FUTURE INTEL® XEON® SCALABLE PROCESSOR (CODENAME: CASCADE LAKE-SP)

Akhilesh Kumar, Sailesh Kottapalli, Ian M Steiner, Bob Valentine, Israel Hirsh, Geetha Vedaraman, Lily P Looi, Mohamed Arafa, Andy Rudoff, Sreenivas Mandava, Bahaa Fahim, Sujal A Vora

Intel Corporation, 2018
Notices and Disclaimers

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

No computer system can be absolutely secure.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

Optimization Notice: Intel'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.

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 http://www.intel.com/performance.

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

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 Xeon are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

* Other names and brands may be claimed as the property of others. © 2017 Intel Corporation.
Outline

- Intel® Xeon® Scalable Processor Roadmap
- Focus Areas for Cascade Lake-SP
  - Instruction Enhancement for AI/Deep Learning Inference
  - Intel® Optane™ DC Persistent Memory
  - Side Channel Mitigations
- Wrap up
First Generation Intel® Xeon® Scalable Processor

Introduced in July 2017

- Skylake-SP core microarchitecture with data center specific enhancements
- Intel® AVX-512 with 32 DP flops per cycle per core
- Data center optimized cache hierarchy – 1MB L2 per core, non-inclusive L3
- New Intel® Mesh architecture
- Enhanced 6 channel memory subsystem
- 48 lanes of PCIe Gen3 with integrated DMA, NTB, and VMD devices
- New Intel® Ultra Path Interconnect (Intel® UPI)

### Features

<table>
<thead>
<tr>
<th>Features</th>
<th>Intel® Xeon® Scalable Processor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cores and Threads Per CPU</td>
<td>Up to 28 cores and 56 threads</td>
</tr>
<tr>
<td>Last-level Cache (LLC)</td>
<td>Up to 38.5 MB (non-inclusive)</td>
</tr>
<tr>
<td>QPI/UPI Speed (GT/s)</td>
<td>Up to 3x UPI @ 10.4 GT/s</td>
</tr>
<tr>
<td>PCIe* Lanes/ Controllers</td>
<td>Up to 48 / 12 / PCIe 3.0 (2.5, 5, 8 GT/s)</td>
</tr>
<tr>
<td>Memory Population</td>
<td>Up to 6 channels of up to 2 RDIMMs, LRDIMMs, or 3DS LRDIMMs</td>
</tr>
<tr>
<td>Max Memory Speed</td>
<td>Up to 2666 MHz</td>
</tr>
</tbody>
</table>

Foundation for Accelerating Data Center Innovations
Cascade Lake CPU is designed to be compatible with first-gen Intel® Xeon® Scalable platform

- Same core count, cache size, and I/O speeds as first-gen
- Process tuning, frequency push, targeted performance improvements
- Architectural improvements through targeted instruction set enhancements
- New platform capabilities with support for Intel® Optane™ DC persistent memory
- Hardware enhancements for protection against side-channel methods

<table>
<thead>
<tr>
<th>Features</th>
<th>Cascade Lake CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cores and Threads</td>
<td>Up to 28 Cores and 56 Threads</td>
</tr>
<tr>
<td>Last-level Cache</td>
<td>Up to 38.5 MB (non-inclusive)</td>
</tr>
<tr>
<td>UPI Speed (GT/s)</td>
<td>Up to 3x UPI @ 10.4 GT/s</td>
</tr>
<tr>
<td>PCIe® 3.0 Lanes</td>
<td>Up to 48 lanes with 12 controllers</td>
</tr>
<tr>
<td>Memory Speed</td>
<td>Up to 6 channels @ up to 2666 MHz</td>
</tr>
</tbody>
</table>

Next Step in the Intel® Xeon® Scalable Processor
AI/DEEP LEARNING ENHANCEMENTS
AI/Deep Learning Software Optimizations on first generation Intel® Xeon® Scalable Processor

Inference Throughput on Intel Caffe ResNet50

(1) Up to 5.4X performance improvement with software optimizations on Caffe ResNet-50 in 10 months with 2 socket Intel® Xeon® Scalable Processor, Configuration Details 1, 2. Performance measurements 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. - SSE2: Intel Xeon® Scalable Platinum 8180 Processor

- SSE3: Intel Xeon® Scalable Platinum 8180 Processor

- SSSE3: Intel Xeon® Scalable Platinum 8180 Processor

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 changes 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 April 2018.
Neural Machine Translation Software Optimization on first generation Intel® Xeon® Scalable Processor

MxNet Amazon® C5 (Intel® Xeon® Processor)
NMT (German to English)

Up to 14X higher inference performance

Configuration Details

Performance measurements 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.

