



Big Data Analytics & Machine Learning on Intel architecture

Igor Freitas
Developer Relations Division
Software & Services Group

Legal Notices and Disclaimers

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. 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, OR LIFE SUSTAINING APPLICATIONS.

Intel may make changes to specifications and product descriptions at any time, without notice.

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

Intel, processors, chipsets, and desktop boards 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.

Any code names featured are used internally within Intel to identify products that are in development and not yet publicly announced for release. Customers, licensees and other third parties are not authorized by Intel to use code names in advertising, promotion or marketing of any product or services and any such use of Intel's internal code names is at the sole risk of the user.

Intel product plans in this presentation do not constitute Intel plan of record product roadmaps. Please contact your Intel representative to obtain Intel's current plan of record product roadmaps.

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

Intel, the Intel logo, Xeon and Xeon logo, Xeon Phi and Xeon Phi 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.

All products, dates, and figures are preliminary and are subject to change without any notice. Copyright © 2016, Intel Corporation.

This document contains information on products in the design phase of development.

Optimization Notice

Copyright © 2016, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.



Optimization Notice

Optimization Notice

Intel® compilers, associated libraries and associated development tools may include or utilize options that optimize for instruction sets that are available in both Intel® and non-Intel microprocessors (for example SIMD instruction sets), but do not optimize equally for non-Intel microprocessors. In addition, certain compiler options for Intel compilers, including some that are not specific to Intel micro-architecture, are reserved for Intel microprocessors. For a detailed description of Intel compiler options, including the instruction sets and specific microprocessors they implicate, please refer to the “Intel® Compiler User and Reference Guides” under “Compiler Options.” Many library routines that are part of Intel® compiler products are more highly optimized for Intel microprocessors than for other microprocessors. While the compilers and libraries in Intel® compiler products offer optimizations for both Intel and Intel-compatible microprocessors, depending on the options you select, your code and other factors, you likely will get extra performance on Intel microprocessors.

Intel® compilers, associated libraries and associated development tools 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 Intel® Streaming SIMD Extensions 2 (Intel® SSE2), Intel® Streaming SIMD Extensions 3 (Intel® SSE3), and Supplemental Streaming SIMD Extensions 3 (Intel® 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.

While Intel believes our compilers and libraries are excellent choices to assist in obtaining the best performance on Intel® and non-Intel microprocessors, Intel recommends that you evaluate other compilers and libraries to determine which best meet your requirements. We hope to win your business by striving to offer the best performance of any compiler or library; please let us know if you find we do not.

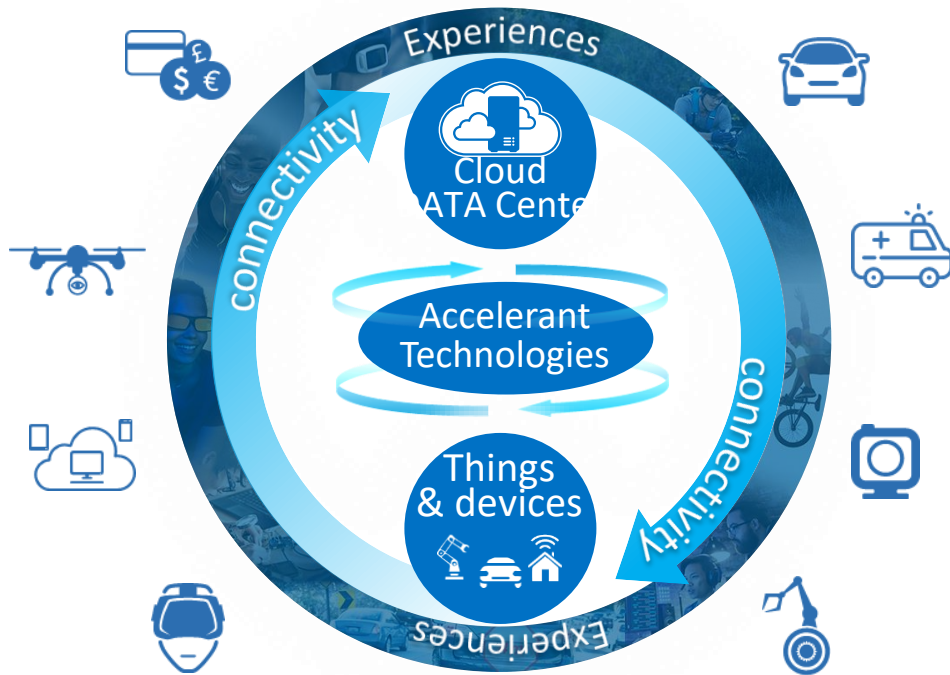
Notice revision #20101101

Agenda

- Intel Machine Learning Strategy
- Software and Hardware Architecture
- Get started today
- Backup – More Architecture Details

Artificial intelligence on intel architecture

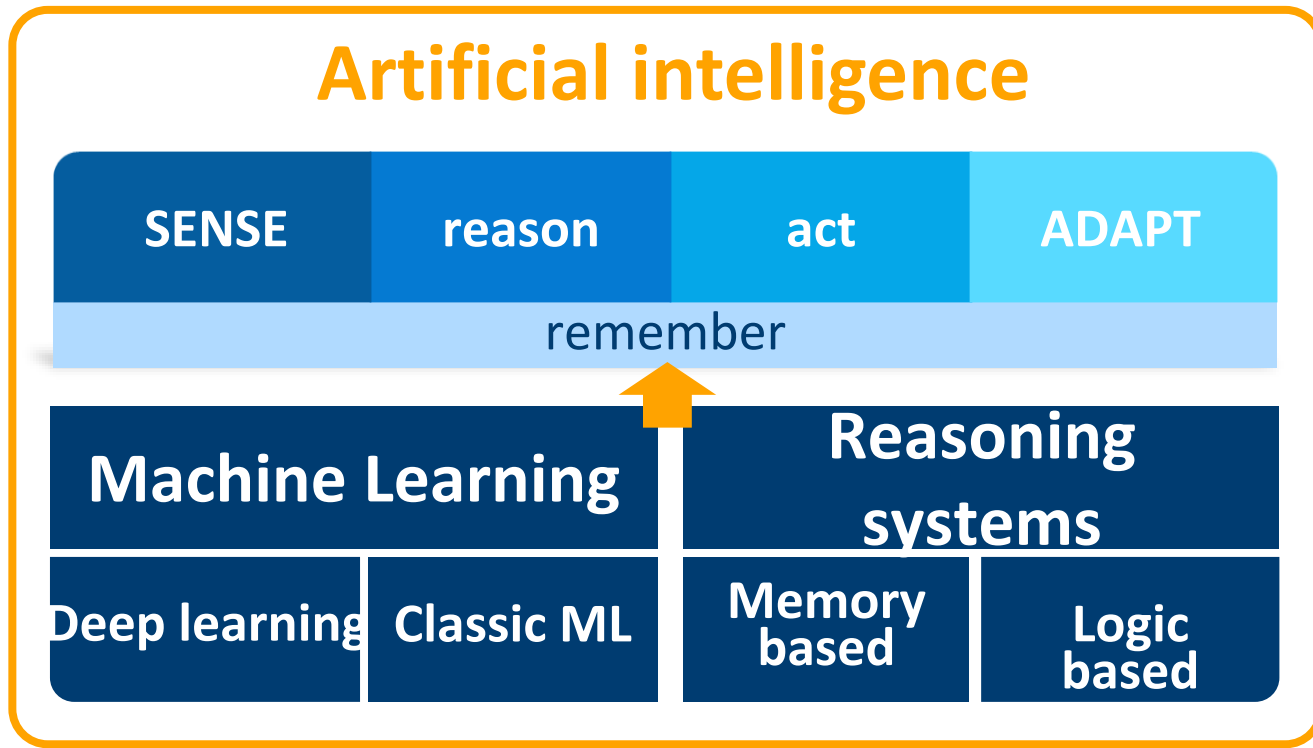
Artificial intelligence @ Intel



- ✓ MACHINE/DEEP LEARNING 
- ✓ REASONING SYSTEMS 
- ✓ Programmable solutions 
- ✓ COMPUTER VISION 
- ✓ TOOLS & STANDARDS 
- ✓ Memory/storage 
- ✓ Networking 
- ✓ communications 

Unleash Your Potential with Intel's Complete AI Portfolio

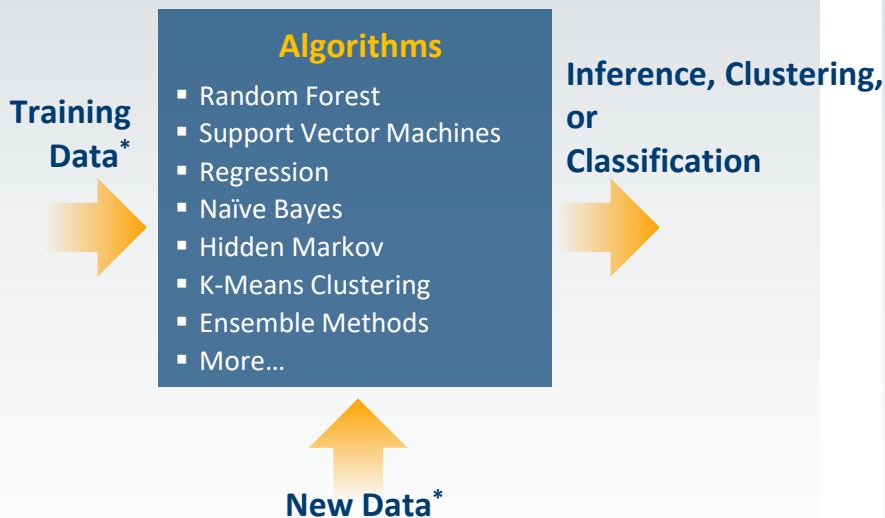
A common language for AI Today



What is Machine Learning?

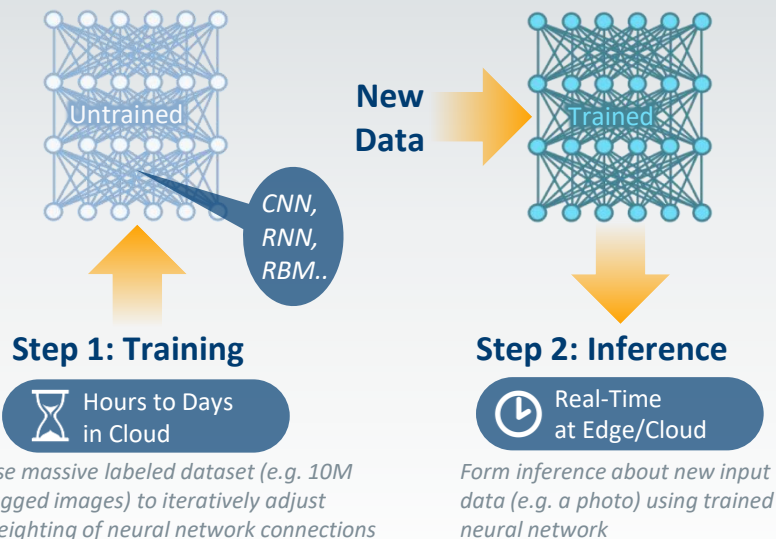
Classic ML

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



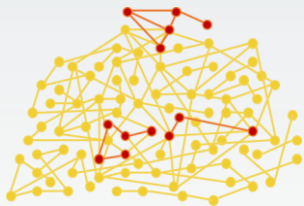
*Note: not all classic machine learning functions require training

What is a Reasoning system?

