An Introduction to Habana AI Processors

Sree Ganesan
AI Software Product Management @ Habana Labs
www.habana.ai
Agenda

- Introduction
- Habana Gaudi AI processor
- Cloud & Datacenter solutions powered by Gaudi
- Software & Support
- Summary
Demand for compute for ML training doubles every 3.4 months

- **Increasing Complexity**
  - Businesses need higher precision in their model predictions
  - Results in larger and more complex models
  - Requires frequent retraining of models
- **Increasing Costs**
  - Increasing compute power required for frequent training of larger models drives up cost to train
  - Becomes a barrier for innovation and growth

Need for dedicated AI processors to address the compute, memory and communication challenges

Habana Labs © 2021
A little about Habana

• Founded in 2016 to develop purpose-built AI processors
• Launched inference processor in 2018, training processor in 2019
• Acquired by Intel in late-2019
• Fully leveraging Intel’s scale, resources and infrastructure
• Accessing Intel ecosystem and customer partnerships
• Delivering aggressive roadmap optimized for AI data center performance and efficiency
Habana’s Dedicated Focus: AI Training and Inference

Training Solution

GAUDI®

Inference Solution

GOYA™
Designed for AI training efficiency, flexibility and scale
A new class of AI Training: Habana Gaudi

Purpose-designed for data center AI Training efficiency

• Cost-efficient AI Training

• Flexibility to ease model migration

• Hardware and software architected for scalability
Gaudi: Architected for performance and efficiency

- Fully programmable Tensor Processing Cores (TPC) with tools & libraries
- Configurable Matrix Math Engine (GEMM)
- Multi-stage memory hierarchy with 32GB HBM2 memory
- Integrated 10 x 100 Gigabit Ethernet for multi-chip scale-out training
- Delivers higher efficiency than traditional CPUs and GPUs
# Designed for flexible and easy model migration

## Ease of use
Integrated with TensorFlow and PyTorch; minimal code changes to get started
- SynapseAI maps model topology onto Gaudi devices

## Customization
SynapseAI TPC SDK facilitates development of custom kernels

## Balanced compute & memory
32GB HBM2 memories similar to GPUs, so existing DL models will fit into Gaudi memory

Developers can enjoy the **same abstraction** they are accustomed to today

Developers can **customize** models to extract best performance

Developers can spend **less effort** to port their models to Gaudi
Designed for Scaling Efficiency

The industry’s FIRST: Native integration of 10 x 100 Gigabit Ethernet RoCE ports onto every Gaudi

• Eliminates network bottlenecks

• Standard Ethernet inside the server and across nodes

• Eliminates lock-in with proprietary interfaces

• Lowers total system cost and power by reducing discrete components
Scaling within a Gaudi Server

- 8 Gaudi OCP OAM cards
- 24 x 100GbE RDMA RoCE for scale-out
- Non-blocking, all-2-all internal interconnect across Gaudi AI processors
- Separate PCIe ports for external Host CPU traffic

Example of Integrated Server with eight Gaudi AI processors, two Xeon CPU and multiple Ethernet Interfaces
Rack and Pod Level Scaling

Easily build rack and pod-scale training systems with off-the-shelf standard ethernet switches

Example of rack configuration with four Gaudi servers (eight Gaudi processors per server) connected to a single Ethernet switch
Gaudi-based
Amazon EC2 AI Training Instances
Gaudi-Based AWS EC2 Instances Coming Soon

“The new EC2 instances will leverage up to 8 Gaudi accelerators and deliver up to 40% better price/performance than current GPU-based EC2 instances for training DL models.”

Andy Jassy, re:Invent 2020

- Amazon’s first non-GPU instances based on Habana Gaudi AI processors
- Improved cost-efficiency makes AI Training accessible to more customers
Gaudi-Based EC2 Instances

Benefit from full stack of Amazon EC2 services:
• AWS DLAMI, DLC for Gaudi
• AWS ECS and EKS orchestration for containerized applications
• Integration with Amazon SageMaker
• Efficient scaling across multiple Gaudi-based EC2 Instances
On-Premise Solution
Partnering with Supermicro
Solutions available now for on-premises customers

Featured Servers:

• Supermicro X12 Gaudi® AI Training System
  ▪ Eight Gaudi HL-205 AI processors
  ▪ Dual-socket 3rd Gen Intel® Xeon® Scalable processors

• Supermicro SuperServer 4029GP-T
  ▪ Eight Goya™ HL-100 PCIe cards
  ▪ Dual-socket 2nd Gen Intel® Xeon® Scalable processors
Habana AI to power SDSC’s Voyager Research Program

336 Gaudi Training accelerators with native RoCE scaling and 16 Goya Inference processors

- Voyager to go into service Fall 2021
- Funded by $5M grant from National Science Foundation
  - Matching funds targeting community support and operation
- AI research conducted across range of science and engineering domains
  - Astronomy, climate sciences, chemistry, particle physics,
- Announced by SDSC in July 2020, more information [here](#).
Software, Resources & Support
SynapseAI Software Suite

