Habana Labs Launches Gaudi2 Deep Learning Training Processor

Habana Gaudi2 processor demonstrates 2x throughput performance over Nvidia A100 for training popular computer vision and NLP models.

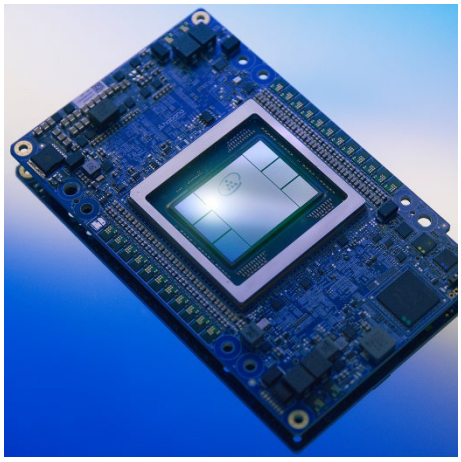
May 10, 2022 — Today at the Intel Vision conference, Habana Labs, an Intel company, launched the Gaudi®2 processor, its second-generation Gaudi® processor for training, and for inference deployments introduced the Greco processor, the soon-to-market successor to the Goya™ processor. The processors are purpose-built for AI deep learning applications. Implemented in 7 nanometer, they build upon Habana's high-efficiency architecture to provide customers with higher-performance model training and inferencing for computer vision and natural language applications in the data center.

At the conference, Habana demonstrated Gaudi2 training throughput performance on computer vision – ResNet-50 (v1.1) – and natural language processing – BERT Phase-1 and Phase-2— workloads, nearly twice that of the Nvidia A100 80GB processor.

More: For more information about the Habana® Gaudi®2 Processor, including the launch news, please visit the Intel Newsroom, Habana Training Solutions, Habana Gaudi2 whitepaper and Deep-Dive on Habana® Gaudi® Processor Video.

Gaudi2: Purpose-Designed for Deep Learning Training

For data center customers, the task of training deep learning models is increasingly time-consuming and costly due to the growing size and complexity of datasets and AI workloads. Gaudi2 was designed to bring improved deep learning performance and efficiency – and choice – to cloud and on-premises customers.



Habana Gaudi2 Mezzanine Card

To increase model accuracy and recency, customers require more frequent training. According to IDC, 74% of machine learning (ML) practitioners surveyed in 2020 run five to 10 training iterations of their models, more than 50% rebuild models weekly or more often, and 26% rebuild models daily or even hourly. And 56% of those surveyed cited cost-to-train as the No. 1 obstacle to their organizations taking advantage of the insights, innovations and enhanced end-customer experiences that AI can provide. The Gaudi platform solutions, first-gen Gaudi and Gaudi2, were born to address this growing need.

More: to see what customers and partners are saying about the deep learning advantages of Gaudi2, see our site.

Born for Deep Learning, now Raised to a New Level

The Habana Gaudi2 processor significantly increases training performance, building on the same high-efficiency first-generation Gaudi architecture that enables customers to enjoy up to 40% better price performance over existing GPU-based solutions in the cloud with Amazon EC2 DL1 instances and on-premises with the Supermicro X12 Gaudi Training Server.
Architectural advances from first-gen Gaudi to Gaudi2 feature:

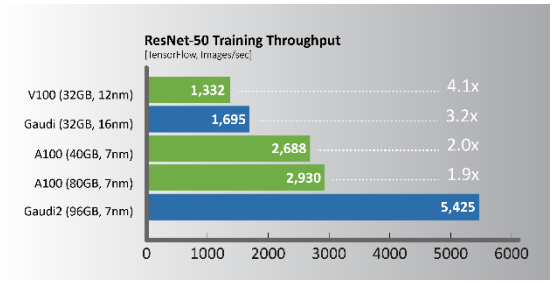
- Leap in process technology from 16 nm to 7 nm
- Introduction of new data types, including FP8, in the Matrix Multiplication Engine (MME) and Tensor Processor Core compute engines
- Increase from eight Tensor processor cores to 24
- Integration of on-chip media processing engine for offloading the host subsystem
- Triple in-package memory capacity from 32GB to 96GB HBM2E at 2.45TB/sec bandwidth
- Double on-board SRAM to 48MB
- Increased integrated RDMA over Converged Ethernet (RoCE2) from 10 integrated NICs to 24 for high-efficiency scale-up and scale-out on industry-standard networking.

100% Designed for AI, 200% Performance

Customers looking to increase time-to-train and operational efficiencies look to out-of-the-box training metrics to assess deep learning performance and value. At Intel Vision, Habana showed them just that – with the performance of the Gaudi2 processor relative to other leading solutions in the marketplace. The following charts are training results for popular computer vision and natural language processing models compared to published metrics of alternative solutions.

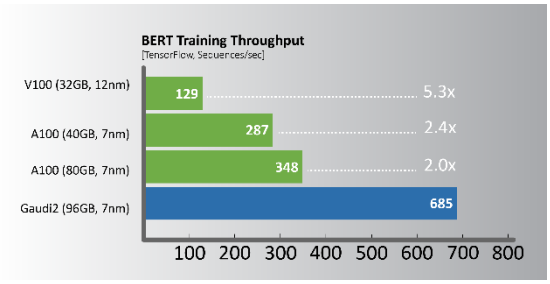
Compared with the A100 GPU, implemented in the same process node, Gaudi2 delivers clear leadership training performance – approximately 2x – as demonstrated with comparison on the following key workloads, including the full software integrated with the framework. These results suggest the Gaudi2 purpose-designed deep-learning acceleration architecture is fundamentally more efficient.

