



INVEST

# THE RACE FOR FASTER MACHINE LEARNING – INTEL ARTIFICIAL INTELLIGENCE TECHNICAL UPDATE

Robert Adamski

PCSS May 2017

OPTIMIZE

ACCELERATE

# Legal Notices and Disclaimers

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY RELATING TO SALE AND/OR USE OF INTEL PRODUCTS, INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT, OR OTHER INTELLECTUAL PROPERTY RIGHT. Intel products are not intended for use in medical, life-saving, life-sustaining, critical control or safety systems, or in nuclear facility applications.

Intel products 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 may make changes to dates, specifications, product descriptions, and plans referenced in this document at any time, without notice.

This document may contain information on products in the design phase of development. The information herein is subject to change without notice. Do not finalize a design with this information.

Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them.

Intel Corporation or its subsidiaries in the United States and other countries may have patents or pending patent applications, trademarks, copyrights, or other intellectual property rights that relate to the presented subject matter. The furnishing of documents and other materials and information does not provide any license, express or implied, by estoppel or otherwise, to any such patents, trademarks, copyrights, or other intellectual property rights.

Wireless connectivity and some features may require you to purchase additional software, services or external hardware.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations

Intel, the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

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

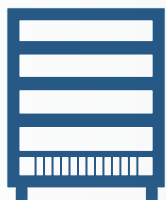
Copyright © 2016 Intel Corporation. All rights reserved.

# Agenda

- Why AI?
- Intel AI portfolio
- MKL-DNN
- Reinforcement Learning on IA
  - Atari Games experiment on Xeon/Xeon Phi
  - Environment - Open AI Gym, PLGRID



# THE NEXT BIG WAVE OF COMPUTING



MAINFRAMES



STANDARDS-BASED SERVERS



CLOUD COMPUTING

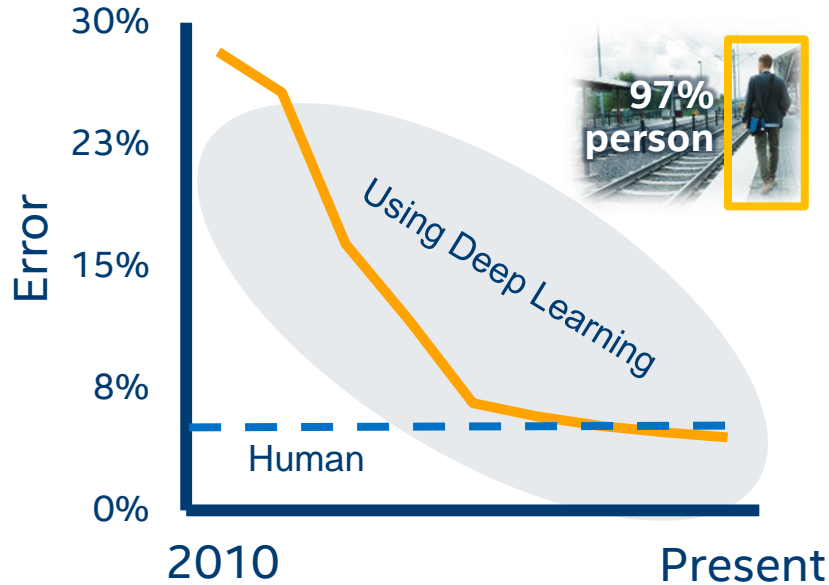
- ✓ DATA DELUGE
- ✓ COMPUTE BREAKTHROUGH
- ✓ INNOVATION SURGE

**ARTIFICIAL  
INTELLIGENCE**

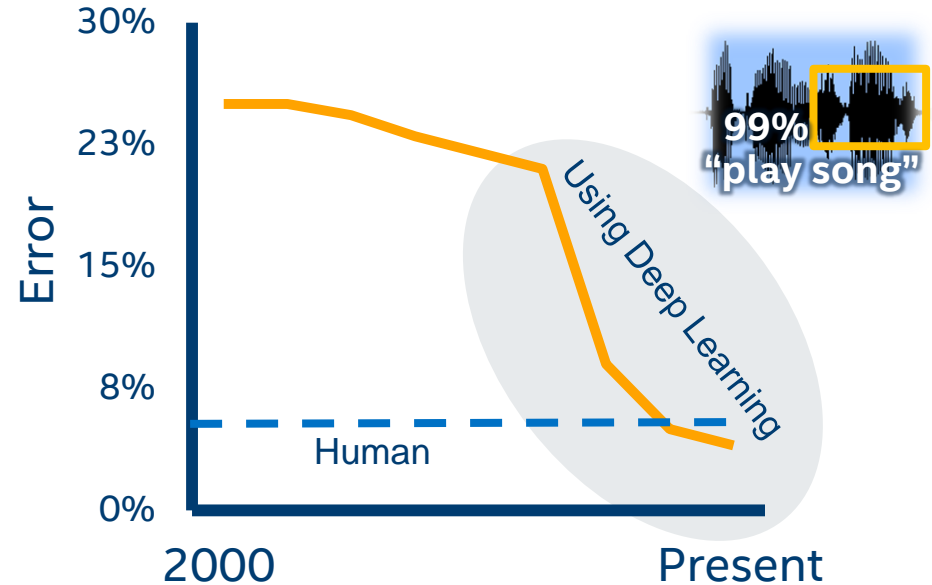
AI COMPUTE CYCLES WILL GROW **12X** BY 2020