Memory based

Using associations between concepts from multiple data types to make sense of complex situations

e.g. *under what system conditions should I perform preventive maintenance to avoid a failure?*



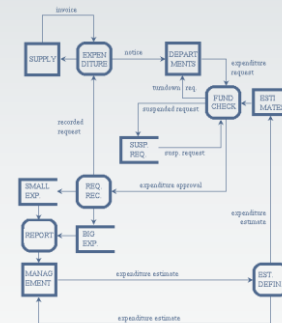
STAT	1	1	1		
06/02	1	1	1	1	
reset	1	1	5	4	1
a1	1	1		1	1
a0	1	1		1	1
habel	1	1		1	1
open.new	1	1		1	1
completed	1	1		1	1
p1	1	1		1	1
	120	121	chv	bxt	06/02
					set

- ✓ Flexibility to handle ALL data types at once
- ✓ Incorporate new data in real-time
- ✓ Transparent and explainable

Logic based

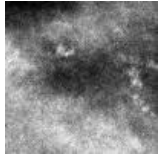
Using a rule-based reasoning engine, usually hand-created or maintained, to perform logical inferencing steps

e.g. *should I maintain or alter my equity portfolio given my risk profile?*

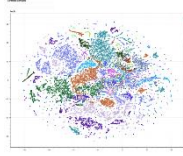


- ✓ Explicit encoding of knowledge
- ✓ Repeatable, reversible, deterministic
- ✓ Transparent and explainable

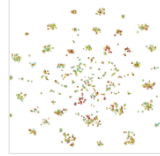
Customer examples



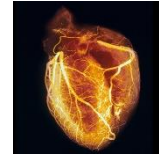
health



Finance



industrial



Health



Finance



industrial

Early Tumor Detection

Leading medical imaging company



Early detection of malignant tumors in mammograms



Millions of “Diagnosed” Mammograms



Deep Learning (CNN) tumor image recognition



Higher accuracy and earlier breast cancer detection

Data Synthesis

Financial services institution with >\$750B assets



Parse info to reduce portfolio manager time to insight



Vast stores of documents (news, emails, research, social)



Deep Learning (RNN w/ encoder/decoder)



Faster and more informed investment decisions

Smart Agriculture

World leader in agricultural biotech



Accelerate hybrid plant development



Large dataset of hybrid plant performance based on genotype markers



Deep Learning (CNN) to detect favorable interactions between genotype markers



More accurate selection leading to cost reduction

Personalized Care

Renowned US Hospital system



Accurately diagnose fatal heart conditions



10,000 health attributes used



Saffron memory-based reasoning



Increased accuracy to 94% compared with 54% for average cardiologist

Customer Personalization

Leading Insurance Group



Increase product recommendation accuracy



5 Product Levels
1,353 Products
12M Members



Saffron memory-based reasoning



50% increase in product recommendation accuracy

Supply Chain Logistics

Multinational Aerospace Corp



Reduce time to select aircraft replacement part



15,000 Aircraft
\$1M/day idle



Saffron memory-based reasoning



Reduced part selection from 4 hours to 5 mins

Deep Learning

Memory-based Reasoning

intel AI portfolio

experiences



tools



Frameworks



E2E Tool

libraries



hardware



Compute



Memory & Storage



Networking



Visual Intelligence

Unleash
Full
potential

*Coming 2017

Optimization Notice

Copyright © 2016 Intel Corporation. All rights reserved.
Other names and brands may be claimed as the property of others.



MACHINE Learning: SW & HW Architecture on Intel[®]

End-to-end use case

Artificial Intelligence For Automated Driving



vehicle

Driving Functions

Environment Modeling

Sensor Fusion

Anomaly Detection



Movidius | intel REALSENSE

ML DL
PP LP

Network

Captured
Sensor Data

Real Time
Model, SW, FW
Updates

Data
Formatting,
Storage,
Management,
Traceability

5G | intel XEON | LB

Cloud

Model
Training

Model
Inference

Compress
Model

Universal Models

Reasoning Systems



saffron TECHNOLOGY

ML DL
RB LB

Intel® Nervana™ portfolio (Detail)



Batch



Many batch models

Train machine learning models across a diverse set of dense and sparse data



Train large deep neural networks
Train large models as fast as possible



Future



Batch



Stream

...



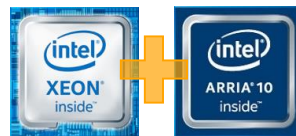
Edge

Infer billions of data samples at a time and feed applications within ~1 day



Option for higher throughput/watt

Infer deep data streams with low latency in order to take action within milliseconds



Required for low latency

Power-constrained environments

Movidius
or other Intel® edge processor



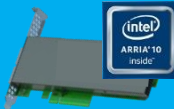




Training

inference

e

roadmap: Intel® nervana™ platform

Shipping
Coming Soon

	Today	2017	2018+	
<p>↑ Targeted acceleration</p>	Crest family (nervana)	 Lake Crest	TBA	
	altera FPGA	 Arria 10 FPGA	 Canyon Vista	TBA
	Intel® Xeon Phi™ processor	 Knights Landing	 Knights Mill	TBA
	Intel® Xeon® processor	 Broadwell	 Skylake, +FPGA	TBA

Optimization Notice

Copyright © 2016 Intel Corporation. All rights reserved.
Other names and brands may be claimed as the property of others.

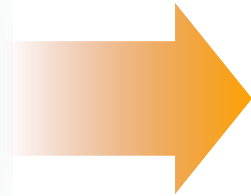


Intel® Nervana™ Platform

For Deep Learning

Lake Crest

Discrete accelerator
First silicon 1H'2017



Knights Crest

Bootable Intel Xeon Processor
with integrated acceleration

Delivering **100X** Reduction in time to train by 2020
COMPARED TO TODAY'S FASTEST SOLUTION¹

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

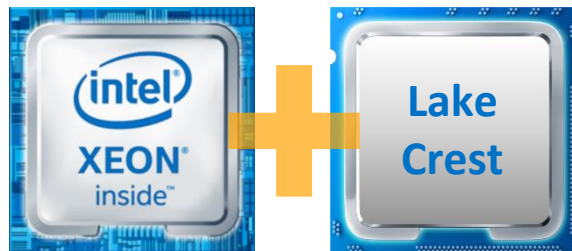
Copyright © 2016 Intel Corporation. All rights reserved.
Other names and brands may be claimed as the property of others.



Lake crest



Deep Learning by Design



Add-in card for unprecedented 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

Enables Shorter Time to Train Using General Purpose Infrastructure



Processor for HPC & enterprise customers 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)

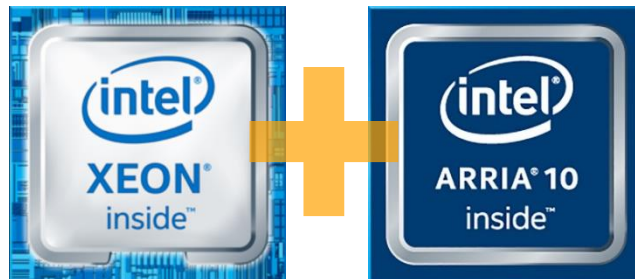
Easier Programmability

- 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

Intel® Arria® 10 FPGA

Superior Inference Capabilities



Add-in card for higher performance/watt inference with low latency and flexible precision

Energy Efficient Inference with Infrastructure Flexibility

- Excellent energy efficiency up to **25** images/sec/watt inference on Caffe/Alexnet
- Reconfigurable accelerator can be used for variety of data center workloads
- Integrated FPGA with Intel® Xeon® processor fits in standard server infrastructure
-OR- Discrete FPGA fits in PCIe card and embedded applications*

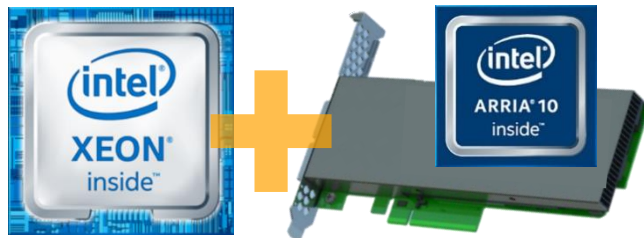
*Xeon with Integrated FPGA refers to Broadwell Proof of Concept Configuration details on slide: 44

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

Canyon Vista

Turnkey Deep Learning Inference Solution



Pre-configured add-in card for higher performance/watt inference for image recognition

Energy Efficient Inference

- Accelerates image recognition using convolutional neural networks (CNN)
- Excellent energy efficient inference up to **25** images/s/w on Caffe/Alexnet
- Fits in standard server infrastructure*

Accelerate Time to Market

- Simplify deployment with preloaded optimized CNN algorithms
- Integrated software ecosystem: optimized libraries, frameworks and APIs

*Standard server infrastructure – verified with Broadwell platform

Configuration details on slide: 44

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

Optimization Notice

Copyright © 2016 Intel Corporation. All rights reserved.
Other names and brands may be claimed as the property of others.



Intel® Xeon® Processor Family

Most Widely Deployed Machine Learning Platform (97% share)*



Processor optimized for a wide variety of datacenter workloads enabling flexible infrastructure

Lowest TCO With Superior Infrastructure Flexibility

- Standard server infrastructure
- Open standards, libraries & frameworks
- Optimized to run wide variety of data center workloads

Server Class Reliability

- Industry standard server features: high reliability, hardware enhanced security

Leadership Throughput

- Industry leading inference performance
- Up to **18X** performance on existing hardware via Intel software optimizations

Configuration details on slide: 30

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

Optimization Notice

Copyright © 2016, Intel Corporation. All rights reserved.
Other names and brands may be claimed as the property of others.



Intel® OMNI-path Architecture

World-Class Interconnect Solution for Shorter Time to Train

HFI Adapters

Single port
x8 and x16



Edge Switches

1U Form Factor
24 and 48 port



Director Switches

QSFP-based
192 and 768 port



Software

Open Source
Host Software and
Fabric Manager



Cables

Third Party Vendors
Passive Copper
Active Optical



Fabric interconnect for breakthrough performance on scale-out apps like deep learning training

Building on some of Industry's best technologies

- Highly leverage existing Aries & Intel True Scale fabrics
- Excellent price/performance \leftrightarrow price/port, 48 radix
- Re-use of existing OpenFabrics Alliance Software
- Over 80+ Fabric Builder Members

Breakthrough Performance

- Increases price performance, reduces communication latency compared to InfiniBand EDR¹:
 - Up to **21%** Higher Performance, lower latency at scale
 - Up to **17%** higher messaging rate
 - Up to **9%** higher application performance

Innovative Features

- Improve performance, reliability and QoS through:
 - Traffic Flow Optimization to maximize QoS in mixed traffic
 - Packet Integrity Protection for rapid and transparent recovery of transmission errors
 - Dynamic lane scaling to maintain link continuity

¹Intel® Xeon® Processor E5-2697A v4 dual-socket servers with 2133 MHz DDR4 memory. Intel® Turbo Boost Technology and Intel® Hyper Threading Technology enabled. BIOS: Early snoop disabled, Cluster on Die disabled, IOU non-posted prefetch disabled, Snoop hold-off timer=9. Red Hat Enterprise Linux Server release 7.2 (Maipo). Intel® OPA testing performed with Intel Corporation Device 24f0 – Series 100 HFI ASIC (B0 silicon). OPA Switch: Series 100 Edge Switch – 48 port (B0 silicon). Intel® OPA host software 10.1 or newer using Open MPI 1.10.x contained within host software package. EDR IB* testing performed with Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 - 36 Port EDR Infiniband switch. EDR tested with MLNX_OFED_Linux-3.2.x. OpenMPI 1.10.x contained within MLNX HPC-X. Message rate claim: Ohio State Micro Benchmarks v. 5.0. osu_mbw_mr, 8 B message (uni-directional), 32 MPI rank pairs. Maximum rank pair communication time used instead of average time, average timing introduced into Ohio State Micro Benchmarks as of v3.9 (2/28/13). Best of default, MXM_TLS=self,rc, and -mca pml yalla tunings. All measurements include one switch hop. Latency claim: HPCC 1.4.3 Random order ring latency using 16 nodes, 32 MPI ranks per node, 512 total MPI ranks. Application claim: GROMACS version 5.0.4 ion_channel benchmark. 16 nodes, 32 MPI ranks per node, 512 total MPI ranks. Intel® MPI Library 2017.0.064. Additional configuration details available upon request.

intel AI portfolio

experiences



tools



Intel® Deep Learning SDK

Intel® Computer Vision SDK

Movidius Neural Compute Stick

saffron
TECHNOLOGY
an intel company

Frameworks

APACHE
Spark
MLlib
BigDL

TensorFlow

mxnet

theano

neon

torch

Caffe

E2E Tool

libraries



Intel Dist

Intel® DAAL

Intel® Nervana™ Graph*

Intel® MKL MKL-DNN Intel® MSL

Movidius
MvTensor
Library

Associative
Memory Base

hardware



Compute

Memory & Storage



Networking

intel REALSENSE
TECHNOLOGY
Movidius
an intel company

Visual Intelligence

Unleash
Full
potential


*Coming 2017

Optimization Notice

Copyright © 2016 Intel Corporation. All rights reserved.
Other names and brands may be claimed as the property of others.





Tools & Frameworks portfolio

Intel® Math Kernel Library
Intel® MKL  MKL-DNN

Intel® Data Analytics Acceleration Library (DAAL)

Intel® Distribution for python™ 

Open Source Frameworks
 theano  Spark
Caffe Mor e...

