# Understand & Integrate INTEL® DL Boost

## Overview

<table>
<thead>
<tr>
<th>Strategy</th>
<th>DEEP DIVE</th>
<th>Steps To Deploy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intel AI</strong></td>
<td><strong>DL BOOST</strong></td>
<td><strong>DEPLOY</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Portfolio</th>
<th>Hardware</th>
<th>Use Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute</td>
<td>Software</td>
<td>Develop</td>
</tr>
<tr>
<td></td>
<td>Performance</td>
<td>Partner</td>
</tr>
</tbody>
</table>
THE DATA-CENTRIC ERA

DEVICE

INTELLIGENT EDGE

MULTI-CLOUD

© 2019 Intel Corporation
DELIVER THE BEST AI PLATFORMS
EXTEND CPU
MOST COMPLETE PORTFOLIO
BEST INTEGRATED PLATFORMS

SHAPE & WIN INDUSTRY OPEN SOFTWARE STACKS
OPTIMIZE CUSTOMER SOFTWARE
BUILD “ONE API” UNIFIED PLATFORM
EVANGELIZE TO DEVELOPERS

SEED & DRIVE THE ECOSYSTEM
SEED EMERGING USE CASES
ATTRACTION & DEVELOP TOP TALENT
PIONEER LEADING EDGE AI
INTEL AI HARDWARE

MOVE FASTER
- ETHERNET
- SILICON PHOTONICS
- OMNI-PATH FABRIC

STORE MORE
- OPTANE DC
- OPTANE DC
- DC SERIES SSD

PROCESS EVERYTHING
- INTEL® XEON® SCALABLE PROCESSORS
- INTEL® XEON® D PROCESSORS
- INTEL® ATOM® C PROCESSORS
- INTEL® FPGA
- INTEL® NERVANA™ NNP
- INTEL® MOVIDIUS™ TECHNOLOGY

© 2019 Intel Corporation
Multi models – high performance across broad set of existing and emerging DL models for applications like Natural Language Processing, Speech & Text, Recommendation Engines, and Video & Image search & filtering

A+ - Combined workflows where use cases span AI plus additional workloads such as - Immersive Media - Media + Graphics

© 2019 Intel Corporation
INTEL AI COMPUTE
DEPLOY AI ANYWHERE FROM DEVICE TO INTELLIGENT EDGE & MULTI-CLOUD

CPU

ACCELERATORS

DEVICE

INTELLIGENT EDGE

MULTI-CLOUD

© 2019 Intel Corporation
EXPLOSION IN DEMAND FOR COMPUTE

INCREASING COMPUTE DEMAND
DIVERSIFYING WORKLOAD NEEDS

Source: Intel analysis
© 2019 Intel Corporation
## UNDERSTAND & INTEGRATE INTEL® DL BOOST

