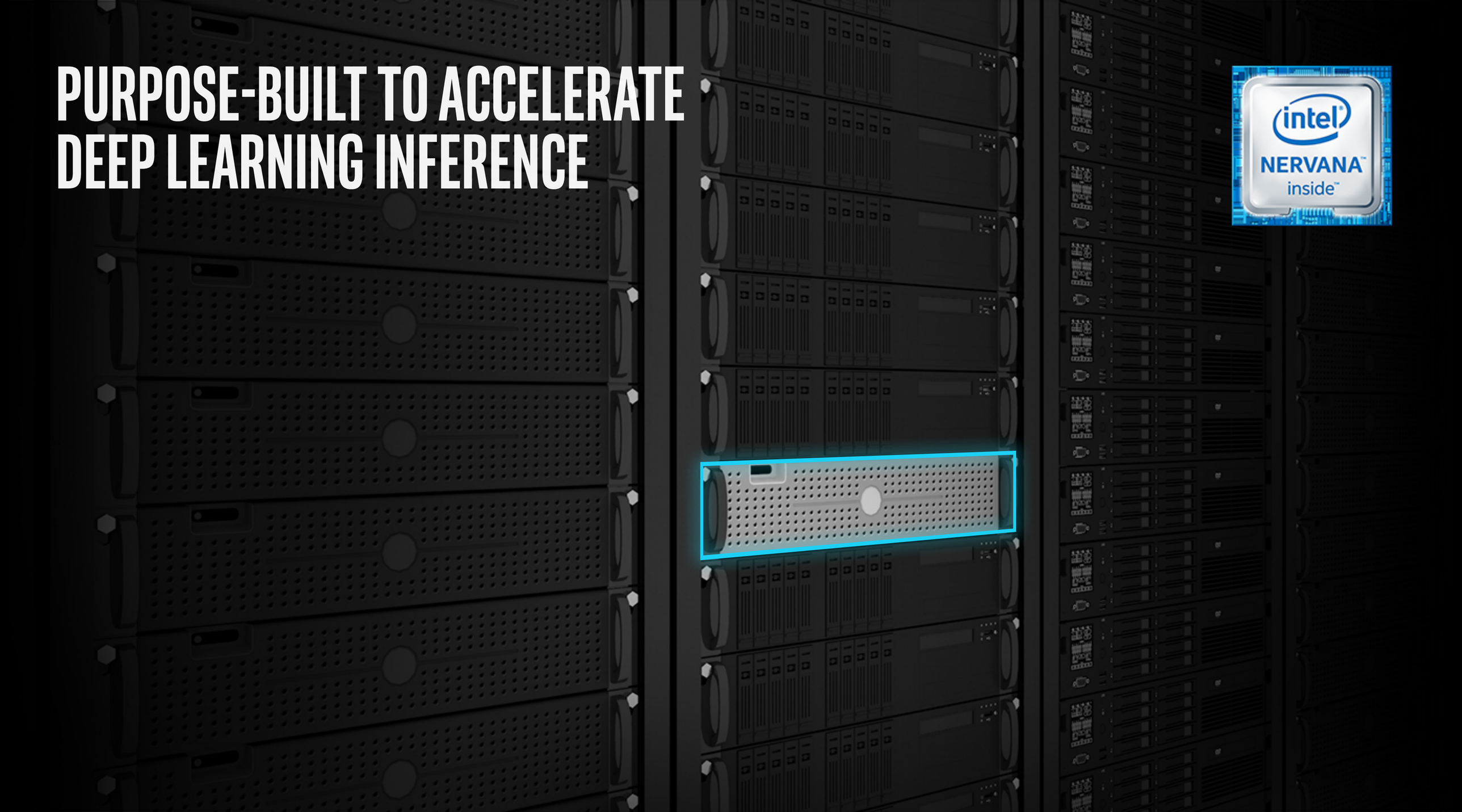
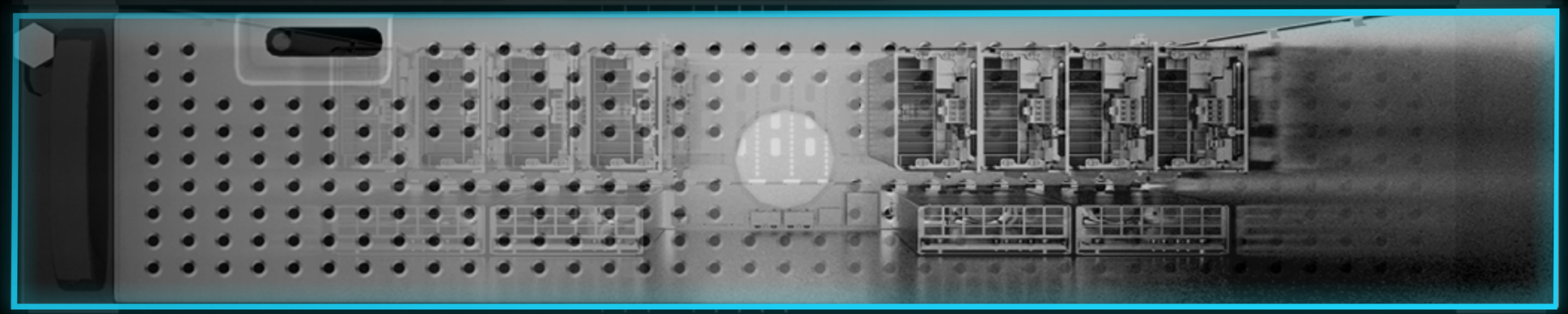
PURPOSE-BUILT TO ACCELERATE DEEP LEARNING INFERENCE



