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.

Outline

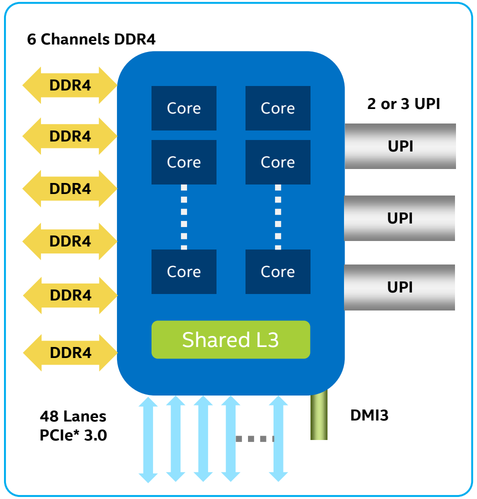
- Intel® Xeon® Scalable Processor Roadmap
- Focus Areas for Cascade Lake-SP
  - Instruction Enhancement for AI/Deep Learning Inference
  - Intel® Optane™ DC Persistent Memory
  - Side Channel Mitigations
- Wrap up

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



Foundation for Accelerating Data Center Innovations

Next Step in the Intel® Xeon® Scalable Processor

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

Grantley Platform		Purley Platform	
Intel® Microarchitecture Codenamed Haswell		Intel® Microarchitecture Codenamed Skylake	
Haswell 22nm New Micro-architecture	Broadwell 14nm	Skylake-SP 14nm New Micro-architecture	Cascade Lake-SP 14nm

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

AI/DEEP LEARNING ENHANCEMENTS

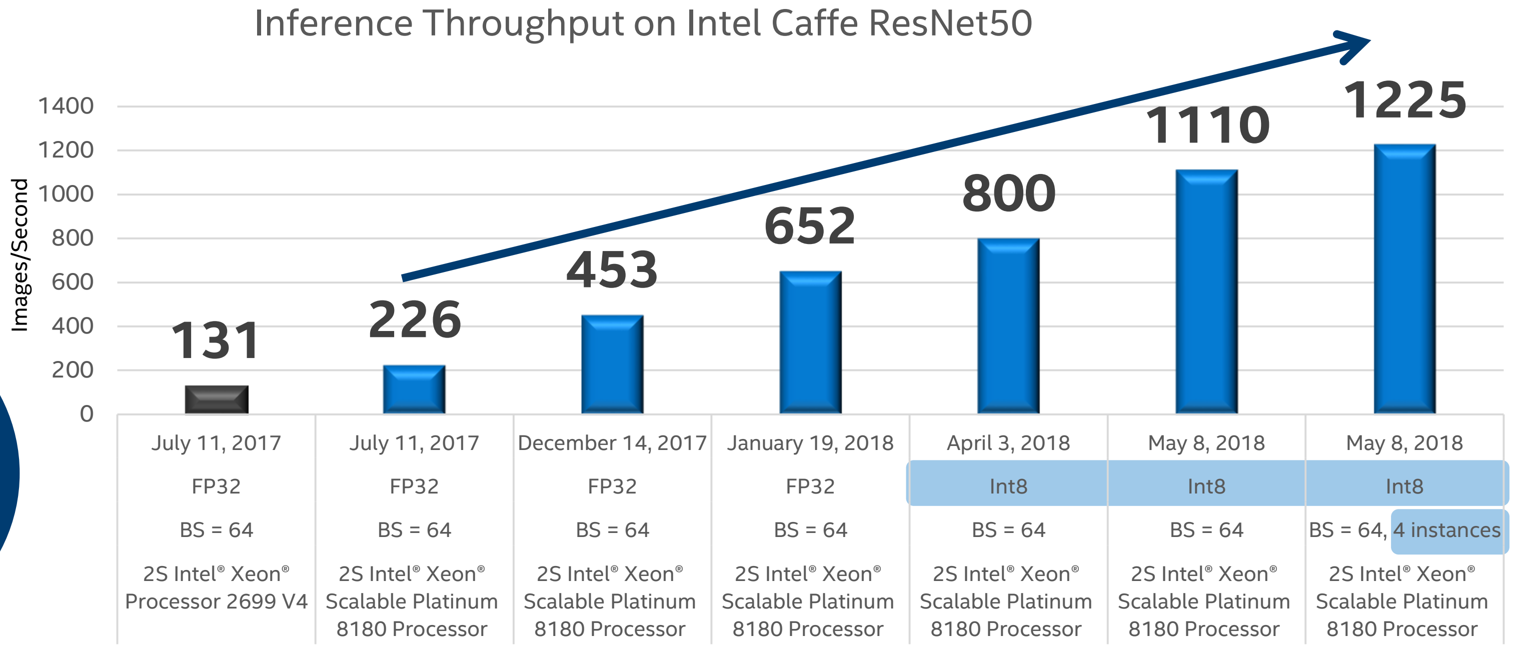
AI/Deep Learning Software Optimizations

on first generation Intel® Xeon® Scalable Processor

5.4X(1)

