on the opportunities that have emerged in the marketplace and their associated challenges. Such include AI tiers, AI tasks, and AI tolerance.

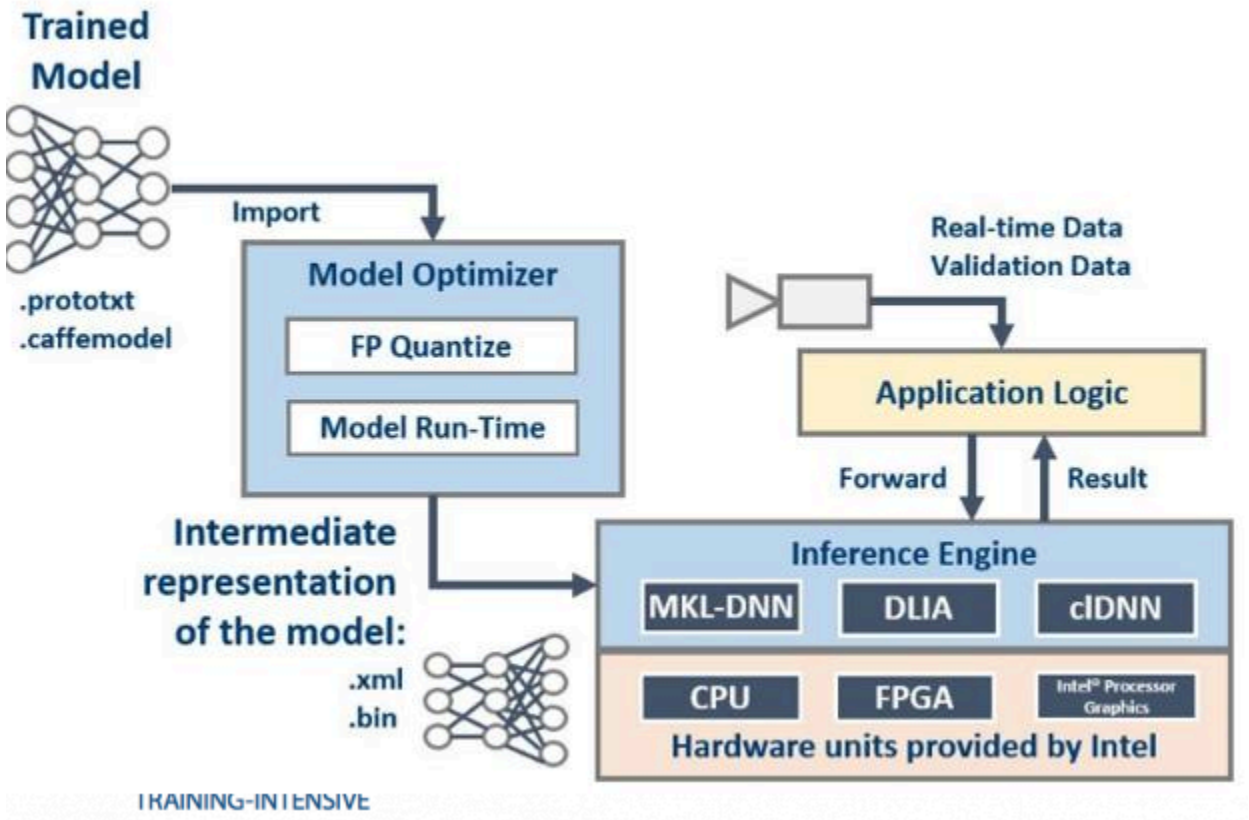
Artificial intelligence tiers

AI accelerators are fast evolving to match the edge where new hardware platform optimization is aimed at enabling a greater autonomy for smart handheld devices, internet of things (IoT) , and embedded devices. The innovation of AI robots has gone beyond the proliferation of Smartphone embedded AI processors to permit almost all activities that have entirely depended on human control such as self-driving vehicles, drones, industrial internet of things, and smart appliances.