# DEEP LEARNING BREAKTHROUGHS

## IMAGE RECOGNITION



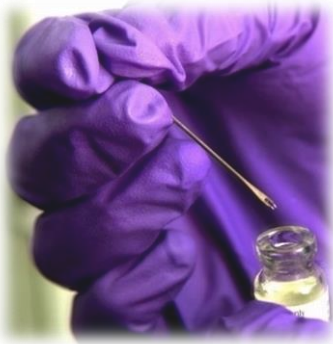
## SPEECH RECOGNITION



These and more are enabling new & improved applications

# AI WILL USHER IN A BETTER WORLD

On the Scale of the Agricultural, Industrial and Digital Revolutions



## ACCELERATE

Large scale solutions

- Cure Diseases
- Prevent Crime
- Unlock Dark Data



## UNLEASH

Scientific Discovery

- Explore New Worlds
- Decode the Brain
- Uncover New Theories



## EXTEND

Human Capabilities

- Personalize Learning
- Enhance Decisions
- Optimize Time



## AUTOMATE

Undesirable Tasks

- Automate Driving
- Save Lives in Danger
- Perform Chores

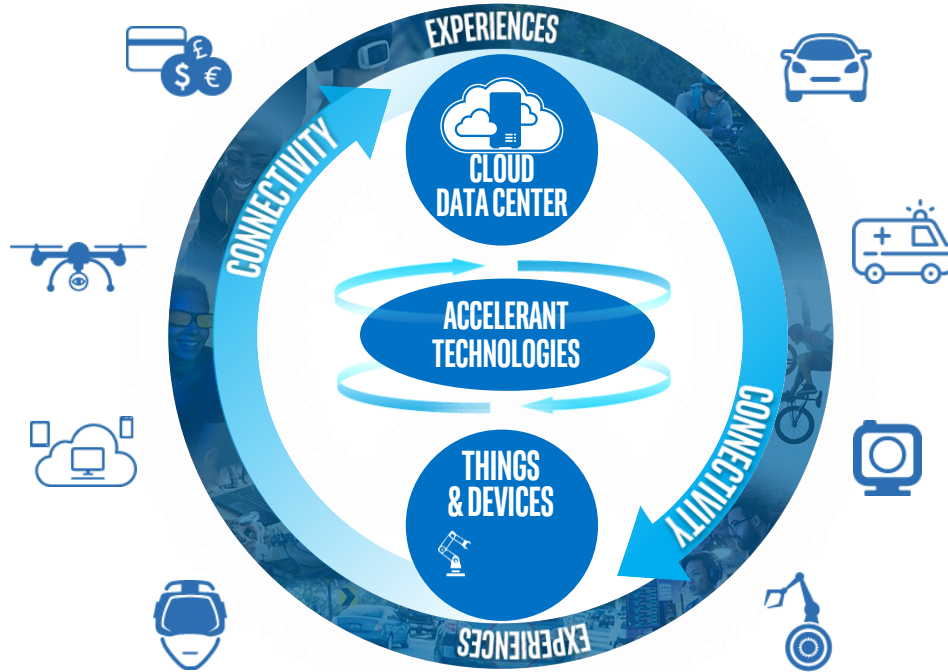
Source: Intel











@IntelAI



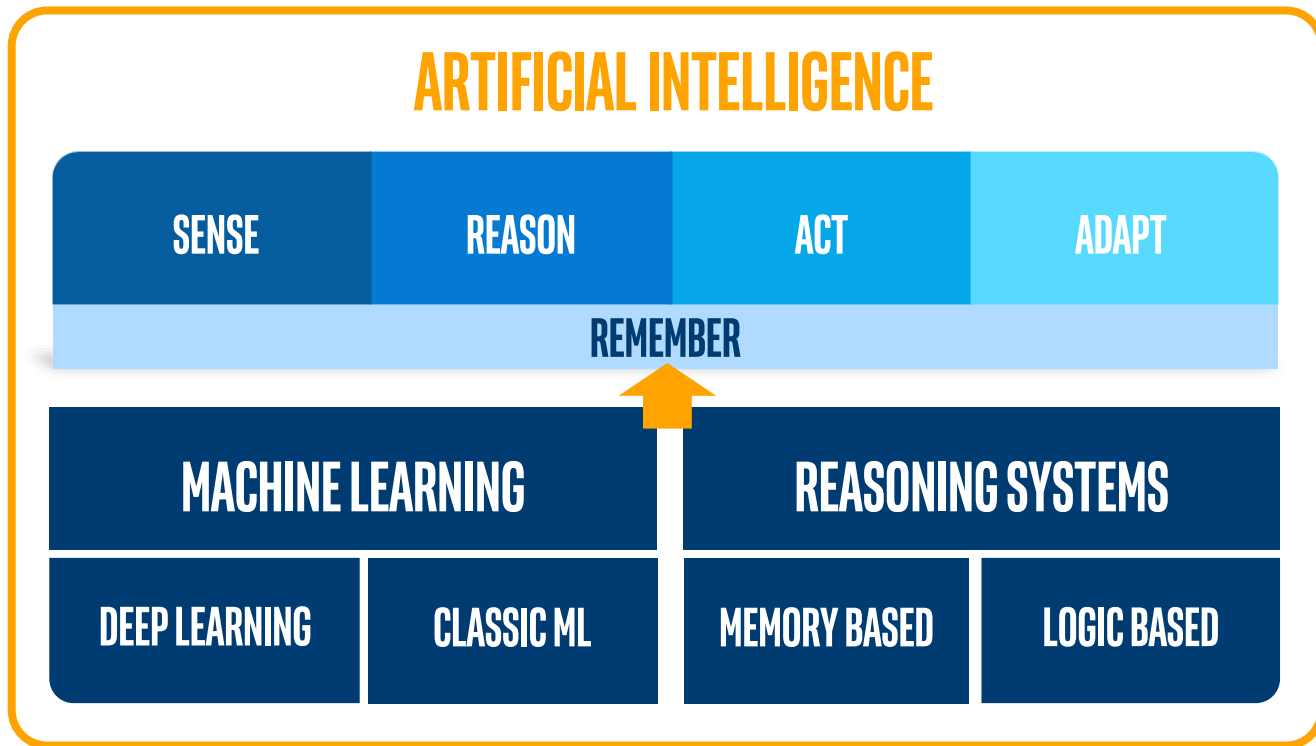
# ARTIFICIAL INTELLIGENCE @ INTEL



- ✓ MACHINE/DEEP LEARNING 
- ✓ REASONING SYSTEMS 
- ✓ PROGRAMMABLE SOLUTIONS 
- ✓ COMPUTER VISION 
- ✓ TOOLS & STANDARDS 
- ✓ MEMORY/STORAGE 
- ✓ NETWORKING 
- ✓ COMMUNICATIONS 

Unleash Your Potential with Intel's Complete AI Portfolio

# A COMMON LANGUAGE FOR AI TODAY

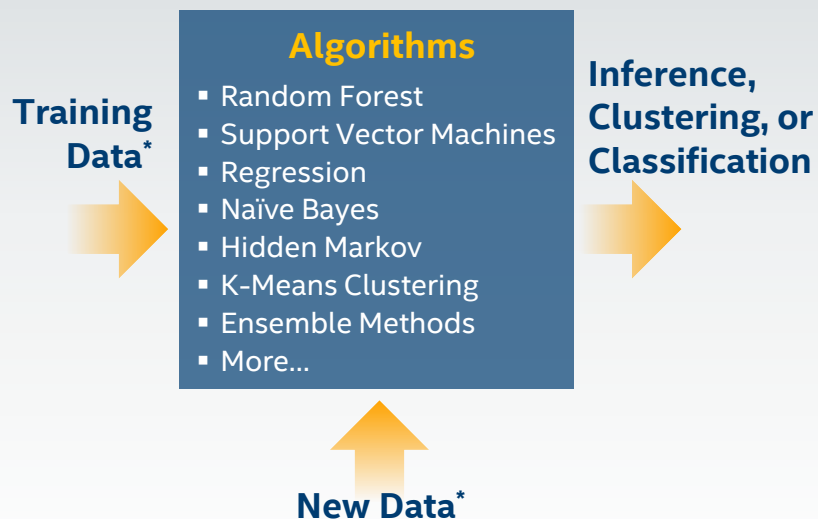




# WHAT IS MACHINE LEARNING?

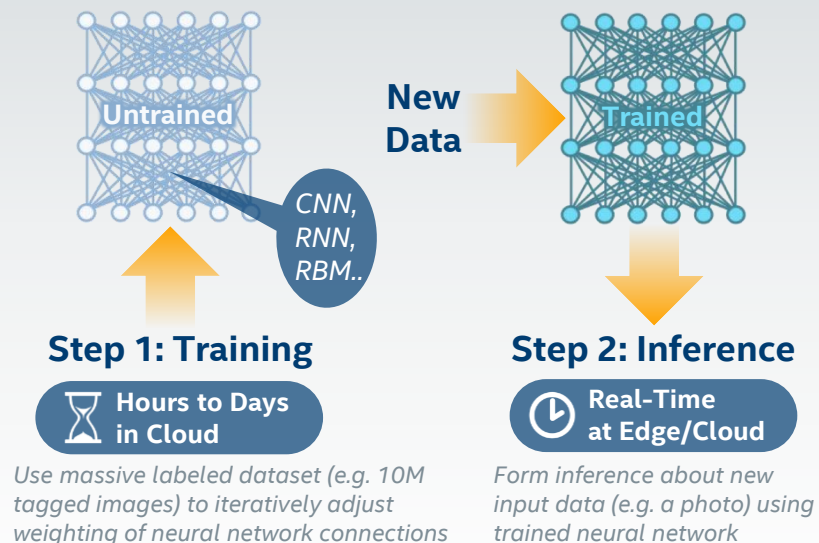
## CLASSIC ML

Using optimized functions or algorithms to extract insights from data



## DEEP LEARNING

Using massive labeled data sets to train deep (neural) graphs that can make inferences about new data



\*Note: not all classic machine learning functions require training

# INTEL AI PORTFOLIO

## EXPERIENCES



## TOOLS



Intel® Deep Learning SDK

Intel® Computer Vision SDK

Movidius Neural Compute Stick



## FRAMEWORKS



theano



Caffe

E2E Tool

## LIBRARIES



Intel Dist  
Intel® DAAL

Intel® Nervana™ Graph\*

Intel® MKL MKL-DNN Intel® MLSL

Movidius MvTensor Library

Associative Memory Base

## HARDWARE



Compute



Memory & Storage



Networking



Visual Intelligence

UNLEASH  
FULL  
POTENTIAL

\*Coming 2017

# LIBRARIES, FRAMEWORKS & TOOLS

## Intel® Math Kernel Library



MKL-DNN



Intel® MLSL

## Intel® Data Analytics Acceleration Library (DAAL)



python  
Intel® Distribution



Open Source Frameworks



Intel Deep Learning SDK



Intel® Computer Vision SDK

High Level Overview

Computation primitives; high performance math primitives granting low level of control

Computation primitives; free open source DNN functions for high-velocity integration with deep learning frameworks

Communication primitives; building blocks to scale deep learning framework performance over a cluster

Broad data analytics acceleration object oriented library supporting distributed ML at the algorithm level

Most popular and fastest growing language for machine learning

Toolkits driven by academia and industry for training machine learning algorithms

Accelerate deep learning model design, training and deployment

Toolkit to develop & deploying vision-oriented solutions that harness the full performance of Intel CPUs and SOC accelerators

Primary Audience

Consumed by developers of higher level libraries and Applications

Consumed by developers of the next generation of deep learning frameworks

Deep learning framework developers and optimizers

Wider Data Analytics and ML audience, Algorithm level development for all stages of data analytics

Application Developers and Data Scientists

Machine Learning App Developers, Researchers and Data Scientists.

Application Developers and Data Scientists

Developers who create vision-oriented solutions

Example Usage

Framework developers call matrix multiplication, convolution functions

New framework with functions developers call for max CPU performance

Framework developer calls functions to distribute Caffe training compute across an Intel® Xeon Phi™ cluster

Call distributed alternating least squares algorithm for a recommendation system

Call scikit-learn k-means function for credit card fraud detection

Script and train a convolution neural network for image recognition

Deep Learning training and model creation, with optimization for deployment on constrained end device

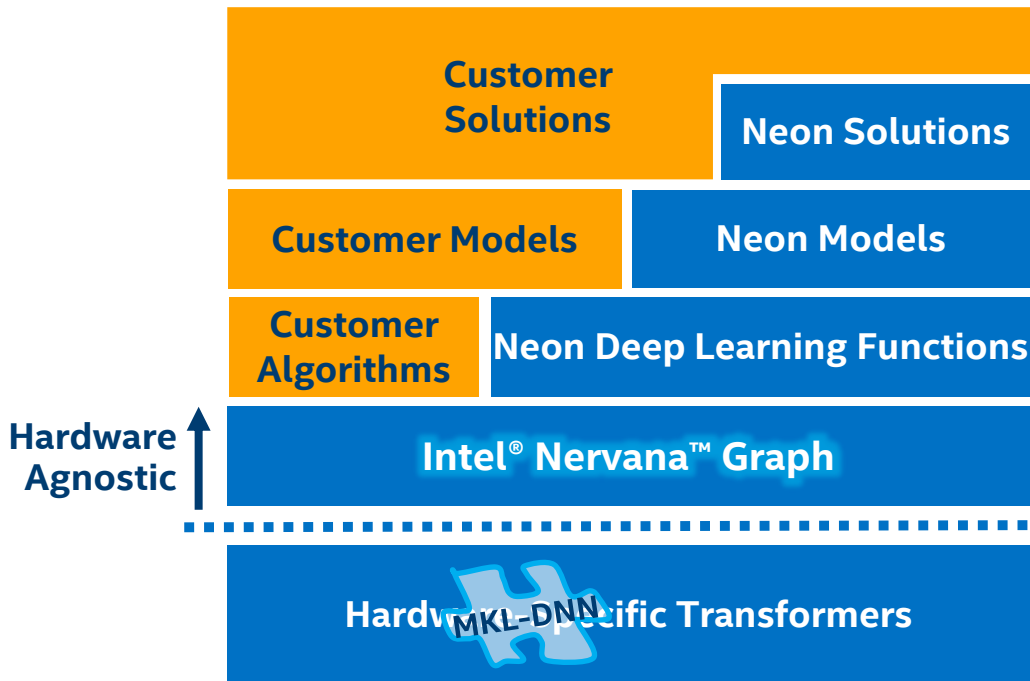
Use deep learning to do pedestrian detection

Find out more at [software.intel.com/ai](https://software.intel.com/ai)

# INTEL® NERVANA™ GRAPH

COMING  
SOON

## High-Performance Execution Graph for Neural Networks



### Intel® Nervana™ Graph

enables optimizations that are applicable across multiple HW targets.

