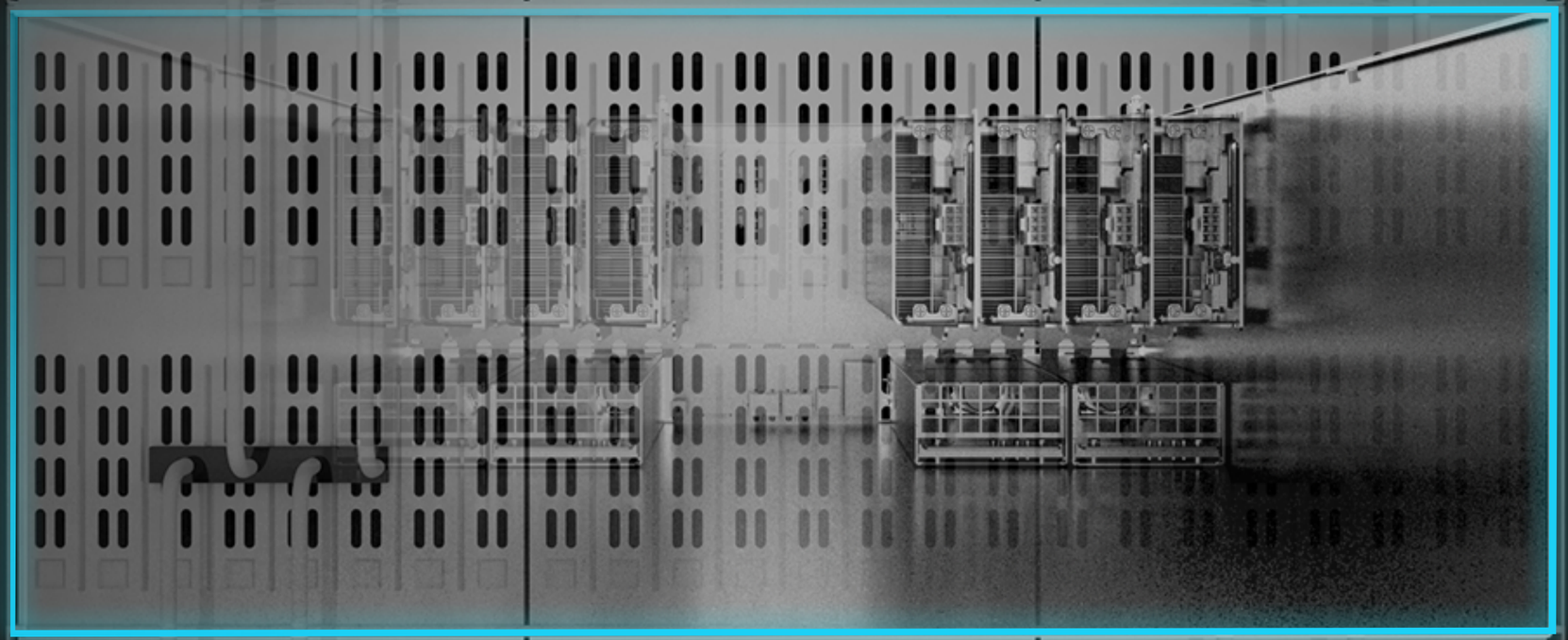
PURPOSE-BUILT TO ACCELERATE DEEP LEARNING TRAINING

