
intel AI
WELCOME TO A NEW INTELLIGENCE
AI DELIVERED AT INTEL SCALE
NAVEEN RAO
CORPORATE VICE PRESIDENT
GM, ARTIFICIAL INTELLIGENCE
Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

$3.5+ BILLION

AI Revenue in 2019

There is no single:

Approach
Budget
Chip
System

DATA READINESS, EXPERTISE, AND USE CASE DETERMINE AI SOLUTION

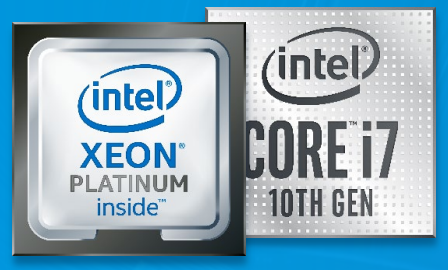
AI WILL INFUSE EVERYTHING...
...SO WE PUT IT EVERYWHERE

OPTIMIZED SOFTWARE

Workload breadth

AI-Specific

CPU



Multi-Purpose,
Foundation for Analytics & AI

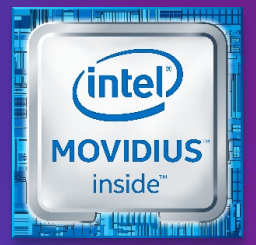
GPU

Data-Parallel Media,
Graphics, HPC & AI

FPGA

Real-Time &
Multi-Function Inference

ASIC



Edge Media, CV,
and Inference

NNP-I

Network Edge-to-Data
Center Inference

NNP-T

Fast Distributed
Training

BUILT-IN SECURITY

READY-MADE FOR AI

Up to 30x AI performance improvement with Intel® Deep Learning Boost (Intel DL Boost) compared to Intel® Xeon® Platinum 8180 processor (July 2017).

30X INFERENCE PERFORMANCE

Intel® Deep Learning Boost

BFLOAT SUPPORT

Demonstrating today
First to provide on multiple products

SOFTWARE OPTIMIZATIONS + EXPERTISE

Direct deep learning framework support
New libraries make hardware more AI-performant

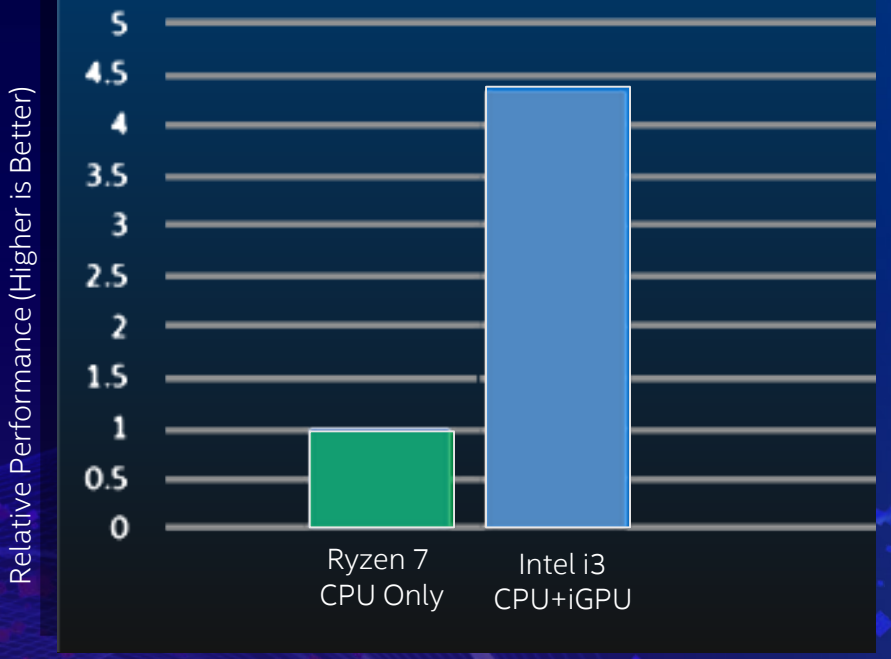
intel AI

MAKING EVERYONE A CREATOR

intel CORE i3 10TH GEN

UP TO 4.3X HIGHER INFERENCE THROUGHPUT

on ResNet 50 vs. AMD Ryzen 7 3700U



Throughput
(Batch size: 1; Precision: INT8)

AI SUMMIT 2019

See backup for configuration details. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.
Other names and brands may be claimed as the property of others.