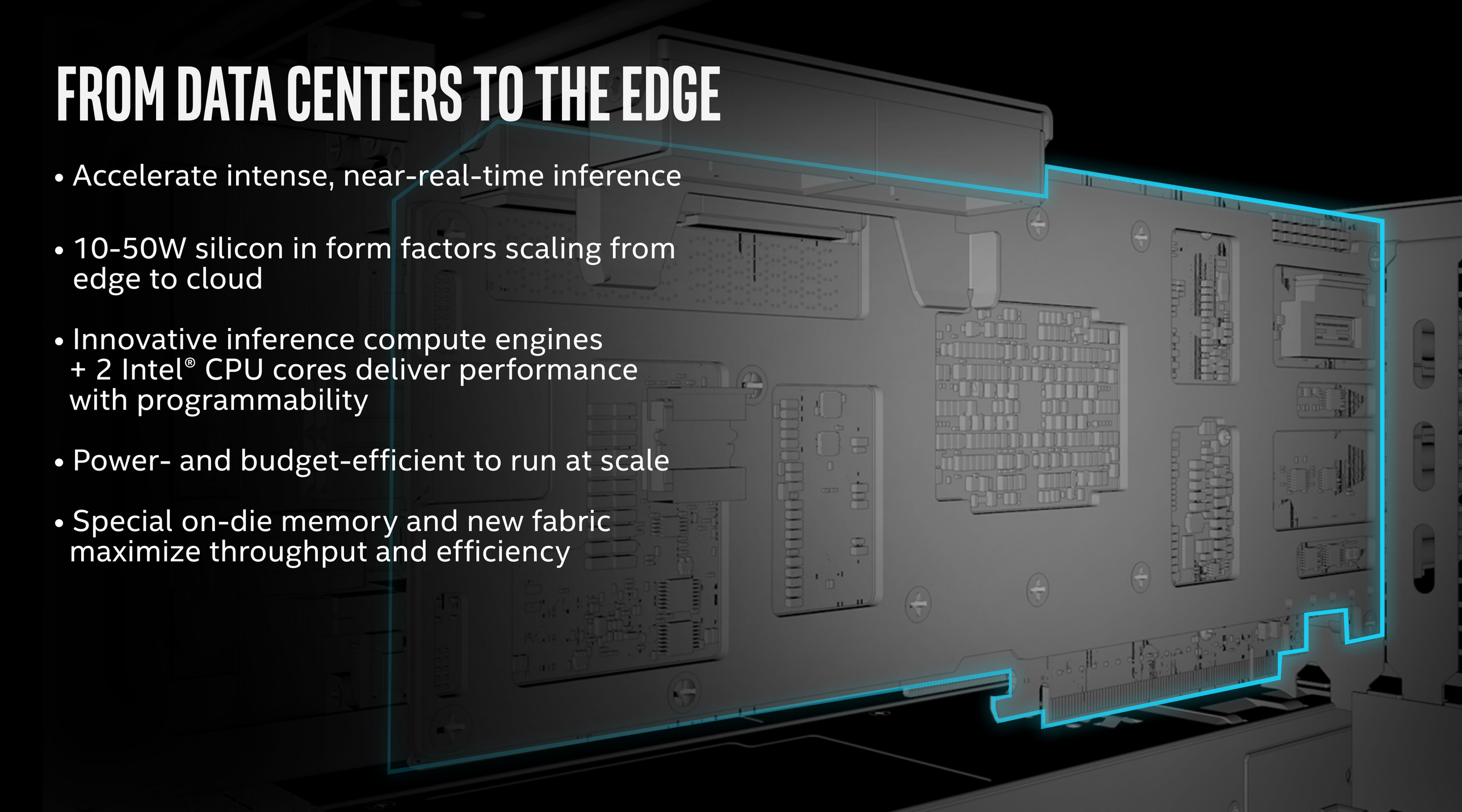
INTEL® NERVANA™ NEURAL NETWORK PROCESSOR FOR INFERENCE

(Intel® Nervana™ NNP-I)



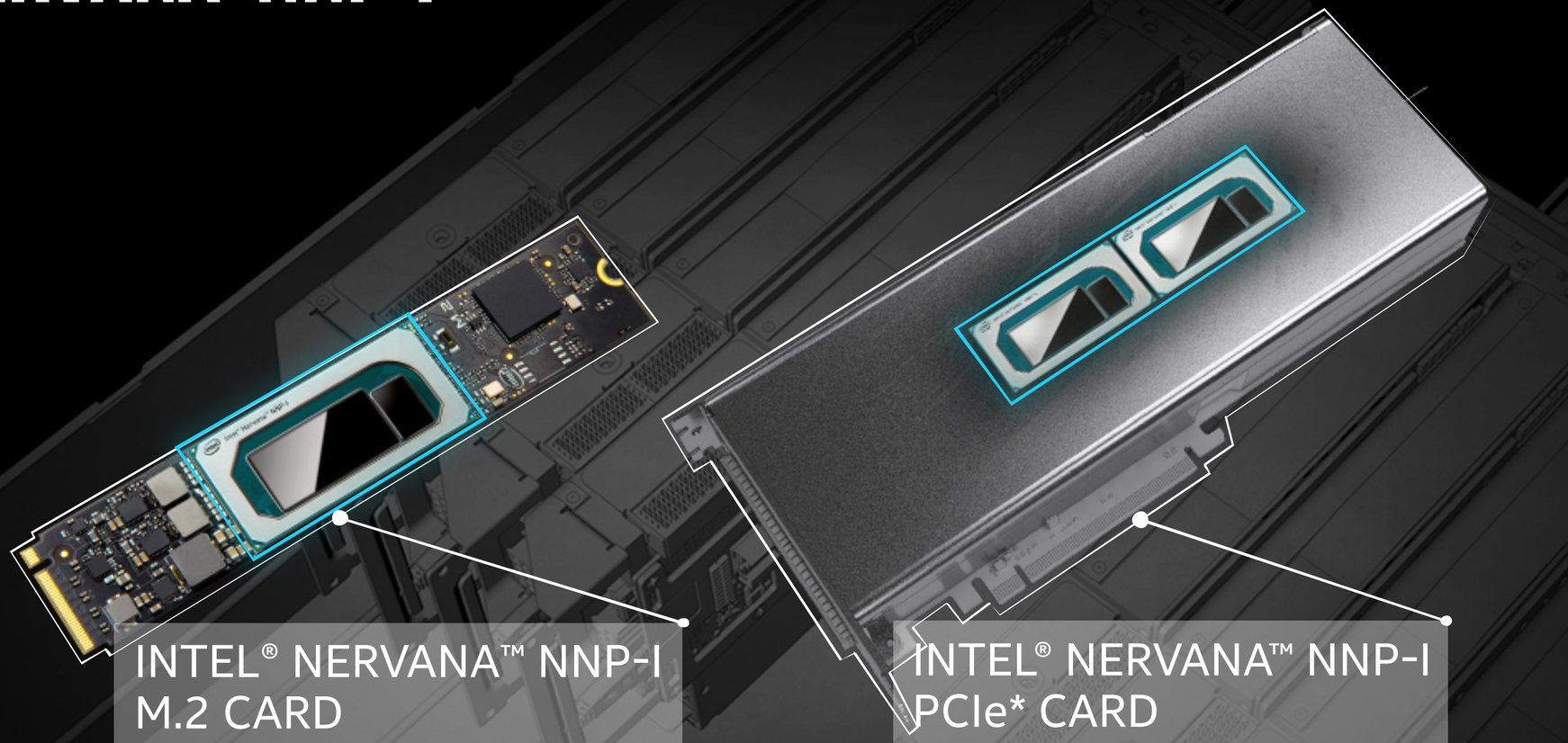
FROM DATA CENTERS TO THE EDGE

- Accelerate intense, near-real-time inference
- 10-50W silicon in form factors scaling from edge to cloud
- Innovative inference compute engines + 2 Intel® CPU cores deliver performance with programmability
- Power- and budget-efficient to run at scale
- Special on-die memory and new fabric maximize throughput and efficiency