### OVERVIEW
- **INTEL AI**
  - STRATEGY
  - PORTFOLIO
  - COMPUTE

### DEEP DIVE
- **DL BOOST**
  - HARDWARE
  - SOFTWARE
  - PERFORMANCE

### STEPS TO DEPLOY
- USE CASES
- DEVELOP
- PARTNER

© 2019 Intel Corporation
INTEL® DEEP LEARNING BOOST

OPTIMIZING AI INFERENC

INTEL OPTIMIZATION FOR CAFFE RESNET-50

INTEL AVX-512

INTEL DL BOOST

14X INTEL® XEON® PLATINUM 8200 PROCESSOR

2X MORE INTEL® XEON® PLATINUM 9200 PROCESSOR

© 2019 Intel Corporation
INTEL® DEEP LEARNING BOOST

OPTIMIZING AI INFERENCE

Microarchitecture view
In a given clock cycle

Port0
Port5

FMA
VNNI
FMA
VNNI

1st gen Intel® Xeon® Scalable processor without Intel® DL Boost

FP32

vpmadd231ps
OUTPUT FP32

INPUT FP32
INPUT FP32

1st gen Intel® Xeon® Scalable processor without Intel® DL Boost

INT8

vpmaddubsw
OUTPUT INT16

INPUT INT8
INPUT INT8

INT8

vpaddd
OUTPUT INT32

INPUT INT8
INPUT INT8

2nd gen Intel® Xeon® Scalable processor with Intel® DL Boost

INT8
VNNI

vpdpbusd
OUTPUT INT32

INPUT INT8
INPUT INT8

Accumulate INT32

Accumulate INT32

© 2019 Intel Corporation
INTEL® DEEP LEARNING BOOST

Vector Neural Network Instruction

PRECISION IS A MEASURE OF THE DETAIL USED FOR NUMERICAL VALUES PRIMARILY MEASURED IN BITS

<table>
<thead>
<tr>
<th>Format</th>
<th>Sign Bit</th>
<th>Exponent Bits</th>
<th>Mantissa Bits</th>
<th>Signed Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP32</td>
<td>s</td>
<td>8</td>
<td>23</td>
<td>±1.18×10⁻³⁸ to ±3.4×10³⁸</td>
</tr>
<tr>
<td>BF16</td>
<td>s</td>
<td>8</td>
<td>7</td>
<td>±1.18×10⁻³⁸ to ±3.4×10³⁸</td>
</tr>
<tr>
<td>FP16</td>
<td>s</td>
<td>5</td>
<td>10</td>
<td>±6.10×10⁻⁵ to ±65504</td>
</tr>
<tr>
<td>INT16</td>
<td>s</td>
<td></td>
<td>15</td>
<td>-32,768 to +32,767</td>
</tr>
<tr>
<td>INT8</td>
<td>s</td>
<td></td>
<td>7</td>
<td>-128 to 127</td>
</tr>
</tbody>
</table>

DL COMPUTE IN HIGH PRECISION - HIGH PROCESSING TIME & HIGH ACCURACY IN RESULTS
DL COMPUTE IN LOWER PRECISION - FASTER RESULTS BUT POTENTIALLY LESS ACCURATE
SPEED ACCURACY TRADE OFF

© 2019 Intel Corporation
**PRECISION FOR DEEP LEARNING**

- **FP32 TRAINING** → **BF16 BIT TRAINING**
  - Better cache usage
  - Avoids bandwidth bottlenecks
  - Maximizes compute resources
  - Lesser silicon area & power

- **FP32 INFERENC** → **8 BIT INFERENC**
  - NEED - FP32 WEIGHTS → INT 8 WEIGHTS FOR INFEERENCE
  - BENEFIT - REDUCE MODEL SIZE AND ENERGY CONSUMPTION
  - CHALLENGE - POSSIBLE DEGRADATION IN PREDICTIVE PERFORMANCE
  - SOLUTION - QUANTIZATION WITH MINIMAL LOSS OF INFORMATION
  - DEPLOYMENT - OPENVINO TOOLKIT OR DIRECT FRAMEWORK (TF)

© 2019 Intel Corporation
**Quantization**

**Post Training Quantization**
Train normally, capture FP32 weights; convert to low precision before running inference and calibrate to improve accuracy.

**Quantization Aware Training**
Simulate the effect of quantization in the forward and backward passes using FAKE quantization.

**Reduce Precision & Keep Model Accuracy**
Collect statistics for the activation to find quantization factor.
Calibrate using subset data to improve accuracy, Convert FP32 weights to INT8.
Captured FP32 weights are quantized to int8 at each iteration after the weight updates.
Convert FP32 weights to INT8.
POST TRAINING QUANTIZATION

POST TRAINING QUANTIZATION
Train normally, capture FP32 weights; convert to low precision before running inference and calibrate to improve accuracy

DIRECT FRAMEWORK [TF]
Training with standard TF → Freeze graph → Calibrate → Convert FP32 weights to INT8 → Quantized graph → Inference with standard TF

FP32 WEIGHTS TRAINING → Collect statistics for the activation to find quantization factor → Calibrate using subset data to improve accuracy, Convert FP32 weights to INT8

8 BIT INFERENCE

OPENVINO TOOL

Off-line stage
Trained model - DL Framework → Model Optimizer → IR → Calibration Tool → Calibrated IR + statistics | + 8B profile → Run-time stage
Inference Engine Read Network → CPU plugin Load Network → I8 inference

https://github.com/IntelAI/models/tree/master/benchmarks

https://docs.openvinotoolkit.org/

© 2019 Intel Corporation
### DL BOOST (VNNI) ON 2ND GEN XEON VS FP32 ON 1ST GEN XEON

