An Overview of Artificial Intelligence Accelerators

Harshita Chadha

The George Washington University

CSCI 6461

Abstract

Artificial intelligence (AI) is a field of computer science that, due to its large scale utility, is becoming fast ubiquitous in today's world. While algorithmic inventions and new software paradigms have had an important part to play in this recent AI revolution, one of the major factors that have allowed this technology to permeate everyday life is the improvement in underlying computer hardware. With the advent of time, new heterogenous computer processor architectures were adopted to better suit the demands of AI based algorithms. Most of these involved the adoption of specialised chips that were designed not to perform general purpose tasks but be efficient in their processing of specific AI related operations such as matrix and vector multiplications. Such processors are rightly entitled AI accelerators as they bring about optimization in the execution of these algorithms. The objective of this article is to study in detail AI accelerators and the evolution of traditional processor systems into AI optimised heterogenous systems of today. A brief survey of existing industry processors is presented as well along with recent metric trends.

An Overview of Artificial Intelligence Accelerators

Artificial Intelligence is that branch of computer science that deals with the development of computer systems that emulate cognitive behavior to simulate complex abilities such as perception of vision, recognition of speech, logical deduction, reasoning, etc. The scientific work that can now be recognized as the earliest example of the demonstration of artificial intelligence fundamentals is the McCulloh and Pitts (1943) artificial neuron that characterized a formal turing complete design. Today, this technology is ubiquitous and intimately intertwined with the human experience. The salient ideas of artificial intelligence and associated fields have been here for more than 50 years. But despite the obvious utility of this technology their use only become prominent in everyday life in the past decade.

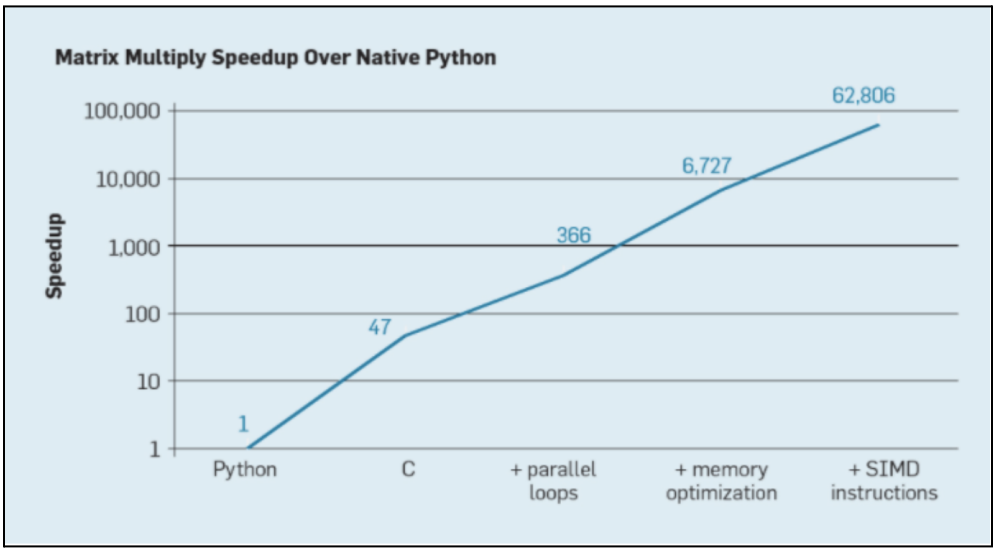
This latency can be attributed, in part, to the lack of computer infrastructure equipped to deal with the massive amount of processing power required by artificial neural networks to generate useful results. Thus, while algorithmic innovations were at the center of the AI revolution, one of the most important contributors to the success of the field was the improvements in the underlying architecture capable of performing the required calculations, given algorithmic constraints. For instance, it is a widely held belief that the research work titled Alexnet as presented by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton (2017) from the University of Toronto rang in a new era for the AI community. They utilized a specialized hardware unit - the GPU to compute the neural network more time efficiently. While their algorithm itself was not revolutionary, their creation of an innovative mapping of the convolution operations set their solution apart. Hence, without the ability to learn complex data relations in a reasonable time frame - a capability imparted because of the improvement in the way the

networks are computed - we would perhaps not have seen such widespread adoption of AI techniques, useful as they may be.

Changing the way in which an AI algorithm is implemented makes the algorithm run more efficiently and saves time especially when it comes to massive real-life computations. An example of this can be demonstrated by taking the case of matrix multiplication - a ubiquitous operation in AI. In Figure 1 we see how using different computation methods to perform this operation can help improve processor performance.

Figure 1

Trends in processor performance for matrix multiplication using multiple techniques (Hennessy, 2019)



This is where AI accelerators come into play. An artificial intelligence accelerator may be defined as a high-efficiency computing device that is capable of handling large-scale neural network workloads due to its specialized design and parallel processing capabilities. These are specialized hardware units present in computer systems that help speed up AI operations by virtue of the highly specific nature of their architecture.

They are a special category of hardware accelerators - processors that are designed to increase the efficiency of the CPU by taking away some of the specialized task load. This helps in increasing the performance of the computer and the efficiency with which the task is executed. Hardware accelerators combine the flexibility that is afforded by the presence of general-purpose CPUs with the performance efficiency that specialized hardware offers to optimize task performance. A few examples of processors being used as AI accelerators are GPU or graphical processing units that are most adept at dealing with parallelized processing of massive amounts of data. Other devices include FPGAs or field programmable gate arrays and ASICs or application-specific integrated circuits.

In this article, a comprehensive survey of the field of artificial intelligence accelerators has been undertaken. The first section highlights the pivotal role these specialized processors have played in the progress of the field of artificial intelligence. Following this, the evolution of computer processors and the advancements that brought about the need for the development of accelerators has been traced. Further, a detailed discussion of the major categories of AI accelerators has been conducted. The last two sections of the article present a detailed overview of the various leading industry processors that are driving innovations in the field of AI and summarise the general industry trends prevalent.

