THE NEW INTEL® XEON® SCALABLE PROCESSOR (FORMERLY SKYLAKE-SP)

Akhilesh Kumar

Intel Corporation, 2017

Authors: Don Soltis, Irma Esmer, Adi Yoaz, Sailesh Kottapalli

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.

INTEL® XEON® SCALABLE PROCESSORS

THE FOUNDATION OF DATA CENTER INNOVATION

WORKLOAD DRIVEN PERFORMANCE

COMPUTE, STORAGE, NETWORK OPTIMIZED

SIMPLE AND EASY TO DEPLOY

ARCHITECTED FOR EFFICIENT, SECURE, AND AGILE DATA CENTER

Agenda

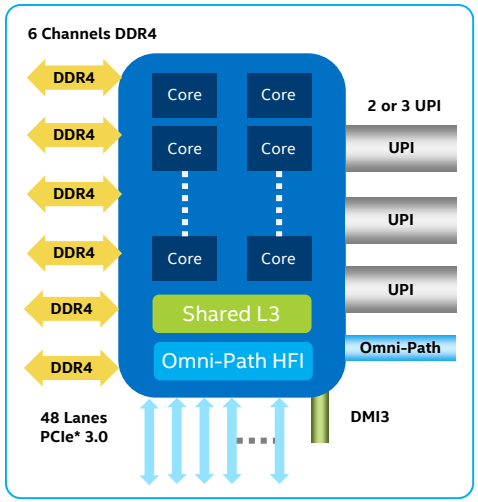
- Intel® Xeon® Scalable Processor Overview
- Processor Architecture Details
  - Core Architecture
  - Interconnect and Cache Architecture
  - Memory Subsystem
  - IO Subsystem
- Performance Benchmarks
- Wrap UP

Intel® Xeon® Scalable Processor Overview

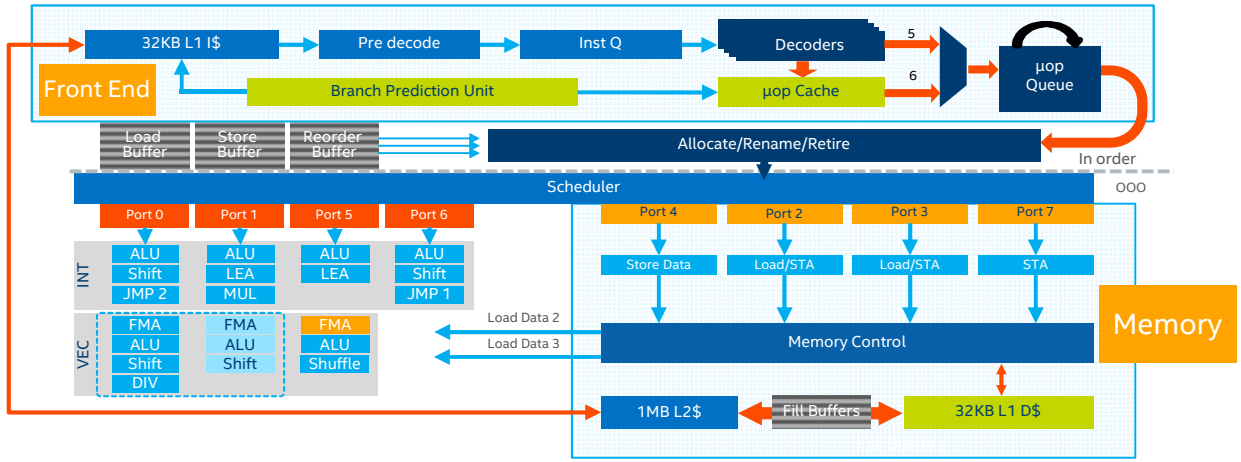
Re-architected from the ground up

- Skylake core microarchitecture with data center specific enhancements
- Intel® AVX-512 with 32 DP flops per cycle per core
- Datacenter optimized cache hierarchy – 1MB L2 per core, non-inclusive L3
- New mesh interconnect architecture
- Enhanced memory subsystem
- Modular IO with integrated devices
- New Intel® Ultra Path Interconnect (Intel® UPI)
- Intel® Speed Shift Technology
- Security & Virtualization enhancements (MBE, PPK, MPX)
- Optional Integrated Intel® Omni-Path Fabric (Intel® OPA)

