The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

HABANA OFFERS GAUDI FOR AI TRAINING

Startup Expects to Top Nvidia V100 Performance at Half the Power

By Linley Gwennap (June 17, 2019)

Habana Labs has achieved first silicon of its initial accelerator for neural-network training, outperforming Nvidia's fastest chip on at least one benchmark. The startup claims the new Gaudi chip will exceed 1,650 images per second (IPS) when training the popular ResNet-50 model. This performance is slightly better than what Nvidia reports for its flagship V100, as Figure 1 shows. Habana says Gaudi will use only 140W when running this benchmark, half the V100's power. These results would make Gaudi twice as power efficient as the V100 and even its little brother, the Tesla T4.

The 16nm Gaudi builds on the same basic architecture as Habana's earlier Goya inference accelerator, which is already in production. Whereas Goya focuses on integer computation, Gaudi fully supports the floating-point formats that most training uses. Gaudi integrates High Bandwidth Memory (HBM2), and to enable large chip clusters, it features 100G Ethernet with remote-DMA (RDMA) capability. For ResNet-50, Habana expects clusters of up to 640 Gaudi chips to deliver near-linear performance scaling. Nvidia, by contrast, sees a severe efficiency drop beyond 16 GPUs.

Whereas Habana sells a single Goya-based product—a PCIe accelerator card—it plans to offer three Gaudi form factors. In addition to a 200W PCIe card, Gaudi will come in an OCP-compliant accelerator module that dissipates up to 300W. Facebook originated this Open Compute Project module design, and several chip providers (but not Nvidia) plan to support it. Habana is also developing a rack-mountable system, the HLS-1, that contains eight Gaudi chips and can serve as an element of a large cluster. The company is testing first silicon and expects all three Gaudi products to sample by the end of this year, which should lead to volume production by mid-2020.

One area where Nvidia retains a sizable lead is in software: its GPUs work with all leading deep-learning frameworks, and many developers employ its Cuda software stack. To help fill this gap, Facebook last year created an open-source project called Glow to convert network graphs from popular frameworks into optimized code for deep-learning accelerators. Habana is the first hardware accelerator that supports Glow; Cadence also supports the compiler for its DSP cores. Esperanto, Intel, Marvell (Cavium), and Qualcomm intend to provide Glow capability as well, building support for the new software.

Big Bandwidth Boost

Habana remains tight-lipped about its deep-learning architecture. We believe the initial Gaudi chip, designated the HL-2000, uses essentially the same compute architecture as

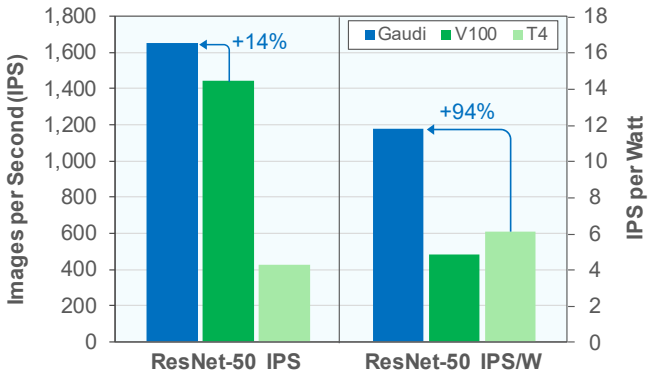

Figure 1. Gaudi emulated performance. For training the simple ResNet-50 model, Habana's Gaudi card offers throughput similar to that of Nvidia's high-end V100 GPU at half the power. It also beats Nvidia's Tesla T4 card in performance per watt. (Data source: vendors)