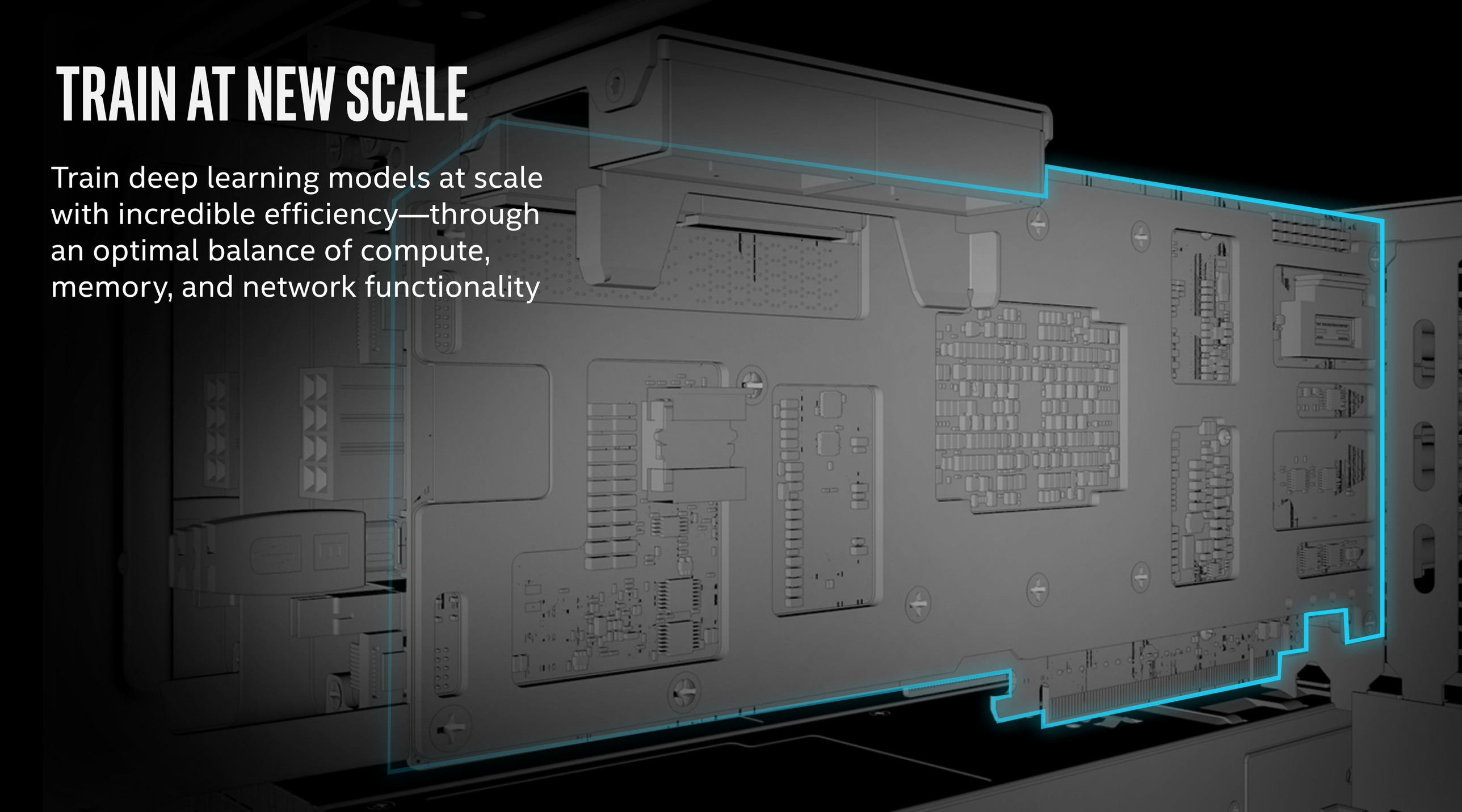
INTEL® NERVANA™ NEURAL NETWORK PROCESSOR FOR TRAINING

(Intel® Nervana™ NNP-T)

