August 2, 2021

An Introduction to Habana AI Processors

Sree Ganesan

AI Software Product Management @ Habana Labs

www.habana.ai

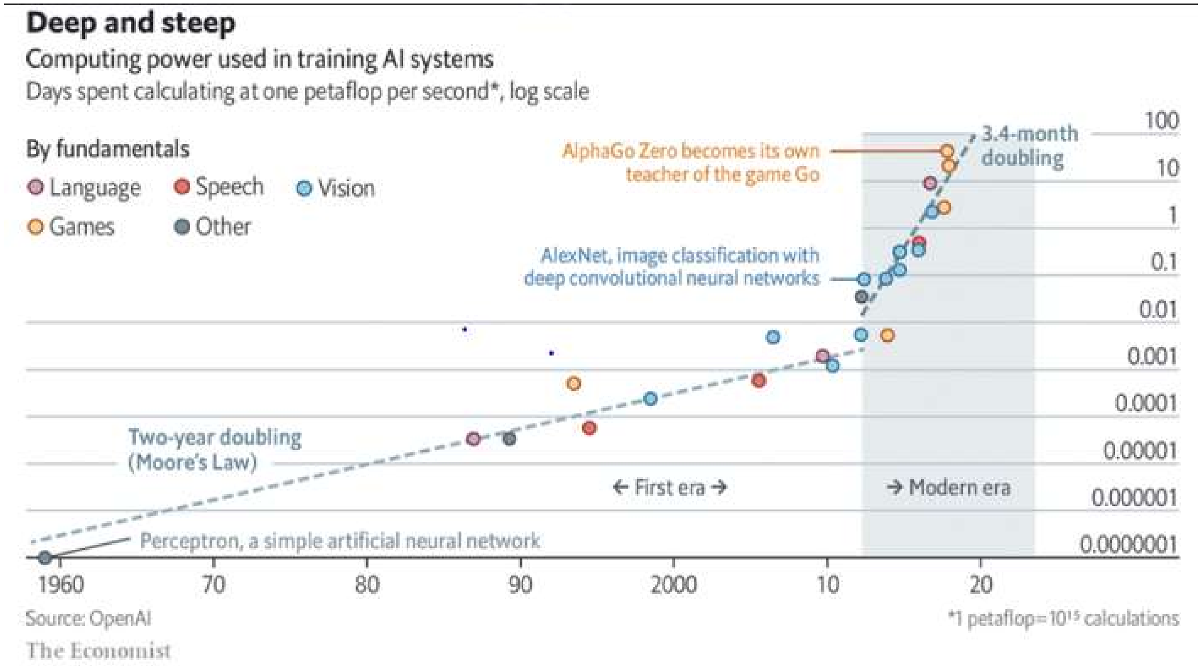
habana
An Intel Company

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



Deep and steep

Computing power used in training AI systems

Days spent calculating at one petaflop per second*, log scale

Source: OpenAI

The Economist

*1 petaflop=10^{15} calculations

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

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

intel

&

habana®

Habana’s Dedicated Focus: AI Training and Inference

Training Solution

GAUDI®

SynapseAI®

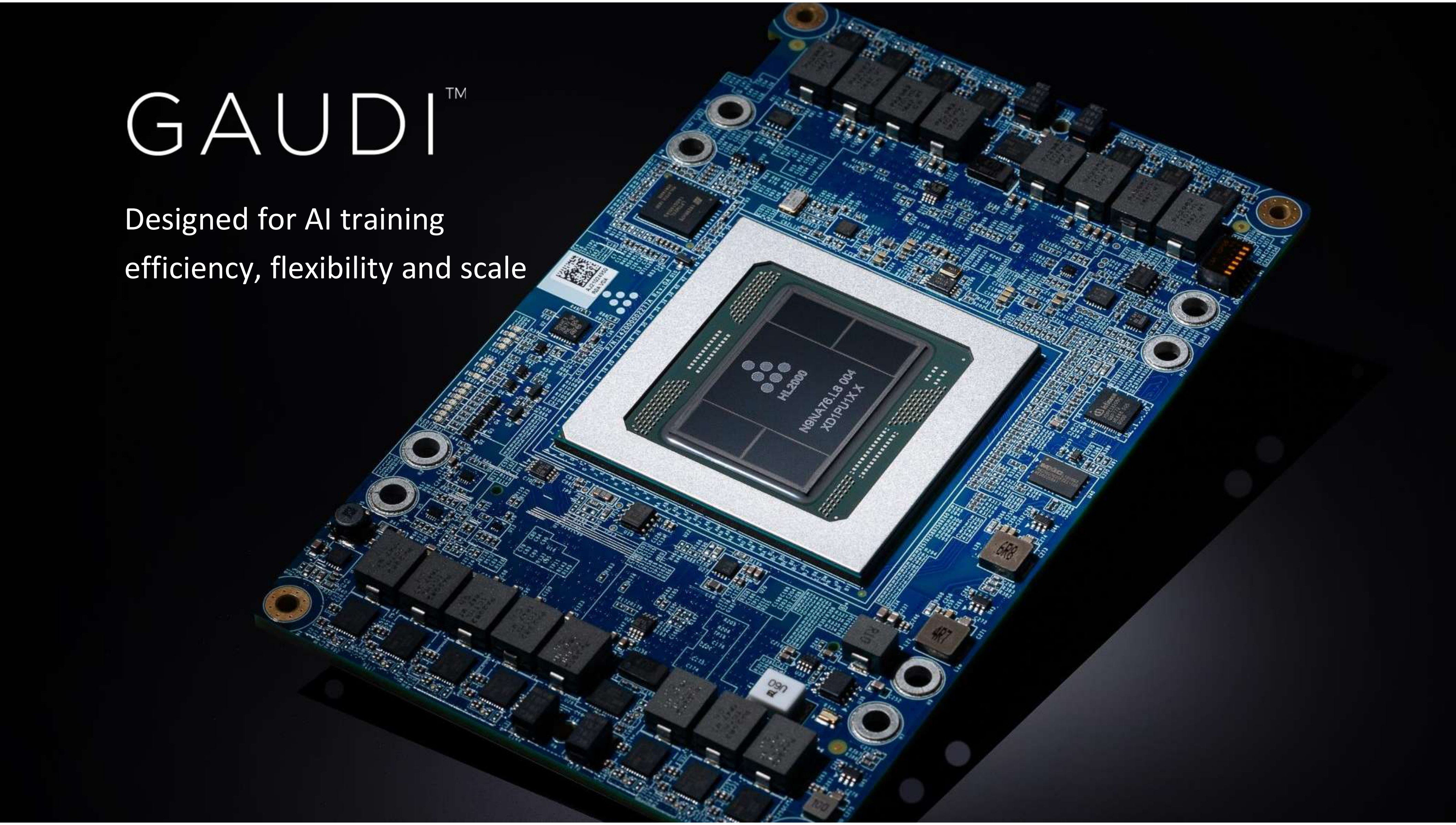
Inference Solution

GOYA™

SynapseAI®

GAUDI™

Designed for AI training efficiency, flexibility and scale



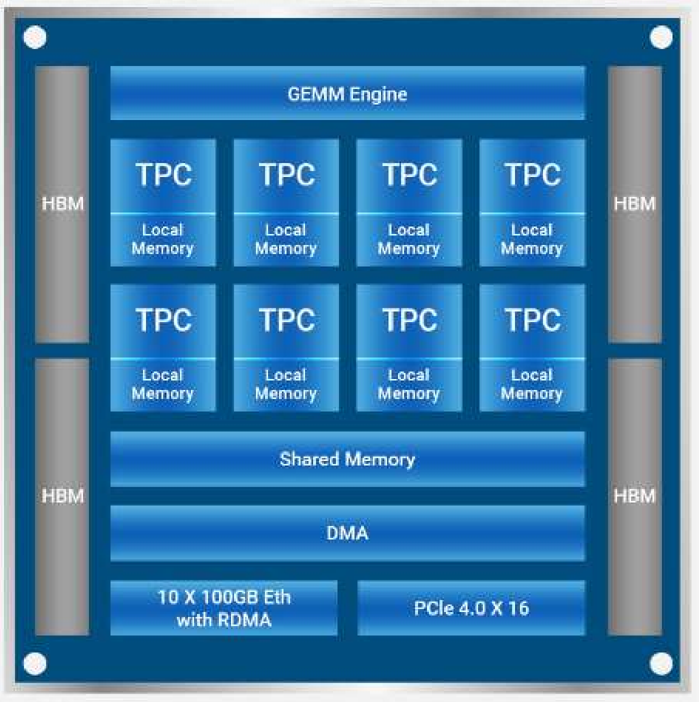
A new class of AI Training: Habana Gaudi

