LEGAL NOTICES & 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.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENCE IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the presence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document 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.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm

Intel, Intel Xeon, Intel Xeon Phi™, Intel® Atom™ are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States or other countries.

Copyright © 2018, Intel Corporation

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

Intel does not control or audit the design or implementation of third party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase. The cost reduction scenarios described in this document are intended to enable you to get a better understanding of how the purchase of a given Intel product, combined with a number of situation-specific variables, might affect your future cost and savings. Nothing in this document should be interpreted as either a promise of or contract for a given level of costs.

Intel® Advanced Vector Extensions (Intel® AVX)* are designed to achieve higher throughput to certain integer and floating point operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you should consult your system manufacturer for more information.

Intel® Advanced Vector Extensions refers to Intel® AVX, Intel® AVX2 or Intel® AVX-512. For more information on Intel® Turbo Boost Technology 2.0, visit http://www.intel.com/go/turbo

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

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 information go to http://www.intel.com/design/literature.htm

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.
RISK FACTORS

The above statements and any others in this document that refer to plans and expectations for the second quarter, the year and the future are forward looking statements that involve a number of risks and uncertainties. Words such as “anticipates,” “expects,” “intends,” “plans,” “believes,” “seeks,” “estimates,” “may,” “will,” “should” and their variations identify forward-looking statements. Statements that refer to or are based on projections, uncertain events or assumptions also identify forward-looking statements. Many factors could affect Intel’s actual results, and variances from Intel’s current expectations regarding such factors could cause actual results to differ materially from those expressed in these forward-looking statements. Intel presently considers the following to be important factors that could cause actual results to differ materially from the company’s expectations. Demand for Intel’s products is highly variable and, in recent years, Intel has experienced declining orders in the traditional PC market segment. Demand could be different from Intel's expectations due to factors including changes in business and economic conditions; consumer confidence or income levels; customer acceptance of Intel's and competitors' products; competitive and pricing pressures, including actions taken by competitors; supply constraints and other disruptions affecting customers; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Intel operates in highly competitive industries and its operations have high costs that are either fixed or difficult to reduce in the short term. Intel's gross margin percentage could vary significantly from expectations based on capacity utilization; variations in inventory valuation, including variations related to the timing of qualifying products for sale; changes in revenue levels; segment product mix; the timing and execution of the manufacturing ramp and associated costs; excess or obsolete inventory; changes in unit costs; defects or disruptions in the supply of materials or resources; and product manufacturing quality/yields. Variations in gross margin may also be caused by the timing of Intel product introductions and related expenses, including marketing expenses, and Intel's ability to respond quickly to technological developments and to introduce new products or incorporate new features into existing products, which may result in restructuring and asset impairment charges. Intel's results could be affected by adverse economic, social, political and physical/infrastructure conditions in countries where Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Intel's results could be affected by the timing of closing of acquisitions, divestitures and other significant transactions. Intel's results could be affected by adverse effects associated with product defects and errata (deviations from published specifications), and by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust, disclosure and other issues, such as the litigation and regulatory matters described in Intel's SEC filings. An unfavorable ruling could include monetary damages or an injunction prohibiting Intel from manufacturing or selling one or more products, precluding particular business practices, impacting Intel’s ability to design its products, or requiring other remedies such as compulsory licensing of intellectual property. A detailed discussion of these and other factors that could affect Intel's results is included in Intel's SEC filings, including the company's most recent reports on Form 10-Q, Form 10-K and earnings release.

Rev. 4/15/14
• Deep Learning in practice

• Intel solutions

• Multi-node training
• Deep Learning in practice

• Intel solutions

• Multi-node training
## Why Now?

<table>
<thead>
<tr>
<th>Bigger Data</th>
<th>Better Hardware</th>
<th>Smarter Algorithms</th>
</tr>
</thead>
</table>
| Image: 1000 KB / picture  
Audio: 5000 KB / song  
Video: 5,000,000 KB / movie | Transistor density doubles every 18 months  
Cost / GB in 1995: $1000.00  
Cost / GB in 2017: $0.02 | Advances in algorithm innovation, including neural networks, leading to better accuracy in training models |
Training

Forward Propagation

Output

Expected

Penalty (error or cost)

Data

0.10 0.15 0.20 ...

0 1 0 ...