Artificial intelligence tasks

AI accelerators are on the edge of penetrating into every tier in the distributed cloud to edge, hyper converged servers, cloud storage, high-performance computing architectures. Fresh hardware innovations are emerging every day and finding improvements into these segments to support more efficient, accurate, rapid AI processing. Such innovations in AI hardware are designed to specifically accelerate data-driven tasks of such distinct application environments. These chipset architectures on the modern market are a reflection of how diverse AI applications have become in a range of machine learning, natural language processing, deep learning, and other AI workloads ranging from computer intensive interference to storage-intensive training. Such involve varying degrees of device autonomy and the person in the loop interaction.

Figure 2 Model flow through the deep learning deployment toolkit. source (Kumar, Severtson and Glonrund)



Artificial intelligence tolerance

Any hardware acceleration for AI innovation applications must be robust enough to survive various environments based on its ability to attain metrics as defined within its specified applications and economic constraints. In terms of operations metrics, the AI hardware acceleration architecture must be competitive in both performance and ownership costs for the tiers and tasks into which it is designed to perform. Comparative industrial benchmarks have become a major factor in determining the precision in performance and determine the price performance profile that will survive in a highly competitive market environment. The

modern-day industry is moving towards the workload optimized artificial intelligence architecture. This architecture will allow users to adopt the most scalable, fastest, low-cost cloud platforms, hardware, and software to run a wide rand of AI tasks such as inference, training, development, and operationalization in any tier.

State of the art

One of the most common developments in AI tiers in the state of art appliances is the Nvidia's latest enhancement to the Jetson Xaviour AI line of AI systems on a chip (SOSs). The Isaac software development kit recently released by Nvidia has been instrumental in the building of robotic algorithms that will run on the dedicated hardware of new robotics. The Jetson Xavier chip is comprised of six processing units that include 512 core GPU, a dual Nvidia VoltaTensor Core GPU, an eight-core Carmel Arm 64 CPU, a dual Nvidia deep learning accelerator, and s vision, image, and video processors. The processing units have enabled the intelligent robotics to handle various algorithms to enable it sense environments autonomously, respond efficiently and effectively and safely operate alongside human engineers.

Vendors are invoking a wide range of technologies in coming up with their product portfolios to address a range of workloads being supported by Ai chipsets. This has also been the case in the architecture of specific embedded artificial intelligence deployments such as SOCs that drive mobile apps and intelligent robotics. For example the Xeon Phi CPU architecture by Intel which has been used for accelerating AI tasks. However, Intel has acknowledged that it will be challenging to put up with the emerging demands without specialized AI accelerator chips that can enable the company to compete effectively with Nvidia Volta in GPUs and other vendors manufacturing NNPUs and other AI chips. As a countermeasure, Intel has put in place a dedicated team that is working on developing a new GPU that will be released in the near future.