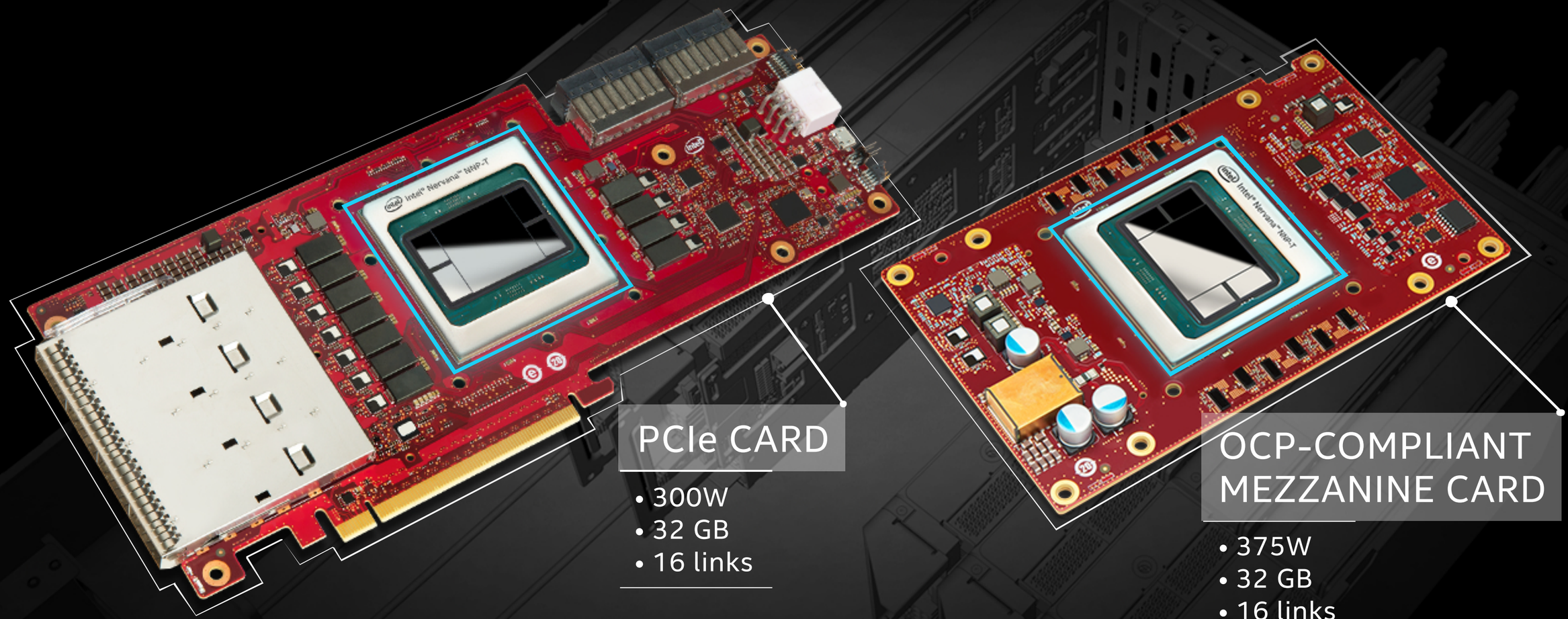


TRAIN AT NEW SCALE

Train deep learning models at scale with incredible efficiency—through an optimal balance of compute, memory, and network functionality