Features	Intel® Xeon® Processor E5/E7 v4	Intel® Xeon® Scalable Processor
Availability	Q2 2016	**Q3 2017**
Cores Per Socket	Up to 22 or 24	**Up to 28**
Threads Per Socket	Up to 44 or 48 threads	**Up to 56 threads**
Last-level Cache (LLC)	Up to 55MB or 60 MB	**Up to 38.5 MB (non-inclusive)**
QPI/UPI Speed (GT/s)	2x or 3x QPI channels @ 9.6 GT/s	**Up to 3x UPI @ 10.4 GT/s**
PCIe* Lanes/ Controllers	40 / 10 / PCIe* 3.0 (2.5, 5, 8 GT/s)	**48 / 12 / PCIe 3.0 (2.5, 5, 8 GT/s)**
Memory Population	4 channels of up to 3 RDIMMs, LRDIMMs, or 3DS LRDIMMs	**6 channels of up to 2 RDIMMs, LRDIMMs, or 3DS LRDIMMs**
Max Memory Speed	Up to 2400	**Up to 2666**
TDP (W)	55W-145W	**70W-205W**



Core Microarchitecture Enhancements



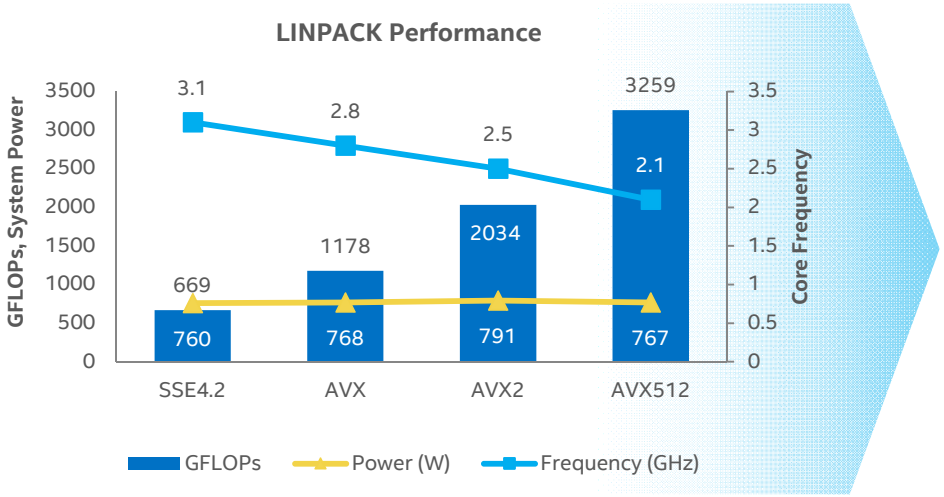
	Broadwell uArch	Skylake uArch
Out-of-order Window	192	**224**
In-flight Loads + Stores	72 + 42	72 + **56**
Scheduler Entries	60	**97**
Registers – Integer + FP	168 + 168	**180** + 168
Allocation Queue	56	**64/thread**
L1D BW (B/Cyc) – Load + Store	64 + 32	**128 + 64**
L2 Unified TLB	4K+2M: 1024	4K+2M: **1536** **1G: 16**

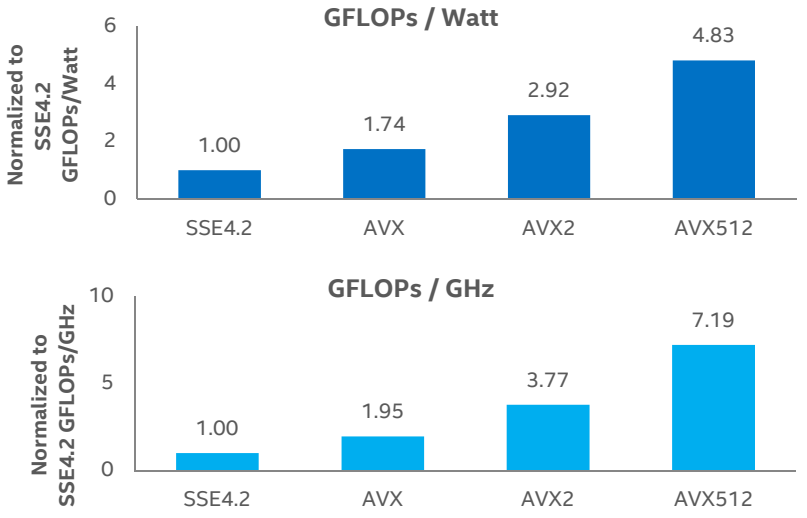
- Larger and improved branch predictor, higher throughput decoder, larger window to extract ILP
- Improved scheduler and execution engine, improved throughput and latency of divide/sqrt
- More load/store bandwidth, deeper load/store buffers, improved prefetcher
- **Data center specific enhancements: Intel® AVX-512 with 2 FMAs per core, larger 1MB L2 cache**

CORE MICROARCHITECTURE ENHANCED FOR DATA CENTER SPECIFIC APPLICATIONS

Intel® Advanced Vector Extensions 512 (Intel® AVX-512)

512-bit wide vectors, 32 operand registers, 8 64b mask registers, Embedded broadcast & rounding



