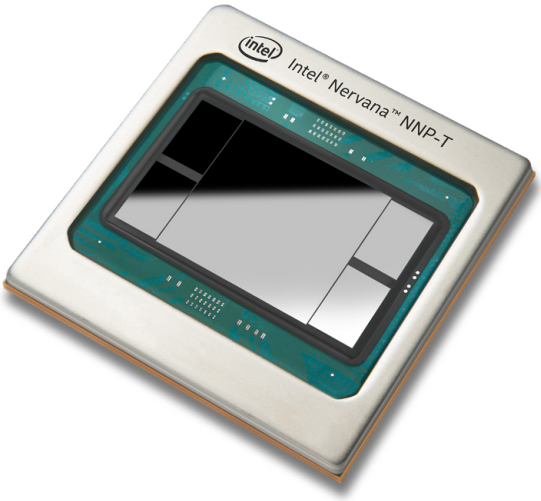
PRODUCT BRIEF

Intel® Nervana™ Neural Network Processor for Training
(Intel® Nervana™ NNP-T)

Built to Accelerate Distributed Deep Learning Training at Near-Linear Scale

The Intel® Nervana™ Neural Network Processor for Training (Intel® Nervana™ NNP-T) enables advanced AI systems for large-scale deep learning training.



Deep learning (DL) models are getting larger and more complex to help machines understand context and make decisions. This next-level reasoning requires new training architectures that efficiently scale out across numerous servers, allowing rapid experimentation, even in models requiring billions or trillions of parameters. The Intel® Nervana™ NNP-T processor, taking its inspiration from the brain, was designed to address this need by enabling the construction of tightly integrated server PODs for efficient and rapid training of the next wave of large DL models.

The Intel Nervana NNP-T design

Intel Nervana NNP-T with inter-chip links (ICLs) is designed with a unique balance of compute, memory, and communications specifically to process DL workloads and move large amounts of data. NNP-T's high-speed ICL communications fabric enables customers to achieve near-linear scale by directly connecting NNP-T cards within servers, between servers, and inside and across racks, creating high performance computing PODs. NNP-T maximizes processor and server POD utilization while reducing time to train, making it a highly energy-efficient alternative to general-purpose compute.

Purpose-built to accelerate training

- Provides high compute utilization via Tensor Processing Clusters (TPCs) with bfloat16 numeric format
- Balances compute, memory, and communication architecture to optimally utilize the silicon
- Leverages both on-die SRAM and on-package high bandwidth memory (HBM) to keep data local, reducing data movement

Designed for scale-out

- Uses ICLs in a glueless fabric architecture for efficient distributed training
- Offers ICL fabric with a fully programmable router and supports reliable transmission
- Achieves near-linear scaling with ICLs across multiple cards, systems, and PODs

Open standards and software, plus programmability

- Offers PCIe and OCP Accelerator Module (OAM) form factors to enable system vendors to innovate on hardware designs
- Supports common open source DL frameworks such as TensorFlow, PaddlePaddle, and PyTorch
- Provides a flexible, programmable Tensor-based instruction set architecture (ISA)

INTEL® NERVANA™ NNP-T PRODUCT PLATFORM

TensorFlow | PyTorch | PaddlePaddle

nGraph

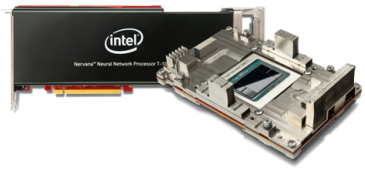
NNP-T DL Library

Kubernetes orchestration manager with NNP-T plugin

NNP-T Management Layer

Virtual and Docker images, pip installer binaries, tarballs, source access

Tools, profilers, configuration tools



INTEL NERVANA NNP-T PCIE AND OAM MEZZ CARDS

For server solutions

SYSTEMS WITH INTER-CHASSIS FABRIC

For OEM server solutions

POD REFERENCE DESIGN

For OEM cloud scale solutions

The product platform offers a full software stack along with a management suite that includes tools and profilers for building and scaling system solutions from small-scale data centers to hyperscale cloud deployments.

Each NNP-T is powered by up to 24 TPCs to perform DL training operations. Each TPC implements a combination of the brain floating-point (bfloat16) and 32-bit floating-point (FP32) numeric formats. Tensor-based bfloat16 architecture brings flexibility to support multiple deep learning primitives while making hardware components as efficient as possible. Each NNP-T also has 16 bidirectional high-speed ICLs, resulting in near-linear scaling across multiple cards in a system, multiple systems in a rack, and multiple racks in a POD. The ICL fabric implements a fully programmable router and support for reliable transmission. TPCs can directly transfer data to the links, rather than taking up precious bandwidth from the HBM subsystem, ensuring lower latency and greater efficiency.

