The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

Data Centers Accelerate AI Processing

Vendors Large and Small Optimize Their Designs for Deep Learning

By Bob Wheeler (December 24, 2018)

In a year dominated by AI, you'd be excused for thinking traditional server CPUs were dying. Yet Intel's data-center revenue (mainly server processors) grew 26% in the first nine months of 2018, reaching a record $16.9 billion. By contrast, Nvidia's data-center revenue (mainly accelerators) rose an astounding 70% from a much smaller base, reaching $2.3 billion in that period. Although AI-accelerator growth slowed in 2018, both established vendors and startups continue to invest heavily in this technology. Companies are pitching everything from simple ISA extensions to purpose-built accelerator chips and cards.

Although cloud operators such as Google have deployed AI-specific hardware, merchant vendors have mostly repurposed existing designs for AI. Recently, the horde of startups developing AI-specific architectures began to deliver production devices. Graphcore, Habana, and Wave are three such companies targeting data centers. Meanwhile, Intel plans to ship its second purpose-built AI silicon for data centers in 2019 after its first effort flopped. These vendors are all chasing Nvidia, which continues to optimize its Tesla GPU cards for AI, where they primarily accelerate deep-neural-network (DNN) training.

Intel and Nvidia continue to benefit from the growth in cloud data centers, but in-house ASICs can displace their products. Examples include Amazon's processors, Google's TPUs, and Huawei's AI chips. Merchant processors and accelerators must demonstrate superior efficiency if they hope to succeed at hyperscalers. In the longer term, a big shakeout in AI-chip vendors is coming, but next year should see more new entrants than failures.

Training for Dollars

In June, two distinguished computer architects—John Hennessy and David Patterson—raised the profile of domain-specific architectures through their ACM Turing Award lecture. They argued that domain-specific architectures (DSAs) combined with domain-specific languages (DSLs) are the path to higher efficiency as Moore's Law slows (we posited the same approach back in *MPR 8/26/13,* "What Comes After the End").

A canonical DSA/DSL example in data centers is Nvidia's Tesla programmed using Cuda to accelerate HPC. Cuda DNN (cuDNN) adapts this model for DNN training, supporting frameworks that include Caffe, TensorFlow, and others. In 2017, Nvidia optimized its Tesla accelerators for convolutional neural networks by adding matrix-multiplication units ("tensor cores") to its 12nm Volta generation. For ResNet-50 training on FP16 data, the Volta-based Tesla V100 card delivers 1,350 images per second (IPS), 4x the performance of the Pascal generation. Volta has dominated the market for training neural networks, offering at least 10x better performance than mainstream Skylake-SP processors.

Hoping to take share from Nvidia, AMD recently introduced Radeon Instinct cards using its 7nm Vega20 GPU. Although Vega20 handles FP16 data, it lacks Volta's matrix-multiply hardware. In ResNet-50 training, AMD's MI60 card delivers only 499 IPS, falling short of Volta at the same power (300W). Nvidia's technology leadership allowed it to optimize its hardware and software for superior efficiency, more than offsetting AMD's process advantage.

One of the first AI startups to target training is Graphcore, which is shipping a processor purpose-built for

DNNs. Its 16nm GC2 approaches the FP16 performance of Volta but consumes half the power. The company can then pack two of its processors into a 300W card. In simulations, the card delivers about 1.5x the ResNet-50 throughput of the V100, but Graphcore has yet to publish benchmark data. Still, the GC2 demonstrates the advantage of a clean-sheet design for DNN acceleration: it offers 4x the performance per watt of AMD's GPU while using process technology that trails by a full node.

Another startup, Wave Computing, is implementing its training technology in an appliance that competes with Nvidia's DGX systems. Wave is shipping evaluation systems that integrate four of its DPU ASICs. That system could deliver performance in the same range as Nvidia's $400,000 DGX-2, which integrates 16 V100 modules. The startup, however, has thus far withheld benchmarks.

Intel is conspicuous by its absence, having decided to use its first-generation Nervana product as only a development platform after it fell far behind the V100 in performance. It plans to ship in late 2019 the NNP L-1000, code-named Spring Crest. On the basis of the company's performance targets, that chip is unlikely to significantly outperform the V100, much less Nvidia's next-generation Ampere design.

Habana Smokes Inference Incumbents

Whereas Nvidia dominates DNN training, data-center operators employ multiple architectures for inference tasks. The Intel Xeon processors are the most popular, but operators also employ custom DNN accelerators, and Nvidia is making inroads with new products. Inference offers more room for optimization owing to the use of lower-precision fixed-point (INT8) math. Last quarter, Nvidia introduced new Tesla T4 cards based on its Turing architecture, its first to include tensor cores that handle integer data. The result is a 70W card that delivers 27TMAC/s for INT8 data. Using the Caffe framework with its TensorRT libraries, the company measured ResNet-50 inference throughput of 3,920 IPS, or 56 IPS per watt, as Figure 2 shows.

Intel's response is to improve Xeon's performance. It added to Cascade Lake new fused-multiply-accumulate instructions for INT8 and INT16 data types under the Vector Neural Network Instruction (VNNI) umbrella. Its simulations show Cascade Lake doubles ResNet-50 inference throughput when using INT8, which should yield about 2,500 IPS for the top-end Xeon (i.e., Platinum 8180 "v2") 2S platform. Two processors plus a south bridge consume about 435W TDP, yielding around 5.7 IPS per watt, or 10% of the Tesla T4's efficiency. Carrying a list price of more than $20,000, the dual-8180 configuration is hardly mainstream.