In 10 months

since Intel® Xeon® Scalable Processor launch



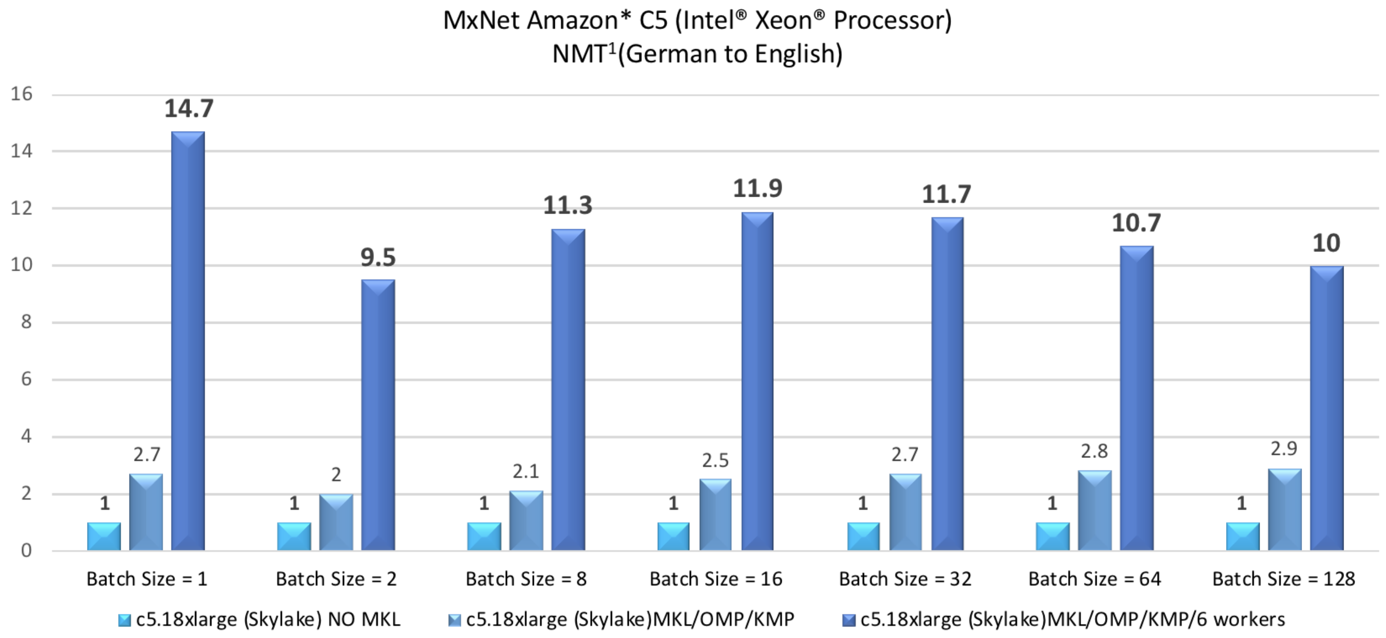
(1) Up to 5.4X performance improvement with software optimizations on Caffe Resnet-50 in 10 months with 2 socket Intel® Xeon® Scalable Processor, Configuration Details 1, 2. Performance measurements were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

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 April 2018.

Neural Machine Translation Software Optimization

on first generation Intel® Xeon® Scalable Processor



Up to

14X (1)

higher inference performance

Configuration Details 3 4

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.

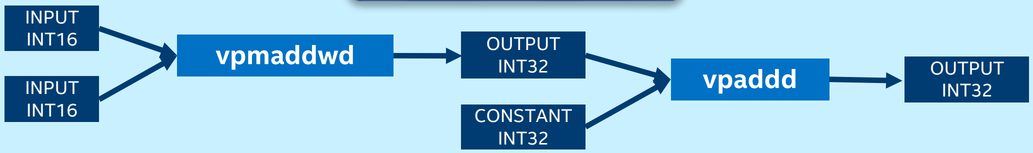
Cascade Lake Vector Neural Network Instructions

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

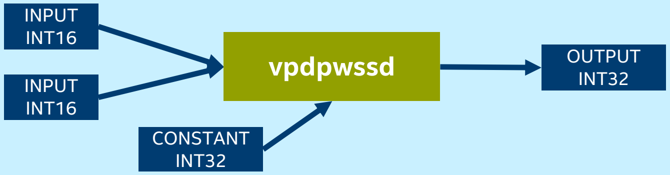
Xeon Scalable AVX-512



Current AVX-512 instructions to perform INT16 convolutions: **vpmaddwd, vpaddd**

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

Cascade Lake SP VNNI



VNNI instruction to accelerate INT16 convolutions: **vpdpwssd**

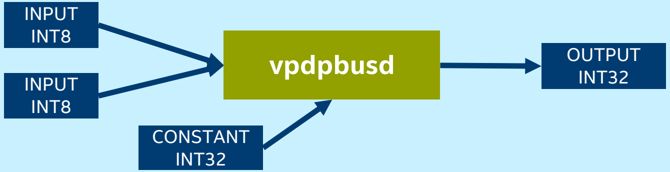
AI/DL Inference Enhancements on INT8 with VNNI

Xeon Scalable AVX-512

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

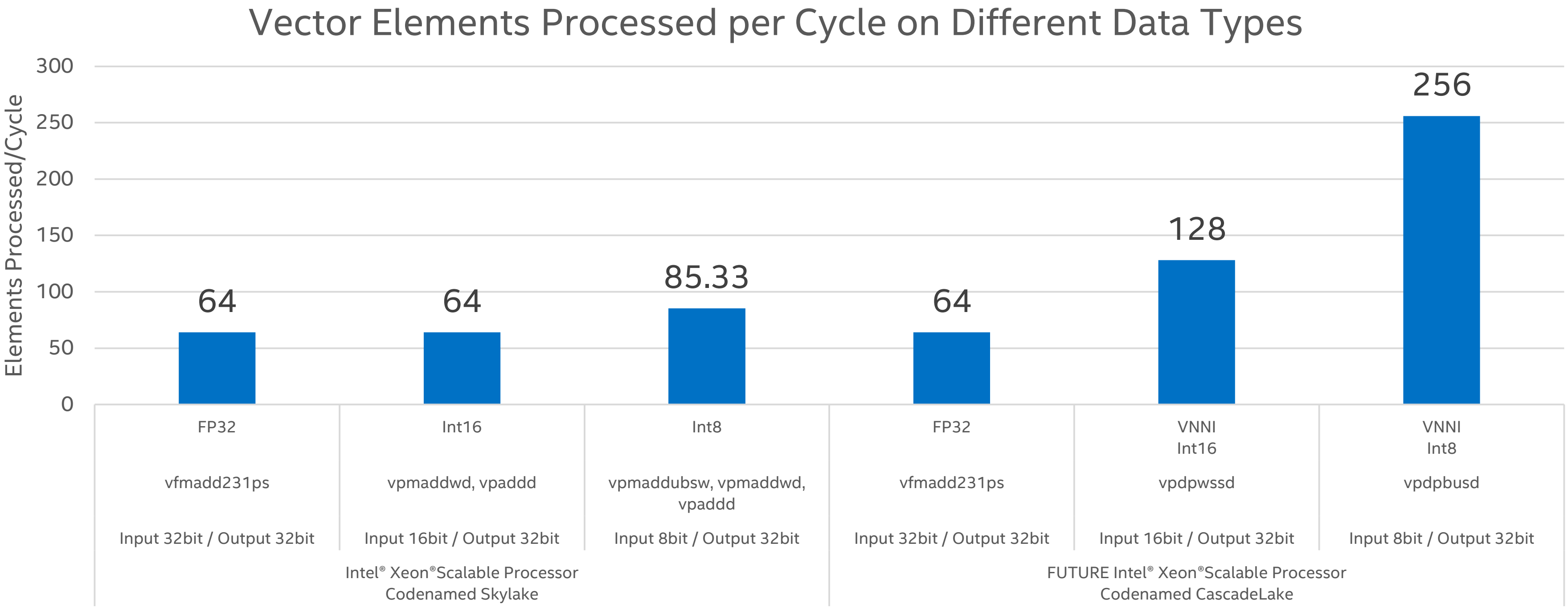
New instructions for accelerating AI on Intel® Xeon® Scalable processors using int8 data