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/benchmarks. Source: Intel measured as of May 2018.
Cascade Lake Vector Neural Network Instructions

Vector Neural Network Instruction (VNNI) on Cascade Lake accelerates Deep Learning and AI inference workloads

- **VNNI**: A new set of Intel® Advanced Vector Extension (Intel® AVX-512) instructions
  - 8-bit (int8) new instruction (VPDPBUSD)
    - Fuses 3 instructions in inner convolution loop using int8 data type
  - 16-bit (int16) new instruction (VPDPWSSD)
    - Fuses 2 instructions in inner convolution loop using int16 data type
AI/DL Inference Enhancements on INT16 with VNNI

Current AVX-512 instructions to perform INT16 convolutions: \texttt{vpmaddwd}, \texttt{vpadd}

New instructions for accelerating AI on Intel® Xeon® Scalable processors using int16 data

VNNI instruction to accelerate INT16 convolutions: \texttt{vpdpwssd}
AI/DL Inference Enhancements on INT8 with VNNI

Current AVX-512 instructions to perform INT8 convolutions: **vpmaddubsw, vpmaddwd, vpaddd**

New instructions for accelerating AI on Intel® Xeon® Scalable processors using int8 data

VNNI instruction to accelerate INT8 convolutions: **vpdpbusd**
VNNI Per Core Throughput

Vector Elements Processed per Cycle on Different Data Types

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Elements Processed/Cycle</th>
<th>Instruction(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP32</td>
<td>64</td>
<td>vfmadd231ps</td>
</tr>
<tr>
<td>Int16</td>
<td>64</td>
<td>vpmaddw, vpadd</td>
</tr>
<tr>
<td>Int8</td>
<td>85.33</td>
<td>vpmaddubsw, vpmaddw, vpadd</td>
</tr>
<tr>
<td>FP32</td>
<td>64</td>
<td>vfmadd231ps</td>
</tr>
<tr>
<td>VNNI Int16</td>
<td>128</td>
<td>vdpwssd</td>
</tr>
<tr>
<td>VNNI Int8</td>
<td>256</td>
<td>vdpbusd</td>
</tr>
</tbody>
</table>

Input 32bit / Output 32bit
Input 16bit / Output 32bit
Input 8bit / Output 32bit
Input 32bit / Output 32bit
Input 16bit / Output 32bit
Input 8bit / Output 32bit

Performance measurements 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 as of May 2018.
Inference Throughput with VNNI

Intel optimization for Caffe ResNet-50

Estimated Throughput on Cascade Lake with VNNI

Framework & Library support
Caffe mxnet TensorFlow

1.0 FP32

1.0X Jul’17
2.8X Jan’18
5.4X Aug’18

Intel® Xeon® Scalable Processor – Hot Chips 2018

1 Intel® Optimization for Caffe Resnet-50 performance does not necessarily represent other framework performance. 2 Based on Intel internal testing: 1X (7/11/2017), 2.8X (1/19/2018) and 5.4X (7/26/2018) performance improvement based on Intel® Optimization for Caffe ResNet-50 inference throughput performance on Intel® Xeon® Scalable Processor. 3 11X (7/25/2018) Results have been estimated using internal Intel analysis, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance. Performance results are based on testing as of 7/11/2017(1x), 1/19/2018(2.8x) & 7/26/2018(5.4) and may not reflect all publicly available security update. See configuration disclosure for details (config 5). No product can be absolutely secure. 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. Other names and brands may be claimed as the property of others.
REIMAGINING DATA CENTER MEMORY HIERARCHY
Growing Gap Between Memory Hierarchy

