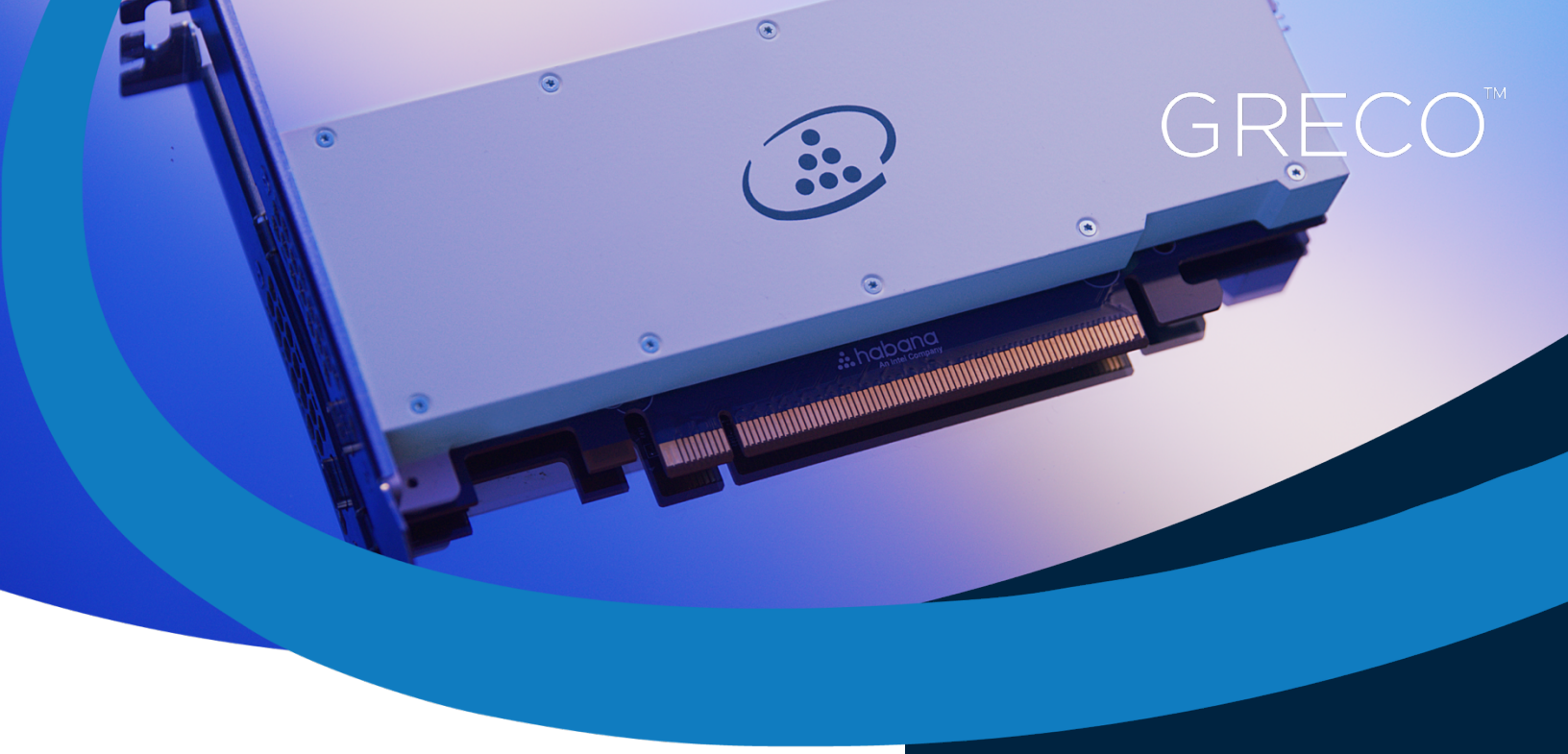


GRECO™

Habana introduces its second-generation AI inference processor, Greco™, built on the company's first-gen programmable architecture with performance and efficiency advancements.

Greco is designed to optimize data center inference deployments — in the cloud and on-premises — with compute density and flexibility, higher performance, and lower power.

