THE AI ERA INFRASTRUCTURE
INDUSTRY TRANSFORMATION & INNOVATION OPPORTUNITIES

Meena Arunachalam
Principal Engineer, Intel Architecture and Graphics Software
DISCLAIMERS:

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

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

No computer system can be absolutely secure.

Intel, the Intel logo, Xeon, Mobileye, DL Boost, Nervana, Atom, Arria, Stratix and Movidius are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

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

© 2019 Intel Corporation.
INDUSTRIAL REVOLUTION -- FROM PHYSICAL TO DIGITAL

Fourth Industrial Revolution represents entirely new ways in which technology becomes embedded within industries, societies and even our human bodies.

18th Century
Mechanization
Steam power
Water Power

19th Century
Mass production
Assembly line
Electricity

20th Century
Electronics
Internet & IT Automation

21st Century
Cyber Physical Systems

WE ARE WITNESSING THIS DIGITAL TRANSFORMATION
CREATING
Data defines the future

Competitiveness and business growth are increasingly determined by the power of data.

*Other names and brands may be claimed as property of others.
INNOVATION ACROSS ALL INDUSTRIES

AI / ML

Processing Power

“DIGITAL FUSION”
Blending of Traditional & Digital Business Models

Data

Smart Cities & Surveillance
Retail: Real-Time Pricing & Inventory
Enterprise/Consumer Analytics
Precision Medicine & Genomic Analytics
Autonomous Cars & 5G Connectivity
Virtual/Augmented Reality

Traditional Business
Digital Business

AI / ML

Power

Data

Smart Cities & Surveillance
Retail: Real-Time Pricing & Inventory
Enterprise/Consumer Analytics
Precision Medicine & Genomic Analytics
Autonomous Cars & 5G Connectivity
Virtual/Augmented Reality
AI IS DRIVING ADVANCED ANALYTICS

DATA DELUGE (2019)
- 25 GB\(^1\) per month
  - Internet User
- 50 GB\(^2\) per day
  - Smart Car
- 3 TB\(^2\) per day
  - Smart Hospital
- 40 TB\(^2\) per day
  - Airplane Data
- 1 PB\(^2\) per day
  - Smart Factory
- 50 PB\(^2\) per day
  - City Safety

ANALYTICS CURVE
- Hindsight: Descriptive Analytics
- Insight: Diagnostic Analytics
- Foresight: Predictive Analytics
- Forecast: Prescriptive Analytics
- Act/Adapt: Cognitive Analytics

INSIGHTS
- Business
- Operational
- Security

AI IS THE DRIVING FORCE

Copyright © 2018 Intel Corporation
INTEL® IS INFUSING AI INTO EVERYTHING WE DO
End to End Data Analytics Flow

Start with Data
ETL → Feature Engr, Classical ML → Data Prep

Data Prep, ETL, Dimension Reduction

Variety of data formats and structures
- Decompression, Filtering, Normalization
- Aggregation, Dimension Reduction
- Summary Statistics Clustering, etc.
- Machine Learning Parameter Estimation Simulation

Standard data and structures
- Hypothesis testing Model errors
- Forecasting Decision Trees, etc.

Pre-processing, Transformation, Analysis, Modeling, Validation, Decision Making
Deep Learning Flow

Artificial Intelligence → Machine Learning → Deep Learning

Training:
- fp32 => fp16 => bfp16
- Time to Train, accuracy, model size
- Lots of labeled data!
- Human
- Bicycle
- Strawberry

Inference:
- fp32 => fp16 => int8 => int4...
- Inference throughput, latency
- Error
- Human Bicycle
- Strawberry
- ?????
AI IS EXPANDING

Deploy AI anywhere with unprecedented hardware choice

Visit: [www.intel.ai/technology](http://www.intel.ai/technology)

