habana

An Intel Company

habana
An Intel Company

Sheik Mohamed Imran
AI – Technical Solutions Specialist
14 March 2022

Train More, Spend Less
with Habana® Gaudi-based Amazon EC2 DL1 Instances

Explosive demand for Deep Learning Training

- More Applications for AI
- More complex models
- Many Iterations
  - 74% of IDC respondents indicate running 5 – 10 iterations of training
  - >50% of respondents rebuild models weekly or more often; 26% rebuilding daily or hourly

Source: IDC Semiannual Artificial Intelligence Tracker (2020H1, published Jan 2021)

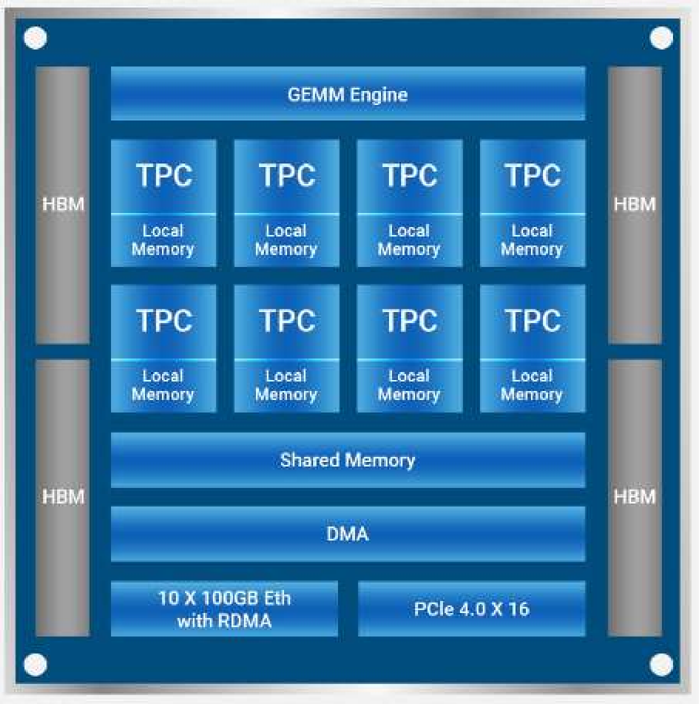
GAUDI™

Designed to advance AI compute efficiency

Gaudi: architected for efficiency

Designed to optimize AI performance, delivering higher efficiency than traditional CPUs & GPUs

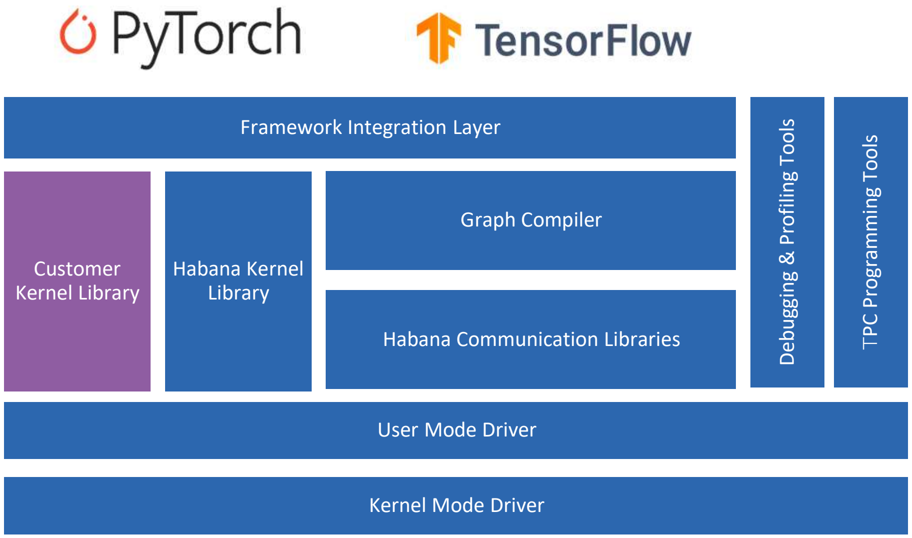
- Heterogeneous compute architecture
  - Configurable centralized GEMM engine (MME)
  - Fully programmable, AI-customized Tensor Processing Cores
- Software-managed memory architecture
  - 32 GB of HBM2 memory
- Natively integrated 10 x 100Gb Ethernet RoCE for scaling



SynapseAI® Software Suite: designed for performance and ease of use

Driving end-user efficiency for model build and migration

- Train deep learning models on Gaudi with minimal code changes
- Integrated with TensorFlow & PyTorch
- Habana Developer Site & GitHub
- Support with reference models, kernel libraries, documentation and “how tos”
- Advanced users can write their own custom kernels