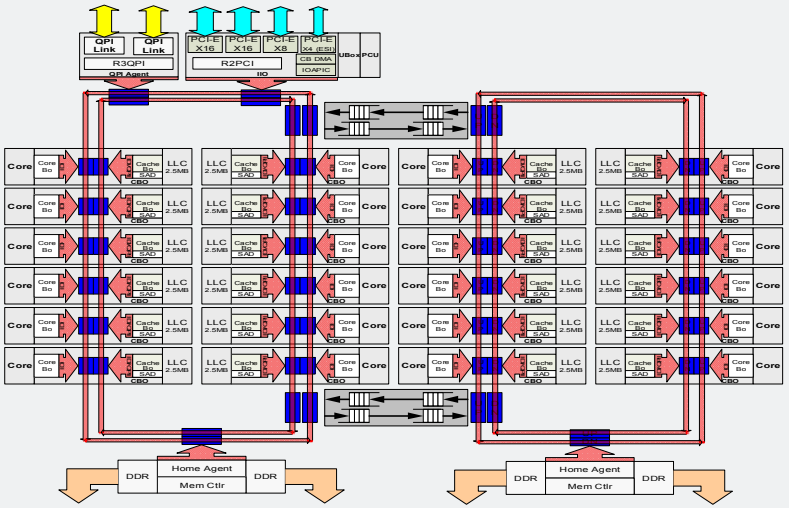


INTEL® AVX-512 DELIVERS SIGNIFICANT PERFORMANCE AND EFFICIENCY GAINS

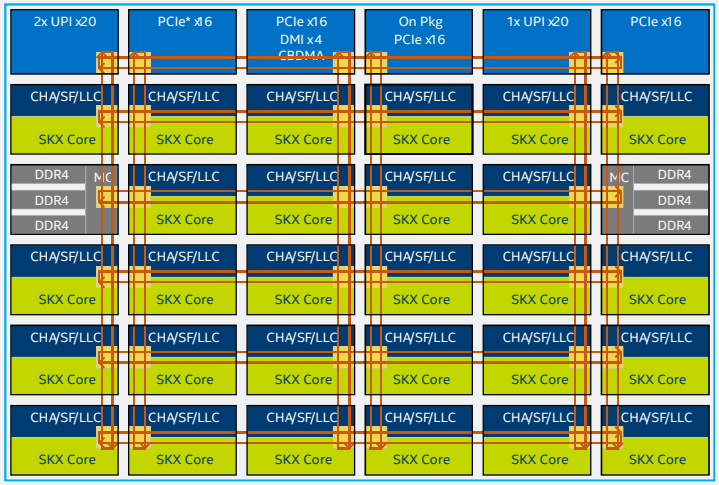
Source as of June 2017: Intel internal measurements. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations 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. Configuration Summary: 1-Node, 2 x Intel® Xeon® Platinum 8180 Processor on Purley-EP (Lewisburg) (S2600WF) with 384 GB (12x32GB DDR4-2666) Total Memory, Intel S3610 800GB SSD, BIOS: SE5C620.86B.01.00.0471.040720170924, 04/07/2017, RHEL Kernel: 3.10.0-514.16.1.el7.x86_64 x86_64, Benchmark: Intel® Optimized MP LINPACK

New Intel® Mesh Interconnect Architecture

2016: Intel® Xeon® Processor E7 v4, 14nm (Broadwell EX 24-Core Die)



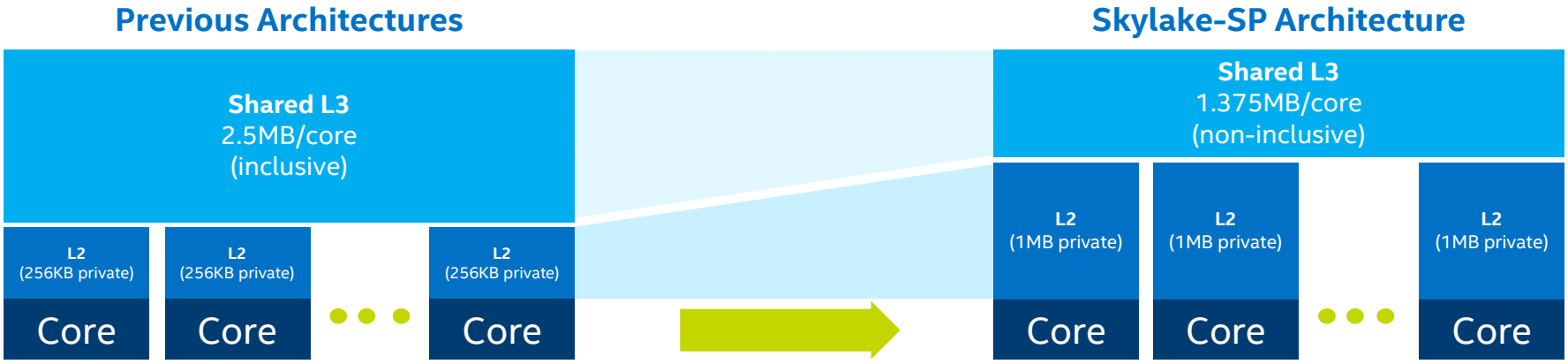
2017: Intel® Xeon® Scalable Processor, 14nm (Skylake-SP 28-Core Die)



