Abstract

This session reviews Intel’s hardware and software strategy for machine learning. Intel is offering a range of tools and partnering with the open source community to help developers deliver optimized machine learning applications for Intel Architecture systems. This session spans a range of software development frameworks, libraries and other tools for machine learning with a focus on performance. Topics include IA-optimized frameworks such as Apache Spark*, Berkeley Caffe*, Google Tensorflow, and Nervana Neon and related performance benchmarks on the latest Intel Xeon and Intel Xeon Phi™ processors. Machine learning libraries Intel® Math Kernel Library and Intel® Data Analytics Acceleration Library will also be highlighted along with the new Intel® Distribution for Python and Intel® Deep Learning SDK.
Speaker Bio

Ananth Sankaranarayanan is currently the Director of Engineering with Intel in the Analytics and Artificial Intelligence Solutions group. His team is responsible for performance, partner and solution engineering functions, driving new platform initiatives jointly with leading Cloud Service Providers, Hardware Manufacturers and Software Vendors worldwide delivering engineered solutions to simplify implementations. Ananth previously led the HPC program for Intel silicon design/manufacturing and delivered 5 successful generations of supercomputers that directly contributed to reducing Intel Silicon Time to Market, and he has been with Intel since 2001. Ananth received his bachelors in computer science and engineering from Bharathidasan University in India and his Masters in Business Administration from City University of Seattle, USA.
Drivers for fast emergence of AI

<table>
<thead>
<tr>
<th>Bigger Data</th>
<th>Better Hardware</th>
<th>Smarter Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image: 1 MB / picture</td>
<td>Significant compute performance increases year over year</td>
<td>Advances in algorithm innovation, including neural networks, leading to better accuracy in training models</td>
</tr>
<tr>
<td>Audio: 5 MB / song</td>
<td>Parallel processing norm now</td>
<td></td>
</tr>
<tr>
<td>Video: 5 GB / movie</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data + Analytics Creates Unique Opportunities

Companies that use analytics best are...

2X 5X 3X 2X

...more likely to

- Make data-driven decisions
- Make decisions faster than others
- Execute on decisions faster
- Have top-quartile financial results

Competitive Advantage
Impactful Insights
Measurable Business Value

Source: Bain
The Evolution of Artificial Intelligence

How can we achieve the best outcome?
What will happen in the future?
Why is this happening?
What happened?
Where is the problem?

Business value

Late
Mainstream
Early

Spread sheets
Transactional Reporting
Early BI
Integrated OLTP and Analytics
Advanced Operational Analytics
Predictive Analytics
Prescriptive Analytics
Cognitive Analytics

Complexity/ Opportunity

Analytics using Artificial Intelligence

Business value
Artificial Intelligence (AI)
Machines that can sense, reason, act without explicit programming

Machine Learning (ML), a key tool for AI, is the development, and application of algorithms that improve their performance at some task based on experience (previous iterations)

Deep Learning (DL)
Algorithms where abstract ideas are represented by multiple (deep) layers of graphs

Classical Machine Learning
Algorithms based on statistical or other techniques for estimating functions from examples

Training: Build a mathematical model based on a data set
Inference: Use trained model to make predictions about new data
AI: Deep Learning Example:

**Step 1: Training**
(In Data Center – Over Hours/Days/Weeks)

- Massive data sets: labeled or tagged input data
- Create “Deep neural net” math model
- Output Classification

  90% person
  8% traffic light

**Step 2: Inference**
(At the Edge or in the Data Center - Instantaneous)

- New input from camera and sensors
- Trained neural network model
- Output Classification

  97% person
AI: Example Use Cases

Cloud Service Providers

- Image classification and detection for accurate diagnosis
- Image recognition/tagging for defect identification
- Natural language recognition (digital assistants)
- Big data pattern detection
- Targeted ads to increase revenue

Financial Services

- Fraud prevention/face detection
- Gaming, check processing
- Computer server monitoring
- Safe navigation for autonomous vehicles
- Financial forecasting and prediction to avoid risk
- Network intrusion detection