INTEL® NERVANA™ NNP T-1000



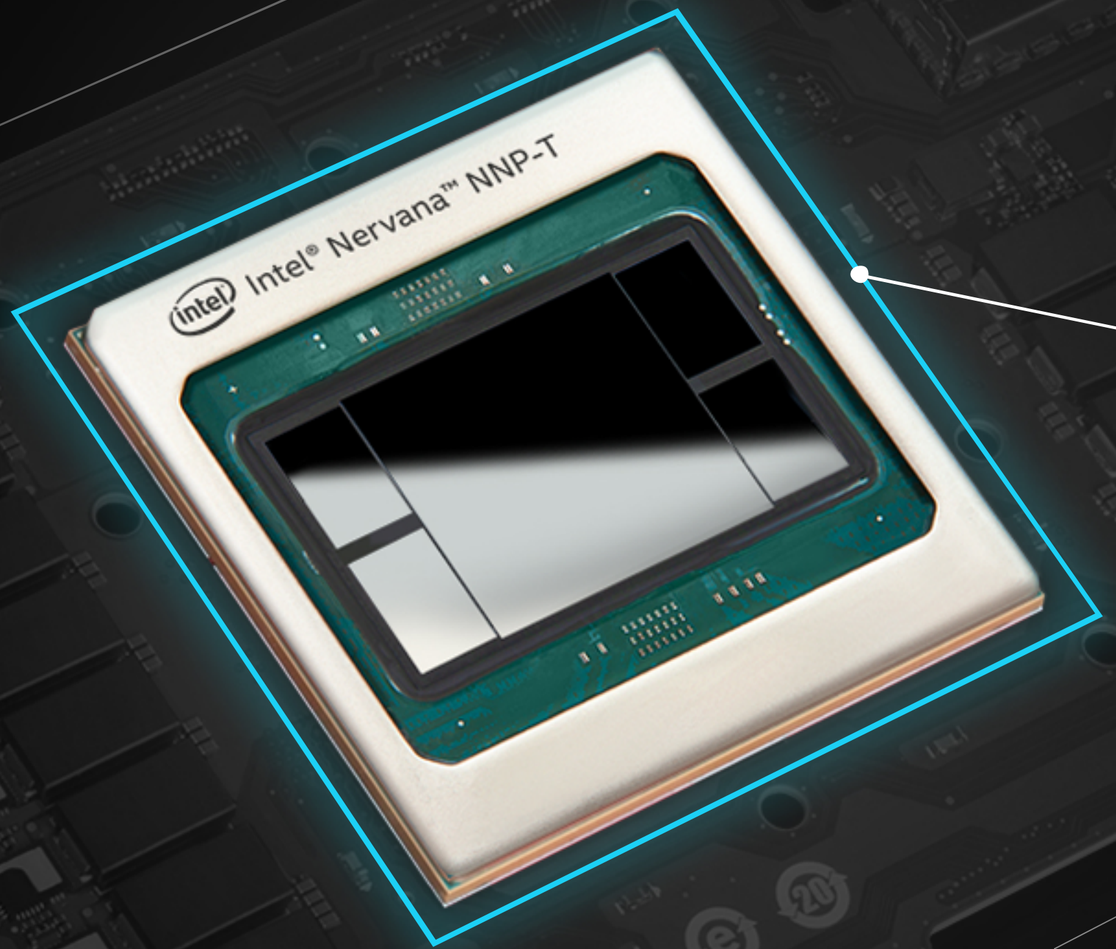
PCIe CARD

- 300W
- 32 GB
- 16 links

OCP-COMPLIANT MEZZANINE CARD

- 375W
- 32 GB
- 16 links

INTEL® NERVANA™ NNP T-1000



QUICK SPECS

- Bfloat16: 119 TOPS