- Train deep learning models on Gaudi with minimal code changes
- Natively integrated with TensorFlow & PyTorch
- Reference models, kernel libraries, software and docs available on GitHub
- Advanced users can write their own custom software kernels
Software Suite Detail

- **Software Suite**
  - **PyTorch**
  - **TensorFlow**

- **TPC Programming Tools**
  - Compiler, Debugger, Simulator

- **Habanalab API**
  - **GC-API, HCL-API and HCCL-API**
    - **Graph Compiler (GC)**
    - **Habana Communication Libraries (HCL, HCCL)**

- **User Mode Driver**
  - **Kernel Mode Driver**

- **Embedded & Management Tools**
  - Firmware update, management & monitoring

- **Debug & Profiling Tools**
  - **Profiler**

- **TensorFlow and PyTorch** integrated with SynapseAI and optimized for Gaudi. Developers enjoy the same abstraction that they are accustomed to today.

- Graph compiler generates optimized binary code that implements model topology on Gaudi.

- Implements efficient scale-up between Gaudi cards within a single node and scale-out across nodes for distributed training. HCL uses Gaudi integrated NICs for both scale-up and scale-out. HCCL allows users to enable Gaudi integrated NIC for scale-up and host NIC for scale-out, using NCCL-compatible API.


- Rich TPC kernel library, with wide variety of operations (for example, elementwise, non-linear, non-GEMM operators).
SynapseAI receives a computational graph of the model from the framework.

It identifies subgraphs (blue nodes) that can be accelerated by Gaudi.

The rest of the graph runs on CPU (yellow node).

The original graph is modified to replace the Gaudi subgraphs with encapsulated nodes (blue).

The framework runtime executes the modified graph.

For each encapsulated node, SynapseAI generates optimized binary code that runs on Gaudi.

TensorFlow integration with SynapseAI
import tensorflow as tf

from TensorFlow.common.library_loader import load_habana_module
load_habana_module()

(x_train, y_train), (x_test, y_test) = tf.keras.datasets.mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0

model = tf.keras.models.Sequential([tf.keras.layers.Flatten(input_shape=(28, 28)), tf.keras.layers.Dense(10),])

loss = tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True)

optimizer = tf.keras.optimizers.SGD(learning_rate=0.01)

model.compile(optimizer=optimizer, loss=loss, metrics=['accuracy'])

model.fit(x_train, y_train, epochs=5, batch_size=128)

model.evaluate(x_test, y_test)

Load the Habana libraries needed to use Gaudi aka HPU device.

Once loaded, the HPU device is registered in TensorFlow and prioritized over CPU.

When an Op is available for both CPU and HPU, the Op is assigned to the HPU.

When an Op is not supported on HPU, it runs on the CPU.
Habana Developer Platform

Welcome to Habana’s developer site.

Here you will find the content, guidance, tools and support needed to easily and flexibly build new or migrate existing AI models and optimize their performance to meet your AI requirements. You can also access the latest Satori software to build or update your infrastructure.

Get access to Habana’s programmable Tensor Processor Core and SatoriAI software stack with support for TensorFlow and PyTorch frameworks, along with our model garden, libraries, containers and tools that enable you to build popular AI models.

For Habana’s latest performance numbers on our reference models, please go to our Models and Performance page.

Docs
Access user guides, release notes, installation guides and more.
Go to documentation →

Resources
Get detailed technical guides to start running models on Satori.
Go to Resources →

Community Forum
Be among the first to join Habana’s developer community on our Forum.
Go to Forum →

Update Your Software
Download the latest software, drivers and tools.
Download →

Habana Models and Performance Data
Access Habana’s popular model scripts and performance data.
Learn more →

Habana GitHub
Access reference models, support roadmap, setup and installation guides and more.
Habana GitHub →

https://developer.habana.ai
Habana Developer Resources

For Data Scientists

Get Started with Gaudi
Build or migrate models with our comprehensive set of guides and tools.
Get started →

Kernel Libraries
Extensive TPC kernel libraries, firmware, drivers and TPC SDK for custom kernel development.
Learn more →

Debugging and Optimization
Access documents for profiling and troubleshooting.
Learn more →

Habana Models
Access Habana's popular model scripts.
Learn more →

For IT and System Admin

Gaudi On-Premise
Set up and manage your on-premise servers and environments.
Learn more →

Update Your Software
Download the latest software, drivers and tools.
Click to access →

Account Specific Software
Access your login and data storage for users and devices, and global code.
Customer login →
Habana’s Developer Documentation

Getting Started
- Gaudi Architecture and Software Overview - Describes Gaudi architecture and the SynapseAI software suite.
- SynapseAI Installation Guide - Describes how to obtain and install the SynapseAI software package.
- Gaudi Migration Guide -

SynapseAI User Guides
- TensorFlow User Guide - Provides guidelines on how to modify existing models to run on TensorFlow platform.
- Horovod-based Scaling for TensorFlow - Provides guidelines on how to run distributed training using Gaudi.
- Gaudi Qualification Library - Provides information on the Habana Labs Qualification Tool (H_qual) for Gaudi.
- Profiler User Guide - Provides information about the SynapseAI Profiling Subsystem.
- System Management Interface Tool User Guide - Describes the system management interface tool.
- Debugging Guide - Provides recommendations and tips for debugging mode functionality and performance.