FLEXIBLE REAL-WORLD DEPLOYMENT AT SCALE

INTEL® NERVANA™ NNP-I



INTEL® NERVANA™ NNP-I M.2 CARD

- 12W card with 1x Intel® Nervana™ NNP-I
- Up to 50 TOPS

INTEL® NERVANA™ NNP-I PCIe* CARD

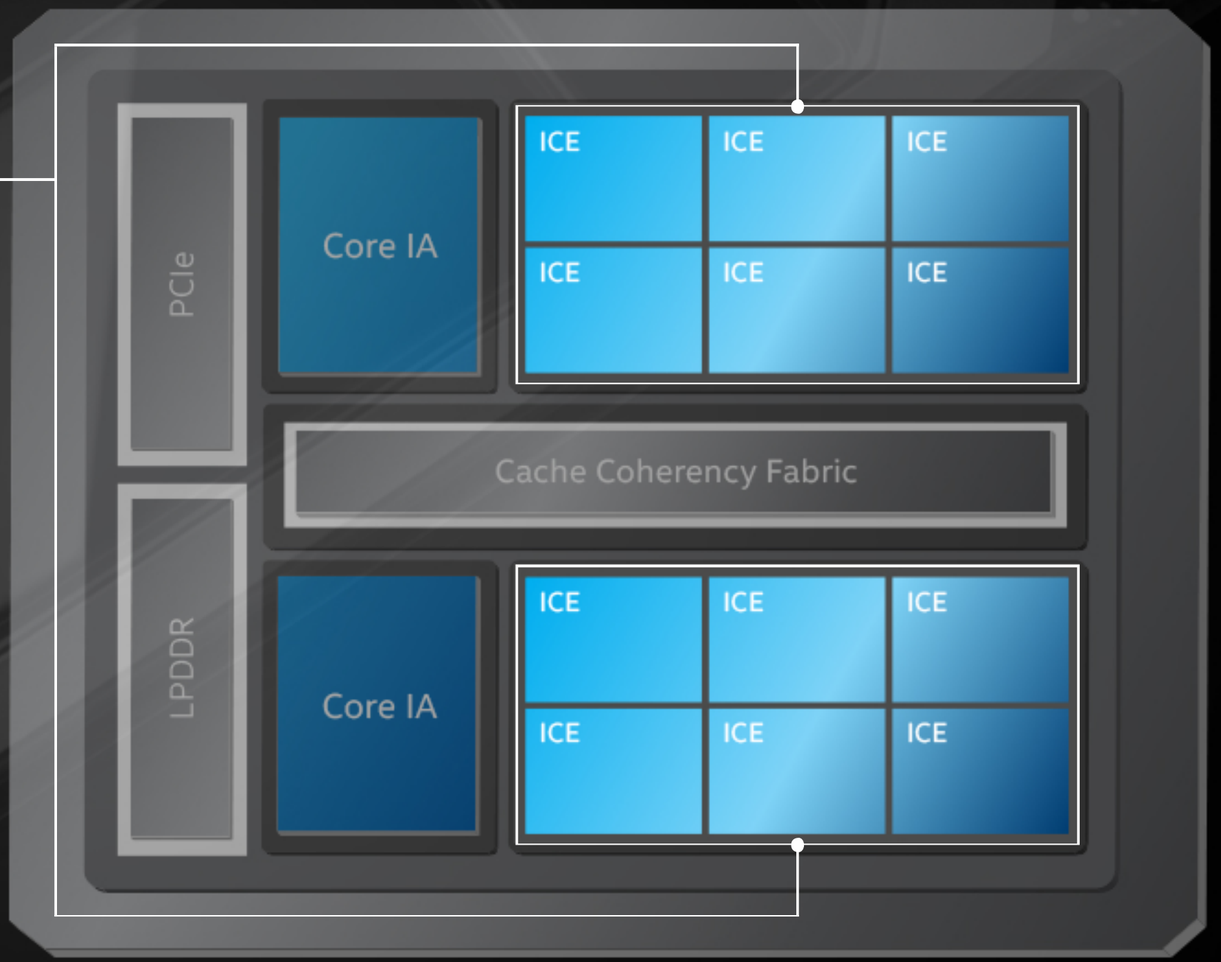
- 75W card with 2x Intel® Nervana™ NNP-I
- Up to 170 TOPS

INTEL® NERVANA™ NNP-I

Highly programmable, performant, and efficient

12 INFERENCE COMPUTE ENGINES (ICE) + 2 INTEL® CPU CORES

Novel combination of hardware flexibility and high-throughput, low-latency performance

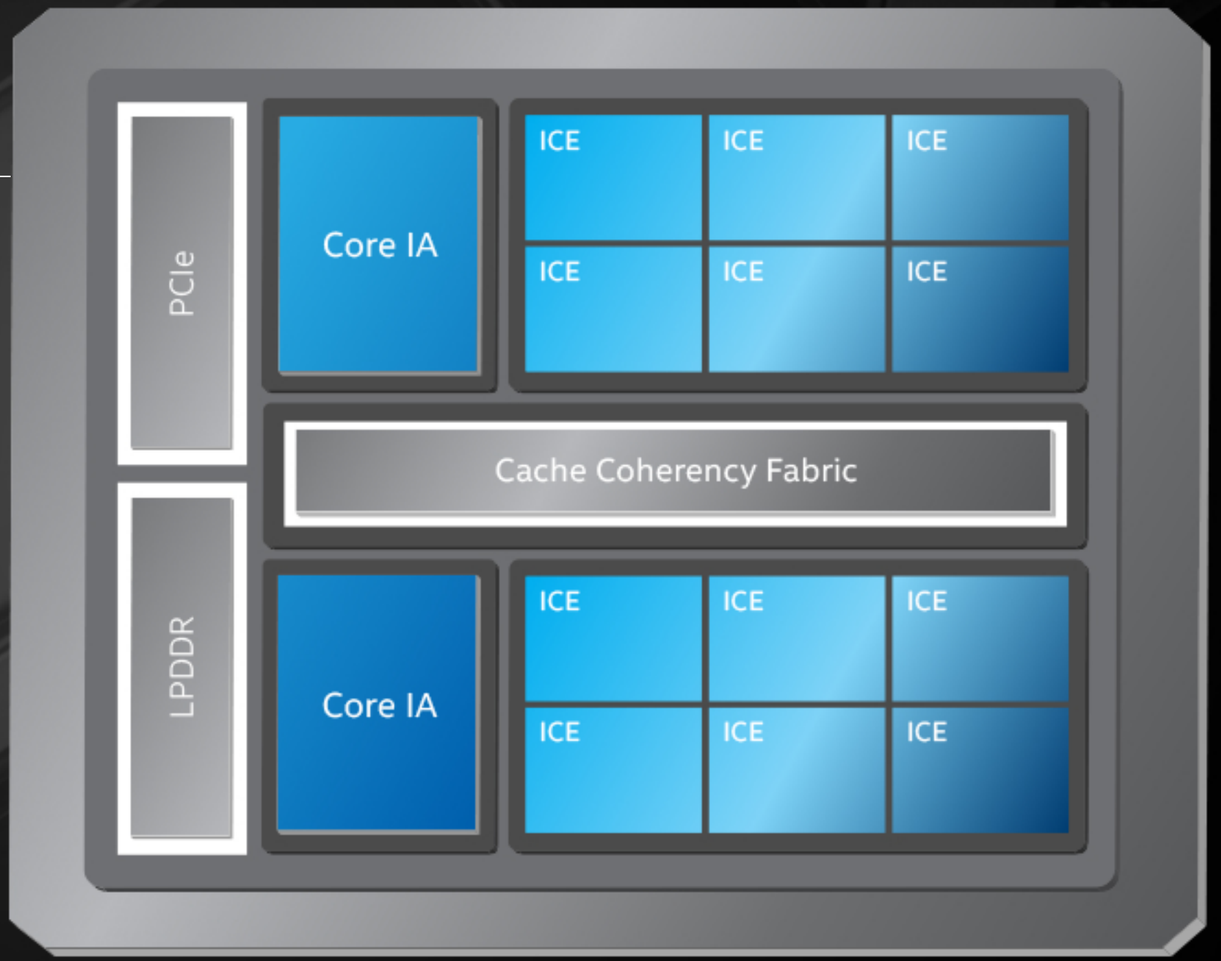


INTEL® NERVANA™ NNP-I

Highly programmable, performant, and efficient

DYNAMIC POWER MANAGEMENT

Fully integrated voltage regulator (FIVR) technology optimizes SoC performance at different power envelopes