- 60 MB on-chip SRAM

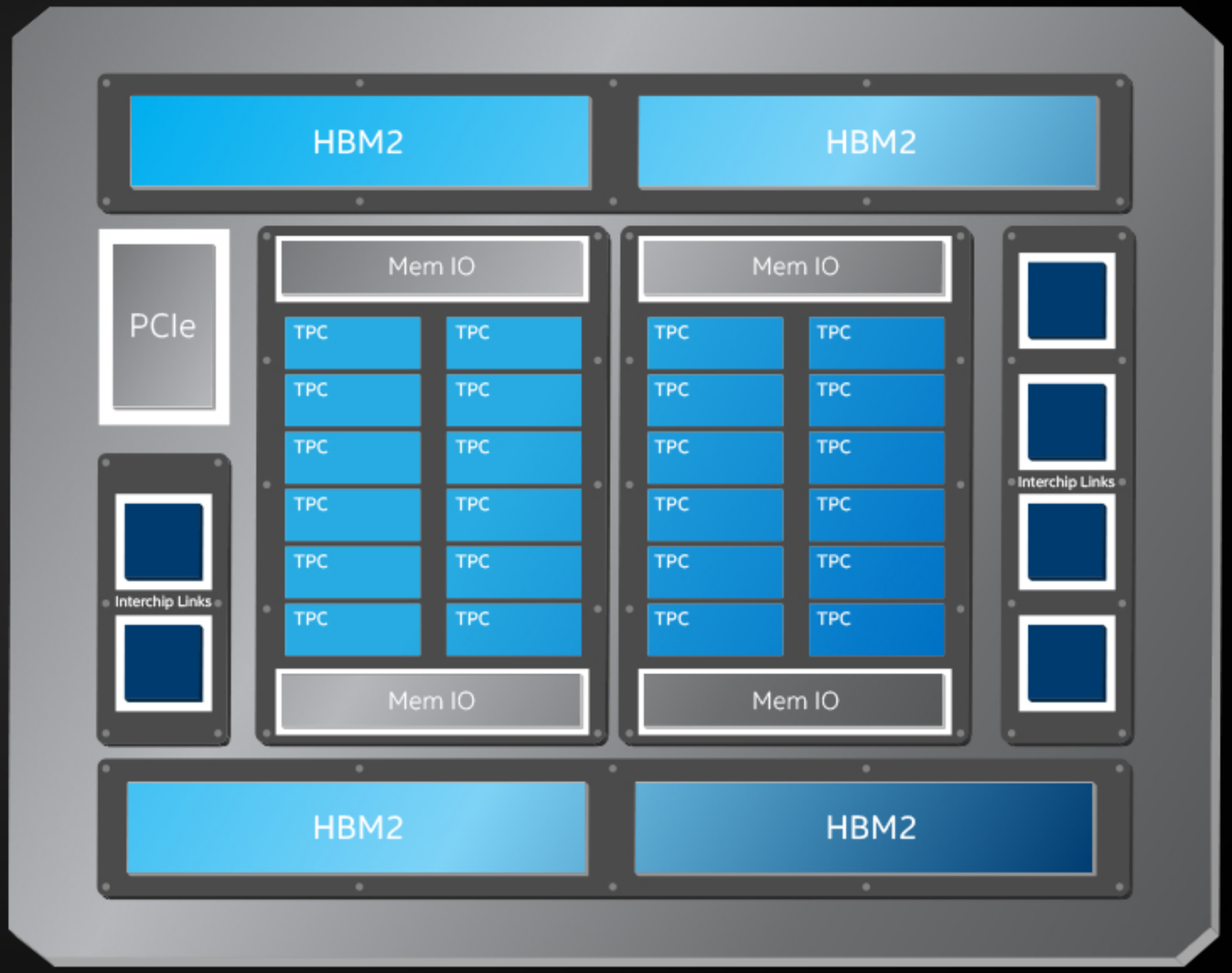
INTEL® NERVANA™ NNP T-1000

Balanced compute, memory, and communication for near-linear scaling of even the most complex models

DEDICATED TENSOR PROCESSING CLUSTERS (TPCs)