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

Copyright © 2018 Intel Corporation
CPU FOUNDATION FOR ARTIFICIAL INTELLIGENCE

MATRIX OPERATIONS
Intel® Advanced Vector Extensions

LOWER & MIXED PRECISION
Intel® Deep Learning Boost

LARGER CACHES, MEMORY LATENCY & BANDWIDTH

OPTIMAL DATA MOVEMENT & TRANSFORMATIONS

OPTIMIZED LIBRARIES AND FRAMEWORKS

INTEL® XEON® SCALABLE PROCESSOR: ENABLES INFRASTRUCTURE-WIDE AI READINESS

Illustration: Intel® Xeon® Scalable Processor
3X peak operations providing significant improvement in inferencing performance
**REINVENTING XEON FOR AI**

**Intel® Optimization for Caffe ResNet-50**

1. **Inference Throughput**

- **1.0X**
  - **Intel® Optimized Caffe**
  - At launch, July 11th 2017

- **2.8X**
  - With new library and framework optimizations
  - Jan 19th 2018

- **5.7X**
  - Enabling Lower precision & System optimizations for higher throughput
  - March 6th 2019

- **14X**
  - **2S Intel® Xeon® Platinum 8280 processor (28 cores/S)**

- **30X**
  - **2S Intel® Xeon® Platinum 9282 processor (56 cores/S)**

---

**Intel® Optimization for Caffe ResNet-50 Performance**

- Performance does not necessarily represent other Framework performance.
- Based on Intel internal testing: 1X, 2.8X, 5.7X, 14X, and 30X performance improvement based on Intel® Optimization for Caffe ResNet-50 inference throughput performance on Intel® Xeon® Scalable Processor. See Configuration Details.
- Performance results are based on testing as of 7/11/2017 (1X), 11/8/2018 (5.7X), 2/20/2019 (14X) and 2/26/2019 (30X) and may not reflect all publically available security updates. No product can be absolutely secure. See configuration disclosure for details.

---

**Intel® Deep Learning Boost**

**Relative Inference Throughput (images/sec)**

(higher is better)

- **2.8X**
  - FP32
  - With new library and framework optimizations
  - Jan 19th 2018

- **5.7X**
  - INT8
  - Enabling Lower precision & System optimizations for higher throughput
  - March 6th 2019

---

**Intel® Xeon® Platinum 8180 Processor (Codename: Skylake)**

---

**2nd Generation Intel® Xeon® Scalable Processor (Cascade Lake)**

---


[3] Performance results are based on testing as of 7/11/2017 (1X), 11/8/2018 (5.7X), 2/20/2019 (14X) and 2/26/2019 (30X) and may not reflect all publically available security updates. No product can be absolutely secure. See configuration disclosure for details.

[4] 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. 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](http://www.intel.com/performance)
SOFTWARE IS ESSENTIAL

**TOOLKITS**
Application Developers

**LIBRARIES**
Data Scientists

**FOUNDATION**
Library Developers

**OPENVINO™ TOOLKIT**

**INTEL® MOVIDIUS™ SDK**

**MACHINE LEARNING LIBRARIES**
- Scikit-Learn
- NumPy
- MLlib

**DEEP LEARNING FRAMEWORKS**
- TensorFlow
- mxnet
- Caffe
- PyTorch
- CNTK
- Fastai

**DEEP LEARNING GRAPH COMPILER**
Intel® nGraph™ Compiler

**HARDWARE**

Other names and brands may be claimed as the property of others.
## INTEL® DISTRIBUTION FOR PYTHON*