Intel Deep Learning Tools 

High Level Overview

High performance math primitives granting low level of control

Free open source DNN functions for high-velocity integration with deep learning frameworks

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

Primary Audience

Consumed by developers of higher level libraries and Applications

Consumed by developers of the next generation of deep learning frameworks

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

Example Usage

Framework developers call matrix multiplication, convolution functions

New framework with functions developers call for max CPU performance

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

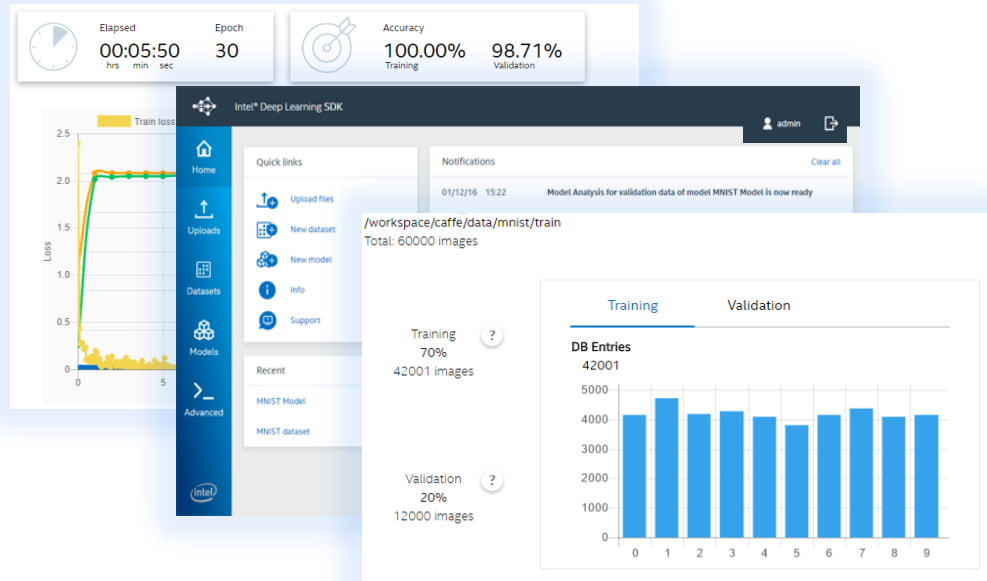
Intel® Deep Learning SDK

BETA: Now
GOLD: Q1'17

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

Optimization Notice

Copyright © 2016, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

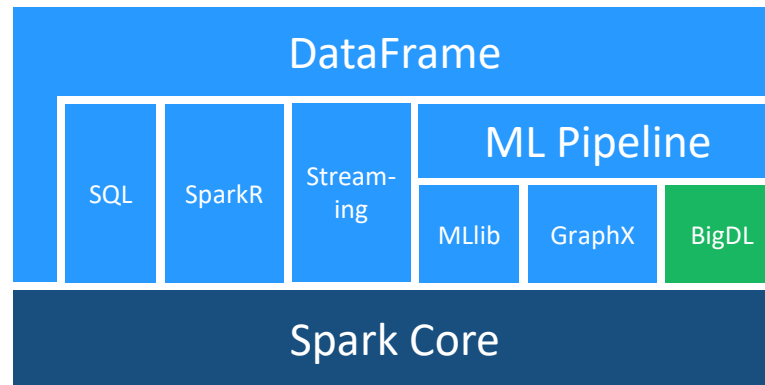


BIGDL

Bringing Deep Learning to Big Data

For developers looking to run deep learning on Hadoop/Spark due to familiarity or analytics use

- Open Sourced Deep Learning Library for Apache Spark*
- Make Deep learning more Accessible to Big data users and data scientists.
- Feature Parity with popular DL frameworks like Caffe, Torch, Tensorflow etc.
- Easy Customer and Developer Experience
 - Run Deep learning Applications as Standard Spark programs;
 - Run on top of existing Spark/Hadoop clusters (No Cluster change)
- High Performance powered by Intel MKL and Multi-threaded programming.
- Efficient Scale out leveraging Spark architecture.



github.com/intel-analytics/BigDL

Optimization Notice

Copyright © 2016, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.



Intel distribution for python

Advancing Python Performance Closer to Native Speeds



For developers using the most popular and fastest growing programming language for AI

Easy, Out-of-the-box Access to High Performance Python

- Prebuilt, optimized for numerical computing, data analytics, HPC
- Drop in replacement for your existing Python (no code changes required)

Drive Performance with Multiple Optimization Techniques

- Accelerated NumPy/SciPy/Scikit-Learn with Intel® MKL
- Data analytics with pyDAAL, enhanced thread scheduling with TBB, Jupyter* Notebook interface, Numba, Cython
- Scale easily with optimized MPI4Py and Jupyter notebooks

Faster Access to Latest Optimizations for Intel Architecture

- Distribution and individual optimized packages available through conda and Anaconda Cloud
- Optimizations upstreamed back to main Python trunk

software.intel.com/intel-distribution-for-python

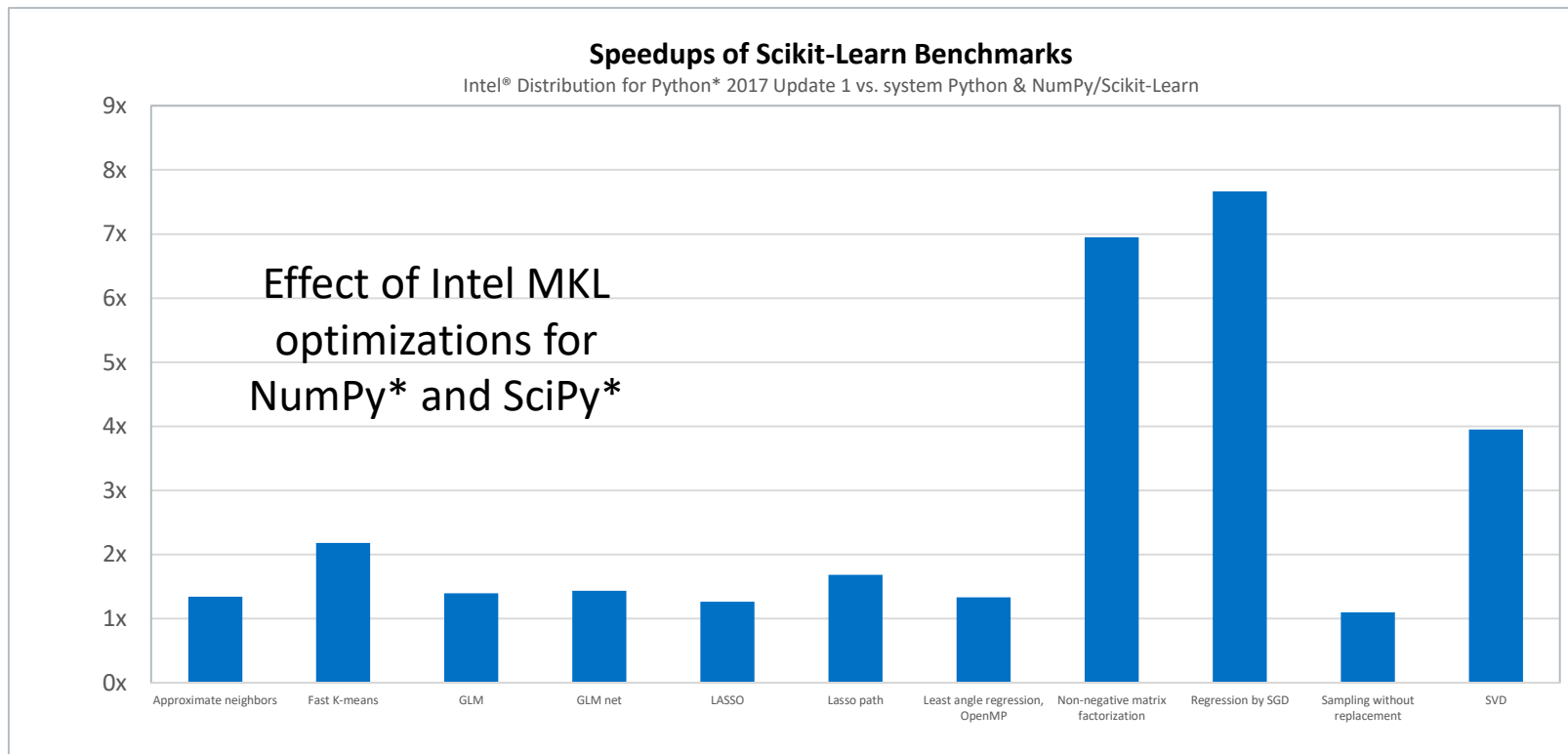
Easy to install with Anaconda*
<https://anaconda.org/intel/>

Optimization Notice

Copyright © 2016, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.



Sket-Learn* Optimizations With Intel® MKL



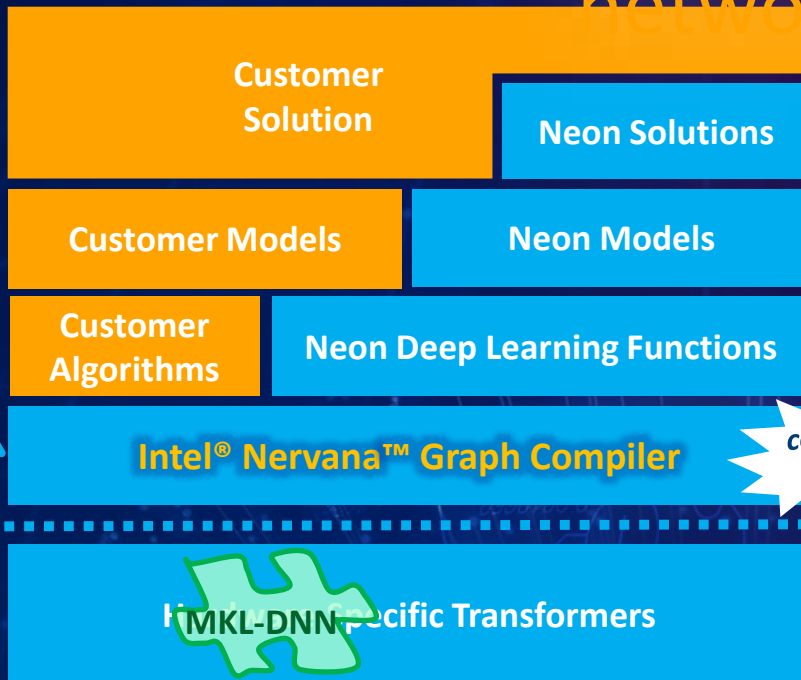
System info: 32x Intel® Xeon® CPU E5-2698 v3 @ 2.30GHz, disabled HT, 64GB RAM; Intel® Distribution for Python* 2017 Gold; Intel® MKL 2017.0.0; Ubuntu 14.04.4 LTS; Numpy 1.11.1; scikit-learn 0.17.1.

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. * Other brands and names are the property of their respective owners. Benchmark Source: Intel Corporation

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 intend to ensure the 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. All other trademarks are the property of their respective owners. Intel Confidential | NDA Required

Coming soon: intel[®] Nervana[™]

graph compiler High-level execution graph for neural networks



Intel[®] Nervana[™] Graph Compiler

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

Intel® Nervana™ AI academy

hone your skills and build the future of ai



Frameworks, Tools and
Libraries
Software innovators and
Black Belts
Workshops, webinars,
meetups

in partnership with

coursera

kaggle™

software.intel.com/ai

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

Optimized Mathematical Building Blocks

Intel® MKL

Linear Algebra

- BLAS
- LAPACK
- ScaLAPACK
- Sparse BLAS
- PARDISO* SMP & Cluster
- Iterative sparse solvers

Fast Fourier Transforms

- Multidimensional
- FFTW interfaces
- Cluster FFT

Vector Math

- Trigonometric
- Hyperbolic
- Exponential
- Log
- Power
- Root

Vector RNGs

- Congruential
- Wichmann-Hill
- Mersenne Twister
- Sobol
- Neiderreiter
- Non-deterministic

Summary Statistics

- Kurtosis
- Variation coefficient
- Order statistics
- Min/max
- Variance-covariance

And More

- Splines
- Interpolation
- Trust Region
- Fast Poisson Solver

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

Intel[®] MKL-DNN

Math Kernel Library for Deep Neural Networks

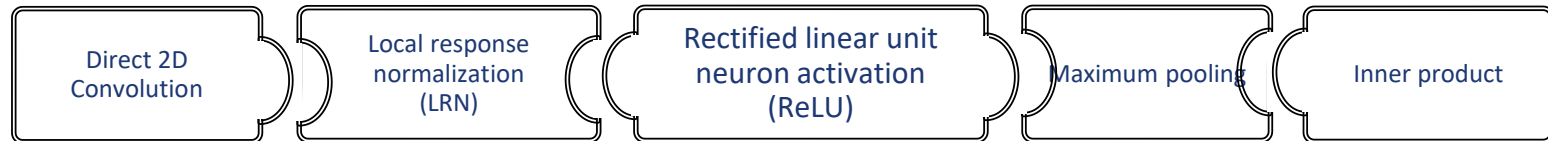
For developers of deep learning frameworks featuring optimized performance on Intel hardware

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[®] MKL library.



github.com/01org/mkl-dnn



Intel® Machine learning scaling library (MLSL)

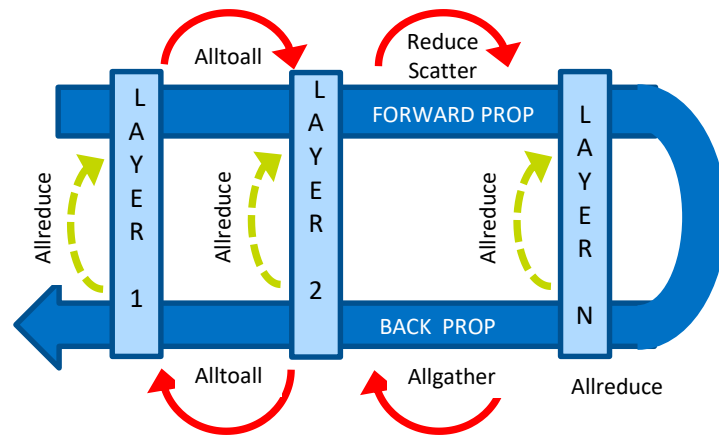
Scaling Deep Learning to 32 Nodes and Beyond

For maximum deep learning scale-out performance on Intel® architecture

BETA Now Available!