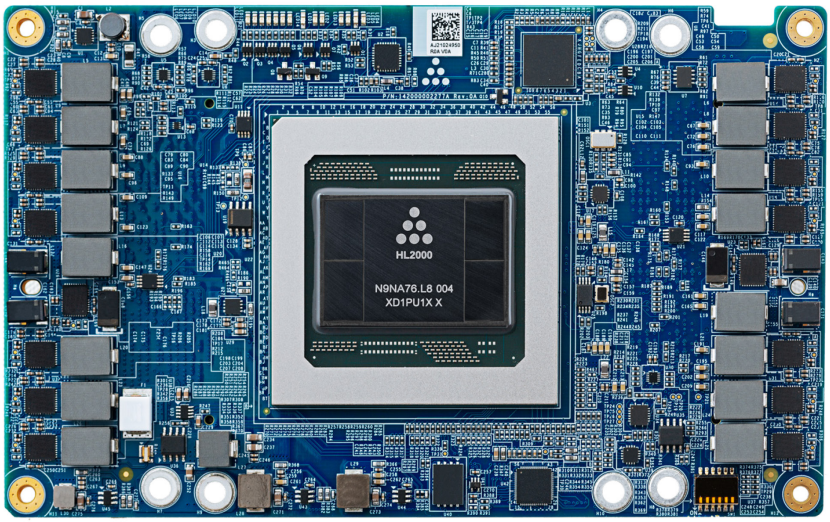

Figure 2. Habana HL-205 module. This OCP-compliant daughtercard contains the Gaudi ASIC (HL-2000), which is copackaged with 32GB of HBM2 in four stacks. (Photo source: Habana)

the HL-1000 Goya chip. The original chip features eight programmable VLIW cores plus a large matrix-multiply (GEMM) accelerator (see *MPR 2/18/19,* "Habana Wins Cigar for AI Inference"). Such a design is well suited to convolutional neural networks (CNNs), and the company has published impressive inference scores for ResNet-50 and other CNNs. These benchmarks rely on the 8-bit-integer (INT8) data type that Goya emphasizes.

Simply adding floating-point capability to the GEMM accelerator would be a good start for training. The Nvidia chips perform half-precision floating-point (FP16) operations at half the rate of INT8 operations, and Gaudi is likely the same. Gaudi's performance advantage over the T4 in FP16 training is about the same as Goya's advantage over the T4 in INT8 inference, bolstering our view that the two Habana chips have the same basic compute structures.

For external memory, the inference chip supports DDR4 DRAM, whereas Gaudi moves to HBM2. Habana withheld details on Goya's DRAM subsystem, but a four-channel DDR4-3200 implementation would produce 136GB/s of peak bandwidth. As Figure 2 shows, Gaudi employs four HBM2 stacks, delivering a total of 1,024GB/s. This dramatic bandwidth increase maintains compute efficiency for the more stressful training function, which requires back propagation of weights. The greater bandwidth also enables larger models, but HBM2 is much more expensive than DDR4 DRAM. The Gaudi products provide 32GB of memory per chip—twice the capacity of the Goya card.

The older Goya has a x16 PCIe Gen 4 interface but no other high-speed I/O. For large neural networks, however, the single 252Gbps connection can become a bottleneck. Nvidia addresses this problem using its proprietary NVLink, a 200Gbps memory-coherent interface; each V100 chip offers six NVLink connections. Habana prefers an open standard, implementing 10 ports of 100Gbps Ethernet (100GbE) on Gaudi to provide a similar amount of per-chip bandwidth. Each port is divisible into two 50GbE connections. To implement these links, the chip integrates 56Gbps PAM4 serdes, which can drive signals across a backplane or through cables up to three meters long, enough for connections within a rack.

Each port enables the RoCEv2 (RDMA Over Converged Ethernet) protocol, which performs low-latency memory-to-memory transfers across standard Ethernet. But RoCE (pronounced "rocky") isn't coherent, meaning the devices don't share a single memory space as they do in an Nvidia DGX system. The shared-memory model is easier to program when dividing a large model across multiple accelerators. But NVLink doesn't scale beyond 16 processors, whereas Ethernet can form arbitrarily large networks.

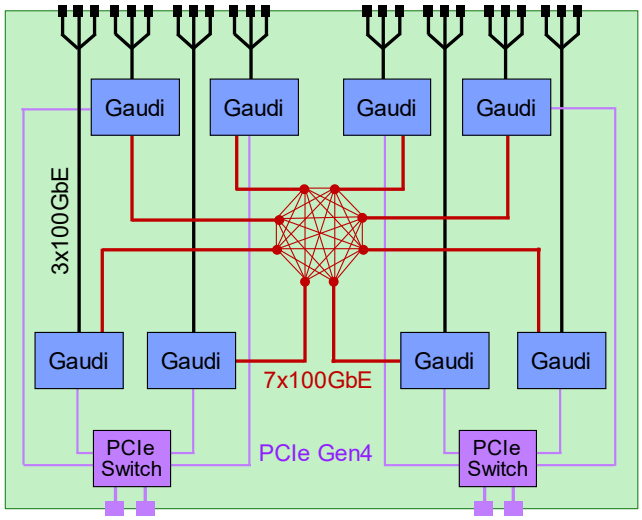

Figure 3. HLS-1 functional diagram. Habana offers a system with eight Gaudi modules that all directly connect via 100G Ethernet, leaving three 100GbE ports per chip for external connections. The accelerator system connects to a standard server through four PCIe ports.