Evolution of Processors and Accelerators

A processor is the most integral part of the computer system - one that actually performs computations on a set of inputs to produce useful results. They usually are composed of some arithmetic and logical unit that takes the charge of executing the instructions to generate computational outputs. With the advent of modern technology and the increasing complexity of applications, we are in constant need of creating better processors. In the beginning, the focus

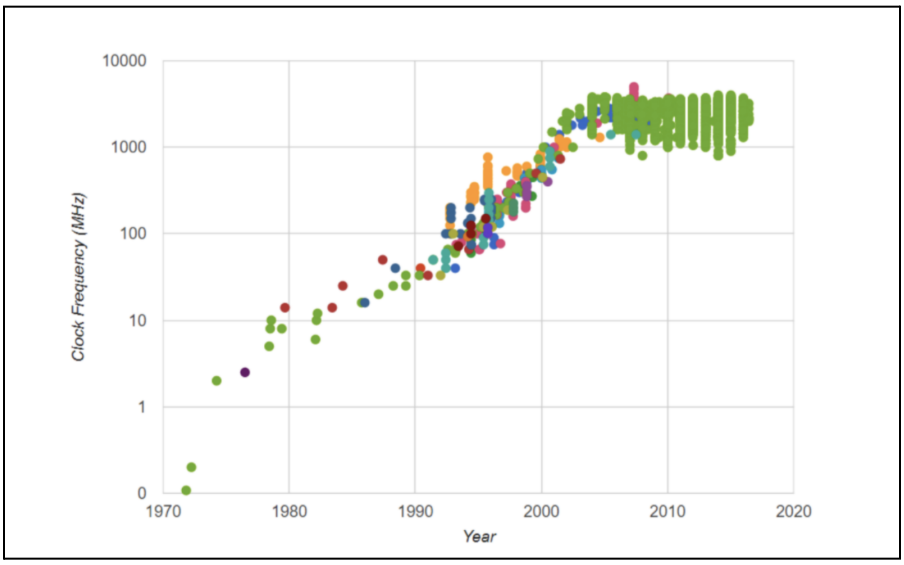
was on creating devices that took less time to generate results but with the progression of time, these efforts were refocused towards creating CPUs that were more efficient. In other words, the emphasis on the time constraint was lifted and was instead placed on energy efficiency and cost optimization. In the past few decades, we have come from using large room-sized computers to microchips with exponentially more processing power. The reason behind this growth is the headways in the semiconductor industry.

One of the key building blocks of continued processor improvement is transistors whose evolution was, for a long duration, governed by Moore's Law. According to this law, every year the number of transistors that can be fit onto an integrated chip or IC doubles. This allowed computer architects to, for a time, downscale transistors so as to fit more of these devices onto processor chips and thereby extract increased computational power. This law lasted roughly 5 decades and then hit a plateau as there is a lower limit to the size of transistors. Another important law governing processors is Dennard's Scaling Law which observed that due to downscaling of transistors by a factor K, if K^2 more transistors are fit onto a processor, then the processing power increases by a factor of K but the power consumption does not go up since electrons are traveling much smaller distances due to increased transistor density on chips. Around the turn of the 21st century, however, this law too hit a plateau. Due to the increasing power density rate, the chips would theoretically end up becoming unsustainably hot and require an additional cooling setup that added costs in terms of invention and setup. Thus, increasing the density stopped being a viable solution to increase CPU frequency.

As stated earlier, initially, the focus was on creating high-frequency general-purpose CPUs so the scaling laws - both Denard's and Moore's - were fully exploited to realize this. A plot of this improvement over the years can be seen in Figure 2.

Figure 2

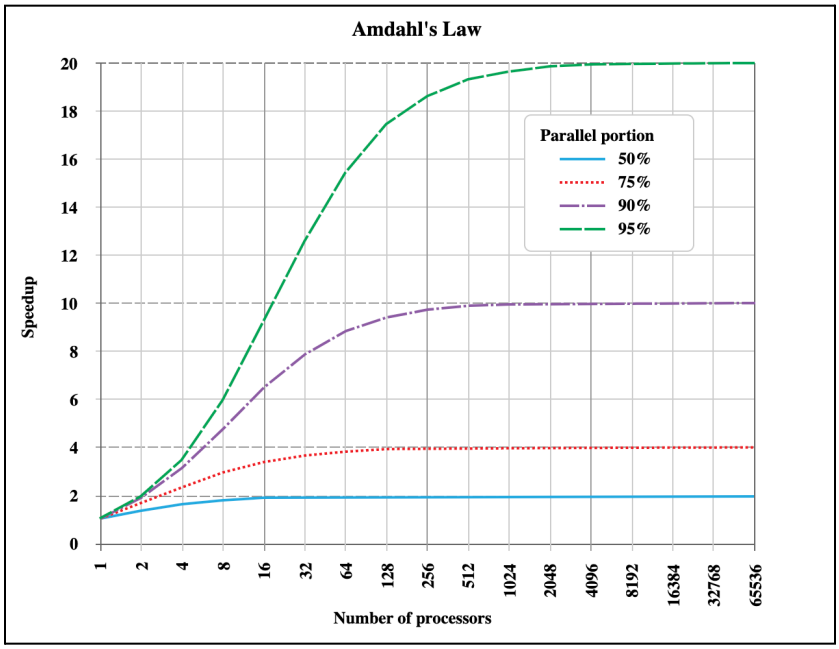
Evolution in processor frequency over the years (Fuchs, 2022)



The non-applicability of the scaling laws gave rise to the age of multicore processors and concepts like multi-threading and parallelism were adopted. One example of this multicore system is the combined use of a GPU and CPU. These concepts, however, had an upper limit when it came to the amount of parallelism that could be done. In accordance with Amdahl's Law which emulates the law of diminishing returns, it is important to carefully study the returns that are obtained by adding more processors to the machine. The law was named after its inventor computer scientist Gene Amdahl(a computer architect from IBM and Amdahl corporation) and was presented at the AFIPS Spring Joint Computer Conference in 1967. It is also known as Amdahl's argument and helps to calculate the processing latency reduction that would materialize by utilizing multiple processing cores. A visualization of this concept in Figure 3.

Figure 3

Speedup Vs Number of Processors in accordance with Amdahl's Law (Wikipedia contributors, 2022)



One of the major implications of Amdahl's law is that if you have N number of processors running the same task in parallel, then adding more processors of the same kind adds less usable power to the system and as a consequence the speedup ratio, define as the ratio of the old unenhanced execution time to the new enhanced execution time, diminishes. Thus, to achieve optimality, the solution is to employ heterogeneous computing techniques instead of just bunching the same identical processors together and hoping for the best outcome. In other words, since adding more transistors was no longer an option, we moved towards specializing processors and using a certain category of processors to perform specific tasks instead of having one all-encompassing general-purpose CPU and thus hardware accelerators came into the picture.

Heterogenous computing employs a number of distinct co-processors capable of adding special computational capabilities to perform some specific tasks for instance - computation of

neural networks and this is what AI accelerator systems are. One of the many ways in which accelerator-based heterogeneous systems are better than traditional general-purpose architecture is in their handling of data. Von Neuman architecture (or even the improved Harvard architecture) that most CPU systems are based on suffers from memory fetch latency problems that cause processors to sit idle. AI accelerators overcome this disadvantage via their in-memory nature which means that they do not have read-write cycles from secondary memory and the memory action is done entirely and directly on the RAM.

In fact, the AI accelerator's boost of performance is so utilitarian that major programming languages such as python now support many dedicated libraries to exploit this functionality and many mainstream corporations are making use of these libraries to compute useful information. For instance, the German automobile manufacturing company Volkswagon or VW makes use of python's special libraries along with GPU's AI acceleration to calculate customer churn rate (Viswanath, 2018).