- Efficient buffer allocation
- Training vs inference optimizations
- Efficient scaling across multiple nodes
- Efficient partitioning of subgraphs
- Compounding of ops

*The Intel® Nervana™ Graph will scale performance across hundreds of machine and deep learning frameworks*

# INTEL® NERVANA™ PORTFOLIO

Common Architecture for Machine & Deep Learning



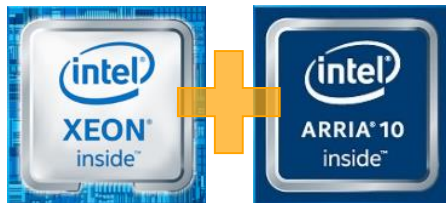
**INTEL® XEON®  
PROCESSORS**

Most Widely Deployed  
Machine Learning  
Platform (>97%\*)



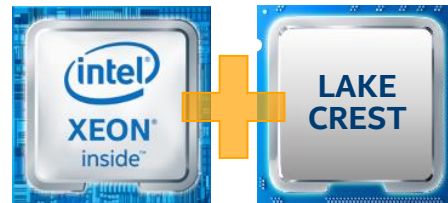
**INTEL® XEON PHI™  
PROCESSORS**

Higher Performance  
Machine Learning,  
General Purpose



**INTEL® XEON® PROCESSOR  
+FPGA**

Breakthrough Deep  
Learning Inference &  
Workload Flexibility



**INTEL® XEON® PROCESSOR  
+ LAKE CREST**

Best-in-Class Neural  
Network Training  
Performance

**TARGETED ACCELERATION**

\*Intel® Xeon® processors are used in 97% of servers that are running machine learning workloads today (Source: Intel)

# ROADMAP: INTEL® NERVANA™ PLATFORM

■ Shipping  
■ Coming Soon

Today

2017

Future

TARGETED ACCELERATION ↑

CREST FAMILY (NERVANA)



Lake Crest

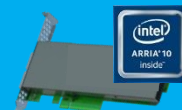
TBA

ALTERA FPGA



Arria 10 FPGA

TBA



Canyon Vista

INTEL® XEON PHI™ PROCESSOR



Knights Landing

TBA



Knights Mill

INTEL® XEON® PROCESSOR



Broadwell

TBA



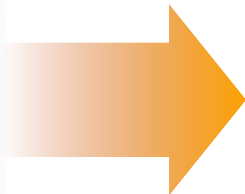
Skylake, +FPGA

# INTEL® NERVANA™ PLATFORM

For Deep Learning

## LAKE CREST

Discrete accelerator  
Coming 2017



## KNIGHTS CREST

Bootable Intel Xeon Processor  
with integrated acceleration

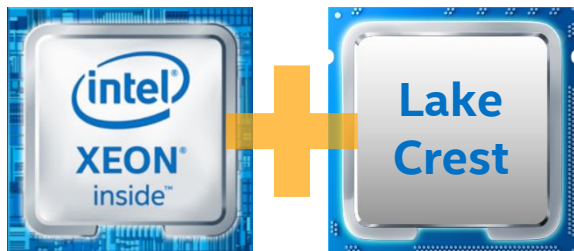
DELIVERING **100X REDUCTION** IN TIME TO TRAIN  
COMPARED TO TODAY'S FASTEST SOLUTION<sup>1</sup> **BY 2020**

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

# LAKE CREST

Best-in-Class Deep Learning Training Performance

COMING  
2017



Accelerator for unprecedented training compute density in deep learning centric environments

## Hardware for DL Workloads

- Custom-designed for deep learning
- Unprecedented compute density
- More raw computing power than today's state-of-the-art GPUs

## Blazingly Fast Data Access

- 32 GB of in package memory via HBM2 technology
- 8 Tera-bits/s of memory access speed

## High Speed Scalability

- 12 bi-directional high-bandwidth links
- Seamless data transfer via interconnects

*Everything needed for deep learning and nothing more!*





# INTEL® XEON PHI™ PROCESSOR FAMILY

Higher Performance Machine Learning, General Purpose



Processor for HPC & enterprises running scale-out, highly-parallel, memory intensive apps

## Removing IO and Memory Barriers

- Integrated Intel® Omni-Path fabric increases price-performance and reduces communication latency
- Direct access of up to **400 GB** of memory with no PCIe performance lag (vs. GPU:16GB)

## Breakthrough Highly Parallel Performance

- Up to **400X** deep learning performance on existing hardware via Intel software optimizations
- Up to **4X** deep learning performance increase estimated (Knights Mill, 2017)

## Easier Programmability

- Binary-compatible with Intel® Xeon® processors
- Open standards, libraries and frameworks

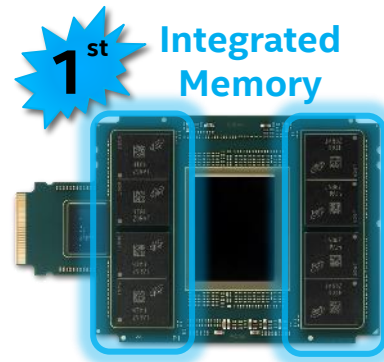
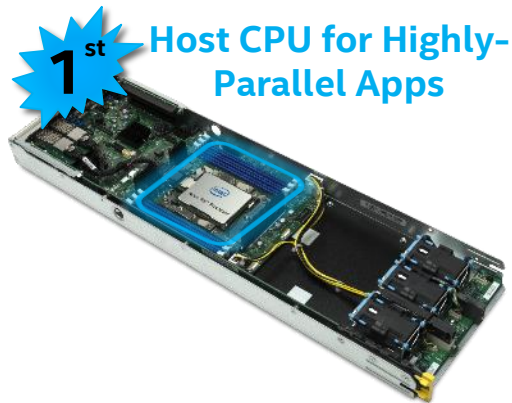
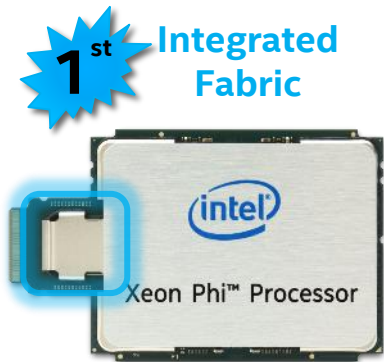
Configuration details on slide: 30

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 November 2016

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.

Notice Revision #20110804

# Introducing the Intel® Xeon Phi™ Processor



## LEADERSHIP PERFORMANCE ... WITH ALL THE BENEFITS OF A CPU

vs. GPU  
Accelerator

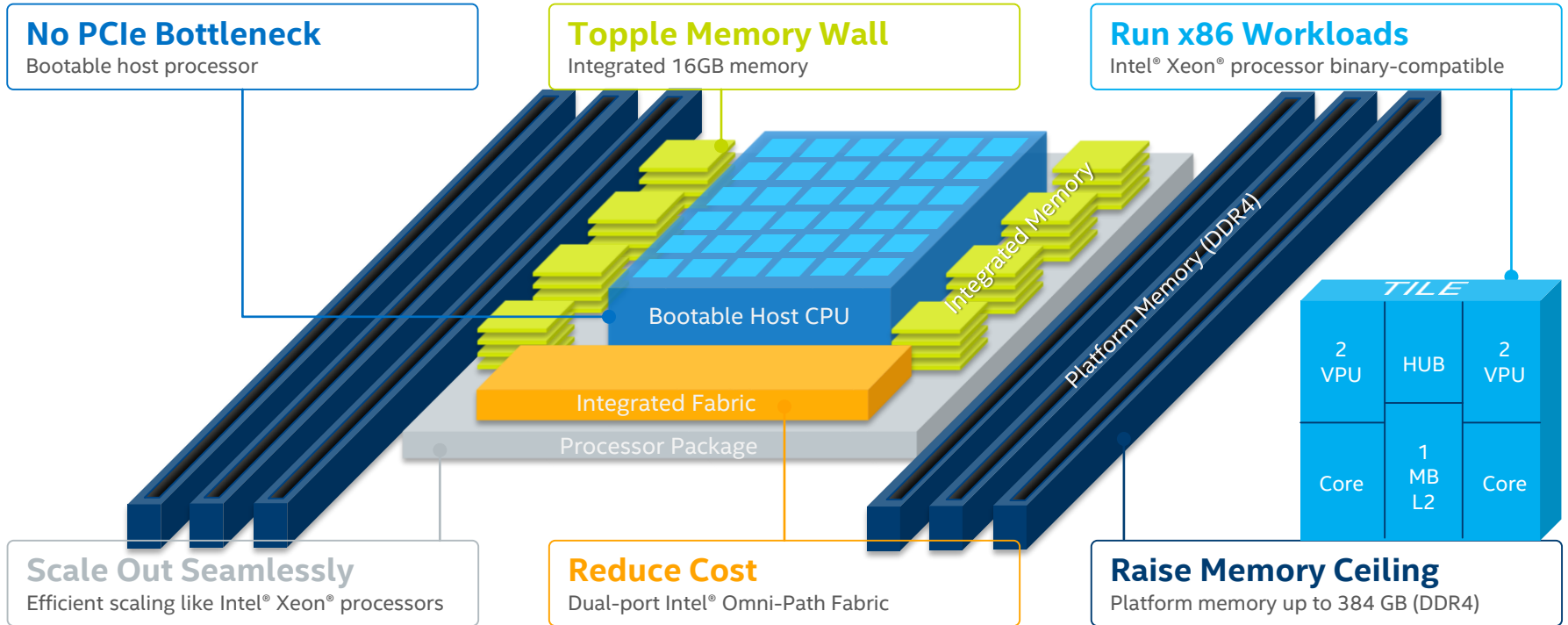


- ✓ Run x86 Workloads
- ✓ Programmability
- ✓ Power Efficient
- ✓ No PCIe Bottleneck
- ✓ Large Memory Footprint
- ✓ Scalability & Future-Ready

\*Intel measured results as of April 2016; see speakers notes for full configuration and performance disclaimers