Cascade Lake SP VNNI

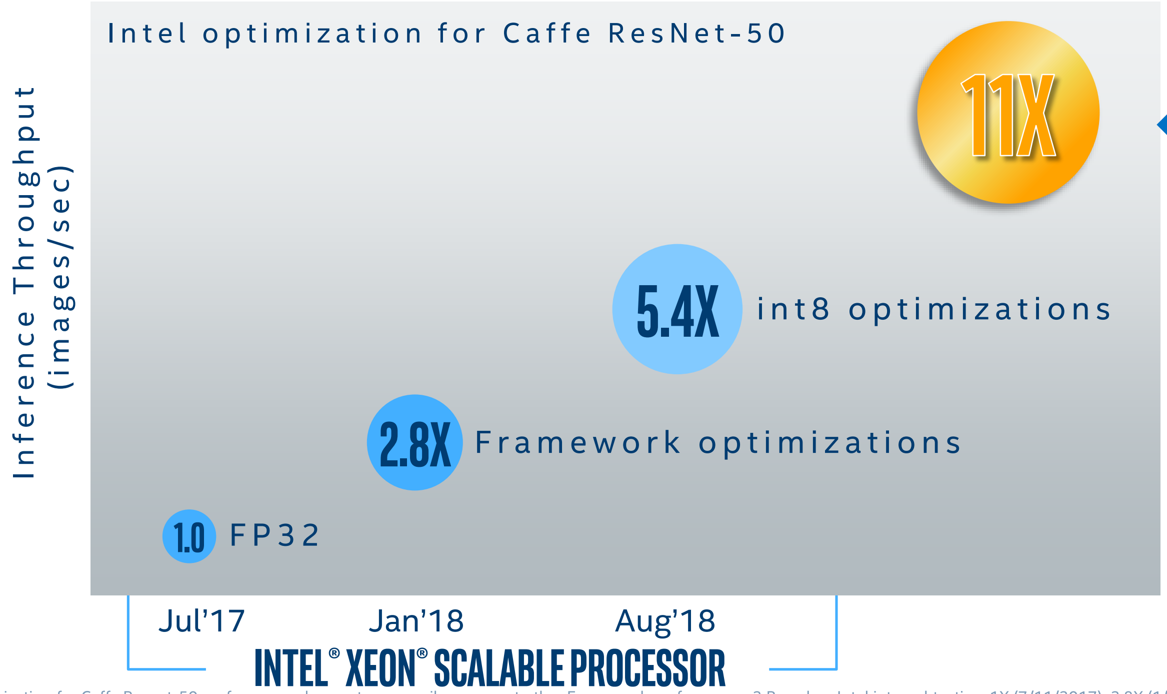


VNNI instruction to accelerate INT8 convolutions: **vpdpbusd**

VNNI Per Core Throughput



Inference Throughput with VNNI



Estimated Throughput on Cascade Lake with VNNI

Framework & Library support

Caffe mxnet TensorFlow

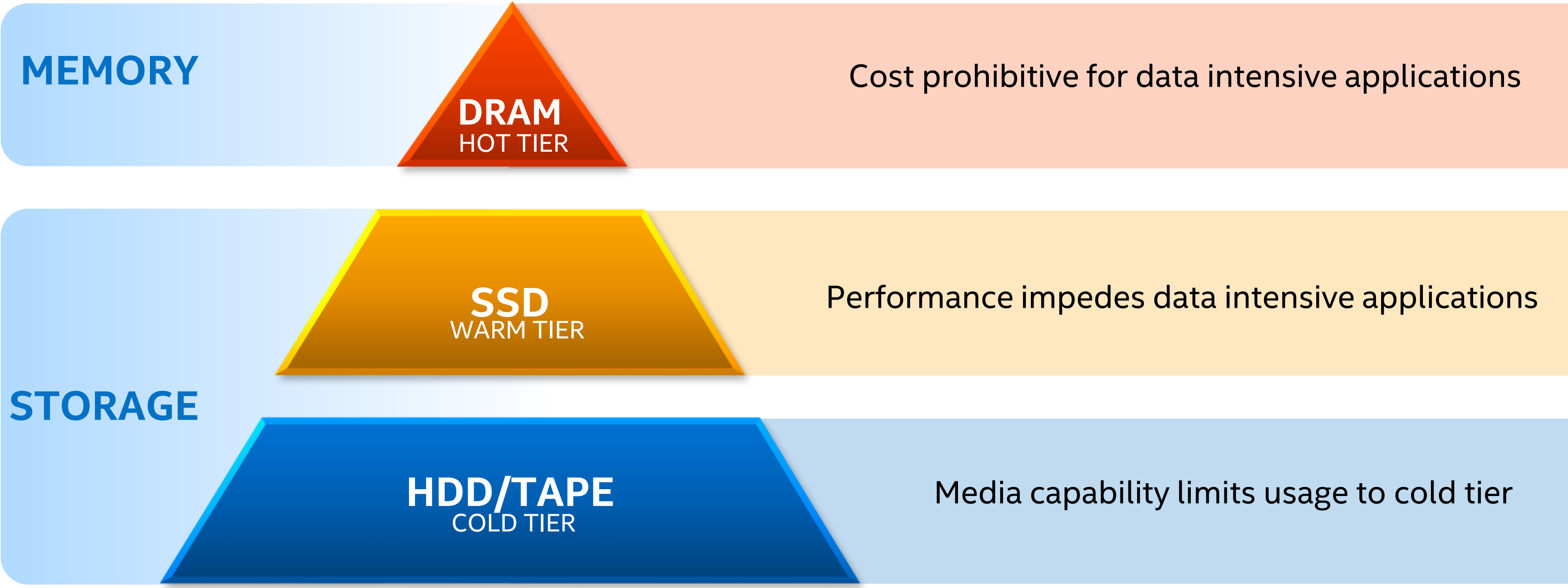
intel MKL-DNN

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 Café 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 publically 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