Now since AI and big data often go hand in hand, it may be argued that the best way to process such large amounts of information is with large machines made up of multiple processors housed within data centers with state-of-the-art computing capabilities. However, because many AI applications are contingent on IoT systems, such dependency spells out issues such as transportation latency, data security, etc. Hence, it may be argued that for most practical applications, it is better to process the massive amounts of data being generated on the edge, closer to devices that sense it. This fact warrants the need for hardware accelerating chips that can supplement the capabilities of a CPU to help process data and is driving much of the recent advances in accelerator innovation.

Classification of Artificial Intelligence Acceleration Systems

Based upon the nature of their capabilities, artificial intelligence accelerators may be classified into four major categories namely, graphics processing units or GPUs, visual processing units or VPUs, field programmable gate array or FPGAs and applications specified integrated circuits or ASICs. Each of these categories and their unique underlying architecture has been explored in detail in the sections that follow.

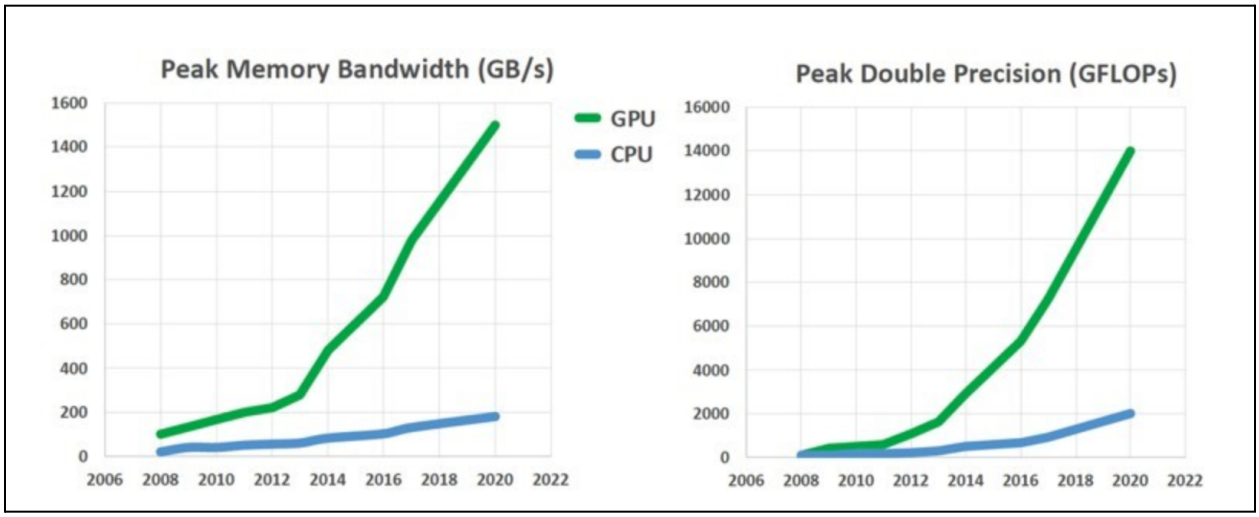
In addition to these four major categories, another classification of accelerators based on the nature of the computation they perform can also be done. The two major categories that arise are inference-based accelerators and training-based accelerators. Inference-based accelerators are typically low-power-consuming processors that store the biases of a pre-trained neural network as constants and help make predictions in real time with minimal latency. Whereas training-based accelerators are powerful processors capable of handling and processing large-scale data to generate the weights and biases that inference accelerators use in their computations.

Graphics Processing Units or GPUs

GPU or graphics processing units are the most ubiquitous artificial intelligence accelerators that exist today. These were initially designed for rendering computer graphics on devices and exploiting parallelism to increase system throughput but nowadays are increasingly turning into programmable processors for parallel data parsing. Today, there exist several state-of-the-art GPU processors in the market that can be utilized to optimize the training of deep neural networks for artificial intelligence implementations. The utilization of GPUs as processors put an end to Moore's law and the CPU optimization plateau implied by it by increasing the processing power manifold. Figure 4 illustrates this fact.

Figure 4

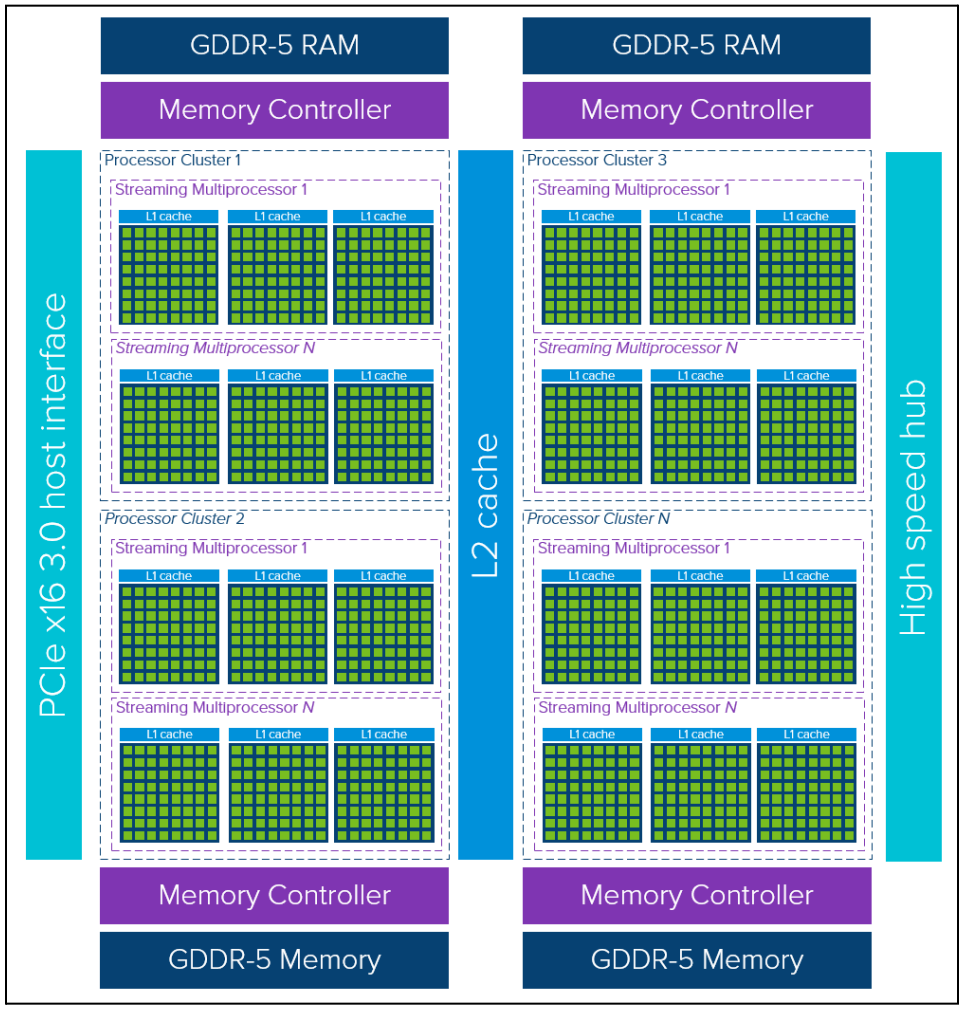
Illustration of GPU-based heterogeneous processors ending Moore's plateau (Wikipedia, 2022)



