Micron Accelerator Bumps Up Memory Bandwidth

By George Leopold

February 26, 2020

Deep learning accelerators based on chip architectures coupled with high-bandwidth memory are emerging to enable near real-time processing of machine learning algorithms. Memory chip specialist Micron Technology argues that hardware accelerators linked to higher memory densities are the best way to accelerate “memory bound” AI applications.

Micron announced a partnership this week with German automotive supplier Continental to adapt the memory maker’s deep learning accelerator to machine learning auto applications. The partners said they would focus on vehicle automation systems.

The accelerator also is being tested at CERN, the Swiss research center.

Micron’s deep learning accelerator has machine learning applications for other edge deployments, as well, where memory bandwidth used for local data processing has so far failed to keep pace with microprocessor core growth.

Micron’s approach combines hardware and software to accelerate and reduce power consumption in FPGAs backed by high-bandwidth memory. The package also includes a machine learning software development kit that abstracts underlying hardware. The scheme eliminates the need for further FPGA programming.

The Micron accelerator runs on a Xilinx Virtex Ultrascale+ FPGA, and can accommodate up to 512 Gb of DDR4 memory with memory bandwidth up to 68 Gb/s. A pre-loaded inference engine supports several neural network types while the FPGA can be programmed in Python and C++.

Meanwhile, the software kit supports a batch of frameworks for training neural networks, including Caffe 2, PyTorch and TensorFlow.

“Working together with Micron to build a scalable and flexible solution for edge inference that supports multiple networks and frameworks will enable us to efficiently deploy machine learning across our platforms,” said Dirk Remde, vice president of Continental’s Innovation Center Silicon Valley.

“One of our key goals in collaborating with Continental is to create an agile edge-inference solution that uses machine learning and delivers the ease of use, scalability, low power and high performance the automotive sector requires,” added Steve Pawlowski, Micron’s corporate vice president of advanced computing solutions.

Micron’s is the latest in a flood of purpose-built AI hardware that includes Tensor processors from Google, GPU accelerators from Nvidia and deep learning engines from AMD and Intel. Micron’s twist involves embedding an inference engine in an accelerator and tying it to a larger memory pipeline.

Google and Micron also have been working with scientists at CERN, the European nuclear research organization, on applying their frameworks to high-energy physics problems. CERN researchers are testing Micron’s deep learning accelerator as part of two projects that support the laboratory’s Large Hadron Collider experiments. The memory-boosted neural network approach is being used to test data acquisition systems.

“By just changing a couple lines of code, researchers target the Micron accelerator as they would target a GPU,” Micron’s Mark Hur noted on recent blog post.