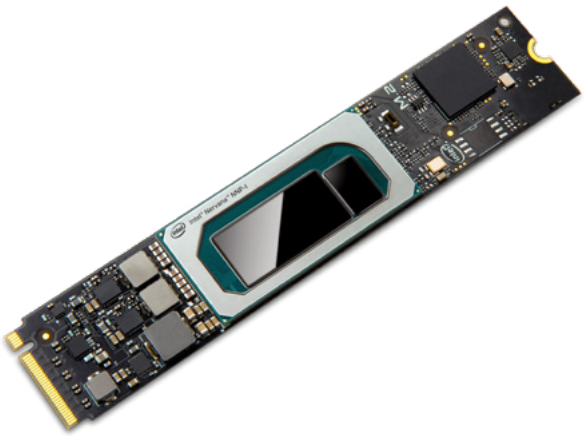
# PRODUCT BRIEF

Intel® Nervana™ Neural Network Processor for Inference
(Intel® Nervana™ NNP-I)

# Built to Accelerate Intense Multimodal Inferencing from Network Edge to Cloud

## The Intel® Nervana™ Neural Network Processor for Inference (Intel® Nervana™ NNP-I) accelerates complex deep learning with incredible scale and efficiency.



**Intel® Nervana™ NNP-I is available in multiple configurations as shown in the table below.**

| | INTEL NERVANA NNP I-1100 | INTEL NERVANA NNP I-1300 |
|---|---|---|
| **Form factor** | M.2 card | PCIe card |
| **Max TDP (card)** | 12W card with 1x Intel Nervana NNP I-1000 | 75W card with 2x Intel Nervana NNP I-1000 |
| **TOPS, Int8** | Up to 50 TOPS | Up to 170 TOPS |

Enterprise-scale AI deployments are significantly increasing the number of inference cycles for a diverse set of applications at various latency and power requirements. Intel has developed a novel ASIC that is purpose-built for ultra-efficient, multimodal inferencing. Intel® Nervana™ NNP-I was designed for intense, near-real-time, high-volume, low-latency compute. It can accommodate exponentially larger, more complex models and run dozens of models and networks in parallel.

### Highly programmable, performant, and efficient

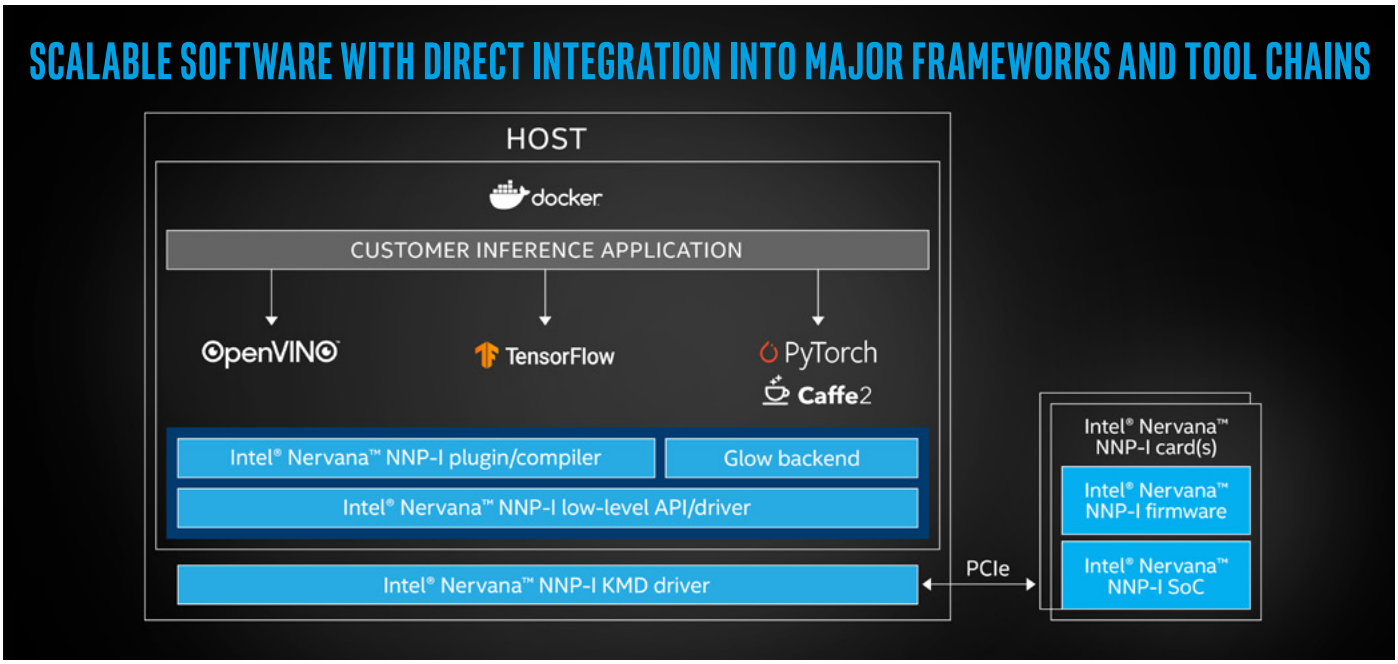
Intel Nervana NNP-I was designed to provide high inference throughput and power efficiency, plus programmable control for flexibility. Fully integrated voltage regulator (FIVR) technology optimizes SoC performance at different power envelopes for dynamic power management. On-die latest generation Intel® architecture (IA) cores that include Intel® Advanced Vector Extensions (Intel® AVX) and Vector Neural Network Instructions (VNNI) enable high levels of programmability, so that AI practitioners can optimize for the next generation of models.

### Optimized for low-precision inference and massive parallelism

Intel Nervana NNP I-1000 is the first generation of the Intel Nervana NNP-I family. Each Intel Nervana NNP I-1000 features 12 of our new Inference Compute Engines (ICEs). By combining these engines with two Intel CPU cores, we can achieve programmability with optimized throughput, allowing numeric flexibility and fast code porting to adjust for many different applications. This purpose-built ASIC offers mixed-precision support with a special focus on low-precision applications for near-real-time performance.

### Load once, infer many times

Intel Nervana NNP-I's memory allows many large models to be loaded once, and then service a large number of inference requests. Large on-die SRAM and an optimized on-die coherent network-on-chip (NoC) enable Intel Nervana NNP-I 1000 to deliver energy-efficient performance for deep learning inference. This design enables low-latency operations by minimizing memory access, utilizing multiple memory hierarchies for fast data sharing, and minimizing data travel with broadcast and data re-use schemes.




## Deployment for the real world

Intel Nervana NNP-I supports deployment in both data centers and at the network edge, with multiple form factors and power levels for flexible ways of deploying inference at scale. It operates as a complete system on an M.2 or PCIe card that includes flash components, BIOS, OS, full driver, software stack, and an Intel® FPGA to help control and reboot the system.

## Open, flexible software

Our full software stack for Intel Nervana NNP-I is built with open components, allowing developers to work the way they want: via direct integration with deep learning frameworks, a graph compiler, or low-level kernel programmability. Our comprehensive, standards-based software support includes integration for all major deep learning frameworks, ONNX, Intel® nGraph, the Intel® Distribution of OpenVINO™ toolkit, and C++.

## ACCELERATE WITH PURPOSE

**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

**Learn more about Intel® Nervana™ Neural Network Processors for Inference at intel.ai/nervana-nnp/nnpi**

intel AI

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product or component can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

1119/RD/CMD/PDF