The architecture of a standard GPU with N processor clusters, represented as a PC, is illustrated in figure 4. Each of these N clusters constitutes n number of streaming multiprocessors represented as SM. Each of these N streaming multiprocessors has assigned to it an L1 cache that contains instructions. The L2 cache is a shared resource amongst the streaming multiprocessors. The data to be processed is grabbed directly from the GDDR-5 RAM. As can be observed from the architecture, the number of memory cache layers in a typical GPU architecture is lesser in comparison to typical general-purpose CPUs. This is because, evidently, GPUs pay more attention to the computation of the data and care less about the latency of data retrieval. Since these devices also employ massive-scale parallelization, the latency caused due to reduced cache layers is made up for.

Figure 5

Architecture of a typical GPU (Hagoort, n.d.)



Thus GPUs support high-performance computing and acceleration of artificial intelligence operations due to the large-scale parallelism that their architecture provides. However, these are still general-purpose devices that are, in most cases, interfaced (eg - CUDA in the case of NVIDIA), to apply for AI applications.

Vision Processing Unit or VPUs

Another class of processors called the vision processing units are also being used to accelerate artificial intelligence algorithms. As their name suggests, these processors are

increasingly being applied in the sector of computer vision and provide means to efficiently process graphical data and aid in image processing problems. Similar to a GPU, VPUs also take the load off the CPU to perform tasks specific to computer vision computations but are more efficient and less power intensive than GPUs. Visual processing units are analogous to system-on-chip and help in the efficient collection and processing of spatial data such as images.

These processors are a perfect solution for edge devices wherein the application of machine learning must be done remotely and within system processing constraints. An example of their power is Intel's Movidius Myriad 2 VPU which when interfaced with a CMOS image sensor can take an image, preprocess it, run it through a pre-trained neural network, and output results all the while only consuming about 1 W of energy.

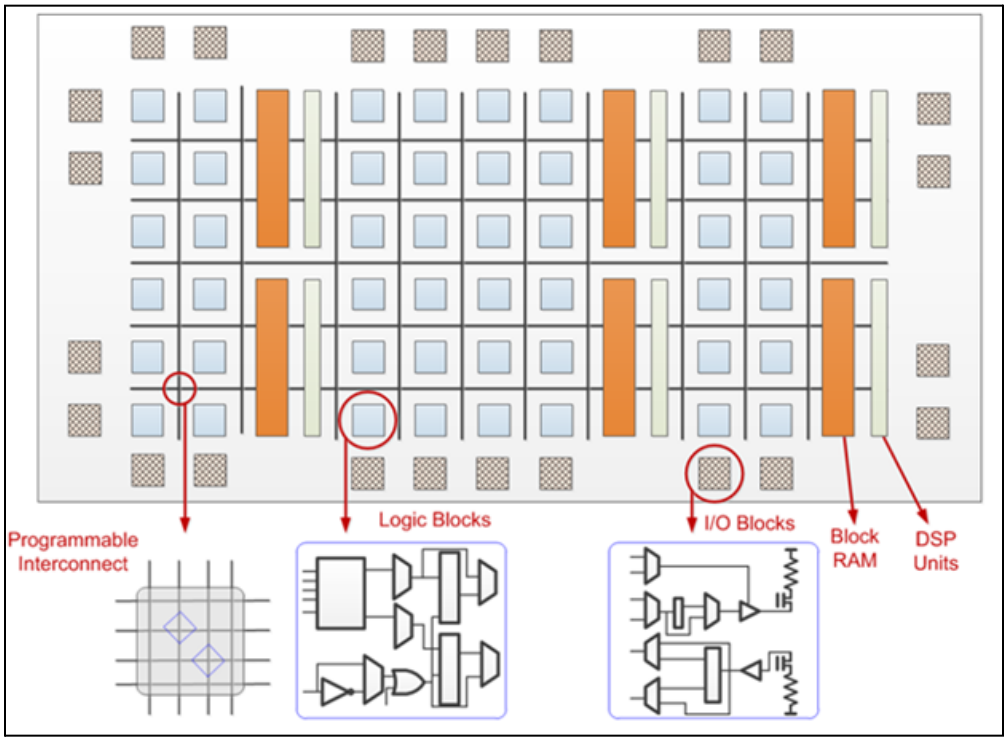
VPUs perform low-precision floating point computation for image processing tasks and just like manycore digital signal processors, their architecture plays emphasis on optimizing the flow of data between the many parallelly executing operations via a scratchpad memory.

Field Programmable Gate Array or FPGA

Field Programmable Gate Array or FPGA is a category of artificial intelligence accelerators that is reprogrammable N number of times. Any arbitrary equation corresponding to a neural network can be represented in the boolean form and then implemented. They can represent combinational as well as sequential logic i.e, clock registers can also be simulated in addition to the boolean form of equations. The modern-day FPGA architecture rests upon 3 salient building blocks namely the interconnect architecture, the I/O block, and the CLB or configurable logic block. These are illustrated in Figure 6.

Figure 6

Architecture of a typical FPGA (HardwareBee, 2021)



