Achronix FPGAs Optimize AI in Industry 4.0 and 5.0 (WP027)

Achronix
Data Acceleration

December 21, 2021 **White Paper**

Introduction

Industry has come a long way over in the last three hundred years. Machines were first introduced in the 1700s, mainly water and steam driven, introducing the Industrial Revolution in the late 1700s (often referred to as Industry 1.0). Although the concept of assembly lines dates back to ancient China's blue and white porcelain manufacturing, Henry Ford was the first to introduce a powered assembly line in the late 1800s setting the framework for Industry 2.0. Automation and computer technology would enter the picture in the late 1960's, framing Industry 3.0 which paved the way for the eventual automation, artificial intelligence (AI) and networked solutions driving Industry 4.0 of today. Although it might appear that humans are no longer in the picture, Industry 5.0 is bringing us full circle by combining the precision and efficiency of robotic systems, driven largely by AI, with the ingenuity and real-time thought of the human mind — all leading to more optimal manufacturing environments.

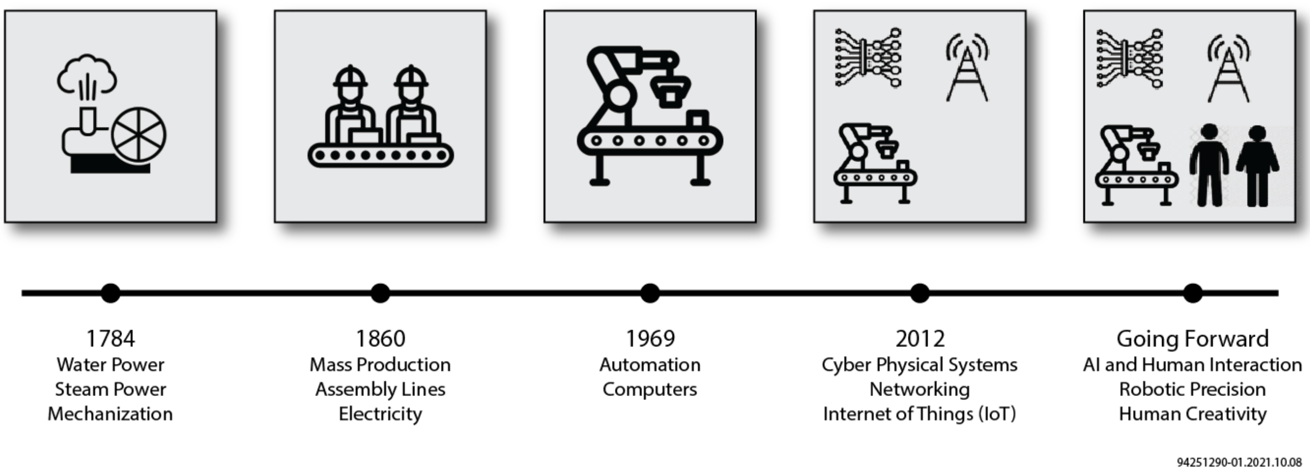

Figure 1: ***Evolution of Industry***

Artificial Intelligence

AI is a branch of computer science focused on the development of machines which emulate human behavior. These machines range on a spectrum from simple execution of an algorithm to learning from the environment and adapting their algorithms with no human intervention. Machine learning (ML) is a subset of AI which uses statistical models derived from datasets to improve specific tasks. As a subset of ML, deep learning (DL) leverages multi-layer neural networks to not only perform basic ML inferencing, but is capable of learning from new data yielding a higher level of cognitive ability (see figure below). In this white paper, any instance of ML or DL is referred to simply as ML.