Deep learning abstraction of message-passing implementation

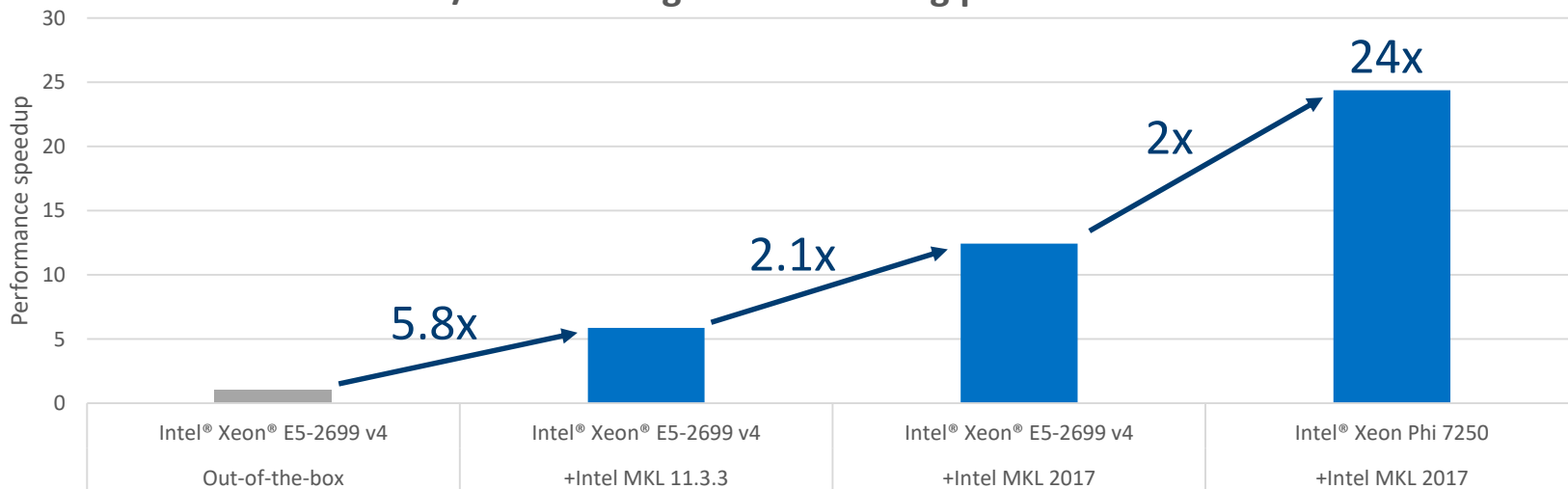
- Built on top of MPI; allows other communication libraries to be used as well
- Optimized to drive scalability of communication patterns
- Works across various interconnects: Intel® Omni-Path Architecture, InfiniBand, and Ethernet
- Common API to support Deep Learning frameworks (Caffe, Theano, Torch etc.)



github.com/01org/MLSL/releases

Better performance in Deep Neural Network workloads with Intel® Math Kernel Library (Intel® MKL)

Caffe/AlexNet single node training 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 information go to <http://www.intel.com/performance>. *Other names and brands may be property of others

Configurations:

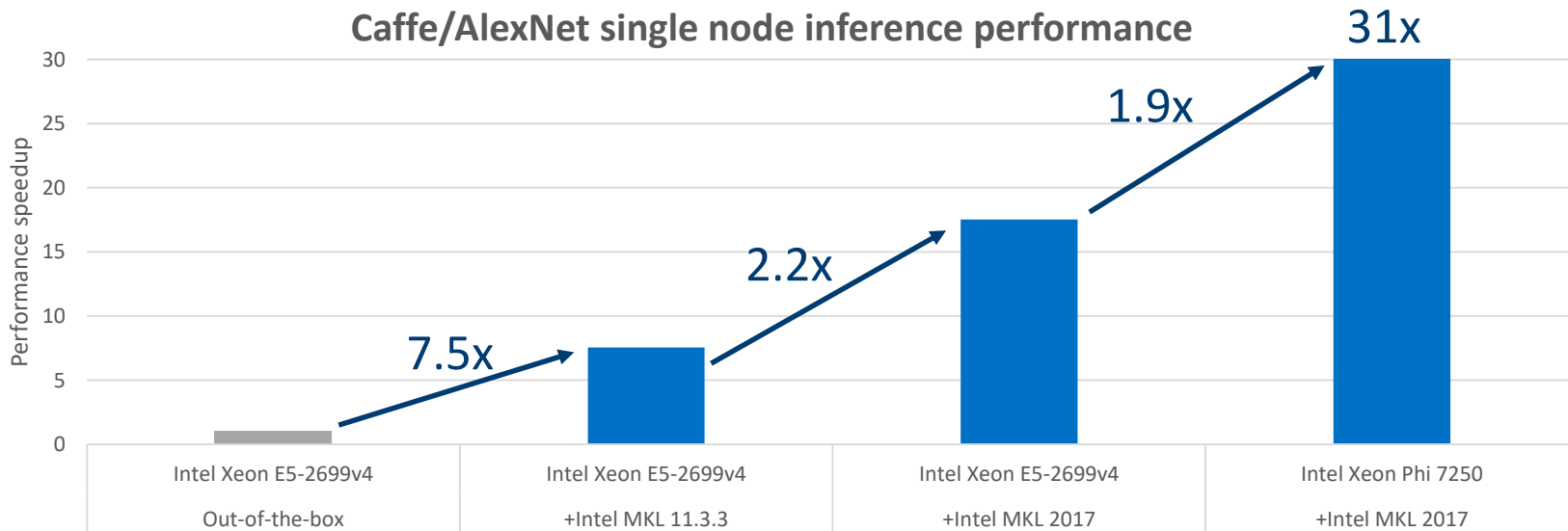
- 2 socket system with Intel® Xeon Processor E5-2699 v4 (22 Cores, 2.2 GHz.), 128 GB memory, Red Hat® Enterprise Linux 6.7, [BVLC Caffe](#), [Intel Optimized Caffe framework](#), Intel® MKL 11.3.3, Intel® MKL 2017
- Intel® Xeon Phi™ Processor 7250 (68 Cores, 1.4 GHz, 16GB MCDRAM), 128 GB memory, Red Hat® Enterprise Linux 6.7, [Intel® Optimized Caffe framework](#), Intel® MKL 2017

Optimization Notice

Copyright © 2016, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.



Better performance in Deep Neural Network workloads with Intel® Math Kernel Library (Intel® MKL)



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:

- 2 socket system with Intel® Xeon® Processor E5-2699 v4 (22 Cores, 2.2 GHz), 128 GB memory, Red Hat® Enterprise Linux 6.7, [BVLC Caffe](#), [Intel Optimized Caffe framework](#), Intel® MKL 11.3.3, Intel® MKL 2017
- Intel® Xeon Phi™ Processor 7250 (68 Cores, 1.4 GHz, 16GB MCDRAM), 128 GB memory, Red Hat® Enterprise Linux 6.7, [Intel® Optimized Caffe framework](#), Intel® MKL 2017

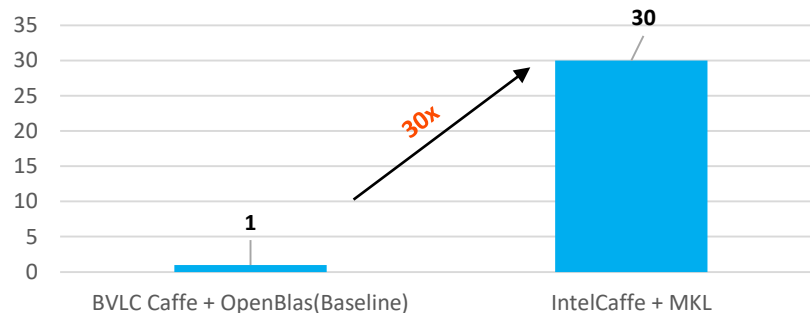
All numbers measured without taking data manipulation into account.

Optimization Notice

Case Study: LeTV Cloud Illegal Video Detection

- LeTV Cloud (www.lecloud.com) is a leading video cloud provider in China
- LeTV Cloud provides illegal video detection service to 3rd party video cloud customers to help them detect illegal videos
- Originally, LeTV adopted open source BVLC Caffe plus OpenBlas as CNN framework, but the performance was poor
- By using Caffe + Intel MKL, they gained up to 30x performance improvement on training in production environment

LeTV Cloud Caffe Optimization - higher is better



* The test data is based on Intel Xeon E5 2680 V3 processor

LeTV Cloud Illegal Video Detection Process Flow

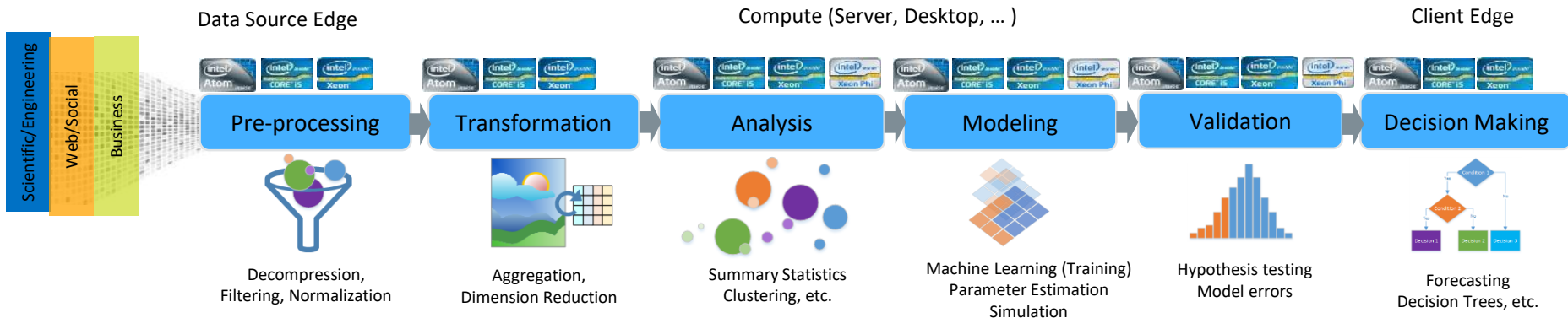


Other names and brands may be claimed as property of others. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations 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>

Optimization Notice

Intel® DAAL: high level view

- Library of optimized building blocks covering all stages of the data analysis, from data extraction till data-driven decisions
- Targets both data centers (Intel® Xeon® and Intel® Xeon Phi™) and edge-devices (Intel® Atom)
 - Perform analysis close to data source (sensor/client/server) to optimize response latency, decrease network bandwidth utilization, and maximize security.
 - Offload data to server/cluster for complex and large-scale analytics only.



Optimization Notice

Copyright © 2016, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.



Get started today

Get started today

Frameworks optimized for Intel

Caffe

[Build Faster Deep Learning Applications with Caffe*](#)

The popular open-source development framework for image recognition is now optimized for Intel® Architecture.

[Get the framework](#)

[Learn how to install Caffe*](#)

theano

[Delving Into Deep Learning](#)

The Python library designed to help write deep learning models is now optimized for Intel® Architecture.

[Visit the library](#)

[Getting started with Theano*](#)



BigDL: Distributed Deep learning on Apache Spark



[Speed Up Your Spark Analytics](#)

Apache Spark* MLlib, the open-source data processing framework's machine learning library, now includes Intel® Architecture support.

[Get the library](#)

[Building faster applications on Spark clusters](#)

[BigDL on GitHub](#)

[BigDL on Spark video](#)

[Running on EC2 Page](#)

Optimization Notice

Copyright © 2016, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.