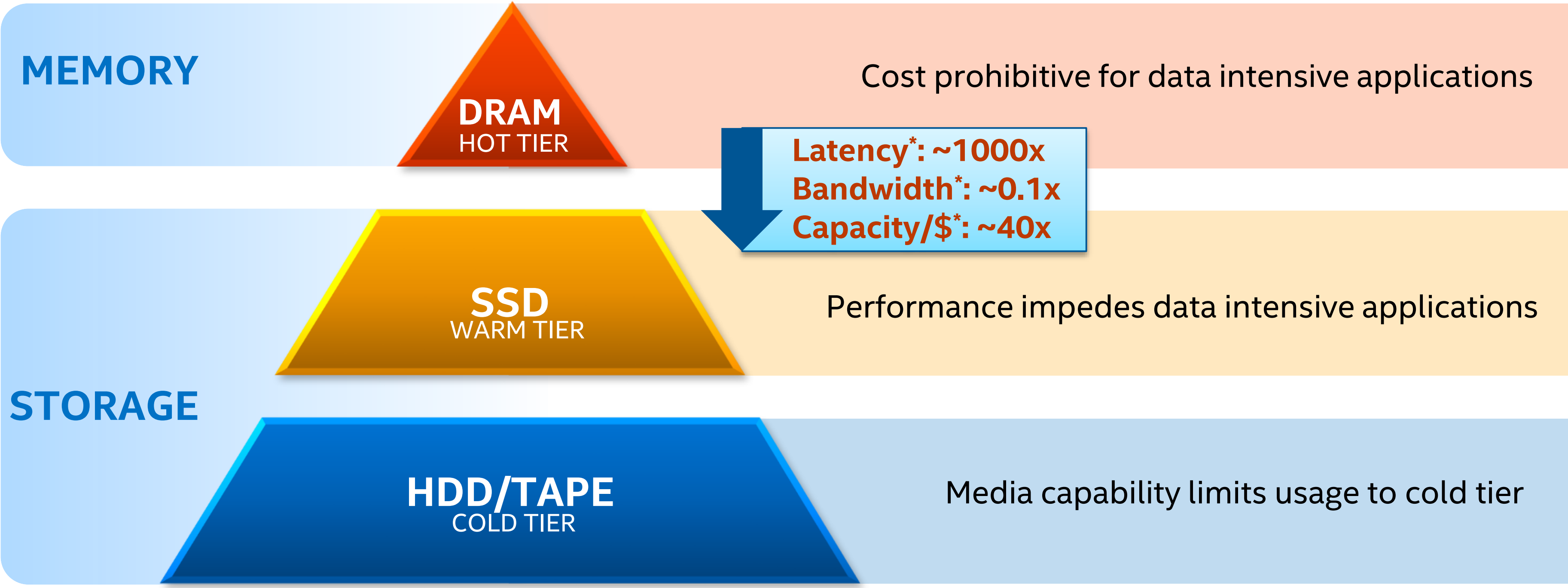
Growing Gap Between Memory Hierarchy

Limitations to traditional architecture impede unified data management



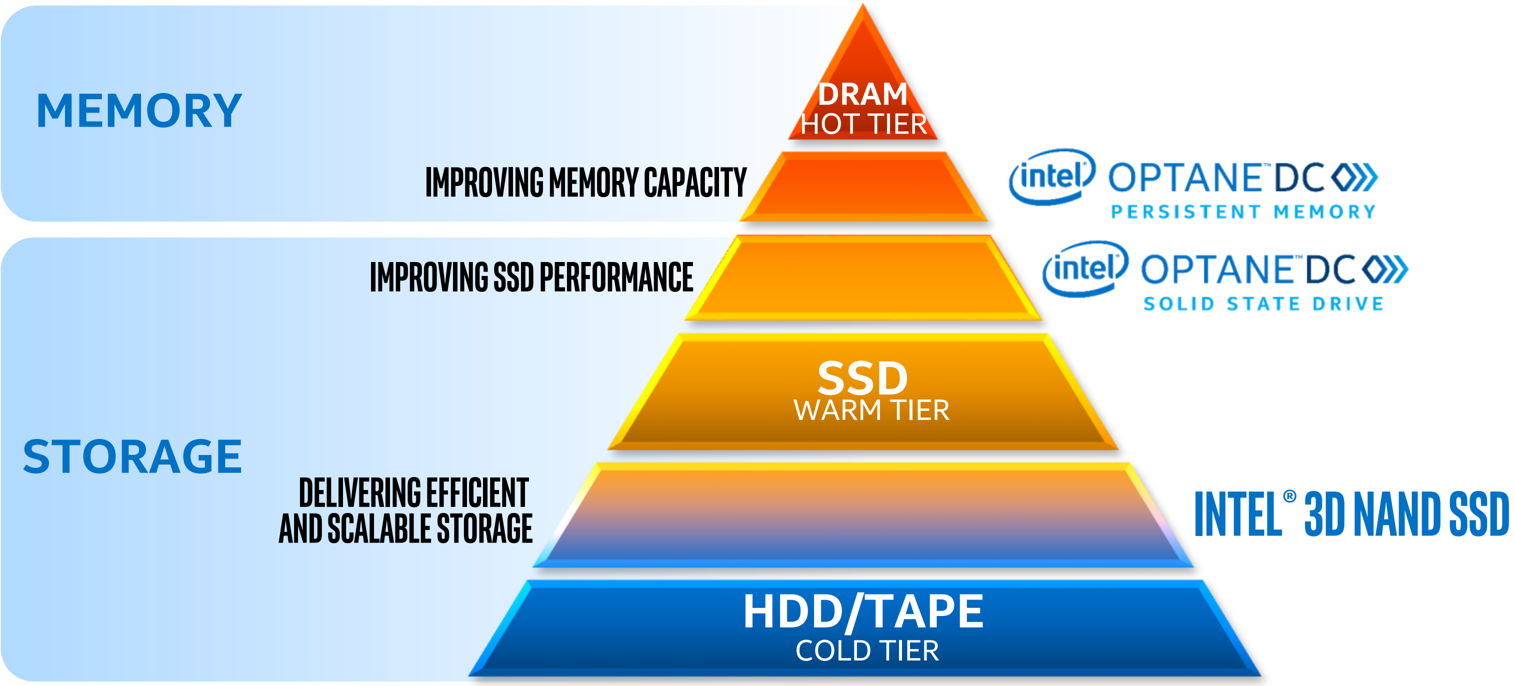
Growing Gap Between Memory Hierarchy

Limitations to traditional architecture impede unified data management



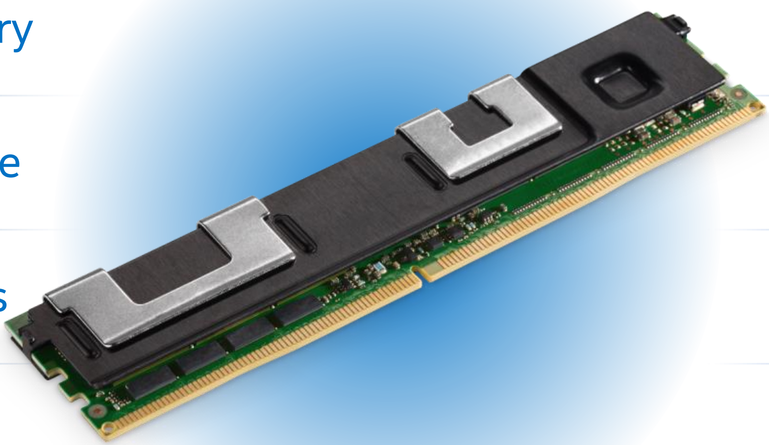
* Actual performance and price may vary

Intel Innovations Address These Gaps



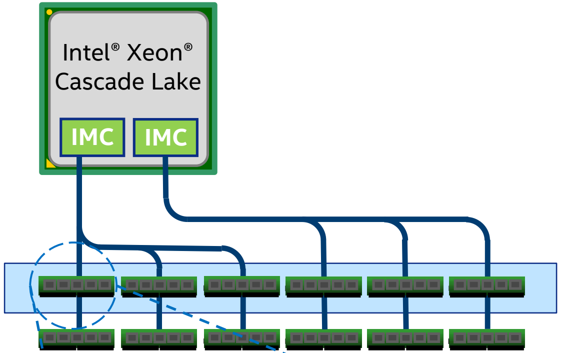
intel OPTANE™ DC PERSISTENT MEMORY

Big and Affordable Memory		128, 256, 512GB
High Performance Storage		DDR4 Pin Compatible
Direct Load/Store Access		Hardware Encryption
