intel AI

DEMOCRATIZING AI

Carlos Morales, Intel AI Products Group

Compute for Everyone

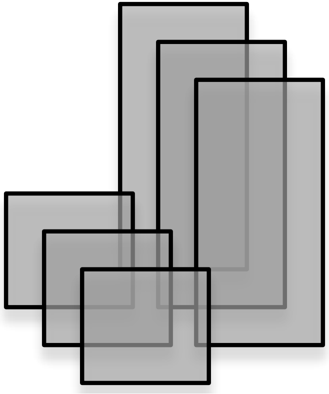
Intel has a long tradition of democratizing compute

by

Making it **easier**

Making it **powerful**

Making it **accessible**



Mainframes
a few thousand users

Compute for Everyone

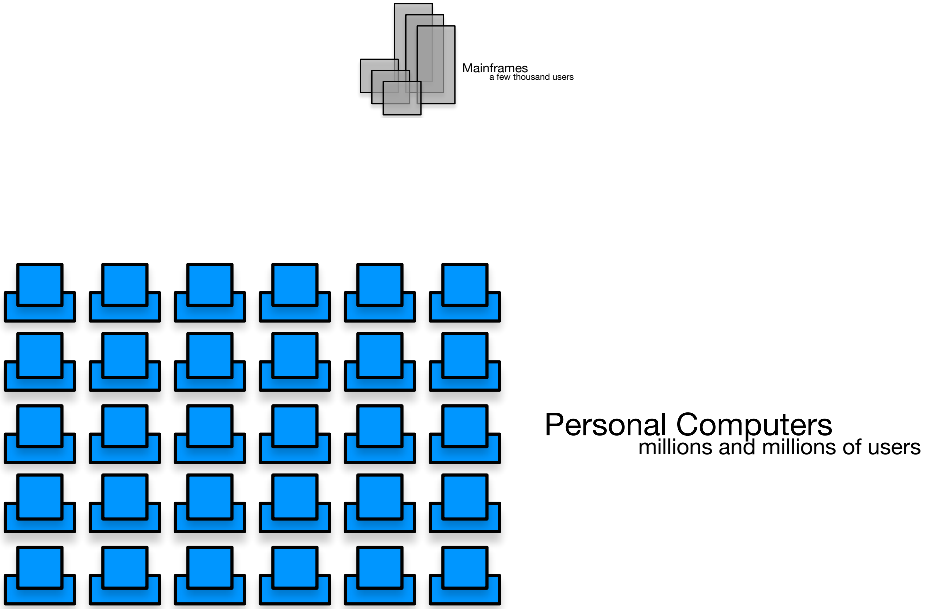
Intel has a long tradition of democratizing compute