HIGH-BANDWIDTH MEMORY

SCALABLE DESIGN

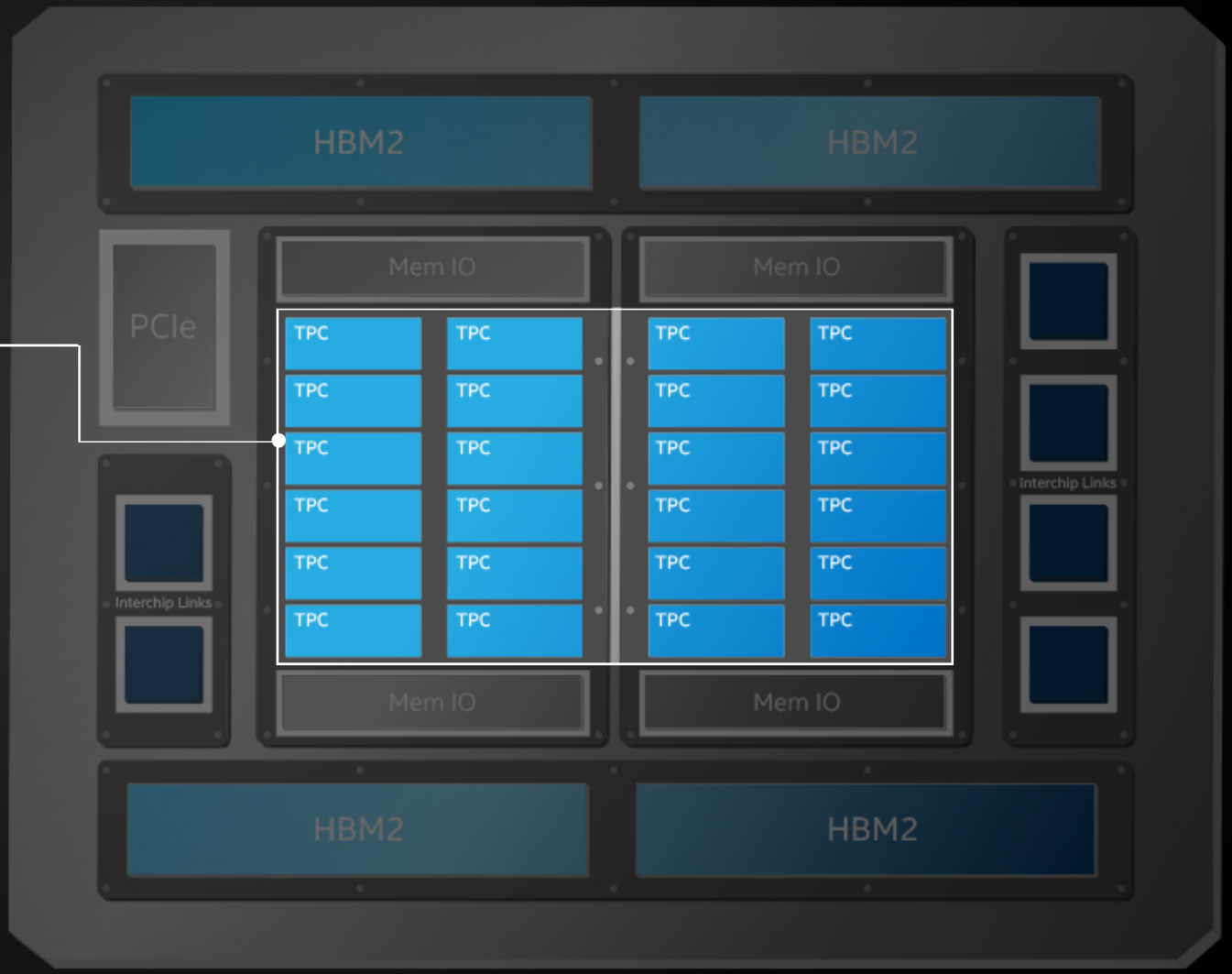


INTEL® NERVANA™ NNP T-1000

Balanced compute, memory, and communication for near-linear scaling of even the most complex models

DEDICATED TENSOR PROCESSING CLUSTERS (TPCs)

Specialized TPCs provide high utilization of underlying compute



INTEL® NERVANA™ NNP T-1000

Balanced compute, memory, and communication for near-linear scaling of even the most complex models