<table>
<thead>
<tr>
<th>Model</th>
<th>FP32 Performance</th>
<th>DL Boost (VNNI)</th>
<th>Relative Inference Throughput Performance [Higher Is Better]</th>
</tr>
</thead>
<tbody>
<tr>
<td>TensorFlow</td>
<td></td>
<td>4.0x @ 0.580 % accuracy loss</td>
<td>Tensorflow: ResNet-101</td>
</tr>
<tr>
<td>TensorFlow</td>
<td></td>
<td>3.9x @ 0.450 % accuracy loss</td>
<td>Tensorflow: ResNet-50</td>
</tr>
<tr>
<td>OpenVINO</td>
<td></td>
<td>3.9x @ 0.209 % accuracy loss</td>
<td>OpenVINO: ResNet-101</td>
</tr>
<tr>
<td>mxnet</td>
<td></td>
<td>3.8x @ 0.560 % accuracy loss</td>
<td>mxnet: ResNet-50</td>
</tr>
<tr>
<td>PyTorch</td>
<td></td>
<td>3.7x @ 0.609 % accuracy loss</td>
<td>PyTorch: SSD-VGG16</td>
</tr>
<tr>
<td>Caffe</td>
<td></td>
<td>3.0x @ 0.120 % accuracy loss</td>
<td>Caffe: SSD-MobileNet</td>
</tr>
<tr>
<td>ResNet-101</td>
<td></td>
<td>2.6x @ 0.003 mAP accuracy loss</td>
<td>ResNet-101</td>
</tr>
<tr>
<td>ResNet-50</td>
<td></td>
<td>2.5x @ 0.00019 mAP accuracy loss</td>
<td>ResNet-50</td>
</tr>
<tr>
<td>ResNet-101</td>
<td></td>
<td>2.2x @ 0.009 mAP accuracy loss</td>
<td>ResNet-101</td>
</tr>
<tr>
<td>ResNet-50</td>
<td></td>
<td>2.1x @ 0.007 % accuracy loss</td>
<td>ResNet-50</td>
</tr>
<tr>
<td>ResNet-101</td>
<td></td>
<td>2.1x @ 0.239 % accuracy loss</td>
<td>ResNet-101</td>
</tr>
</tbody>
</table>

**Baseline:** FP32 performance on Intel® Platinum 8180 Processor

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 SSE4 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

© 2019 Intel Corporation
DATA CENTRIC APPLICATION INNOVATION

white paper: https://www.intel.ai/papers/siemens-healthineers-ai-cardiac-imaging/
CASE STUDY: MEDICAL IMAGING (AI-BASED CARDIAC MRI SEGMENTATION)

Client: Siemens Healthineers is the global market leader in diagnostic imaging. Artificial intelligence (AI) is transforming care delivery and expanding precision medicine. Siemens Healthineers utilizes AI to help automate and standardize complex diagnostics to improve patient outcomes.

Challenge: Segmentation of anatomical regions for medical imaging requires hardware that delivers increased throughput for Deep Learning inferencing. This is used to automate and standardize clinical measurements, such as measurement of heart chambers, ejection fraction, and strain. Due to increased data generation and workflow demands, high performance Deep Learning inference is required to help reduce the burden on radiologists and cardiologists, enabling improved clinical decisions in near real-time.

Solution: The next generation Intel® Xeon® Scalable processors with Intel® Deep Learning Boost, in combination with the Intel® Distribution of OpenVINO™ toolkit, automate anatomical measurements in near real-time and improve workflow efficiency, while maintaining accuracy of the model without the need for discrete GPU investment.

RESULT

5.5X FASTER (NOW 201 FPS)

Using 2nd Generation Intel® Xeon® Scalable with Intel® DL Boost (INT8 precision) compared to FP32 precision

Left Ventricle, Right Ventricle (short-axis), Left Ventricle, Right Atrium and Left Atrium (long-axis)
## UNDERSTAND & INTEGRATE INTEL DL BOOST

### OVERVIEW
- **INTEL AI**
- **STRATEGY**
- **PORTFOLIO**
- **COMPUTE**

### DEEP DIVE
- **DL BOOST**
- **HARDWARE**
- **SOFTWARE**
- **PERFORMANCE**

### STEPS TO DEPLOY
- **USE CASES**
- **DEVELOP**
- **PARTNER**
DEEP LEARNING ON XEON

IMAGE
ResNet-50 Inference performance
Max Inference Throughput
At >7ms latency (images/sec)

| Nvidia V100 | 5925 |
| Nvidia T4 | 4189 |


Config 2: See Config slide 22 for Xeon and Nvidia configuration slide number 37 in this deck.

Config 3: See Config slide 23 for Xeon and Nvidia configuration slide number 37 in this deck.

Config 4: See Config slide 24 for Xeon and Nvidia configuration slide number 37 in this deck.

LANGUAGE
Machine Translation
TensorFlow Transformer-LT
English-German Translation Throughput (FP32, samples/sec)

| Nvidia T4 | 28.0 |

Recommendation Systems
mxNet Wide}&Deep
Throughput (FP32, samples/sec)

| Nvidia T4 | 997,987.0 |

RECOMMENDATION

Config 1: See Config slide 12 for detailed Xeon configuration.

Config 2: 1.87x performance improvement in MLPerf v0.5 MiniGo performance on 2S Intel® Xeon® Platinum 9282 processor compared to published MLPerf v0.5 MiniGo results on 2S Intel® Xeon® Platinum 8180 processor.