# Intel® Xeon Phi™ Processor

A Highly-Parallel CPU that Transcends GPU Accelerators



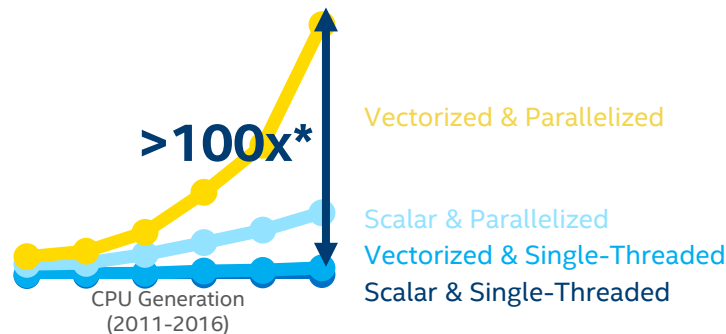
<sup>1</sup>Reduced cost based on Intel internal estimate comparing cost of discrete networking components with the integrated fabric solution



# Solve Biggest Challenges Faster

Highly-Parallel

Intel® Xeon® processors are increasingly parallel and require modern code



Intel® Xeon Phi™ processors are extremely parallel and use general purpose programming



Up to 72 cores (288 threads)



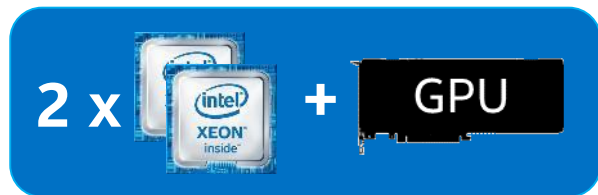
Intel® Advanced Vector Extensions 512 (AVX-512)

\*Binomial Options DP simulation performed on Intel® Xeon® processor X5570 (formerly codenamed Nehalem), Intel® Xeon® processor x5680 (formerly codenamed Westmere), and Intel® Xeon® processor E5 2600 families v1 through v4 for 4 sets of code with varying levels of vectorization and threading optimization

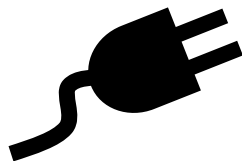


# Realize Compelling Value

Power Efficiency & Cost Savings



683W\*

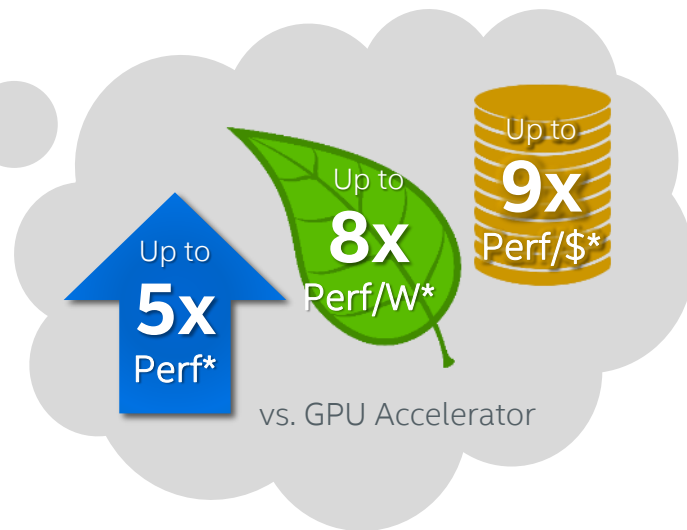


\$13,750\*



378W\*

\$7,300\*



\*Intel measured results as of April 2016; see speakers notes for full configuration and performance disclaimers



# Maximize Your Potential

## Broad Ecosystem



**>30 Systems Providers<sup>1</sup>**

**>15 ISV Application Partners<sup>1</sup>**

**>60 Intel® Parallel Computing Centers<sup>1</sup>**

## Intel® Xeon Phi™ Processor: Broad Ecosystem Support



[www.intel.com/xeonphi/partners](http://www.intel.com/xeonphi/partners)

## Intel® Parallel Computing Centers (IPCC)



[software.intel.com/en-us/ipcc](https://software.intel.com/en-us/ipcc)

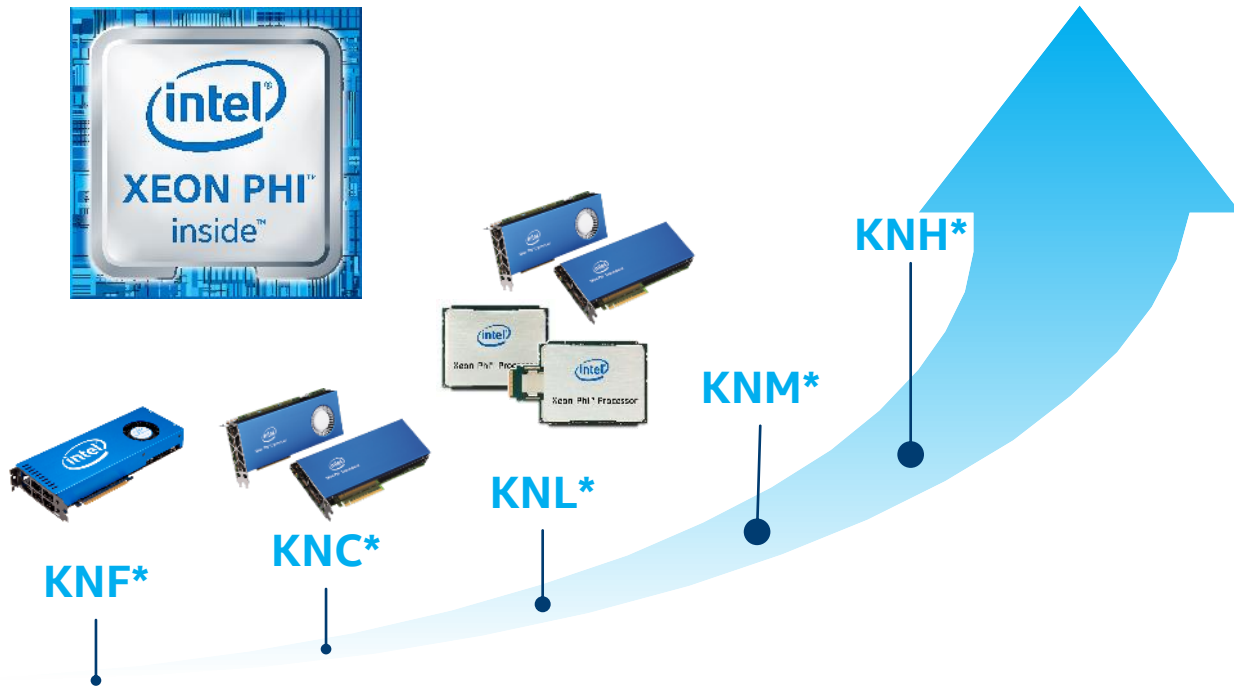
\*Other names, brands and logos may be claimed as the property of others

<sup>1</sup> As of June 2016

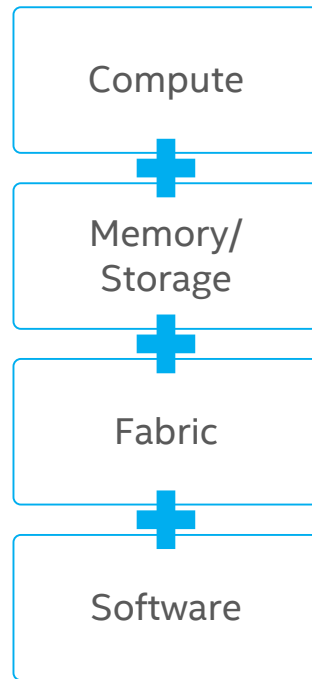


# Maximize Your Potential

## Robust Roadmap



## Intel® Scalable System Framework



\*KNF (Knights Ferry), KNC (Knights Corner), KNL (Knights Landing) are abbreviations for former codenames for Intel® Xeon Phi™ product family products. KNM is the abbreviation for the Knights Mill codename for a future Intel® Xeon Phi™ product. KNH is the abbreviation for the Knights Hill codename of a future Intel® Xeon Phi™ product

# Intel Strategy: Intel® Nervana™ Portfolio

Machine Learning  
Framework  
Optimizations

Spark MLlib

Intel® Distribution for  
Python

Deep Learning  
Framework  
Optimizations



Caffe theano

Low Level  
Software  
Primitives

Intel® DAAL

Intel® MKL

Nervana Graph

Intel® MKL-DNN

Intel® Silicon



+ Storage, Network



# INTEL<sup>®</sup> MKL-DNN

## Math Kernel Library for Deep Neural Networks

For developers of deep learning frameworks featuring optimized performance on Intel hardware

### Distribution Details

- Open Source
- Apache 2.0 License
- Common DNN APIs across all Intel hardware.
- Rapid release cycles, iterated with the DL community, to best support industry framework integration.
- Highly vectorized & threaded for maximal performance, based on the popular Intel<sup>®</sup> MKL library.

BETA Now  
Available!

[github.com/01org/mkl-dnn](https://github.com/01org/mkl-dnn)

Direct 2D  
Convolution

Local response  
normalization  
(LRN)

Rectified linear unit  
neuron activation  
(ReLU)

Maximum  
pooling

Inner product

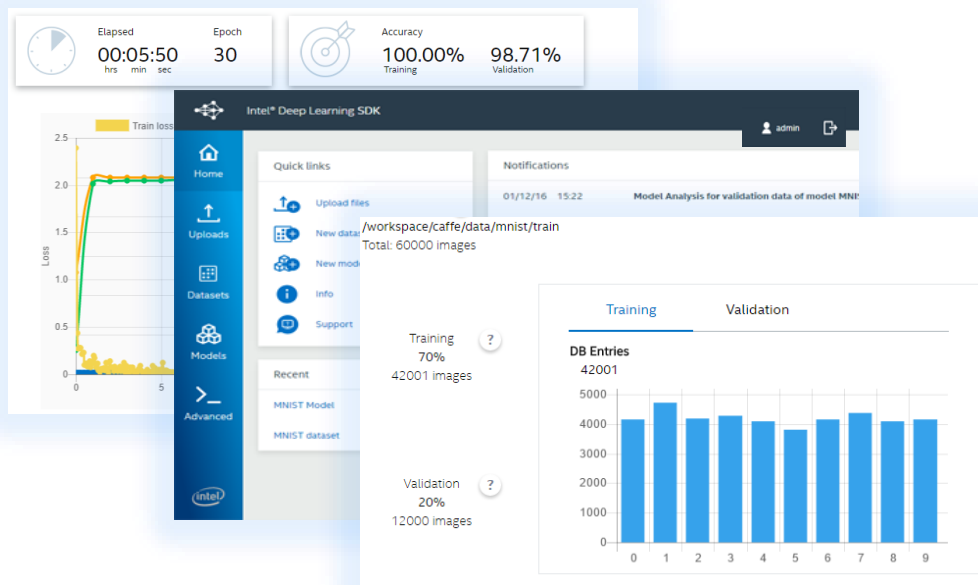
# INTEL® DEEP LEARNING TRAINING TOOL

## Accelerate Deep Learning Development



For developers looking to accelerate deep learning model design, training & deployment

- **FREE** for data scientists and software developers to develop, train & deploy deep learning
- **Simplify installation** of Intel optimized frameworks and libraries
- **Increase productivity** through simple and highly-visual interface
- **Enhance deployment** through model compression and normalization
- **Facilitate integration** with full software stack via inference engine



[software.intel.com/deep-learning-sdk](https://software.intel.com/deep-learning-sdk)



# INTEL® NERVANA™ AI ACADEMY

Hone Your Skills and Build the Future of AI



Benefit from expert-led trainings, hands-on workshops, exclusive remote access, and more.