by

Making it **easier**

Making it **powerful**

Making it **accessible**



Mainframes
a few thousand users

Personal Computers
millions and millions of users

Compute for Everyone

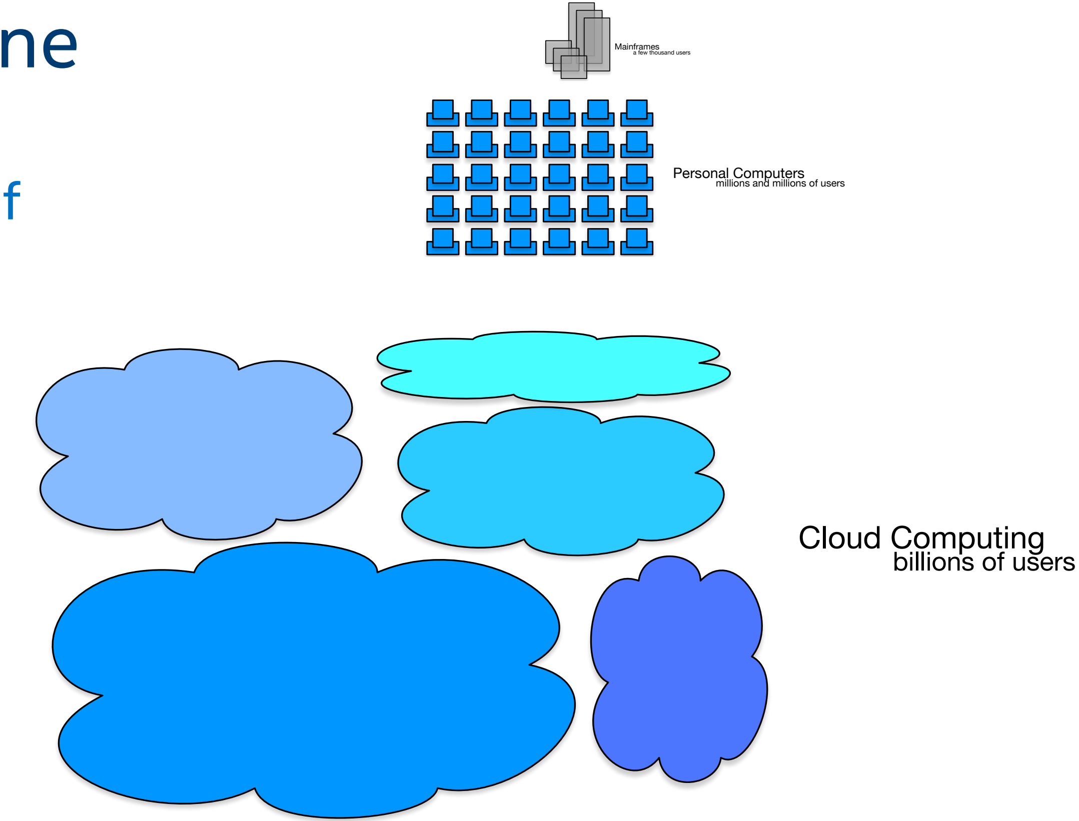
Intel has a long tradition of democratizing compute

by

Making it **easier**

Making it **powerful**

Making it **accessible**



Compute for Everyone

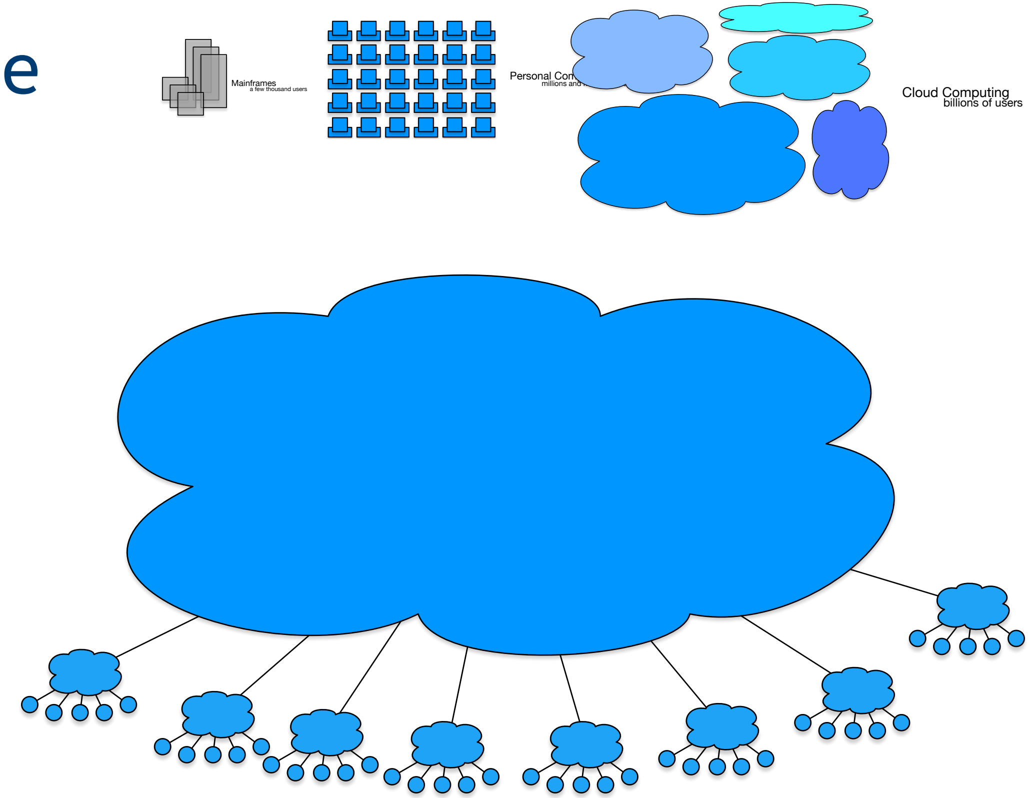
Intel has a long tradition of democratizing compute