Config 3: Estimated score of 11.81 on the MLPerf v0.5 Training Closed Division – Reinforcement benchmark (MiniGo). Result not verified by MLPerf. MLPerf name and logo are trademarks. See [www.mlperf.org](http://www.mlperf.org) for more information.

Config 4: CPU MLPerf MiniGo Result Relative to MLPerf Baseline

| 2S Intel® Xeon® Platinum 9282 Processor (56 cores) | 11.8 |
| 2S Intel® Xeon® Platinum 8280 Processor (28 cores) | 6.3 |

REINFORCEMENT


Config 2: See Config slide 22 for Xeon and Nvidia configuration slide number 37 in this deck.

Config 3: See Config slide 23 for Xeon and Nvidia configuration slide number 37 in this deck.

Config 4: See Config slide 24 for Xeon and Nvidia configuration slide number 37 in this deck.

Performance results are based on testing as of 17/12/2019 and may not reflect all publicly available security updates. No product can be absolutely secure. See configuration disclosure for details. 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. Software and workload 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: [www.intel.com/content/www/us/en/re一心/technologies/indicators.html](http://www.intel.com/content/www/us/en/re一心/technologies/indicators.html). Other names and brands may be claimed as the property of others.
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 workloads. 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 names and brands may be claimed as the property of others.
DEVELOP THE FUTURE OF AI

OPTIMIZED SOFTWARE FOR THE BEST AI PERFORMANCE

**TOOLS**

- OpenVINO™
- nGraph

software.intel.com/OpenVINO/toolkit
github.com/NervanaSystems/ngraph

**LIBRARIES**

- MKL® DNN
- DAAL
- PYTHON

**FRAMEWORKS**

- TensorFlow
- PyTorch
- mxnet
- PaddlePaddle
- Caffe
- ONNX

github.com/intel/mkl-dnn
github.com/intelAI/models/tree/master/benchmarks
software.intel.com/en-us/intel-daal
software.intel.com/en-us/distribution-for-python

© 2019 Intel Corporation
Other names and brands may be claimed as the property of others.
**MAKING AI ON IA EASY**

**INTEL® DL BOOST ECOSYSTEM SUPPORT**

<table>
<thead>
<tr>
<th>Software Vendors</th>
<th>Cloud Service Providers</th>
<th>Enterprises</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffe, mxnet, OpenVINO, ONNX, PaddlePaddle, PyTorch, TensorFlow</td>
<td>Alibaba Group, AWS, Baidu Cloud, JD.COM, Microsoft, Tencent Cloud</td>
<td>AIhua, 联想, 海思科金, Neusoft, SIEMENS Healthineers</td>
</tr>
</tbody>
</table>

Performance results are based on testing as of dates shown in configuration and may not reflect all publicly available security updates. See configuration disclosure for details. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.
INTEL AI ECOSYSTEM

DEPLOY AI SOLUTIONS FASTER THROUGH INTEL'S STELLAR ECOSYSTEM

PARTNER INTEL AI BUILDERS

builders.intel.com/ai

LEADING AI SYSTEMS AND SOFTWARE SOLUTIONS >80 AVAILABLE FROM 200+ PARTNERS

BUILD INTEL AI PORTFOLIO

DELL* EMC* READY SOLUTIONS FOR AI

INSPIX FPGA SOLUTION FOR AI

LENOVO AI OFFERINGS

AND MORE

INNOVATE INTEL AI ACADEMY

software.intel.com/ai

ACCELERATING AI INNOVATION ACROSS CRITICAL BUSINESS WORKLOADS

© 2019 Intel Corporation
POWERING THE AI TRANSFORMATION
AI INFUSED INTO HARDWARE, SOFTWARE & ECOSYSTEM

SPANNING DEVICE TO INTELLIGENT EDGE & MULTI-CLOUD
UNMATCHED PORTFOLIO TO MOVE, STORE, PROCESS DATA

CPU, GPU, FPGA & ACCELERATOR OPTIMIZED FOR PROCESSING YOUR AI
AI COMPUTE PORTFOLIO FOR THE DATA-CENTRIC ERA

2ND GENERATION INTEL® XEON® SCALABLE PROCESSORS WITH INTEL® DEEP LEARNING BOOST
THE ONLY CPU WITH INTEGRATED AI ACCELERATION

© 2019 Intel Corporation
THANKS!
NOTICES AND DISCLAIMERS

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 product or component can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks.

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/benchmarks.

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.

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.

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.

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

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

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

Intel and the Intel logo are trademarks of Intel Corporation in the United States and other countries.

© 2019 Intel Corporation.