Native Persistence		High Reliability



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

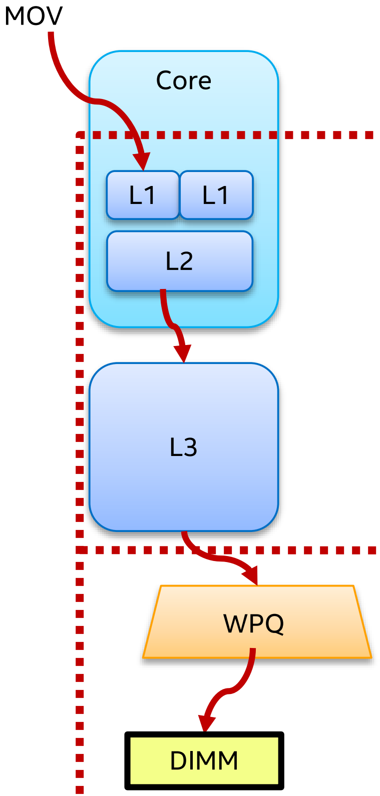
Intel® Optane™ DC Persistent Memory Hardware Interface



Intel® Optane™ DC Persistent Memory Module

- 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

Hardware Interface: Persistence Domain



CPU CACHES

CLWB + fence
-or-
CLFLUSHOPT + fence
-or-
CLFLUSH
-or-
NT stores + fence
-or-
WBINVD (kernel only)

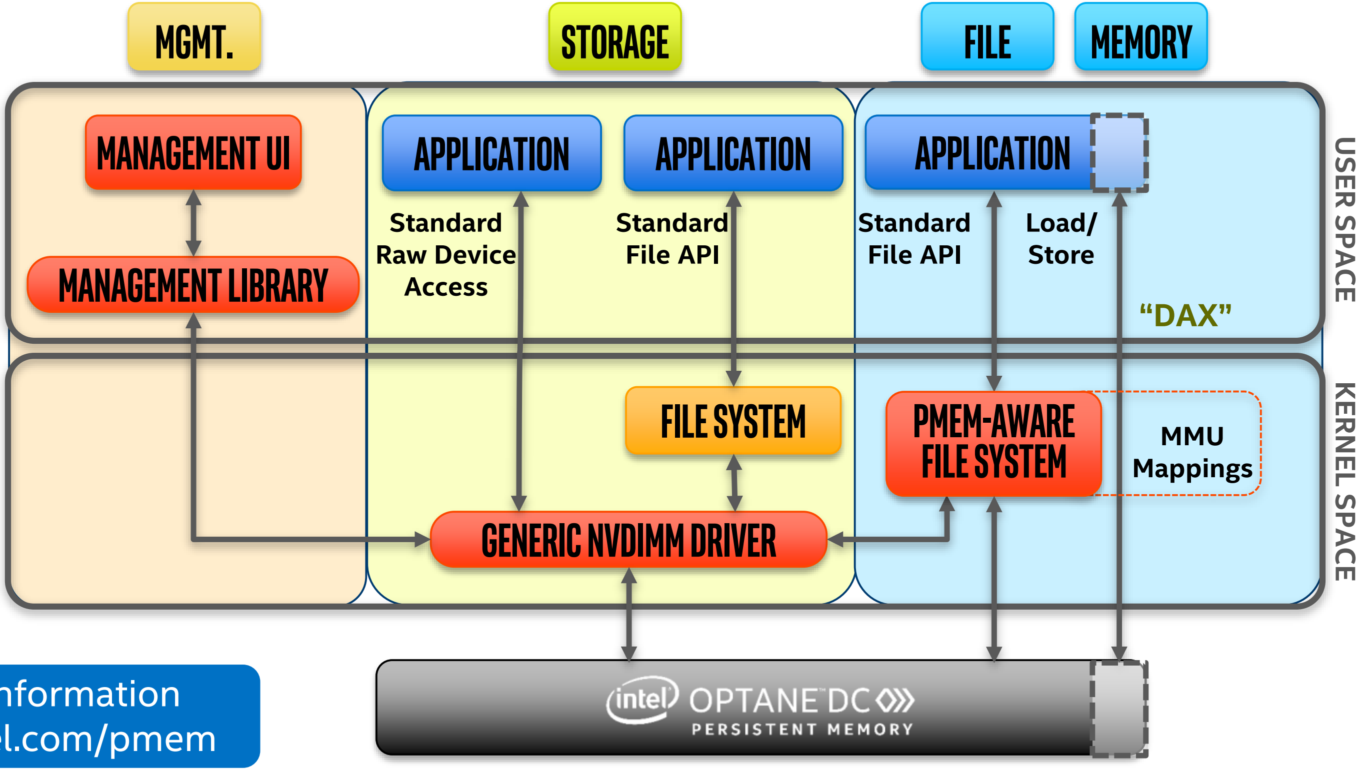
Custom
Power fail protected domain
indicated by ACPI property:
CPU Cache Hierarchy

WPQ

ADR
-or-
WPQ Flush (kernel only)