Person Cat Dog Bike

Person Cat Dog Bike

0.05
Inference

Forward Propagation

data

output

0.02 0.85 0.07 ⋮ 0.01

person cat dog bike

Back Propagation

expected

0 1 0 ⋮ 0

penalty (error or cost)

person cat dog bike
Typical Deep Learning kernels

Key computational kernels for a full-connected 2-layer neural network

In practice (1/2)

Innovation Cycle
- Label data: 15%
- Load data: 15%
- Augment data: 23%
- Experiment with topologies: 15%
- Tune hyper-parameters: 15%
- Support inference inputs: 8%
- Document results: 8%
- Labor-intensive: 20%
- Compute-intensive (Model Training): 80%
- Labor-intensive: 20%

Time-to-Solution
- Source Data: 1%
- Development Cycle: 1%
- Inferencing: 1%
- Inference within broader application: ∞

Production Deployment
- Data Integration & Management: 3%
- Data Processing: 2%
- Data Store: 95%
- AI: 2%
- Decision Process: 1%
- Broader Application: 1%
- Refresh: 1%

Source: Intel customer engagements

End-to-end Intel deployment solution
In practice (2/2)

Drive for better accuracy leads to

- More input data: more computation, training on a single server takes too long
- Re-training after deployment: training should take hours, not days
• Deep Learning in practice
• Intel solutions
• Multi-node training
### AI portfolio

**SOLUTIONS**
- Data Scientists
- Technical Services
- Reference Solutions

**PLATFORMS**
- Intel® AI DevCloud
- Intel® Deep Learning System
- Intel® Saffron

**TOOLS**
- Intel® Deep Learning Studio
- Intel® Deep Learning Deployment Toolkit
- Intel® Computer Vision SDK
- Intel® Movidius™ Software Development Kit (SDK)

**FRAMEWORKS**
- TensorFlow
- Caffe
- MKL
- OpenCV
- OpenVINO
- ONNX
- PyTorch
- CNTK
- Others

**LIBRARIES**
- Intel® MKL/MKL-DNN, cDNN, DAAL, Intel Python Distribution, etc.

**TECHNOLOGY**
- END-TO-END COMPUTE
- SYSTEMS & COMPONENTS

---

*Beta available
† Future
*Other names and brands may be claimed as the property of others.
All purpose

**Intel® Xeon® Scalable Processors**

*Known Compute for AI*

Scalable performance for widest variety of AI & other datacenter workloads – including breakthrough deep learning training & inference

**Flexible Acceleration**

**Intel® FPGA**

*Enhanced DL Inference*

Scalable acceleration for deep learning inference in real-time with higher efficiency, and wide range of workloads & configurations

**Deep Learning**

**Intel® Nervana™ Neural Network Processor**

*Deep Learning by Design*

Scalable acceleration with best performance for intensive deep learning training & inference, period

---

**AI Datacenter**
Performance Drivers for AI Workloads

**COMPUTE**

- **Higher number of operations per second**
  - Intel® Xeon® Platinum 8180 Processor (1-socket)
    - up to 3570 GFLOPS on SGEMM (FP32)
    - up to 5185 GOPS on IGEMM (Int8)

- **Increased parallelism and vectorization**
  - Intel® Xeon® Scalable Processor offers Intel® AVX-512 with up to 2 512bit FMA units computing in parallel per core

- **Higher number of cores**
  - Up to 28 core Intel® Xeon® Scalable Processors

**BANDWIDTH**

- **High Throughput, Low Latency**
  - Intel® Xeon® Scalable Processors offer up to 6 DDR4 channels per socket and new mesh architecture
  - Intel® Xeon® Processor 8180 Up to 199GB/s of STREAM Triad performance on a 2 socket system

- **Efficient Large Sized Caches**
  - Intel® Xeon® Scalable Processors offer increased private local Mid-Level Cache MLC up to 1MB per core

---


Configuration Details on Slide 47 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 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](http://www.intel.com/performance). Source: Intel measured as of June 2017. 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.
Up to 3.4x Integer Matrix Multiply Performance on Intel® Xeon® Platinum 8180 Processor

Matrix Multiply Performance on Intel® Xeon® Platinum 8180 Processor compared to Intel® Xeon® Processor E5-2699 v4

GEMM performance (Measured in GFLOPS) Higher is Better

<table>
<thead>
<tr>
<th>Single Precision Floating Point General Matrix Multiply</th>
<th>Integer General Matrix Multiply</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGEMM (FP32)</td>
<td>IGEMM (INT8)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2.3</td>
<td>3.4</td>
</tr>
</tbody>
</table>

1S Intel® Xeon® Processor E5-2699 v4
1S Intel® Xeon® Platinum 8180 Processor

8bit IGEMM will be available in Intel® Math Kernel Library (Intel® MKL) 2018 Gold to be released by end of Q3 2017

Enhanced matrix multiply performance on Intel® Xeon® Scalable Processor

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as “Spectre” and “Meltdown.” Implementation of these updates may make these results inapplicable to your device or system.