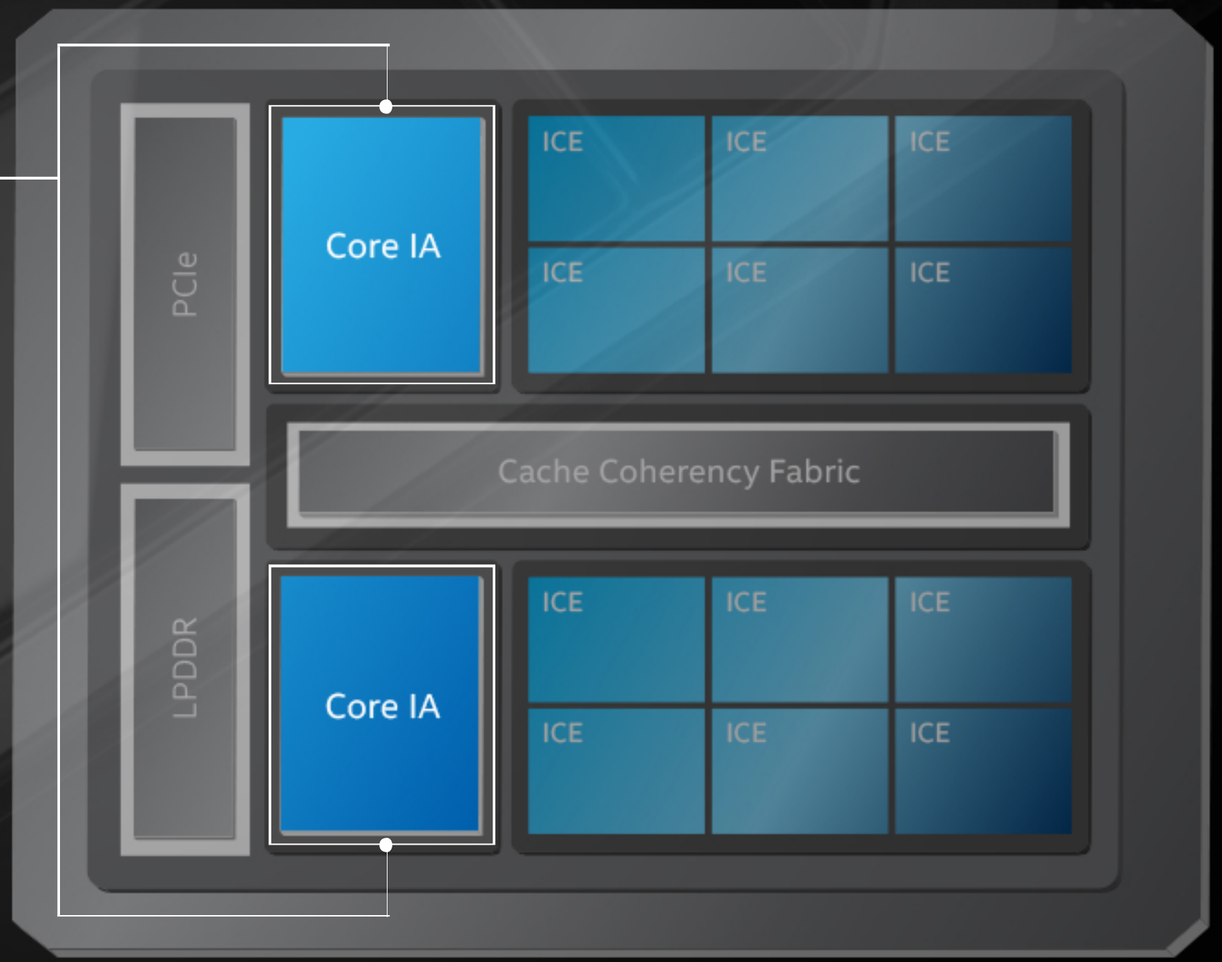


INTEL® NERVANA™ NNP-I

Highly programmable, performant, and efficient

ON-DIE INTEL® ARCHITECTURE CORES

Programmability with Intel® Advanced Vector Extensions (Intel® AVX) and Vector Neural Network Instructions (VNNI)