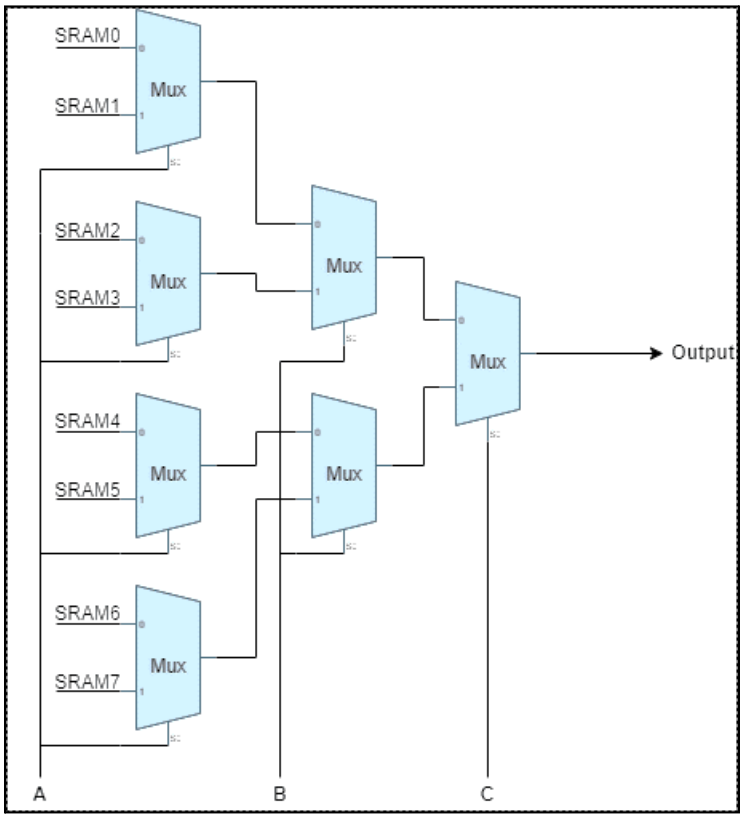
The configurable logic block is made up of 3 main elements. These are multiplexers, flip flops, and a look-up table which is the main element responsible for implementing the logical function. The multiplexers and flip flops are used to choose between combinational logic or sequential logic's data output and implementation of sequential logic respectively. The lookup table or the LUT is the most important component of the FPGA architecture as this is the part that helps implement any equation thereby giving these processors their reprogrammable ability. The LUT is composed of multiplexers and SRAM cells. A sample lookup table architecture is shown in Figure 6.

The I/O block, as its name suggests, is the block that can be used for data interfacing. All of these three salient components of the FPGA are laid out with respect to an interconnected grid that ensures communication within the processor. This interconnection is a matrix-like grid made

up of wiring and various switches that connects together the I/O black, the logic blocks, and other FPGA components.

Figure 7

FPGA Look-up Table (LUT) (HardwareBee, 2021)



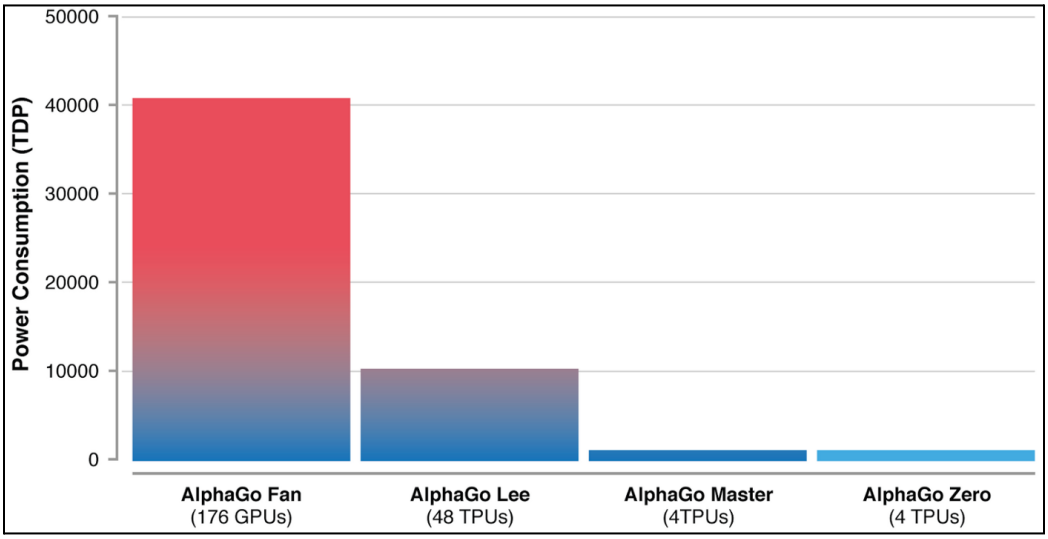
As AI accelerators, FPGAs offer minimal latency and high throughput. They combine optimal performance with great flexibility because of their reprogrammable nature. They are also extremely cost-effective and have low power consumption characteristics. They are highly efficient when it comes to overcoming memory buffering bottlenecks and are sensor fusion enabled making them extremely integral in terms of the modern AI landscape.

Application Specified Integrated Circuit or ASIC

ASIC stands for application-specified integrated circuit which is a highly specialized chip designed to perform a specific set of operations. This is in direct contrast with most modern general-purpose CPUs as well as the FPGAs that are reprogrammable. Since most artificial intelligence algorithms involve performing a given fixed set of operations on massive data, ASIC chips that are designed specifically to handle the processing of data for a specific AI task are highly beneficial and effective.

Figure 8

Performance evolution of ASIC chips (AlphaGo Zero, n.d.)



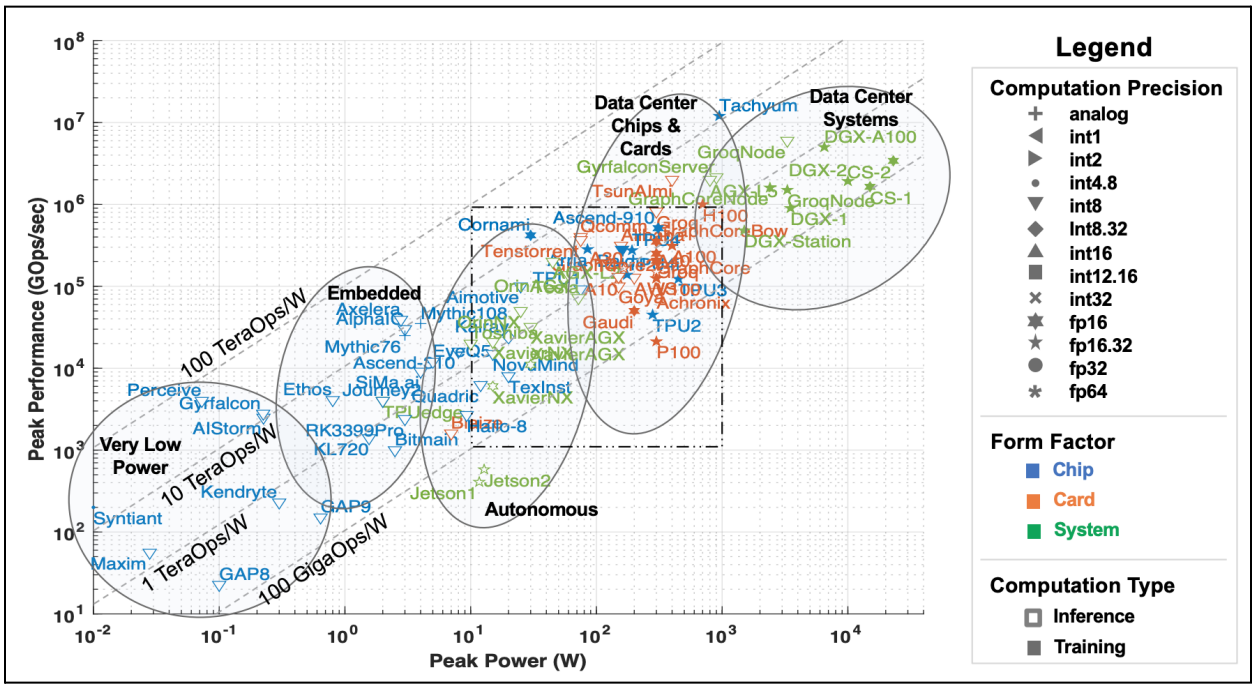
What they lack in rigidity, ASICs make up for in performance by consuming low amounts of energy to perform highly specialized tasks. An example of this capability is illustrated in figure 8 using a TPU or tensor processing unit, a very common and up-and-coming example of ASIC chips (Sapunov, 2021).