Get started today

Get Intel® Libraries (Community License) for Free



Intel® Data Analytics Acceleration Library [Intel® DAAL](#)

Highly optimized library that helps speed big data analytics by providing algorithmic building blocks for all data analysis stages and for offline, streaming, and distributed analytics usages.

[Open-source options for Intel® DAAL](#)

[Learn more about Intel® DAAL](#)



Intel® Math Kernel Library [Intel® MKL](#)

A high-performance library with assets to help accelerate math processing routines and increase application performance.

[Deep neural network technical preview for Intel® MKL](#)

[Get the library](#)

Training

- [Accelerating Deep Learning and Machine Learning](#)

This talk focuses on two Intel performance libraries, MKL and DAAL, which offer optimized building blocks for data analytics and machine learning.

- [Remove Python Performance Barriers for Machine Learning](#)

This webinar highlights significant performance speed-ups achieved by implementing multiple Intel tools and techniques for high performance Python.

- [Analyze Python* App Performance with Intel® VTune™ Amplifier](#)

Efficient profiling techniques can help dramatically improve the performance of your Python* code. Learn how Intel® VTune Amplifier can help.

- [Boost Python* Performance with Intel® Math Kernel Library](#)

Meet Intel® Distribution for Python*, an easy-to-install, optimized Python distribution that can help you optimize your app's performance.

- [Building Faster Data Applications on Spark* Clusters](#)

Apache Spark* is big for big data processing apps. Intel® Data Analytics Acceleration Library (Intel® DAAL) can help optimize performance. Learn how.

- [Faster Big Data Analytics Using New Intel® Data Analytics Acceleration Library](#)

Big data is BIG. And you need information faster. New Intel® Data Analytics Acceleration Library (Intel® DAAL) speeds data processing for data mining, statistical analysis, and machine learning.

Resources

Intel® Machine Learning

- <http://www.intel.com/content/www/us/en/analytics/machine-learning/overview.html>
- Intel Caffe* fork, <https://github.com/intelcaffe/caffe>
- Intel Theano* fork, <https://github.com/intel/theano>
- Intel® Deep Learning SDK: <https://software.intel.com/deep-learning-SDK>

Intel® DAAL

- <https://software.intel.com/en-us/intel-daal>

Intel® Omni-Path Architecture

- <http://www.intel.com/content/www/us/en/high-performance-computing-fabrics/omni-path-architecture-fabric-overview.html>

Intel(R) MKL Resources

Intel® MKL website

- <https://software.intel.com/en-us/intel-mkl>

Intel® MKL forum

- <https://software.intel.com/en-us/forums/intel-math-kernel-library>

Intel® MKL benchmarks

- <https://software.intel.com/en-us/intel-mkl/benchmarks#>

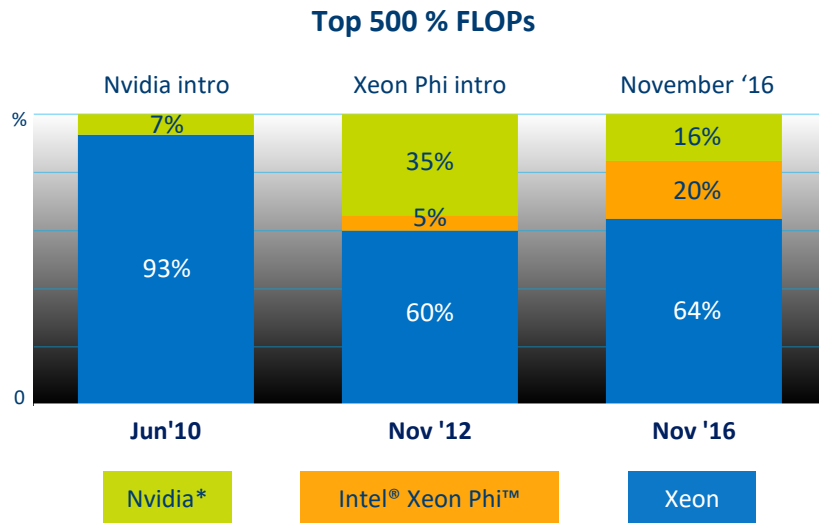
Intel® MKL link line advisor

- <http://software.intel.com/en-us/articles/intel-mkl-link-line-advisor/>

performance

Artificial Intelligence Plan

Bringing the HPC Strategy to AI



Intel® Nervana™ Portfolio



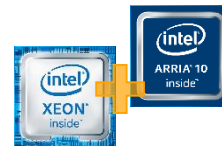
Most widely deployed machine learning solution

COMING 2017
SKYLAKE



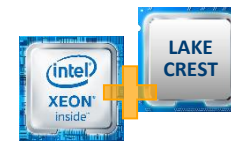
High performance, classic machine learning

COMING 2017
KNIGHTS MILL



Programmable, low-latency inference

SDVs SHIPPING TODAY
BROADWELL + ARRIA 10



Best in class neural network performance

COMING 2017
LAKE CREST



Optimization Notice

Copyright © 2016 Intel Corporation. All rights reserved. Other names and brands may be claimed as the property of others.

Intel® Xeon® Processor Family

Most Widely Deployed Machine Learning



Lowest TCO With Superior Infrastructure Flexibility

- Standard server infrastructure
- Open standards, libraries & frameworks
- Optimized to run wide variety of data center workloads

Server Class Reliability

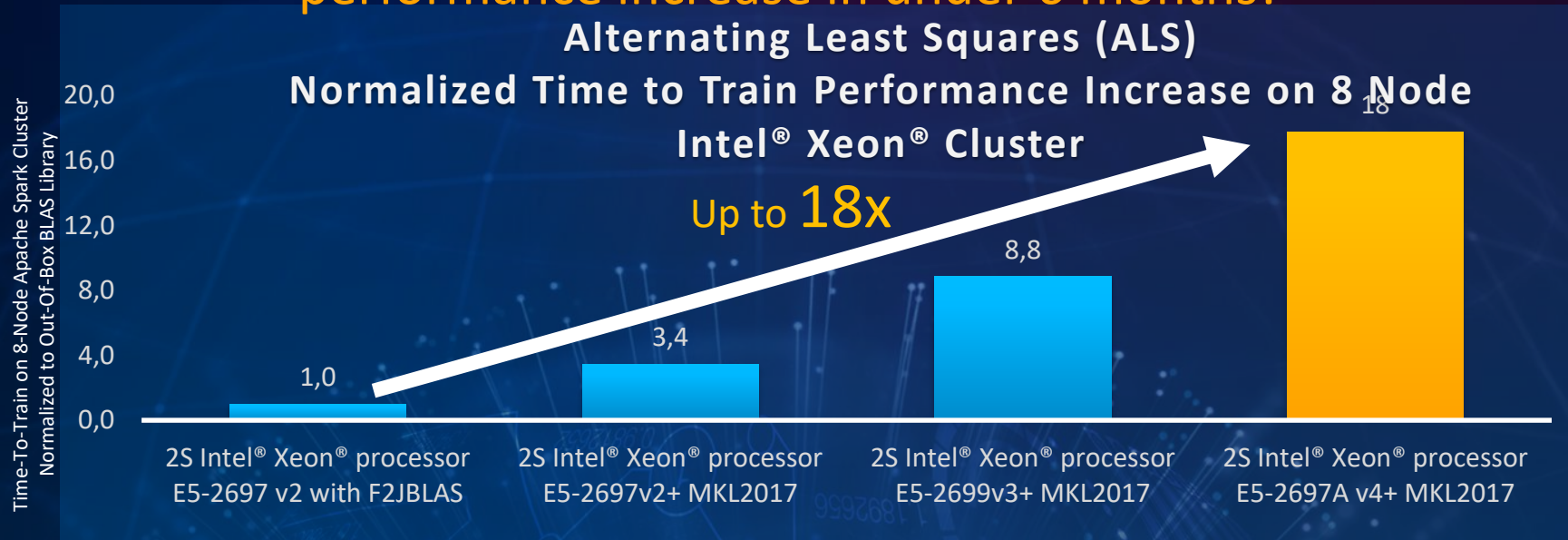
- Industry standard server features: high reliability, hardware enhanced security

Leadership Throughput

- Industry leading inference performance

Intel® Xeon® processor Performance

Increasing customer value on existing systems through optimizing machine learning algorithms for standard cpu architecture: up to 18x performance increase in under 6 months!



Higher is better

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 Family

Enables shorter time to train using general purpose
infrastructure



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

- Near linear scaling with **31X** reduction in time to train when scaling to 32 nodes
- Up to **400X** performance on existing hardware via Intel software optimizations
- Up to **4X** deep learning performance increase estimated on Knights Mill (2017)

Easier Programmability

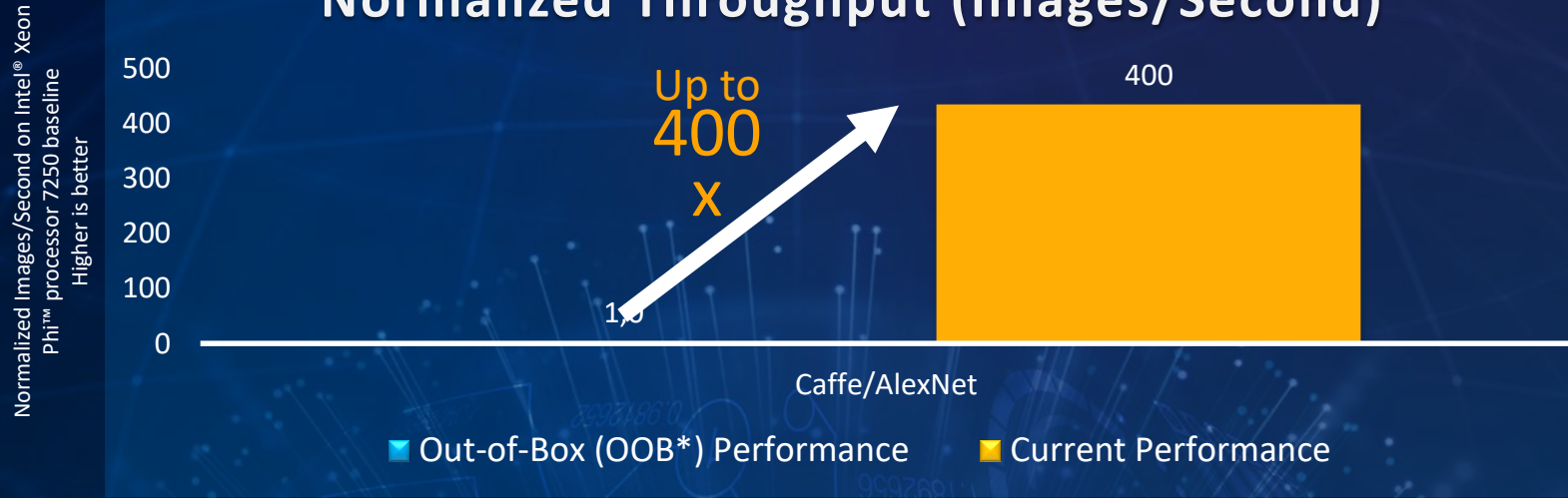
- Binary-compatible with Intel® Xeon® processors
- Open standards, libraries and frameworks

Intel® Xeon Phi™ processor

Shattering misconceptions that cpus are not well-suited for deep learning:
performance

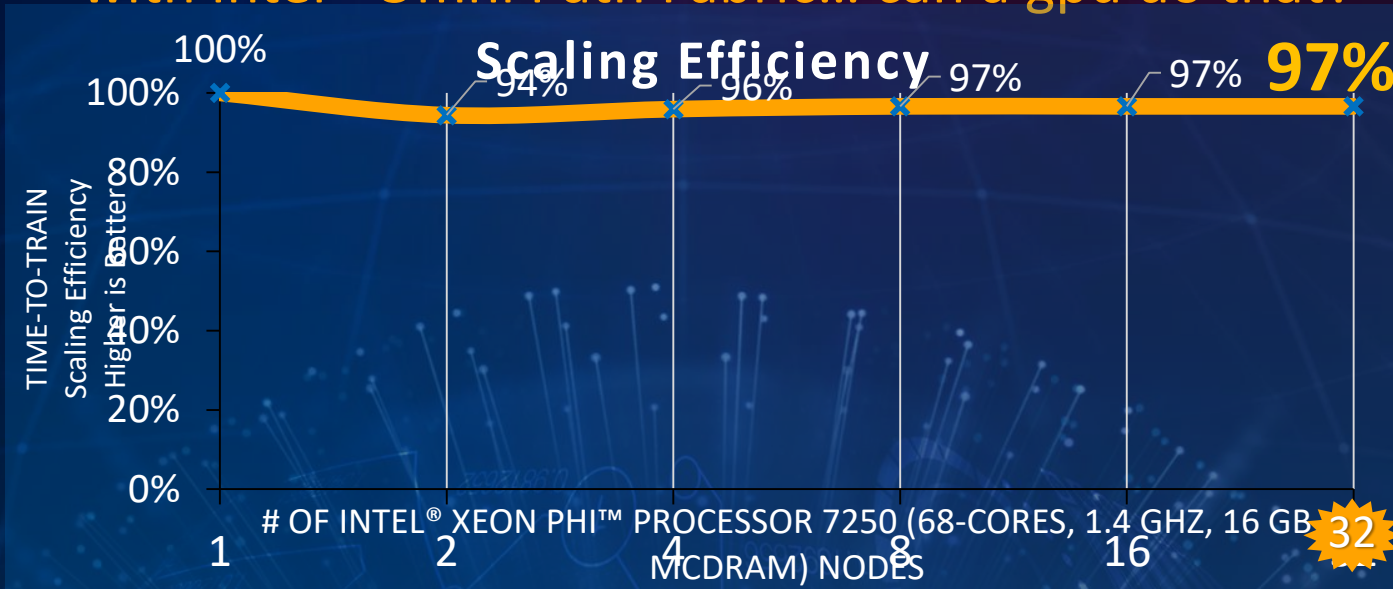
SW optimization for cpu deliver up to 400X performance gain on existing Hardware in <6 months

