



Inference Card

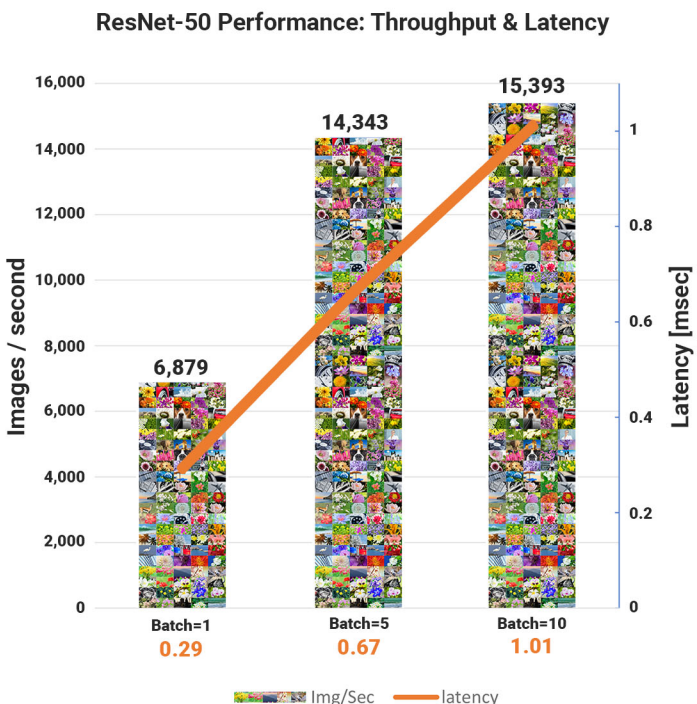
Habana Labs Goya™ is the industry's first commercially available deep learning inference processor product-line designed specifically to deliver superior performance, power efficiency and cost savings.

Habana Labs HL-10x PCIe cards incorporate a single Goya HL-1000 Processor and are designed to accelerate various AI inferencing workloads, such as image recognition, neural machine translation, sentiment analysis, recommender systems and many others.

Goya HL-1000 processor is a ground up design for neural network processing that incorporates a fully programmable TPC™ (Tensor Processing Core) along with its SynapseAI™ software stack.



SPECIFICATIONS



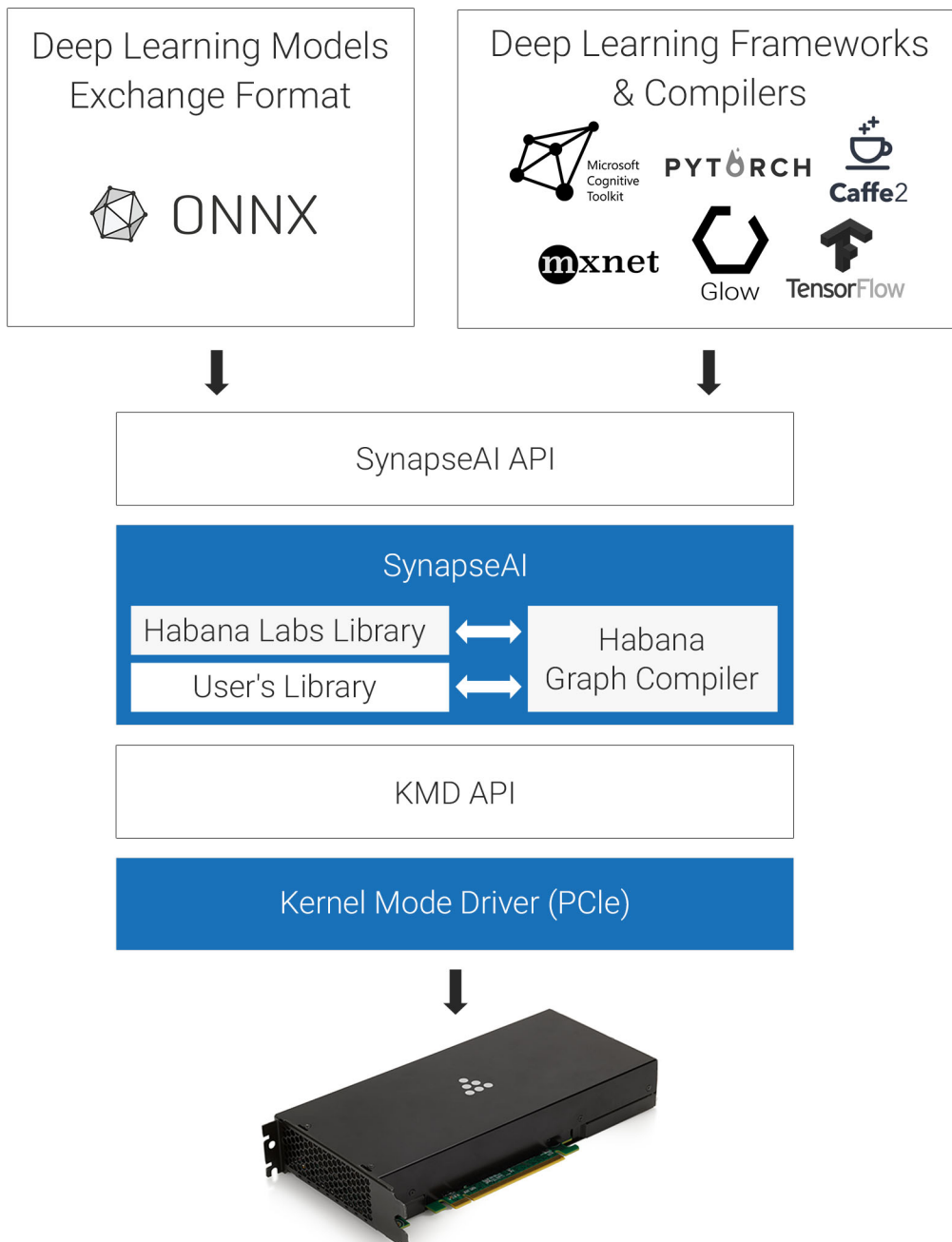
PROCESSOR TECHNOLOGY	Goya HL-1000
SYSTEM INTERFACE	PCIe Gen 4.0 x 16 8-pin auxillary power connector full length, full height
FORM FACTORS & PRODUCT SKUS	HL-100: dual-slot, passive cooling HL-102: dual-slot, active cooling
MEMORY	4/8/16GB DDR4 memory ECC protected
THERMAL DESIGN POWER	TDP - 200W
SCENARIO POWER & EFFICIENCY	ResNet-50: 15,393 images/second @103W (=149 images/sec/watt)

Fast Deployment with SynapseAI:

Habana Labs' SynapseAI is a comprehensive inference software toolkit that simplifies the development and deployment of deep learning models. SynapseAI provides inference network model compilation and runtime, eliminating the need of low level programming.

The Goya platform is training-platform-agnostic: the deep neural network can be trained on any hardware platform. The SynapseAI then compiles the trained model for use on the Goya processor and the result is an optimized execution code in terms of accuracy, latency, throughput and efficiency for inferencing usage.

Goya Software Stack



© 2019 Habana Labs Ltd. All rights reserved. Habana Labs, Habana, the Habana Labs logo, Goya, Gaudi, Pure AI, TPC and SynapseAI are trademarks or registered trademarks of Habana Labs Ltd. All other trademarks or registered trademarks and copyrights are the property of their respective owners.