Healthcare

Automotive
AI Market Opportunity

**PRESENT**

Server Market (2015)¹

- 93% AI servers
- 7% Other servers

**Architecture MSS**

- 94.3% Intel
- 2.4% Intel+GPU
- 3.3% Other

**FUTURE**

Data is the next disrupter
By 2020, Machine to Machine connections will be 47% of total devices & connections

- 1PB/DAY Connected Factory
- 40TB/DAY Connected Plane
- 4TB/DAY Connected Car
- 90MB/DAY PC
- 90MB/DAY Connected Car
- 30MB/DAY Smartphone

AI-based analytics market²

- $8.2B 2013
- $70B 2020

---

¹Source: DCG Market Intelligence team

²Source: IDC, IOT market related to analytics
INTEL AI STRATEGY
Intel AI strategy
Making AI more pervasive by enabling deployment ready AI solutions through a large, open ecosystem

Solution blueprints
for reference across industries

Tools/Platforms
to accelerate deployment of IA solution stack

Optimized Open Frameworks
that scale to multi-node and deliver best performance

Free Libraries/Languages
featuring optimized ML/DL building blocks to enable developers

Best in class hardware
Cross compatible portfolio spanning from data center to edge
delivering high perf, perf/TCO, perf/w
AI: HARDWARE
Intel AI Products for the Datacenter

**Training**
- Intel® Xeon Phi™ Processors
  - Optimized for performance
  - Scales with cluster size for shorter time to model
  - x86 architecture, consistent programming model for training and inference

**Inference**
- Intel® Xeon® Processors
  - Optimized for performance/TCO
  - Most widely deployed inference solution
- Intel® Xeon® Processors + FPGA (discrete)
  - Optimized for performance/watt
  - Reconfigurable – can be used to accelerate many DC workloads
  - Programmable with industry standard OpenCL
Intel® Xeon Phi™ Processor Family
Enables shorter time to train

Breakthrough Highly-Parallel Performance
- Up to ~6 SGEMM TFLOPs$^3$ per socket
- Great scaling efficiency resulting in lower time to train for multi-node
- Eliminates add-in card PCIe* offload bottleneck and utilization constraints

Removes Barriers through Integration
- Integrated Intel® Omni-Path fabric (dual-port; 50 GB/s) increases price-performance and reduces communication latency for deep learning networks

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

Configurations: 1-8 see page 16.

All specifications refer to the future Intel® Xeon Phi™ processor (Knights Landing) unless otherwise noted.

All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance. *Other brands and names are the property of others.
Mainstream Training: Intel® Xeon Phi™ Processor 7250

Competitive Deep Learning Image Classification SINGLE-NODE Training With Mainstream cuDNN

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 information go to http://www.intel.com/performance. *Other names and brands may be property of others

Configurations:
1. Nvidia Tesla M40* (core@923MHz, 12GB, mem@3004MHz, 250W), DIGITS Deep Learning Machine* hosted on 2S Intel® Xeon® processor E5-2620 v3, 64GB memory, Ubuntu* 14.04, Nvidia* Driver v352.41, cuDNN v4, BVLC:Caffe cuDNN v5 or NVDA/Caffe cuDNN v5
2. Intel® Xeon Phi™ processor 7250 (68 Cores, 1.4 GHz, 16GB), 128GB memory, Red Hat* Enterprise Linux 6.7, Intel* Caffe: https://github.com/intelcaffe


2S Intel® Xeon® processor E5-2620 v3 +
Nvidia® Tesla M40

Intel® Xeon Phi™ processor 7250

Normalized Images/Second
Higher is better

1.04 1.29
1.03 1.10

Dataset: Large image database

Topology: Caffe*/AlexNet1
Why does Scaling Matter?

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

Deep Learning Image Classification Training Performance - MULTI-NODE Scaling

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.