Gain access to the latest libraries, frameworks, tools and technologies from Intel to accelerate your AI project.

Collaborate with industry luminaries, developers, students, and Intel engineers.

[software.intel.com/ai/academy](https://software.intel.com/ai/academy)

# Reinforcement Learning on IA

Experiments on **Xeon/Xeon Phi**

Team: **deepsense.io, Intel**

Platform: **PLGRID**

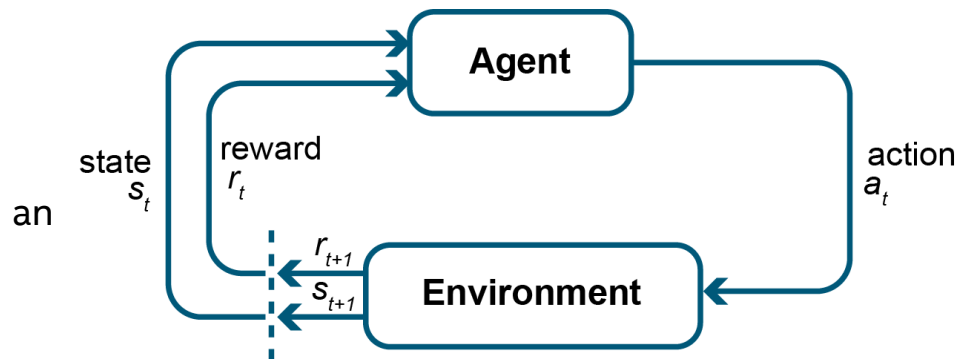
# Reinforcement Learning

**Agent** learns from interaction  
with an **Environment**.

Very general problem

## Examples:

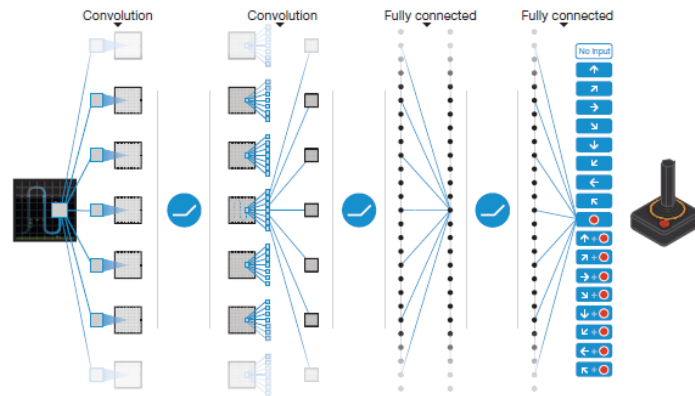
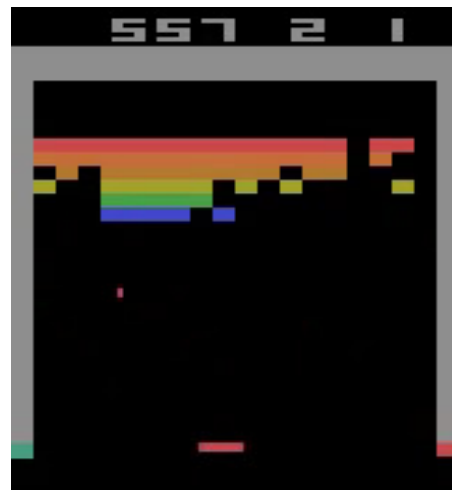
- Robot learning to move items in real environment - a reward is given when item is moved from A to B.
- Robot learning the same task in a simulator.
- An agent playing a board game like Chess - reward for winning a game.
- An agent playing a video game – rewards like in the actual game.



# Reinforcement Learning

## The task: Atari games on CPU

- Train agents for playing **Atari games** from pixel information
- The agent should maximize their score in the game
- Async A2C as an RL algorithm
- 4-layer ConvNet for processing input images



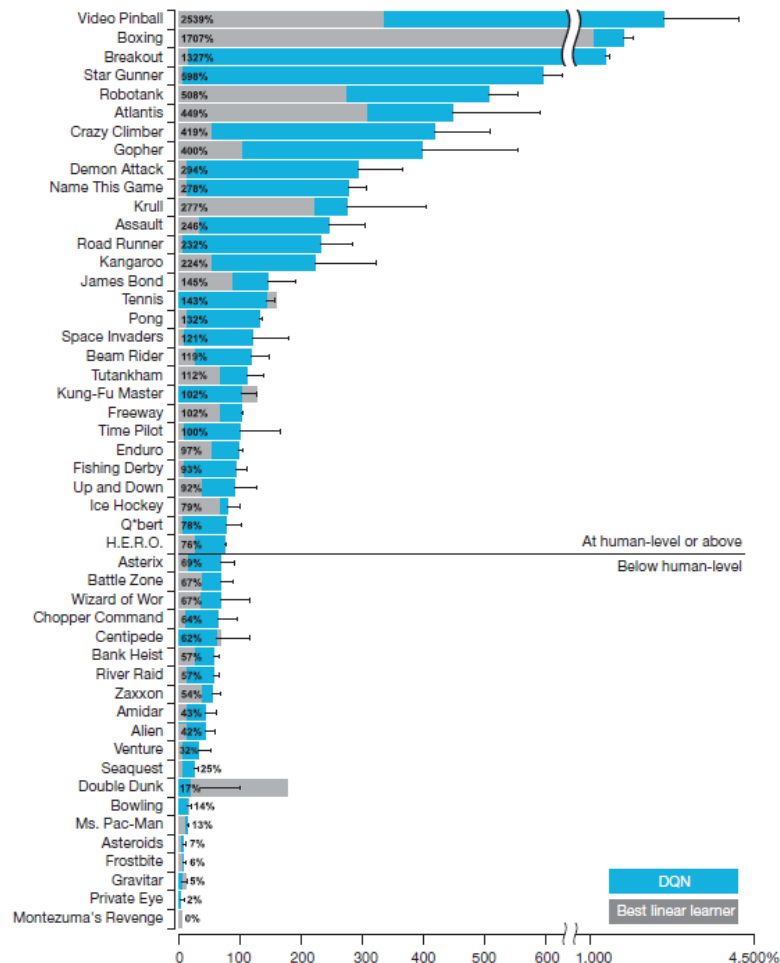
# Reinforcement Learning

**Benchmark games:** Atari 2600 classics

**Environment:** Open AI Gym

**Input:** game screens, 210 x 160 pixels, 3 color channels

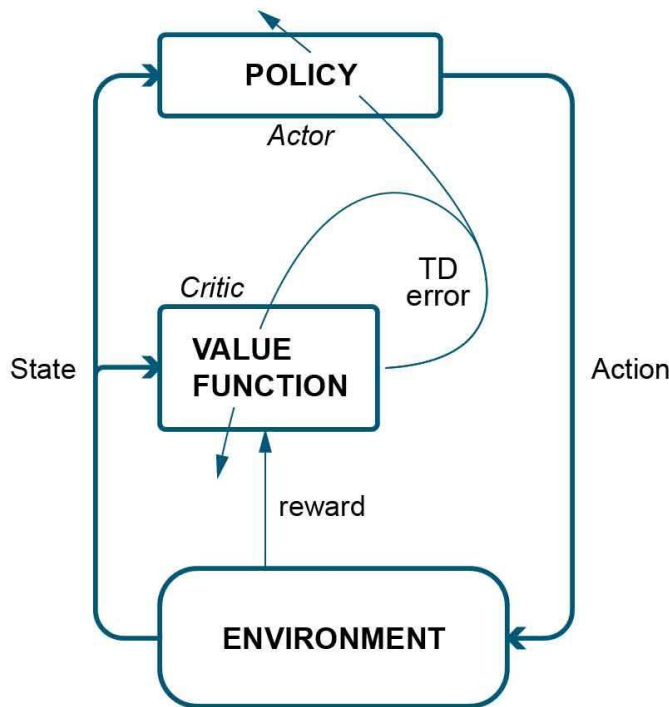
**Output:** one of 18 controller actions



# Reinforcement Learning

## Features of the Batch Asynchronous Advantage Actor-Critic Algorithm (BA3C):

- Hundreds of game simulators are running **in parallel** on a single machine
- The simulators use a shared model to evaluate actions
- The model can batch predictions from multiple simulators to increase efficiency
- The games played by the simulators are also batched and used for training the model



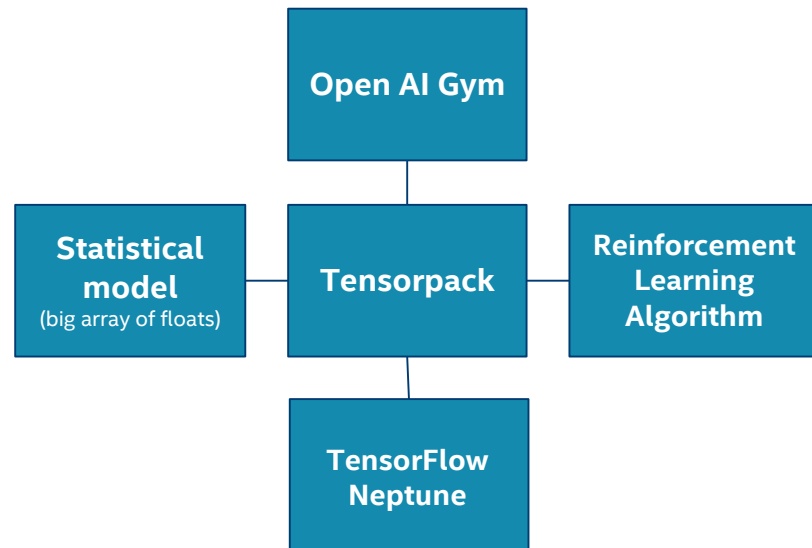
Source: Ben Lau, Using Keras...