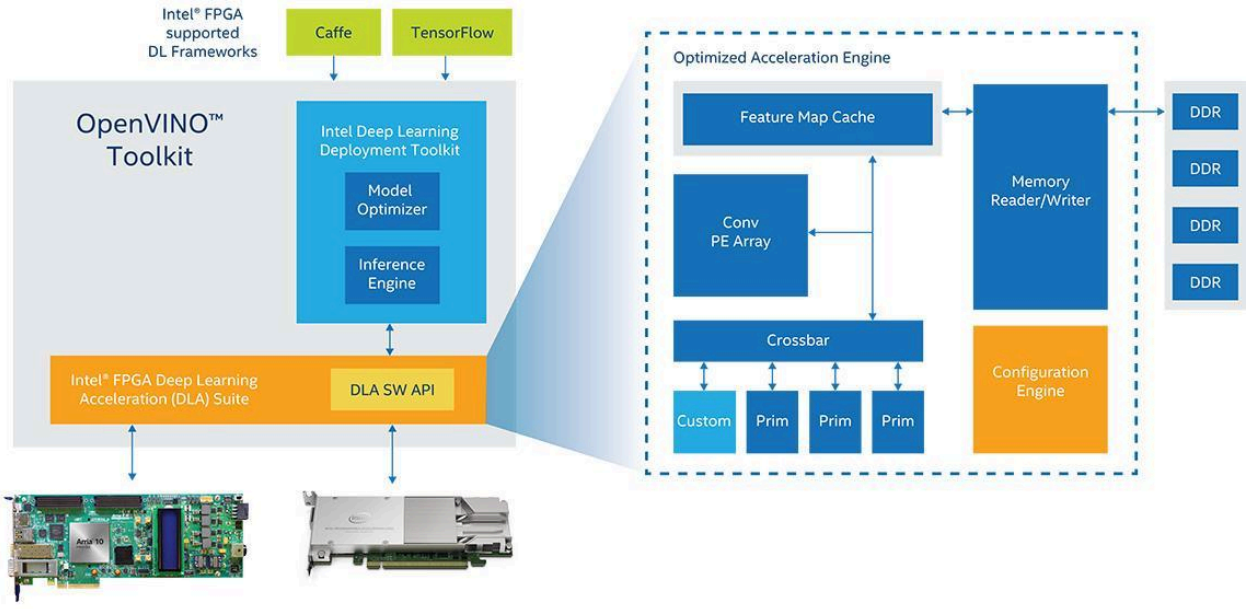


Figure 3 AI inference for vision solution using Intel FPGA supported DL frameworks source (4)

The company has also continued to hedge its competitive edge with AI optimized chipsets architecture categories such as neural network processors (Nervana), computer vision ASCIs (Movidious, autonomous-vehicle ASCIs (mobilEye), and FPGAs (Altera). The company is also aimed at building a quantum computing chip and self-learning neuromorphic for handling the future AI generation challenges.

Methodologies/Best practices

To improve the performance of AI algorithms, there is the need for high-performance computation in limited memory resources. To fuel the development of AI applications, hardware friendly algorithms, domain-specific architecture, and emerging technologies are necessary. This research paper has focused on addressing the challenges of AI acceleration and neuriophonic computing in three aspects. These aspects are; (1) solving the memory challenges for AI

hardware acceleration, (2) Solving the computing challenges for AI applications, and (3) novel architecture design for neuromorphic computing with the emerging technologies.