**Topology:** AlexNet*

**Dataset:** Large image database

<table>
<thead>
<tr>
<th>Normalized Training Time - Higher is better</th>
<th>1 node</th>
<th>2 nodes</th>
<th>4 nodes</th>
<th>8 nodes</th>
<th>16 nodes</th>
<th>32 nodes</th>
<th>64 nodes</th>
<th>128 nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0x</td>
<td>1.0x</td>
<td>1.9x</td>
<td>3.7x</td>
<td>6.6x</td>
<td>12.8x</td>
<td>23.5x</td>
<td>33.7x</td>
<td>52.2x</td>
</tr>
</tbody>
</table>

*Note: Performance data is based on specific system configurations using benchmark applications. Performance in other applications may vary. Some benchmarks may use hardware accelerators for certain functionalities.
Great 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, 16 GB MCDRAM), 128 GB memory, Red Hat® Enterprise Linux 6.7, Intel® Optimized Frameworks
Knights Mill: Next Gen Intel® Xeon Phi™ processor

Enables shorter time to train

- **Trains Machines Faster**
  - >2X* Single Precision & >4X* 16-bit Mixed Precision faster deep learning training performance
  - Highly distributed processing with efficient scaling over multi-node offers flexible infrastructure for ML/DL workloads

- **Consistent Programming Model**
  - Common Xeon & Xeon Phi programming for developers
  - Optimized for industry standard Open Source frameworks
  - Bootable Host-CPU avoids offloading latency & bottleneck

- **Memory Flexibility**
  - High memory bandwidth with integrated DRAM increases performance for complex neural datasets by reducing latency
  - Large DDR4 memory capacity for massive AI use cases

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

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](http://www.intel.com/benchmarks). Performance estimate wrt KNL 7250 SKU SGEMM.

Performance Calculation= AVX freq X Cores X Flops per Core X Efficiency
Intel® Xeon® Processor E5 Family
High throughput inference on existing server class infrastructure

Leadership Throughput
• Classify 1115 images/second

Server Class Reliability
• Industry standard server features: high reliability, hardware enhanced security

Lowest TCO With Good Infrastructure Flexibility
• Standard server infrastructure
• Open standards, libraries & frameworks
• Optimized to run wide 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
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. 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. *Other names and brands may be claimed as property of others.
Scoring on Intel® Xeon® Processor E5-2699 v4
Deep Learning Image Classification SINGLE-NODE SCORING Performance

Results have been estimated or simulated using internal 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 information go to http://www.intel.com/performance. *Other names and brands may be property of others

Configurations (more details see slide 10):
1. Nvidia Tesla M4* (core@923MHz, 4GB, mem@3004MHz, 75W), DIGITS* Deep Learning Machine, 1S Intel® Xeon® Processor E5-2620 v3, 2.4GHz, 64GB, Ubuntu 14.04, Nvidia* Driver version 352.68, cuDNN v4, BVLC/Caffe cuDNN v5 or NVIDIA/Caffe cuDNN v5
2. 2S Intel® Xeon® Processor E5-2699 v4, 22C, 2.3GHz, 128GB, Red Hat Enterprise Linux* 6.7, Intel* Caffe : https://github.com/intelcaffe

Batch Size:256; ** Q4’16 based on MKL engineering version

Topologies: Caffe*/AlexNet

Dataset: Large image database

Normalized Images/Second
Higher is better

<table>
<thead>
<tr>
<th>Topology</th>
<th>Dataset</th>
<th>Normalized Images/Second</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffe*/AlexNet</td>
<td>Large image database</td>
<td>1.54</td>
</tr>
<tr>
<td>Caffe*/AlexNet</td>
<td>Large image database</td>
<td>1.59</td>
</tr>
<tr>
<td>Caffe*/AlexNet</td>
<td>Large image database</td>
<td>1.54</td>
</tr>
<tr>
<td>Caffe*/AlexNet</td>
<td>Large image database</td>
<td>1.59</td>
</tr>
</tbody>
</table>

cuDNN v4   | BVLC/Caffe cuDNN v5 | NVDA/Caffe cuDNN v5 | Intel Caffe - MKL 2017 Beta update | Intel Caffe - MKL 2017 Gold**
INTEL® DAAL
Intel® Data Analytics Acceleration Library (Intel® DAAL)

An Intel-optimized library that provides building blocks for all data analytics stages, from data preparation to data mining & machine learning

- Python, Java & C++ APIs
- Can be used with many platforms (Hadoop*, Spark*, R*, ...) but not tied to any of them
- Flexible interface to connect to different data sources (CSV, SQL, HDFS, ...)
- Windows*, Linux*, and OS X*

- Developed by same team as the industry-leading Intel® Math Kernel Library
- Open source, Free community-supported and commercial premium-supported options
- Also included in Parallel Studio XE suites
Intel DAAL Overview

Industry leading performance, C++/Java/Python library for machine learning and deep learning optimized for Intel® Architectures.