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	Customization	Balanced compute & memory
Integrated with TensorFlow and PyTorch; minimal code changes to get started ➞ SynapseAI maps model topology onto Gaudi devices	SynapseAI TPC SDK facilitates development of custom kernels	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

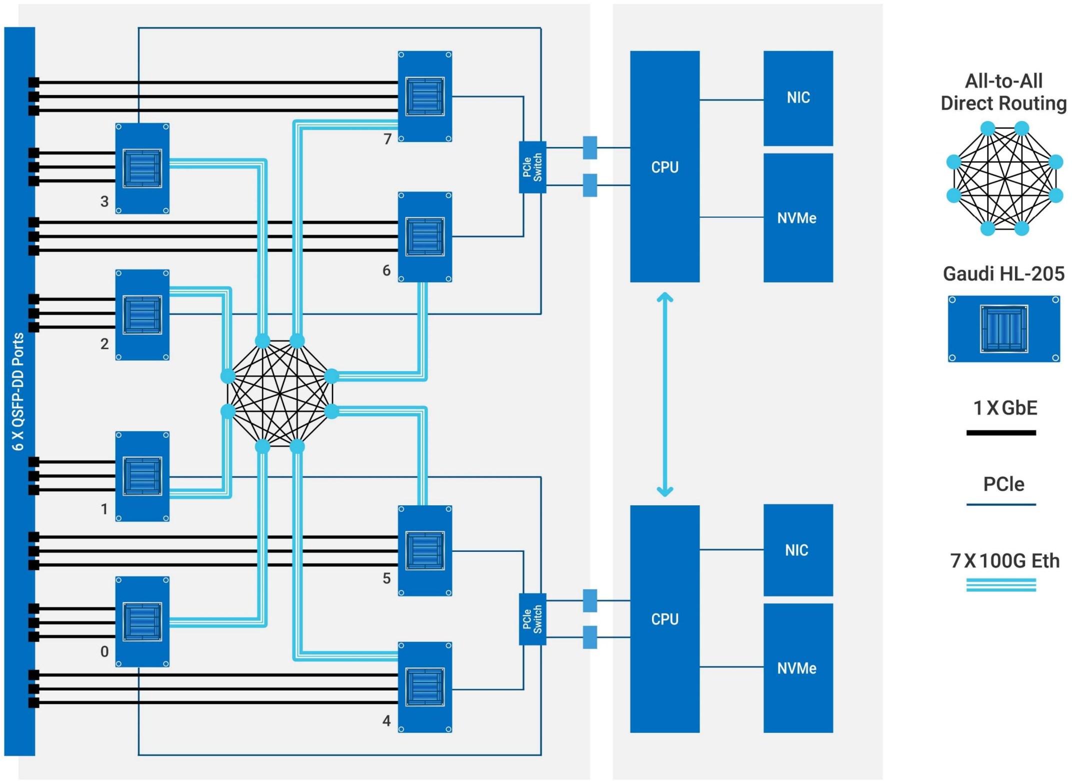
Designed for Scaling Efficiency

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

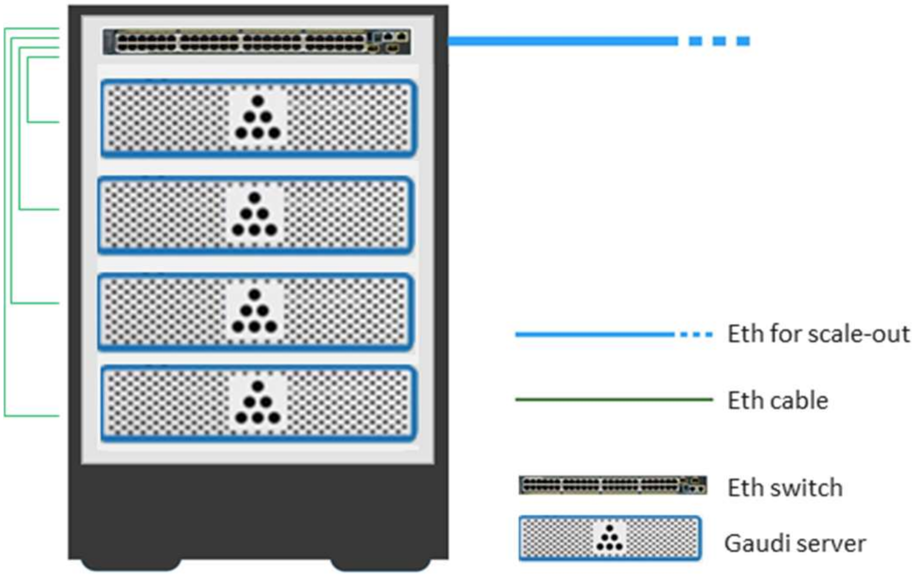
- 8 Gaudi OCP OAM cards
- 24 x 100GbE RDMA RoCE for scale-out
- Non-blocking, all-2-all internal interconnect across Gaudi AI processors
- Separate PCIe ports for external Host CPU traffic



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

NEW!

Habana Gaudi-based Amazon EC2 instances

Built specifically for ML training and powered by new Habana Gaudi processors from Intel

AVAILABLE FIRST HALF OF 2021

Up to 40% better price/performance over current GPU-based EC2 ML training instances

Support for TensorFlow and PyTorch

> *"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
  - https://www.supermicro.com/en/products/system/AI/4U/SYS-420GH-TNGR

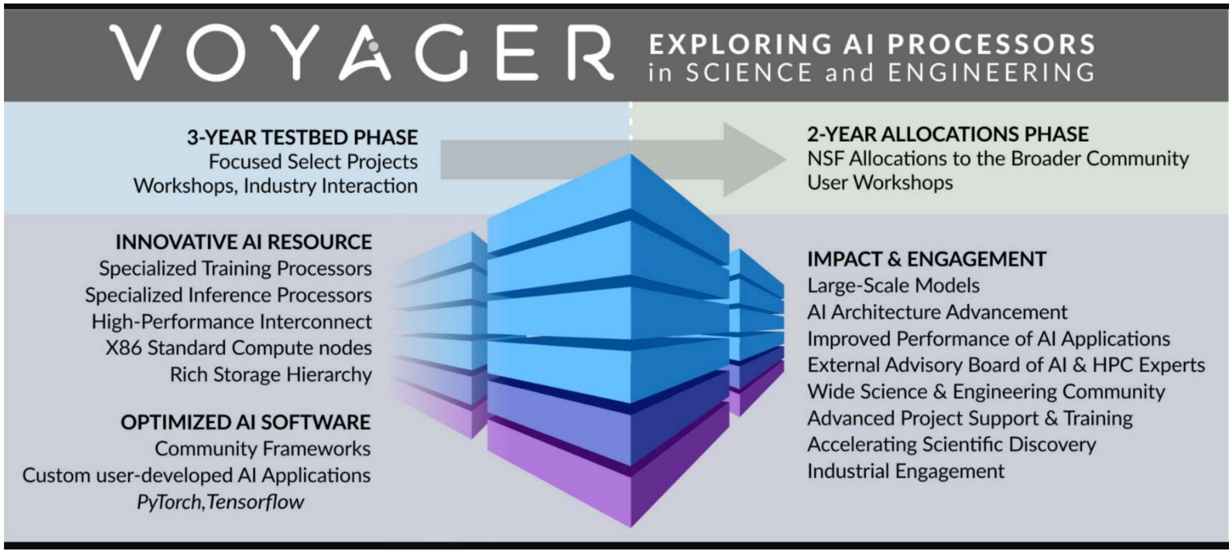
- Supermicro SuperServer 4029GP-T
  - Eight Goya™ HL-100 PCIe cards
  - Dual-socket 2nd Gen Intel® Xeon® Scalable processors
  - https://www.supermicro.com/en/products/system/4U/4029/SYS-4029GP-TRT.cfm