Minimum Required
Power fail protected domain:
Memory subsystem

The SNIA NVM Programming Model



For more information
software.intel.com/pmem

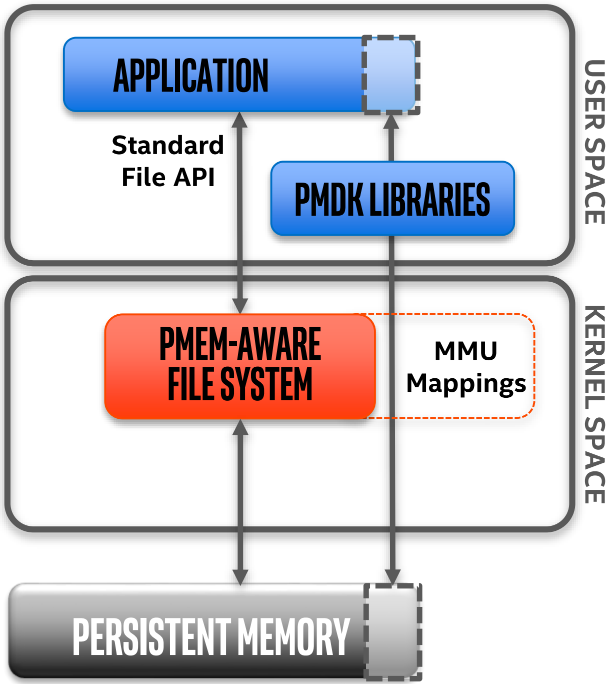
The Persistent Memory Development Kit - pmdk

PMDK is a collection of libraries

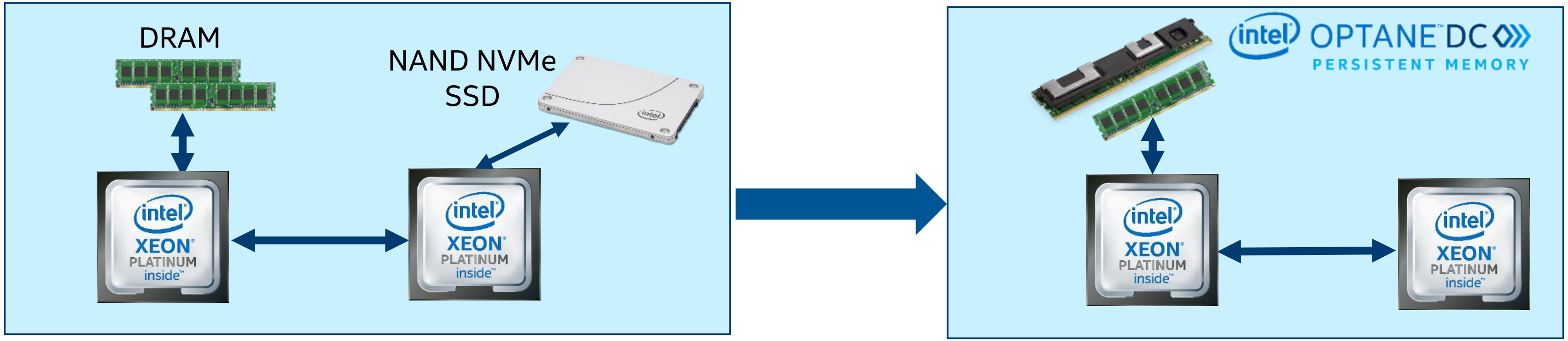
- Developers pull only what they need
  - Low level programming support
  - Transaction APIs
- Fully validated
- Performance tuned

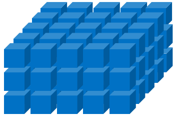
Open source & product neutral

software.intel.com/pmem



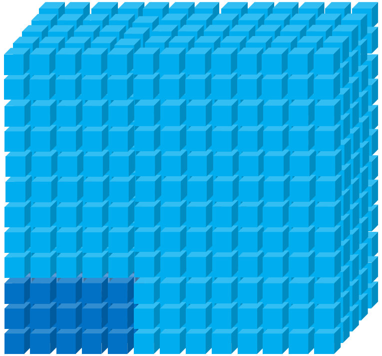
Usage Example: High Performance Storage





9X MORE READ TRANSACTIONS (OPS/SEC)