by

Making it **easier**

Making it **powerful**

Making it **accessible**



Fog and IoT Computing
so many billions of users

Compute for Everyone

Intel has a long tradition of democratizing compute

by

Making it **easier**

Making it **powerful**

Making it **accessible**

but

What does

Democratizing AI

actually mean?

Democratizing AI

What does that actually mean?

Making it **easier**

by

Automating and abstracting anything that is not AI

Making it **powerful**

Making it **accessible**

Democratizing AI

What does that actually mean?

Making it **easier**

Making it **powerful**

Making it **accessible**

by

Automating and abstracting anything that is not AI

Enabling scale up, scale out and novel AI techniques for *everyone*

Democratizing AI

What does that actually mean?

Making it **easier**

Automating and abstracting anything that is not AI

Making it **powerful**

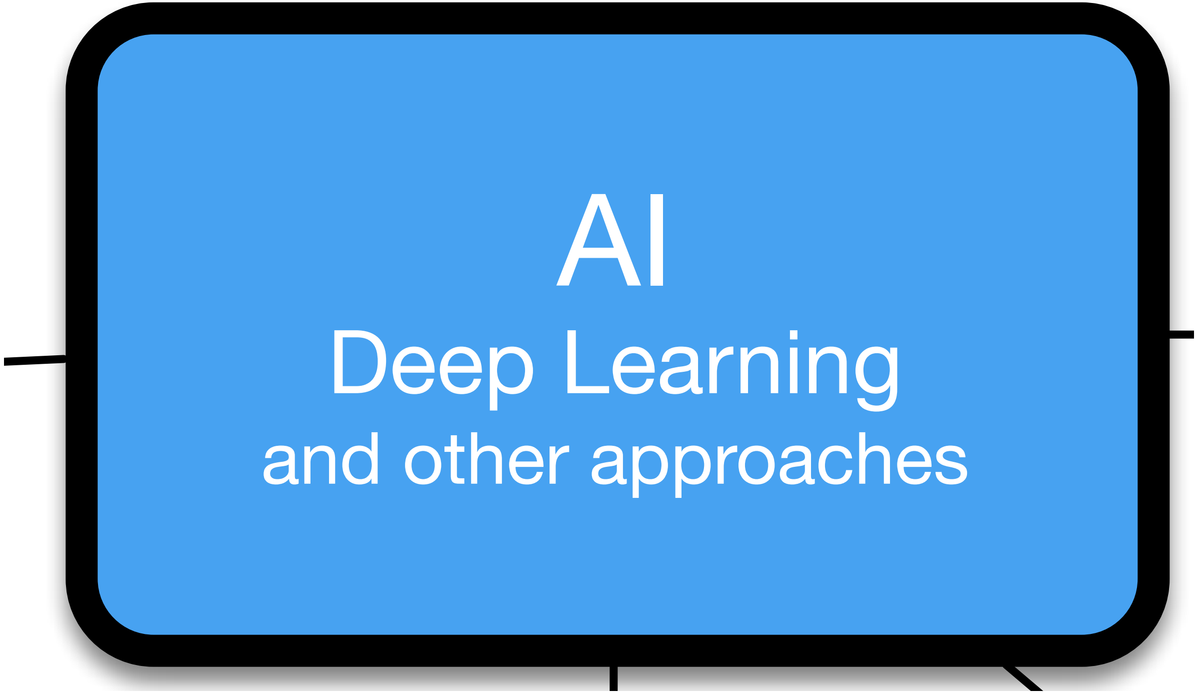
Enabling scale up, scale out and novel AI techniques for *everyone*