TPC User Guides
- TPC Tools Installation Guide - Provides installation instructions for the TPC-C compiler, assembler, disassembler and all necessary heaters.
- TPC Tools Debugger - Provides debugger and simulator installation instructions as well as user guidelines.

API Reference Guides
- Habana Communication Library (HCL) API Reference - Provides the list of Habana Communication Library APIs.
- Habana Collective Communications Library (HCCCL) API Reference - Provides the list of Habana Collective Communication Library APIs.
- Habana Labs Management Library (HLM) API Reference - Provides the list of APIs for monitoring and managing various states within Habana Labs’ AI accelerators.
- TensorFlow CustomOp API Reference - Provides documentation for TF CustomOp APIs.

Operators
- TensorFlow Operators - Summarizes the SynapseAI TensorFlow supported operators for Gaudi.
- PyTorch Operators - Summarizes the SynapseAI PyTorch supported operators for Gaudi.

© Copyright 2021, Habana Labs Revision Boosdia7.
# Habana Developer Software Vault

## Update Your Software

Download the latest SynapseAI(R) Software, including Habana's graph compiler and runtime, TPC kernel library, firmware and drivers, and tools. These components are needed to update an existing system to the latest drivers and firmware. For more information on how to install this content, please refer to the Installation Guide.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Download</th>
</tr>
</thead>
</table>
| habanaloa-graph | Installs the Graph Compiler and the run-time. | Ubuntu18.04  
Ubuntu20.04  
AmazonLinux2  
Centos7.5 |
| habanaloa-thunk | Installs the thunk library.         | Ubuntu18.04  
Ubuntu20.04  
AmazonLinux2  
Centos7.5 |
| habanaloa-dkms  | Installs the PCIe driver.           | Ubuntu18.04  
Ubuntu20.04  
AmazonLinux2  
Centos7.5 |
| habanaloa-fw-tools | Installs various Firmware embedded tools (etm, n-smi, etc.) | Ubuntu18.04  
Ubuntu20.04  
AmazonLinux2  
Centos7.5 |
| habanaloa-oseen | Installs hyperv level demo's data loader. | Ubuntu18.04  
Ubuntu20.04  
AmazonLinux2  
Centos7.5 |
| habanaloa-qual  | Installs the qualification application package. | Ubuntu18.04  
Ubuntu20.04  
AmazonLinux2 |

https://vault.habana.ai
Habana GitHub

https://github.com/HabanaAI
Software Installation and Deployment

Setup_and_Install repository on Habana GitHub provides instructions on how to setup your environment with the SynapseAI software stack

- **SynapseAI Orchestration**
  - (Kubernetes Gaudi plugin, Kubeflow mpi-operator)

- **SynapseAI TensorFlow Container Image**
  - (TensorFlow frontend, horovod, open-mpi)

- **SynapseAI Base Installation Image**
  - (OS, Gaudi linux kernel driver, user mode driver, graph compiler, HCL/HCLC & embedded tools)

- **Gaudi Server**

Gaudi-optimized Docker container images with all necessary dependences*

**Official releases publicly available on Habana Vault**

<table>
<thead>
<tr>
<th>Orchestration</th>
<th>Kubernetes (1.19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frameworks</td>
<td>TensorFlow2 and PyTorch</td>
</tr>
<tr>
<td>Operating Systems</td>
<td>Ubuntu 18.04 and 20.04</td>
</tr>
<tr>
<td>Container Runtimes</td>
<td>Docker (Docker CE version 18.09)</td>
</tr>
<tr>
<td>Distributed Training Schemes</td>
<td>TensorFlow with Horovod and tf.distribute PyTorch distributed (native)</td>
</tr>
</tbody>
</table>

*Habana GitHub will have repository with Dockerfiles to “build your own” Docker images*
### Gaudi Reference Models – July 2021

- Gaudi reference model [roadmap](https://habanaai/synapsai-roadmap) on Habana GitHub
- Scripts and detailed instructions to enable the reference models on Gaudi available on the [Model-References](https://model-references) repository
Habana’s Developer Forum

Welcome to Habana's Developer Forum

We are creating a new community for developers and data scientists who wish to post questions and solutions, and share in creating the next new thing in AI, leveraging a new class of AI accelerators from Habana. Be among the first to join this new way to AI.

- **Announcements**: 0 topics
- **Training**: 2 topics
  - PyTorch
  - TensorFlow
- **Inference**: 0 topics
- **Uncategorized**: 1 topic
  - Topics that don’t need a category, or don’t fit into any other existing category.
- **Site Feedback**: 0 topics
  - Discussion about this site, its organization, how it works, and how we can improve it.

https://forum.habana.ai
There are many exciting opportunities for deep learning in scientific research.

Habana Labs invites you to explore the possibilities with our Gaudi AI Training Processor!
THANK YOU

www.habana.ai