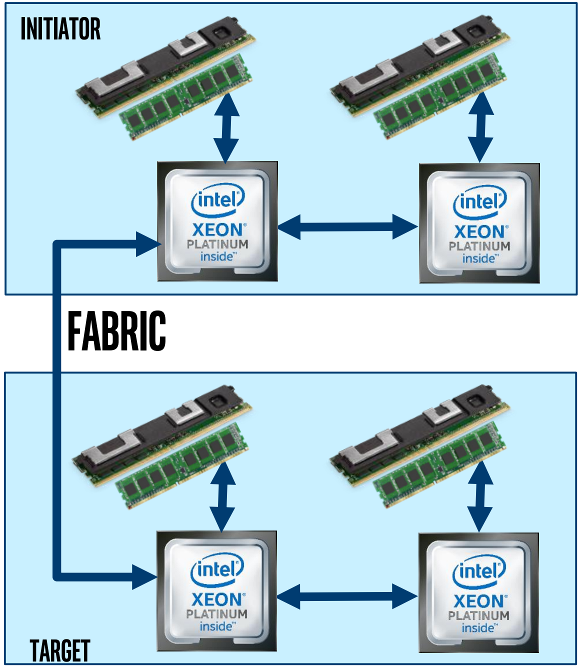
11X MORE USERS PER SYSTEM

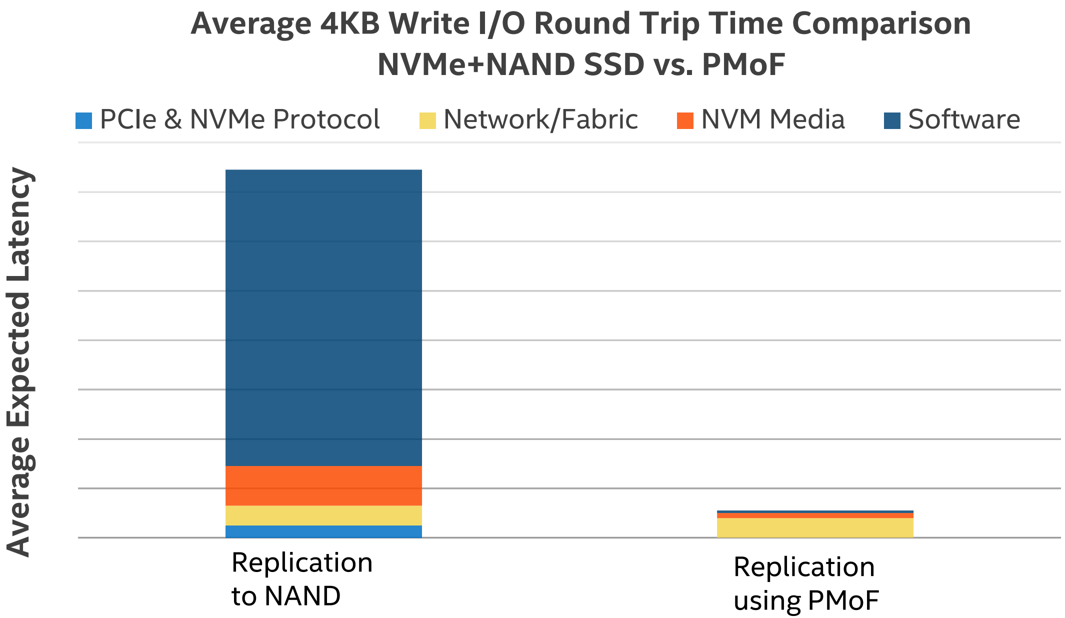


VS. COMPARABLE SERVER SYSTEM WITH DRAM AND NAND NVMe DRIVES WHEN USING APACHE* CASSANDRA-4.0

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





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

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

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

WRAP UP

Future Intel® Xeon® Scalable Processor (Codename: Cascade Lake-SP)

Software Libraries and Optimizations

AI/DL Enhancement through VNNI

intel OPTANE™ DC PERSISTENT MEMORY

Side-Channel Analysis Mitigations

Process Tuning, Frequency Boost, Targeted Performance Improvements

Intel® Xeon® Scalable Platform

Further Accelerating Data Center Innovations

intel®

Config Details for Skylake Inference throughput (March 2018)

Config 1

Framework	caffe branch: master version: f6d01efbe93f70726ea3796a4b89c612365a6341	caffe branch: master version: f6d01efbe93f70726ea3796a4b89c612365a6341	caffe branch: master version: f6d01efbe93f70726ea3796a4b89c612365a6341	caffe branch: master version: f6d01efbe93f70726ea3796a4b89c612365a6341	caffe branch: master version: f6d01efbe93f70726ea3796a4b89c612365a6341	caffe branch: master version: f6d01efbe93f70726ea3796a4b89c612365a6341	caffe branch: master version: f6d01efbe93f70726ea3796a4b89c612365a6341	caffe branch: master version: f6d01efbe93f70726ea3796a4b89c612365a6341	neon branch: (HEAD version: f43cfa2e26f9c84b0f42fcda50b6b83104623223	neon branch: master version: f43cfa2e26f9c84b0f42fcda50b6b83104623223
Platform	SKX_8180	SKX_8180	SKX_8180	SKX_8180	SKX_8180	SKX_8180	SKX_8180	SKX_8180	SKX_8180	SKX_8180
Sockets	2	2	2	2	2	2	2	2	2	2
Processor	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores
BIOS	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215
Enabled Cores	56	56	56	56	56	56	56	56	56	56
Slots	12	12	12	12	12	12	12	12	12	12
Total Memory	376.46GB	376.28GB	376.28GB	376.46GB	376.28GB	376.46GB	376.46GB	376.28GB	376.28GB	376.46GB
Memory Configuration	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz
Memory Comments	Micron	Micron	Micron	Micron	Micron	Micron	Micron	Micron	Micron	Micron
Disks	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB
OS	CentOS Linux-7.3.1611-Core	CentOS Linux-7.3.1611-Core	CentOS Linux-7.3.1611-Core	CentOS Linux-7.3.1611-Core	CentOS Linux-7.3.1611-Core	CentOS Linux-7.3.1611-Core	CentOS Linux-7.3.1611-Core	CentOS Linux-7.3.1611-Core	Ubuntu-14.04-trusty	Ubuntu-14.04-trusty
Hyper Threading	ON	ON	ON	ON	ON	ON	ON	ON	ON	ON
Turbo	ON	ON	ON	ON	ON	ON	ON	ON	ON	ON
Topology	resnet_50_v1	resnet_50_v1	vgg16	vgg16	vgg16	inception_v3	inception_v3	inception_v3	resnet_50_v2	resnet_50_v2
Batchsize	1	64	1	64	128	1	64	128	1	64
Dataset	NoDataLayer	NoDataLayer	NoDataLayer	NoDataLayer	NoDataLayer	NoDataLayer	NoDataLayer	NoDataLayer	NoDataLayer	NoDataLayer
Engine	MKLDNN version: ae00102be506ed0fe2099c6557df2aa88ad57ec1	MKLDNN version: ae00102be506ed0fe2099c6557df2aa88ad57ec1	MKLDNN version: ae00102be506ed0fe2099c6557df2aa88ad57ec1	MKLDNN version: ae00102be506ed0fe2099c6557df2aa88ad57ec1	MKLDNN version: ae00102be506ed0fe2099c6557df2aa88ad57ec1	MKLDNN version: ae00102be506ed0fe2099c6557df2aa88ad57ec1	MKLDNN version: ae00102be506ed0fe2099c6557df2aa88ad57ec1	MKLDNN version: ae00102be506ed0fe2099c6557df2aa88ad57ec1	MKL2017 version: mklml_lnx_2018.0.1.20171227	MKL2017 version: mklml_lnx_2018.0.1.20171227
IP	172.18.0.2	172.18.0.2	172.18.0.2	172.17.0.2	172.18.0.2	172.18.0.2	172.17.0.2	172.18.0.2	172.17.0.3	172.17.0.2
Kernel version	4.4.0-109-generic	3.10.0-693.11.6.el7.x86_64	3.10.0-693.11.6.el7.x86_64	4.4.0-109-generic	3.10.0-693.11.6.el7.x86_64	4.4.0-109-generic	4.4.0-109-generic	3.10.0-693.11.6.el7.x86_64	3.10.0-693.11.6.el7.x86_64	4.4.0-109-generic

Config Details for Skylake Inference throughput (April 2018)