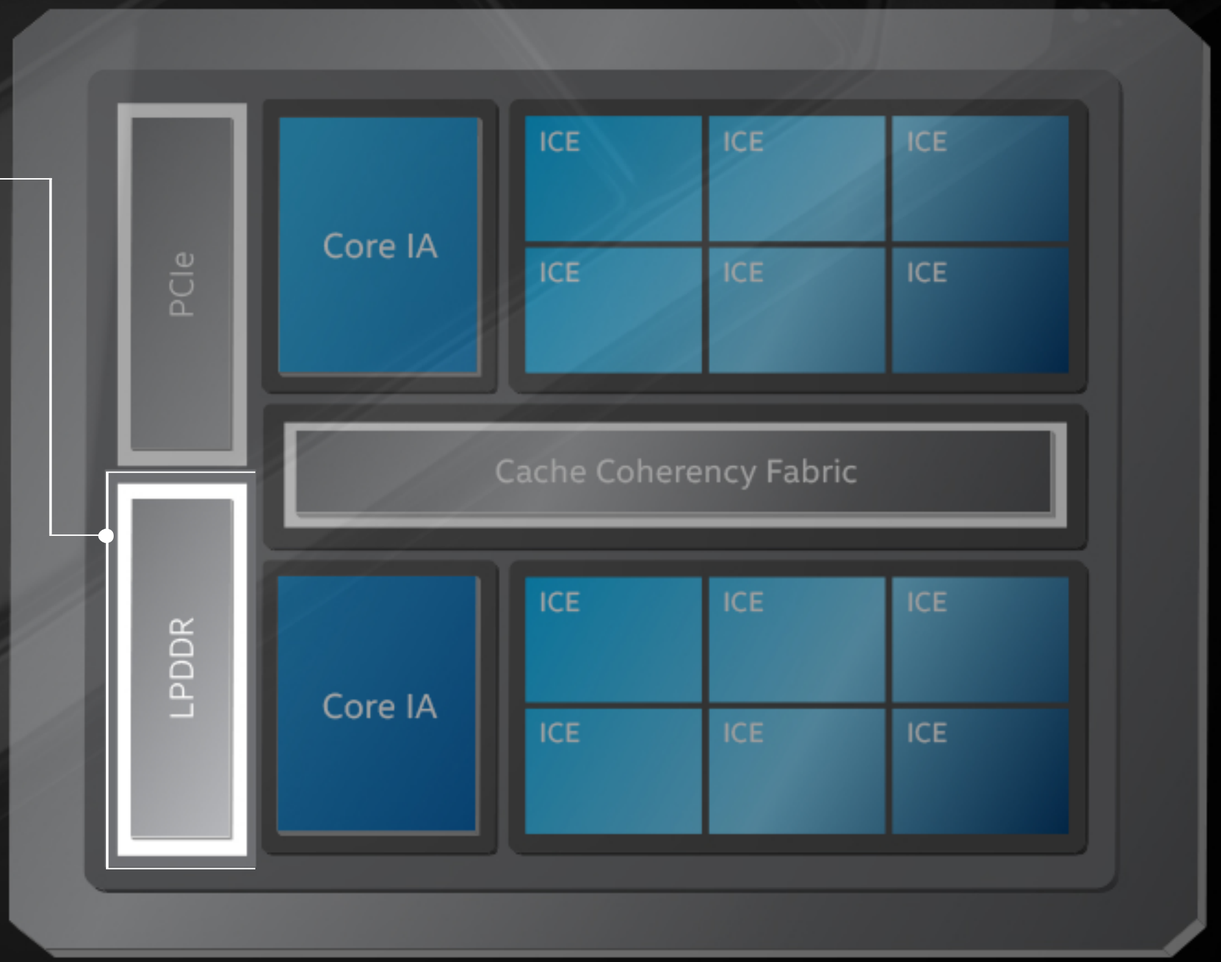


INTEL® NERVANA™ NNP-I

Highly programmable, performant, and efficient

75 MB SRAM

On-die SRAM and fabric deliver high performance for deep learning models

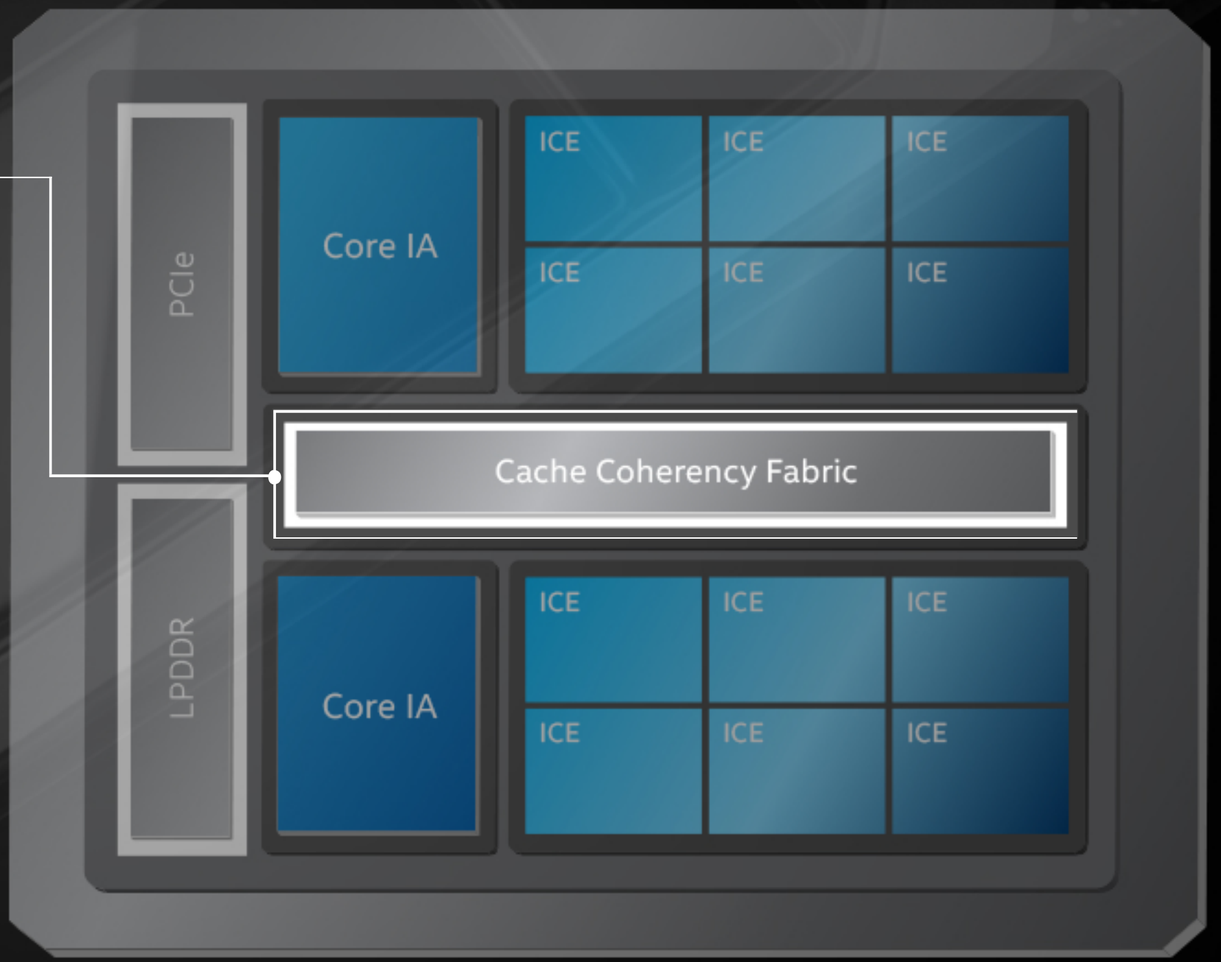


INTEL® NERVANA™ NNP-I

Highly programmable, performant, and efficient

CACHE COHERENCY FABRIC

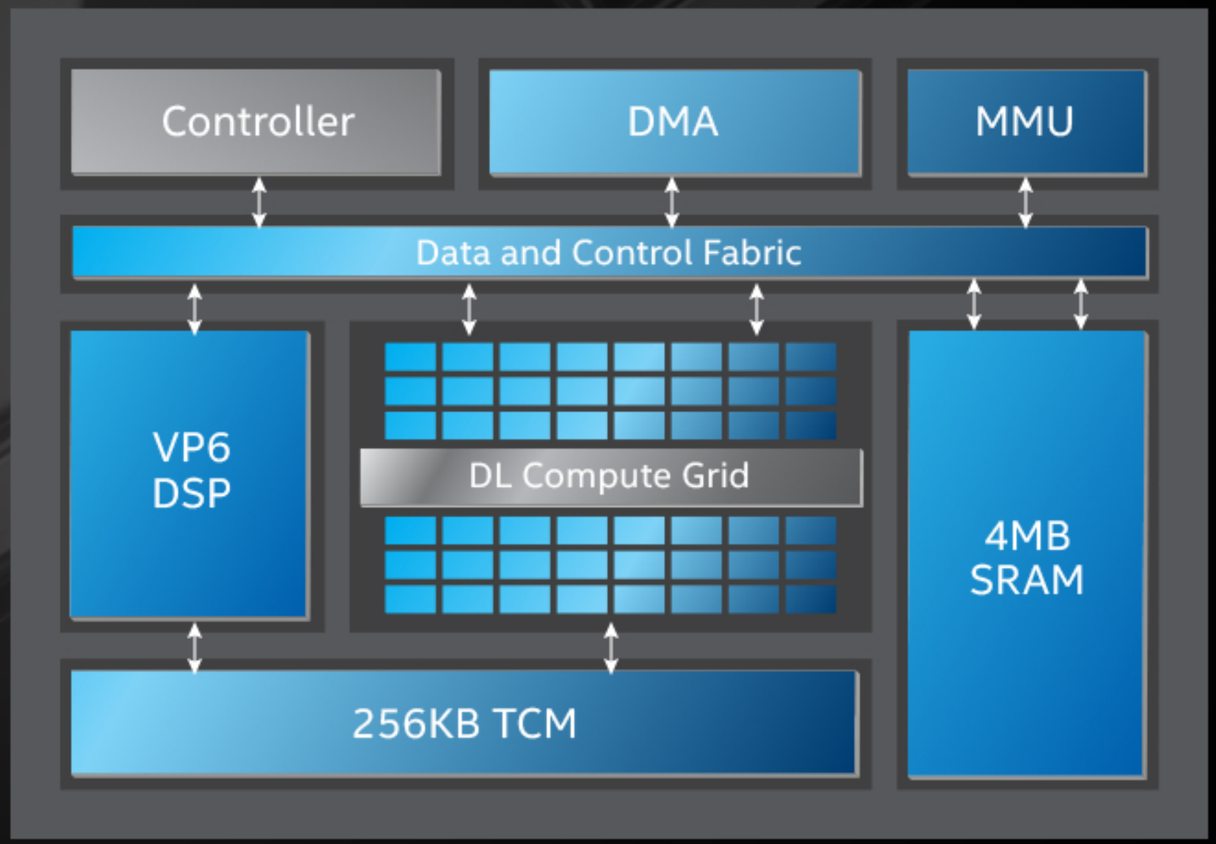
24 MB hardware-managed, high-performance shared cache