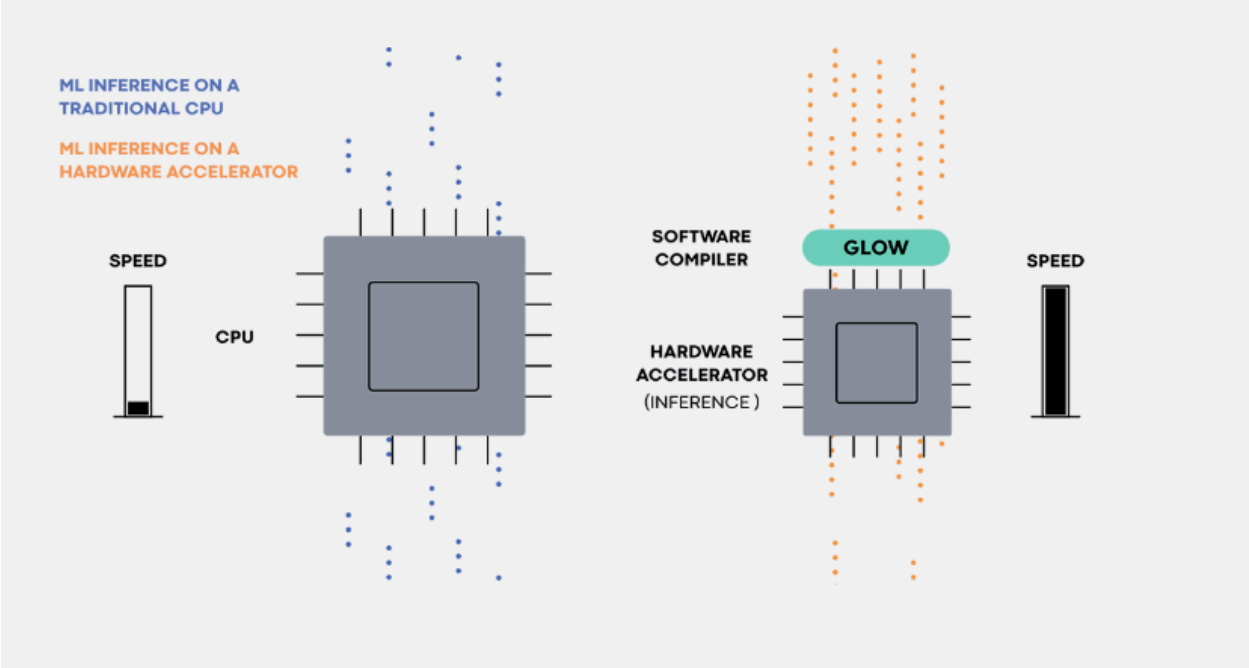


Figure 4 AI hardware acceleration infrastructure (Mensor)

Solving the memory challenges for AI hardware acceleration

Various methods are available to boost energy efficiency and memory performance in the modern AI hardware acceleration platform. Such include process in memory (PIM) accelerator using emerging nonvolatile devices (PRIME). PIM is one of the promising solutions for addressing memory wall challenges in the emerging and the future AI architecture systems. The emergence of metal oxide resistance random access memory (ReRAM) indicating the potential of its use in main memory. HBM enabled GPU for data-intensive applications gives room for a software pipeline to alleviate the limitations in capacity for the HBM for CNN. DRAM-based reconfigurable in-situ processing architecture DRISA which is majorly composed DRAM

memory arrays where each particular memory bit line designed to perform bitwise Boolean logic operations such as NOR and support in situ computing.

Solving the computing challenges for AI applications

Computing challenges for AI hardware acceleration can be solved through various approaches such as instruction set architecture design for a neural network, the use of a novel domain-specific instruction set architecture (ISA) for NN accelerators through a comprehensive analysis of existing NN technologies. SEAL has been on the lead in studying FPGA based Deep Neural network accelerators designed to attain high-level performance while maintaining low power consumption. Deep learning accelerator unit accelerator (DLAU) employs three pipelined processing units to utilize tile technique and improve throughput for deep learning techniques. Software optimization techniques such as model compression by Delphi lead to significant performance improvements and power saving AI architecture products. CNNLab provides a novel deep learning framework using FPGA and GPU based accelerators. The platform provides a uniform programming model for users in a way that the AI hardware implementation and scheduling are not visible to the programmers. The heterogeneous architecture HEMERA is considered the most ideal for achieving the required energy efficient and improved performance. A runtime framework is also proposed to manage acceleration of different workloads efficiency. A runtime framework is also essential for effective management of acceleration in various workloads.

Novel architecture design for neuromorphic computing with the emerging technologies

Neuromorphic computing learns from the human brain through high energy efficient information processing capability. The scalable energy-efficient architecture Lab (SEAL) has proposed a number of novel architectures for neuromorphic computing that enabled by the