Config 2

Framework	tensorflow branch: master version: 024aecf414941e11eb643e29ceed3e1c47a115ad	tensorflow branch: master version: 024aecf414941e11eb643e29ceed3e1c47a115ad	tensorflow branch: master version: 024aecf414941e11eb643e29ceed3e1c47a115ad	tensorflow branch: master version: 024aecf414941e11eb643e29ceed3e1c47a115ad	tensorflow branch: master version: 024aecf414941e11eb643e29ceed3e1c47a115ad	tensorflow branch: master version: 024aecf414941e11eb643e29ceed3e1c47a115ad	tensorflow branch: master version: 024aecf414941e11eb643e29ceed3e1c47a115ad	tensorflow branch: master version: 024aecf414941e11eb643e29ceed3e1c47a115ad	tensorflow branch: master version: 024aecf414941e11eb643e29ceed3e1c47a115ad	mxnet branch: master version: fdb5664900682c8173a50826acec8b1111d9cd6f	mxnet branch: master version: fdb5664900682c8173a50826acec8b1111d9cd6f	mxnet branch: master version: fbbc080d47323dbc23eef4a1453452624cea859b	mxnet branch: master version: 9a0d0028695cbc3553ffadf3537b1003bb0006c4	mxnet branch: master version: fdb5664900682c8173a50826acec8b1111d9cd6f	mxnet branch: master version: fdb5664900682c8173a50826acec8b1111d9cd6f
Sockets	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Processor	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores	Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores
BIOS	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215	SE5C620.86B.00.01.0004.071220170215
Enabled Cores	56	56	56	56	56	56	56	56	56	56	56	56	56	56	56
Slots	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
Total Memory	376.28GB	376.46GB	376.46GB	376.46GB	376.46GB	376.46GB	376.46GB	376.46GB	376.46GB	376.28GB	376.28GB	376.28GB	376.28GB	376.28GB	376.28GB
Memory Configuration	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz	12slots / 32 GB / 2666 MHz
Memory Comments	Micron	Micron	Micron	Micron	Micron	Micron	Micron	Micron	Micron	Micron	Micron	Micron	Micron	Micron	Micron
Disks	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB	sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB
OS	CentOS Linux-7.3.1611-Core	CentOS Linux-7.3.1611-Core	CentOS Linux-7.3.1611-Core	CentOS Linux-7.3.1611-Core	CentOS Linux-7.3.1611-Core	CentOS Linux-7.3.1611-Core	CentOS Linux-7.3.1611-Core	CentOS Linux-7.3.1611-Core	CentOS Linux-7.3.1611-Core	Ubuntu-16.04-xenial	Ubuntu-16.04-xenial	Ubuntu-16.04-xenial	Ubuntu-16.04-xenial	Ubuntu-16.04-xenial	Ubuntu-16.04-xenial
Hyper Threading	ON	ON	ON	ON	ON	ON	ON	ON	ON	ON	ON	ON	ON	ON	ON
Turbo	ON	ON	ON	ON	ON	ON	ON	ON	ON	ON	ON	ON	ON	ON	ON
Topology	resnet_50_v1	resnet_50_v1	resnet_50_v1	vgg16	vgg16	vgg16	inception_v3	inception_v3	inception_v3	resnet_50_v2	vgg16	vgg16	inception_v3	inception_v3	inception_v3
Batchsize	1	64	128	1	64	128	1	64	128	1,64,128	1	64, 128	1	64	128
Instances/ Streams on 2 sockets	8	8	8	8	8	8	8	8	8	1	1	1	1	1	1
Dataset	NoDataLayer	NoDataLayer	NoDataLayer	NoDataLayer	NoDataLayer	NoDataLayer	NoDataLayer	NoDataLayer	NoDataLayer	Imagenet	Imagenet	NoDataLayer	Imagenet	NoDataLayer	Imagenet
Engine	MKLDNN version: e0bfcaa7fcb2b1e1558f5f0676933c1db807a729	MKLDNN version: e0bfcaa7fcb2b1e1558f5f0676933c1db807a729	MKLDNN version: e0bfcaa7fcb2b1e1558f5f0676933c1db807a729	MKLDNN version: e0bfcaa7fcb2b1e1558f5f0676933c1db807a729	MKLDNN version: e0bfcaa7fcb2b1e1558f5f0676933c1db807a729	MKLDNN version: e0bfcaa7fcb2b1e1558f5f0676933c1db807a729	MKLDNN version: e0bfcaa7fcb2b1e1558f5f0676933c1db807a729	MKLDNN version: e0bfcaa7fcb2b1e1558f5f0676933c1db807a729	MKLDNN version: e0bfcaa7fcb2b1e1558f5f0676933c1db807a729	MKLDNN version: f5218ff4fd2d16d13aada2e632afd18f2514fee3	MKLDNN version: f5218ff4fd2d16d13aada2e632afd18f2514fee3	MKLDNN version: 283c4a8c24b4e1dea05fbc7941594a7375babad1	MKLDNN version: f5218ff4fd2d16d13aada2e632afd18f2514fee3	MKLDNN version: f5218ff4fd2d16d13aada2e632afd18f2514fee3	MKLDNN version: f5218ff4fd2d16d13aada2e632afd18f2514fee3
Kernel version	3.10.0-693.11.6.el7.x86_64	4.4.0-109-generic	4.4.0-109-generic	4.4.0-109-generic	4.4.0-109-generic	4.4.0-109-generic	4.4.0-109-generic	4.4.0-109-generic	4.4.0-109-generic	3.10.0-693.11.6.el7.x86_64	3.10.0-693.11.6.el7.x86_64	3.10.0-693.11.6.el7.x86_64	3.10.0-693.11.6.el7.x86_64	3.10.0-693.11.6.el7.x86_64	3.10.0-693.11.6.el7.x86_64

Config 3

Configuration details of Amazon EC2 C5.18xlarge 1 node systems

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

Configuration details of Amazon EC2 C5.18xlarge 1 node systems

Framework Version	mxnet mkldnn : https://github.com/apache/incubator-mxnet/ 4950f6649e329b23a1efdc40aaa25260d47b4195
Topology Version	GNMT: https://github.com/awslabs/sockeye/tree/master/tutorials/wmt
Batch size	GNMT:1 2 8 16 32 64 128
Dataset, version	GNMT: WMT 2017 (http://data.statmt.org/wmt17/translation-task/preprocessed/)
MKLDNN	F5218ff4fd2d16d13aada2e632afd18f2514fee3
MKL	Version: parallel_studio_xe_2018_update1 http://registrationcenterdownload.intel.com/akdlm/irc_nas/tec/12374/parallel_studio_xe_2018_update1_cluster_edition_online.tgz
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 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),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). 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: ae00102be506ed0fe2099c6557df2aa88ad57ec1 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 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),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). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

Configuration Details for Inference Throughput with VNNI

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:464c268e544bae26f9b85a2acb9122c766a4c396 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.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 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 2018.0.20170425. Caffe run with “numactl -l“.

11X inference thoughput improvement with CascadeLake:

Future Intel Xeon Scalable processor (codename Cascade Lake) 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 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.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 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 2018.0.20170425. Caffe run with “numactl -l“.