ICE

Inference compute engines (ICE) optimized
for maximum performance and efficiency

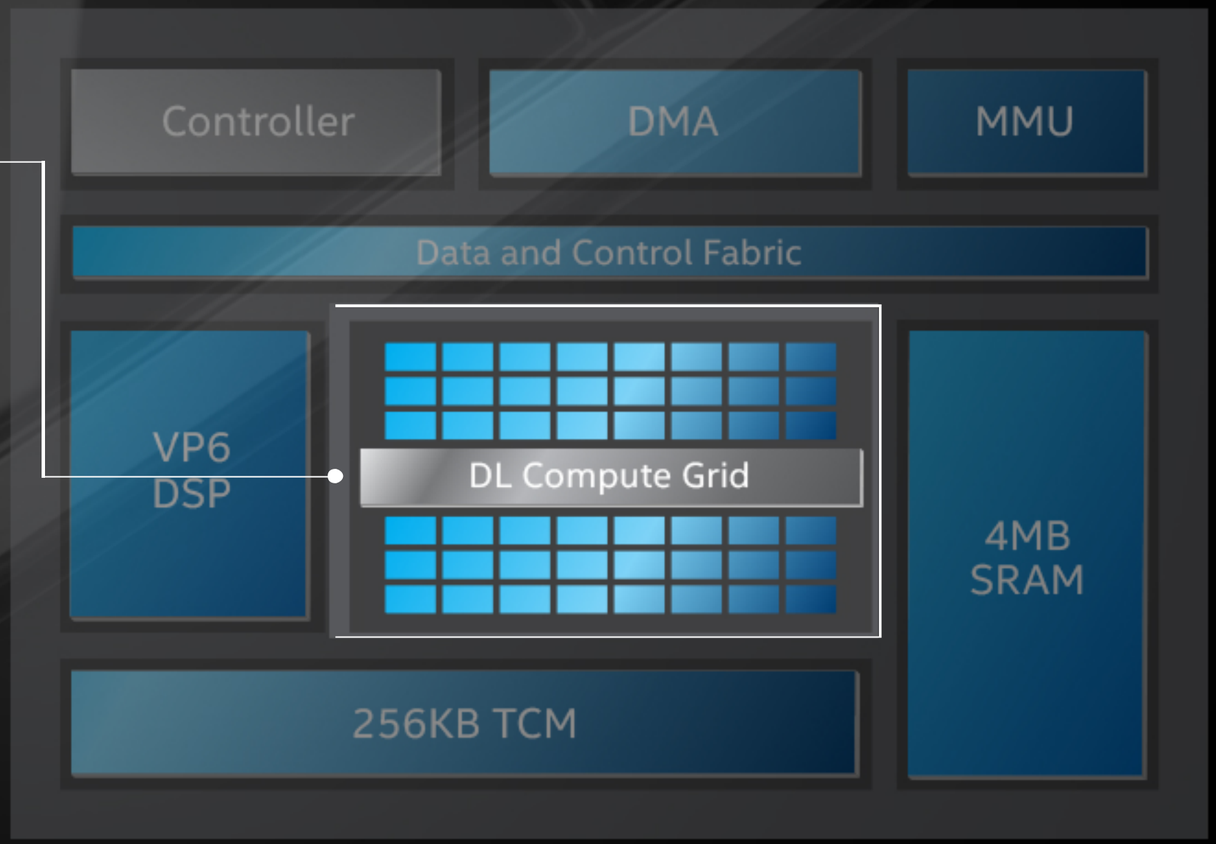
- DEEP LEARNING COMPUTE GRID
- PROGRAMMABLE VECTOR PROCESSOR (DSP)
- LARGE LOCAL SRAM
- HIGH-BANDWIDTH DATA MEMORY ACCESS



ICE

Inference compute engines (ICE) optimized
for maximum performance and efficiency

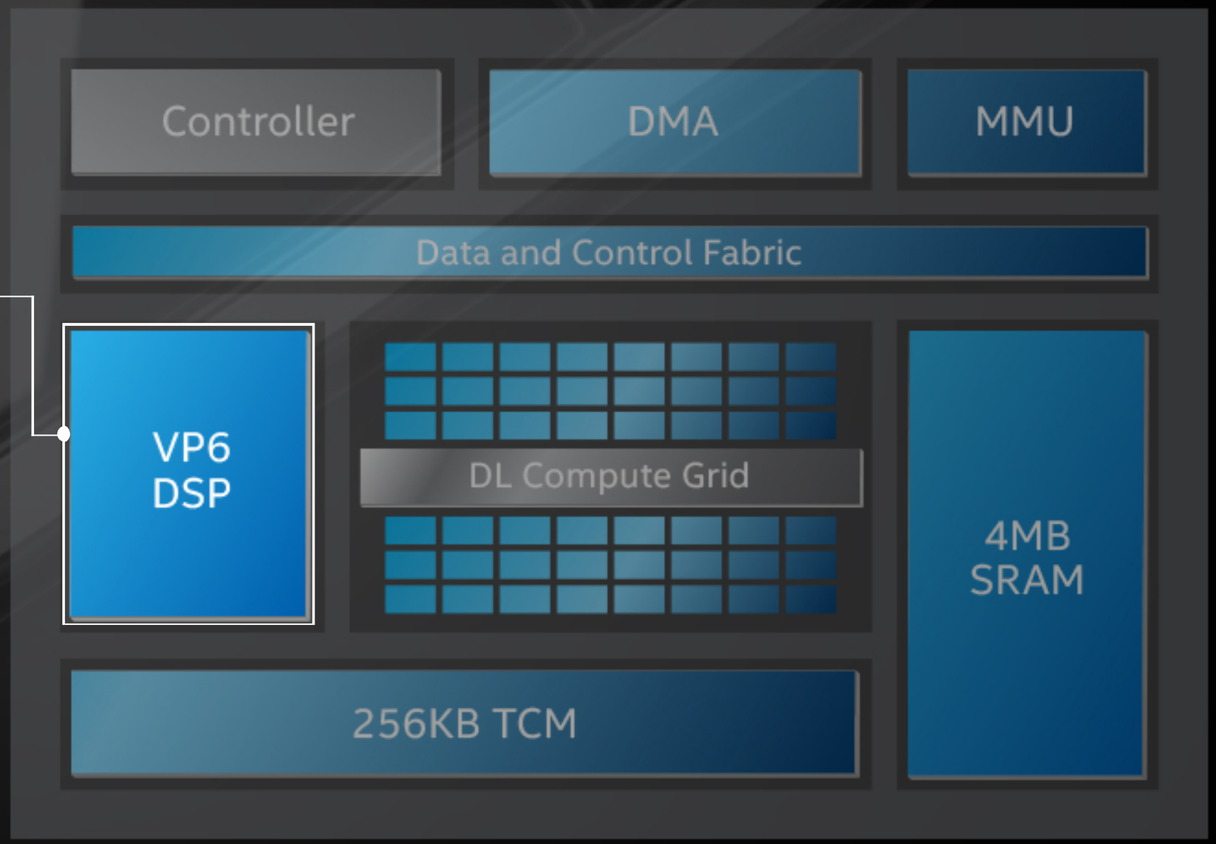
- DEEP LEARNING COMPUTE GRID
- PROGRAMMABLE VECTOR PROCESSOR (DSP)
- LARGE LOCAL SRAM
- HIGH-BANDWIDTH DATA MEMORY ACCESS



ICE

Inference compute engines (ICE) optimized
for maximum performance and efficiency

- DEEP LEARNING COMPUTE GRID
- PROGRAMMABLE VECTOR PROCESSOR (DSP)
- LARGE LOCAL SRAM
- HIGH-BANDWIDTH DATA MEMORY ACCESS