PyTorch

TensorFlow

Framework Integration Layer

Customer Kernel Library

Habana Kernel Library

Graph Compiler

Habana Communication Libraries

Debugging & Profiling Tools

TPC Programming Tools

User Mode Driver

Kernel Mode Driver

DL1 instances powered by Gaudi processors features

AWS Custom
2nd Gen Xeon Scalable Processors

Up to 8 Habana Gaudi
accelerators with 32GB
HBM per processor

400Gbps Networking & 4TB
of NVMe Storage

All-to-all 100Gbps
interconnect

SynapseAI SDK integrated
with TensorFlow and PyTorch

Support for developing
custom kernels

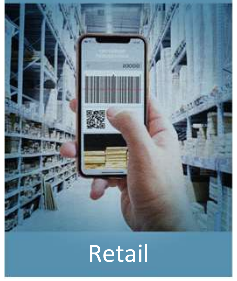
Use cases

Object Detection & Segmentation

Defect detection

Manufacturing

Fraud detection & inventory management



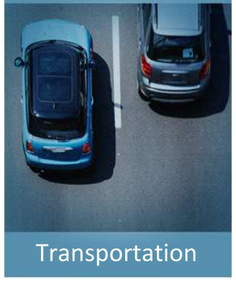
Retail

2D/3D Scanning & medical imaging

Medical

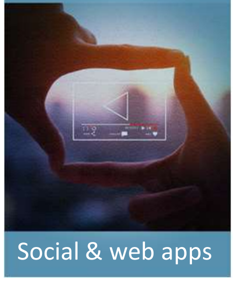
Image Classification

Autonomous Vehicle segmentation



Transportation

Photo & video identification



Social & web apps

Natural Language Processing

Subject matter query

Question/ answer

Sentiment analysis

Visualize Performance and Build Custom Kernels

Habana Profiling Tools

Monitor core utilization, enabling performance analysis and optimization

Habana Kernel Library

Rich TPC kernel library with support for wide variety of operators such as non-linear, elementwise, non-GEMM

TPC Programming Tools

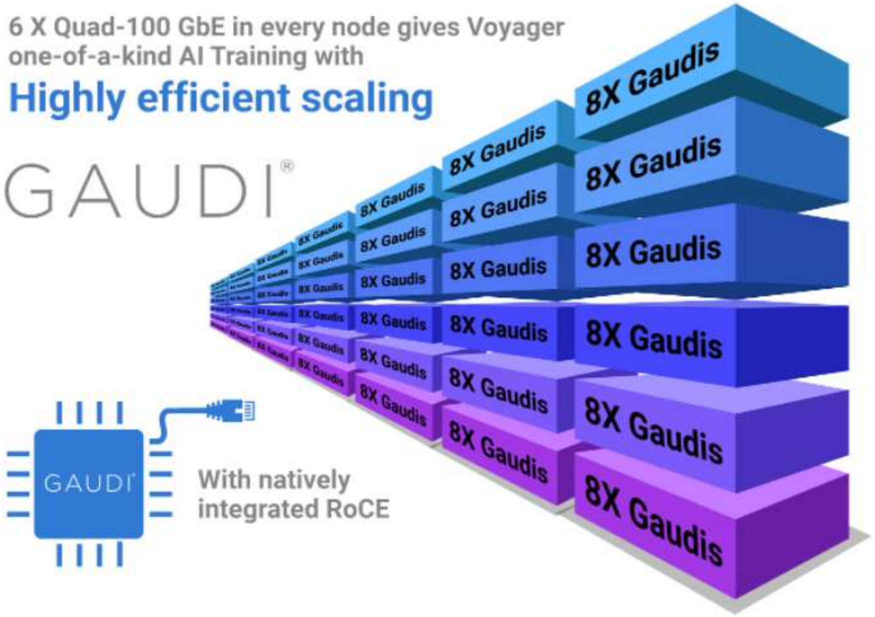
Build custom kernels using LLVM-based TPC-C compiler, simulator, and debugger

Habana Communication Libraries

Scale up to multiple Gaudi cards within a node or scale out across nodes for distributed training

Gaudi is also driving efficiencies in HPC

SDSC Voyager Supercomputer powered by 336 Gaudi training processors


6 X Quad-100 GbE in every node gives Voyager one-of-a-kind AI Training with
Highly efficient scaling
GAUDI®
8X Gaudis
GAUDI®
With natively integrated RoCE

Voyager goes into service this fall

Supermicro X12 8-Gaudi Server powering Voyager

Funded by the National Science Foundation

AI research conducted across range of science and engineering domains

habana
An Intel Company