Supermicro
X12 Gaudi
AI Training System

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.



VOYAGER EXPLORING AI PROCESSORS in SCIENCE and ENGINEERING

3-YEAR TESTBED PHASE
Focused Select Projects
Workshops, Industry Interaction

2-YEAR ALLOCATIONS PHASE
NSF Allocations to the Broader Community
User Workshops

INNOVATIVE AI RESOURCE
Specialized Training Processors
Specialized Inference Processors
High-Performance Interconnect
X86 Standard Compute nodes
Rich Storage Hierarchy

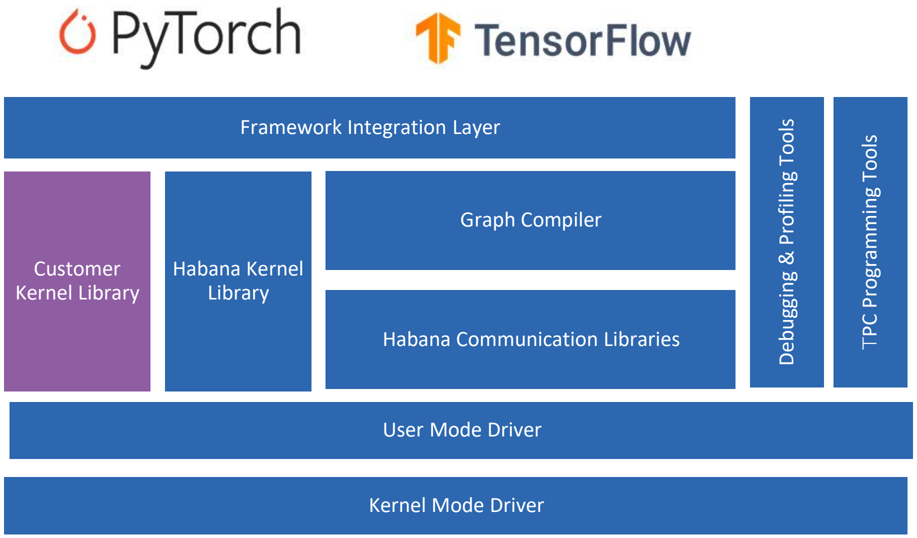
OPTIMIZED AI SOFTWARE
Community Frameworks
Custom user-developed AI Applications
PyTorch,Tensorflow

IMPACT & ENGAGEMENT
Large-Scale Models
AI Architecture Advancement
Improved Performance of AI Applications
External Advisory Board of AI & HPC Experts
Wide Science & Engineering Community
Advanced Project Support & Training
Accelerating Scientific Discovery
Industrial Engagement

Software,
Resources & Support

SynapseAI Software Suite

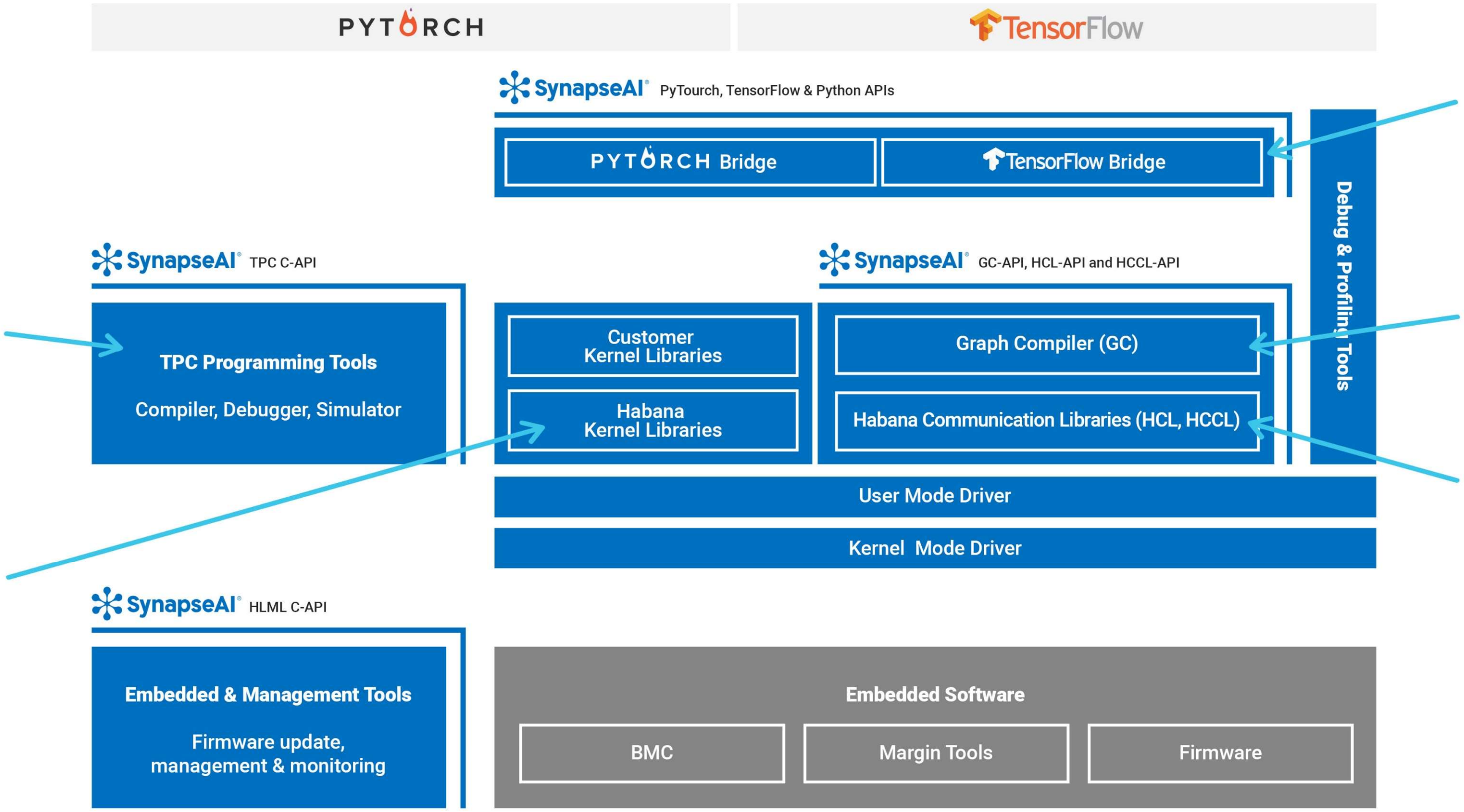
- Train deep learning models on Gaudi with minimal code changes
- Natively integrated with TensorFlow & PyTorch
- Reference models, kernel libraries, software and docs available on GitHub
- Advanced users can write their own custom software kernels



Software Suite Detail

LLVM-based TPC-C compiler, simulator and debugger facilitate development of custom TPC kernels. TensorFlow and PyTorch.