# Reinforcement Learning

## Software and hardware stack:

- **Tensorpack** Framework implementing selected learning algorithms in TensorFlow ([Yuxin Wu](#)). Provides an efficient implementation of Async A2C algorithm
- **TensorFlow** General framework for machine learning
- **OpenAI Gym** Framework providing standard environments for reinforcement learning
- **Neptune** Tool for monitoring and managing experiments ([deepsense.io](#))
- **PLGRID** and Xeon Phi server (KNL)



# DNN functions from Math Kernel Library

- We discovered that some TF convolutions were significantly slowing down the training.
- We used [MKL](#) (version 2017.0.098) for better performance
- We forked TensorFlow and provided alternative implementation of convolution using MKL primitives

# MKL convolution - backpropagation

Input shape	Kernel shape	Default TF time	MKL TF time	Default TF time	MKL TF time
		(Xeon) [ms]	(Xeon) [ms]	(Xeon Phi) [ms]	(Xeon Phi) [ms]
128x84x84x16,	5x5x16x32	368.18	29.63	1,236.98	8.97
128x40x40x32,	5x5x32x32	114.72	19.55	343.73	6.33
128x18x18x32	5x5x32x64	28.82	6.07	36.74	2.52
128x7x7x64	3x3x64x64	5.57	3.18	7.38	2.31

# Results

# Reinforcement Learning

Top performance in Breakout

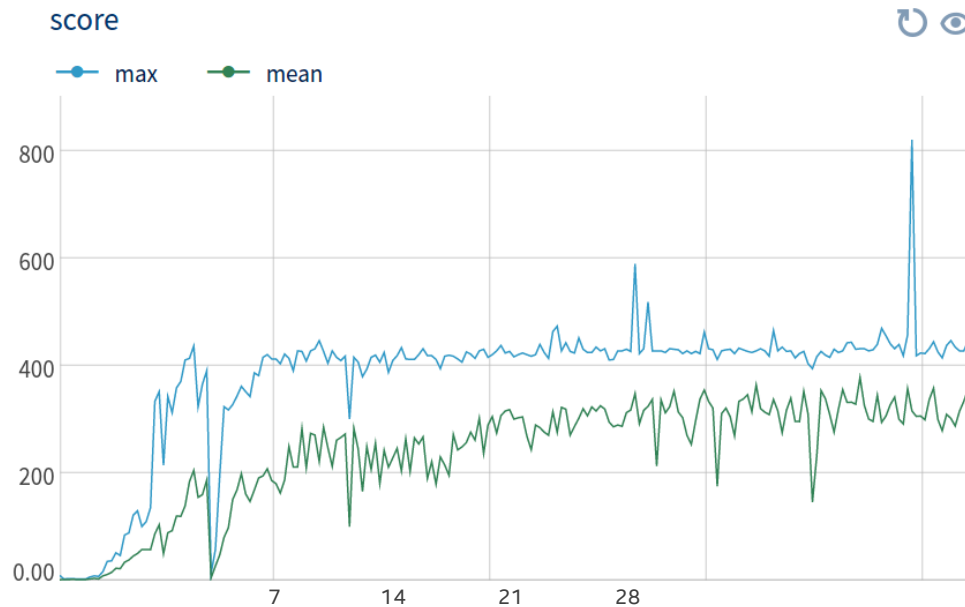


Top performance in River Raid

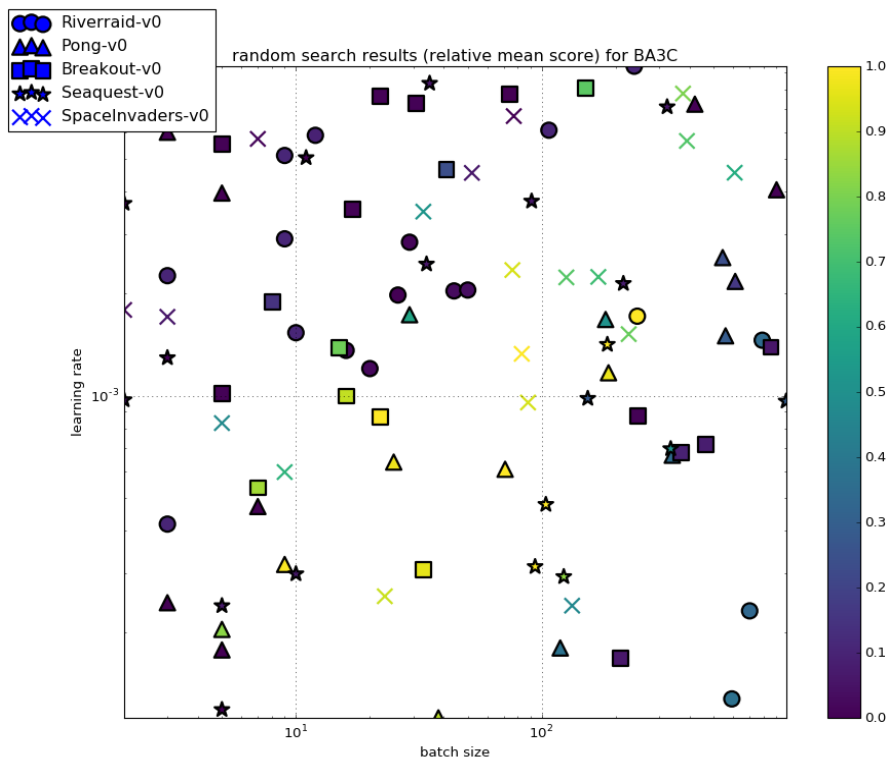


# Reinforcement Learning

## Monitoring the learning process using the Neptune tool (Breakout on Xeon)



# Experiments on PLGRID - Results



# Reinforcement Learning

## Summary

- RL agents trained on **CPU** in just a few hours
- 10x performance gain with MKL DNN implementation, 2.5x for convolutions only
- Atari games and Intel processors, <https://arxiv.org/abs/1705.06936>

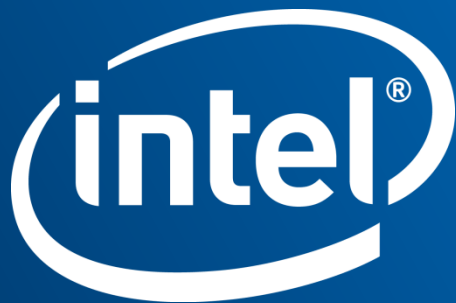
## Challenges and future work:

- Multinode implementation of code on optimized TensorFlow



# THANK YOU

ROBERT.ADAMSKI@INTEL.COM



# LEGAL NOTICES & DISCLAIMERS

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at [intel.com](http://intel.com), or from the OEM or retailer. No computer system can be absolutely secure.

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 <http://www.intel.com/performance>.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

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.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel, the Intel logo, Pentium, Celeron, Atom, Core, Xeon and others 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.

© 2016 Intel Corporation.

# CONFIGURATION DETAILS

## **2S Intel® Xeon® processor E5-2697A v4 on Apache Spark™ with MKL2017 up to 18x performance increase compared to 2S E5-2697 v2 + F2JBLAS machine learning training**

**BASELINE:** Intel® Xeon® Processor E5-2697 v2 (12 Cores, 2.7 GHz), 256GB memory, CentOS 6.6, F2JBLAS: <https://github.com/tomml/netlib-java>, Relative performance 1.0

Intel® Xeon® processor E5-2697 v2 Apache® Spark® Cluster: 1-Master + 8-Workers, 10Gbit/sec Ethernet fabric, Each system with 2 Processors, Intel® Xeon® processor E5-2697 v2 (12 Cores, 2.7 GHz), Hyper-Threading Enabled, 256GB RAM per System, 1-240GB SSD OS Drive, 12-3TB HDDs Data Drives Per System, CentOS® 6.6, Linux 2.6.32-642.1.1.el6.x86\_64, Intel® MKL 2017 build U1\_20160808, Cloudera Distribution for Hadoop (CDH) 5.7, Apache® Spark® 1.6.1 standalone, OMP\_NUM\_THREADS=1 set in CDH®, Total Java Heap Size of 200GB for Spark® Master and Workers, Relative performance up to 3.4x

Intel® Xeon® processor E5-2699 v3 Apache® Spark® Cluster: 1-Master + 8-Workers, 10Gbit/sec Ethernet fabric, Each system with 2 Processors, Intel® Xeon® processor E5-2699 v3 (18 Cores, 2.3 GHz), Hyper-Threading Enabled, 256GB RAM per System, 1-480GB SSD OS Drive, 12-4TB HDDs Data Drives Per System, CentOS® 7.0, Linux 3.10.0-229.el7.x86\_64, Intel® MKL 2017 build U1\_20160808, Cloudera Distribution for Hadoop (CDH) 5.7, Apache® Spark® 1.6.1 standalone, OMP\_NUM\_THREADS=1 set in CDH®, Total Java Heap Size of 200GB for Spark® Master and Workers, Relative performance up to 8.8x

Intel® Xeon® processor E5-2697A v4 Apache® Spark® Cluster: 1-Master + 8-Workers, 10Gbit/sec Ethernet fabric, Each system with 2 Processors, Intel® Xeon® processor E5-2697A v4 (16 Cores, 2.6 GHz), Hyper-Threading Enabled, 256GB RAM per System, 1-800GB SSD OS Drive, 10-240GB SSDs Data Drives Per System, CentOS® 6.7, Linux 2.6.32-573.12.1.el6.x86\_64, Intel® MKL 2017 build U1\_20160808, Cloudera Distribution for Hadoop (CDH) 5.7, Apache® Spark® 1.6.1 standalone, OMP\_NUM\_THREADS=1 set in CDH®, Total Java Heap Size of 200GB for Spark® Master and Workers, Relative performance up to 18x