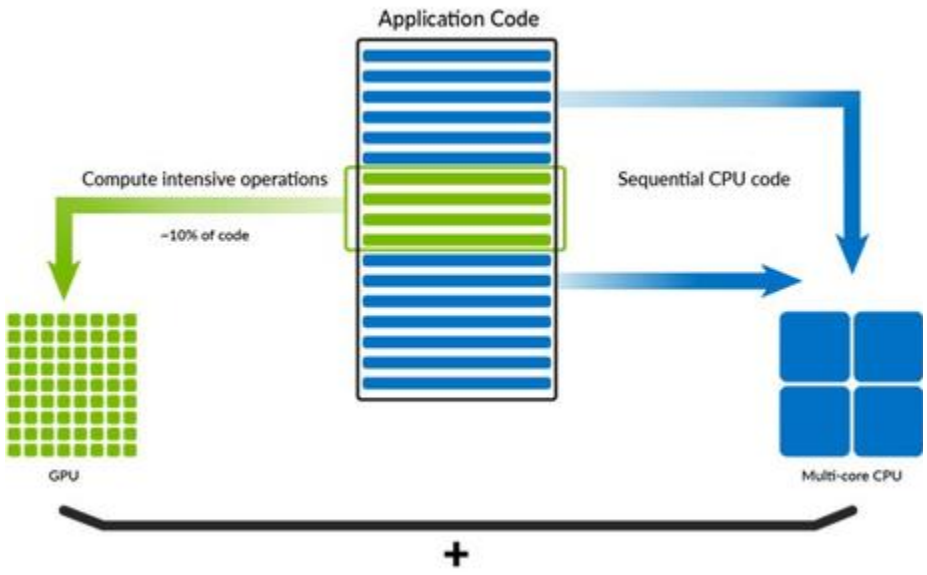
emerging circuit technologies and devices. The simulation platform for the memristor-based neuromorphic system called MNSIM. The system is a hierarchical structure for memristor-based neuromorphic computing acceleration to provide a flexible interface for customization. Neural network transformation and co-design under neuromorphic hardware constraints is another efficient novel architecture design for neuromorphic. A toolset referred to as neural network transformation mapping and simulation (NEUTRAMS) which includes three key components including a configurable clock driven simulator of neuromorphic chips, a neural network transformation algorithm, and an optimized runtime tool that maps neural networks onto the target hardware for better resource utilization. More study should be done on how to utilize 3D technology in a model that will efficiently support Neural Networks.

Experiments, Use cases, Examples

Hardware acceleration for the artificial intelligence is so diverse and evolving so rapidly that it is a challenge to keep up with ever emerging innovation in the market. Apart from the core AI chipset manufacturers such as Nvidia and Intel ASCIs for specific based platforms, several new items have also emerged. Google has developed a special NNPU, a tensor processing unit which is available for the AI apps on Google Cloud Platform. Microsoft is also setting up an AI chip for use in its HoloLens augmented reality headset as observed in the Forbes report by (2). Amazon is also on the leading technologies with its reportedly working on an AI chip for its Alexa home assistance. Tesla is also building an AI processor for their self-driven electric cars and Apple is also working on an AI processor that is designed for powering Siri and Face ID. Common AI hardware accelerator technologies are discussed below

GPU

The Graphics Processing Unit (GPU) is a single chip processor whose primary role is to manage and accelerate the performance of graphics and video. Some features of GPU include 3D or 2D graphics, Rendering polygons, digital output for monitor display, and application support for hyperactive graphics software. In 1999 Nvidia released the GeForce 256 GPU model that could process 10 million polygons per second and was comprised of more than 22 million transistors according to a journal report by (3). In (1), it was also noted that the GeForce has 256 single chip processor with an integrated transform, lighting effects, drawing and BitBLT support, triangle clipping and rendering engines. However, (2) notes that although users could achieve an unprecedented performance of over 100X as compared to the ordinary CPU, there was a challenge in the use of GPGPU graphics programming API's like OpenGL and Cg to program the GPU. This limitation to the fdprogramming abilities limited the accessibility to the tremendous capabilities of GPU for scientific applications.



igure 5 How GPU acceleration works ((11)

GPUs programming limitations inspired the innovation of the compute unified Device Architecture by Nvidia as a parallel programming platform. The manufacture of millions of CUDA enabled GPUs have enabled software developers, researchers and scientists find broad-ranging uses for GPU computing with CUDA. Examples of their various uses include; identifying hidden plaque in arteries, analyze air traffic flow, and visualize molecules. GPUs have been used to stimulate blood flow and identify the hidden arterial plaque without the use of exploratory surgery or the invasive imaging techniques. Computer models have also been useful in alleviating congestions and control airplane traffic move freely. The computational power of GPU has also been useful to the NASA in obtaining large performance gain, and reduces the analysis time from ten minutes to two seconds. The molecular simulation referred to as NAMD (nanoscale molecular dynamics) has improved the performance boost from the use of GPUs. As noted by Mensor, The parallel architecture of the GPUs have speeded up the performance and enabled NAMD developers to port computer intensive portions of the application to the GPU using the CUDA toolkit.