Rich TPC kernel library, with wide variety of operations (for example, elementwise, non-linear, non-GEMM operators).



PYTORCH

TensorFlow

SynapseAI® PyTourch, TensorFlow & Python APIs

PYTORCH Bridge

TensorFlow Bridge

Debug & Profiling Tools

SynapseAI® TPC C-API

TPC Programming Tools

Compiler, Debugger, Simulator

Customer Kernel Libraries

Habana Kernel Libraries

SynapseAI® GC-API, HCL-API and HCCL-API

Graph Compiler (GC)

Habana Communication Libraries (HCL, HCCL)

User Mode Driver

Kernel Mode Driver

SynapseAI® HLML C-API

Embedded & Management Tools

Firmware update, management & monitoring

Embedded Software

BMC

Margin Tools

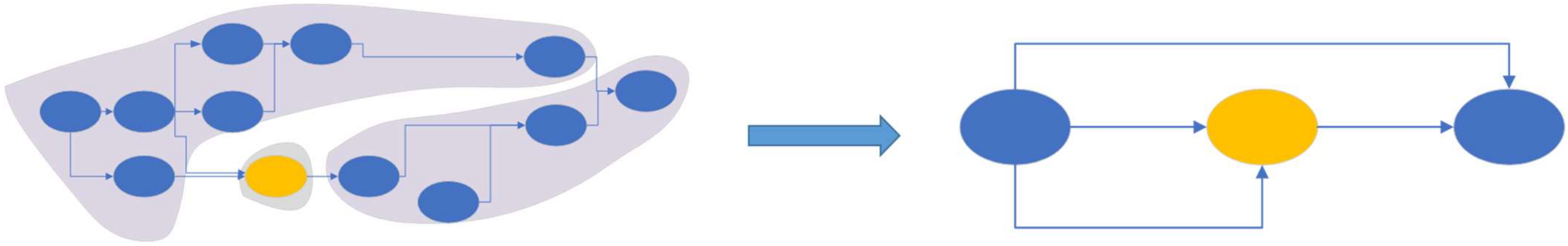
Firmware

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

TensorFlow integration with SynapseAI



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

Getting Started with TensorFlow on Gaudi

```
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.

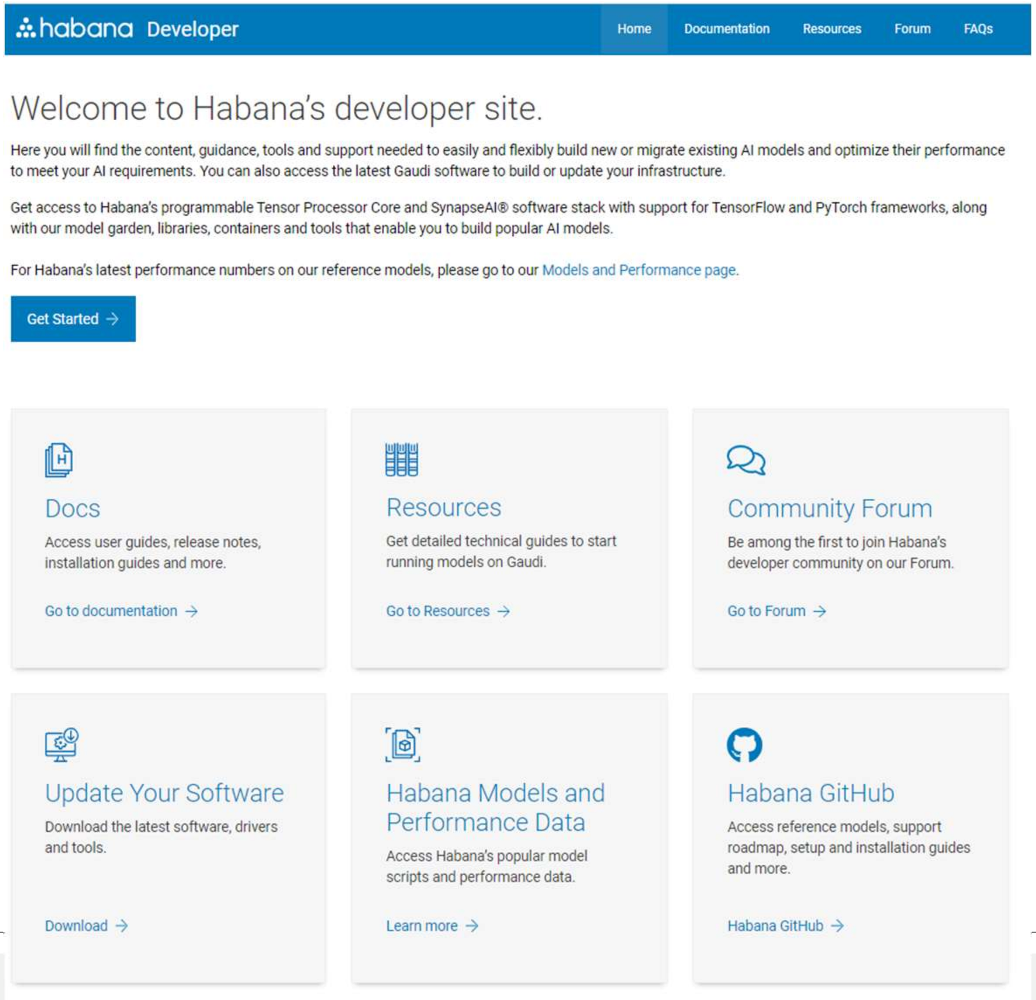
Once loaded, the **HPU device** is registered in TensorFlow and prioritized over CPU.

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

https://developer.habana.ai



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 Gaudi software to build or update your infrastructure.

Get access to Habana's programmable Tensor Processor Core and SynapseAI® 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 Gaudi.

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.

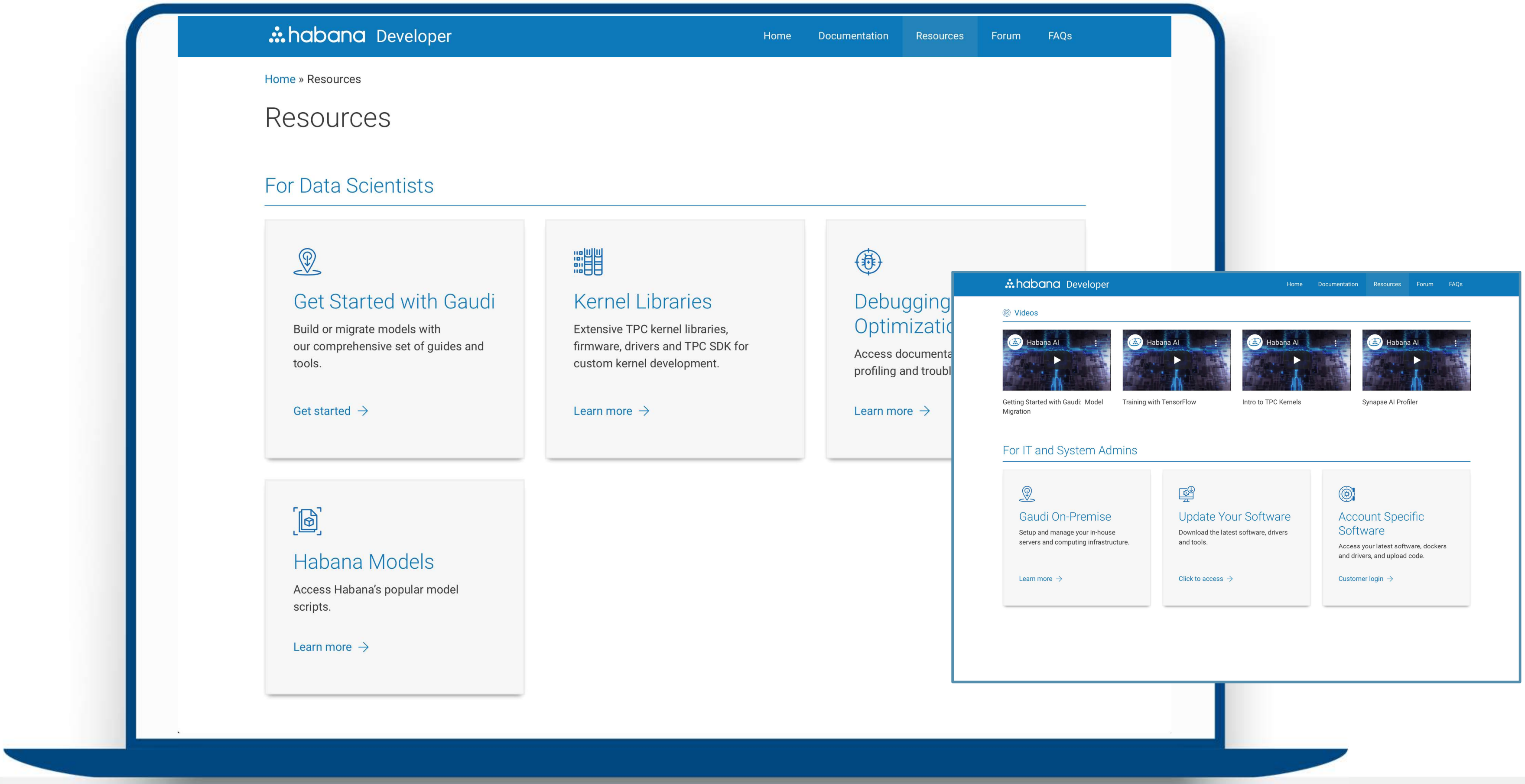
Learn more →

Habana GitHub

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

Habana GitHub →

Habana Developer Resources



Home » Resources

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 Optimizatio

Access documenta profiling and troubl

Learn more →

Habana Models

Access Habana's popular model scripts.

Learn more →

Videos

Getting Started with Gaudi: Model Migration