Configuration Details on Slide: 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. For more complete information visit: http://www.intel.com/performance. Source: Intel measured as of June 2017 Optimization Notice: Intel’s compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.
Up to 65% Performance Boost with Intel® AVX-512 on Intel® Xeon® Platinum 8180 processor

![Graph showing performance boost](image)

Test results above quantify the value add of Intel® AVX-512 to Convolution layer performance. All results shown above are measured on Intel® Xeon® Platinum 8180 Processor running AI topologies on Caffe framework with and without Intel® AVX-512 enabled.

Enhanced compute performance with Intel® AVX-512 on Intel® Xeon® Scalable Processor

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

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). Source: Intel measured as of June 2017.
## Inference Throughput Performance Images/Second

### 2 Socket Intel® Xeon® Scalable Processors Inference Throughput Performance Images/Second

<table>
<thead>
<tr>
<th></th>
<th>Images/Second</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Caffe</strong></td>
<td></td>
</tr>
<tr>
<td>ResNet50 BS = 1</td>
<td>165</td>
</tr>
<tr>
<td>ResNet50 BS = 64</td>
<td>448</td>
</tr>
<tr>
<td>VGG16 BS = 1</td>
<td>88</td>
</tr>
<tr>
<td>VGG16 BS = 64</td>
<td>176</td>
</tr>
<tr>
<td>VGG16 BS = 128</td>
<td>178</td>
</tr>
<tr>
<td><strong>TensorFlow</strong></td>
<td></td>
</tr>
<tr>
<td>ResNet-50 BS = 1</td>
<td>22</td>
</tr>
<tr>
<td>ResNet-50 BS = 64</td>
<td>117</td>
</tr>
<tr>
<td>ResNet-50 BS = 128</td>
<td>192</td>
</tr>
<tr>
<td>ResNet-50 BS = 1</td>
<td>195</td>
</tr>
<tr>
<td>VGG16 BS = 1</td>
<td>38</td>
</tr>
<tr>
<td>VGG16 BS = 64</td>
<td>139</td>
</tr>
<tr>
<td>VGG16 BS = 128</td>
<td>69</td>
</tr>
<tr>
<td><strong>Neon</strong></td>
<td></td>
</tr>
<tr>
<td>ResNet-50 BS = 1</td>
<td>39</td>
</tr>
<tr>
<td>ResNet-50 BS = 64</td>
<td>142</td>
</tr>
<tr>
<td>ResNet-50 BS = 1</td>
<td>233</td>
</tr>
</tbody>
</table>

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

Refer Configuration detail 4. Source: Intel measured or estimated as of November 2017.

Caffe, TF framework optimized to use MKLDNN libraries, Neon results shown here uses MKL2017.

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.
Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

Refer Configuration detail 5 Source: Intel measured or estimated as of November 2017.
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-dependant 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.
Intel® Nervana™ Neural Network Processor (NNP)

**Deep Learning by Design**

**Custom Hardware**
- Large reduction in time-to-train
- High compute density for DL workloads

**Blazing 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
Intel® Nervana™ NNP: Lake Crest Architecture

Interposer

Floorplan not to scale
FlexPoint™ Numerical Format Designed

**Float16**
- 11 bit mantissa precision (-1024 to 1023)
- Individual 5-bit exponents

**Flex16**
- 16 bit mantissa 45% more precision than Float16 (-32,768 to 32,767)
- Tensor-wide shared 5-bit exponent

*Flex16 accuracy on par with Float32 but with much smaller cores*

Intel® nGraph™ Library

High-Performance Execution Graph for Neural Networks

Use Cases

Models

Frameworks

Hardware

nGraph

NNP

Intel® Xeon® Processor

Integrated Graphics

FPGA

Intel® Movidus™ VPU
• Deep Learning in practice
• Intel solutions
• Multi-node training
Training parallelism options
Training parallelism options