CHA– Caching and Home Agent; SF– Snoop Filter; LLC– Last Level Cache
SKX Core– Skylake Server Core; UPI – Intel® UltraPath Interconnect

Intel® Mesh Improves Scalability with Higher Bandwidth and Reduced Latencies

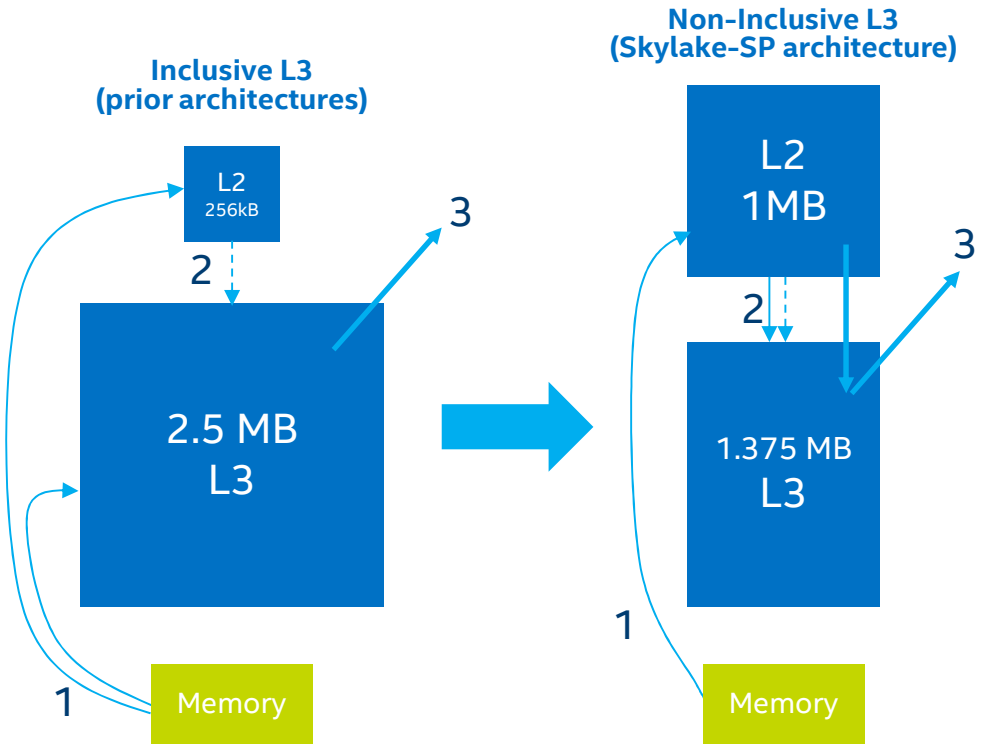
Re-Architected L2 & L3 Cache Hierarchy



- On-chip cache balance shifted from shared-distributed (prior architectures) to private-local (Skylake architecture):
  - Shared-distributed ➔ shared-distributed L3 is primary cache
  - Private-local ➔ private L2 becomes primary cache with shared L3 used as overflow cache
- Shared L3 changed from inclusive to non-inclusive:
  - Inclusive (prior architectures) ➔ L3 has copies of all lines in L2
  - Non-inclusive (Skylake architecture) ➔ lines in L2 ***may not*** exist in L3