Machine learning algorithm used for all configurations: Alternating Least Squares ALS Machine Learning Algorithm <https://github.com/databricks/spark-perf>

## **Intel® Xeon Phi™ Processor 7250 GoogleNet V1 Time-To-Train Scaling Efficiency up to 97% on 32 nodes**

32 nodes of Intel® Xeon Phi™ processor 7250 (68 Cores, 1.4 GHz, 16GB MCDRAM: flat mode), 96GB DDR4 memory, Red Hat® Enterprise Linux 6.7, export OMP\_NUM\_THREADS=64 (the remaining 4 cores are used for driving communication) MKL 2017 Update 1, MPI: 2017.1.132, Endeavor KNL bin1 nodes, export I\_MPI\_FABRICS=tmi, export I\_MPI\_TMI\_PROVIDER=psm2, Throughput is measured using "train" command. Data pre-partitioned across all nodes in the cluster before training. There is no data transferred over the fabric while training. Scaling efficiency computed as: (Single node performance / (N \* Performance measured with N nodes)) \* 100, where N = Number of nodes

Intel® Caffe: Intel internal version of Caffe

GoogLeNetV1: <http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43022.pdf>, batch size 1536

## **Intel® Xeon Phi™ processor 7250 up to 400x performance increase with Intel Optimized Frameworks compared to baseline out of box performance**

**BASELINE:** Caffe Out Of The Box, Intel® Xeon Phi™ processor 7250 (68 Cores, 1.4 GHz, 16GB MCDRAM: cache mode), 96GB memory, Centos 7.2 based on Red Hat® Enterprise Linux 7.2, BVLC-Caffe: [https://github.com/BVLC/caffe\\_with\\_OpenBLAS](https://github.com/BVLC/caffe_with_OpenBLAS), Relative performance 1.0

**NEW:** Caffe: Intel® Xeon Phi™ processor 7250 (68 Cores, 1.4 GHz, 16GB MCDRAM: cache mode), 96GB memory, Centos 7.2 based on Red Hat® Enterprise Linux 7.2, Intel® Caffe: <https://github.com/intel/caffe> based on BVLC Caffe as of Jul 16, 2016, MKL GOLD UPDATE1, Relative performance up to 400x

AlexNet used for both configuration as per <https://papers.nips.cc/paper/4824-Large-image-database-classification-with-deep-convolutional-neural-networks.pdf>, Batch Size: 256

## **Intel® Xeon Phi™ Processor 7250, 32 node cluster with Intel® Omni Path Fabric up to 97% GoogleNetV1 Time-To-Train Scaling Efficiency**

Caffe: Intel® Xeon Phi™ processor 7250 (68 Cores, 1.4 GHz, 16GB MCDRAM: flat mode), 96GB DDR4 memory, Red Hat® Enterprise Linux 6.7, Intel® Caffe: <https://github.com/intel/caffe>, not publicly available yet

export OMP\_NUM\_THREADS=64 (the remaining 4 cores are used for driving communication)

MKL 2017 Update 1, MPI: 2017.1.132, Endeavor KNL bin1 nodes, export I\_MPI\_FABRICS=tmi, export I\_MPI\_TMI\_PROVIDER=psm2, Throughput is measured using "train" command. Split the images across nodes and copied locally on each node at the beginning of training. No IO happens over fabric while training.

GoogLeNetV1: <http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43022.pdf>, batch size 1536

## **Intel® Xeon Phi™ processor Knights Mill up to 4x estimated performance improvement over Intel® Xeon Phi™ processor 7290**

**BASELINE:** Intel® Xeon Phi™ Processor 7290 (16GB, 1.50 GHz, 72 core) with 192 GB Total Memory on Red Hat Enterprise Linux® 6.7 kernel 2.6.32-573 using MKL 11.3 Update 4, Relative performance 1.0

**NEW:** Intel® Xeon phi™ processor family – Knights Mill, Relative performance up to 4x

## **Intel® Arria 10 – 1150 FPGA energy efficiency on Caffe/AlexNet up to 25 img/s/w with FP16 at 297MHz**

Intel® AlexNet Classification Implementation as specified by <http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf>, Training Parameters taken from Caffe open-source Framework are 224x224x3 Input, 1000x1 Output, FP16 with Shared Block-Exponents. All compute layers (incl. Fully Connected) done on the FPGA except for Softmax. Arria 10-1150 FPGA, -1 Speed Grade on Altera PCIe DevKit with x72 DDR4 @ 1333 MHz, Power measured through on-board power monitor (FPGA POWER ONLY), ACDS 16.1 Internal Builds + OpenCL SDK 16.1 Internal Build, Compute machine is an HP Z620 Workstation, Xeon E5-1660 at 3.3 GHz with 32GB RAM. The Xeon is not used for compute.

Knights Mill performance: 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.

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 SSE3 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. Notice Revision #20110804

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 everything except Knights Mill which is estimated as of November 2016



# CONFIGURATION DETAILS (CONT'D)

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. Notice Revision #20110804.

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.

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/datacenter>. Tested by Intel as of 14 June 2016. Configurations:

Faster and more scalable than GPU claim based on Intel analysis and testing

Up to 2.3x faster training per system claim based on AlexNet\* topology workload (batch size = 1024) using a large image database running 4-nodes Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat Enterprise Linux\* 6.7 (Santiago), 1.0 TB SATA drive WD1003FZEX-00MK2A0 System Disk, running Intel® Optimized DNN Framework (internal development version) training 1.33 million images in 10.5 hours compared to 1-node host with four NVIDIA "Maxwell" GPUs training 1.33 million images in 25 hours (source: <http://www.slideshare.net/NVIDIA/gtc-2016-opening-keynote> slide 32).

Up to 38% better scaling efficiency at 32-nodes claim based on GoogLeNet deep learning image classification training topology using a large image database comparing one node Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, DDR4 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat\* Enterprise Linux 6.7, Intel® Optimized DNN Framework with 87% efficiency to unknown hosts running 32 each NVIDIA Tesla\* K20 GPUs with a 62% efficiency (Source: <http://arxiv.org/pdf/1511.00175v2.pdf> showing FireCaffe\* with 32 NVIDIA Tesla\* K20s (Titan Supercomputer\*) running GoogleNet\* at 20x speedup over Caffe\* with 1 K20).

Up to 6 SP TFLOPS based on the Intel Xeon Phi processor peak theoretical single-precision performance is preliminary and based on current expectations of cores, clock frequency and floating point operations per cycle. FLOPS = cores x clock frequency x floating-point operations per second per cycle

Up to 3x faster single-threaded performance claim based on Intel estimates of Intel Xeon Phi processor 7290 vs. coprocessor 7120 running XYZ workload.

Up to 2.3x faster training per system claim based on AlexNet\* topology workload (batch size = 1024) using a large image database running 4-nodes Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat Enterprise Linux\* 6.7 (Santiago), 1.0 TB SATA drive WD1003FZEX-00MK2A0 System Disk, running Intel® Optimized DNN Framework, Intel® Optimized Caffe (internal development version) training 1.33 billion images in 10.5 hours compared to 1-node host with four NVIDIA "Maxwell" GPUs training 1.33 billion images in 25 hours (source: <http://www.slideshare.net/NVIDIA/gtc-2016-opening-keynote> slide 32).

Up to 38% better scaling efficiency at 32-nodes claim based on GoogLeNet deep learning image classification training topology using a large image database comparing one node Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, DDR4 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat\* Enterprise Linux 6.7, Intel® Optimized DNN Framework with 87% efficiency to unknown hosts running 32 each NVIDIA Tesla\* K20 GPUs with a 62% efficiency (Source: <http://arxiv.org/pdf/1511.00175v2.pdf> showing FireCaffe\* with 32 NVIDIA Tesla\* K20s (Titan Supercomputer\*) running GoogleNet\* at 20x speedup over Caffe\* with 1 K20).

Up to 50x faster training on 128-node as compared to single-node based on AlexNet\* topology workload (batch size = 1024) training time using a large image database running one node Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat Enterprise Linux\* 6.7 (Santiago), 1.0 TB SATA drive WD1003FZEX-00MK2A0 System Disk, running Intel® Optimized DNN Framework, training in 39.17 hours compared to 128-node identically configured with Intel® Omni-Path Host Fabric Interface Adapter 100 Series 1 Port PCIe x16 connectors training in 0.75 hours. Contact your Intel representative for more information on how to obtain the binary. For information on workload, see <https://papers.nips.cc/paper/4824-Large-image-database-classification-with-deep-convolutional-neural-networks.pdf>.

Up to 30x software optimization improvement claim based on customer CNN training workload running 2S Intel® Xeon® processor E5-2680 v3 running Berkeley Vision and Learning Center\* (BVLC) Caffe + OpenBlas\* library and then run tuned on the Intel® Optimized Caffe (internal development version) + Intel® Math Kernel Library (Intel® MKL).