Making it **accessible**

by

Bringing it to the compute platform you already have

Making AI Easier *by* **Automating** and abstracting anything that is not AI

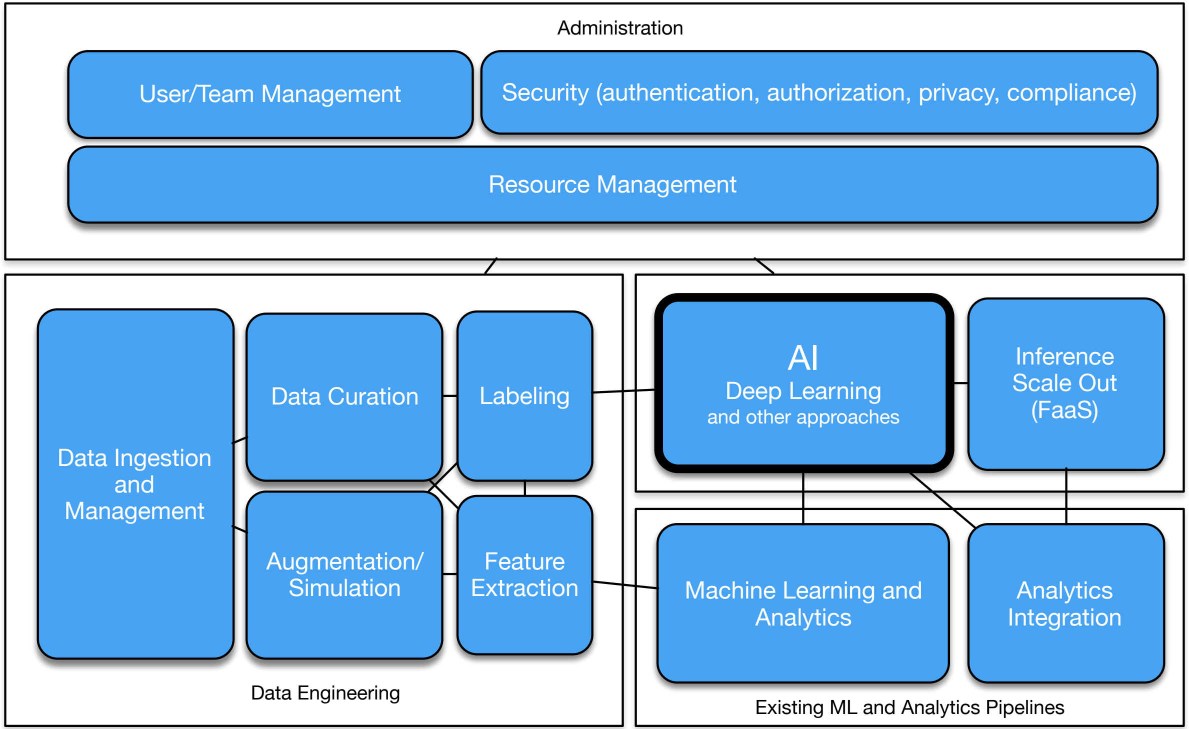


AI

Deep Learning

and other approaches

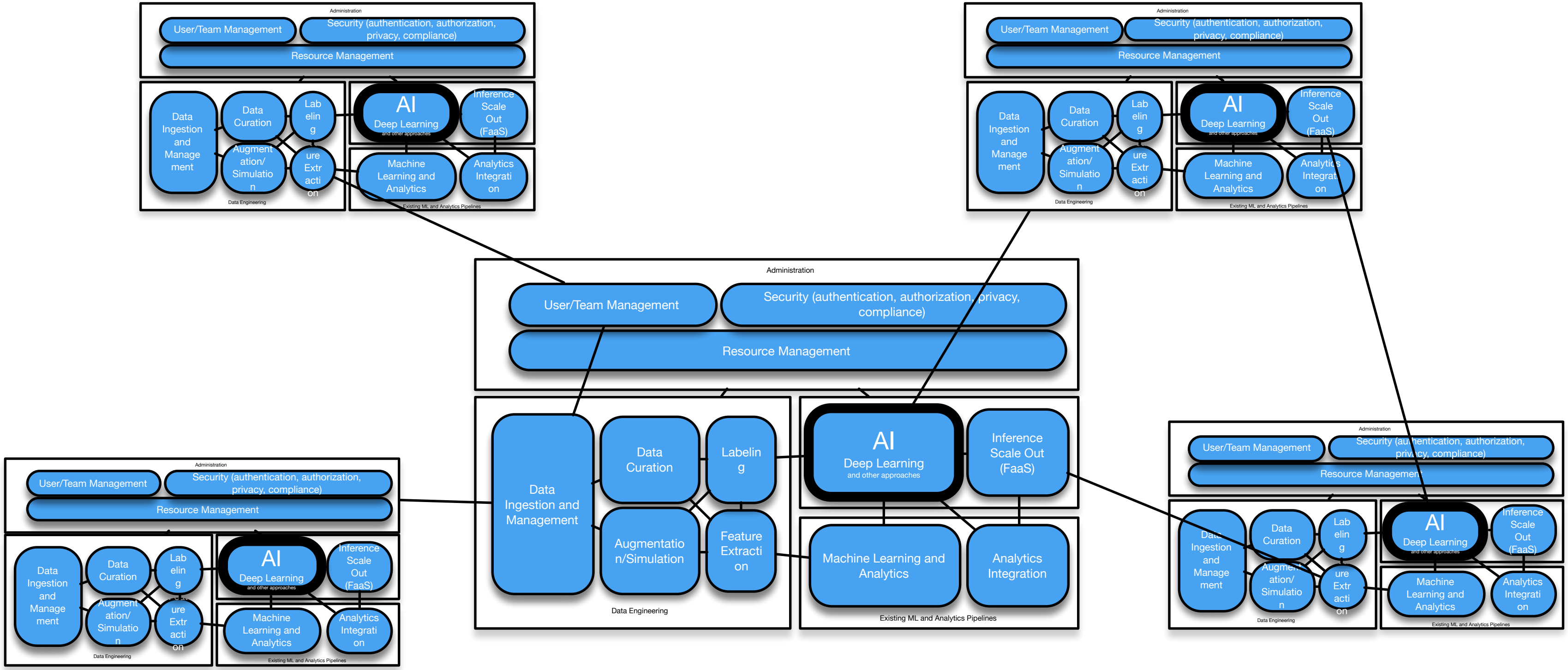
Making AI Easier *by* **Automating** and abstracting anything that is not AI



Making AI Easier

by

Automating and abstracting anything that is not AI



Administration

User/Team Management

Security (authentication, authorization, privacy, compliance)

Resource Management

Data Ingestion and Management

Data Curation

Labeling

Augmentation/Simulation

Feature Extraction

Data Engineering

AI

Deep Learning

and other approaches

Inference Scale Out (FaaS)

Machine Learning and Analytics

Analytics Integration

Existing ML and Analytics Pipelines

Making AI Easier *by* **Automating** and abstracting anything that is not AI

How do we solve messy problem?

Making AI Easier *by* **Automating** and abstracting anything that is not AI

How do we solve messy problem?

The open source community, with Intel's support, is converging on solutions.

DLaaS offerings are flourishing

Kubernetes is the API

Democratizing AI

What does that actually mean?