A well-funded Israeli startup, Habana is attacking high-throughput inference with its Goya HL-1000 card, which sampled in 3Q18. It measured ResNet-50 throughput of 15,012 IPS at 100W (typical), yielding 150 IPS per watt. Goya uses C-programmable tensor cores that support integer data types as well as FP32. In 2Q19, the company plans to sample Gaudi, a training processor designed to scale to large arrays using a 2Tbps interconnect. Habana has raised $120 million with Intel Capital as the lead investor.

Another approach to inference acceleration is to employ FPGAs, as Microsoft has done in its Project Brainwave with Intel Stratix 10. For customers needing off-the-shelf hardware, Intel unveiled accelerator cards that sport Arria 10 FPGAs and are available through server OEMs. The one based on the Arria 10 GX, for example, is a PCIe Gen3 x8 card with an FPGA, 8GB of DDR4 SDRAM, and a 40G Ethernet port; it has a 70W TDP rating. Although it provides an SDK and libraries for Arria 10, Intel doesn't directly support DNN frameworks such as TensorFlow and Caffe.

Xilinx is more aggressive, offering FPGA cards directly to end users and adding its xDNN software to handle frameworks. In October, it introduced the Alveo line of PCIe accelerator cards, which employ 16nm UltraScale+ FPGAs. It rates the top-end Alveo U250 at 33.3 TOPS for INT8 data at 225W (TDP). Xilinx and a partner recently demonstrated the U250 running ResNet-50 inference workloads at 3,700 IPS. We estimate ResNet-50 efficiency at 33 IPS per watt (typical), falling short of Tesla's T4. The company's 7nm generation (Versal) adds programmable AI engines optimized for integer data, an improvement that

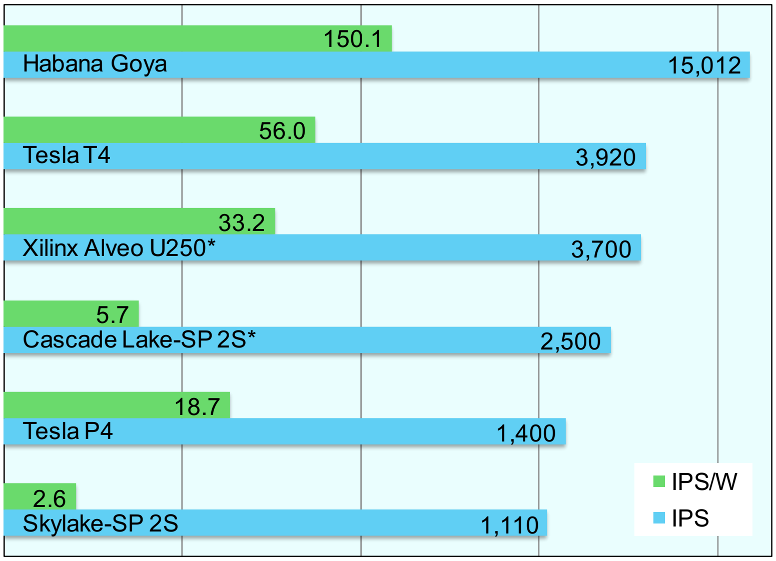

Figure 2. ResNet-50 inference throughput and efficiency (log scale). Habana's Goya sets a new benchmark, demonstrating the advantage of a clean-sheet DNN design. For Intel, we show the top-end Xeon Platinum 8180 model of both Skylake and Cascade Lake in dual-socket (2S) configurations; other products are single-chip designs. (Data source: vendors, except *The Linley Group estimate)

should greatly boost performance per watt. Samples of the first Versal models are due in 2H19. Customers can use the first-generation Alveo cards in the interim, particularly for software development.

Competing With Cloud Customers

A rising tide should lift all boats, but hyperscale cloud operators and OEMs are creating dangerous currents for merchant vendors. Google is the prime example with its TPUs, now entering their third generation. Whereas the TPUv1 (circa 2015) was an inference coprocessor, the TPUv2 (deployed in 2017) is a standalone processor designed for training. Its matrix unit handles floating point, and it's designed to work in clusters. At the end of 2018, TPUv2 clusters (or pods) were in alpha testing, while the TPUv3 had entered cloud beta testing. Google is ramping TPUv2 deployment for training, and Nvidia's data-center growth has slowed to less than 10% per quarter.

Among the server OEMs, Huawei is the most aggressive in developing in-house silicon. Its chip arm, HiSilicon, has created ASICs and SoCs for many markets. Now, the company has developed the Ascend 910 chip for data-center training and inference. It's also investing in its own cloud-services business, which is likely to consume internally developed chips where possible. Intel appears most exposed in this case.

In 2019, Nvidia will finally see competition from third-party AI accelerators as multiple startups begin shipping their first designs. These startups have all promised massive performance and efficiency improvements, but once their products reach the market, we'll see which ones are delivering on these promises. Some will, but the ones that don't will begin to fall by the wayside as early as 2020. This consolidation phase will be similar to the boom-and-bust cycle of GPU startups in the 1990s and NPU startups in the 2000s. Ultimately, only a few AI-chip startups will succeed in the data center. Intel and Nvidia will push their current efforts, but they may end up acquiring successful startups. One way or another, these big vendors will continue to profit from AI's rapid growth. ♦