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

**FOR DEVELOPERS USING THE MOST POPULAR AND FASTEST GROWING PROGRAMMING LANGUAGE FOR AI**

<table>
<thead>
<tr>
<th>EASY, OUT-OF-THE-BOX ACCESS TO HIGH PERFORMANCE PYTHON</th>
<th>DRIVE PERFORMANCE WITH MULTIPLE OPTIMIZATION TECHNIQUES</th>
<th>FASTER ACCESS TO LATEST OPTIMIZATIONS FOR INTEL® ARCHITECTURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ Prebuilt, optimized for numerical computing, data analytics, HPC</td>
<td>▪ Accelerated NumPy/SciPy/Scikit-Learn with Intel® MKL</td>
<td>▪ Distribution and individual optimized packages available through conda and Anaconda Cloud</td>
</tr>
<tr>
<td>▪ Drop in replacement for your existing Python (no code changes required)</td>
<td>▪ Data analytics with pyDAAL, enhanced thread scheduling with TBB, Jupyter* Notebook interface, Numba, Cython</td>
<td>▪ Optimizations upstreamed back to main Python trunk</td>
</tr>
<tr>
<td></td>
<td>▪ Scale easily with optimized MPI4Py and Jupyter notebooks</td>
<td></td>
</tr>
</tbody>
</table>

### ADVANCING PYTHON* PERFORMANCE CLOSER TO NATIVE SPEEDS

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice. Other names and brands may be claimed as the property of others.

© 2019 Intel Corporation | Optimization Notice
INTEL® DATA ANALYTICS ACCELERATION LIBRARY (INTEL® DAAL)
BUILDING BLOCKS FOR ALL DATA ANALYTICS STAGES, INCLUDING DATA PREPARATION, DATA MINING & MACHINE LEARNING

Pre-processing ➔ Transformation ➔ Analysis ➔ Modeling ➔ Validation ➔ Decision Making

Open Source | Apache* 2.0 License
Common Python, Java and C++ APIs across all Intel hardware
Optimized for large data sets including streaming and distributed processing
Flexible interfaces to leading big data platforms including Spark* and range of data formats (CSV, SQL, etc.)

HIGH PERFORMANCE MACHINE LEARNING AND DATA ANALYTICS LIBRARY

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice. Other names and brands may be claimed as the property of others.
INTEL® MATH KERNEL FOR DEEP LEARNING NEURAL NETWORKS (INTEL® MKL-DNN)
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.

EXEMPLARY:
- Direct 2D Convolution
- Local response normalization (LRN)
- Rectified linear unit neuron activation (ReLU)
- Maximum pooling
- Inner product

Accelerate Performance of Deep Learning Models

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice. Other names and brands may be claimed as the property of others.
INTEL® DISTRIBUTION OF OPENVINO TOOLKIT

OPTIMIZE EXISTING MODELS, RUN INFERENCE WHERE YOU NEED IT

Build high performance deep learning inference and computer vision

A toolkit to accelerate development of high performance computer vision & deep learning inference into vision/AI applications from edge to cloud. It enables deep learning on hardware accelerators and easy deployment across multiple types of Intel® platforms (CPU, GPU/Intel® Processor Graphics, FPGA, VPU).

Who needs it?
• Computer vision, deep learning developers
• Data scientists
• OEMs, ISVs, system integrators

Usages
Security surveillance, robotics, retail, healthcare, AI, office automation, transportation, non-vision use cases (speech, text) & more.

Free Download ➤ software.intel.com/openvino-toolkit
Open Source version ➤ 01.org/openvinotoolkit

© 2019 Intel Corporation
HPC ↔ AI: IMAGE ANALYSIS FOR DRUG DISCOVERY

NOVARTIS

Joint Intel & Novartis collaboration

224 X 224 X 3
ImageNet

26X Larger

1024 X 1280 X 3
Microscopic Images

Customer: Novartis Inst. of Biomedical Research (Switzerland) is one of the largest pharmaceutical companies in the world

Challenge: High content screening of cellular phenotypes is a fundamental tool supporting early stage drug discovery. While analyzing whole microscopic images are desirable, these images are 26X larger than benchmark dataset such as ImageNet^1K. As a result, the high computational workload with high memory requirement would be prohibitive for deep learning model training.

Solution: Intel and Novartis teams were able to scale and train the model with 32 TensorFlow* workers in 31 minutes.

RESULTS

Processing 1024x1280 large image dataset, reducing the training time to 31 minutes to >99% accuracy on 2S Intel® Xeon® processor based cluster.