Limitations to traditional architecture impede unified data management

**MEMORY**

- DRAM (HOT TIER)
  - Cost prohibitive for data intensive applications

**STORAGE**

- SSD (WARM TIER)
  - Performance impedes data intensive applications
- HDD/TAPE (COLD TIER)
  - Media capability limits usage to cold tier
Growing Gap Between Memory Hierarchy

Limitations to traditional architecture impede unified data management

- **MEMORY**
  - **DRAM**
    - Hot Tier
    - Latency: \(~1000\)x
    - Bandwidth: \(~0.1\)x
    - Capacity/$: \(~40\)x
  - Cost prohibitive for data intensive applications

- **STORAGE**
  - **SSD**
    - Warm Tier
  - Performance impedes data intensive applications
  - **HDD/TAPE**
    - Cold Tier
    - Media capability limits usage to cold tier

*Actual performance and price may vary
Intel Innovations Address These Gaps

**MEMORY**
- Improving Memory Capacity

**STORAGE**
- Improving SSD Performance
- Delivering Efficient and Scalable Storage

**Diagram:***
- DRAM Hot Tier
- SSD Warm Tier
- HDD/Tape Cold Tier
- Intel® 3D NAND SSD

*Intel® Xeon® Scalable Processor – Hot Chips 2018*
Big and Affordable Memory

High Performance Storage

Direct Load/Store Access

Native Persistence

Supported on future Intel® Xeon® Scalable Processors (Cascade Lake)

128, 256, 512GB

DDR4 Pin Compatible

Hardware Encryption

High Reliability
Intel® Optane™ DC Persistent Memory Hardware Interface

- DDR4 electrical and physical interface with proprietary protocol extensions
- Memory channel can be shared between DDR4 and Intel® Optane™ DC persistent memory modules
  - Enables systems to support greater than 3TB of system memory per CPU socket
- Cache line size accesses
- Idle latency close to DDR4 DIMMs

Performance measurements 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® 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/benchmarks. Source: Intel measured as of May 2018.
Hardware Interface: Persistence Domain

- Core: L1, L1, L2, L3
- CPU CACHES: CLWB + fence, CLFLUSHOPT + fence, CLFLUSH, NT stores + fence, WBINVD (kernel only)
- Custom Power fail protected domain indicated by ACPI property: CPU Cache Hierarchy
- Minimum Required Power fail protected domain: Memory subsystem
- WPQ: ADR, WPQ Flush (kernel only)
- DIMM
The SNIA NVM Programming Model

For more information, visit software.intel.com/pmem
The Persistent Memory Development Kit - pmdk

PMDK is a collection of libraries
- Developers pull only what they need
  - Low level programming support
  - Transaction APIs
- Fully validated
- Performance tuned

Open source & product neutral

software.intel.com/pmem
Usage Example: High Performance Storage

Results have been estimated based on tests conducted on pre-production systems, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to www.intel.com/benchmarks.
Usage Example: Data Replication with Persistent Memory over Fabric

Average 4KB Write I/O Round Trip Time Comparison
NVMe+NAND SSD vs. PMoF

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to www.intel.com/benchmarks. *Three 9s and five 9s availability assumes bi-weekly maintenance restarts.
HARDWARE MITIGATION FOR SIDE CHANNEL
Cascade Lake Mitigations for Side-Channel Methods

Cascade Lake implements hardware mitigations against targeted side-channel methods

<table>
<thead>
<tr>
<th>Variant</th>
<th>Side-Channel Method</th>
<th>Mitigation on Cascade Lake</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variant 1</td>
<td>Bounds Check Bypass</td>
<td>OS/VMM</td>
</tr>
<tr>
<td>Variant 2</td>
<td>Branch Target Injection</td>
<td>Hardware + OS/VMM</td>
</tr>
<tr>
<td>Variant 3</td>
<td>Rogue Data Cache Load</td>
<td>Hardware</td>
</tr>
<tr>
<td>Variant 3a</td>
<td>Rogue System Register Read</td>
<td>Firmware</td>
</tr>
<tr>
<td>Variant 4</td>
<td>Speculative Store Bypass</td>
<td>Firmware + OS/VMM or runtime</td>
</tr>
<tr>
<td></td>
<td>L1 Terminal Fault</td>
<td>Hardware</td>
</tr>
</tbody>
</table>