Model

I
Input data

W
Weights or model

O
Output or activations
Training parallelism options

Hybrid
Addressing the scaling challenge

Hybrid parallelism to improve compute efficiency:

- Partition across activations and weights to minimize skewed matrices

Avoid global transfers via node groups:

- Activations transfer within a group
- Weight transfer across groups
Communication patterns in DL

Reduce the activations from layer N-1 and scatter at layer N

Common MPI collectives in DL
- Reduce Scatter
- AllGather
- AllReduce
- AlltoAll
**Intel® Omni-Path Architecture**

**World-Class Interconnect Solution for Shorter Time to Train**

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

---

**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

---

**STRONG FOUNDATION**
- Highly leverage existing Aries & Intel True Scale fabrics
- Excellent price/performance 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 EDR\(^{1}\):
  - 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

---

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

\(^{1}\)Intel® 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, IDU non-posted prefetch disabled, Snipp hold-off timers, Red Hat Enterprise Linux Server release 7.2 (Maipo), Intel® OPA testing performed with Intel Corporation Device 2410 – Series 100 HFI ASIC (80 silicon), OPA Switch; Series 100 Edge Switch – 48 port (80 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 587700 – 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, MXML_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® - SURFsara* Research Collaboration
Multi-Node Intel® Caffe ResNet-50
Scaling Efficiency on 2S Intel® Xeon® Platinum 8160 Processor Cluster

90% scaling efficiency with up to 74% Top-1 accuracy on 256 nodes

Configuration Details
Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

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. Source: Intel measured or estimated as of November 2017.
Intel® - SURFsara* Research Collaboration – Multi-Node Intel® Caffe ResNet-50
Time to Train on 2S Intel® Xeon® Platinum 8160 Cluster with Imagenet –1K

**Scaling Efficiency**

![Graph showing time to train on different node counts](image)

**Configuration Details**

- **Global BS= 4096**
  - 15170 Images/sec
  - Time: 70 mins
  - 74% Top-1

- **Global BS= 6400**
  - 19047 Images/sec
  - Time: 56 mins
  - 74% Top-1

- **Global BS= 8192**
  - 24240 Images/sec
  - Time: 44 mins
  - 74% Top-1

- **78 mins**
  - 76.6% Top-1

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). Source: Intel measured or estimated as of November 2017.
*MareNostrum4/Barcelona Supercomputing Center: https://www.bsc.es/

**Compute Nodes:** 2 sockets Intel® Xeon® Platinum 8160 CPU with 24 cores each @ 2.10GHz for a total of 48 cores per node, 2 Threads per core, L1d 32K; L1i cache 32K; L2 cache 1024K; L3 cache 33792K, 96 GB of DDR4, Intel® Omni-Path Host Fabric Interface, dual-rail. Software: Intel® MPI Library 2017 Update 4Intel® MPI Library 2019 Technical Preview OFI 1.5.0PSM2 w/ Multi-EP, 10 Gbit Ethernet, 200 GB local SSD, Red Hat® Enterprise Linux 6.7.

Intel® Caffe: Intel® version of Caffe; http://github.com/intel/caffe/, revision 8012927bf2bf70231bc7ff55de0b1bc11de4a69.

Intel® MKL version: mklml_lnx_2018.0.20170425; Intel® MLSL version: l_mlsl_2017.1.016

**Model:** Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50) and modified for wide-RedNet-50. Batch size as stated in the performance chart.

**Time-To-Train:** measured using "train" command. Data copied to memory on all nodes in the cluster before training. No input image data transferred over the fabric while training;

**Performance measured with:**
export OMP_NUM_THREADS=44 (the remaining 4 cores are used for driving communication), export I_MPI_FABRICS=tmi, export I_MPI_TMI_PROVIDER=psm2