Normalized Throughput (Images/Second)



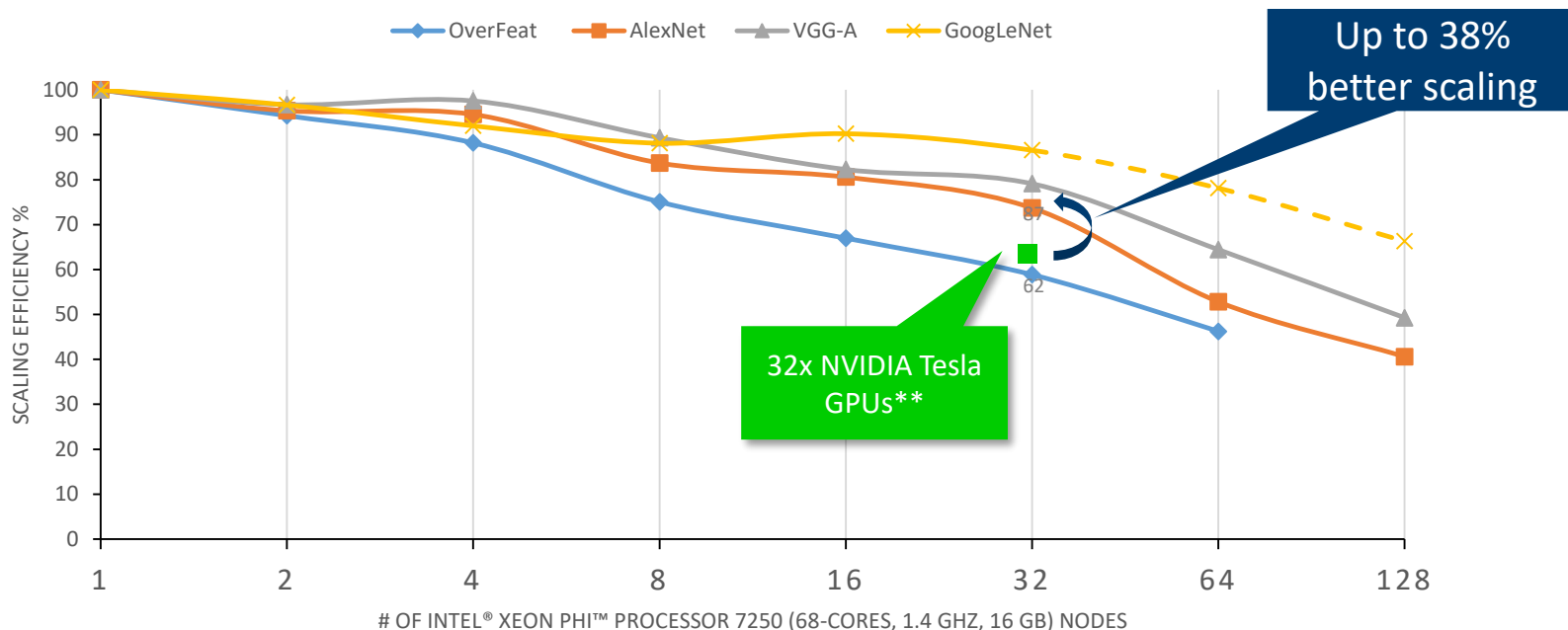
Intel® Xeon Phi™ processor

performance
up to 97% Scaling Efficiency enables faster training on GoogleNet V1 using a 32 node cluster of Intel® Xeon Phi™ Processor 7250 with Intel® Omni Path Fabric... can a gpu do that?



Better Scaling Efficiency: Intel® Xeon Phi™ Processor

Deep Learning Image Classification Training Performance - MULTI-NODE Scaling



Dataset: Large image database

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:
* Intel® Xeon Phi™ Processor 7250 (68 Cores, 1.4 GHz, 16GB MCDRAM), 128 GB memory, Red Hat® Enterprise Linux 6.7, Intel® Optimized Frameworks
**Source: <http://arxiv.org/abs/1511.00175> showing FireCaffe* with 32x NVIDIA® K20s (Titan Supercomputer*) running GoogLeNet* at 20x speedup over Caffe* with 1x K20

Optimization Notice

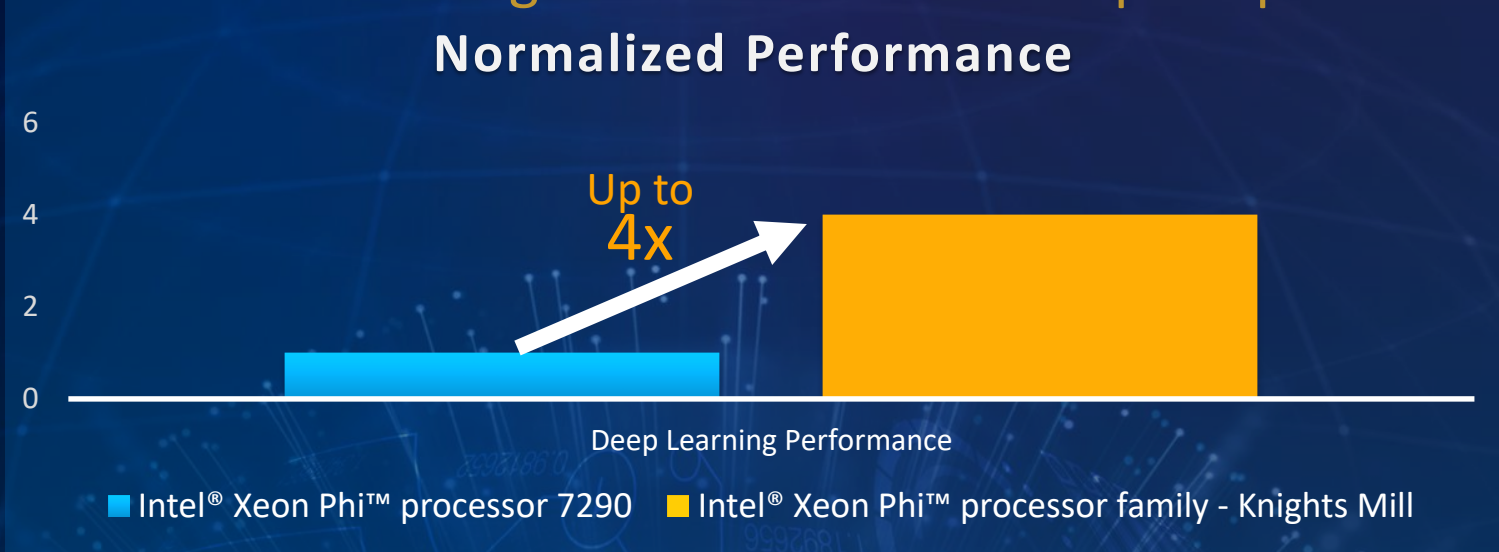
Copyright © 2016, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.



Intel® Xeon Phi™ processor

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

Estimated normalized performance on Intel® Xeon Phi™ processor 7290 compared to Intel® Xeon Phi™ Knights Mill



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.

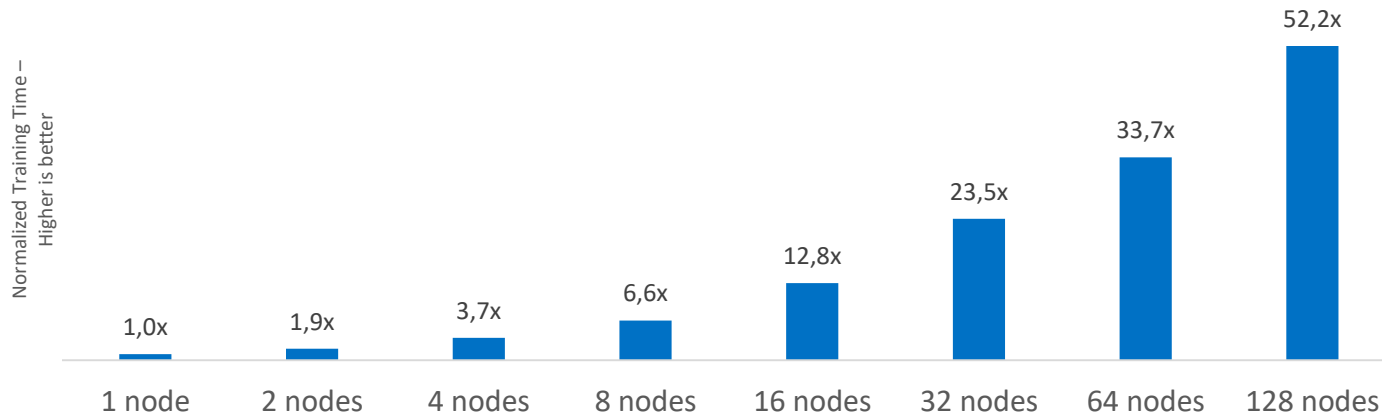
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobliMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may 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

Train Up to 50x faster with Intel® Xeon Phi™ Processor

Deep Learning Image Classification Training Performance - MULTI-NODE Scaling



Topology: **AlexNet***

Dataset: **Large image database**

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/datacenter>. Configurations: 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>.

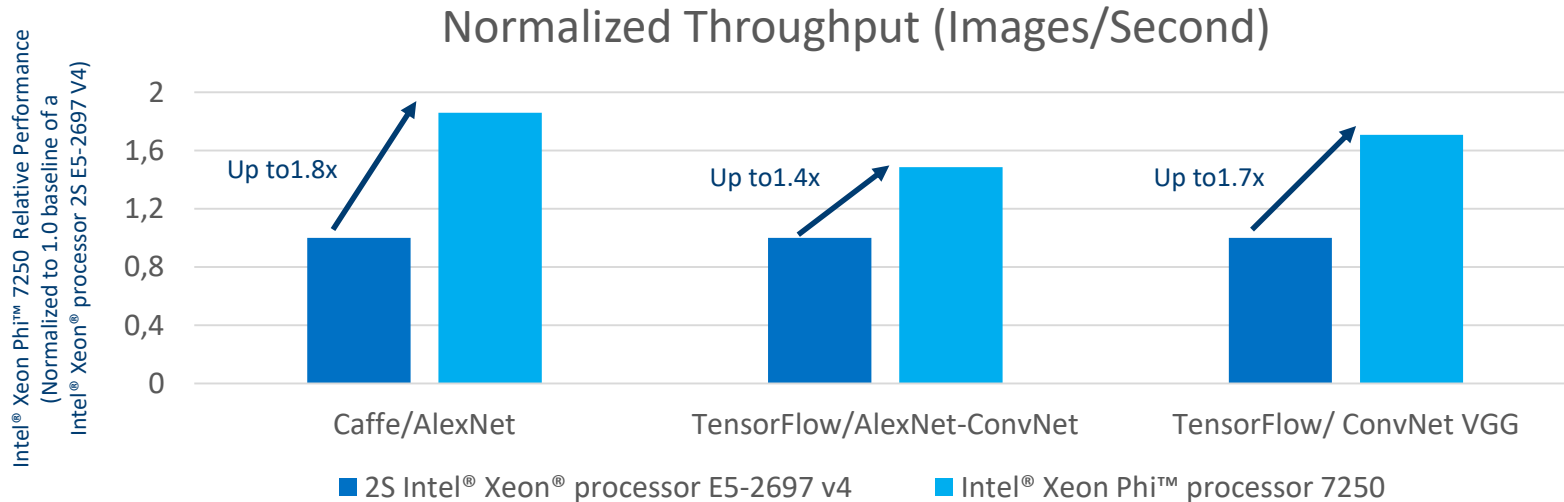
Optimization Notice

Copyright © 2016, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.