Pre-processing

(De-)Compression

Transformation

PCA
Statistical moments
Variance matrix
QR, SVD, Cholesky
Apriori

Analysis

Linear regression
Naïve Bayes
SVM
Classifier boosting
Kmeans
EM GMM

Modeling

Collaborative filtering
Neural Networks

Validation

Decision Making

Scientific/Engineering
Web/Social
Business
Problem Statement

Big data analytics
*Current common practice*
- Run on state-of-art hardware
- Built with a patchwork of math libs
- Under-exploiting hardware performance features

- Limited performance
- Many layers of dependencies
- Low ROI on HW investment

Data sources
- Finance
- Social media
- Marketing
- IoT
- Mfg
...

Connectors
- SQL stores
- NoSQL stores
- In-memory stores

Big data frameworks: Hadoop, Spark, Cassandra, etc.
Big data frameworks: Hadoop, Spark, Cassandra, etc.

Problem Statement

Data sources
- Finance
- Social media
- Marketing
- IoT
- Mfg
...

SQL stores

NoSQL stores

In-memory stores

Connectors

Data mining

Recommendation engines

Customer behavior modeling

BI analytics

Real time analytics

- Limited performance
- Many layers of dependencies
- Low ROI on HW investment

Data mining
Spark* MLLib

Recommendation engines
Breeze

Connectors
Netlib-Java

Customer behavior modeling
JVM

BI analytics

Real time analytics

JNI

Netlib BLAS

Big data frameworks: Hadoop, Spark, Cassandra, etc.
Ideal Solution

**Big data analytics Desired practice**

- Run on state-of-art hardware
- Single library to cover all stages of data analytics
- Fully optimized for underlying hardware

**Data sources**
- SQL stores
- NoSQL stores
- In-memory stores
- Connectors

**Data mining**
**Recommendation engines**
**Customer behavior modeling**
**BI analytics**
**Real time analytics**

**Desired Solution**
- Optimized performance
- Simpler development & deployment
- High ROI on HW investment

**Big data frameworks:** Hadoop, Spark, Cassandra, etc.
Intel® DAAL vs. Spark® Mllib
K-means Performance Comparison on Eight-node Cluster

<table>
<thead>
<tr>
<th>Speedup</th>
<th>300K rows, 4K columns, 120 clusters</th>
<th>3000K rows, 4K columns, 120 clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7x</td>
<td>4x</td>
</tr>
</tbody>
</table>

Configuration Info - Versions: Intel® Data Analytics Acceleration Library 2017, Spark 1.2; Hardware: Intel® Xeon® Processor E5-2699 v3, 2 Eighteen-core CPUs (40MB LLC, 2.3GHz), 128GB of RAM per node; Operating System: CentOS 6.6 x86_64.

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 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 DAAL Components

Data Processing Algorithms
Optimized analytics building blocks for all data analysis stages, from data acquisition to data mining and machine learning

Data Modeling Algorithms
Data structures for model representation, and operations to derive model-based predictions and conclusions

Data Management
Interfaces for data representation and access. Connectors to a variety of data sources and data formats, such as HDFS, SQL, CSV, ARFF, and user-defined data source/format

Data Sources
Numeric Tables
Compression / Decompression
Serialization / Deserialization
Data Transformation and Analysis in Intel® DAAL

- **Basic statistics for datasets**
  - Low order moments
  - Quantiles

- **Correlation and dependence**
  - Cosine distance
  - Correlation distance
  - Variance-Covariance matrix

- **Matrix factorizations**
  - SVD
  - QR
  - Cholesky

- **Dimensionality reduction**
  - PCA

- **Outlier detection**
  - Association rule mining (Apriori)
  - Univariate
  - Multivariate

*Algorithms supporting streaming and distributed processing in initial release*
Machine Learning in Intel® DAAL