Training with TensorFlow

Intro to TPC Kernels

Synapse AI Profiler

For IT and System Admins

Gaudi On-Premise

Setup and manage your in-house servers and computing infrastructure.

Learn more →

Update Your Software

Download the latest software, drivers and tools.

Click to access →

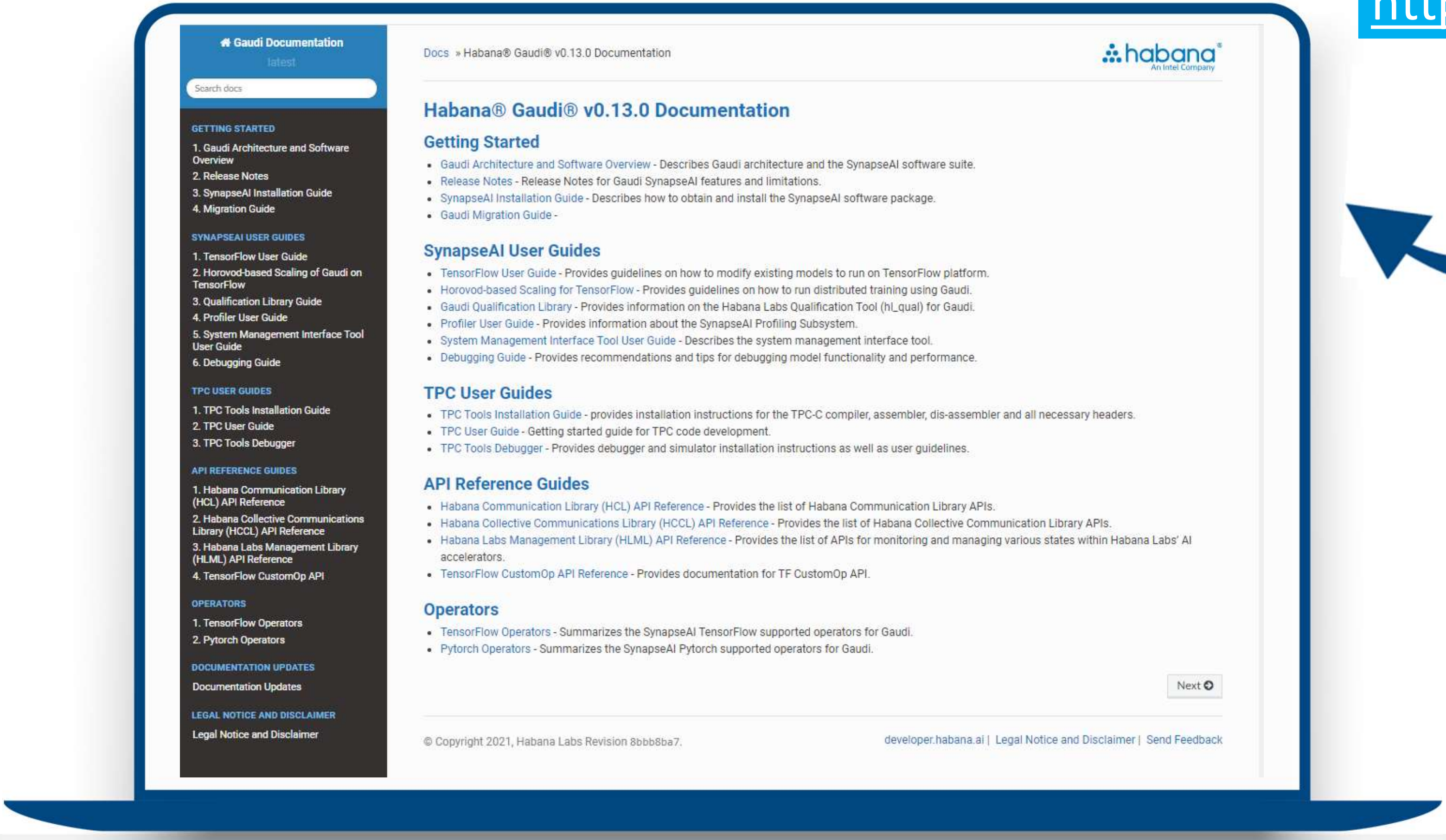
Account Specific Software

Access your latest software, dockers and drivers, and upload code.

Customer login →

Habana’s Developer Documentation

https://docs.habana.ai


Gaudi Documentation
latest
Search docs
GETTING STARTED
1. Gaudi Architecture and Software Overview
2. Release Notes
3. SynapseAI Installation Guide
4. Migration Guide
SYNAPSEAI USER GUIDES
1. TensorFlow User Guide
2. Horovod-based Scaling of Gaudi on TensorFlow
3. Qualification Library Guide
4. Profiler User Guide
5. System Management Interface Tool User Guide
6. Debugging Guide
TPC USER GUIDES
1. TPC Tools Installation Guide
2. TPC User Guide
3. TPC Tools Debugger
API REFERENCE GUIDES
1. Habana Communication Library (HCL) API Reference
2. Habana Collective Communications Library (HCCL) API Reference
3. Habana Labs Management Library (HLML) API Reference
4. TensorFlow CustomOp API
OPERATORS
1. TensorFlow Operators
2. Pytorch Operators
DOCUMENTATION UPDATES
Documentation Updates
LEGAL NOTICE AND DISCLAIMER
Legal Notice and Disclaimer
Docs » Habana® Gaudi® v0.13.0 Documentation
habana An Intel Company

Habana® Gaudi® v0.13.0 Documentation

Getting Started

- Gaudi Architecture and Software Overview - Describes Gaudi architecture and the SynapseAI software suite.
- Release Notes - Release Notes for Gaudi SynapseAI features and limitations.
- 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 (hl_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 model functionality and performance.

TPC User Guides

- TPC Tools Installation Guide - provides installation instructions for the TPC-C compiler, assembler, dis-assembler and all necessary headers.