Making it **easier**

Making it **powerful**

Making it **accessible**

by

Automating and abstracting anything that is not AI

Enabling scale up, scale out and novel AI techniques for *everyone*

Bringing it to the compute platform you already have

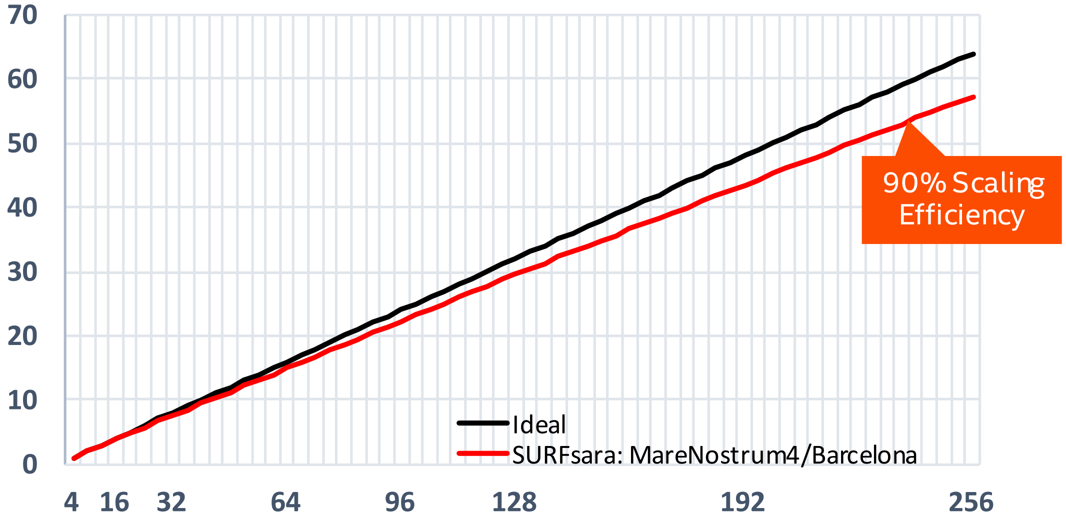
Making it Powerful

Xeon Democratizes AI

by

Enabling scale up, scale out and novel AI techniques for *everyone*

Intel® - SURFsara* Research Collaboration - Multi-Node Intel® Caffe ResNet-50 Scaling Efficiency on 2S Intel® Xeon® Platinum 8160 Processor Cluster



- MareNostrum4 Barcelona Supercomputing Center
- ImageNet-1K
- 256 nodes
- 90% scaling efficiency
- Top-1/Top-5 > 74%/92%
- Batch size of 32 per node
- Global BS=8192
- Throughput: 15170 Images/sec

Time-To-Train: 70 minutes (50 Epochs)

Configuration Details 2:Slide 127

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system. 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 complete information visit: http://www.intel.com/performance Source: Intel measured as of June 2017

Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Xeon Democratizes AI

INFERENCE THROUGHPUT

Up to

198x

Intel® Xeon® Platinum 8180 Processor
higher Intel optimized Caffe GoogleNet v1 with Intel® MKL
inference throughput compared to
Intel® Xeon® Processor E5-2699 v3 with BVLC-Caffe

TRAINING THROUGHPUT

Up to

127x

Intel® Xeon® Platinum 8180 Processor
higher Intel Optimized Caffe AlexNet with Intel® MKL
training throughput compared to
Intel® Xeon® Processor E5-2699 v3 with BVLC-Caffe

Intel® Xeon® Platinum 8180 Processor higher Intel optimized Caffe Resnet50 with Intel® MKL inference throughput 133X and training throughput 73X compared to Intel® Xeon® Processor E5-2699 v3 with BVLC-Caffe

Inference and training throughput measured with FP32 instructions. Inference performance with INT8 is expected to be higher

AI performance is constantly improving with hardware and software optimizations on Intel® Xeon® Scalable Processors

INFERENCE using FP32 Batch Size Caffe GoogleNet v1 256 AlexNet 256 Configuration Details on Configs: 18, 25
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 complete information visit: http://www.intel.com/performance Source: Intel measured as of June 2017 Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Xeon Democratizes AI: Case Study

GE Healthcare

Intel's Solution Stack includes

Intel® Xeon® Scalable processors

Intel® Solid State Drives

Intel Deep Learning Deployment Toolkit

Intel® Math Kernel Library for Deep Neural Networks

OPTIMIZED MODEL

Exceeds GE Inferencing Target

14x FASTER | **5.9x** ABOVE TARGET

Democratizing AI

What does that actually mean?