FPGA

FPGAs have been around since the 1980s. As opposed to other chips, they are programmable on demand. FPGAs can be thought of as boards containing low-level chip basics such as AND and OR gates as noted by (1). These chips are configured using hardware description language (HDL) fundamentals. The chips are programmed to match specifications of a dedicated application or task. Intel was originally the interested party in FPGAs by acquiring Altera, the key manufacturer. Intel recently completed a research on two generations of Intel FPGAs; Intel Arria10 and Intel Stratix10, while Nvidia manufactured the Titan X pascal GPU,>

the Intel Stratix 10 FPGA outperformed the GPU and at the same time used pruned compact data sets as compared to the full 32-bit floating point data (FP32).

ASICs

Nirvana's first generation of AI ASIC called lake crest became one of the most reliable custom designs for AI accelerators. The machine learning AI platform called neuASIC was released on June 2018 by a California based Esilicon Company called Santa Clara. According to the intel AI department head, the first commercial AI ASIC called the NNP-L1000 will be launched in 2019 according to (9). ASICs are similar to the GPUs except that ASICs offer an instruction set and libraries that allow the GPU to be programmed in such a way that it can operate on data stored locally to accelerate other parallel algorithms. However, designing ASICs can be an expensive endeavor which in addition requires highly qualified and expensive engineers. The chip would also require frequent updates to keep up with the emerging technologies and manufacturing process.

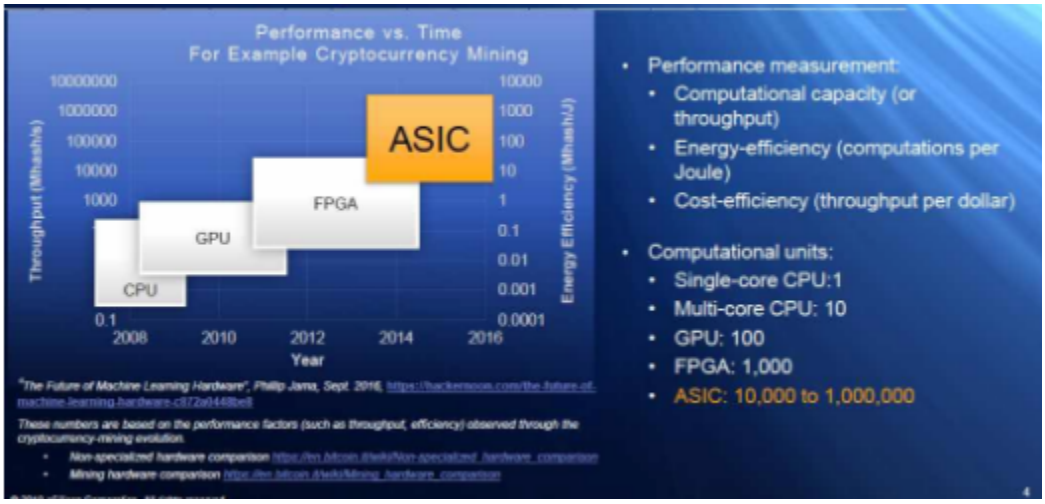


Figure 6

The ASIC advantage source ((8))

Heterogeneous computing

Heterogeneous computing refers to the systems that can use more than one core or processors. According to the report presented in Zahran, It plays an important role in high

performance and power efficient embedded systems dedicated to performing various data processing tasks such as computer vision. FPGAs are used for accelerating the application where it is used as a core processor for standard CPUs. However, the problem of interfacing the FPGAs and CPUs is yet to be completely solved although the process of hardware design has become comparatively easier through the use of High-Level Synthesis tools.