intel AI

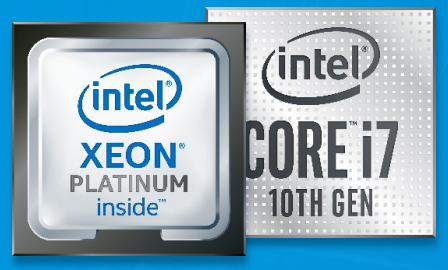
AI WILL INFUSE EVERYTHING...
...SO WE PUT IT EVERYWHERE

OPTIMIZED SOFTWARE

Workload breadth

AI-Specific

CPU



Multi-Purpose,
Foundation for Analytics & AI

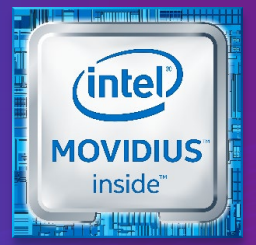
GPU

Data-Parallel Media,
Graphics, HPC & AI

FPGA

Real-Time &
Multi-Function Inference

ASIC



Edge Media, CV,
and Inference

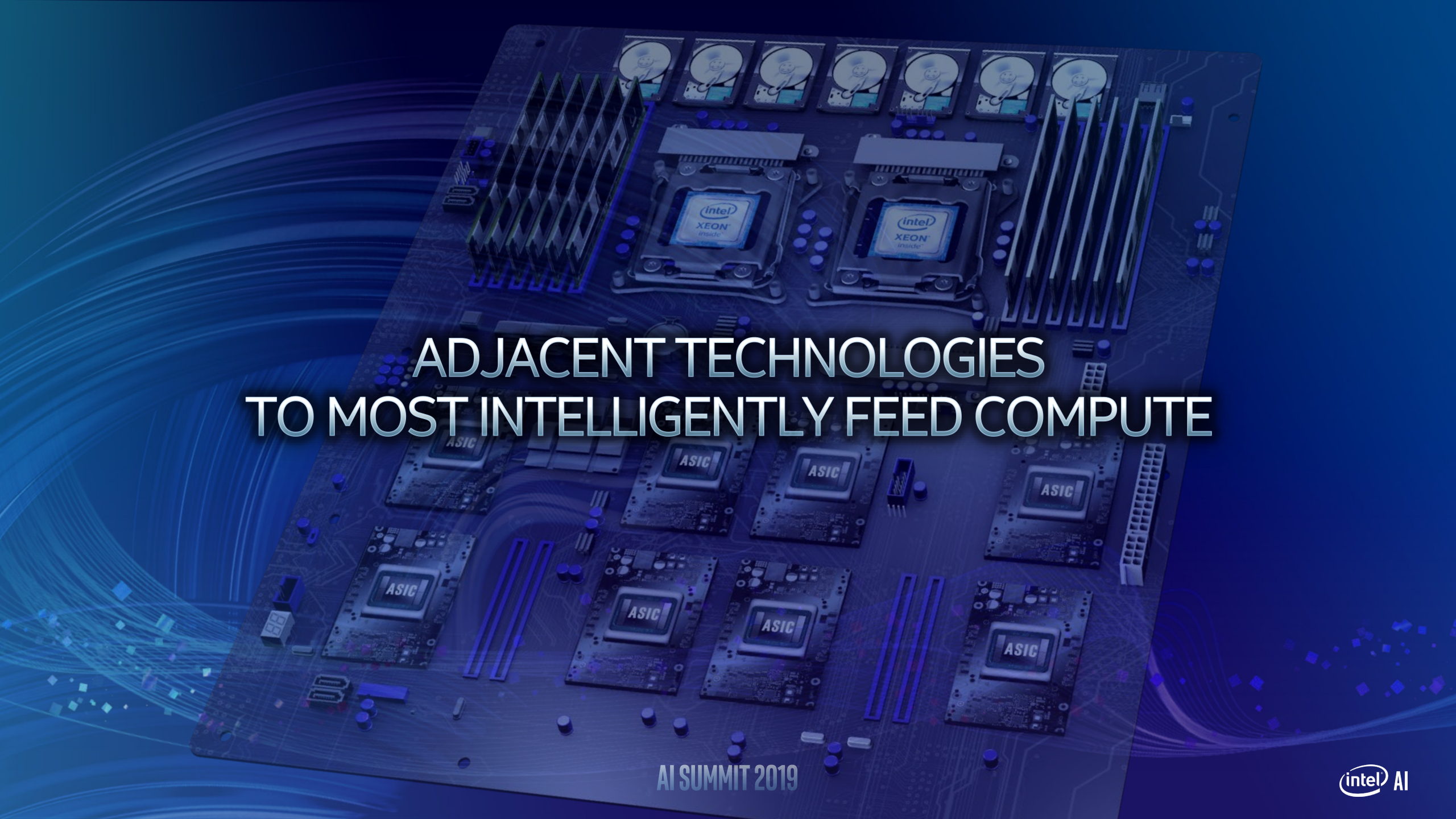
NNP-I

Network Edge-to-Data
Center Inference

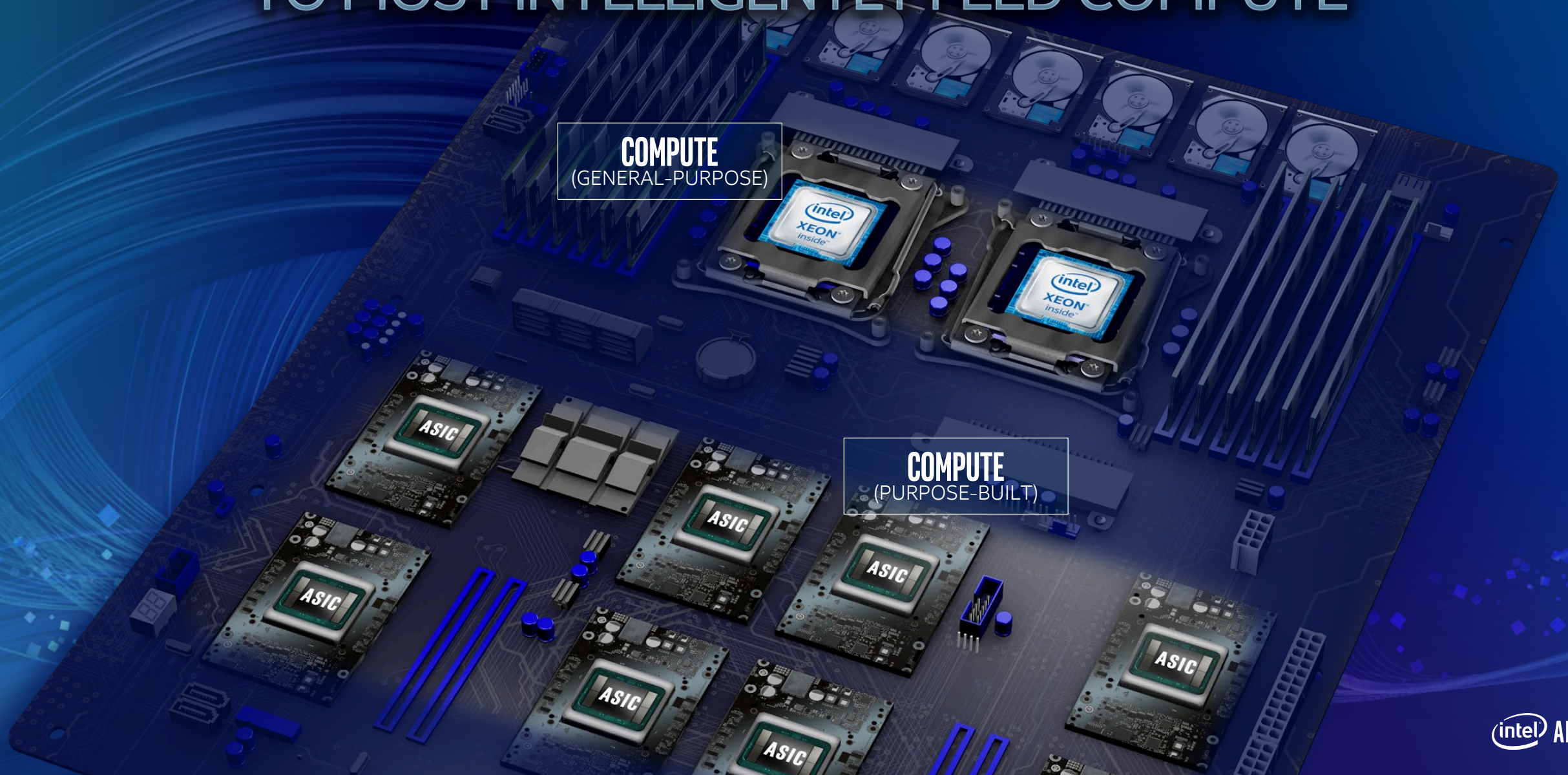
NNP-T

Fast Distributed
Training

BUILT-IN SECURITY

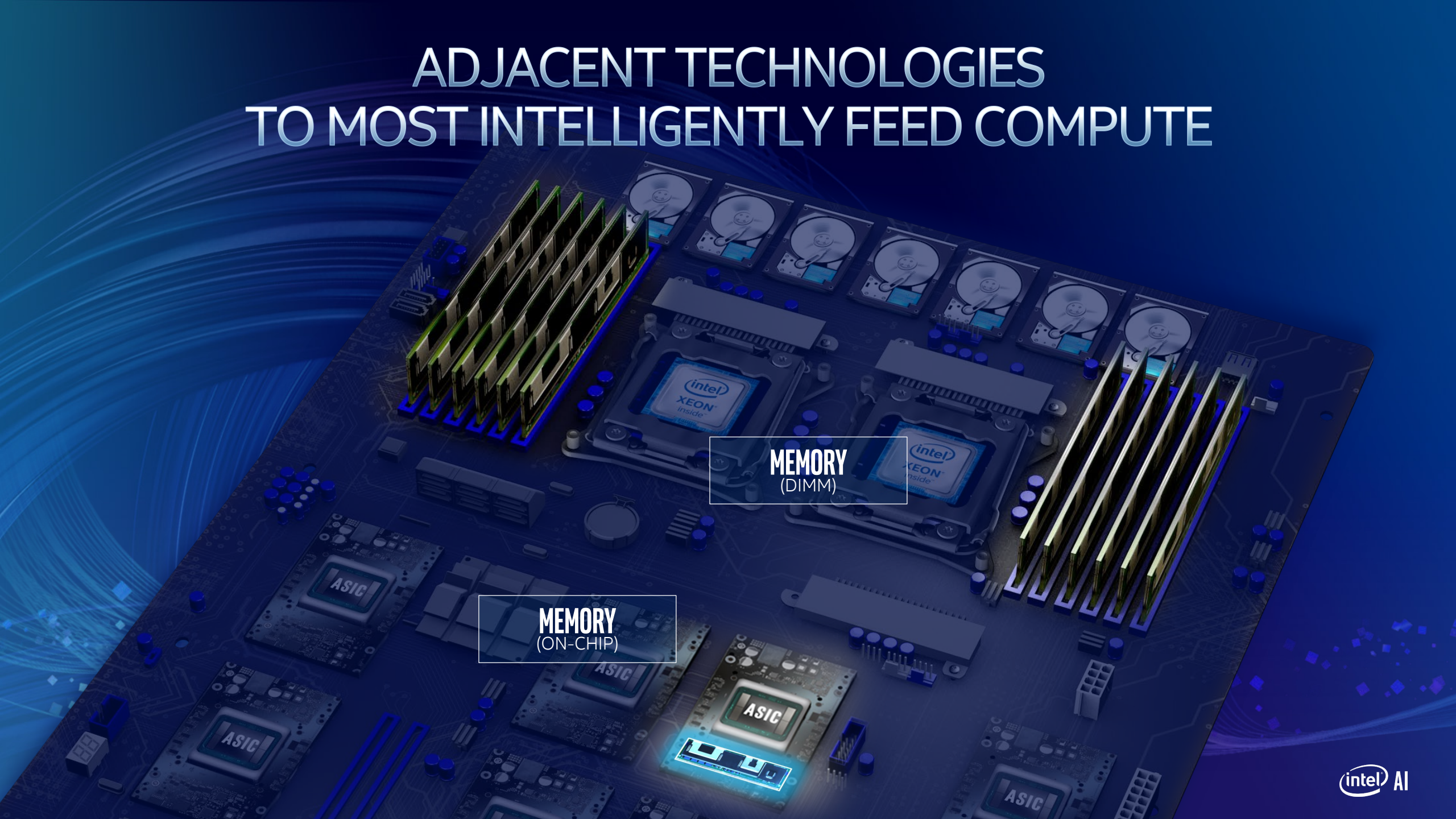

ADJACENT TECHNOLOGIES
TO MOST INTELLIGENTLY FEED COMPUTE
AI SUMMIT 2019
intel AI