OMP_NUM_THREADS=44 KMP AFFINITY="(S,L)-{0-87}" KMP HW USUBSET=11 MLSL NUM SERVERS=4 mpiexec.hydra -PSM2 -t n $SLURM JOB NUM NODES -ppn 1 -f hosts2 -genv OMP_NUM_THREADS 44 -env KMP AFFINITY "(S,L)-{0-87}" -genv KMP HW USUBSET 11 -genv I_MPI FABRICS tmi -genv I_MPI HYDRA BRANCH COUNT $SLURM JOB NUM NODES -genv I_MPI HYDRA PMI CONNECT alltoall sh -c 'cat /ilsvrc12_train_lmdb_stripped_64/data.mdb > /dev/null ; cat /ilsvrc12_val_lmdb_stripped_64/data.mdb > /dev/null ; ulimit -u 8192 ; ulimit -a ; numactl -H ; /caffe/build/tools/caffe train --solver=/caffe/models/intel_optimized_models/multinode/resnet_50_256_nodes_8k_batch/solver_poly_quick_large.prototxt -engine "MKL2017"


*SURFsara B.V. is the Dutch national high-performance computing and e-Science support center. Amsterdam Science Park, Amsterdam, The Netherlands.

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.
SGEMM, IGEMM, STREAM Configuration Details as of July 11th, 2017

SGEMM: System Summary 1-Node, 1 x Intel® Xeon® Platinum 8180 Processor GEMM - GF/s 3570.48 Processor Intel® Xeon® Platinum 8180 Processor (38.5M Cache, 2.50 GHz)Vendor Intel Nodes 1 Sockets 1 Cores 28 Logical Processors 56 Platform Lightning Ridge SKX Platform Comments Slots 12 Total Memory 384 GB Memory Configuration 12 slots / 32 GB / 2666 MT/s / DDR4 RDIMM Memory Comments OS Red Hat Enterprise Linux* 7.3 OS/Kernel Comments kernel 3.10.0-514.el7.x86_64 Primary / Secondary Software ic17 update2 Other Configurations BIOS Version: SE5C620.86B.01.00.0412.020920172159 HT No Turbo Yes 1-Node, 1 x Intel® Xeon® Platinum 8180 Processor on Lightning Ridge SKX with 384 GB Total Memory on Red Hat Enterprise Linux* 7.3 using ic17 update2. Data Source: Request Number: 2594, Benchmark: SGEMM, Score: 3570.48 Higher is better

SGEMM, IGEMM proof point: SKX: Intel(R) Xeon(R) Platinum 8180 CPU Cores per Socket 28 Number of Sockets 2 (only 1 socket was used for experiments) TDP Frequency 2.5 GHz BIOS Version SE5C620.86B.01.00.0412.020920172159 Platform Wolf PASS Ubuntu 16.04 Memory 384 GB Memory Speed Achieved 2666 MHz BDX: Intel(R) Xeon(R) CPU E5-2699v4 Cores per Socket 22 Number of Sockets 2 (only 1 socket was used for experiments) TDP Frequency 2.2 GHz BIOS Version GRRFSDP1.86B.0271.R00.1510301446 Platform Cottonwood Pass OS Red Hat 7.0 Memory 64 GB Memory Speed Achieved 2400 MHz

STREAM: 1-Node, 2 x Intel® Xeon® Platinum 8180 Processor on Neon City with 384 GB Total Memory System Configuration CFG1; Platform Wolf-Pass Qual; Number of Sockets 2; Motherboard Intel Corporation, S2600WFD; Memory 12x32GB DDR4 2666MH; OS Distribution "RHEL 7.3 Kernel: 3.10.0-514.el7.x86_64 x86_64" BIOS Version S5C620.86B.01.00.0470.040720170855 Storage Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC)

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.
Skylake AI Configuration Details as of July 11th, 2017

Platform: 2S Intel® Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to “performance” via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC).

Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance

Deep Learning Frameworks:

- **Caffe:** ([http://github.com/intel/caffe/](http://github.com/intel/caffe/)), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For “ConvNet” topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from [https://github.com/intel/caffe/tree/master/models/intel_optimized_models](https://github.com/intel/caffe/tree/master/models/intel_optimized_models) (GoogLeNet, AlexNet, and ResNet-50), [https://github.com/intel/caffe/tree/master/models/default_vgg_19](https://github.com/intel/caffe/tree/master/models/default_vgg_19) (VGG-19), and [https://github.com/soumith/convnet_benchmarks/tree/master/caffe/imagenet_winners](https://github.com/soumith/convnet_benchmarks/tree/master/caffe/imagenet_winners) (ConvNet benchmarks; files were updated to use newer Caffe prototxt format but are functionally equivalent). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