SKYLAKE-SP CACHE HIERARCHY ARCHITECTED SPECIFICALLY FOR DATACENTER USE CASE

Inclusive vs Non-Inclusive L3

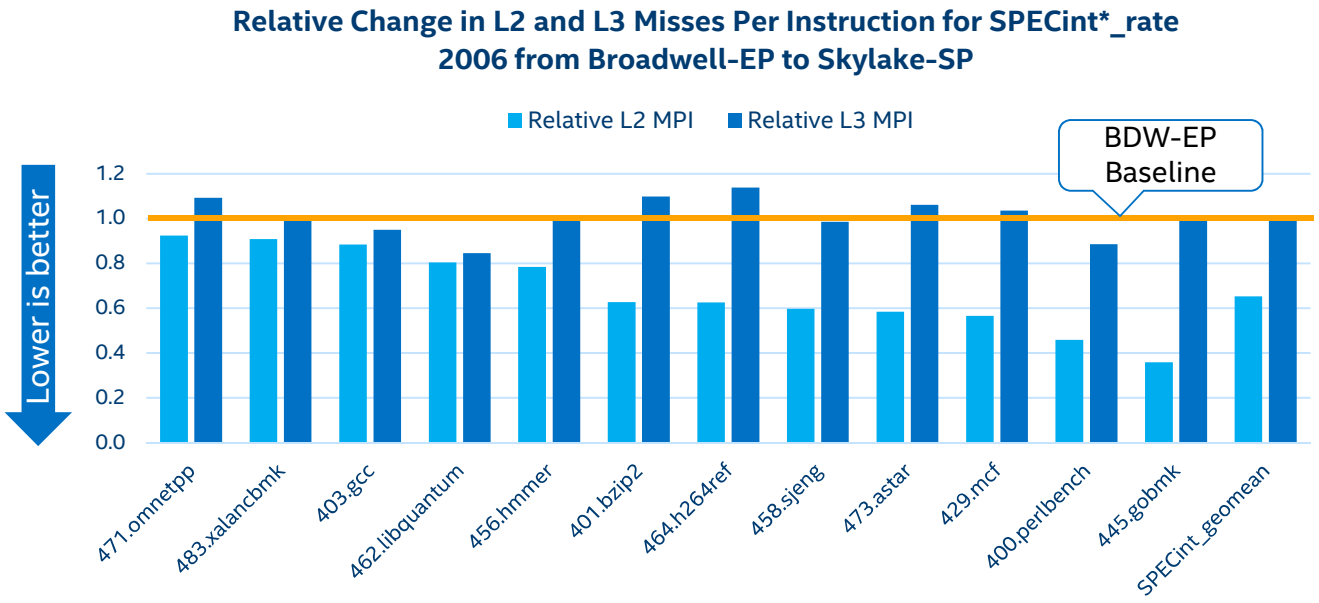


1. Memory reads fill directly to the L2, no longer to both the L2 and L3
2. When a L2 line needs to be evicted, both modified and unmodified lines are written back
3. Data shared across cores are copied into the L3 for servicing future L2 misses

Cache hierarchy architected and optimized for data center use cases:

- Virtualized use cases get larger private L2 cache free from interference
- Multithreaded workloads can operate on larger data per thread (due to increased L2 size) and reduce interconnect and L3 activity

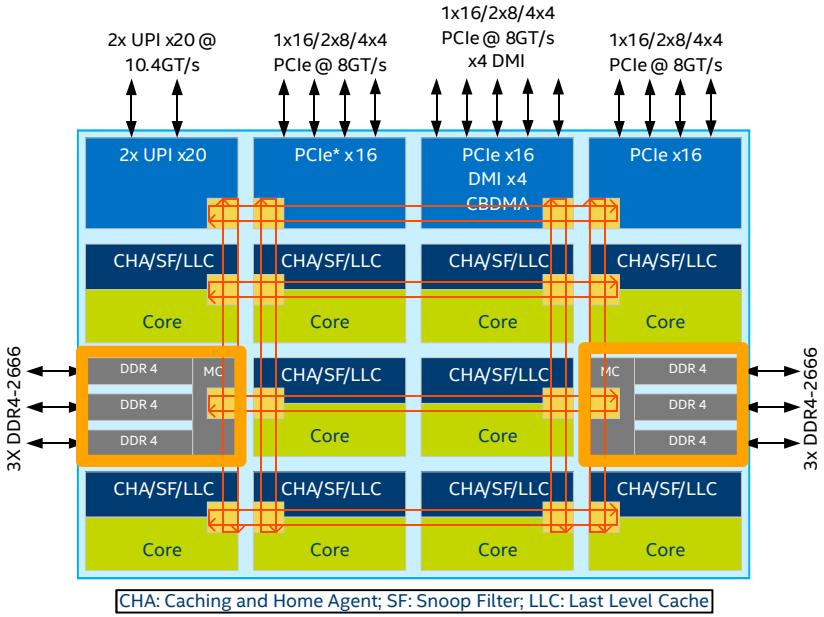
Cache Performance



Skylake-SP cache hierarchy significantly reduces L2 misses without increasing L3 misses compared to Broadwell-EP

Source as of June 2017: Intel internal measurements on platform with Xeon Platinum 8180, Turbo enabled, UPI=10.4, SNC1, 6x32GB DDR4-2666 per CPU, 1 DPC, and platform with E5-2699 v4, Turbo enabled, 4x32GB DDR4-2400, RHEL 7.0. 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.