ADJACENT TECHNOLOGIES TO MOST INTELLIGENTLY FEED COMPUTE



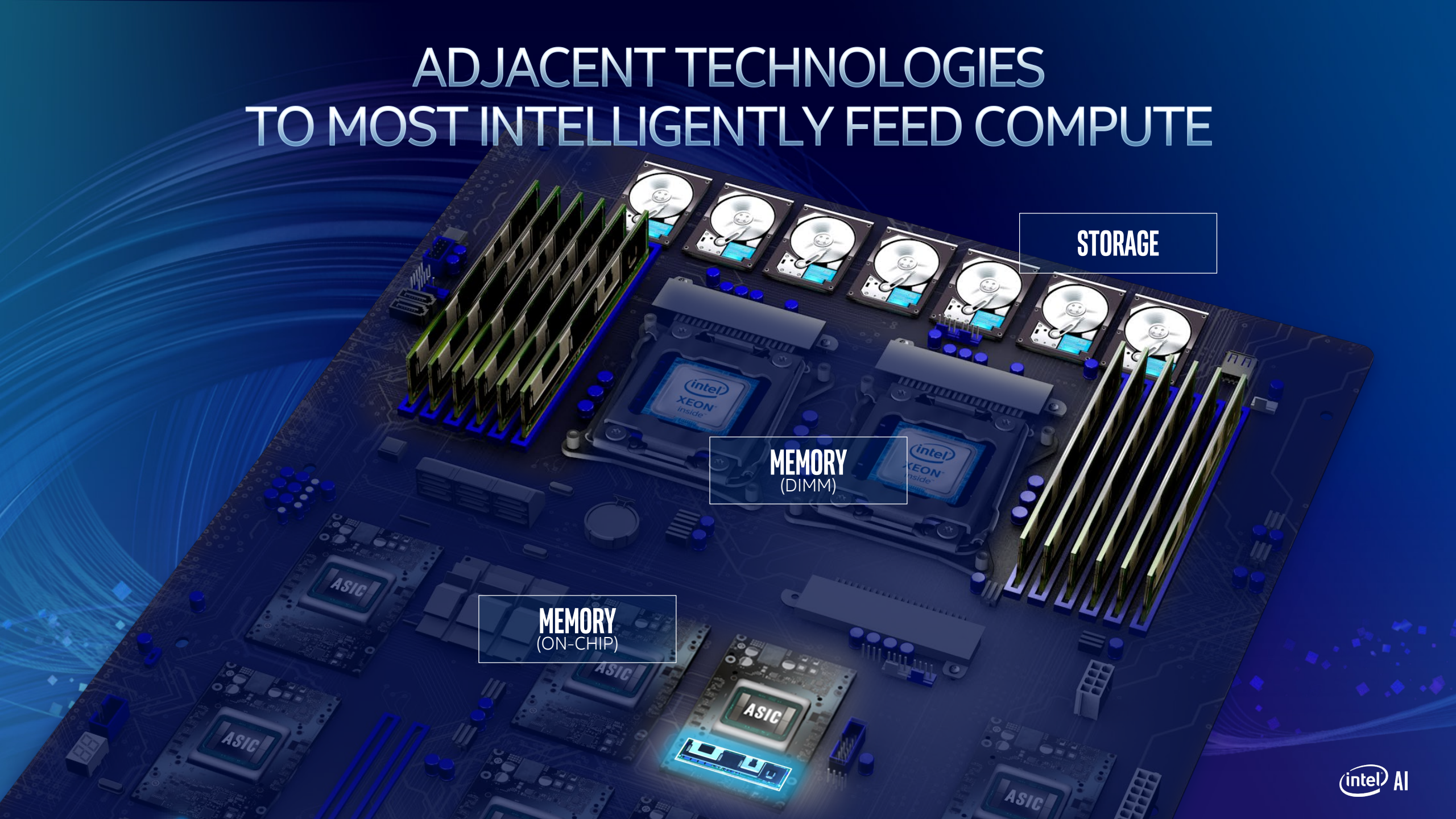
intel AI

ADJACENT TECHNOLOGIES TO MOST INTELLIGENTLY FEED COMPUTE



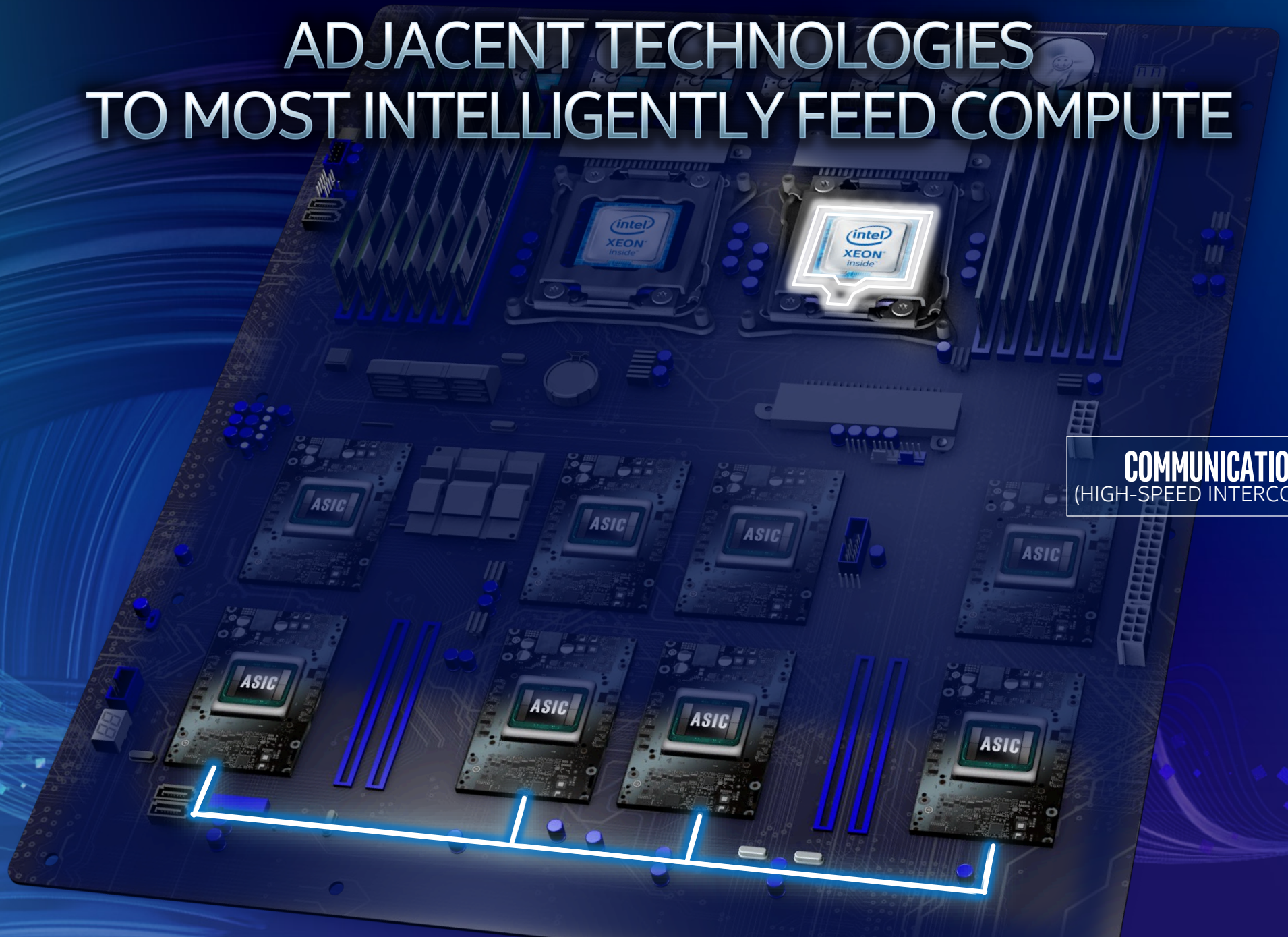
intel AI

ADJACENT TECHNOLOGIES TO MOST INTELLIGENTLY FEED COMPUTE



intel AI

ADJACENT TECHNOLOGIES TO MOST INTELLIGENTLY FEED COMPUTE



COMMUNICATIONS
(HIGH-SPEED INTERCONNECT)

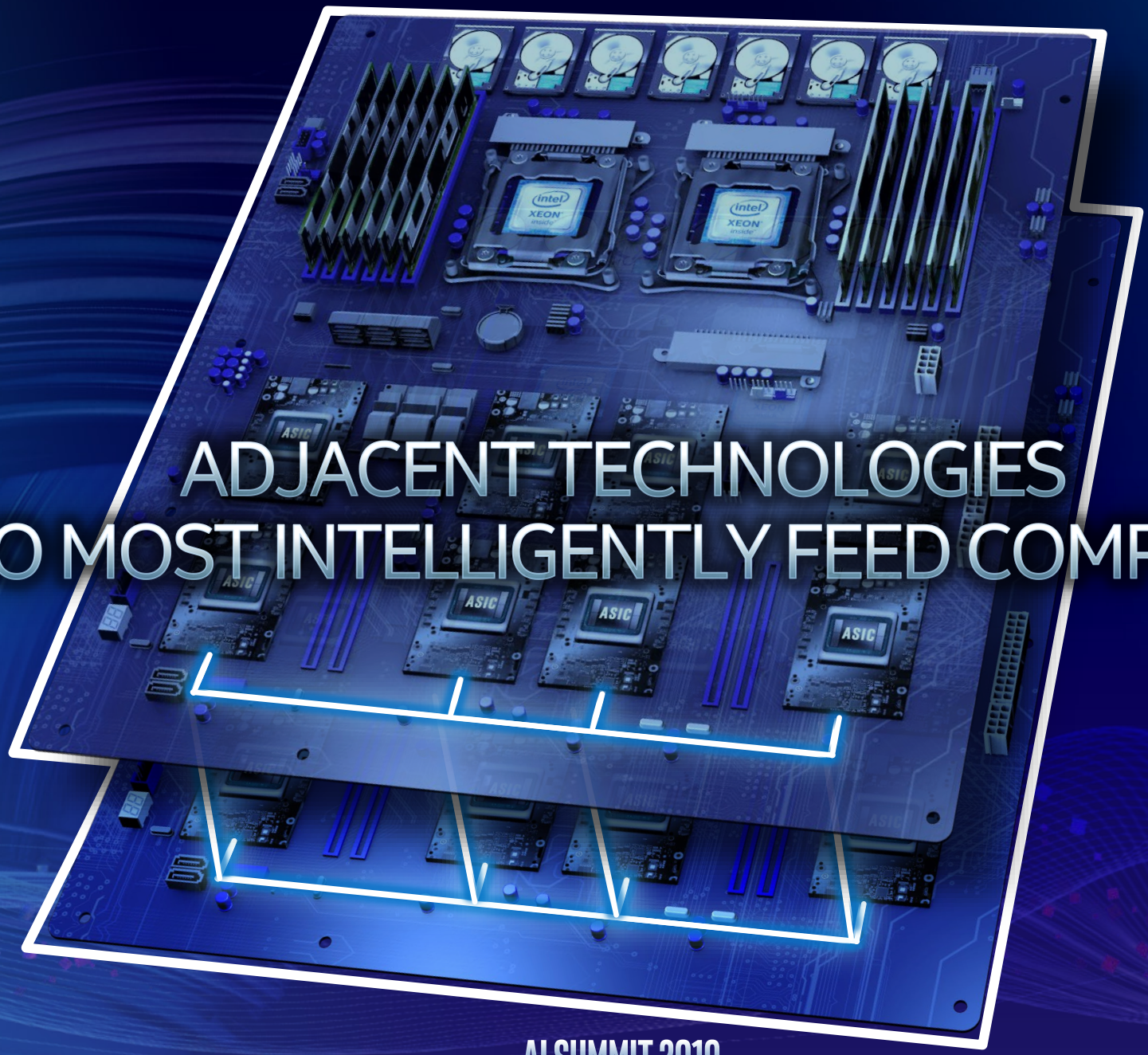
intel AI

ADJACENT TECHNOLOGIES TO MOST INTELLIGENTLY FEED COMPUTE

COMMUNICATIONS
(HIGH-SPEED INTERCONNECT)

COMMUNICATIONS
(SCALE INTRA-CHASSIS)

intel AI



ADJACENT TECHNOLOGIES TO MOST INTELLIGENTLY FEED COMPUTE

AI SUMMIT 2019

intel AI

Open software to keep hardware nimble and working better together.