Performance results are based on testing as of May 17, 2018 and may not reflect all publicly available security update. See configuration disclosure for details. No product can be absolutely secure.

Optimization notice: while some compilers may or may not continue to the same degree for optimizing microarchitectures that are not unique to Intel microcircuits. 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 may become optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are not derived from any benchmarking standards.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Other processors may have lower performance. See performance test methodology. The results presented are for a specific computer configuration as of the date of publication of this document, and represent the performance of this computer configuration. Other computers, components, software, and/or configurations (e.g., product version, processor speed, size and functions) may achieve different results. Any change to any of these 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

High Content Screening Training with 313K Images on 64-NODE Intel® 2S Xeon® Scalable processor 6148, TensorFlow*, ‘horovod’*, OpenMPI*, Batch Size=32/Node, Intel® Omni-Path™ Fabric
**USE CASE**

**RESULT**

94% scaling efficiency up to 128 nodes, with a significant reduction in training time per epoch for 3D-GANs & >2500X Inference

<table>
<thead>
<tr>
<th>Method</th>
<th>Machine</th>
<th>Time/Shower (msec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Simulation</td>
<td>2S Intel® Xeon® Platinum 8180</td>
<td>17000</td>
</tr>
<tr>
<td>(geant4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3D GAN (batch size 128)</td>
<td>2S Intel® Xeon® Platinum 8180</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Inference Perf:**

>2500X

---

**CUSTOMER:** CERN, the European Organization for Nuclear Research, which operates the Large Hadron Collider (LHC), the world’s largest and most powerful particle accelerator

**Challenge:** CERN currently uses Monte Carlo simulations for complex physics and geometry modeling, which is a heavy computational load that consumes up to >50% of the Worldwide LHC (Large Hadron Collider) Computing Grid (WLCG) power for electron shower simulations.

**Solution:** Distributed training using 128 nodes of the TACC Stampede 2 cluster (Intel® Xeon® Platinum 8160 processor, Intel® OPA) and a 3D Generative Adversarial Network (3D GAN). Performance was first optimized on a single node then scaled using TensorFlow® optimized with Intel® MKL-DNN, using 4 workers/node and an optimized number of convolutional filters.

---


---

**HPC ↔ AI: DIS/REPLACING MONTE CARLO SIM.**

**CERN HIGH ENERGY PHYSICS**

**JOINT COLLABORATION WITH INTEL AND SURFSARA**

---

**Wall Clock time per Activity**

**MC - related**

- MC Simulation
- MC Event Generation
- Analysis
- Group Production
- Data Processing
- TO Processing
- Others

**3D-Generative Adversarial Networks (GANs)**

**3D Generative Adversarial Networks (GANs)**

**Customer:** CERN, the European Organization for Nuclear Research, which operates the Large Hadron Collider (LHC), the world’s largest and most powerful particle accelerator

**Challenge:** CERN currently uses Monte Carlo simulations for complex physics and geometry modeling, which is a heavy computational load that consumes up to >50% of the Worldwide LHC (Large Hadron Collider) Computing Grid (WLCG) power for electron shower simulations.

**Solution:** Distributed training using 128 nodes of the TACC Stampede 2 cluster (Intel® Xeon® Platinum 8160 processor, Intel® OPA) and a 3D Generative Adversarial Network (3D GAN). Performance was first optimized on a single node then scaled using TensorFlow® optimized with Intel® MKL-DNN, using 4 workers/node and an optimized number of convolutional filters.
Multi-Node Training Performance & Accuracy (2018)

Distributed training using data parallelism

94% Scaling efficiency up to 128
HPC ↔ AI: CHEST X-RAY IMAGE CLASSIFICATION

DellEMC

Joint collaboration with SURFsara, & DellEMC

Identifying thoracic pathologies from the NIH ChestXray14 dataset

Emphysema affects more than: 3 Mil U.S & 65 Mil Worldwide

Pneumonia affects more than: 1 Mil US & 450 Mil Worldwide

Customer: DellEMC*

Research on AI applications on Intel® Xeon® CPUs: Medical, Cloud, HPC, etc.