- **Supervised learning**
  - Regression
    - Linear Regression
  - Classification
    - Weak learner
      - Boosting (Ada, Brown, Logit)
    - Naïve Bayes
    - Support Vector Machine

- **Unsupervised learning**
  - Neural Networks
  - K-Means Clustering
  - EM for GMM
  - Collaborative filtering
    - Alternating Least Squares

*Algorithms supporting streaming and distributed processing*
What's New: Intel DAAL 2017

• Neural Networks
• Python API (a.k.a. PyDAAL)
  – Easy installation through Anaconda or pip
• Open source project on GitHub

Fork me on GitHub: [https://github.com/01org/daal](https://github.com/01org/daal)
INTEL® MKL AND MKL-DNN
Highly optimized threaded math routines
- Performance, Performance, Performance!

Industry’s leading math library
- Widely used in science, engineering, data processing

Tuned for Intel® processors – current and next generation

Be multiprocessor aware
- Cross-Platform Support
- Be vectorised, threaded, and distributed multiprocessor aware
## Components of Intel MKL 2017

### Linear Algebra
- BLAS
- LAPACK
- ScaLAPACK
- Sparse BLAS
- Sparse Solvers
- Iterative
- PARDISO*
- Cluster Sparse Solver

### Fast Fourier Transforms
- Multidimensional
- FFTW interfaces
- Cluster FFT

### Vector Math
- Trigonometric
- Hyperbolic
- Exponential
- Log
- Power
- Root
- Vector RNGs

### Summary Statistics
- Kurtosis
- Variation coefficient
- Order statistics
- Min/max
- Variance-covariance

### And More...
- Splines
- Interpolation
- Trust Region
- Fast Poisson Solver

### Deep Neural Networks
- Convolution
- Pooling
- Normalization
- ReLU
- Inner Product

---

New
Intel® Math Kernel Library and Intel® MKL-DNN for Deep Learning Framework Optimization

Deep Learning Frameworks

- Caffe
- theano
- neon
- Google Tensorflow
- torch

Intel® Math Kernel Library (Intel® MKL)

Intel® MKL-DNN

<table>
<thead>
<tr>
<th>Intel® MKL</th>
<th>Intel® MKL-DNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN primitives + wide variety of other math functions</td>
<td>DNN primitives</td>
</tr>
<tr>
<td>C DNN APIs</td>
<td>C/C++ DNN APIs</td>
</tr>
<tr>
<td>Binary distribution</td>
<td>Open source DNN code*</td>
</tr>
<tr>
<td>Free community license, Premium support available as part of Parallel Studio XE</td>
<td>Apache 2.0 license</td>
</tr>
<tr>
<td>Broad usage DNN primitives; not specific to individual frameworks</td>
<td>Multiple variants of DNN primitives as required for framework integrations</td>
</tr>
<tr>
<td>Quarterly update releases</td>
<td>Rapid development ahead of Intel MKL releases</td>
</tr>
</tbody>
</table>

* GEMM matrix multiply building blocks are binary
**Improved Deep Neural Network training performance using Intel® Math Kernel Library (Intel® MKL)**

Caffe/AlexNet single node training performance

![Chart showing performance speedup](chart)

5.8x Intel® Xeon® E5-2699 v4
2.1x Intel® Xeon® E5-2699 v4 +Intel MKL 11.3.3
2x Intel® Xeon® E5-2699 v4 +Intel MKL 2017
24x Intel® Xeon® 7250

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](http://www.intel.com/performance). *Other names and brands may be property of others.

- Z 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: 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.
Improved Deep Neural Network inference performance using Intel® Math Kernel Library (Intel® MKL)

Caffe/AlexNet single node inference performance

- Intel Xeon E5-2699v4
  - Out-of-the-box: 7.5x
  - +Intel MKL 11.3.3: 2.2x
  - +Intel MKL 2017: 1.9x
- Intel Xeon Phi 7250
  - +Intel MKL 2017: 31x

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 E5-2699v4 (22 Cores, 2.2 GHz), 128 GB memory, Red Hat® Enterprise Linux 6.7, BVL 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: 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.
Case Study I: Deep Learning