Eight in a Box

Figure 3 shows how Gaudi employs the new I/O capabilities in a multichip configuration: specifically, the eight-chip HLS-1 system. In this design, each Gaudi chip connects directly to each of the seven others using one of its 100GbE ports, creating a fully connected nonblocking mesh. The remaining three Ethernet ports from each chip go to the front panel; they can connect to additional HLS-1 systems. To save space, these ports are arranged as six QSFP-DD connectors. The complete system fits in three standard rack units (3U) and draws a maximum of 3,000W—roughly the same size and power as Nvidia's DGX-1.

The HLS-1 motherboard includes two six-port PCIe switches that connect the eight Gaudi chips to four PCIe ports on the backplane, as the figure shows. These ports can connect to any standard server, allowing customers to choose the type and amount of CPU performance for their application (model). Although the HLS-1 is an off-the-shelf scalable solution, Habana expects many customers to design their own systems, taking advantage of standard PCIe and Ethernet components.

By comparison, the V100 has only six NVLink ports, so the eight-chip DGX-1 design creates a hypercube mesh that requires two hops to move data between certain pairs of chips. On the other hand, NVLink provides twice the bandwidth of 100GbE, ultimately providing greater bisection bandwidth for an eight-chip board. To create the 16-chip DGX-2 system, Nvidia had to design a switch chip for its proprietary protocol. Other companies can't build similar systems without this proprietary chip, although they can build eight-chip designs. A Gaudi customer could build a 16-chip system without a switch chip by splitting the Ethernet ports into 50Gbps links and fully interconnecting the processors.

Scaling to Massive Clusters

Habana customers can combine HLS-1 systems using their 24 external 100GbE ports. For example, two systems could connect using all 24 links for maximum bandwidth. In a large configuration, 24 HLS-1 systems could be fully connected with a single 100GbE link between each, joining a total of 192 Gaudi chips. Splitting the links into 50GbE doubles the number of directly connected systems. Customers can create even larger HLS-1 clusters using commodity 64-port Ethernet switches to connect the systems. As the bandwidth between systems diminishes, however, so does the scaling efficiency, particularly for more-complex models. Ethernet switches, especially when heavily loaded, have nondeterministic latency that can bog down large networks.

Nvidia's DGX-1 system features four 100Gbps InfiniBand cards, providing 50Gbps per GPU; the DGX-2 offers eight links for 16 GPUs, yielding the same per-chip bandwidth. Although InfiniBand is a standard, few vendors support it. In fact, Nvidia recently acquired the primary vendor, Mellanox, to avoid any supply disruption (see *MPR 4/8/19,* "Nvidia to Acquire Mellanox"). InfiniBand natively supports RDMA, allowing customers to build large clusters of Nvidia GPUs. Nvidia created the NCCL ("nickel") software stack to optimize internode communication latency.

Unfortunately, this approach works poorly in very large clusters. According to Nvidia's benchmarks, a single DGX-2 system with 16 V100 chips achieves 91% scaling on ResNet-50 training (that is, relative to the ideal of 16x the performance of a single V100). The company recently published MLPerf scores for clusters of 512 and 640 V100 chips; these clusters comprised DGX-2 systems. The scaling for these extremely large configurations was only 27% and 24%, respectively, indicating that most of the GPU performance is lost to communication bottlenecks.

Price and Availability

Habana plans to sample the Gaudi module and PCIe card, along with the HLS-1 system, by the end of this year. We expect production versions by mid-2020. The company withheld pricing. For more information, access *www.habana.ai*.

Having six times more bandwidth per chip than the DGX systems, the HLS-1 scales more effectively. As Figure 4 shows, Habana estimates 90% scalability for a 64-chip cluster and 80% scalability for a 512-chip cluster, but it has yet to build enough chips to verify these estimates. Because a single Gaudi performs similarly to a single V100 on ResNet-50 training, Habana's performance advantage in the largest clusters grows to more than 3x on the basis of these estimates. Note that the size and type of model will affect the efficiency rating; ResNet-50 is a relatively simple model that requires interchip communication mainly for the fully connected layers, but more-complex models may scale less well.