Cascade Lake SP expected to provide higher performance over software mitigations available for existing products

For additional information related to security updates and side channel methods on Intel® products, please visit https://www.intel.com/content/www/us/en/architecture-and-technology/facts-about-side-channel-analysis-and-intel-products.html
Future Intel® Xeon® Scalable Processor (Codename: Cascade Lake-SP)

- Process Tuning, Frequency Boost, Targeted Performance Improvements
- AI/DL Enhancement through VNNI
- Side-Channel Analysis Mitigations
- Software Libraries and Optimizations

Intel® Xeon® Scalable Platform

Further Accelerating Data Center Innovations
<table>
<thead>
<tr>
<th>Framework</th>
<th>Caffe</th>
<th>Caffe</th>
<th>Caffe</th>
<th>Caffe</th>
<th>Caffe</th>
<th>Caffe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Branch</td>
<td>master</td>
<td>master</td>
<td>master</td>
<td>master</td>
<td>master</td>
<td>master</td>
</tr>
<tr>
<td>Version</td>
<td>f6d01efeb93f7072deca3796a4b89c5457fd5daf2aa8a567dc1f2aa8a567e1</td>
<td>f6d01efeb93f7072deca3796a4b89c5457fd5daf2aa8a567dc1f2aa8a567e1</td>
<td>f6d01efeb93f7072deca3796a4b89c5457fd5daf2aa8a567dc1f2aa8a567e1</td>
<td>f6d01efeb93f7072deca3796a4b89c5457fd5daf2aa8a567dc1f2aa8a567e1</td>
<td>f6d01efeb93f7072deca3796a4b89c5457fd5daf2aa8a567dc1f2aa8a567e1</td>
<td>f6d01efeb93f7072deca3796a4b89c5457fd5daf2aa8a567dc1f2aa8a567e1</td>
</tr>
<tr>
<td>Platform</td>
<td>SKX_8180</td>
<td>SKX_8180</td>
<td>SKX_8180</td>
<td>SKX_8180</td>
<td>SKX_8180</td>
<td>SKX_8180</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>Processor</td>
<td>Intel(R) Xeon(R) Platinum 8100 CPU @ 2.50GHz / 28 cores</td>
<td>Intel(R) Xeon(R) Platinum 8100 CPU @ 2.50GHz / 28 cores</td>
<td>Intel(R) Xeon(R) Platinum 8100 CPU @ 2.50GHz / 28 cores</td>
<td>Intel(R) Xeon(R) Platinum 8100 CPU @ 2.50GHz / 28 cores</td>
<td>Intel(R) Xeon(R) Platinum 8100 CPU @ 2.50GHz / 28 cores</td>
<td>Intel(R) Xeon(R) Platinum 8100 CPU @ 2.50GHz / 28 cores</td>
</tr>
<tr>
<td>BIOS</td>
<td>SE6C620.868.000.0100.0004.071220170215</td>
<td>SE6C620.868.000.0100.0004.071220170215</td>
<td>SE6C620.868.000.0100.0004.071220170215</td>
<td>SE6C620.868.000.0100.0004.071220170215</td>
<td>SE6C620.868.000.0100.0004.071220170215</td>
<td>SE6C620.868.000.0100.0004.071220170215</td>
</tr>
<tr>
<td>Disks</td>
<td>sda RS3WC080 HHD 744.1GB,sdb RS3WC080 HHD 1.5TB,sdc RS3WC080 HHD 5.5TB</td>
<td>sda RS3WC080 HHD 744.1GB,sdb RS3WC080 HHD 1.5TB,sdc RS3WC080 HHD 5.5TB</td>
<td>sda RS3WC080 HHD 744.1GB,sdb RS3WC080 HHD 1.5TB,sdc RS3WC080 HHD 5.5TB</td>
<td>sda RS3WC080 HHD 744.1GB,sdb RS3WC080 HHD 1.5TB,sdc RS3WC080 HHD 5.5TB</td>
<td>sda RS3WC080 HHD 744.1GB,sdb RS3WC080 HHD 1.5TB,sdc RS3WC080 HHD 5.5TB</td>
<td>sda RS3WC080 HHD 744.1GB,sdb RS3WC080 HHD 1.5TB,sdc RS3WC080 HHD 5.5TB</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>Memory</td>
<td>Micron</td>
<td>Micron</td>
<td>Micron</td>
<td>Micron</td>
<td>Micron</td>
<td>Micron</td>
</tr>
<tr>
<td>Memory Configuration</td>
<td>12slots / 32 GB / 2666 MHz</td>
<td>12slots / 32 GB / 2666 MHz</td>
<td>12slots / 32 GB / 2666 MHz</td>
<td>12slots / 32 GB / 2666 MHz</td>
<td>12slots / 32 GB / 2666 MHz</td>
<td>12slots / 32 GB / 2666 MHz</td>
</tr>
<tr>
<td>Hyper-Threading</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
<td>ON</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>Topology</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
</tr>
<tr>
<td>Batchsize</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
</tr>
<tr>
<td>Dataset</td>
<td>NoDataLayer</td>
<td>NoDataLayer</td>
<td>NoDataLayer</td>
<td>NoDataLayer</td>
<td>NoDataLayer</td>
<td>NoDataLayer</td>
</tr>
<tr>
<td>Engine</td>
<td>version: ae00102be506ed00fe2099565757d3f2aa8a567e1</td>
<td>version: ae00102be506ed00fe2099565757d3f2aa8a567e1</td>
<td>version: ae00102be506ed00fe2099565757d3f2aa8a567e1</td>
<td>version: ae00102be506ed00fe2099565757d3f2aa8a567e1</td>
<td>version: ae00102be506ed00fe2099565757d3f2aa8a567e1</td>
<td>version: ae00102be506ed00fe2099565757d3f2aa8a567e1</td>
</tr>
<tr>
<td>IP</td>
<td>172.18.0.2</td>
<td>172.18.0.2</td>
<td>172.18.0.2</td>
<td>172.18.0.2</td>
<td>172.18.0.2</td>
<td>172.18.0.2</td>
</tr>
<tr>
<td>Kernel</td>
<td>4.4.0-109-generic</td>
<td>4.4.0-109-generic</td>
<td>4.4.0-109-generic</td>
<td>4.4.0-109-generic</td>
<td>4.4.0-109-generic</td>
<td>4.4.0-109-generic</td>
</tr>
</tbody>
</table>
## Configuration details of Amazon EC2 C5.18xlarge 1 node systems