Challenge: Train a chest X-ray model that delivers highly-efficient scaling performance on Intel® Xeon® processor nodes, while also delivering higher accuracy than the existing ChexNet-121 model


RESULT

Training time reduced to 11 mins while increasing the accuracy across 10 categories & 4% (>90%) better relative to the existing DenseNet-121 model

Transfer Learning

Categorical Accuracies in identifying diseases using ResNet-50 vs CheXNet-121

- Better accuracy in 10 categories with scaled out trained ResNet50!

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. 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 intended to measure relative performance and not necessarily absolute performance. Performance tests are based on Intel engineering measurements and tests. 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.

Transfer Learning

Use Case

Identifying thoracic pathologies from the NIH ChestXray14 dataset

Emphysema affects more than: 3 Mil U.S & 65 Mil Worldwide

Pneumonia affects more than: 1 Mil US & 450 Mil Worldwide

Customer: DellEMC*

Research on AI applications on Intel® Xeon® CPUs: Medical, Cloud, HPC, etc.

Challenge: Train a chest X-ray model that delivers highly-efficient scaling performance on Intel® Xeon® processor nodes, while also delivering higher accuracy than the existing ChexNet-121 model

Training Performance: ResNet-50 on ChestXRay14

Intel® 2S Xeon® Gold 6148F processor based DellEMC® PowerEdge C6420 Zenith® Cluster on OPA™ Fabric

TensorFlow®* + “horovod”*, IMPI

Relative Training Throughput

(Images/sec)

(Higher is Better)

- TensorFlow ResNet-50 Node=1, Workers=4...
- TensorFlow ResNet-50 Nodes=128, Workers=512...
- TensorFlow ResNet-50 Nodes=256, Workers=1024...

Training Time

8 MINUTES to reach a solution with 256 25k Intel® Xeon® Gold 6148 processor

104x faster using 128 Intel® Xeon® nodes!

187x faster using 256 Intel® Xeon® nodes!

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, including the specific instruction sets covered by this 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

Thanks to DellEMC for Access to Internal Cluster
Case Study: Image Recognition

World Bank

Client: The International Comparison Program (ICP) in the World Bank Development Data Group

Challenge: The World Bank team needed to automate the process of confirming that the crowd-sourced photos, gathered from cellphone contributors from 15 countries, were accurately classified into one of 162 categories ranging from food to footwear, and to remove personally identifiable information (PII) from the photos.

Solution: Utilized Intel's BigDL framework (a distributed deep-learning library for Apache Spark*) and an AWS Databricks* platform running on Intel® Xeon® Processors (AWS R4.8xlarge instance with 20 nodes) to help classify more than 1 million crowdsourced photos before sharing the dataset with the public.

RESULT

High accuracy results using an AWS Databricks* platform to train a dataset consisting of almost 1 million images in 69 categories, with near linear scaling on a partial dataset.
Client: Center for Digital Health Innovation (CDHI) at UCSF, leveraging new digital health technologies to transform healthcare.

Challenge: Projected by 2040 – 78M adults with doctor-diagnosed OA & 35M with arthritis-attributable activity limitations. Need automated system that classifies menisci based on presence/absence of lesions, provides immediate objective results at MRI scan, & eliminates intra-user variability.

Solution: Apache Spark* with BigDL on CDH 5.9*, on Intel® Xeon® servers from Dell*. With 3D image convolution in BigDL, the CDHI team built a MRI classification system & deployed it on their CDH Dell cluster.

RESULT

“Dataset, model development and training [...] implementing 3D CNN in BigDL to analyze MRI scans and classify OA (osteoarthritis) [...] provides rich 3D imaging support [...] on the same cluster where data is stored”
Artificial Intelligence will empower TRANSFORMATIVE INNOVATIONS

WE ARE WITNESSING THIS DIGITAL TRANSFORMATION CREATING