- TPC User Guide - Getting started guide for TPC code development.
- 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 (HCCL) API Reference - Provides the list of Habana Collective Communication Library APIs.
- Habana Labs Management Library (HLML) 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 API.

Operators

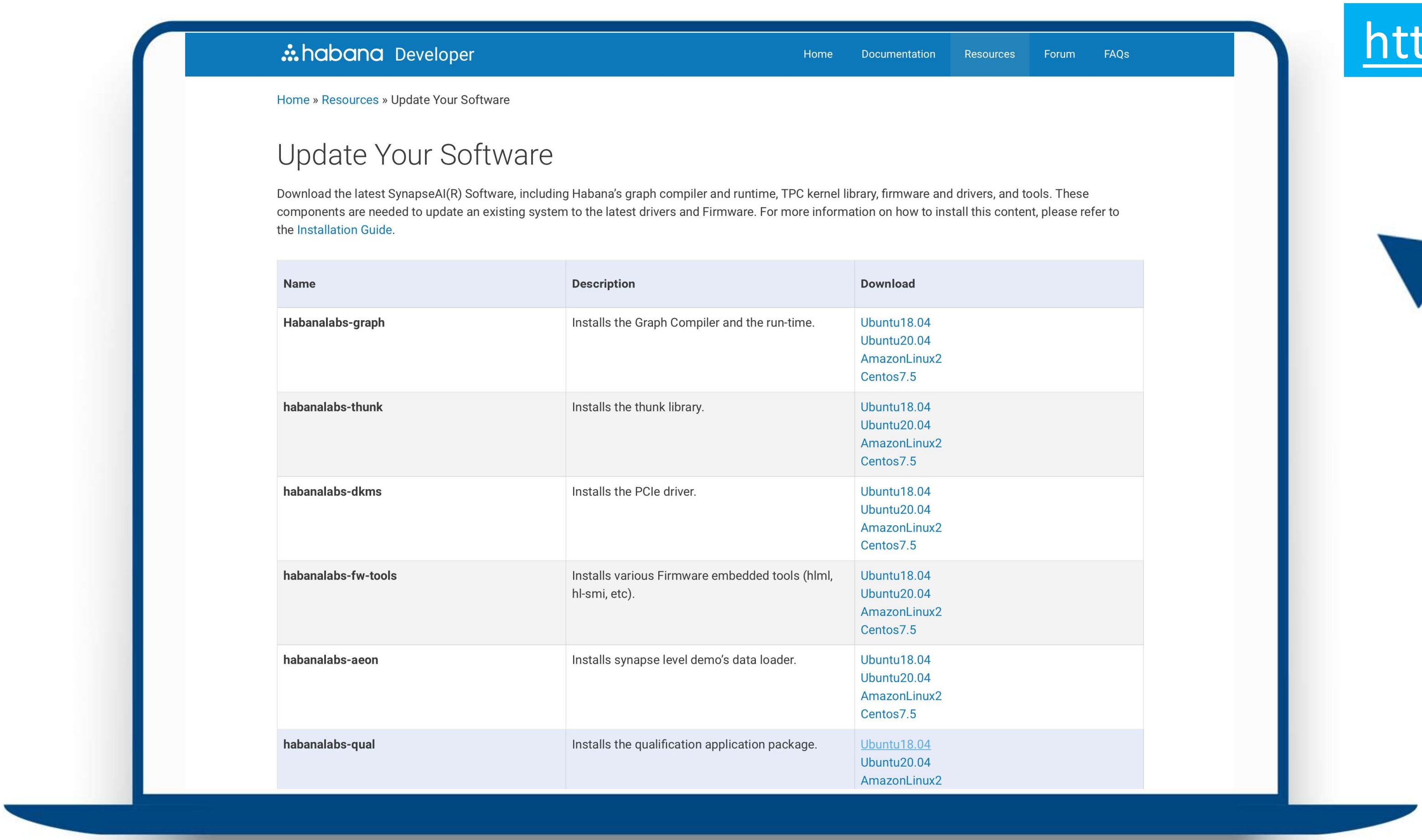
- TensorFlow Operators - Summarizes the SynapseAI TensorFlow supported operators for Gaudi.
- Pytorch Operators - Summarizes the SynapseAI Pytorch supported operators for Gaudi.

© Copyright 2021, Habana Labs Revision 8bbb8ba7.

developer.habana.ai | Legal Notice and Disclaimer | Send Feedback

Habana Developer Software Vault

https://vault.habana.ai



Home » Resources » Update Your Software

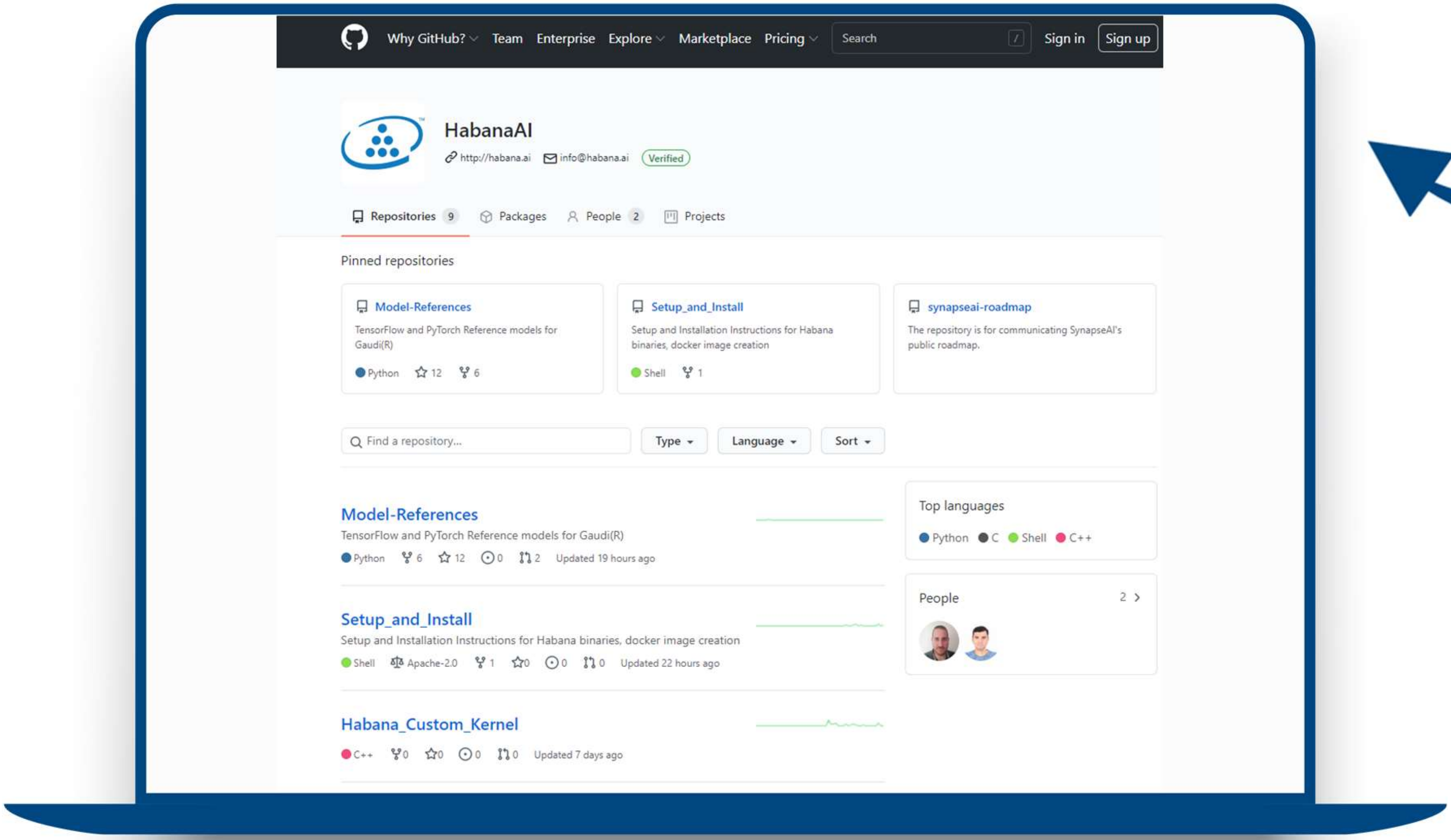
Update Your Software

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.

Name	Description	Download
Habanalabs-graph	Installs the Graph Compiler and the run-time.	Ubuntu18.04 Ubuntu20.04 AmazonLinux2 Centos7.5
habanalabs-thunk	Installs the thunk library.	Ubuntu18.04 Ubuntu20.04 AmazonLinux2 Centos7.5
habanalabs-dkms	Installs the PCIe driver.	Ubuntu18.04 Ubuntu20.04 AmazonLinux2 Centos7.5
habanalabs-fw-tools	Installs various Firmware embedded tools (hlml, hl-smi, etc).	Ubuntu18.04 Ubuntu20.04 AmazonLinux2 Centos7.5
habanalabs-aeon	Installs synapse level demo's data loader.	Ubuntu18.04 Ubuntu20.04 AmazonLinux2 Centos7.5
habanalabs-qual	Installs the qualification application package.	Ubuntu18.04 Ubuntu20.04 AmazonLinux2

Habana GitHub

https://github.com/HabanaAI



HabanaAI

http://habana.ai info@habana.ai Verified

Pinned repositories

Model-References
TensorFlow and PyTorch Reference models for Gaudi(R)
● Python ☆ 12 6

Setup_and_Install
Setup and Installation Instructions for Habana binaries, docker image creation
● Shell 1

synapseai-roadmap