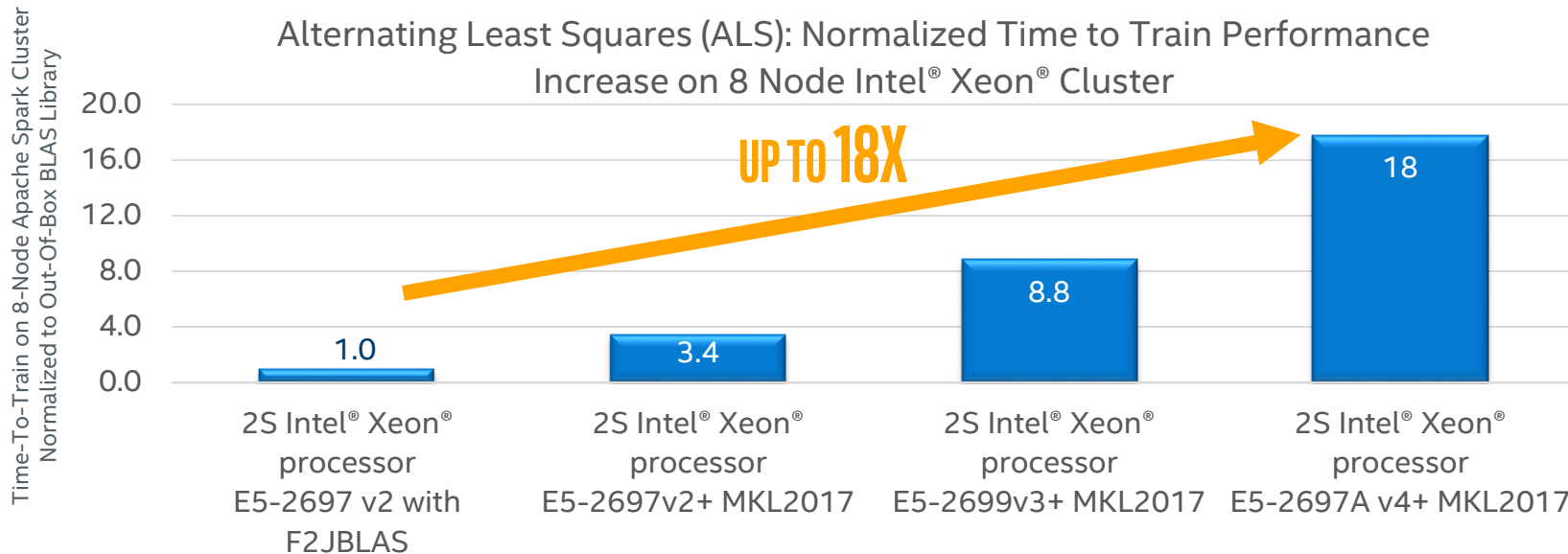


# BENCHMARKS



# INTEL® XEON® PROCESSOR PERFORMANCE

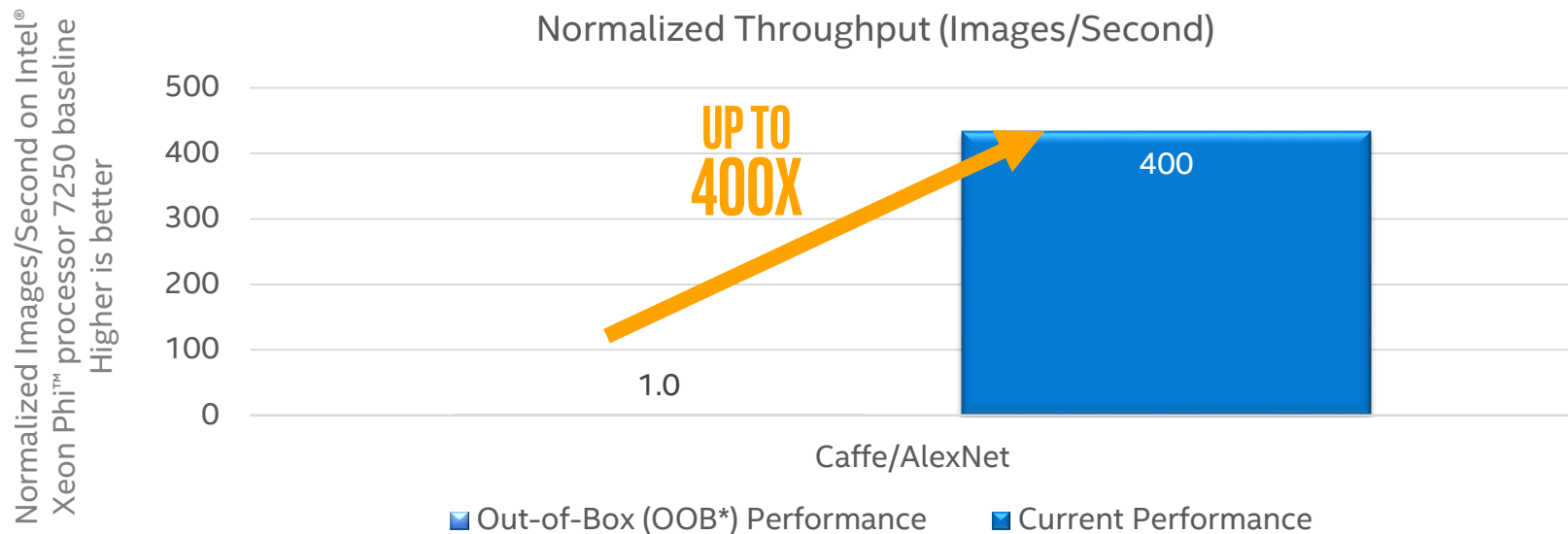
Increasing customer value on existing systems via software optimization for CPU: up to 18x performance increase in under 6 months!



Configuration details on slide: 30  
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 November 2016  
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.  
Notice Revision #20110804

# INTEL® XEON PHI™ PROCESSOR PERFORMANCE

Shattering misconceptions that CPU is not well-suited for deep learning:  
SW optimization delivers up to 400X perf gain on existing HW in <6 months



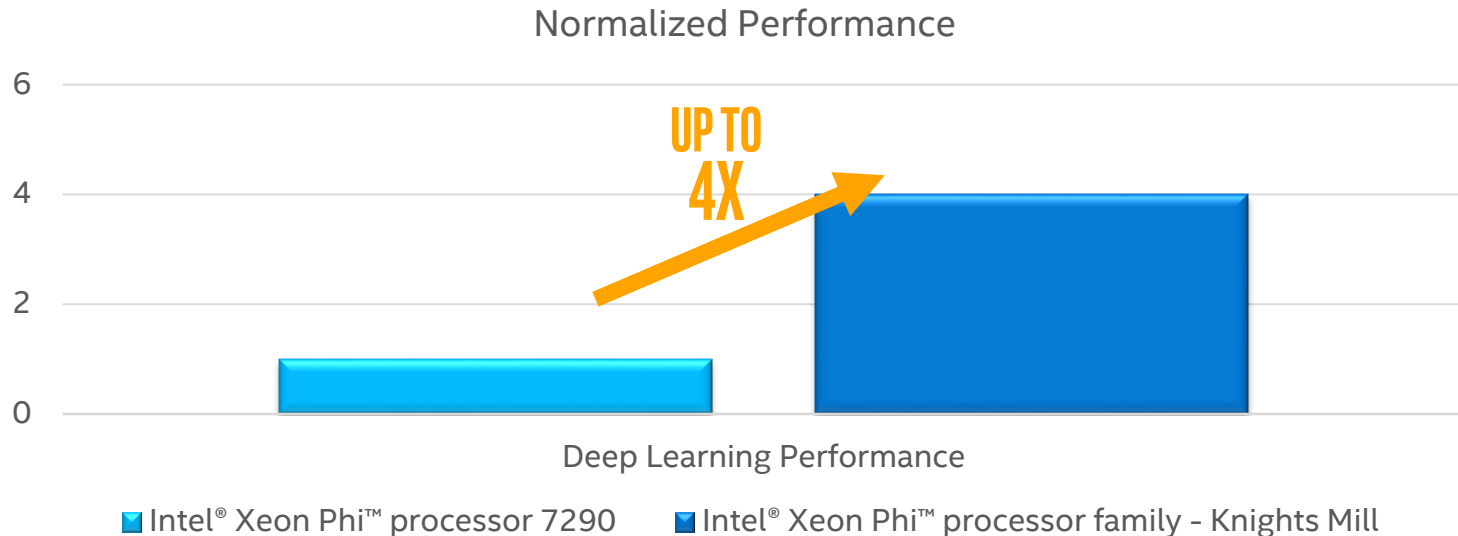
Configuration details on slide: 30  
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. 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 November 2016  
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.  
Notice Revision #20110804





# INTEL® XEON PHI™ PROCESSOR PERFORMANCE

Continued performance breakthroughs: Knights mill (2017) will deliver Up to 4X deep learning performance increase over current generation Intel® Xeon phi™ processor



Configuration details on slide: 30

Knights Mill: 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.

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 November 2016

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.

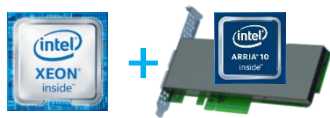
Notice Revision #20110804



# INTEL® ARRIA® 10 FPGA PERFORMANCE

## Deep Learning Image Classification SINGLE-NODE Inference Performance Efficiency

### DISCRETE



System	Throughput	Power	Throughput / Watt
Arria® 10-115 (FP32, Full Size Image, Speed @306Mhz) <sup>1</sup>	575 img/s	~31W	18.5 img/s/W
Arria® 10-115 (FP16, Full Size Image, Speed @297Mhz) <sup>1</sup>	1020 img/s	~40W	25.5 img/s/W
Nvidia* M4 <sup>2</sup>			20 img/s/w

### INTEGRATED\*



Topology	Datatype	Throughput*
AlexNet*	FloatP32	~360 Image/s
	FixedP16	~600 Image/s
	FixedP8	~750 Images/s
	FixedP8 Winograd	~1200 Images/s

**Note:** Intel® Xeon® processor performance is not included in tables

\* Xeon with Integrated FPGA refers to Broadwell Platform Proof of Concept  
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 <http://www.intel.com/performance>. Other names and brands may be property of others.  
Configurations:  
- AlexNet1 Classification Implementation as specified by <http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf>, Training Parameters taken from Caffe open-source Framework are 224x224x3 Input, 1000x1 Output, FP16 with Shared Block Exponents, All compute layers (incl. Fully Connected) done on the FPGA except for Softmax  
- Arria 10-1150 FPGA, -1 Speed Grade on Altera PCIe DevKit with x72 DDR4 @ 1333 MHz, Largest mid-range part from newest family (1518 DSPs, 2800 M20Ks), Power measured through on-board power monitor (FPGA POWER ONLY) ACDS 16.1 Internal Builds  
+ OpenCL SDK 16.1 Internal Build Host system: HP Z620 Workstation, Xeon E5-1660 at 3.3 GHz with 32GB RAM. The Xeon is not used for compute.  
- AlexNet: <https://papers.nips.cc/paper/4824-Imagenet-classification-with-deep-convolutional-neural-networks.pdf>, Batch Size:256 2.From GTC Keynote