ICE

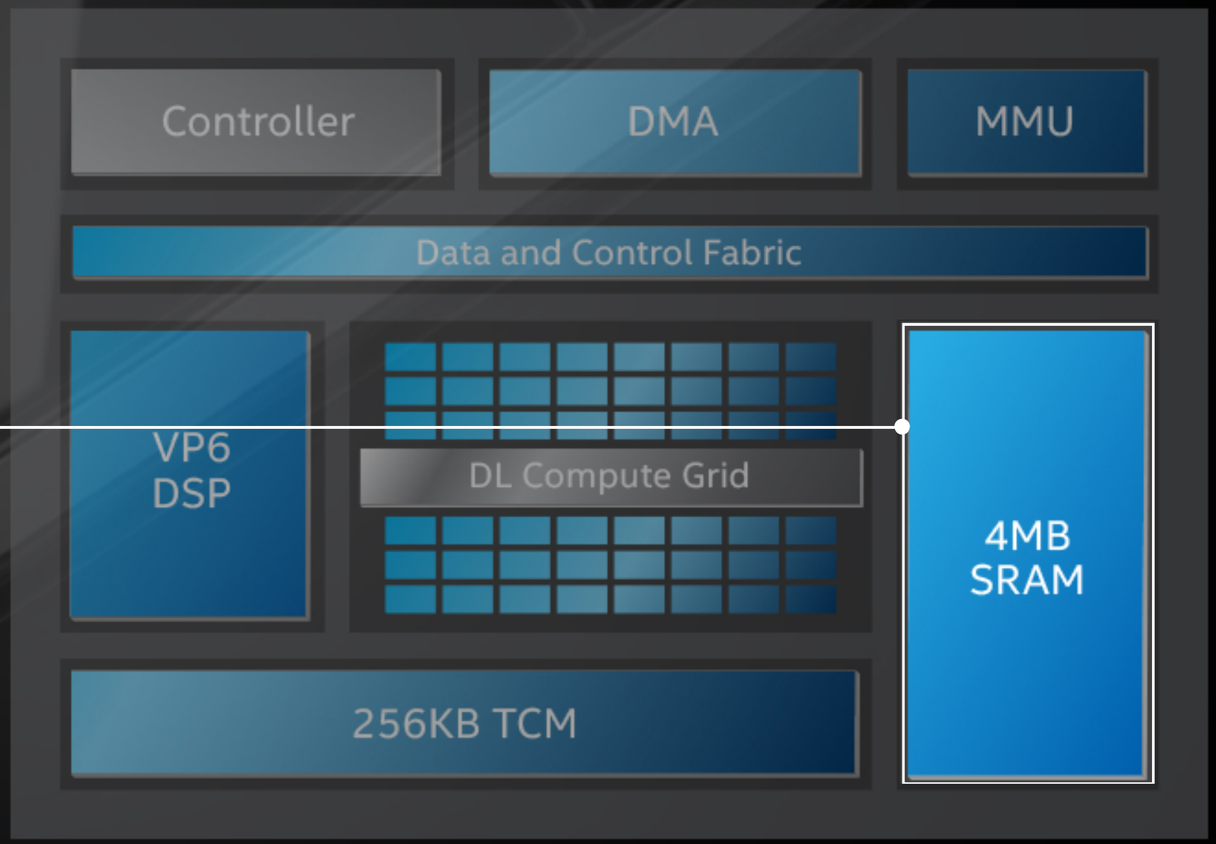
Inference compute engines (ICE) optimized
for maximum performance and efficiency

DEEP LEARNING COMPUTE GRID

PROGRAMMABLE VECTOR PROCESSOR (DSP)