<table>
<thead>
<tr>
<th>Benchmark Segment</th>
<th>AI/ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark type</td>
<td>Inference</td>
</tr>
<tr>
<td>Benchmark Metric</td>
<td>Sentence/Sec</td>
</tr>
<tr>
<td>Framework</td>
<td>Official mxnet</td>
</tr>
<tr>
<td>Topology</td>
<td>GNMT(sockeye)</td>
</tr>
<tr>
<td># of Nodes</td>
<td>1</td>
</tr>
<tr>
<td>Platform</td>
<td>Amazon EC2 C5.18xlarge instance</td>
</tr>
<tr>
<td>Sockets</td>
<td>2S</td>
</tr>
<tr>
<td>Processor</td>
<td>Intel® Xeon® Platinum 8124M CPU @ 3.00GHz (Skylake)</td>
</tr>
<tr>
<td>BIOS</td>
<td>N/A</td>
</tr>
<tr>
<td>Enabled Cores</td>
<td>18 cores / socket</td>
</tr>
<tr>
<td>Platform</td>
<td>N/A</td>
</tr>
<tr>
<td>Slots</td>
<td>N/A</td>
</tr>
<tr>
<td>Total Memory</td>
<td>144GB</td>
</tr>
<tr>
<td>Memory Configuration</td>
<td>N/A</td>
</tr>
<tr>
<td>SSD</td>
<td>EBS Optimized 200GB, Provisioned IOPS SSD</td>
</tr>
<tr>
<td>OS</td>
<td>Red Hat 7.2 (HVM) Amazon Elastic Network Adapter (ENA) Up to 10 Gbps of aggregate network bandwidth</td>
</tr>
<tr>
<td>Network Configurations</td>
<td>Installed Enhanced Networking with ENA on Centos Placed the all instances in the same placement</td>
</tr>
<tr>
<td>HT</td>
<td>ON</td>
</tr>
<tr>
<td>Turbo</td>
<td>ON</td>
</tr>
<tr>
<td>Computer Type</td>
<td>Server</td>
</tr>
</tbody>
</table>
## Configuration details of Amazon EC2 C5.18xlarge 1 node systems