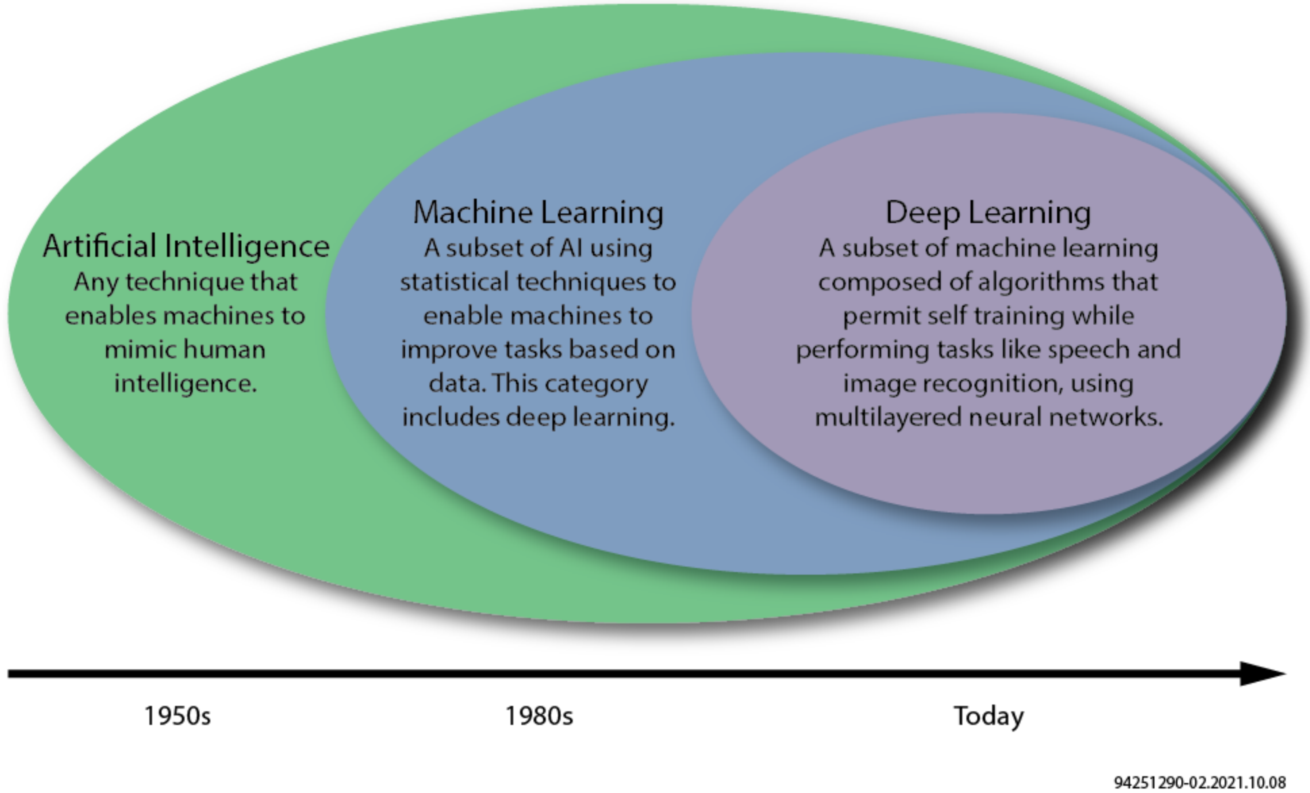

Figure 2: ***AI/ML/DL Spectrum***

Some general examples of AI use cases include advanced driver assistance systems (ADAS), which is the backbone for autonomous vehicles, speech recognition and synthesis (think Huawei's Celia), medical diagnosis, data and network security, predictive models for financial services (such as electronic trading) or recommendation for e-commerce and streaming services and, of course, industrial manufacturing.

AI became relevant in the manufacturing environment as Industry 4.0 evolved in the early 2010s. Many applications leverage AI today to streamline both manufacturing and business operations, processes, security and supply chain. Using predictive algorithms, AI can monitor equipment, optimizing maintenance schedules and ultimately predict mechanical failure.

Management of supply chains for materials associated with manufacturing can also leverage predictive algorithms to keep lines up and running more efficiently. Considering both past and present business demand, AI algorithms can aid in the forecasting of future business. These AI systems can be tied to the supply chain and inventory management systems which leads to faster time to revenue, minimizing overhead costs. Robotics have been an essential component as early as Industry 3.0. But as we move into Industry 5.0, these robotic systems will need to have adaptive AI algorithms, most certainly DL algorithms, which are not only learning on their own, but are also capable of interpreting real-time inputs from human beings. This ability to adapt in real time with minimal latency will be essential.

Eco-system Components Other Than AI

AI will continue as a major component to the continued roll-out of Industry 4.0 and the evolution of Industry 5.0; however, real-time data is needed in order for these AI algorithms to thrive. The Internet of things (IoT) is a system of interrelated electronic devices which source and receive data from both the analog and digital world. Time, pressure, temperature, speed, angles as well as audible and visual sources of data need to be captured and converted to structured data which then can be analyzed and controlled by various AI-based systems. 5G networking has been deployed since 2019 (South Korea was the first) and, compared to 4G networks, is providing up to 100 times the bandwidth (up to 10 Gbps) and up to 500 times the number of unique channels. Combining 5G networks with IoT, the incoming data is overwhelming and creates a new paradigm in computing, namely the need for data accelerators.

Data Accelerators

With this massive amount of data, the burden on the data center to consume and make sense of the data puts a large burden on the traditional compute server model. The old way of managing increases in data was to simply increase the server count in a data center. This increase in server footprint not only increases CAPEX, the OPEX also increases due to an increase in power both for running the machines as well as cooling them. Depending on the data accelerator type and the workload, the computational ability of a single data accelerator on one server can do the work of as many as 15 servers, drastically cutting down the CAPEX and OPEX. Hardware-based data accelerators provide the additional benefits of lower latency and higher determinism which translates to higher application performance for latency-sensitive use cases such as autonomous vehicles, Industry 4.0/5.0, financial services and many others. One final characteristic of the optimal data accelerator is flexibility allowing adaptation to changing ML/DL algorithms, whether that being modification to the algorithm itself, a change in workload and /or updating of the datasets behind the ML/DL algorithms.