OPEN

PROVIDING FLEXIBILITY TO ADVANCE A FAST-MOVING LANDSCAPE AND EMBRACE COMMUNITY INNOVATION

OpenVINO™ nGraph NAUTA ANALYTICS ZOO

COMPLETE

ACCESS TO KERNEL, COMPILER, AND FRAMEWORKS, FOR DEVELOPERS TO WORK HOW THEY WANT

TensorFlow PYTORCH Caffe GLOW mxnet ONNX

INNOVATIVE

TOOLS TO PUSH BOUNDARIES FOR NEXT-LEVEL AI

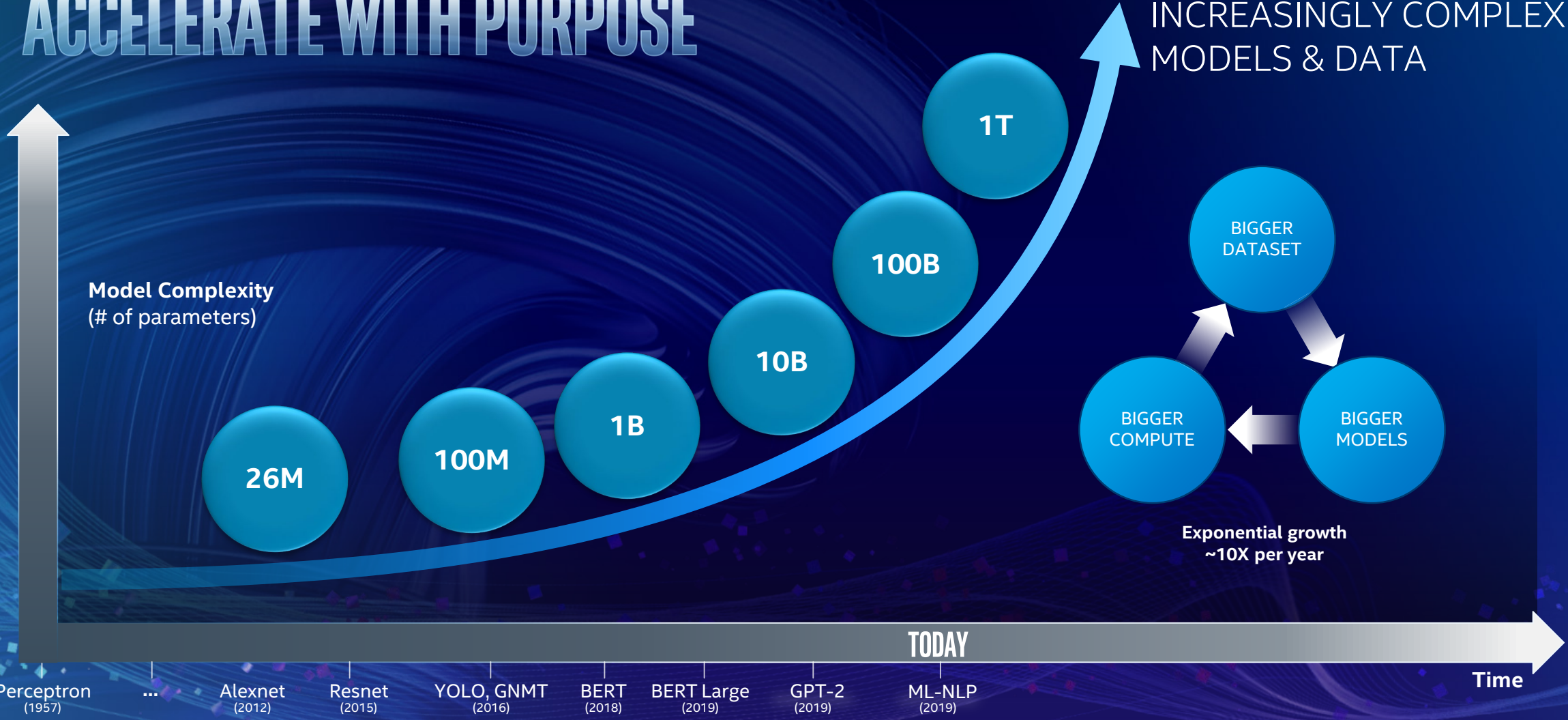
RL COACH NN DISTILLER HOMOMORPHIC ENCRYPTION (HE) TRANSFORMER NLP ARCHITECT

intel AI

AS MODEL COMPLEXITY GROWS, ACCELERATE WITH PURPOSE

NEXT-GEN DL

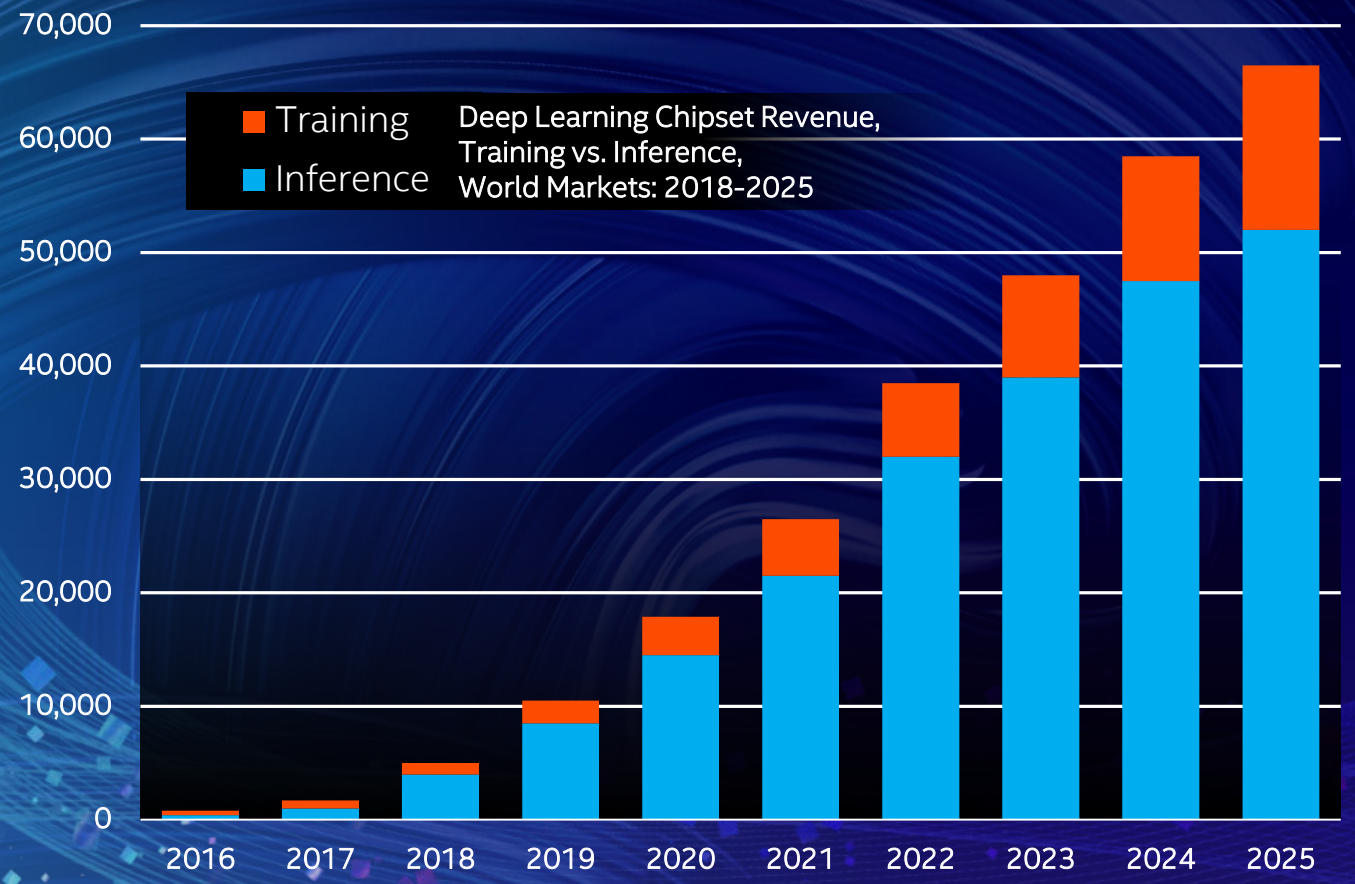
INCREASINGLY COMPLEX MODELS & DATA



AI SUMMIT 2019

intel AI

THE EDGE OPPORTUNITY…AND CHALLENGE



Tractica

"Cameras grow at highest CAGR."

MARKETSANDMARKETS™

"75% of AI hardware will be at the edge."

