Machine learning is a frequent buzzword heard throughout the technology space and has promises to revolutionize the way data is collected, understood, and used. For this all to happen, the algorithms used to interpret this data can be run on various types of hardware from general purpose CPUs, to more specialized GPUs, Application Specific Integrated Circuits or ASICs, and more frequently Field Programmable Gate Arrays or FPGAs. What is unique about FPGAs compared to the other types of hardware listed is that these devices allow users to use a hardware description language to reconfigure their integrated circuit for numerous purposes, configurations and optimizations up to the limits of the FPGA itself. This analysis will take a deeper look at the comparison of GPUs and FPGAs, some current implementations, and the advantages and disadvantages of their use for cloud-based machine learning.

Many of today's compute intensive needs have been moved into the cloud where these services can be sold to consumers in the form of Infrastructure as a Service (IaaS). This includes being able to use the servers in a virtual machine, store data and is priced as a pay-as-you-go structure, meaning that a user only pays for the server resources used and nothing more. Some providers of these services, such as Amazon or Microsoft Azure, have started expanding their offerings to include FPGAs for many applications, including machine learning. These providers are realizing that FPGA maybe be a more cost-effective way to deliver what users and developers are requesting. Some of the benefits that FPGAs bring over GPUs in this area is within the name itself – "field programmable." In one such example of how FPGAs can be changed to fit changing needs of a company, Dell's artificial intelligence neural net provides the flexibility not typically provided by GPUs. The ability to change the underlying hardware architecture allows FPGA-based platforms to use state-of-the-art deep learning innovations as they emerge (Morss, 2019).

Another large benefit of FPGAs over GPU options in the area of machine learning is the power efficiencies gained with the use of FPGAs. One of the biggest concerns with warehouse-scale computing is the efficiency of the devices and the power consumed by the systems that are used in computations and not for other things such as cooling. FPGAs are well known for their low power use and efficiency. For example, in a research project conducted by Microsoft, the found that in an image classification situation the Arria 10 FPGA performed almost 10 times better in power consumption than GPUs. Additionally, Xilinx, a company the manufactures several models of FPGAs, found an almost four times less consumption in power in their comparisons (Fallahlalehzari, 2020).

Further, the throughput and latency of FPGAs are also touted as a benefit for machine learning when compared to GPUs. For deep neural applications, FPGAs provide much more throughput with their on-chip memory which is essential to reducing the latency. The high amount of on-chip cache memory reduces the memory bottlenecks associated with external memory access as well as the power and costs of a high memory bandwidth solution. (Fallahlalehzari, 2020). With an FPGA it is possible to achieve a latency of less than 1 nanosecond whereas a GPU could be several times more. One of the main reasons for this low latency is that FPGAs, as mentioned before, can be much more specialized meaning that they do not rely on a generic operating system and communications do not have to go through generic buses, such as USB or PCIe (van der Ploeg, 2018). The benefits of this low latency are numerous and have many different applications. An area where low latency is necessary is automated driver systems, which also employ machine learning. One example could be when a vehicle's front-facing radar sensors detect that a vehicle in front of them has slowed down or stopped. It must relay this information to the FPGA, which then must do quick calculations to then send instructions to the brakes to slow the car down either slowly or rapidly to prevent a collision. Additionally, the low latency of FPGAs has the potential to help transition business and education into the growing online work and learning environments that have come forth due to the COIVD-19 pandemic. Many suppliers of online services, including those who facilitate services such as video and teleconferencing, have felt the strain on their systems and having low latency FPGAs could help these services scale, as opposed to using only GPUs.

Despite these and other benefits of using FPGAs in place of general purpose GPUs in the large computations for machine learning, they do have their setbacks. One such issue is converting code for machine learning algorithms in more widespread languages such as C/C++, Python, Java, and others into a hardware description language (HDL). Because the HDL describes the actual paths and connections of the integrated circuits, it is not always intuitive to write this type of language to optimize the best algorithm. One solution to this problem is to use High Level Synthesis (HLS), which allows users to use a more intuitive, algorithmic programming language like C and let the HLS to "abstract away hardware-level design (Singh, 2020)." Additionally, the time for the code to be compiled into the useable bitfile that is implemented on the FPGA itself can take much longer than is required for compiling software-based languages, which are hardware agnostic. For example, Intel's OpenCL compiler typically takes between 4 and 12 hours to compile a typical program for the FPGA. These long compile times are due to

the place-and-route phase of the compilation. This means that the compiler must map the circuits that the engineer wants to the FPGA resources that are available (van der Ploeg, 2018).

Even with these setbacks, providers have been working to ensure that the benefits of FPGAs are available to developers and programmers who may not have any experience with FPGAs or HDLs. For example, Amazon's AWS has created different development environments for based on the developer's experience with HDLs. They are able to supply developers with a software development kit (SDK) that allows them to "develop, simulate, debug, compile and run hardware accelerated applications on Amazon EC2 F1 instances, EC2 F1 instances are high-performance compute instances with field programmable gate arrays (FPGAs) that enable the development and deployment of custom hardware accelerators on AWS cloud (Amazon Web Services, 2020)." Simply put, Amazon is making it easier for those who wish to use the AWS cloud to accelerate their programs with the FPGAs bringing in a new era of FPGA acceleration as a service (FaaS). This could be a big step for many different developers because of the benefits of FPGA mentioned before. For this course, this could be one area that could be looked into further to investigate the claims of better performance.

FPGAs have already made their impact on the cloud computing space and it is evident that there will be more areas where their presence will be felt. As autonomous systems become more prevalent the need for low latency systems with big computing power will be required. As has been discussed, FPGAs will be a necessary inclusion in these embedded systems. The biggest problems with implementing FPGAs into these and other systems will be providing developers with an easy-to-use platform for development. Allowing developers to write code in languages they already know and using HLS to abstract the code into the hardware description language needed for the FPGA will be necessary. While more analysis is required to realize the full potential of these devices compared to GPU implementation for similar tasks, it is clear that there will be a broad expansion within the cloud computing space and in other systems more broadly.

References

Amazon Web Services. (2020, September 12). *Overview of AWS EC2 FPGA Development Kit.* Retrieved from Github: https://github.com/aws/aws-fpga

Fallahlalehzari, F. (2020). *FPGA vs GPU for Machine Learning Applications: Which one is better?* Retrieved from ALDEC: The Design Verification Company: https://www.aldec.com/en/company/blog/167--fpgas-vs-gpus-for-machine-learning-applications-which-one-is-better

Morss, J. (2019, January 16). *FPGAs vs. GPUs: A Tale of Two Accelerators*. Retrieved from Dell Technologies: https://blog.dellemc.com/en-us/fpgas-vs-gpus-tale-two-accelerators/

Singh, A. (2020, March 2). *Hardware for Deep Learning: Know Your Options*. Retrieved from Towards Data Science: https://towardsdatascience.com/deep-learning-hardware-know-your-options-9e95026b5d5e

van der Ploeg, A. (2018, August 18). *Why use an FPGA instead of a CPU or GPU?* Retrieved from eScience Center: https://blog.esciencecenter.nl/why-use-an-fpga-instead-of-a-cpu-or-gpu-b234cd4f309c