LARGE LOCAL SRAM

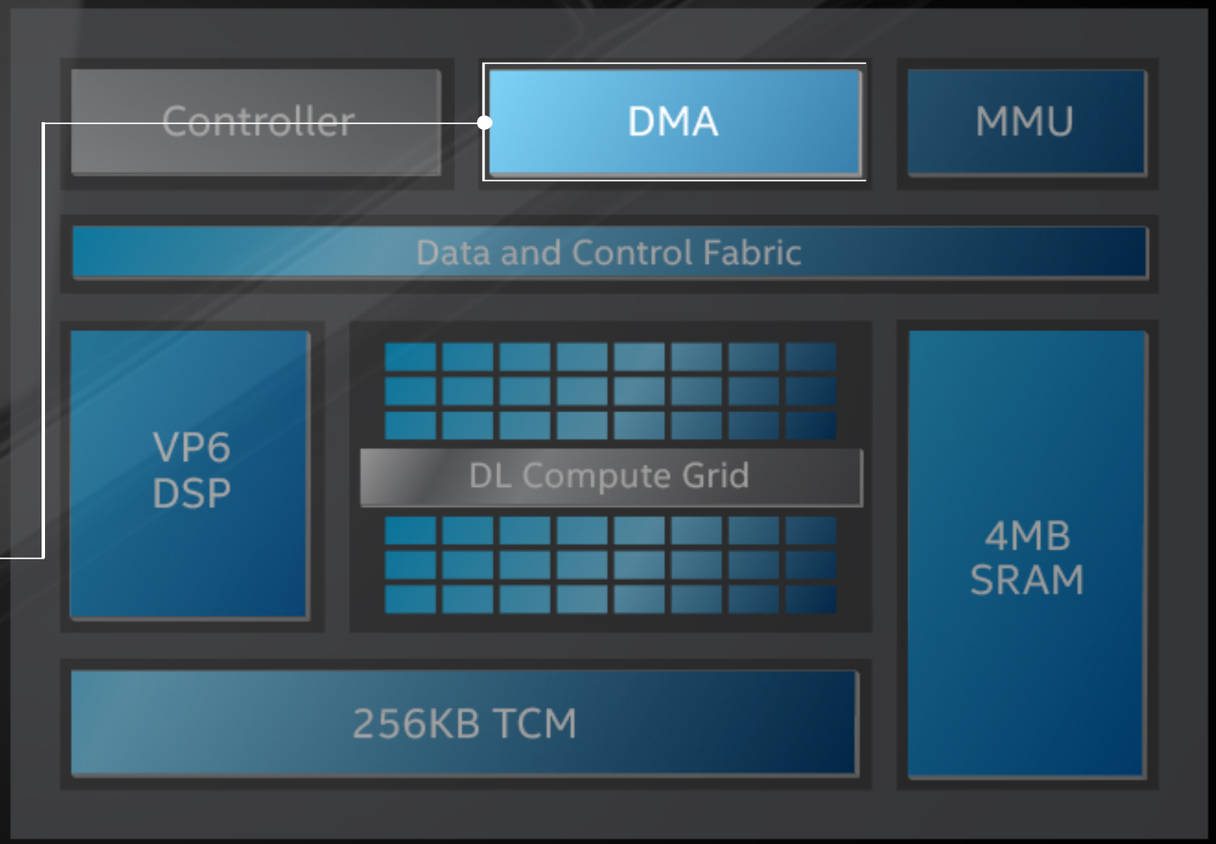
HIGH-BANDWIDTH DATA MEMORY ACCESS



ICE

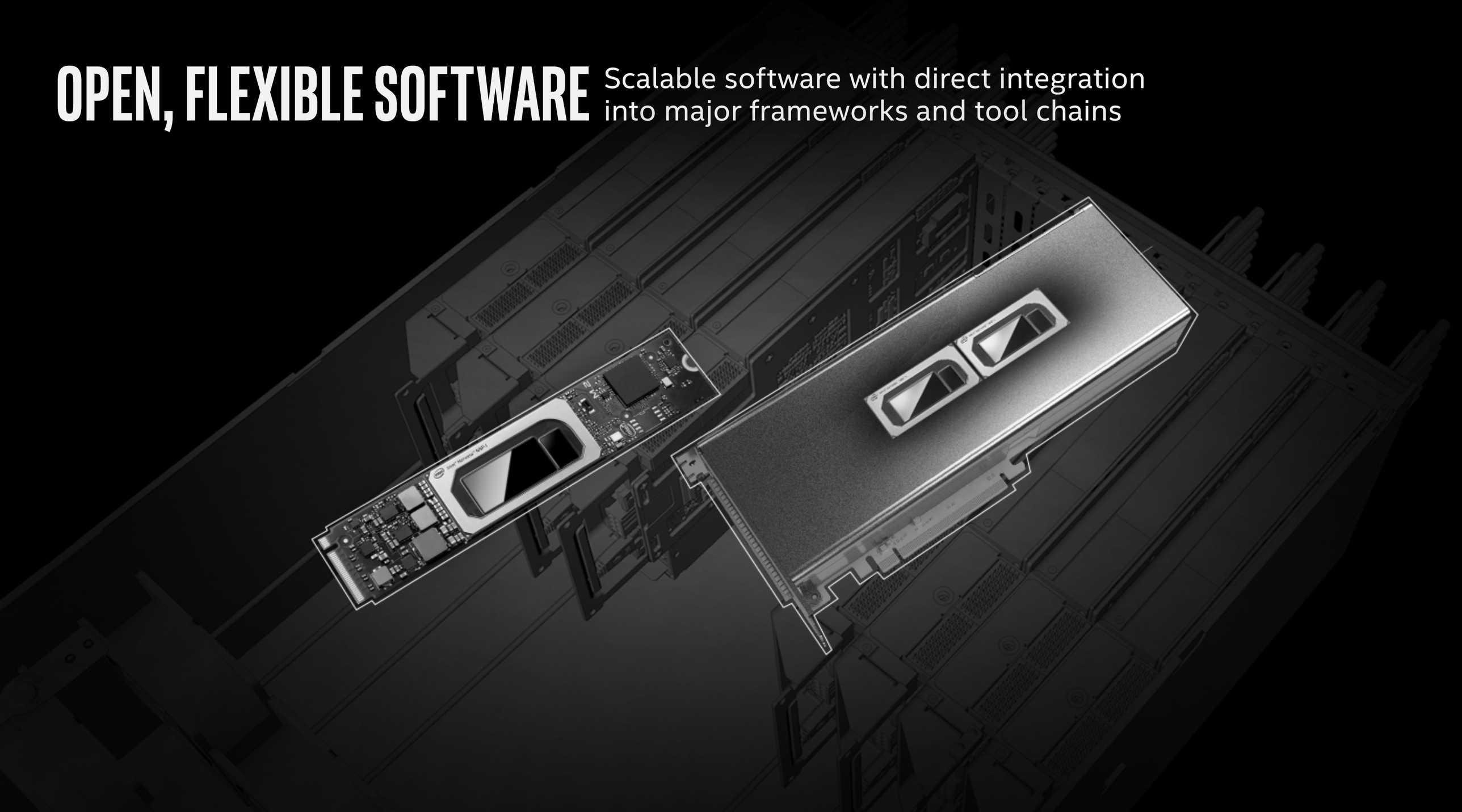
Inference compute engines (ICE) optimized
for maximum performance and efficiency

- DEEP LEARNING COMPUTE GRID
- PROGRAMMABLE VECTOR PROCESSOR (DSP)
- LARGE LOCAL SRAM
- HIGH-BANDWIDTH DATA MEMORY ACCESS



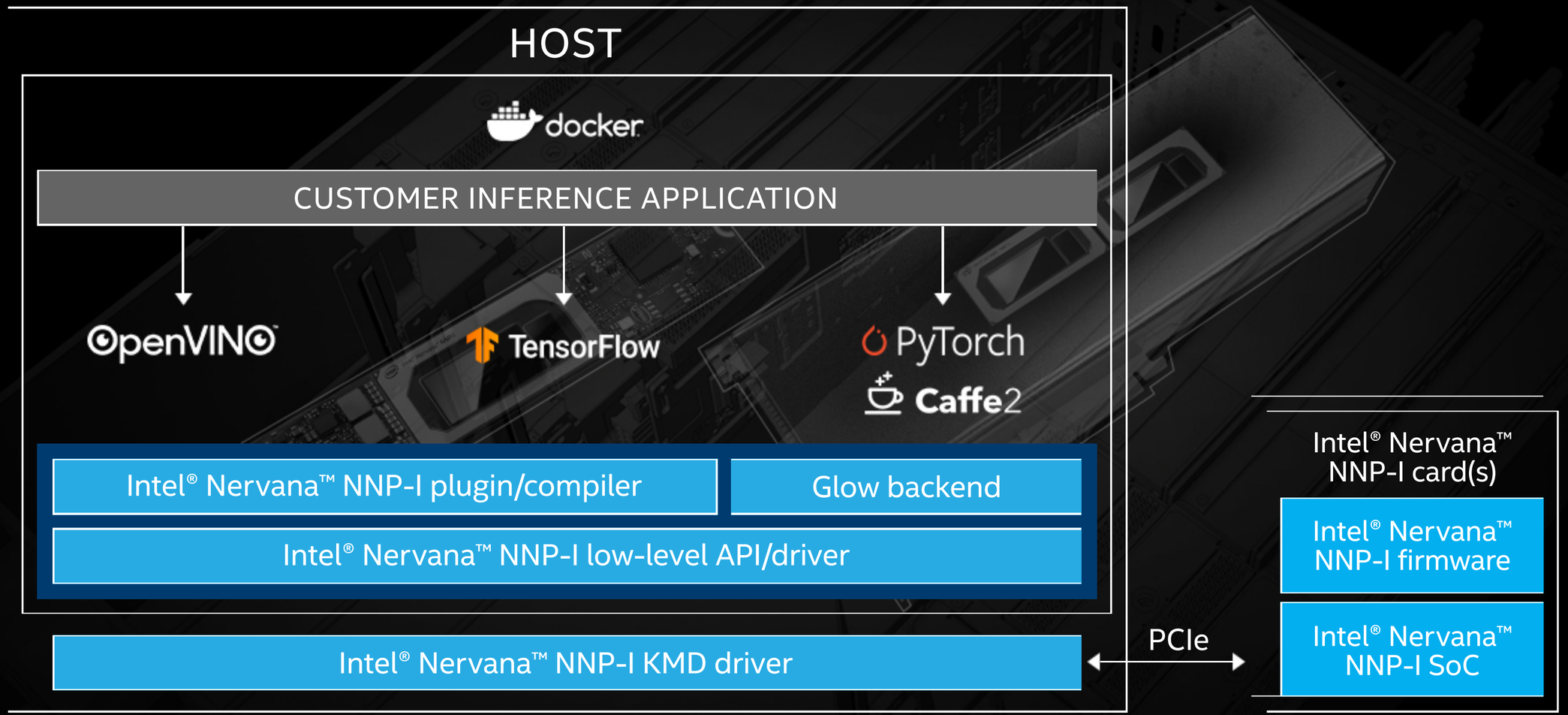
OPEN, FLEXIBLE SOFTWARE

Scalable software with direct integration into major frameworks and tool chains



OPEN, FLEXIBLE SOFTWARE

Scalable software with direct integration into major frameworks and tool chains

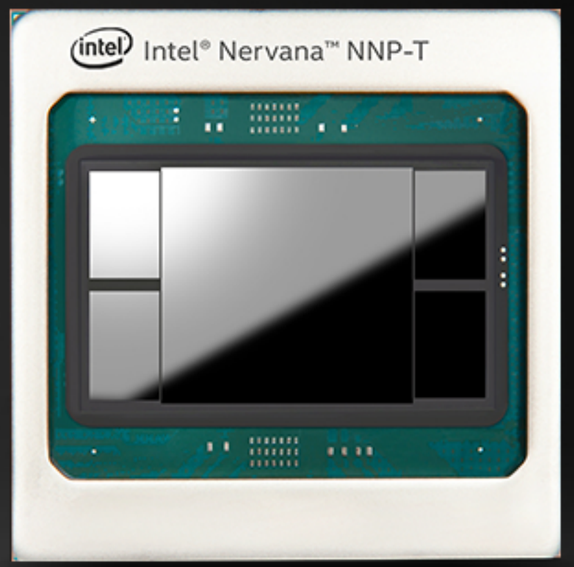


INTEL® NERVANA™ NEURAL NETWORK PROCESSOR FAMILY

Delivering the scale and efficiency demanded
by deep learning model evolution

INTEL® NERVANA™ NNP-I

Intense inference performance scaling for diverse latency and power needs



INTEL® NERVANA™ NNP-T

Deep learning training at incredible scale and efficiency, solving memory constraints and data flow bottlenecks