Intel® Xeon Phi™ Processor Image Classification Training Throughput

Single node: 1s Xeon Phi up to 1.8x better than two Intel® Xeon® processor E5-2697v4



Configuration details on slide: 12

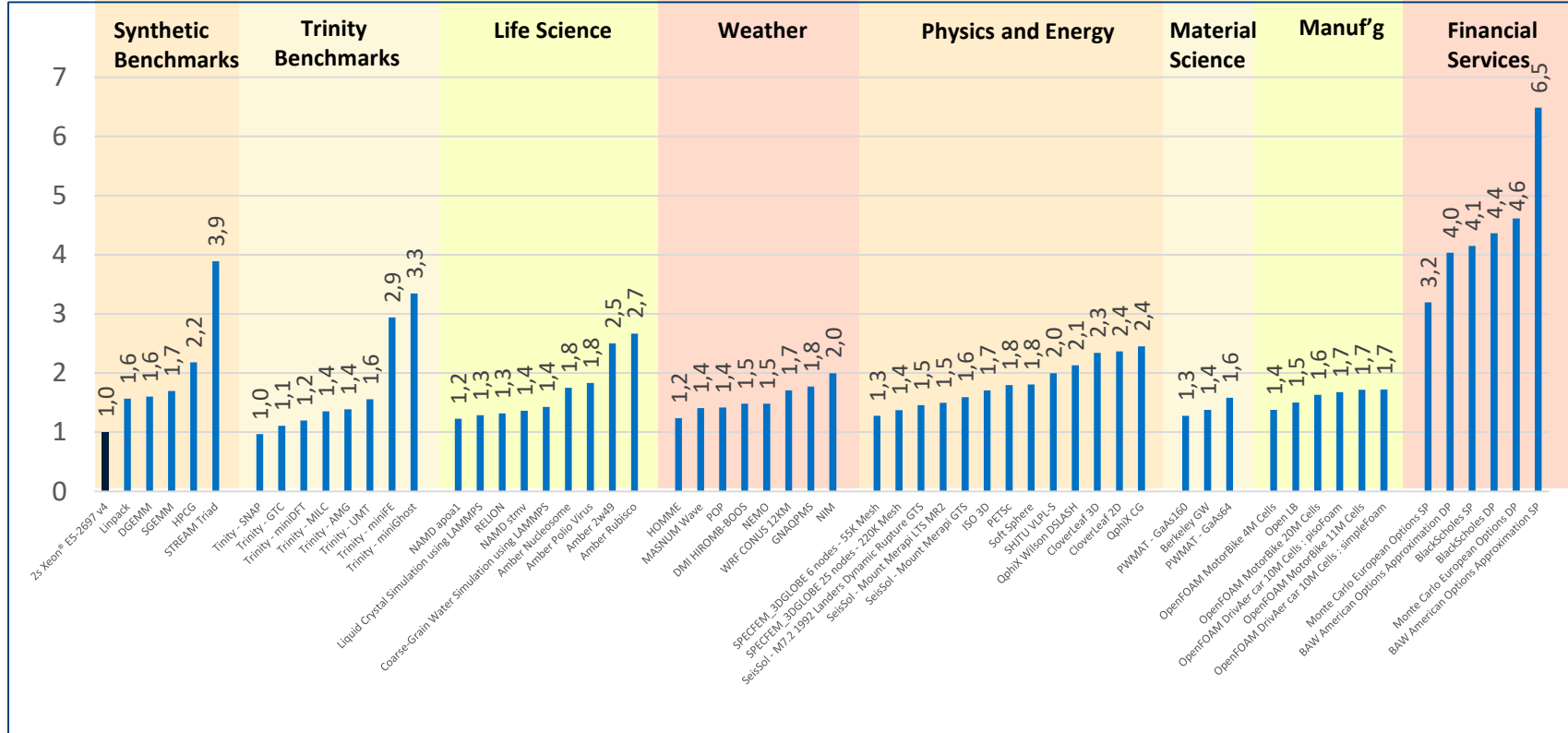
Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, 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

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

Intel® Xeon Phi™ Processor Increases Customer Value through More Cores, Wider Vectors, and Memory BW



Intel® Xeon Phi™ Processor 7250
 Relative Performance (Higher is better)
 (Normalized to 1.0 baseline of a 2 Socket
 Intel® Xeon® processor E5-2697 v4)



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 May 2016.

Optimization Notice

Copyright © 2016, Intel Corporation. All rights reserved.
 *Other names and brands may be claimed as the property of others.





Code Modernization initiatives in the Brazilian HPC Ecosystem

Oil & Gas - Reservoir Simulator at PETROBRAS

- Up to 6.5x performance gains in their Reservoir Simulator software¹



Laboratório Nacional de Computação Científica

Health & Life Sciences

- On-going white-paper in Molecular Dynamics software with LNCC
- Partial results*
- Xeon only:
 - Original code vs Modernized code: up to 11x speedup
- Xeon + 1 Xeon Phi (same optimized code)
 - 1.14x speedup



LNCC - National Laboratory for Scientific Computing
Largest HPC cluster in Latin America

- Up to 30x performance gain in Oil & Gas applications²



Laboratório Nacional de Computação Científica

NCC / UNESP
An Intel® Modern Code Partner



- 11 HPC Hands-on Workshops so far
- 576+ developers trained so far

Authors:

¹CENPES team and Gilvan Vieira - gilvandsv@gmail.com

²LNCC - Frederico Cabral - fredluiscabral@gmail.com

³NCC/UNESP - Silvio Stanzani - silvio.stanzani@gmail.com

Backup – architecture details

Intel® Xeon Phi™ Processor

(Knights Landing)



Self-Boot Processor

Binary-compatibility with Xeon, 3+ TFLOPS¹ (DP)

On-package memory

16GB, Up to 490 GB/s STREAM TRIAD

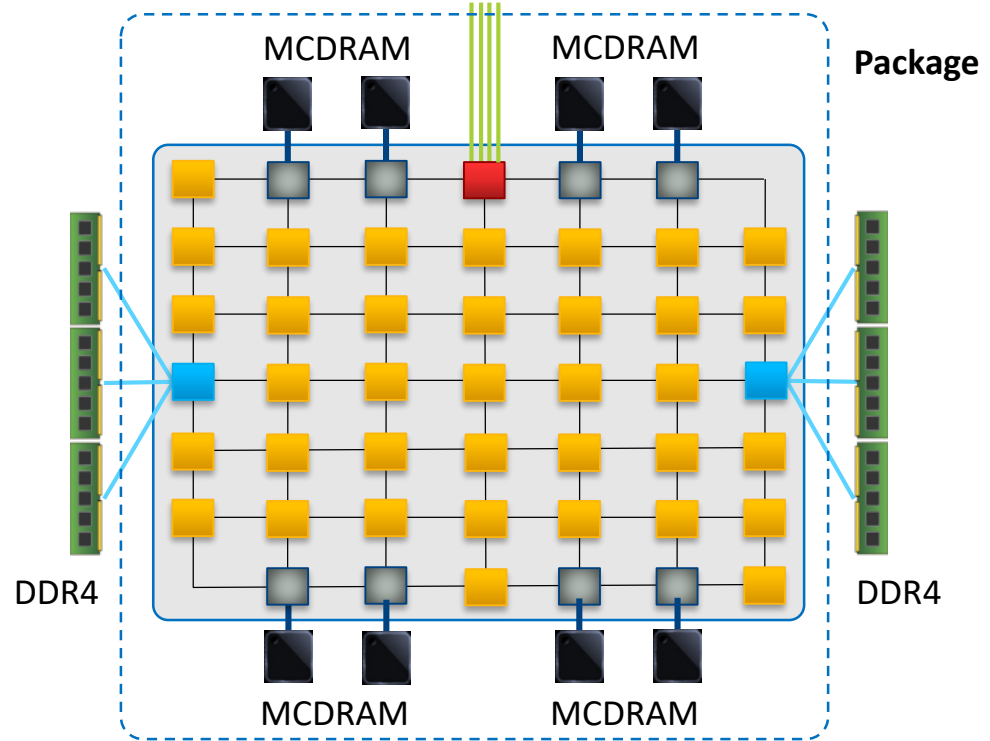
Platform Memory

Up to 384GB (6ch DDR4-2400 MHz)

Other Key Features

- ✓ 2D Mesh Architecture
- ✓ Out-of-Order Cores
- ✓ 3X Single-Thread vs. KNC
- ✓ Intel® AVX-512 Instructions
- ✓ Scatter/Gather Engine
- ✓ Integrated Fabric - OPA

x4 DMI2 to PCH
36 Lanes PCIe* Gen3 (x16, x16, x4)



TILE:
(up to 36)

2VPU	HUB	2VPU
Core	1MB L2	Core



Tile



EDC (Embedded DRAM Controller)



IMC (Integrated Memory Controller)



IIO (Integrated I/O Controller)

¹Theoretical peak performance

Optimization Notice

Copyright © 2016, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.



Knights Mill & Groveport Platform Overview



1 Trains Machines Faster

- Up to 2.5X* Single Precision performance over Knights Landing for deep learning workloads
- Industry leader variable precision QVNNI up to 4X* faster performance
- Highly distributed multi-node scaling

2 Memory Flexibility & Bootable Host-CPU

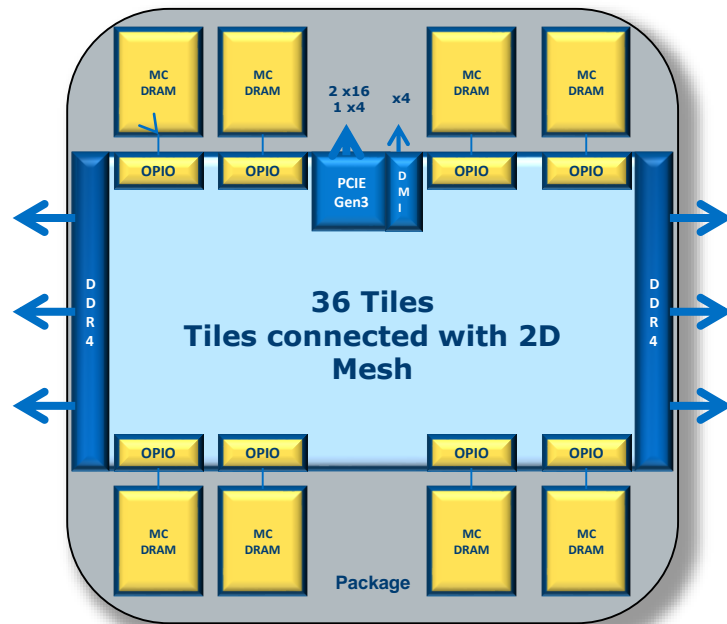
- High memory bandwidth with integrated 16GB MC DRAM and bootable host-CPU reduces offloading & latency challenges
- 384GB 6-channel DDR4 memory capacity for massive AI use cases

3 Consistent Programming Models

- Common Intel® Xeon® & Intel® Xeon Phi™ programming
- Optimized for industry standard Open Source ML frameworks
- Flexibility to run vast workloads across x86 infrastructure

*NOTE: Performance theoretical wrt KNL7250 SKU based on KNM architectural changes.