Accelerators Abound - The Industry Scenario

In this section a comprehensive overview of industry accelerators produced in the last decade has been provided. Figure 9 has been adopted from Reuther et al's (2022) yearly survey and illustrates a plot of the performance metrics for notable accelerator processors manufactured in the past few years. In the plot, as mentioned in the legend, the shapes represent distinct precision levels, the color of the plot point represents the form factor, and the hollowness represents whether the chip is intended for inference or training operations. In the text that follows, a detailed overview of a select few most notable and recent processors depicted in the graph has been carried out.

Figure 9

Peak Performance VS Power plot for Industrial Accelerator Chips (Michaleas et al, 2022)



In 2012, IBM in collaboration with scholars from Yale University and New York University created a vision SoC (system-on-a-chip) entitled NeuFlow by drawing inspiration

from biology (Pham et al, 2012). This chip was primarily designed to accelerate computer vision tasks that usually involve large and complex matrix manipulation operations. As reported by the creators of the chip, the average power consumption was around 0.6W and the system is capable of delivering 320 Giga operations per second.

In 2016, Han et al proposed an energy-efficient inference engine or EIE that was capable of accelerating multiplication matric operations by using the process of weight sharing using SRAM (Han et al, 2016). The processing power of this chip was stated as being 102 Giga operations per second and the inference-based chip can compute the fully connected layers of AlexNet while consuming only 600mW of power. In 2017, Gao et al at Stanford proposed TETRIS which used 3D memory and a high-density partitioning-based architecture that freed up chip infrastructure that could now be used to perform computations - a direct contrast to local SRAM-based systems (Gao et al, 2017).

Chen et al proposed Eyeriss - an accelerator chip that is used to optimize the computation of convolutional neural network inference by enhancing energy efficiency (Chen et al, 2017). It constitutes a 168-processor big spatial array and works to minimize data movement by the concept of reuse. A series of dataflow-based inference chips originating at the Chinese Academy of Sciences termed the Diano series was made up of 4 distinct inference chips each designed for an application-specific machine learning process (Chen et al, 2016). TruNorth is a CMOS-based IC created using the principles of neuromorphic engineering (Esser et al, 2016). It constitutes a processor network based on the manycore paradigm with a basic transistor count of 5.4 billion.

In 2017, Intel launched the Movidius Myriad X processor (Hruska, 2017) which is a deep learning processor capable of performing inference operations at the edge and achieving a performance metric of about 1 trillion operations per second. It is a visual processing unit and

comes bundled with a neural engine to easily perform computer vision-related tasks. Google Edge TPU is a processor (Google Cloud, n.d.) created specifically for inference applications that are embedded in nature and makes use of low-precision parameters. The Kneron KL520 accelerator (Asia, 2020) is a chip specifically designed for edge-based internet of thing inference applications. Mythic's Intelligent processing unit accelerator (Fick, 2018) was created specifically for embedded, computer vision, and data center-based applications and uses a RISC-V-based control processor with flash memory to perform matrix computations.

In 2020, XMOS adapted its existing processing chip entitles XCore (Ward-Foxton, 2020) to include support for machine learning with an aim to provide low-power support for simple applications such as voice-based keyword detection. ARM's Ethos series of ML processors namely the Ethos-N37, Ethos-N57, and Ethos-N77 (Schor, 2021). These neural processing units are composed of the same underlying microarchitecture with slight differences in the configuration of the implements. The most powerful processor of the family, the Ethos-N77, is capable of performing 1 terabyte worth of operations in a second at the frequency of 1.0 GHz.

The NovuTensor Chip by the company NovuMind is designed for neural inference instead of training and is most suited for data center applications. It is a 28 nm CMOS-based chip that has co-processors making it suitable for inference of vision (EETimes, 2018). The Lighspeeur 5801 designed by Gyrfalcon (Word-Foxton, 2019) has an in-memory design to facilitate matric processing for inference tasks.

Infineon, an Indo-US startup launched the AlphaIC processor (Clarke, 2018) that performs agent-based computations wherein each of the kernels that is computed is paired up with an agent. The Israeli AI startup company Hailo released its first chip Hailo-8 which is an

edge inference chip that displays an impressive performance with a power efficiency of 2.8 TOPS/ Watt (Ward-Foxton, 2019).

The CloudBlazer T10 processor (Clarke, 2019) is a data center training card that supports a large range of precision values. It made its debut in the year 2019 and has a maximum power consumption of 225W. In 2020, Untether's TSunAImi card (Clancy, 2022) was introduced. It is an in-memory processor with an SRAM array and is aimed at providing efficient inference of neural network algorithms. The TDA4VM chip (Ward-Foxton, 2020) launched by Texas Instruments (TI) is an accelerating processor for autonomous system applications. It is based on the C7x digital signal processor by TI. The EyeQ5 processor (EETimes, 2018) by Mobileye is an autonomous application-based processor with 8 CPU cores and 18 computer vision processors. The Cloud AI100 accelerator (McGrath, 2019) chip was created by Qualcomm and delivers high performance at low power consumption.

NVIDIA, in 2021, launched several new GPU cards for data center applications namely the Ampere A-40, A-10, and A-30 (Morgan, 2021). In 2022, they announced a new chip titled Hopper (H100) which is a GPU accelerator that promises an increased number of tensor cores and enhanced memory bandwidth. Google, in 2021, launched its inference-only chip - the TPV4i AI accelerator (Weiss, 2021). It is available to be interfaced via Google Cloud Compute (GCC). Since then, the company has gone on to announce the TPV4i's successor chip - the TPV4. Cerebras improved upon their wafer scale engine (WSE) accelerator chip and scaled it down to a size of 7nm. The new and improved accelerator is called the CS-2 and constitutes about 8.5 X 10^5 arithmetic units and exhibits a power consumption of up to 23 KW (Cerebras, 2022).

Trends and Conclusions

In this article we performed a comprehensive review of artificial intelligence accelerators which are specialized computer processors that help optimize the execution of AI operations for enhanced performance. The ubiquity of these processors can be attributed to their energy-efficient nature, their support for massive levels of parallelism, their scalability, etc.