Few customers build such large clusters. The 512-chip V100 cluster can train ResNet-50 across 82 epochs of 1.3 million images in just six minutes. Many popular models are more complex than ResNet-50, but a handful of eight-chip systems can train most models in a few hours. Even large data centers with many training systems will typically install several clusters that can each train different models rather than installing a single massive cluster. But for customers that have outrageously large networks or want to minimize training time, the ability to build large, highly efficient clusters is valuable.

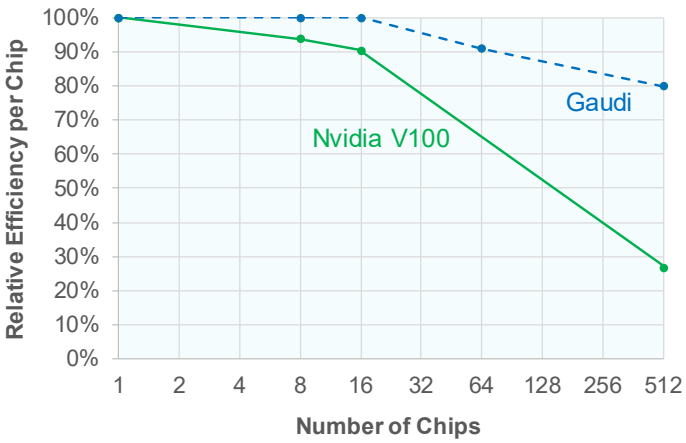

Figure 4. Training scalability on ResNet-50. Both Habana and Nvidia demonstrate high efficiency in clusters of up to 16 chips, but Nvidia loses far more performance in very large clusters. Gaudi data is estimated. (Data source: vendors)

Freedom From Nvidia

On ResNet-50, Habana's first training chip delivers performance similar to that of Nvidia's V100. Gaudi's greater network bandwidth improves scalability for customers that build clusters with more than 16 accelerator chips. Although Habana has withheld pricing for the Gaudi products, we expect they'll sell at a considerable discount relative to Nvidia's high prices. For example, a DGX-1 system has a list price of $129,000, although it includes a complete host server. Habana's HLS-1 instead connects to a separate server, allowing customers to configure more or less CPU performance depending on their needs and the models they run. Gaudi has about the same TDP as the V100 but runs cooler on many models (such as ResNet-50), providing further energy-cost savings. Gaudi's use of standard Ethernet instead of NVLink and InfiniBand simplifies building and scaling systems.

Habana falls short of Nvidia on the software side. Gaudi's rated performance requires the Habana SynapseAI API; the company plans to have its first standard training framework, TensorFlow, running in September, with other frameworks to follow. Even for inference, Habana doesn't support frameworks other than TensorFlow except through the ONNX interface. The open-source Glow software could eventually replicate Nvidia's Cuda, but for now, it's immature and far from complete. Furthermore, Habana's architecture is best suited to CNNs; the company hasn't released benchmarks on any other type of model. Nvidia's GPU-based architecture is a more general-purpose design.

It's somewhat unfair to compare Gaudi to the V100, which has been in production since late 2017. By the time the 16nm Gaudi reaches production, Nvidia aims to release its next-generation GPU. That chip will shrink to 7nm transistors, increasing power efficiency, and will include other performance improvements as part of the new Ampere architecture. Nvidia also plans to boost scalability for large clusters, probably by adding more InfiniBand/Ethernet ports in the next-generation DGX-3. These improvements are likely to overcome Gaudi's small per-chip performance lead on ResNet-50, although Habana should retain power-efficiency and cost advantages. The startup plans to eventually move both Goya and Gaudi to 7nm technology, but its 7nm training product is unlikely to appear sooner than 2021.

Habana's Goya is an easy sell given its big advantages in performance and power efficiency. Instead of a large performance gain, Gaudi provides freedom from high prices as well as freedom from proprietary software and proprietary interfaces. This freedom will appeal to hyperscale customers that currently buy many expensive Nvidia GPUs for neural-network training. Habana is particularly aligned with Facebook, becoming the first vendor to announce hardware for the social network's OCP form factor and Glow software. Because Facebook is working with several other accelerator suppliers, Habana's efforts don't guarantee big shipments. But as the first company to announce a training accelerator that beats Nvidia's performance, Habana is the leading challenger in this category.♦