The software stack supports popular frameworks such as TensorFlow, PaddlePaddle, and PyTorch to enable customers to develop models on the framework of their choice. The software handles memory management, message passing, synchronization, and scheduling of data transfers to ease data-parallel and model-parallel distributed training across hundreds of cards with high communications efficiency.

Using NNP-T, customers can configure a system with up to eight cards and connect multiple training systems to build a training POD. Customers can implement data parallelism and model parallelism on systems or PODs to train large complex models. NNP-T can be connected in a number of topologies, such as ring, hybrid cube mesh, and fully connected, to achieve various throughput and latency requirements. For instance, in a 32-card, 15kW rack preproduction system configuration, Intel Nervana NNP-T demonstrates convergence at state-of-the-art accuracy as measured on ResNet-50, same as with any FP32 compute.[1] Additionally, NNP-T achieves up to 95% scaling on ResNet-50 and BERT as measured on a 32-card rack.[2]

Intel Nervana NNP-T highlighted features

Tensor Processing Clusters (TPCs) support a combination of bfloat16 format for high performance and FP32 for accuracy and flexibility, providing high compute utilization.

Brain floating-point format (bfloat16), a truncated (16-bit) version of the 32-bit single-precision floating-point format (FP32), enables easy conversion to and from FP32. Additionally, bfloat16 has more dynamic range than FP16, making it easy for DL models to converge.

Optimized memory architecture uses a 2-D, on-chip mesh to enable high bandwidth and predictable latency between on-chip SRAM and off-chip HBM memory.

Intel® Nervana™ NNP-T inter-chip links achieve scale-out (high input/output bandwidth), where PODs connected by these links can be further scaled out over commodity data center fabric.

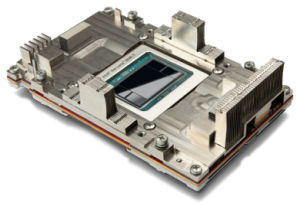
DL framework support leverages nGraph to efficiently optimize models in different supported frameworks such as TensorFlow, PyTorch, and PaddlePaddle.

Programmability with C++ and Python enables users to create a new kernel by implementing a custom operation in C++ or Python, using a Tensor domain-specific language. The custom operation is passed through the DL framework to a Multi-Level Intermediate Representation (MLIR)-based compiler.

An enhanced Kubernetes Scheduler (K8s) plugin makes ICL-aware scheduling decisions and supports workloads spanning ICL and data center fabric.

Intel® Nervana™ NNP-T 1300 PCIe Card

Intel® Nervana™ NNP-T 1400 OAM Mezzanine Card



Form Factor	Dual-slot standard PCIe card	OAM mezzanine card
Spec Compliance	PCIe CEM	OAM 1.0
Dimensions	265.32 mm x 111.15 mm	102 mm x 165 mm
Compute Core	22 TPC	24 TPC
Frequency	950 MHz	1100 MHz
SRAM	55 MB on-chip SRAM w/ ECC	60 MB on-chip SRAM w/ ECC
HBM	32 GB 2.4 Gbps HBM2 w/ ECC	32 GB 2.4 Gbps HBM2 w/ ECC
TDP	300W	375W
Thermal Solution	Passive integrated	Passive cooling
Inter-Chip Link (ICL)	16 x 112 Gbps bidirectional ICL (448 GB/s)	16 x 112 Gbps bidirectional ICL (448 GB/s)
I/O to Host CPU	PCIe Gen3/Gen4 x 16	PCIe Gen3/Gen4 x 16
Supported ICL Interconnect Topology	Ring	Ring, hybrid cube mesh, fully connected
Multichassis/multirack scaling	Yes	Yes

Learn more about Intel® Nervana™ Neural Network Processors for Training at intel.ai/nervana-nnp/nnpt

intel AI

1 Measurements based on Intel internal testing using preproduction hardware/software as of November 2019. Accuracy target as referenced in MLPerf Link: https://mlperf.org/training-overview/. All products, computer systems, dates, and figures are preliminary based on current expectations and are subject to change without notice. For more complete information about performance results, visit www.intel.ai/benchmarks.

2 Measurements based on Intel internal testing using preproduction hardware/software as of November 2019. All products, computer systems, dates, and figures are preliminary based on current expectations and are subject to change without notice. For more complete information about performance results, visit www.intel.ai/benchmarks.

1119/RD/CMD/PDF