The survey of the industrial processors conducted in the section above and figure 9 presents many inferences that help highlight salient industry trends. For instance, for most inference-based processors, Int8 precision is the most common choice. Moreover, the very low-power and embedded processors are system-on-chip solutions that come bundled with add-on features such as analog-to-digital converters, network interfaces, etc. Amongst the chips calibrated specially for autonomous robots and vehicle applications, as well as data centers, the density of the chips, has seen an upwards trend. The high-end, high-capacity chips are increasingly coming bundled with networking capabilities that allow these processors to work in tandem with one another making such systems especially useful for extremely large AI models.

In the past decade, the processing efficiency and the peak power value for processors has seen tremendous growth and the specification released from manufacturers are indicative of the fact that this growth is owed in part to the adaptation of low-precision formats. Generally, it is believed that the more precise the metrics for a model are, the more accurate the predictions. Thus, opting for low-precision processors might seem counterintuitive at first glance. Detailed studies, however, have helped conclude that that is not always the case and there is a tradeoff that exists between precision and performance that if tread carefully can help arrive at a reasonably comfortable balance of values.

Another observable trend is that many recent data flow applications-based accelerators are being designed to implement mathematical kernels that are capable of processing more than just neural networks by virtue of their statically programmed computational hardware. This makes them capable of being used for applications such as computational fluid dynamics, large-scale weather prediction systems, etc.

Thus, in conclusion, it is clear that there exists an abundance of blueprints and possible strategies that can help pioneer new innovation in the sector of AI acceleration and this also is only a fraction of the ideas that exist in academia still yet to be industrially realized.

References

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5(4), 115–133. https://doi.org/10.1007/bf02478259

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017) "ImageNet classification with deep convolutional Neural Networks," Communications of the ACM, 60(6), pp. 84–90. Available at: https://doi.org/10.1145/3065386.

John L Hennessy, D.A.P. (2019) A new golden age for computer architecture, ACM. Available at: https://cacm.acm.org/magazines/2019/2/234352-a-new-golden-age-for-computer-architecture/abstract (Accessed: November 24, 2022).

Fuchs, A. (2022, January 29). AI accelerators - part II: Transistors and pizza (or: Why do we need accelerators)? Medium. Retrieved November 24, 2022, from https://medium.com/@adi.fu7/ai-accelerators-part-ii-transistors-and-pizza-or-why-do-we-need-accelerators-75738642fdaa

How VW predicts churn with GPU-accelerated machine learning and visual analytics. (n.d.). Retrieved November 24, 2022, from https://www.heavy.ai/blog/vw-predicts-churn-with-gpu-accelerated-machine-learning-and-visual-analytics

Hardware acceleration. (2022, July 18). Retrieved November 24, 2022, from https://en.wikipedia.org/wiki/Hardware_acceleration

Farooq, U. (2021, April 12). What is hardware acceleration and when should you use it? Retrieved November 24, 2022, from https://www.makeuseof.com/what-is-hardware-acceleration/

Artificial Intelligence. (2022, November 22). Retrieved November 24, 2022, from https://en.wikipedia.org/wiki/Artificial_intelligence

Asicnorth. (2022, March 18). ASIC vs. FPGA: What's the difference?: ASIC North Inc. Retrieved November 24, 2022, from https://www.asicnorth.com/blog/asic-vs-fpga-difference/

Reuther, A., Michaleas, P., Jones, M., Gadepally, V., Samsi, S., & Kepner, J. (2019). Survey and benchmarking of machine learning accelerators. *2019 IEEE High Performance Extreme Computing Conference (HPEC)*. doi:10.1109/hpec.2019.8916327

What Is an AI Accelerator? – How It Works | Synopsys. (n.d.). https://www.synopsys.com/ai/what-is-an-ai-accelerator.html

Wikipedia contributors. (2022, November 23). *AI accelerator*. Wikipedia. https://en.wikipedia.org/wiki/AI_accelerator

Wikipedia contributors. (2022a, October 27). *Amdahl's law*. Wikipedia. https://en.wikipedia.org/wiki/Amdahl's_law

GeeksforGeeks. (2022, October 4). *Computer Organization | Amdahl's law and its proof*. https://www.geeksforgeeks.org/computer-organization-amdahls-law-and-its-proof/

Wikipedia contributors. (2022a, September 27). *Heterogeneous computing*. Wikipedia. https://en.wikipedia.org/wiki/Heterogeneous_computing

Chawla, V. (2020, December 28). *The Different Types Of Hardware AI Accelerators*. Analytics India Magazine. https://analyticsindiamag.com/the-different-types-of-hardware-ai-accelerators/

Milojicic, D. (2020). Accelerators for Artificial Intelligence and High-Performance Computing. *Computer*, *53*(2), 14–22. https://doi.org/10.1109/mc.2019.2954056

Gupta, N. (2021). Introduction to hardware accelerator systems for artificial intelligence and machine learning. *Advances in Computers*, 1–21. https://doi.org/10.1016/bs.adcom.2020.07.001

Hagoort, N. (n.d.). *Exploring the GPU Architecture | VMware*. The Cloud Platform Tech Zone. https://core.vmware.com/resource/exploring-gpu-architecture

Comparing VPUs, GPUs, and FPGAs for Deep Learning Inference. (n.d.). Teledyne Flir. Retrieved November 24, 2022, from https://www.flir.com/discover/iis/machine-vision/comparing-vpus-gpus-and-fpgas-for-deep-learning-inference/

Wikipedia contributors. (2022a, July 18). *Vision processing unit*. Wikipedia. https://en.wikipedia.org/wiki/Vision_processing_unit

(2021, March 1). *The Ultimate Guide to FPGA Architecture*. HardwareBee. https://hardwarebee.com/the-ultimate-guide-to-fpga-architecture/

FPGA vs. GPU for Deep Learning. (n.d.). Intel. Retrieved November 24, 2022, from https://www.intel.com/content/www/us/en/artificial-intelligence/programmable/fpga-gpu.html

AlphaGo Zero: Starting from scratch. (n.d.). https://www.deepmind.com/blog/alphago-zero-starting-from-scratch

A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi and J. Kepner, "Survey of Machine Learning Accelerators," 2020 IEEE High Performance Extreme Computing Conference (HPEC), 2020, pp. 1-12, doi: 10.1109/HPEC43674.2020.9286149.