- **TensorFlow:** ([https://github.com/tensorflow/tensorflow](https://github.com/tensorflow/tensorflow)), commit id 207203253b6f8ea5e938a512798429f91d5b4e7e. Performance numbers were obtained for three convnet benchmarks: alexnet, googlenetv1, vgg([https://github.com/soumith/convnet_benchmarks/tree/master/master/tensorflow](https://github.com/soumith/convnet_benchmarks/tree/master/master/tensorflow)) using dummy data. GCC 4.8.5, Intel MKL small libraries version 2018.0.20170425, interop parallelism threads set to 1 for alexnet, vgg benchmarks, 2 for googlenet benchmarks, intra op parallelism threads set to 56, data format used is NCHW, KMP_BLOCKTIME set to 1 for googlenet and vgg benchmarks, 30 for the alexnet benchmark. Inference measured with --caffe time --forward_only --engine MKL2017option, training measured with --forward_backward_only option.

- **MxNet:** ([https://github.com/dmlc/mxnet/](https://github.com/dmlc/mxnet/)), revision 5efd91a7f36f6eae483ed82b0358c8d46b5a7aa20. Dummy data was used. Inference was measured with a modified version of benchmark_score.py which also runs backward propagation. Topology specs from [https://github.com/dmlc/mxnet/tree/master/example/image-classification/symbols](https://github.com/dmlc/mxnet/tree/master/example/image-classification/symbols). GCC 4.8.5, Intel MKL small libraries version 2018.0.20170425.

- **Neon:** ZP/MKL_CHWN branch commit id:52bd02d2c947a2adabb8a227166a7da5d9123b6d. Dummy data was used. The main.py script was used for benchmarking, in mkl mode. ICC version used : 17.0.3 20170404. Intel MKL small libraries version 2018.0.20170425.

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.
Broadwell AI Configuration Details as of July 11th, 2017

Platform: 2S Intel® Xeon® CPU E5-2699 v4 @ 2.20GHz (22 cores), HT enabled, turbo disabled, scaling governor set to “performance” via acpi-cpufreq driver, 256GB DDR4-2133 ECC RAM, CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3500 Series (480GB, 2.5in SATA 6Gb/s, 20nm, MLC).

Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact,1,0', OMP_NUM_THREADS=44, CPU Freq set with cpupower frequency-set -d 2.2G -u 2.2G -g performance

Deep Learning Frameworks:

- Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with “caffe time --forward_only” command, training measured with “caffe time” command. For “ConvNet” topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (GoogleLeNet, AlexNet, and ResNet-50), https://github.com/intel/caffe/tree/master/models/default_vgg_19 (VGG-19), and https://github.com/intel/caffe/tree/master/models/intel_optimized_models (GoogLeNet, AlexNet, and ResNet-50), https://github.com/intel/caffe/tree/master/models/default_vgg_19 (VGG-19), and https://github.com/intel/caffe/tree/master/models/default_vgg_19 (VGG-19), and https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners (ConvNet benchmarks; files were updated to use newer Caffe prototxt format but are functionally equivalent). GCC 4.8.5, Intel MKL small libraries version 2017.0.2.20170110.

- TensorFlow: (https://github.com/tensorflow/tensorflow), commit id 207203253b6f8ea5e938a512798429f91d5b4e7e. Performance numbers were obtained for three convnet benchmarks: alexnet, googleenetv1, vgg(https://github.com/soumith/convnet-benchmarks/tree/master/tensorflow) using dummy data. GCC 4.8.5, Intel MKL small libraries version 2018.0.20170425, interop parallelism threads set to 1 for alexnet, vgg benchmarks, 2 for googleenet background benchmarks, intra op parallelism threads set to 44, data format used is NCHW, KMP_BLOCKTIME set to 1 for googlenet and vgg benchmarks, 30 for the alexnet benchmark. Inference measured with --caffe-time-forward_only -engine MKL2017option, training measured with --forward-backward-only option.

- MXNet: (https://github.com/dmlc/mxnet/), revision e9f281a27584cdb78db8ce6b6e648b3dcb10d37. Dummy data was used. Inference was measured with “benchmark_score.py”, training was measured with a modified version of benchmark_score.py which also runs backward propagation. Topology specs from https://github.com/dmlc/mxnet/tree/master/example/image-classification/symbols. GCC 4.8.5, Intel MKL small libraries version 2017.0.2.20170110.

