



Microsoft Targets Deep Learning for the Rest of Us

September 14, 2015 - IDC Link

By: [Robert Sorensen Jr.](#)

At the August 2015 Hot Chips conference in Silicon Valley, [Microsoft](#) presented an update on its effort to make deep learning resources more accessible to a wider range of developers by offering deep learning FPGA-based hardware in their future Open Cloud servers that are designed expressly for hyperscale datacenters, such as those in their Azure cloud[1] on page 0 . Microsoft indicated that the new FPGA chips are specifically designed to extract strong price/performance on key deep learning applications while operating at lower power levels than the more traditional specialized GPU hardware — such as those offered by NVIDIA.

Although NVIDIA's strong focus on deep learning hardware built on its GPU products has placed the firm at the forefront of this rapidly emerging field, standing up a GPU-based development system can require a strong financial and technical commitment from either deep learning developers or aspirants. With this effort, Microsoft is looking to leverage its current efforts to build out its capabilities in public clouds by offering deep learning hardware that is well suited to the unique requirements of a large datacenter — finely targeted and stringent power, thermal, and budget limits — as well as the ability to support a wide range of applications that may or may not need deep learning capabilities on a continued or even regular basis.

Deep learning applications are driving many of the most advanced developments today in natural language processing, computer vision, and video and photo recognition. Firms such as Baidu, Google, and Facebook are leading the charge in using deep learning, and the sector has a wide range of use cases that include fraud detection for the finance sector, recommendation engines for ecommerce advertising, machine vision for real-time threat detection at airports, and image and video searching to support the social media experience. As such, the sector is ripe for new firms to enter the fray and many will be eager to consider any new option that lowers the barriers to entry for exploration and subsequent development.

Although Microsoft is primarily looking to offer an accessible and low-cost deep learning option, the FPGA product it is developing is designed to offer impressive performance. According to Microsoft's claims, its Catapult Server equipped with an Altera Arria FPGA would soon be capable of processing about 800 images per second on the industry standard ImageNet 1K benchmark, compared with a counterpart NVIDIA Titan X-based system that processed 4,129 images per second but required 10 times more power (25W for the FPGA versus 250W for the NVIDIA chip). In addition, Microsoft notes that, in a typical deep learning cluster, the GPUs do most of the computational heavy lifting while the CPUs do little more than oversee data transfer and draw power. With their FPGA design, Microsoft notes that the CPUs that are not being used directly to support a deep learning program can be applied to other open jobs in the datacenter, further increasing the overall utility of the design.

Microsoft has not made any announcements as to when this new FPGA will be available within the Azure — or other — public cloud, but the firm is looking to soon upgrade the FPGA technology to Altera's leading-edge Stratix 10 FPGA — which will use Intel's 14nm process — that could offer significant performance increases. And while it is too soon to tell exactly how many developers will be willing to test the waters of deep learning on a cheaper, more accessible platform, this effort bodes well for potential expansion in the pool of deep learning developers and, perhaps more important, it opens the door for additional purpose-built hardware — not just from Microsoft but from any developer — that can support the business case for application-specific FPGAs targeted for the datacenter.

[1] on page 0 Deep learning deals with the use of a multiple-layer neural network for applications such as natural language processing. Training a deep learning network is mostly supervised in that it allows the computer to learn and "think" by itself through interactions with a set of training data. While training, each layer in the network performs a well-defined subtask, feeding its output to the input of the next layer in the network, and learning progresses through a relatively simple feedback loop based on differences between the neural net output and the data contained in the "training" set. Although highly automated, training a deep neural net can consume significant amounts of computation cycles. In contrast, with unsupervised learning, no labels are given to the algorithm within the training set, leaving it on its own to find structure in the input.

Subscriptions Covered:

[High-Performance Data Analysis](#), [Technical Computing](#)

Please contact the IDC Hotline at 800.343.4952, ext.7988 (or +1.508.988.7988) or sales@idc.com for information on applying the price of this document toward the purchase of an IDC or Industry Insights service or for information on additional copies or Web rights. Visit us on the Web at www.idc.com. To view a list of IDC offices worldwide, visit www.idc.com/offices. Copyright 2015 IDC. Reproduction is forbidden unless authorized. All rights reserved.