| Framework Version | mxnet mkldnn: [https://github.com/apache/incubator-mxnet/](https://github.com/apache/incubator-mxnet/)
|  | 4950f6649e329b23a1efdc40aaa25260d47b4195 |
| Topology Version | GNMT: [https://github.com/awslabs/sockeye/tree/master/tutorials/wmt](https://github.com/awslabs/sockeye/tree/master/tutorials/wmt) |
| Batch size | GNMT: 1 2 8 16 32 64 128 |
| MKLDNN | F5218ff4fd2d16d13aada2e632af82514fee3 |
| MKL | Version: parallel_studio_xe_2018_update1
| Compiler | g++: 4.8.5
gcc: 7.2.1 |
Configuration Details for Inference Throughput with VNNI

1x inference throughput improvement in July 2017:
Tested by Intel 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. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b2b150b5c. 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 (ResNet-50),and https://github.com/soumith/convnet-benchmarks/tree/master/cafe/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”.

2.8x inference throughput improvement in January 2018:
Tested by Intel as of Jan 19th 2018 Processor: 2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON, Turbo ON Total Memory 376.46GB (12slots / 32 GB / 2666 MHz). CentOS Linux-7.3.1611-Core, SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework Intel® Optimization for caffe version:f6d01efbe93f70726ea3796a4b89c612365a6341 Topology::resnet_50_v1 BIOS:SE5C620.86B.00.01.0009.101920170742 MKLDNN: version: ae00102be506ed0fe2099c6557df2aa8ad57ec1 NoDataLayer. . Datatype:FP32 Batchsize=64 Measured: 652.68 imgs/sec vs Tested by Intel 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. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b2b150b5c. 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 (ResNet-50),and https://github.com/soumith/convnet-benchmarks/tree/master/cafe/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”.

Future Intel® Xeon® Scalable Processor – Hot Chips 2018
5.4x inference throughput improvement in August 2018:

Tested by Intel as of measured July 26th 2018 : 2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON, Turbo ON Total Memory 376.46GB (12slots / 32 GB / 2666 MHz). CentOS Linux-7.3.1611-Core, kernel: 3.10.0-862.3.3.el7.x86_64, SSD sda RS3WC080 HDD 744.1GB, sdb RS3WC080 HDD 1.5TB, sdc RS3WC080 HDD 5.5TB, Deep Learning Framework Intel® Optimization for caffe version:a3d5b022fe026e9092fc7abc7654b1162ab9940d Topology:resnet_50_v1 BIOS:SE5C620.86B.00.01.0013.030920180427 MKLDNN: version:464c268e544bae26f9b85a2acb9127c7664a4c93 instances: 2 instances: https://software.intel.com/en-us/articles/boosting-deep-learning-training-inference-performance-on-xeon-and-xeon-phi) NoDatasocket:2 (Results on Intel® Xeon® Scalable Processor were measured running multiple instances of the framework. Methodology described hereLayer. Datatype: INT8 Batchsize=64 Measured: 1233.39 imgs/sec vs Tested by Intel as of July 11th 2017: 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.elic7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC).Performance measured with: Environment variables: KMP_AFFINITY=’granularity=very_fine, compact’, OMP_NUM_THREADS=54, CPU Freq set with cpupower frequency-set -d 2.5G-u 3.8G -g performance. Caffe: (https://github.com/intel/caffe/), revision f96b759f71b2281835f690af26715882b150b5c. 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 (ResNet-50). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.1.20170425. Caffe run with “numactl -l 1”.

11X inference throughput improvement with CascadeLake:
Future Intel Xeon Scalable processor (codename Cascade Lake) results have been estimated or simulated using internal Intel analysis or architecture simulation tools or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance vs Tested by Intel as of July 11th 2017: 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.elic7.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. Caffe: (https://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 (ResNet-50), Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.1.20170425. Caffe run with “numactl -l 1”.

Configuration Details for Inference Throughput with VNNI