Memory Subsystem

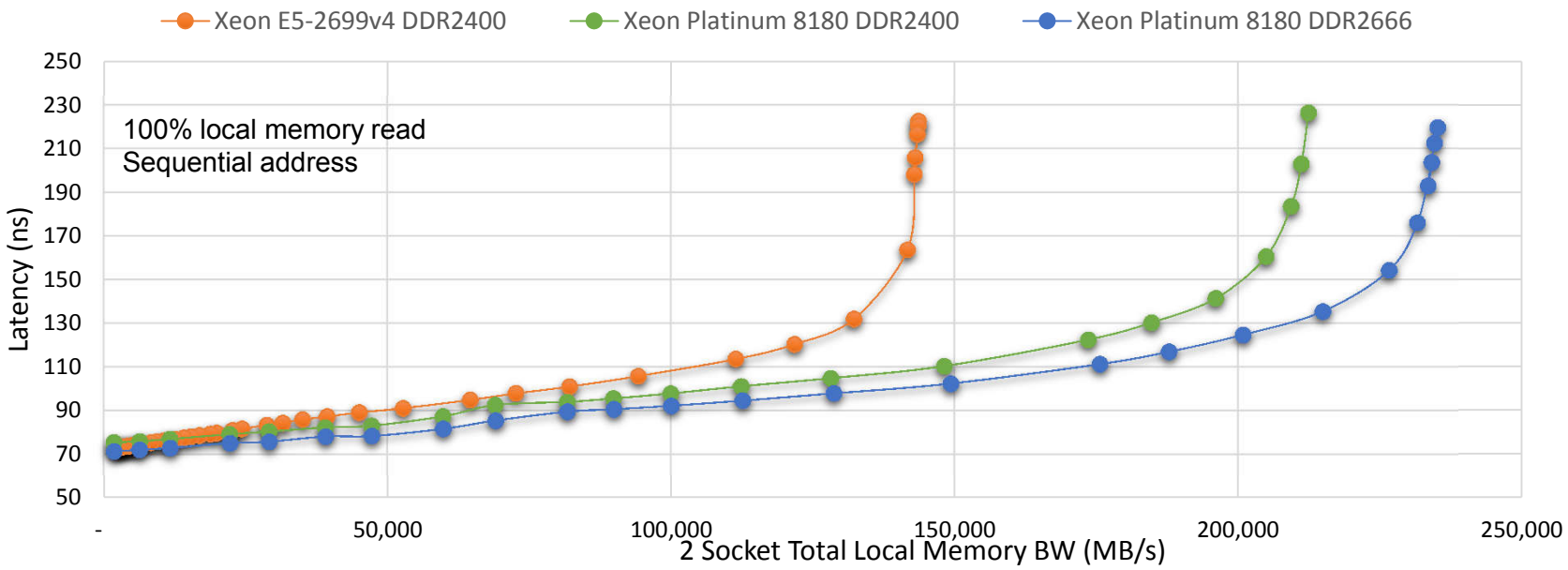


- 2 Memory Controllers, 3 channels each ➔ total of 6 memory channels
  - DDR4 up to 2666, 2 DIMMs per channel
  - Support for RDIMM, LRDIMM, and 3DS-LRDIMM
  - 1.5TB Max Memory Capacity per Socket (2 DPC with 128GB DIMMs)
  - >60% increase in Memory BW per Socket compared to Intel® Xeon® processor E5 v4
- Consistent and low access latency to all memory attached to a socket from any core
- Several optimizations for lower latency and higher bandwidth efficiency
- New memory failure detection and recovery with Adaptive Double Device Data Correction (ADDDC)

SIGNIFICANT MEMORY BANDWIDTH AND CAPACITY IMPROVEMENTS OVER PRIOR GENERATION

Memory Performance

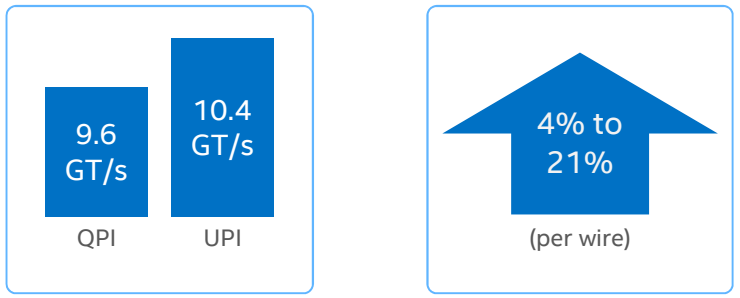
Bandwidth-Latency Profile

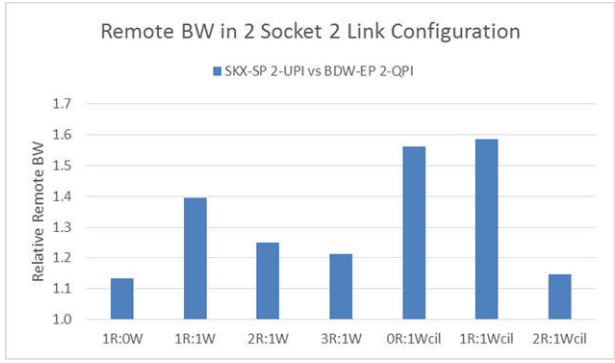


Source as of June 2017: Intel internal measurements on platform with Xeon Platinum 8180, Turbo enabled, UPI=10.4, SNC1, 6x32GB DDR4-2400/2666 per CPU, 1 DPC, and platform with E5-2699 v4, Turbo enabled, 4x32GB DDR4-2400, RHEL 7.0. Software and workloads used in performance tests may have been optimized 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/performance