- Neon: ZP/MKL_CHWN branch commit id:52bd02ac89a72a2adabb8a2a7b166a7da5d9123b6d. Dummy data was used. The main.py script was used for benchmarking, in mkl mode. ICC version used: 17.0.3 20170404, Intel MKL small libraries version 2018.0.20170425.

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

Config 3
### Config Details for Skylake and Broadwell Inference throughput

**Platform**

<table>
<thead>
<tr>
<th>Benchmark Metric</th>
<th>images/s</th>
<th>images/s</th>
<th>images/s</th>
<th>images/s</th>
<th>images/s</th>
<th>images/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Framework</td>
<td>Caffe</td>
<td>Tensorflow</td>
<td>MxNet</td>
<td>Caffe</td>
<td>Tensorflow</td>
<td>MxNet</td>
</tr>
<tr>
<td># of Nodes</td>
<td>2699v4</td>
<td>2699v4</td>
<td>2699v4</td>
<td>2699v4</td>
<td>2699v4</td>
<td>2699v4</td>
</tr>
<tr>
<td>Sockets</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>BIOS</td>
<td>SESC610.86B.01.01.1016D0161452</td>
<td>SESC610.86B.01.01.1016D0161452</td>
<td>SESC610.86B.01.01.1016D0161452</td>
<td>SESC610.86B.01.01.1016D0161452</td>
<td>SESC610.86B.01.01.1016D0161452</td>
<td>SESC610.86B.01.01.1016D0161452</td>
</tr>
<tr>
<td>Enabled Cores</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td># of Nodes</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>Total Memory</td>
<td>384 GB</td>
<td>384 GB</td>
<td>384 GB</td>
<td>384 GB</td>
<td>384 GB</td>
<td>384 GB</td>
</tr>
<tr>
<td>Memory Configuration</td>
<td>Sockets: 2</td>
<td>Sockets: 2</td>
<td>Sockets: 2</td>
<td>Sockets: 2</td>
<td>Sockets: 2</td>
<td>Sockets: 2</td>
</tr>
<tr>
<td>Memory Comments</td>
<td>Kingston</td>
<td>Kingston</td>
<td>Kingston</td>
<td>Kingston</td>
<td>Kingston</td>
<td>Kingston</td>
</tr>
<tr>
<td>SSD</td>
<td>1000GB SSD + 240GB SSD</td>
<td>1000GB SSD + 240GB SSD</td>
<td>1000GB SSD + 240GB SSD</td>
<td>1000GB SSD + 240GB SSD</td>
<td>1000GB SSD + 240GB SSD</td>
<td>1000GB SSD + 240GB SSD</td>
</tr>
<tr>
<td>OS</td>
<td>CentOS Linux 7.3.1611-Core</td>
<td>CentOS Linux 7.3.1611-Core</td>
<td>CentOS Linux 7.3.1611-Core</td>
<td>CentOS Linux 7.3.1611-Core</td>
<td>CentOS Linux 7.3.1611-Core</td>
<td>CentOS Linux 7.3.1611-Core</td>
</tr>
<tr>
<td>HT</td>
<td>OFF</td>
<td>OFF</td>
<td>OFF</td>
<td>OFF</td>
<td>OFF</td>
<td>OFF</td>
</tr>
<tr>
<td>Turbo</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
</tr>
<tr>
<td>Computer Type</td>
<td>Server</td>
<td>Server</td>
<td>Server</td>
<td>Server</td>
<td>Server</td>
<td>Server</td>
</tr>
</tbody>
</table>

**System Configuration**

- **Processor**: Intel(R) Xeon(R) CPU E5-2699 v4 @ 2.20GHz
- **Sockets**: 2
- **Total Memory**: 24GB
- **BIOS**: SESC610.86B.01.01.1016D0161452
- **TORQUE**: Enabled Cores: 22
- **Slots**: 8 / 24
- **Processor**: Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz
- **Sockets**: 8 / 24
- **Total Memory**: 384 GB
- **BIOS**: SESC620.86B.0X.01.0017.070520171742
- **TORQUE**: Enabled Cores: 28
- **Slots**: 12 / 24
- **Total Memory**: 384 GB