Tractica

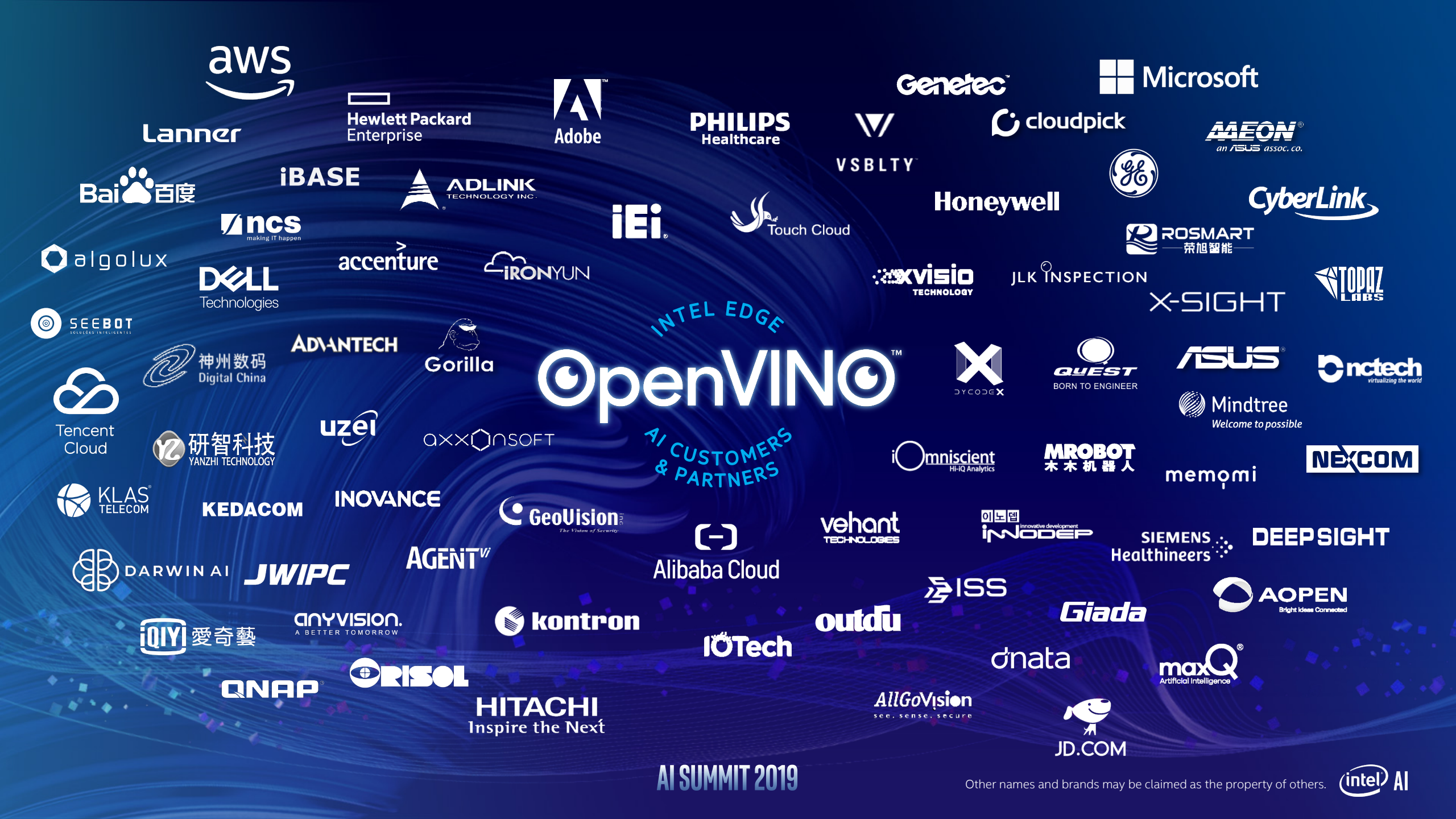
"The success of AI on edge needs clever optimization techniques on limited power."

Gartner

AI SUMMIT 2019

intel AI


intel AI
JONATHAN BALLON
VICE PRESIDENT
INTERNET OF THINGS GROUP

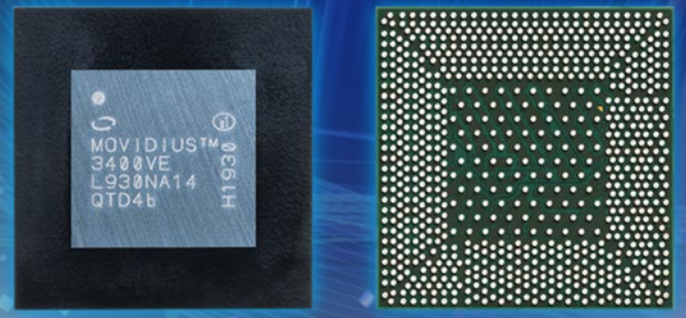

aws
Hewlett Packard Enterprise
Adobe
PHILIPS Healthcare
Genetec
Microsoft
Lanner
VSBLTY
cloudpick
AAEON an ASUS assoc. co.
Baidu 百度
iBASE
ADLINK TECHNOLOGY INC.
iEi
Touch Cloud
Honeywell
CyberLink
ncs making IT happen
ROSMART 荣旭智能
algolux
accenture
IRONYUN
DELL Technologies
XVISIO TECHNOLOGY
JLK INSPECTION
TOPAZ LABS
X-SIGHT
SEEBOT
INTEL EDGE
OpenVINO™
AI CUSTOMERS & PARTNERS
ADVANTECH
Gorilla
神州数码 Digital China
Tencent Cloud
DYCODEX
QUEST BORN TO ENGINEER
ASUS
nctech virtualizing the world
Mindtree Welcome to possible
研智科技 YANZHI TECHNOLOGY
uzei
axxonsoft
iOmniscient Hi-IQ Analytics
MROBOT 未来机器人
NEXCOM
memomi
KLAS TELECOM
KEDACOM
INOVANCE
GeoVision
vehant TECHNOLOGIES
이노뎁 innovative development INODEP
SIEMENS Healthineers
DEEPSIGHT
Alibaba Cloud
DARWIN AI
JWIPC
AGENT VI
ISS
AOPEN Bright Ideas Connected
Giada
iQIYI 愛奇藝
anyvision A BETTER TOMORROW
kontron
outdu
IOTech
dnata
maxQ Artificial Intelligence
QNAP
ORISOL
HITACHI Inspire the Next
AllGoVision see. sense. secure
JD.COM
AI SUMMIT 2019
Other names and brands may be claimed as the property of others.
intel AI

LAUNCHING 1H'20

NEXT-GEN MOVIDIUS™ VPU (KEEM BAY)

BUILT FOR EDGE AI

- DEEP LEARNING INFERENCE + COMPUTER VISION + MEDIA
- FASTER MEMORY BANDWIDTH
- GROUNDBREAKING HIGH-EFFICIENCY ARCHITECTURE
- ACCELERATED WITH OpenVINO™



FLEXIBLE FORM FACTORS

EDGE EXPERIENCES

AI SUMMIT 2019

intel AI

KEEM BAY IS BUILT FOR EDGE AI...

FAST +

4X NVIDIA TX2

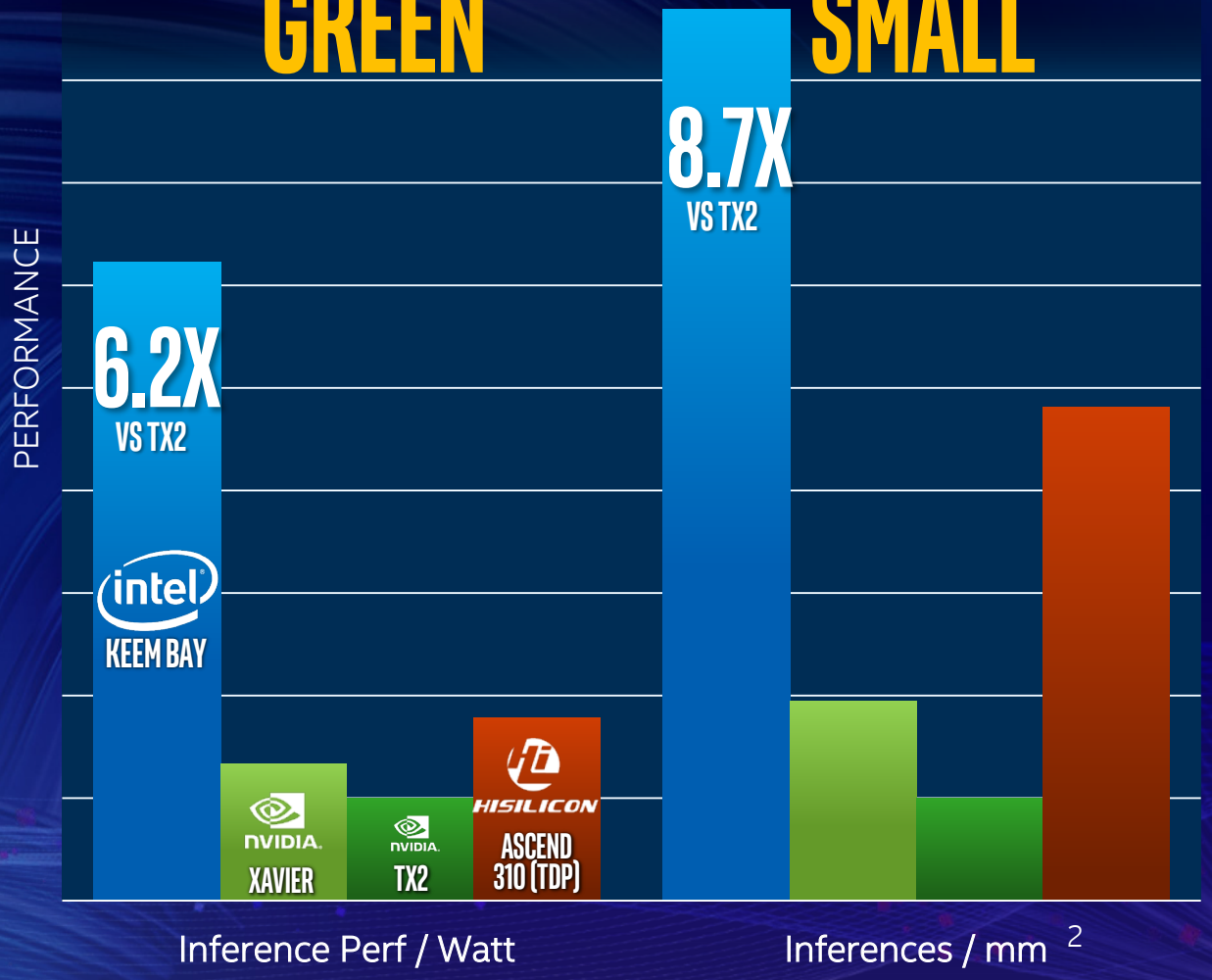
1.25X Ascend 310

vs. NVIDIA Xavier **ON PAR**[1]

@ 1/5TH POWER

GREEN

SMALL



EFFICIENT

4X

Inferences / Sec / TOPS
vs NVIDIA Xavier

The above is preliminary performance data based on pre-production components. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See backup for configuration details.

