WELCOME TO A NEW INTELLIGENCE
AI DELIVERED AT INTEL SCALE

NAVEEN RAO
CORPORATE VICE PRESIDENT
GM, ARTIFICIAL INTELLIGENCE
$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

CPU
Multi-Purpose, Foundation for Analytics & AI

GPU
Data-Parallel Media, Graphics, HPC & AI

FPGA
Real-Time & Multi-Function Inference

ASIC

Built-in security

AI-Specific

CPU

GPU

FPGA

ASIC

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.
30X INFERENC 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

Up to 30x AI performance improvement with Intel® Deep Learning Boost (Intel DL Boost) compared to Intel® Xeon® Platinum 8180 processor (July 2017).
Up to 4.3x higher inference throughput on ResNet 50 vs. AMD Ryzen 7 3700U

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.
AI WILL INFUSE EVERYTHING...

...SO WE PUT IT EVERYWHERE

Workload breadth

CPU
Multi-Purpose, Foundation for Analytics & AI

GPU
Data-Parallel Media, Graphics, HPC & AI

FPGA
Real-Time & Multi-Function Inference

AI-Specific

ASIC

Built-in security

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.
ADJACENT TECHNOLOGIES TO MOST INTELLIGENTLY FEED COMPUTE
Adjacent technologies to most intelligently feed compute.
ADJACENT TECHNOLOGIES TO MOST INTELLIGENTLY FEED COMPUTE
ADJACENT TECHNOLOGIES TO MOST INTELLIGENTLY FEED COMPUTE
ADJACENT TECHNOLOGIES TO MOST INTELLIGENTLY FEED COMPUTE
ADJACENT TECHNOLOGIES TO MOST INTELLIGENTLY FEED COMPUTE
ADJACENT TECHNOLOGIES TO MOST INTELLIGENTLY FEED COMPUTE
Open software to keep hardware nimble and working better together.

**Open**
Providing flexibility to advance a fast-moving landscape and embrace community innovation.

**Complete**
Access to kernel, compiler, and frameworks, for developers to work how they want.

**Innovative**
Tools to push boundaries for next-level AI.

Other names and brands may be claimed as the property of others.
As model complexity grows, accelerate with purpose

Model Complexity
(# of parameters)

- Perceptron (1957)
- Alexnet (2012)
- Resnet (2015)
- YOLO, GNMT (2016)
- BERT (2018)
- BERT Large (2019)
- GPT-2 (2019)
- ML-NLP (2019)
- BERT Large (2019)

- 26M
- 100M
- 1B
- 10B
- 100B
- 1T

Exponential growth
~10X per year

Next-gen DL
Increasingly complex models & data

Bigger dataset
Bigger models
Bigger compute

Other names and brands may be claimed as the property of others.
THE EDGE OPPORTUNITY...AND CHALLENGE


- 75% of AI hardware will be at the edge.
- The success of AI on edge needs clever optimization techniques on limited power.

Cameras grow at highest CAGR.
Other names and brands may be claimed as the property of others.
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

LAUNCHING 1H'20

NEXT-GEN MOVIDIUS™ VPU (KEMM BAY)
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.

4X Inferences / Sec / TOPS vs NVIDIA Xavier

1.25X Ascend 310

vs. NVIDIA Xavier ON PAR\(^1\) at 1/5\(^{1\text{TH}}\) POWER

6.2X vs TX2

8.7X vs TX2

Keem Bay throughput within 10% vs Xavier throughput.
ANNOUNCING TODAY
THE FIRST EDGE AI NANODEGREE

WOMEN WHO CODE SCHOLARSHIPS
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
Deep learning models are quickly growing in complexity, requiring 2X compute power every 3.5 months. Why?
GROWING MODEL COMPLEXITY → RAPIDLY INCREASING COMPUTE

Model Complexity
(# of parameters)

Perceptron (1957)
Alexnet (2012)
Resnet (2015)
YOLO, GNMT (2016)
BERT (2018)
BERT Large (2019)
GPT-2 (2019)
ML-NLP (2019)

Other names and brands may be claimed as the property of others.
THE CONTINUUM OF INTELLIGENCE...

WHAT IS YOUR FAVORITE FOOD?

“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

“Pizza with pepperoni and salad. How about you?”

BIGGER NETWORKS THAT PUSH TODAY’S LIMITS OF COMPUTE

“Chicken soup pizza is a dish food around forks with food.”

LIMITED NEURAL NETWORKS

Other names and brands may be claimed as the property of others.
The next leap forward means not looking back.
INTEL® NERVANA™ NEURAL NETWORK PROCESSOR FAMILY

Intel & Nervana
Innovation & Expertise
Hardware Revolution

SEE THEM IN ACTION FOR THE FIRST TIME TODAY.

AI SUMMIT 2019
Incredibly efficient inference scaling for diverse latency and power needs across multiple topologies.

50 trillion calcs/sec in the palm of your hand

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, 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 of 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/
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.
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
“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 SMELYANSKY
DIRECTOR, 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](https://github.com/mlperf/training/tree/master/image_classification)
3. NNP-T Performance measured on pre-production NNP-T1000 silicon, using 22 TPUCs at 600Mhz 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](http://www.intel.ai/benchmarks). Other names and brands may be claimed as the property of others.
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

Other names and brands may be claimed as the property of others.
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.
BAIDU X-MAN 4.0 ACCELERATING NNP-T TO MARKET

Intel NNP-T

Baidu X-Man 4.0

32 NNP-T/Rack
(actual photo)
Open Accelerator Infrastructure (OAI)
Sub-project within OCP Server Project

MADE POSSIBLE BY INDUSTRY COLLABORATION

Intel NNP-T
(OAM-compliant)

Baidu X-Man 4.0
(OAI-compliant)

Other names and brands may be claimed as the property of others.
PaddlePaddle

User-Friendly DL Framework

High-Performance Inference

Large-Scale Training

ERNIE model $\rightarrow$ 5X speedup

CONTINUED OPTIMIZATION
PaddlePaddle

User-Friendly DL Framework

High-Performance Inference

Large-Scale Training

Cross-Modal Universal Semantic Representation

Other names and brands may be claimed as the property of others.
AI AT INTEL
SCALE AND EFFICIENCY,
FOR AI EVERYWHERE
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.

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.

Intel, the Intel logo, Xeon™ and Movidius™ are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries. *Other names and brands may be claimed as the property of others.

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

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.

Intel, the Intel logo, Xeon™ and Movidius™ are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries. *Other names and brands may be claimed as the property of others.