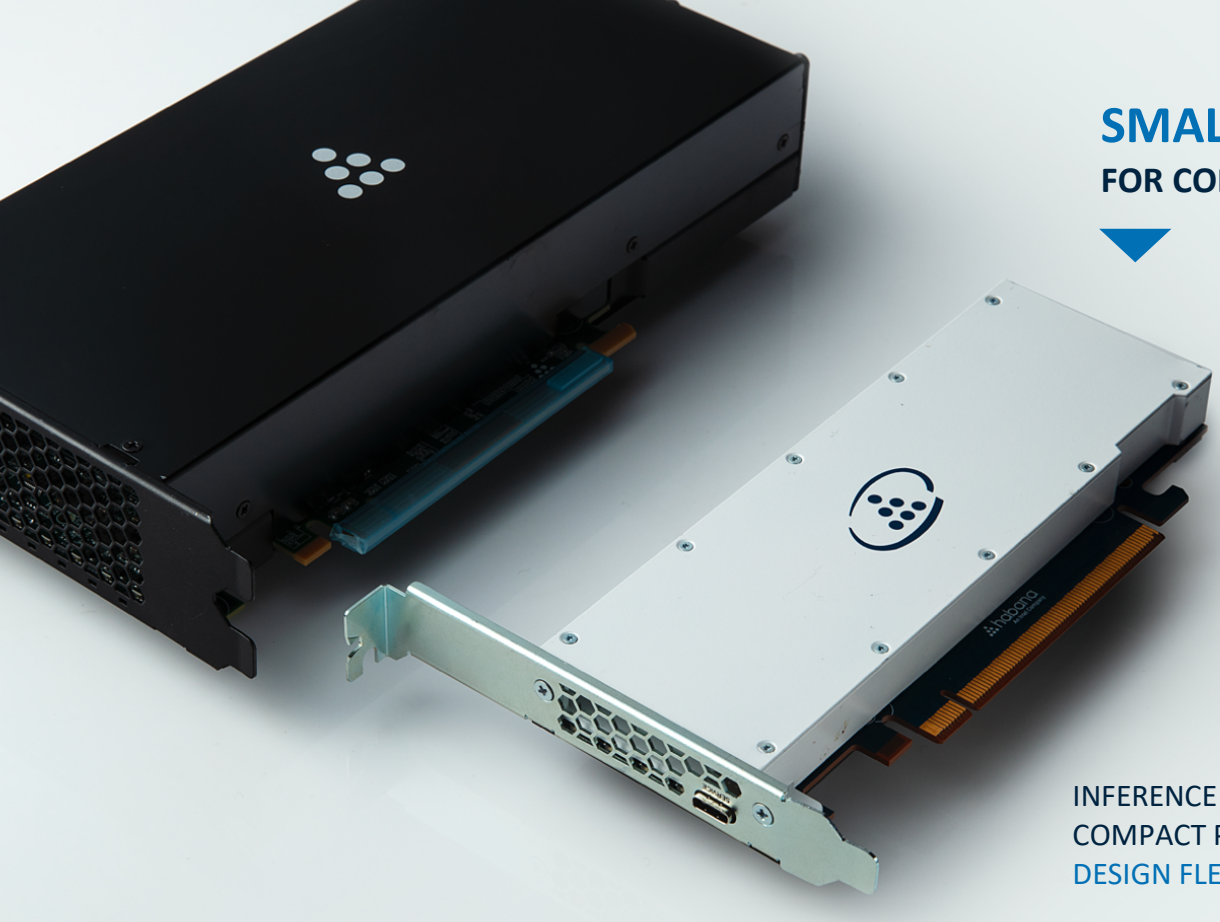
To boost inference speed and efficiency, Greco integrates media decoding, encoding and post-processing on-chip, supporting media formats HEVC, H.264, JPEG and P-JPEG. In addition, Greco supports additional data types, Bfloat16, FP16, and 4-bit Integer, giving customers flexibility in balancing inference speed and accuracy. Greco's memory bandwidth has been increased by 5x over first-generation Goya™ with 16GB LPDDR5 memories and on-chip SRAM has increased from 50 to 128 MB, helping to boost inference performance across all model types.

Greco is offered in small form factor, reduced from the Goya dual-slot PCIe card to single-slot, half-height, half-length (HHHL) PCIe Gen 4 x8 interface, packing the performance into the compact HHHL to deliver improved inference compute density, allowing customers to double the number of cards in the server and reducing the total cost of ownership for high-density inferencing deployments. This compact form enables customers to drop in and replace GPU cards with Greco in existing servers.

Greco targets a wide variety of demanding, high-compute inference workloads for a diverse set of use cases and applications with its low-profile, 75W design, which easily fits into standard PCIe data center servers.

HABANA® GRECO™ INFERENCE PROCESSOR

Form Factors	PCIe Single Slot - HHHL
Thermal Solution	Passive
TDP (WATT)	75
System Interface	PCIe 4x8
Memory	4x64 bit, LPDDR5 16GB
Data Types	INT16, INT8, INT4 FP32, FP16, BF16
Media	Integrated Media: Decoding, encoding and post processing
P/N	HL-110



SMALLER PROFILE
FOR COMPUTE DENSITY

INFERENCE PERFORMANCE IN A COMPACT PACKAGE:
DESIGN FLEXIBILITY & EFFICIENCY

Fast Deployment with SynapseAI® Software Suite

Habana Labs' SynapseAI Software Suite enables efficient mapping of neural network topologies onto Habana's Greco inference processors. Designed to facilitate ease of use and high-performance inference, the software suite includes Habana's graph compiler and runtime, performance-optimized TPC kernel library, firmware and drivers, and developer tools such as the TPC programming tool kit for custom kernel development and SynapseAI Profiler.

The Habana Developer Site is the hub for Habana developers where they can find a wealth of information to get started, including tutorials, reference models, how-to guides, documentation, and so on. It also hosts a Forum for the Habana developer community.

TPC

Polymorphic MME

Sparsity

Media

habana
An Intel Company