Comparison of Frames Per Second utilizing Resnet-50, Batch 1.

1. Keem Bay throughput within 10% vs Xavier throughput.

AI SUMMIT 2019

intel AI

OpenVINO
AI INFERENCE SOFTWARE WORKFLOW

OPTIMIZE
TensorFlow
PYTORCH
mxnet
Keras
Caffe
ONNX
OpenVINO

TEST


DEV CLOUD
FOR THE EDGE
LAUNCHING TODAY

DEPLOY
COMING SOON
intel NERVANA inside
intel CORE i7 10TH GEN
OpenVINO
intel XEON inside
intel FPGA
intel MOVIDIUS inside

SCALE


AI SUMMIT 2019

intel AI

ANNOUNCING TODAY

THE FIRST EDGE AI NANODEGREE

WOMEN WHO CODE

SCHOLARSHIPS

AI SUMMIT 2019

intel AI

KEEM BAY DELIVERS EFFICIENT OUTPERFORMANCE

PURPOSE BUILT PORTFOLIO FOR THE EDGE

OPENVINO & DEV CLOUD FOR THE EDGE DEMOCRATIZING AI

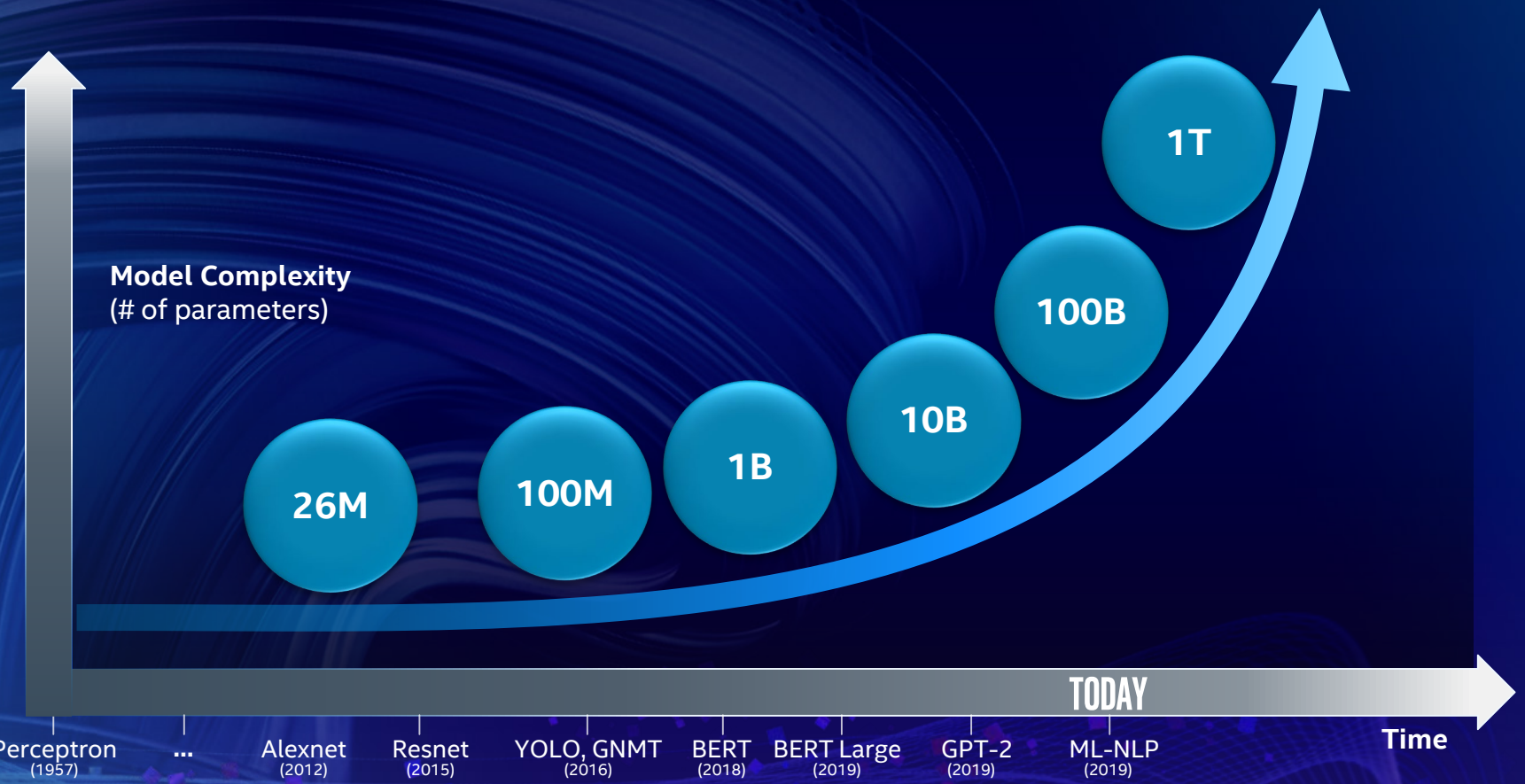
BUILDING THE NEXT GENERATION OF DEVELOPERS UDACITY NANO DEGREE

intel AI

Deep learning models
are quickly growing in complexity,
requiring 2X compute power
every 3.5 months.

Why?

GROWING MODEL COMPLEXITY → RAPIDLY INCREASING COMPUTE



DATA → INFORMATION → KNOWLEDGE

Manipulating Knowledge Effectively Will Be THE Compute Problem

intel AI

THE CONTINUUM OF INTELLIGENCE...

WHAT IS YOUR FAVORITE FOOD?

"Chicken soup pizza is a dish food around forks with food."

LIMITED NEURAL NETWORKS

"Pizza with pepperoni and salad. How about you?"

BIGGER NETWORKS THAT PUSH TODAY'S LIMITS OF COMPUTE

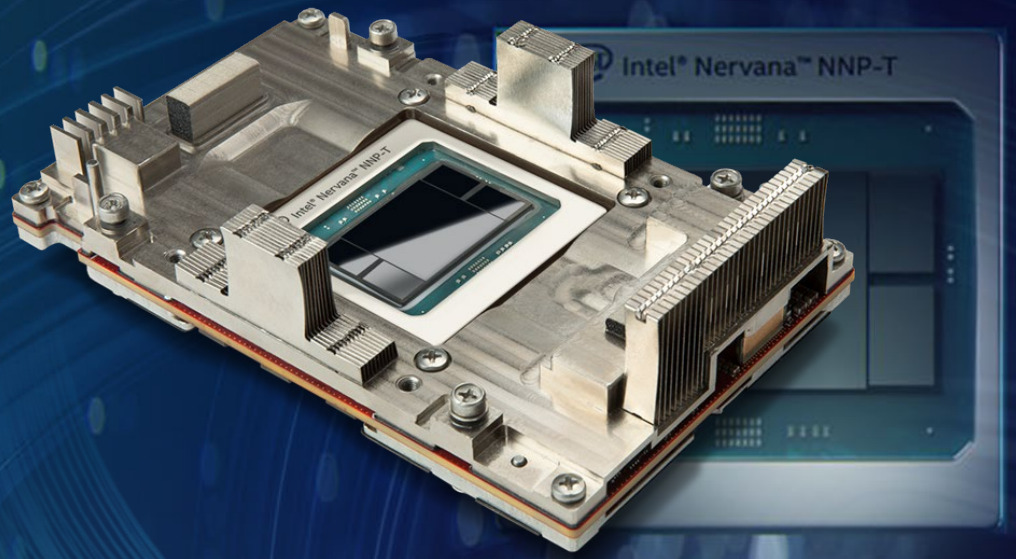
"Sushi, especially toro, because of its exquisite mouthfeel. I used to be averse to raw fish until I first experienced the majesty of Nobu Matsuhisa's first restaurant in Beverly Hills."

WHAT'S POSSIBLE WITH LARGER, MORE COMPLEX MODELS

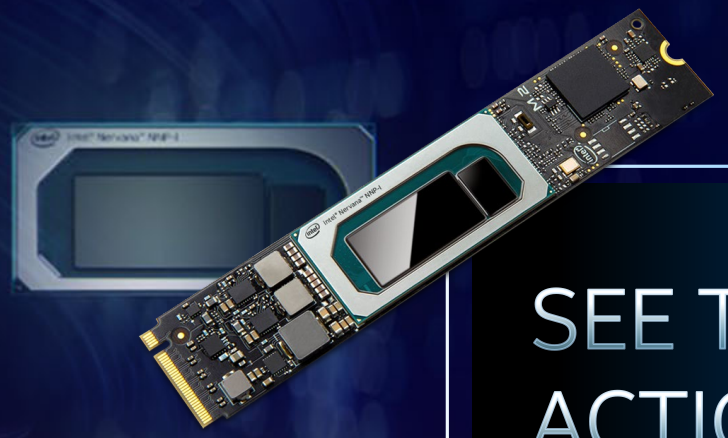
intel AI

The next leap forward means not looking back.