LeCloud* Illegal Video Detection

- LeCloud: leading video cloud provider in China who provides illegal video detection service
- Originally: Adopted open source BVLC Caffe w/OpenBlas as CNN framework
- Now: Using Intel Optimized Caffe plus Intel® Math Kernel Library, achieved 30x performance improvement for training in production

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

![Diagram of Video Classification Process](attachment:image.png)
## Intel® DAAL+ Intel® MKL = Complementary Big Data Libraries Solution

<table>
<thead>
<tr>
<th>Intel MKL</th>
<th>Intel DAAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>C and Fortran API Primitive level</td>
<td>Python, Java &amp; C++ API High-level</td>
</tr>
<tr>
<td>Processing of homogeneous data in single or double precision</td>
<td>Processing heterogeneous data (mix of integers and floating point), internal conversions are hidden in the library</td>
</tr>
<tr>
<td>Type of intermediate computations is defined by type of input data (in some library domains higher precision can be used)</td>
<td>Type of intermediate computations can be configured independently of the type of input data</td>
</tr>
<tr>
<td>Most of MKL supports batch computation mode only</td>
<td>3 computation modes: Batch, streaming and distributed</td>
</tr>
<tr>
<td>Cluster functionality uses MPI internally</td>
<td>Developer chooses communication method for distributed computation (e.g. Spark, MPI, etc.) Code samples provided.</td>
</tr>
</tbody>
</table>

“Initially, the Spark/Shark-based solution required 40 hours to compete a computation. Youku improved performance significantly by implementing Intel® Math Kernel Library (Intel® MKL) into its solution...After implementation of Intel MKL, Youku reduced the computation time to less than three hours.”

Source: Youku Tudou Video Sharing Recommendation Case Study
INTEL® DEEP LEARNING SDK
Intel® Deep Learning SDK
Accelerate Your Deep Learning Solution

A free set of tools for data scientists and software developers to develop, train, and deploy deep learning solutions

“Plug & Train/Deploy”

Simplify installation & preparation of deep learning models using popular deep learning frameworks on Intel hardware

Maximum Performance

Optimized performance for training and inference on Intel® Architecture

Increased Productivity

Faster Time-to-market for training and inference, Improve model accuracy, Reduce total cost of ownership
Deep Learning Training Tool
Intel® Deep Learning SDK

- Simplify installation of Intel optimized Deep Learning Frameworks
- Easy and Visual way to Set-up, Tune and Run Deep Learning Algorithms:
  - Create training dataset
  - Design model with automatically optimized hyper-parameters
  - Launch and monitor training of multiple candidate models
  - Visualize training performance and accuracy

Data Scientist

DL Training Tool

Install
Configure
Run

Accuracy
Utilization
Model

Dataset

DL Framework
MKL-DNN

Trained Model

Datacenter

.label

.Dataset

.Model

.Trained

.Model

Accuracy

Utilization

.Model

.caffemodel

.proto.txt
Deep Learning Deployment Tool
Intel® Deep Learning SDK

Trained Model

- Imports trained models from all popular DL framework regardless of training HW
- Compresses model for improved execution, storage & transmission (pruning, quantization)
- Generates scoring HW-specific code (C/C++, OpenVX graphs, OpenCL, etc.)
- Enables seamless integration with full system / application software stack

Unleash fast scoring performance on Intel products while abstracting the HW from developers

- .prototxt
- .caffemodel

Model Optimizer
  - FP Quantize
  - Model Compress
  - Model Analysis

Application Logic
  - Real-time Data Validation Data
  - Forward
  - Result

Inference Run-Time
  - MKL-DNN
  - OpenVX

End-Point SW Developer

Deploy-ready model

45.
Deep Learning Tools for End-to-End Workflow

Intel® Deep Learning SDK

Intel DL Training Tool

- INSTALL / SELECT IA-Optimized Frameworks
- PREPARE / CREATE Dataset with Ground-truth
- DESIGN / TRAIN Model(s) with IA-Opt. Hyper-Parameters
- MONITOR Training Progress across Candidate Models
- EVALUATE Results and ITERATE

