The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

Habana Wins Cigar for AI Inference

Startup Takes Performance Lead With Custom Architecture

By Linley Gwennap (February 18, 2019)

Startup Habana Labs is in production with its first AI accelerator card, and it's not taking baby steps. Targeting data centers, the Goya accelerator offers roughly 4x the performance of Nvidia's Tesla T4 on the popular ResNet-50 model. It targets AI inference, but the company is also developing an accelerator for training. Goya is well suited to small batch sizes and offers one-third the latency of the T4. It even delivers about 3x better power efficiency than Nvidia. The Habana card plugs into standard servers via PCI Express and runs neural networks developed in TensorFlow, MXNet, and other popular frameworks.

To deliver this performance, the company developed the HL-1000 ASIC, which implements a custom architecture. The secretive startup withheld most details of this 16nm chip other than a high-level block diagram.

Habana is located in Israel, not Cuba. Its founders have a strong record, with ties to Annapurna, which Amazon acquired in 2015 (see *MPR 2/8/16,* "Amazon Exposes Annapurna Chips"); to Ceva, a leader in DSP intellectual property (IP); to 3D-sensor developer PrimeSense, which Apple acquired in 2013; and to Galileo, which Marvell acquired in 2000. Founded in early 2016, Habana has quickly delivered its first product. It recently raised $75 million from Intel Capital and other investors including well-known firms Bessemer and Battery Ventures, bringing its total funding to $120 million. This capital has helped the company increase its staff to more than 130 employees.

Uncovering the Art of Goya

For its chip, Habana designed a VLIW CPU based on its own instruction-set architecture, calling it the TPC (Tensor Processor Core). The VLIW approach eliminates complex pairing and issue logic found in superscalar CPUs while executing many operations per cycle. The instruction set features AI-specific operations.

The TPC enables SIMD vector instructions that operate on both integer and floating-point data—specifically, INT8, INT16, INT32, and FP32 formats, according the company. Each TPC has its own register files and local memory to feed the scalar and vector units. To ensure determinism and reduce latency, the chip omits caches, which typically get in the way of streaming data.

The 8 TPC cores alone, however, can't keep up with Nvidia's Tesla T4. For additional performance, Habana's chip includes a GEMM (general matrix multiply) engine, as Figure 1 shows. To further improve power efficiency, the ASIC includes an on-chip memory shared among the CPUs and the GEMM. This approach avoids off-chip memory accesses, reducing both power and latency. Although some AI accelerators instead employ High Bandwidth Memory

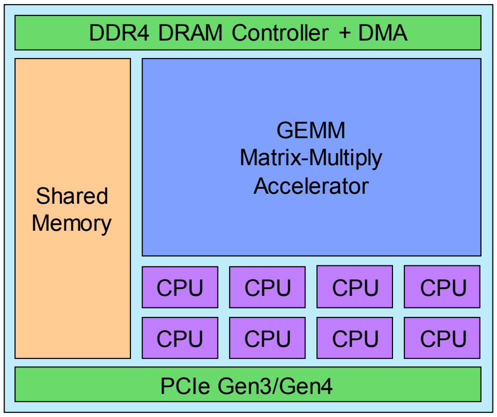

Figure 1. Habana HL-1000 processor. The chip combines a matrix-multiply accelerator with eight custom VLIW CPUs.

For More Information

Habana's Goya HL-100 accelerator card is now in production. The company withheld its price. For more information on the product, including a whitepaper with performance details, access *habana.ai/product*.

to solve the latency problem, HBM adds cost and power. Habana instead provides up to 16GB of relatively slow DDR4 DRAM on the Goya board. The DRAM is ECC protected to boost reliability.

Under the control of its programmable TPCs, Goya can process entire neural networks, communicating with the host processor only to obtain the next item (e.g., an image) for inferencing. The card sports a 16-lane PCIe Gen3 or Gen4 connection for this purpose. It has a 200W TDP, but the company has measured power consumption of 99–120W when running ResNet or Inception and even less for simpler models such as GoogLeNet. Even so, the card requires an external-power connection, as it exceeds the 75W PCI limit.

Habana provides native drivers for TensorFlow and MXNet, and it's developing drivers for other popular frameworks. The company supports the Open Neural Network Exchange (ONNX) format as well, so developers can use other frameworks to access the Goya design. Habana also allows customers to compile Python and C code to run on the TPCs, enabling them to employ Goya for general compute acceleration. As the TPCs are a small fraction of the chip's performance, however, we expect Goya lags well behind Nvidia GPUs for compute workloads that don't use the GEMM engine.