Heterogeneous computing is all about complexity. It is more complex and requires specialized skills in both hardware and software design in all platforms. It also requires a balanced approach to ensure that the right task is performed by the appropriate processor. This process requires significant hardware and software development time and architecture. Example of heterogeneous computing is the Smartphone chip or system on chip (SOC) that implies multiple specialized integrated processing cores with each dedicated to a specific task such as 3D, general computing, video capture, video display, camera, gestures, music, connectivity, and sensors (Mohead 2).

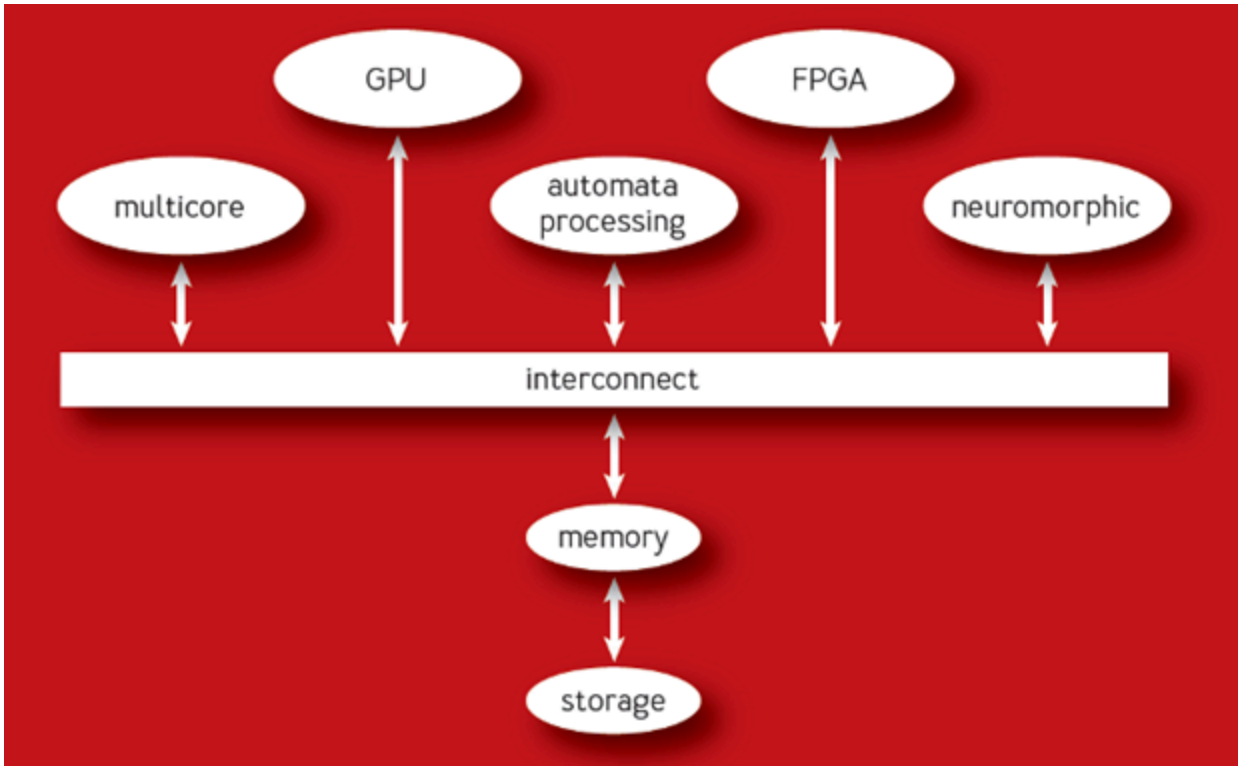


Figure 7 How Heterogeneous computing work source ((3))

Virtual machines and environments for NN acceleration

Deep machine learning in a high-level automated system is based on statistical machine learning techniques and operates in two-stage processes; the deep neural network (DNN) model that is designed specifically for the problem domain and data available is trained on a GPU or high-performance CPU for a few weeks (Kumar, Severtson and Glonrund). The DNN is then deployed into a production environment to take in a continuous stream of input data and run the inference in real time. The DNN yields output either directly as the end result or further fed into downstream systems.