There are three competing hardware approaches to this data acceleration market, the GPU, the FPGA and custom ASICs. Consider the chart in the following figure. CPUs will always have the highest flexibility, but with a tradeoff of power, performance and cost versus dedicated data accelerators. That leaves GPUs, ASICs and FPGAs. ASICs will certainly deliver the highest efficiency and performance; however, ASICs are fixed function and do not provide needed flexibility to adapt to changing AI algorithms, specification changes for newer technologies, vendor-specific requirements and workload optimizations. GPUs, which have been the traditional workhorse in the core data center, are limited to pure computational use cases, excluding in most cases, the ability to accelerate networking and storage, and do so at the expense of power and cost. FPGAs are capable of network, compute and storage acceleration, have the speed of an ASIC and the needed flexibility to deliver the optimal data acceleration in today's core and edge data centers. In addition to all of the data acceleration, FPGAs will play a critical role in sensor fusion and the consolidation of incoming data traffic, laying the groundwork for data consumption.

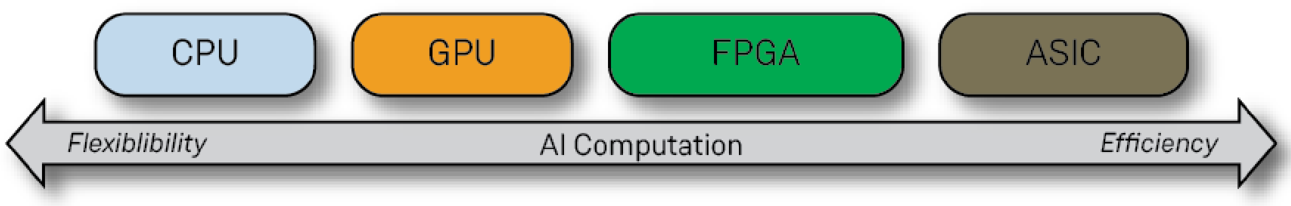

Figure 3: ***CPUs vs GPUs vs FPGAs vs ASICs***

Achronix Offering

Achronix develops FPGA-based, data acceleration products for AI/ML compute, networking and storage applications. Unlike other high-performance FPGA companies, Achronix offers both FPGAs and embedded FPGA IP solutions. In addition to standalone FPGAs and eFPGA IP, Achronix offers PCIe-based acceleration cards, which can be used for development, field trials or production use cases. The Speedster®7t family of 7nm FPGAs offers the highest speed I/O available in the industry, with 400 GbE, PCIe Gen 5 and dual memory interfaces: standard DDR4 and GDDR6 memory interfaces which provides an impressive 600% speed advantage over DDR4. None of these high-speed interfaces would mean much if the data could not flow easily through the FPGA fabric. To prevent such a bottleneck, Achronix architected a two-dimensional network on chip (2D NoC), which effectively acts as a data superhighway for all external I/O, hardened functional blocks within the FPGA and the FPGA fabric itself. This 2D NoC delivers in excess of 20 Tbps of bi-directional bandwidth which well exceeds the aggregate bandwidth needs of the I/O and functional blocks — removing latency concerns from intra-chip communication.

For higher volume use cases that are cost, performance and power sensitive, ASICs are typically developed, but what how to address the need for flexibility? Whether it is algorithm variants, changing requirements, vendor-specific requirements, operator-specific requirements, protocol adaptation or varying interfaces of functional system blocks, there needs to be some level of flexibility. Speedcore™ eFPGA IP is the answer to equipping an ASIC with “just enough” flexibility. The resource amount and mix of LUTs, memory, DSP/MLP and 2D NoC are determined by the ASIC developer, then Achronix delivers the custom IP for monolithic integration into their ASIC or SoC design.

VectorPath™ accelerator cards, which are PCIe form-factor hardware accelerator platforms, can be considered for evaluation, development, field trial vehicles or for production applications. Custom versions of this solution can also be created to address specific user requirements.

Conclusion

AI, ML and DL will continue to drive Industry 4.0 and 5.0 to new levels of productivity and efficiency. Aided by IoT and 5G technology, automation and robotics will merge with human ingenuity and creativity creating manufacturing environments that we humans could have never imagined even 10 years ago. While enabling sensor fusion which enables interfacing to many IoT devices, FPGAs offer a balance of high performance and flexibility needed for AI systems in the manufacturing environment.

Data Acceleration

Achronix Semiconductor Corporation

2903 Bunker Hill Lane
Santa Clara, CA 95054
USA

Website: www.achronix.com
E-mail : info@achronix.com

Notice of Disclaimer