The repository is for communicating SynapseAI's public roadmap.

Model-References
TensorFlow and PyTorch Reference models for Gaudi(R)
● Python 6 ☆ 12 0 2 Updated 19 hours ago

Setup_and_Install
Setup and Installation Instructions for Habana binaries, docker image creation
● Shell Apache-2.0 1 ☆0 0 0 Updated 22 hours ago

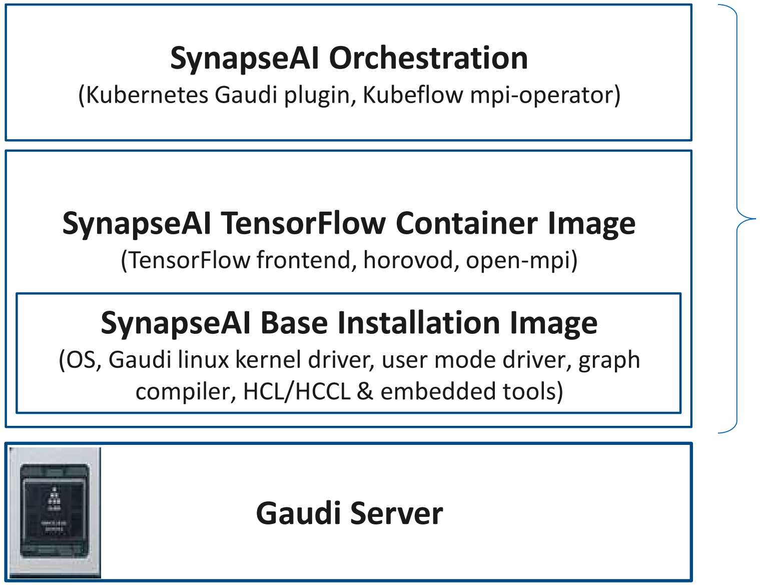
Habana_Custom_Kernel
● C++ 0 ☆0 0 0 Updated 7 days ago

Top languages
● Python ● C ● Shell ● C++

People 2

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/HCCL & embedded tools)

Gaudi Server

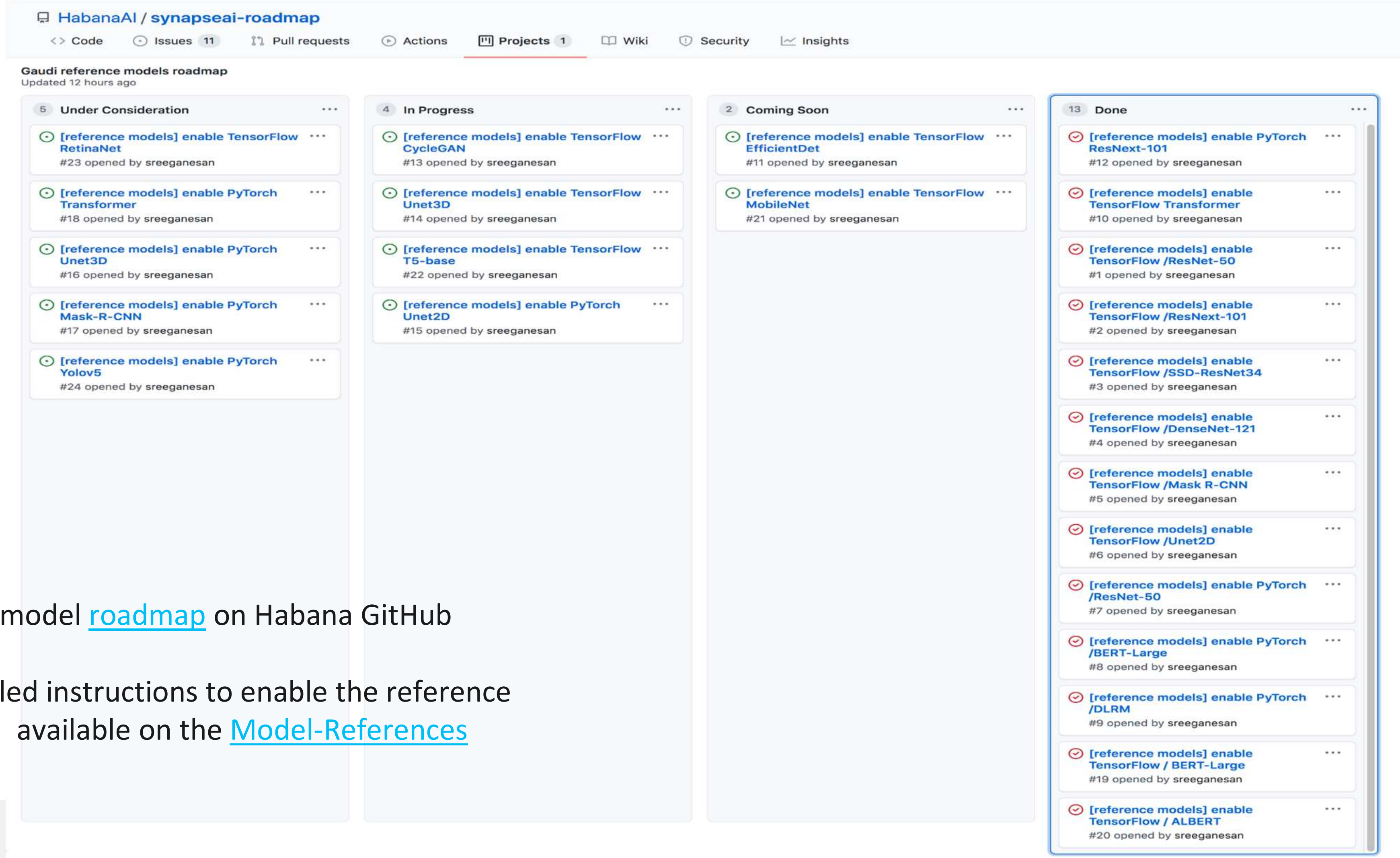
Gaudi-optimized Docker container images with all necessary dependences*

Official releases publicly available on Habana Vault

Orchestration	Kubernetes (1.19)
Frameworks	TensorFlow2 and PyTorch
Operating Systems	Ubuntu 18.04 and 20.04
Container Runtimes	Docker (Docker CE version 18.09)
Distributed Training Schemes	TensorFlow with Horovod and tf.distribute PyTorch distributed (native)

** Habana GitHub will have repository with Dockerfiles to "build your own" Docker images*

Gaudi Reference Models – July 2021



Gaudi reference models roadmap
Updated 12 hours ago

5 Under Consideration

- [reference models] enable TensorFlow RetinaNet
  #23 opened by sreeganesan
- [reference models] enable PyTorch Transformer
  #18 opened by sreeganesan
- [reference models] enable PyTorch Unet3D
  #16 opened by sreeganesan
- [reference models] enable PyTorch Mask-R-CNN
  #17 opened by sreeganesan
- [reference models] enable PyTorch Yolov5
  #24 opened by sreeganesan

4 In Progress

- [reference models] enable TensorFlow CycleGAN
  #13 opened by sreeganesan
- [reference models] enable TensorFlow Unet3D
  #14 opened by sreeganesan
- [reference models] enable TensorFlow T5-base
  #22 opened by sreeganesan
- [reference models] enable PyTorch Unet2D
  #15 opened by sreeganesan

2 Coming Soon