A Great Performance

Habana has released detailed benchmarks showing Goya's performance on several models, including Inception, VGG, and GoogLeNet as well as multiple versions of ResNet. All are convolutional neural networks (CNNs) that perform image recognition. The company imported pretrained networks in TensorFlow or ONNX format from public repositories and ran them using its available drivers. Figure 2 compares Goya and the Tesla T4 on ResNet-50, since Nvidia also publishes data for that model.

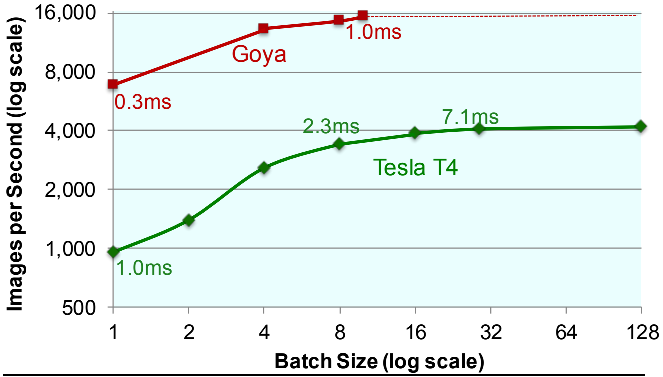

Figure 2. Comparison of ResNet-50 inference throughput. Habana's Goya card outperforms Nvidia's Tesla T4 card across all batch sizes while offering lower latency. (Data source: vendors)

Cloud-service providers often combine multiple items into a single inference pass, called a batch. Larger batches create more opportunities for parallelism and thus can improve MAC utilization. As Figure 2 shows, the Nvidia design requires a batch size of 16 to get close to its maximum throughput, and performance (measured in images per second, or IPS) drops by 75% when processing one item at a time. In contrast, Goya exhibits a flatter performance curve and reaches its maximum performance at a batch size of 10.

Because of the different curves, Goya's advantage is more pronounced at a batch size of one, where it delivers 7x more throughput than the T4. For batches of eight or larger, the advantage is still substantial at about 4x. Goya's strong performance also reduces latency to just 0.29ms for a single item and 1.01ms at a batch size of 10. For any given batch size, Nvidia's latency is about 3x longer.

The T4 draws its power solely from the PCI connector and is rated at just 70W TDP (see *MPR 10/15/18,* "Turing T4 Targets AI Inference"). On ResNet-50, Nvidia measured the card's power at nearly 70W, indicating the design throttles its core to stay under that limit. Habana measured Goya at 103W on this test, about 50% more power than the T4. Thus, Goya still leads in power efficiency (IPS/W), but the gap is smaller: about 2.5x at maximum performance. The T4's 12nm technology helps reduce this gap, as it has a small edge over Goya in transistor power.

First to Market With Real Products

Habana plans to sample a second product, branded Gaudi, in the second quarter. Gaudi's architecture targets AI training; to support this workload, we expect the chip will upgrade Goya's GEMM engine to handle floating-point computation. Gaudi will also introduce a memory-coherent interconnect (similar to NVLink) that allows large models to be trained across multiple accelerators. We expect Habana to carry forward most of the rest of the Goya design to reduce development time. If all goes well, Gaudi could reach production by the end of 2019. Afterwards, the company plans to move its first-generation chips from 16nm to 7nm manufacturing to further improve performance and power efficiency.

Although many startups aimed to deliver data-center AI accelerators to market by the end of 2018, Habana is the only one to demonstrate better-than-Nvidia performance on production silicon. Graphcore also claims to be in production, but it has yet to publish measured performance data, and even its simulations show a relatively narrow (1.5x) advantage over Nvidia. Other startups now aim to deliver products later this year.

Goya's impressive performance shows clear advantages over Nvidia's newest products in throughput, power efficiency, and latency. The card is particularly well suited to cloud services that require low latency or small batches. Many inference deployments operate offline, where latency is moot, or have enough activity to easily fill large batches. But even for these applications, Goya's performance and efficiency could make it a good choice if the price is right. (The company withheld pricing, among other information.)

Habana's performance data covers only CNNs. Many cloud services, particularly voice recognition and language translation, employ recurrent neural networks (RNNs) and other types of models that require more non-MAC operations. These models spend more time on the TPC cores and less time in the GEMM array, reducing their performance and efficiency on the Goya design. Data-center customers should validate Goya's performance on their own workloads to best assess its viability. In any case, the emerging competition among AI-accelerator vendors should spur innovation and reduce prices. ♦

To subscribe to *Microprocessor Report,* access *www.linleygroup.com/mpr* or phone us at 408-270-3772.