HIGH-BANDWIDTH MEMORY

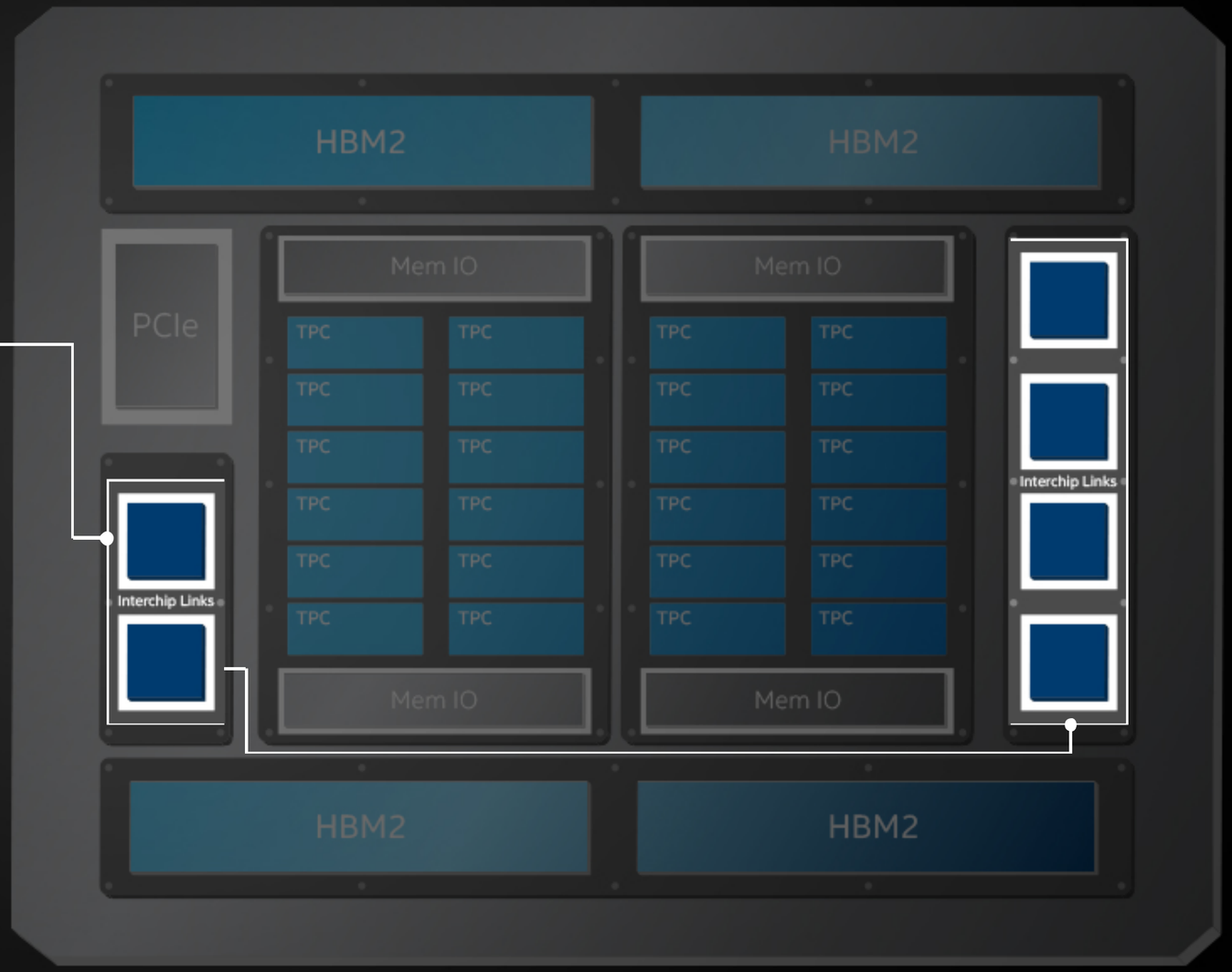
High-efficiency memory architecture with independent HBM and TPC-to-TPC data buses enables scaling of complex training models

INTEL® NERVANA™ NNP T-1000

Balanced compute, memory, and communication for near-linear scaling of even the most complex models

SCALABLE DESIGN

Glueless connectivity with massive scaling via intra- and inter-chassis links, enabling cross-chassis scale-out with the same network connectivity



INTEL® NERVANA™ NNP T-1000 Product Platform

TensorFlow | PyTorch | PaddlePaddle

nGraph

Intel® Nervana™ NNP-T DL library

Kubernetes orchestration manager
with Intel® Nervana™ NNP-T plug-in

Intel® Nervana™ NNP-T management layer

Virtual and Docker* images, PIP installer binaries, tar balls, 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

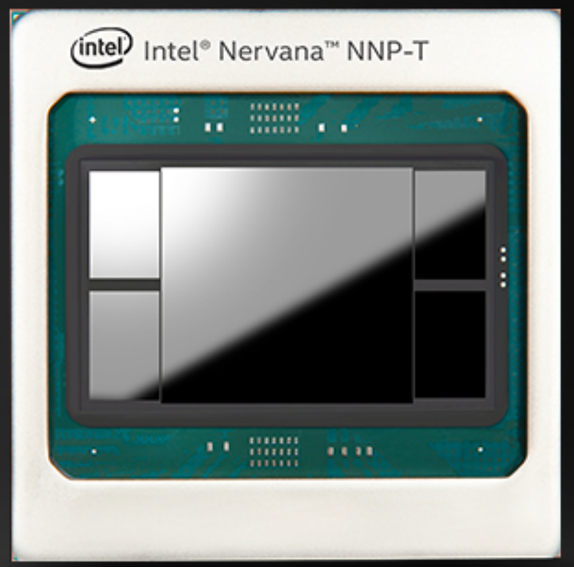
POD
REFERENCE DESIGN
For OEM cloud-scale solutions

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