- [reference models] enable TensorFlow EfficientDet
  #11 opened by sreeganesan
- [reference models] enable TensorFlow MobileNet
  #21 opened by sreeganesan

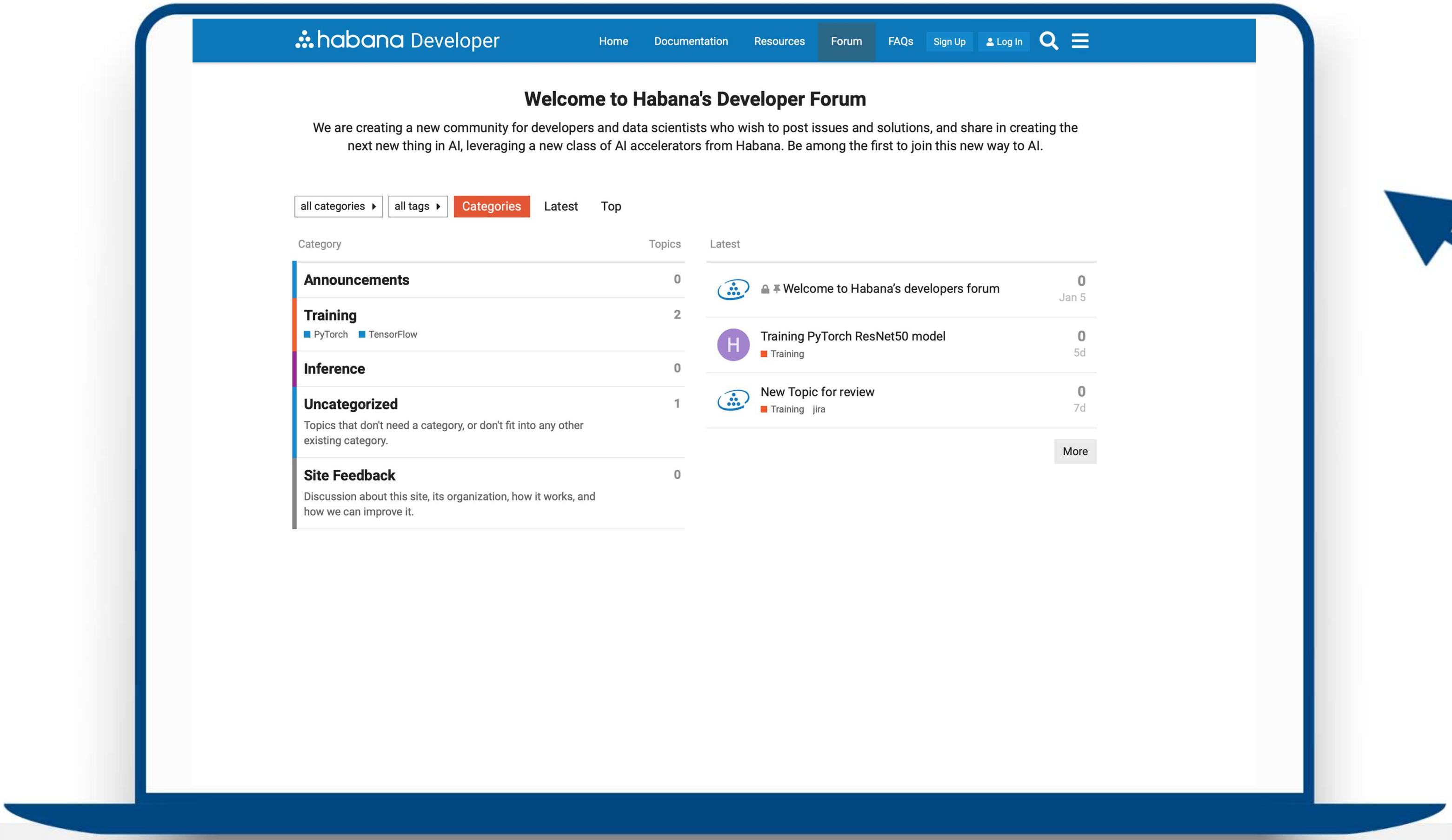
13 Done

- [reference models] enable PyTorch ResNext-101
  #12 opened by sreeganesan
- [reference models] enable TensorFlow Transformer
  #10 opened by sreeganesan
- [reference models] enable TensorFlow /ResNet-50
  #1 opened by sreeganesan
- [reference models] enable TensorFlow /ResNext-101
  #2 opened by sreeganesan
- [reference models] enable TensorFlow /SSD-ResNet34
  #3 opened by sreeganesan
- [reference models] enable TensorFlow /DenseNet-121
  #4 opened by sreeganesan
- [reference models] enable TensorFlow /Mask R-CNN
  #5 opened by sreeganesan
- [reference models] enable TensorFlow /Unet2D
  #6 opened by sreeganesan
- [reference models] enable PyTorch /ResNet-50
  #7 opened by sreeganesan
- [reference models] enable PyTorch /BERT-Large
  #8 opened by sreeganesan
- [reference models] enable PyTorch /DLRM
  #9 opened by sreeganesan
- [reference models] enable TensorFlow / BERT-Large
  #19 opened by sreeganesan
- [reference models] enable TensorFlow / ALBERT
  #20 opened by sreeganesan

- Gaudi reference model roadmap on Habana GitHub
- Scripts and detailed instructions to enable the reference models on Gaudi available on the Model-References repository

Habana’s Developer Forum

https://forum.habana.ai



Welcome to Habana's Developer Forum

We are creating a new community for developers and data scientists who wish to post issues 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.

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

Latest	
Welcome to Habana’s developers forum	0 Jan 5
Training PyTorch ResNet50 model Training	0 5d
New Topic for review Training jira	0 7d

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!

March 2021

THANK YOU

www.habana.ai

habana

An Intel Company