**Framework Version**

- **Framework**: Caffe
- **Version**: 16f8c2b7ebdd6a9b
- **MKL DNN Library Version**: fbd4d7b0e82fe1d
- **MKL Library version**: version: mklml_lnx_2018.0.20
- **Performance Measurement Knobs**
  - **OMP_NUM_THREAD**: S=44, numactl -l
  - **OMP_NUM_THREAD**: S=44, numactl -l
  - **OMP_NUM_THREAD**: S=44, numactl -l
  - **OMP_NUM_THREAD**: S=44, numactl -l
  - **OMP_NUM_THREAD**: S=44, numactl -l
  - **OMP_NUM_THREAD**: S=44, numactl -l

**Performance Measurement Knobs**

- **OMP_NUM_THREAD**: S=44, numactl -l
- **OMP_NUM_THREAD**: S=44, numactl -l
- **OMP_NUM_THREAD**: S=44, numactl -l
- **OMP_NUM_THREAD**: S=44, numactl -l
- **OMP_NUM_THREAD**: S=44, numactl -l

**Compiler**

- **Gcc**: gcc 4.8.5
- **MKL DNN Library Version**: version: mklml_lnx_2018.0.20
- **Performance estimation**: obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.
## Benchmarks Details for Skylake and Broadwell Training throughput

<table>
<thead>
<tr>
<th>Benchmark Metric</th>
<th>Caffe/s</th>
<th>Tensorflow/s</th>
<th>MxNet/s</th>
<th>Caffe/s</th>
<th>Tensorflow/s</th>
<th>MxNet/s</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FrameWork</strong></td>
<td>Caffe</td>
<td>Tensorflow</td>
<td>MxNet</td>
<td>Caffe</td>
<td>Tensorflow</td>
<td>MxNet</td>
</tr>
<tr>
<td><strong>Platform</strong></td>
<td>2699v4</td>
<td>2699v4</td>
<td>2699v4</td>
<td>8180</td>
<td>8180</td>
<td>8180</td>
</tr>
<tr>
<td><strong>Sockets</strong></td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>Processor</strong></td>
<td>Intel(R) Xeon(R) CPU E5-2699 v4 @ 2.20GHz</td>
<td>Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BIOS</strong></td>
<td>SE5C610.86B.01.01.1016.05120162161452</td>
<td>SE5C620.86B.01.0017.070520171742</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>QDF (Indicate QS / ES2)</strong></td>
<td>Enabled Cores</td>
<td>22</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Slots</strong></td>
<td>8 / 24</td>
<td>12 / 24</td>
<td></td>
<td>8 / 24</td>
<td>12 / 24</td>
<td></td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>128 GiB</td>
<td>384 GiB</td>
<td></td>
<td>128 GiB</td>
<td>384 GiB</td>
<td></td>
</tr>
<tr>
<td><strong>OS</strong></td>
<td>CentOS Linux - 7.3.1611-CentOS Linux release 7.3.1611 (Core)</td>
<td>CentOS Linux release 7.3.1611-CentOS Linux release 7.3.1611 (Core)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Other Configurations</strong></td>
<td>HT: OFF</td>
<td>HT: ON</td>
<td>Turbo: ON</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Computer Type</strong></td>
<td>Server</td>
<td>Server</td>
<td>Server</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Framework Version</strong></td>
<td>1686c2b7ebd6a9b645ae6d6d911d</td>
<td>1686c2b7ebd6a9b645ae6d6d911d</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Topology Version, BATCHSIZE</strong></td>
<td>1686c2b7ebd6a9b645ae6d6d911d</td>
<td>1686c2b7ebd6a9b645ae6d6d911d</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Dataset, version</strong></td>
<td>No data layer</td>
<td>No data layer</td>
<td>No data layer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Compiler</strong></td>
<td>gcc 4.8.5</td>
<td>gcc 4.8.5</td>
<td>gcc 4.8.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MKL DNN Library Version</strong></td>
<td>fbd4d7be0e82fe1f1dc6d06a361c6541c65467045fd231c fbd4d7be0e82fe1f1dc6d06a361c6541c65467045fd231c</td>
<td>fbd4d7be0e82fe1f1dc6d06a361c6541c65467045fd231c fbd4d7be0e82fe1f1dc6d06a361c6541c65467045fd231c</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Performance Measurement Knobs</strong></td>
<td>OMP_NUM_THREADS=44, numaclt -1</td>
<td>KMP_BLOCKTIME=1, KMP_SETTINGS=1, KMP AFFINITY:granularity=fin e,verbose,compact,1,0, numaclt -1</td>
<td>OMP_NUM_THREADS=44, numaclt -1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.