Nvidia Volta/Tesla application for NN acceleration

Nvidia Volta/Tesla core GPU is the fastest processor for AI in the world delivering up to 125 teraflops of deep machine learning performance per single chip. The processor can

accelerate hardware in all frameworks such as Caffe2, Kaldi, Cognitive Toolkit, Matlab, Keras, MXNET, Pytorch, TensorFlow, Chainer, and paddle. The Nvidia GPUs work with the rapidly expanding environments of RNN, CNN, RL, GAN as indicated in report (12) and hybrid network architecture.

Conclusion

In modern human society, artificial intelligence algorithms have impacted an evolutionary influence. AI has covered a wide range of applications ranging from image recognition, speech recognition and comprehension, traffic control, data analytics, and robotics intelligence. Hardware acceleration in artificial intelligence applications has been spearheaded by technology giant companies such as Nvidia, Intel, Google, and Microsoft among others. Various companies have come up with their versions of hardware acceleration technologies such as GPU, FPGA, ASICs, among others. These technologies have been useful in machine learning, language recognition, vision, video processing. Efforts have also been made to make machines that can learn their environment on their own although majority learns from human brain and the existing CPUs, and processing environments. Merging some of these technologies such as FPGAs with the CPU has become a challenge. However, the Nvidia volta/ tesla core GPU has proved to be the most robust AI in the world. However, the future of AI hardware acceleration remains uncertain though promising. Some of the notable AI applications have been voice recognition, machine learning, and computer vision which have enabled the construction of self-driven cars, intelligent Smartphone, and automated manufacturing processes.

Works cited

1. Anadiotis, George. *AI chips for big data and machine learning: GPUs, FPGAs, and hard choices in the cloud and on-premise*. 20 August 2018. 15 October 2018.
2. Freund, Karl. "Will ASIC Chips Become The Next Big Thing In AI?" *Forbes* 04 August 2017: 1.
3. Holzinger, Marc ReichenPhilipp, et al. "Heterogeneous Computing Utilizing FPGAs." *Journal of Signal Processing Systems* 31 May 2018: 1-13.
4. Intel. "AI Inferencing for Computer Vision Solutions using OpenVINO™ Toolkit." 2018. *Accelerated deep learning inference for powerful data.* 2018.
5. Kumar, Gopi, Brad Severtson and CJ Glonrund. "Introduction to the Deep Learning Virtual Machine." 03 June 2018. *icrosoft Azure.* 15 October 2018.

6. Mensor, Steve. *How eFPGAs will help build the brave new world of AI*. 23 April 2018. 23 April 2018.
7. Mohead, Patrick. "Tech Giants: Move To Specialized Computing Or Die." 06 August 2018. *MOOR insight and strategy.* 06 October 2018.
8. Murphy, Bernard. *Platform ASICs Target Datacenters, AI*. Conference. NY: IEEE, 2018.
9. Quach, Katyanna. "Intel's latest promise: Our first AI ASIC chips will arrive in 2019." *The Register* 23 May 2018: 1.
10. syncedreview. "Deep Learning in Real Time—Inference Acceleration and Continuous Training." 20 August 2017. *AI Technology & Industry Review.* 15 October 2018.
11. Yoshida, Junko. "AI Comes to ASICs in Data Centers." *EETimes* 06 June 2018: 1.
12. Zhu, Maohua, et al. *Performance evaluation and optimization of HBM-Enabled GPU for data-intensive applications*. Design, Automation & Test in Europe Conference & Exhibition. Lausanne, Switzerland: IEEE, 2017.