New Intel® Ultra Path Interconnect (Intel® UPI)

- Intel® Ultra Path Interconnect (Intel® UPI), replacing Intel® QPI
- Faster link with improved bandwidth for a balanced system design
  - Improved messaging efficiency per packet
- 3 UPI option for 2 socket – additional bandwidth for non-NUMA high bandwidth use cases





INTEL® UPI ENABLES SYSTEM SCALABILITY WITH HIGHER INTER-SOCKET BANDWIDTH

Source as of June 2017: Intel internal measurements on platform with Xeon Platinum 8180, Turbo enabled, UPI=10.4, 6x32GB DDR4-2666, 1 DPC, and platform with E5-2699 v4, Turbo enabled, 4x32GB DDR4-2400, RHEL 7.0. 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.

Processor Integrated I/O

3 independent pipelines of x16 PCIe* Gen3

- Each x16 can be bifurcated into 2x8, 1x8+2x4, or 4x4 root ports
- New traffic controller pipeline improves over prior design
- Contains VTd (Virtualization Technology for Directed IO) engine to support address translation and interrupt remapping

Non-Transparent Bridging (NTB)

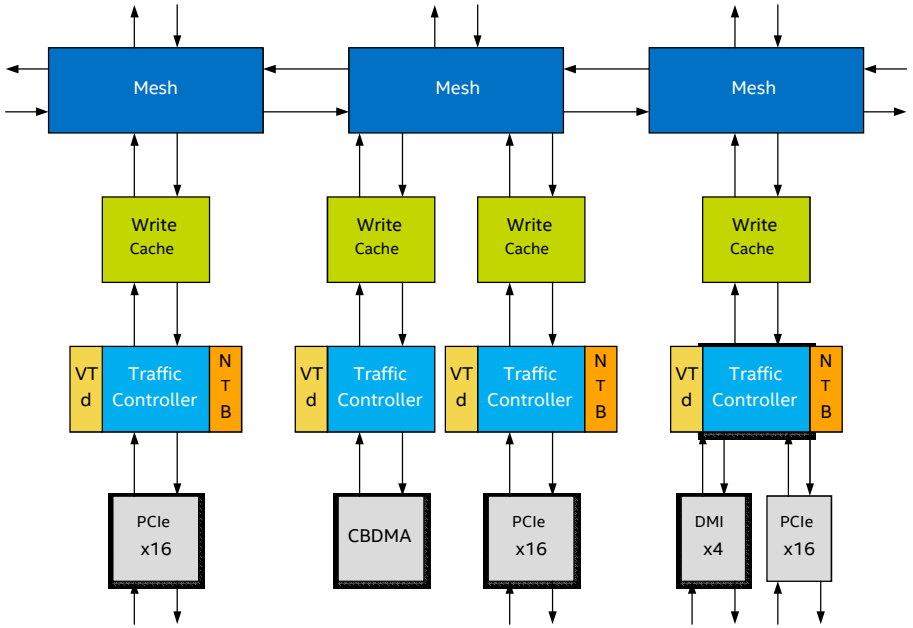
- One NTB per x16 PCIe, which can be configured as 1x8 or 1x4 NTB

Intel® QuickData Technology (CBDMA)