INTEL® NERVANA™ NEURAL NETWORK PROCESSOR FAMILY



INTEL® NERVANA™ NNP-T



INTEL® NERVANA™ NNP-I

SEE THEM IN ACTION FOR THE FIRST TIME TODAY.

Intel & Nervana
Innovation & Expertise
Hardware Revolution

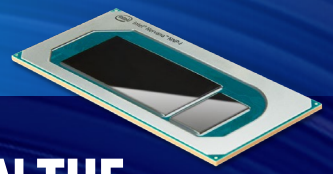
AI SUMMIT 2019

intel AI

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

In production 2019

Incredibly efficient inference scaling for diverse latency and power needs across multiple topologies.



"50 TRILLION CALCS/SEC IN THE PALM OF YOUR HAND"

FORTUNE

EXPECT PERF/WATT LEADERSHIP AT LAUNCH FOR COMMERCIALLY AVAILABLE ACCELERATORS

DENSITY LEADERSHIP

OPEN FULL-STACK SW

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance. Performance claims calculated per node based on Intel and Nvidia submissions to MLPerf Inference v0.5 results published on November 6, 2019 at https://mlperf.org/inference-results/. Measurements based on Intel internal testing and benchmarking using pre-production hardware/software as of October 2019. For more complete information visit intel.ai/benchmarks. All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice. Configuration d **intel.ai/benchmarks** MLPerf v0.5 Inference Closed ResNet-v1.5 Offline, entry Inf-0.5-33.; MLPerf v0.5 Inference Closed ResNet-v1.5 Offline, entry Inf-0.5-25.; MLPerf v0.5 Inference Closed ResNet-v1.5 Offline, entry Inf-0.5-21.

1. Based on results published at https://mlperf.org/inference-results/

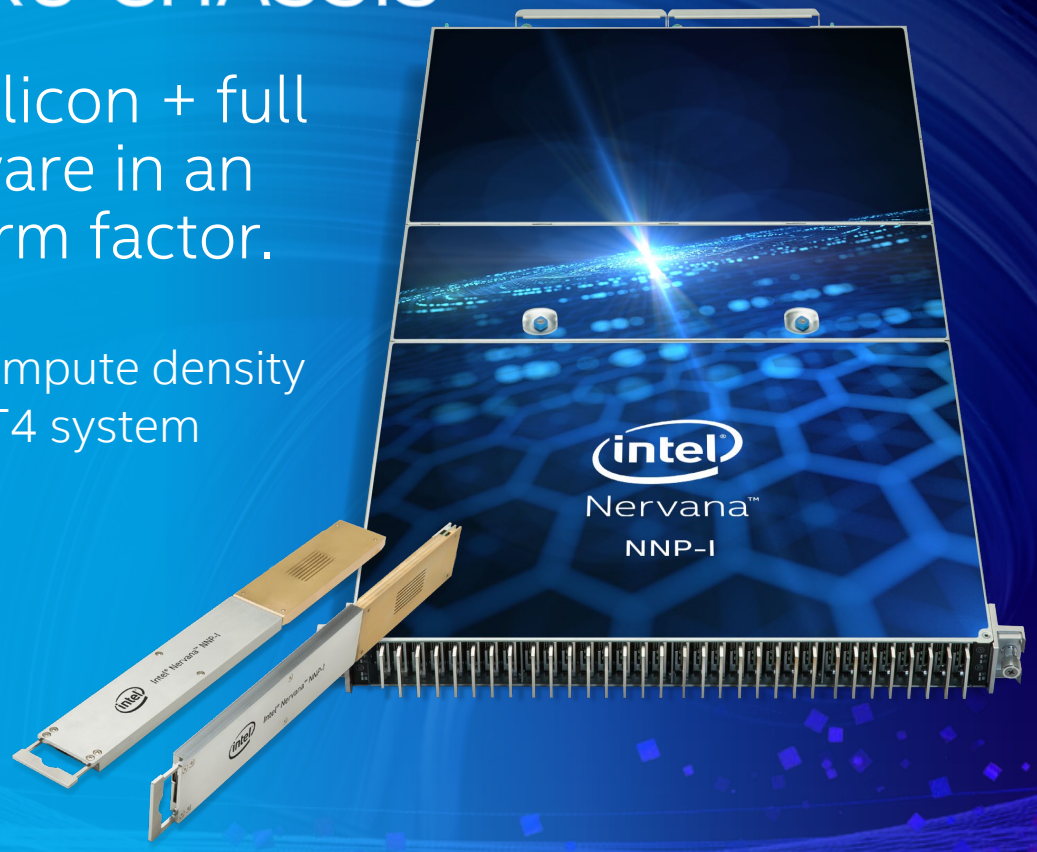
AI SUMMIT 2019

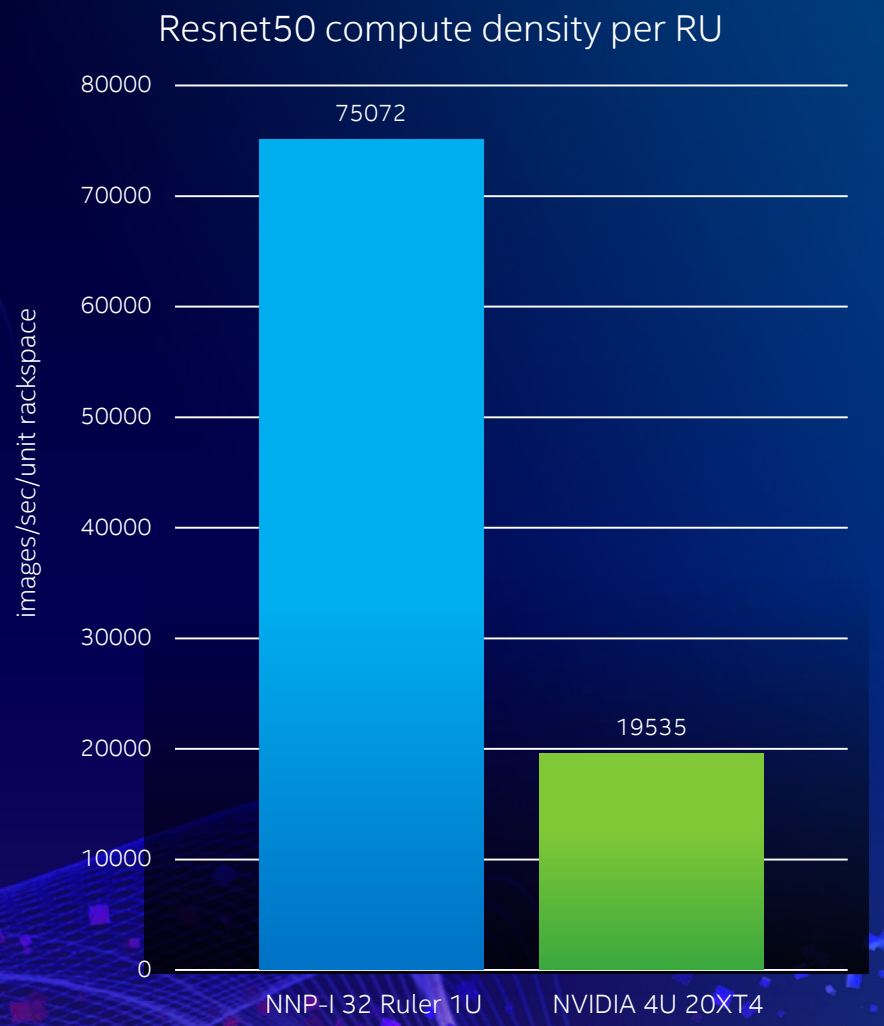
intel AI

SINGLE RU CHASSIS

Pre-prod silicon + full stack software in an industry form factor.

Up to 3.7X compute density over NVIDIA T4 system





Measurements based on Intel internal testing and benchmarking using pre-production hardware/software as of October 2019. For more complete information visit intel.ai/benchmarks. All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice. Configuration details at **intel.ai/benchmarks**
Other names and brands may be claimed as the property of others.

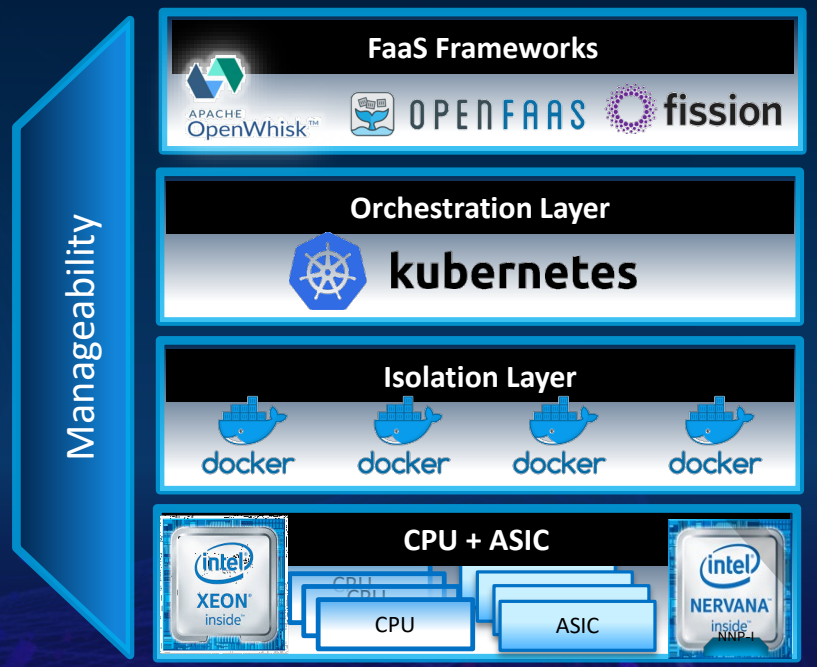
AI SUMMIT 2019

intel AI

DATA CENTER-READY INFERENCE

SEAMLESS CLOUD-NATIVE INFERENCE AT SCALE

- Kubernetes device plugin and management interfaces
- NNPI-enabled containers for ease of development and deployment
- Full reference solution stack with emerging deployment models like FaaS/CaaS