L. Du, and Y. Du, "Hardware Accelerator Design for Machine Learning", in Machine Learning - Advanced Techniques and Emerging Applications. London, United Kingdom: IntechOpen, 2017 [Online]. Available: https://www.intechopen.com/chapters/58659 doi: 10.5772/intechopen.72845

Park, H., & Kim, S. (2021). Hardware accelerator systems for artificial intelligence and machine learning. Hardware Accelerator Systems for Artificial Intelligence and Machine Learning, 51–95. doi:10.1016/bs.adcom.2020.11.005

Chen, Y., Xie, Y., Song, L., Chen, F., & Tang, T. (2020). A Survey of Accelerator Architectures for Deep Neural Networks. Engineering. doi:10.1016/j.eng.2020.01.007

Talib, M.A., Majzoub, S., Nasir, Q. *et al.* A systematic literature review on hardware implementation of artificial intelligence algorithms. *J Supercomput* 77, 1897–1938 (2021). https://doi.org/10.1007/s11227-020-03325-8

Mittal, S. A survey of FPGA-based accelerators for convolutional neural networks. *Neural Comput & Applic* 32, 1109–1139 (2020). https://doi.org/10.1007/s00521-018-3761-1

Lee, K. J. (2020). Architecture of neural processing unit for deep neural networks. Advances in Computers. doi:10.1016/bs.adcom.2020.11.001

Michaleas, P., Jones, M., Gadepally, V., Samsi, S., & Kepner, J. (2022). AI and ML Accelerator Survey and Trends. *2022 IEEE High Performance Extreme Computing Conference (HPEC)*. https://doi.org/10.1109/hpec55821.2022.9926331

Sapunov, G. (2021, December 27). *Hardware for Deep Learning. Part 4: ASIC - Intento*. Medium. https://blog.inten.to/hardware-for-deep-learning-part-4-asic-96a542fe6a81

Pham, P. H., Jelaca, D., Farabet, C., Martini, B., LeCun, Y., & Culurciello, E. (2012). NeuFlow: Dataflow vision processing system-on-a-chip. *2012 IEEE 55th International Midwest Symposium on Circuits and Systems (MWSCAS)*. https://doi.org/10.1109/mwscas.2012.6292202

Han, S., Liu, X., Mao, H., Pu, J., Pedram, A., Horowitz, M. A., & Dally, W. J. (2016). EIE: Efficient Inference Engine on Compressed Deep Neural Network. *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*. https://doi.org/10.1109/isca.2016.30

Gao, M., Pu, J., Yang, X., Horowitz, M., & Kozyrakis, C. (2017). TETRIS. *ACM SIGARCH Computer Architecture News*, *45*(1), 751–764. https://doi.org/10.1145/3093337.3037702

Chen, Y. H., Krishna, T., Emer, J. S., & Sze, V. (2017). Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks. *IEEE Journal of Solid-State Circuits*, *52*(1), 127–138. https://doi.org/10.1109/jssc.2016.2616357

Chen, Y., Chen, T., Xu, Z., Sun, N., & Temam, O. (2016). DianNao family. *Communications of the ACM*, *59*(11), 105–112. https://doi.org/10.1145/2996864

Esser, S. K., Merolla, P. A., Arthur, J. V., Cassidy, A. S., Appuswamy, R., Andreopoulos, A., Berg, D. J., McKinstry, J. L., Melano, T., Barch, D. R., di Nolfo, C., Datta, P., Amir, A., Taba, B., Flickner, M. D., & Modha, D. S. (2016). Convolutional networks for fast, energy-efficient neuromorphic computing. *Proceedings of the National Academy of Sciences*, *113*(41), 11441–11446. https://doi.org/10.1073/pnas.1604850113

Hruska, J. (2017, August 30). New Movidius Myriad X VPU Packs a Custom Neural Compute Engine. *ExtremeTech*. https://www.extremetech.com/computing/254772-new-movidius-myriad-x-vpu-packs-custom-neural-compute-engine

Edge TPU - Run Inference at the Edge |. (n.d.). Google Cloud. https://cloud.google.com/edge-tpu/

Asia, A. (2020, January 30). *Kneron's Next-Gen Edge AI Chip Gets $40m Boost*. EE Times Asia. https://www.eetasia.com/knerons-next-gen-edge-ai-chip-gets-40m-boost/

Fick, D. (2018, September 11). Mythic @ Hot Chips 2018 - Mythic. *Medium*. https://medium.com/mythic-ai/mythic-hot-chips-2018-637dfb9e38b7

Ward-Foxton, S. (2020, February 14). XMOS adapts Xcore into AIoT 'crossover processor.' *EE Times*. https://www.eetimes.com/xmos-adapts-xcore-into-aiot-crossover-processor/

Schor, D., (2021, May 25). Arm Ethos is for Ubiquitous AI At the Edge. *WikiChip Fuse*. https://fuse.wikichip.org/news/3282/arm-ethos-is-for-ubiquitous-ai-at-the-edge/

(2018, October 24). NovuMind's AI Chip Sparks Controversy. *EE Times*. https://www.eetimes.com/novuminds-ai-chip-sparks-controversy/

Ward-Foxton, S. (2019, November 15). *Gyrfalcon Unveils Fourth AI Accelerator Chip*. EE Times. https://www.eetimes.com/gyrfalcon-unveils-fourth-ai-accelerator-chip/

Clarke, P. (2018, August 26). *Indo-US startup preps agent-based AI processor*. EENewsEurope. https://www.eenewseurope.com/en/indo-us-startup-preps-agent-based-ai-processor-2/

Ward-Foxton, S. (2019a, August 24). *Details of Hailo AI Edge Accelerator Emerge*. EE Times. https://www.eetimes.com/details-of-hailo-ai-edge-accelerator-emerge/

Clarke, P. (2019, December 12). Globalfoundries aids launch of Chinese AI startup. *EENewsEurope*. https://www.eenewseurope.com/en/globalfoundries-aids-launch-of-chinese-ai-startup/

Clancy, M. (2022, August 23). Ushers in the PetaOps Era with At-Memory Computation for AI Inference Workloads. *Untether AI*. https://www.untether.ai/inthenews/untether-ai-ushers-in-the-petaops-era-with-at-memory-computation-for-ai-inference-workloads

Ward-Foxton, S. (2020a, February 4). TI's First Automotive SoC with an AI Accelerator Launches. *EE Times*. https://www.eetimes.com/tis-first-automotive-soc-with-an-ai-accelerator-launches/

(2018b, November 19). Mobileye's New EyeQ5: How Open is Open? *EE Times*. https://www.eetimes.com/mobileyes-new-eyeq5-how-open-is-open/

McGrath, D. (2019, April 10). Qualcomm Targets AI Inferencing in the Cloud. *EE Times*. https://www.eetimes.com/qualcomm-targets-ai-inferencing-in-the-cloud/

Morgan, T. P. (2021, May 4). Nvidia Rounds Out "Ampere" Lineup With Two New Accelerators. *The Next Platform*. https://www.nextplatform.com/2021/04/15/nvidia-rounds-out-ampere-lineup-with-two-new-accelerators/

Weiss, T. R. (2021, May 21). Google Launches TPU v4 AI Chips. *HPCwire*. https://www.hpcwire.com/2021/05/20/google-launches-tpu-v4-ai-chips/

Cerebras. (2022, November 21). *Homepage*. https://www.cerebras.net/