Intel DL Deployment Tool

- IMPORT Trained Model (trained on Intel or 3rd Party HW)
- COMPRESS Model for Inference on Target Intel HW
- GENERATE Inference HW-Specific Code (OpenVX, C/C++)
- INTEGRATE with System SW / Application Stack & TUNE
- EVALUATE Results and ITERATE

MKL-DNN Optimized Machine Learning Frameworks
- theano
- Caffe
- Microsoft
- Google
- Tensor
- Torch

Xeon (local or cloud)

Optimized libraries & run-times (MKL-DNN, OpenVX, OpenCL)

Data acquisition (sensors) and acceleration HW (FPGA, etc)

Target Inference Hardware Platform (physical or simulated)
Intel Deep Learning Software Stack

**Intel® Deep Learning SDK**

**Deep Learning Frameworks**
- theano
- Caffe
- Theano
- TensorFlow
- Torch
- Neon

**Intel® MKL-DNN**

**Intel® Math Kernel Library**
- Xeon
- Xeon Phi
- FPGA

Tools to accelerate design, training and deployment of deep learning solutions

**Targeted release: early Q4’2016**

Popular Deep Learning frameworks

Open source Intel x86 optimized DNN APIs, combined with Intel® MKL and build tools designed for scalable, high-velocity integration with ML/DL frameworks.

Includes:
- New algorithms ahead of MKL releases
- IA optimizations contributed by community

SW building block to extract max Intel HW performance and provide common interface to all Intel accelerators.

Intel libraries as path to bring optimized ML/DL frameworks to Intel hardware

Software.intel.com/machine-learning

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

1. Enable hooks to Intel® MKL, Intel® DAAL, Intel® IPP functions in the most popular numerical packages
   - NumPy, SciPy, Scikit-Learn, PyTables, Scikit-Image, ...

2. Available through Intel® Distribution for Python* and as Conda packages
   - Most optimizations eventually upstreamed to home open source projects

3. Provide Python interfaces for Intel® DAAL (a.k.a PyDAAL)
Numpy & Scipy optimizations with Intel® MKL

**Linear Algebra**
- BLAS
- LAPACK
- ScALAPACK
- Sparse BLAS
- Sparse Solvers
  - Iterative
  - PARDISO SMP & Cluster

**Fast Fourier Transforms**
- Multidimensional
- FFTW interfaces
- Cluster FFT

**Vector Math**
- Trigonometric
- Hyperbolic
- Exponential
- Log
- Power
- Root

**Vector RNGs**
- Multiple BRNG
- Support methods for independent streams creation
- Support all key probability distributions

**Summary Statistics**
- Kurtosis
- Variation coefficient
- Order statistics
- Min/max
- Variance-covariance

**And More**
- Splines
- Interpolation
- Trust Region
- Fast Poisson Solver

**Up to**
- 100x faster!
- 10x faster!
- 60x faster!
Near native performance on Intel® Xeon™ and Intel® Xeon Phi™

- Runs out-of-the-box with any Python
- Intel Distribution for Python delivers much greater efficiency than “system” Python
- Potential for future multi-threaded performance tunings in numpy and scipy
Roadmap & Reviews

Available as free standalone download
Commercial support through Intel® Parallel Studio 2017

“I expected Intel’s numpy to be fast but it is significant that plain old python code is much faster with the Intel version too.”

Dr. Donald Kinghorn, Puget Systems Review

InfoWorld
Intel's Python distribution provides a major math boost

The still-in-beta Python distribution uses Math Kernel Library to speed up processing on Intel hardware
The distribution's main touted advantage is speed -- but not a PyPy-style general speedup via a JIT. Instead, the MKL speeds up certain math operations so that they run faster on one thread and multiple threads.

insideHPC
HPC Podcast Looks at Intel's Pending Distribution of Python

Yes, Intel is doing their own Python build! It is still in beta but I think it's a great idea. ..........Yeah, it's important!

Summary

I. Deep Learning framework optimizations on Xeon, Xeon Phi – Session #

II. Intel DAAL

III. Intel MKL, MKL-DNN

IV. Intel Python optimizations – Session #

V. Intel Deep Learning SDK

Learn more at www.intel.com/machinelearning
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.

<table>
<thead>
<tr>
<th>Optimization Notice</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>

Notice revision #20110804