Groveport Platform Bootable Host CPU



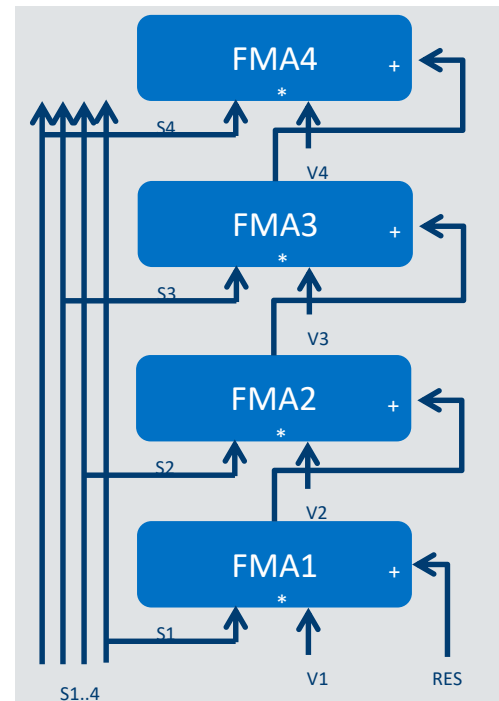
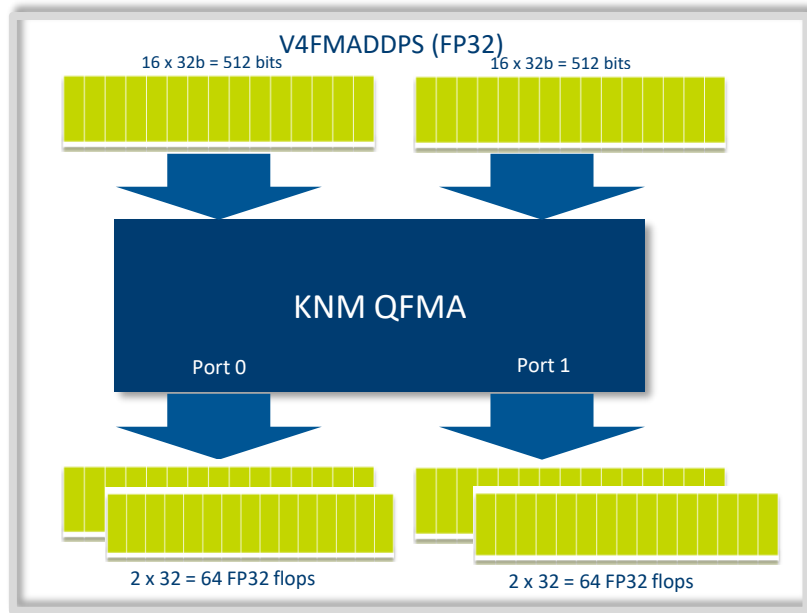
Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks. Performance estimate wrt KNL 7250 SKU SGEMM. Performance Calculation= AVX freq X Cores X Flops per Core X Efficiency

Optimization Notice

Knights Mill QFMA for Faster Performance

Enhanced ISA QFMA instructions in Knights Mill delivers:

- ✓ Higher Peak Flops for CNN, RNN, DNN, LSTM
- ✓ Higher Efficiency (One Quad FMA executed in two cycles)
- ✓ 2X FP operations per cycle



QMADD packs 4 IEEE FMA ops in a single instruction

*2X faster than KNL SP

Optimization Notice

Copyright © 2016, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

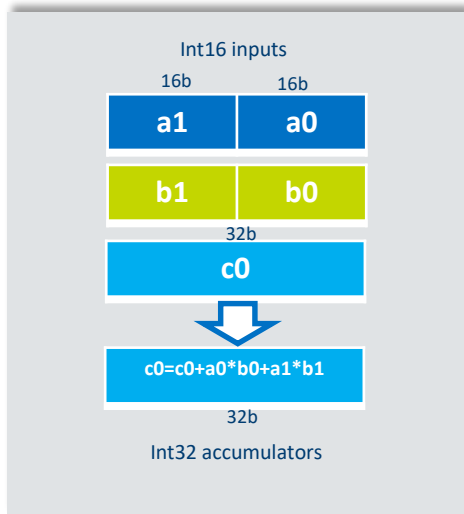


Knights Mill Variable Precision Performance

Enabling Faster Throughput for Machine Learning Training

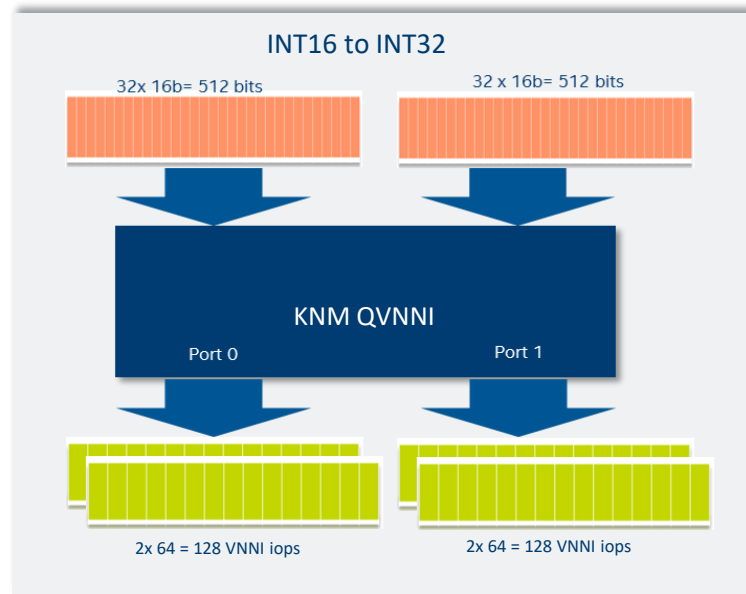
VNNI

- 2x the flops by using INT16 inputs
- Similar accuracy as SP by using INT32 accumulated output



QVNNI

- 2x VNNI operations per port
- 4x* ML performance than regular AVX512-SP



*4x faster than KNL SP

Optimization Notice

Copyright © 2016, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

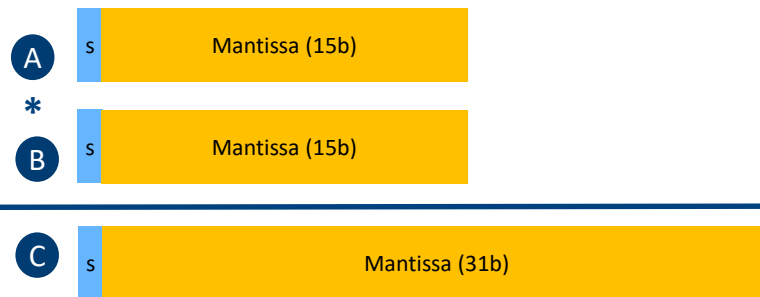


Knights Mill QVNNI Advantages over FP16

FP16



Intel® QVNNI



Flexpoint operation



QVNNI for Higher Accuracy and Faster Operations

Optimization Notice

Copyright © 2016, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.



Knights Hill Processor Developments



CPU – fabric integration

- Direct access to KNH CPU resources
- Improved fabric latency
- Lower cost and improved density opportunity



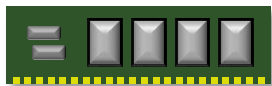
Enhanced performance

- Huge leap in Dual Precision vector performance
- Dramatic leap in Single & 16-bit performance
- High density system options
- Improved Intel® OPA fabric bandwidth



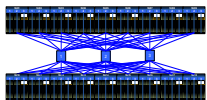
Reduced costs

- TCO via Performance/Watt
- Faster time to solution
- Higher radix Omni-Path switches



Memory

- Higher capacity and bandwidth in package memory
- Innovations in 3D XPoint™ technology support



Scaling

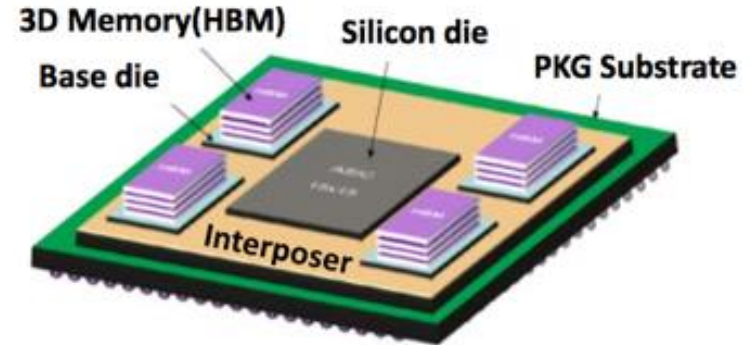
- Emphasis on reliability and resiliency
- Storm Lake 2 scaling support for 100K nodes

Optimization Notice

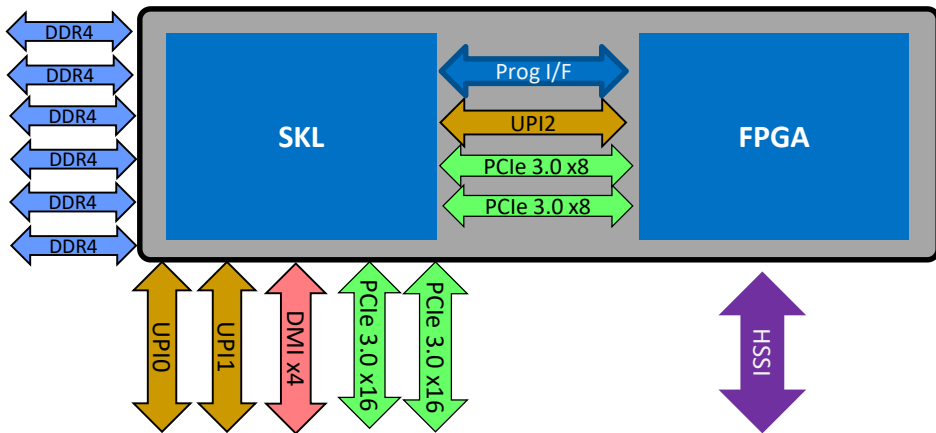
Intel Nervana Engine (Lake Crest)

Nervana Engine

- ~55 Tera Ops
- Patented ***FlexPoint*** precision for maximum Tput and high accuracy
- Over Tera b/s of inter/intra connectivity for optimal scaling
- 32GB HBM
- Standard PCIe Gen3x16 AiC (“like GPU”)
- Platform design for 4-8 AiCs per 3U-4U chassis



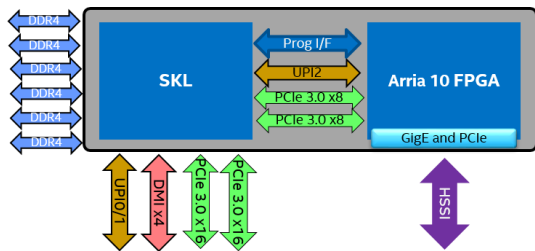
Skylake + FPGA on Purley



- Power for FPGA is drawn from socket & requires modified Purley platform specs
- Platform Modifications include Stackup, Clock, Power Delivery, Debug, Power up/down sequence, Misc IO pins (see BOM cost section)

Cores	Up to 28C with Intel® HT Technology	
FPGA	Altera® Arria 10 GX 1150	
Socket TDP	Shared socket TDP of 165W combined, or Up to 165W SKL & Up to 90W FPGA	
Socket	Socket P	
Scalability	Up to 2S – with SKL-SP or SKL + FPGA SKUs	
PCH	Lewisburg: DMI3 – 4 lanes; 14xUSB2 ports Up to: 10xUSB3; 14xSATA3, 20xPCIe*3 New: Innovation Engine, 4x10GbE ports, Intel® QuickAssist Technology	
	For CPU	For FPGA
Memory	6 channels DDR4 RDIMM, LRDIMM, Apache Pass DIMMs 2666 1DPC, 2133, 2400 2DPC	Low latency access to system memory via UPI & PCIe interconnect
Intel® UPI	2 channels (10.4, 9.6 GT/s)	1 channel (9.6 GT/s)
PCIe*	PCIe* 3.0 (8.0, 5.0, 2.5 GT/s)	PCIe* 3.0 (8.0, 5.0, 2.5 GT/s)
	32 lanes per CPU Bifurcation support: x16, x8, x4	16 lanes per FPGA Bifurcation support: x8
High Speed Serial Interface (Different board design based on HSSI config)	N/A	2xPCIe 3.0 x8
		Direct Ethernet (4x10 GbE, 2x40 GbE, 10x10 GbE, 2x25 GbE)

SKL+FPGA Customer Profile



Application Development Method



New - Library Approach



New - Library Approach



New - Extended IA Flow



Traditional Flow (RTL or OpenCL)

Customer Profile

Target Customer: No previous FPGA or hardware design experience; focus on end user application tuning

Target Segment: Cloud, Enterprise (Health & Science, Analytics)

Target Customer: Customer with RTL expertise (discrete FPGA or ASIC), focus on networking acceleration

Target Segment: Cloud, Networking, Enterprise (Government, FSI)

Target Customer: Customer with RTL expertise (discrete FPGA or ASIC) or OpenCL expertise

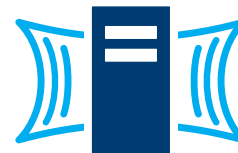
Target Segment: Networking, Enterprise (FSI, Health & Science, Gov't)

Intel® Xeon® Processor E5 Family with Altera Arria® 10 FPGA



Energy Efficient Scoring

- Best in class energy efficiency at 18.1 images/s/w
- Up to 40 percent lower power than previous generation FPGAs and SoCs



Infrastructure Flexibility

- Fits in standard server infrastructure
- Reconfigurable accelerator can be used for variety of data center workloads

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>

Intel® Arria 10- 1150 FPGA energy efficiency up to 25 images/second/watt on Caffe/AlexNet



Arria 10 1150 FP16 @ 297 MHz

Energy efficiency on Caffe/AlexNet up to 25 images/s/w



Configuration Details:

Vanilla 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

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

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

Legal Disclaimer & Optimization Notice

INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS”. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

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.

Copyright © 2016, Intel Corporation. All rights reserved. Intel, Pentium, Xeon, Xeon Phi, Core, VTune, Cilk, and the Intel logo are trademarks of Intel Corporation in the U.S. and other countries.

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