Making it **easier**

Making it **powerful**

Making it **accessible**

by

Automating and abstracting anything that is not AI

Enabling scale up, scale out and novel AI techniques for *everyone*

Bringing it to the compute platform you already have

Democratizing AI

Making it **accessible**

by

Bringing it to the compute platform you already have

Optimizing Xeon AI

Augmenting Xeon with a broad compute portfolio

Enabling End-to-end AI

And most importantly

Making it easier to leverage the full stack

intel® AI PORTFOLIO

SOLUTIONS

Data Scientists

Technical Services

Reference Solutions

PLATFORMS

Intel® AI DevCloud

Intel® Deep Learning

Cloud	System*

intel Saffron™

REASONING

TOOLS

Intel® Deep Learning Studio^

Intel® Deep Learning Deployment Toolkit

Intel® Computer Vision SDK

Intel® Movidius™ SDK

FRAMEWORKS

TensorFlow

Caffe

mxnet

Caffe2*

PYTORCH*

CNTK*

PaddlePaddle*

BigDL ON Spark

neon

LIBRARIES

Intel® MKL/MKL-DNN, clDNN, DAAL, Intel Python Distribution, etc.

DIRECT OPTIMIZATION

Intel® nGraph™ Library

CPU Transformer	NNP Transformer†	More...

TECHNOLOGY

intel ARRIA 10 inside · intel XEON inside

DATACENTER

intel CORE i7 inside · intel ATOM inside · Movidius

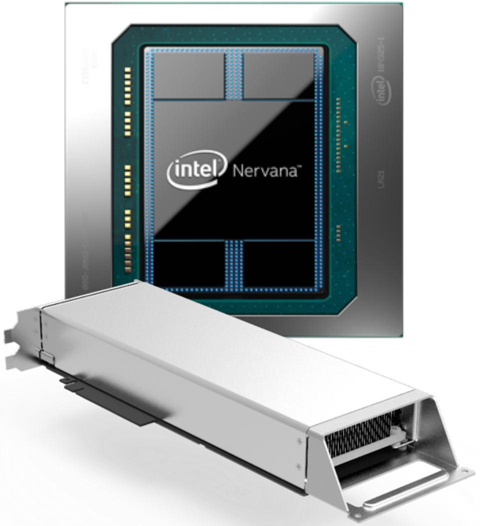
EDGE/GATEWAY

SYSTEMS & COMPONENTS

*Future product
†Beta available
^Available in the Intel® Deep Learning Cloud, coming to other platforms later
Other names and brands may be claimed as the property of others.

INTEL® NERVANA™ NEURAL NETWORK PROCESSOR (NNP)¥

Scalable acceleration with best performance for intensive deep learning



PARALLELISM

Massively-parallel compute

Specialized on-die fabrics

Optimized numerics - Flexpoint

SCALABILITY

Large on-die memory

High speed interconnects

Massive inter-chip data transfer

UTILIZATION

Direct SW control for best on-chip memory usage

Managed data-flow paths

ROADMAP

First silicon in 2017

Product roadmap on track to exceed performance goal[1]

¥ Formerly codenamed as the Crest Family

[1] Source: https://newsroom.intel.com/news-releases/intel-ai-day-news-release/?_ga=2.26542141.1088441208.1508441324-198894050.1498491572.

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

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 complete information visit: http://www.intel.com/performance. Source: Intel measured or estimated as of November 2017

PROJECT BRAINWAVE FOR REAL-TIME AI

"A major leap forward in both performance and flexibility for cloud-based serving of deep learning models."

Doug Burger
Distinguished Engineer

Microsoft

INTEL IS DEMOCRATIZING AI

INTEL IS DEMOCRATIZING AI

by

Offering edge-to-edge AI compute solutions

Developing key AI software with the open source community

and

Making it work better together

intel® AI

Thank you!

Notices and Disclaimers

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 complete information visit: **http://www.intel.com/performance**.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit **www.intel.com/benchmarks**.

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 computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel, the Intel logo, Xeon, Xeon Phi and Nervana are trademarks of Intel Corporation in the U.S. and/or other countries.

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