intel AI

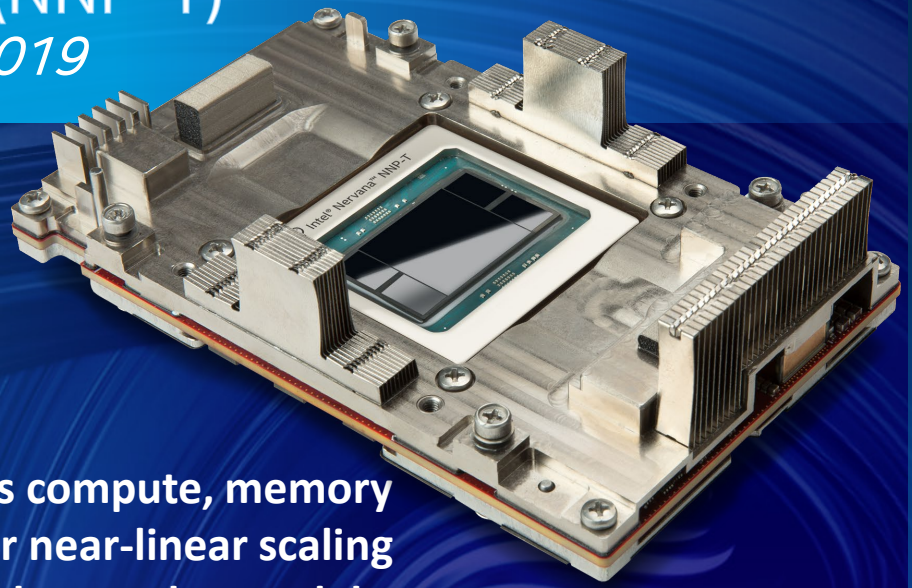
“We are excited to be working with Intel to deploy faster and more efficient inference compute with the Intel® Nervana™ NNP-I and to extend support for our state-of-the-art deep learning compiler, Glow, to the NNP-I.”

MISHA SMELYANSKIY
DIRECTOR, AI

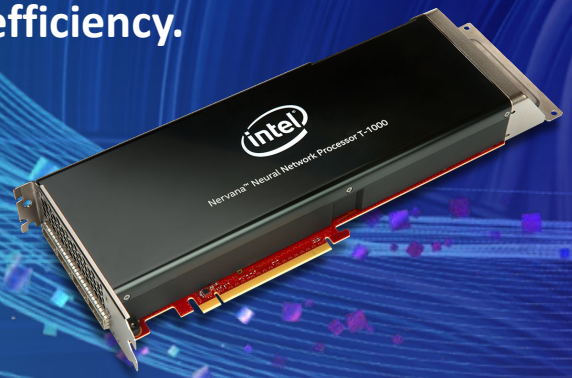
intel AI

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

In production 2019



Carefully balances compute, memory & interconnect for near-linear scaling to train increasingly complex models at high efficiency.



REAL-WORLD READINESS

- Industry-leading scaling, up to 95%on ResNet 50 and BERT, *with* State-Of-The-Art accuracy[1, 2]
  - *Competition observed at 73%*
- Highly energy-efficient solution
- Same data rate on 8 or 32 cards[3]
- Scale well beyond 32 cards
- Glueless fabric for high- performing systems at significant cost savings[4]

1. Measurements based on Intel internal testing using pre-production 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.
2. Accuracy target as referenced in MLPerf Link: https://github.com/mlperf/training/tree/master/image_classification
3. NNP-T Performance measured on pre-production NNP-T1000 silicon, using 22 TPCs at 900MHz core clock and 2GHz HBM clock, Host is an Intel® Xeon® Gold 6130T CPU @ 2.10GHz with 64 GB of system memory
4. No additional switching and NIC costs required.

For more complete information about performance results, visit www.intel.ai/benchmarks. Other names and brands may be claimed as the property of others.

POD REFERENCE DESIGN

GLUELESS PEER-TO-PEER SCALING FABRIC

CHIP TO CHIP

CHASSIS TO CHASSIS

RACK TO RACK

NO OTHER SWITCH REQUIRED

WELCOME TO A NEW INTELLIGENCE

DR. KENNETH CHURCH
BAIDU AI RESEARCH FELLOW

FOLLOW UP ON ANNOUNCEMENT FROM BAIDU CREATE

- This July in Beijing, Baidu and Intel announced collaboration on the Intel® Nervana™ NNP-T
- Enhancing hardware and software designs of the new purpose-built product to train increasingly complex models at maximum efficiency

intel AI

PaddlePaddle

Platform

AI Platform & Ecosystem

Cognition

Language & Knowledge

Perception

Speech

Vision

AR/VR

Infrastructure

Data

Algorithm

Compute

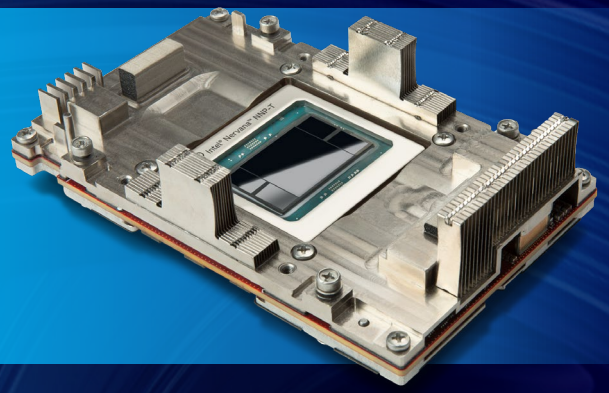
AI Security

NNP-T / X-Man

intel AI

BAIDU X-MAN 4.0 ACCELERATING NNP-T TO MARKET

Intel NNP-T



Baidu X-Man 4.0

32 NNP-T/Rack
(actual photo)

intel AI

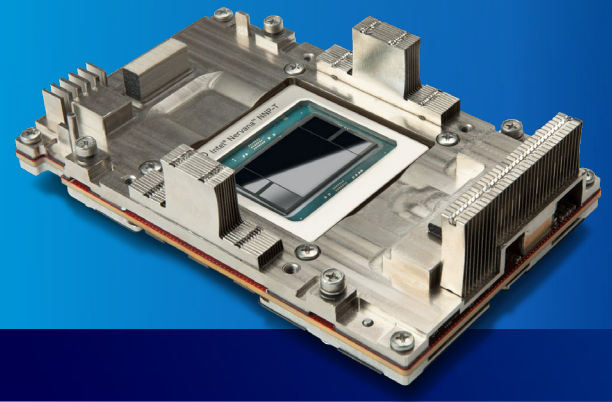
MADE POSSIBLE BY INDUSTRY COLLABORATION

Open Accelerator Infrastructure (OAI)

Sub-project within OCP Server Project

OPEN Compute Project®

Intel NNP-T
(OAM-compliant)



Baidu X-Man 4.0
(OAI-compliant)

intel AI

飞桨 PaddlePaddle

User-Friendly DL Framework

Cross-Modal Universal Semantic Representation

High-Performance Inference

Large-Scale Training

NNP-T

Next-gen

ERNIE model → 5X speedup

CONTINUED OPTIMIZATION

AI SUMMIT 2019

intel AI

PaddlePaddle

User-Friendly DL Framework

High-Performance Inference

Large-Scale Training

Cross-Modal Universal Semantic Representation

NNP-T

Next-gen

intel OPTANE DC PERSISTENT MEMORY

intel OPTANE DC SOLID STATE DRIVE

intel 3D NAND SSD

intel SGX

CONTINUED OPTIMIZATION

AI SUMMIT 2019

intel AI


AI AT INTEL
SCALE AND EFFICIENCY,
FOR AI EVERYWHERE
AI SUMMIT 2019
intel AI

DISCLAIMER

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.

Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

For more information go to www.intel.com/benchmarks. Performance results are based on testing as of Oct 31, 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

Product	Intel Keem Bay VPU	NVIDIA Jetson TX2	Huawei Atlas 200 (Ascend 310)	NVIDIA Xavier AGX
Testing as of	10/31/2019	10/30/19	8/25/19	10/22/19
Precision	INT8	FP16	INT8	INT8
Batch Size	1	1	1	1
Sparsity	50% weight sparsity	N/A	N/A	N/A
Product Type	Keem Bay EA CRB Dev kit (preproduction)	Jetson Developer kit	Atlas 200 Developer kit	Jetson Developer kit
Mode	N/A	nvpmodel 0 Fixed Freq	N/A	nvpmodel 0 Fixed Freq
Memory	4GB	8GB	8GB	16GB
Processor	ARM* A53 x 4	ARM*v8 Processor rev 3 (v8l) × 4	ARM* A53 x 8	ARM*v8 Processor rev 0 (v8l) × 2
Graphics	N/A	NVIDIA Tegra X2 (nvgpu)/integrated	N/A	NVIDIA Tegra Xavier (nvgpu)/integrated
OS	Ubuntu 18.04 Kernel 1.18 (64-bit) on Host Yocto Linux 5.3.0 RC8 on KMB	Ubuntu 18.04 LTS (64-bit)	Ubuntu 16.04	Ubuntu 18.04 LTS (64-bit)
Hard Disk	N/A	32GB	32GB	32GB
Software	Performance demo firmware	JetPack: 4.2.2	MindSpore Studio, DDK B883	JetPack: 4.2.1
Listed TDP	N/A	10W	20W	30W

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. Check with your system manufacturer or retailer or learn more at www.intel.com.