Computer vision – ResNet-50



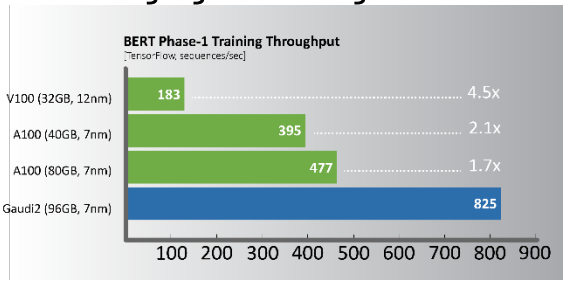
Natural Language Processing – BERT

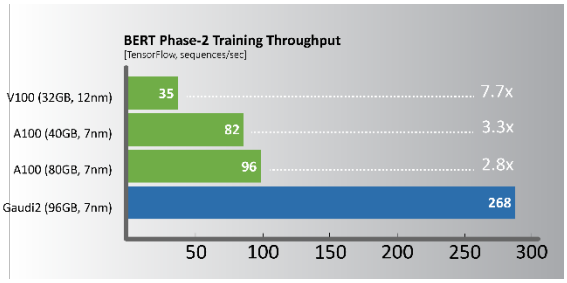
Effective throughput combining Phase-1 and Phase-2



For workloads and configurations, visit the Vision section at intel.com/performanceindex. Results may vary.

Natural Language Processing – BERT





For workloads and configurations, visit the Vision section at intel.com/performanceindex. Results may vary.

Networking Capacity, Flexibility, Efficiency

The integration of 24 100-Gigabit RoCE ports onto every Gaudi2 processor significantly amplifies training bandwidth.

- **Scale-up:** 21 of the ports on every Gaudi2 processor are dedicated to connecting to the other seven processors in the 8-card HLS-Gaudi®2 server, in an all-to-all, non-blocking configuration.
- **Scale-out:** Three of the ports on every processor are dedicated to scale out, providing 2.4 terabits of networking throughput in the 8-card Gaudi server.
- **OCP OAM compliant:** To simplify system design for customers, Habana provides a Universal Baseboard (UBB) compliant with the OCP specification as a product.
- **Ease and flexibility of use:** With the integration of industry-standard RoCE on chip, customers can easily scale and configure Gaudi2 systems to suit their deep learning cluster requirements, from one to thousands of Gaudi2s.
- **Choice for system build-out:** With system implementation on widely used industry-standard Ethernet connectivity, Gaudi2 enables customers to choose from a wide array of Ethernet switching and related networking equipment, enabling cost-savings.
- **Integration efficiency:** On-chip integration of the network interface controller (NIC) ports significantly lowers overall component costs.

HLS-Gaudi2 Server

In addition to the eight Gaudi2 cards, the HLS-Gaudi2 server features a dual-socket Intel® Xeon® Scalable Processor subsystem. Habana offers this server for customer evaluations of Gaudi2, while partnering with system OEMs to bring market solutions for end-customer deployments.

Advancements from the Gaudi2 Data Center:
To date, one thousand HLS-Gaudi2s have been deployed in the Habana Gaudi2 data centers in Israel to support research and development for Gaudi2 software optimization and to inform further advancements in the forthcoming Gaudi®3 processor.

More: to build deep learning training systems with Habana Gaudi2, see the video.

Access to Habana Gaudi2 solutions: Gaudi2 is available to Habana customers. Performance results were measured on the HLS-Gaudi2 server that is available from Habana. Supermicro announced it will be bringing to market soon the Supermicro Gaudi®2 AI Training Server: SYS-820GH-TNR2 that will contain eight Gaudi2 and dual 3rd Gen Intel® Xeon® Scalable processors. Habana is also working with DDN to deliver a turnkey server featuring the Supermicro Gaudi2 server with augmented AI storage with the pairing of the DDN AI400X2 storage solution.

Habana Gaudi2 8-node Cluster

Simplified Model Build and Migration: Meeting Developers Where They are

To support customers as they transition workloads and systems to Gaudi®2 from existing GPU-based models and help them preserve their software development investments, the Habana SynapseAI® Software Suite is optimized for deep learning workloads and designed to ease model

build and migration. Meeting deep learning developers where they are, SynapseAI integrates TensorFlow and PyTorch frameworks and provides over 30 popular computer vision and natural language reference models. Developers are given support with documentation and tools, how-to content and community support on the Habana Developer Site and provided reference models and model roadmap on the Habana GitHub.

More: For a deep dive into developer support on Gaudi® and Gaudi®2 processors, see the Habana Developer Site.

About Intel

Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, go to newsroom.intel.com and intel.com.

About Habana Labs

Habana Labs, an Intel company, is a leading AI Processor company founded in 2016 to develop purpose-built processor platforms optimized for training deep neural networks and for inference deployment in production environments. We are unlocking the true potential of AI with platforms offering orders of magnitude improvements in processing performance, scalability, cost, and power consumption. Acquired by Intel, Inc. in 2019, Habana operates as an independent business unit within the Intel Data Center & AI Products Group.