- 2x bandwidth on Mem-Mem copy
- Supports MMIO-Mem copy

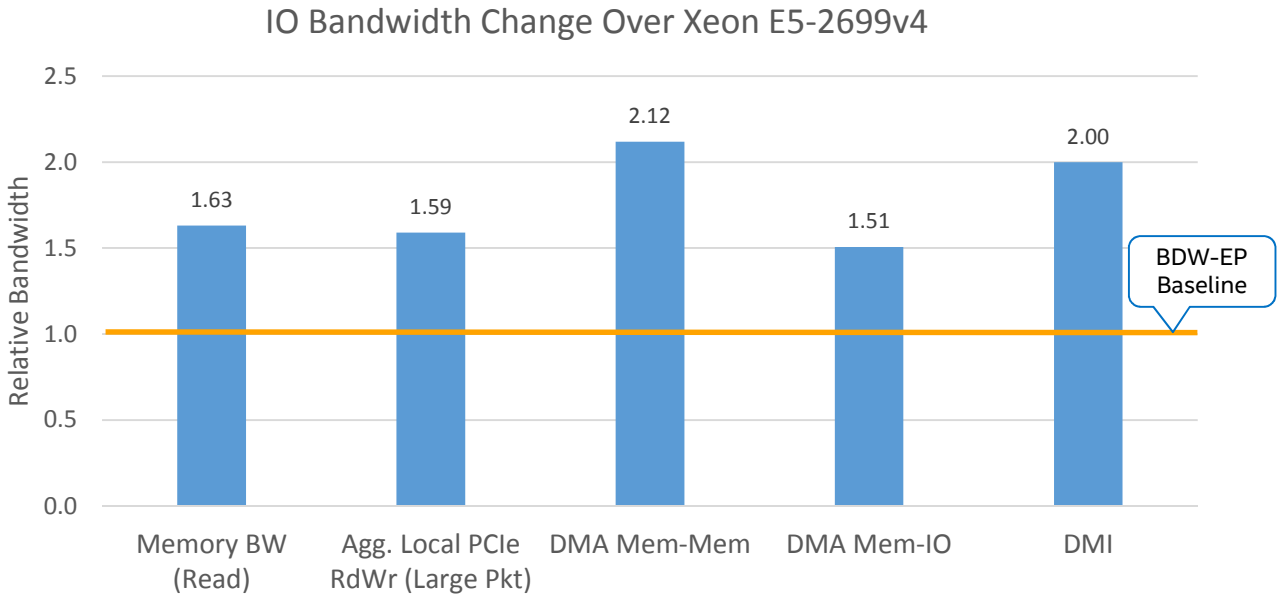
Intel® Volume Management Device (VMD)

- One VMD domain per x16 PCIe



MODULAR IO DESIGN WITH IMPROVED FEATURE SET FOR CONVERGED DATA CENTER

I/O Performance



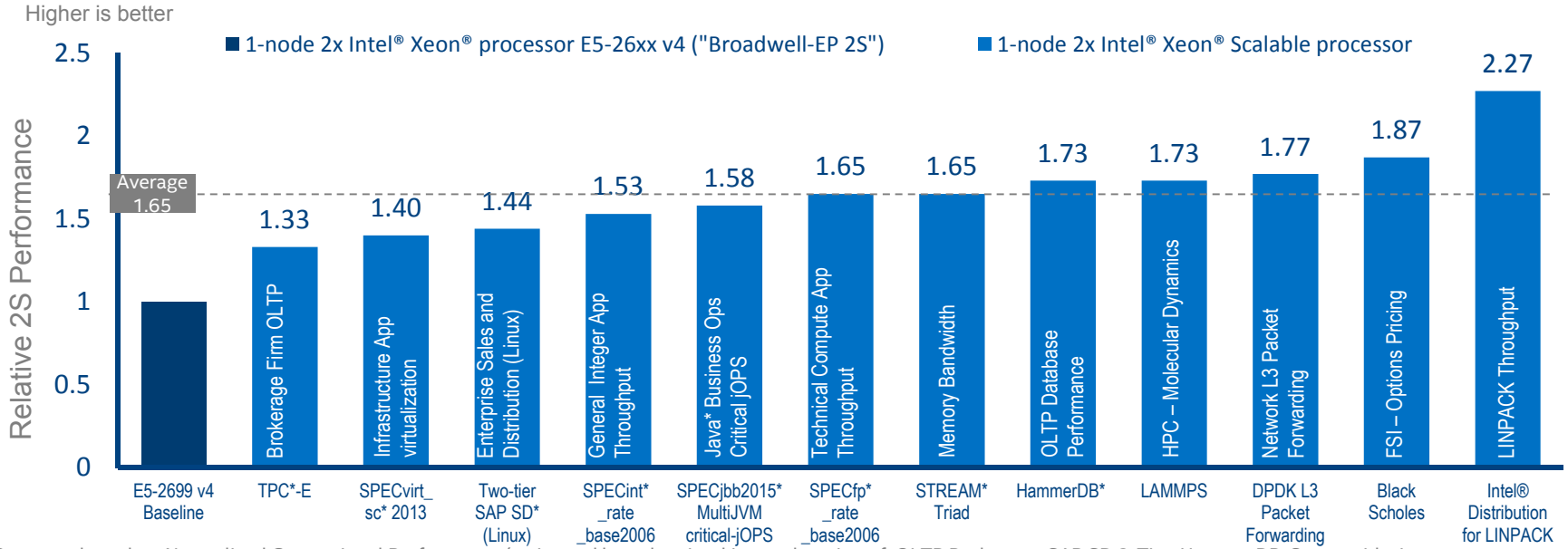
>50% aggregate IO bandwidth improvement in line with memory bandwidth increase for a balanced system performance

DMI: Direct Media Interface, link between CPU and PCH

Source as of June 2017: Intel internal measurements on platform with Xeon Platinum 8180, Turbo enabled, UPI=10.4, SNC1, 6x32GB DDR4-2400/2666 per CPU, 1 DPC, and platform with E5-2699 v4, Turbo enabled, 4x32GB DDR4-2400, RHEL 7.0. 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.

PERFORMANCE BENCHMARKS

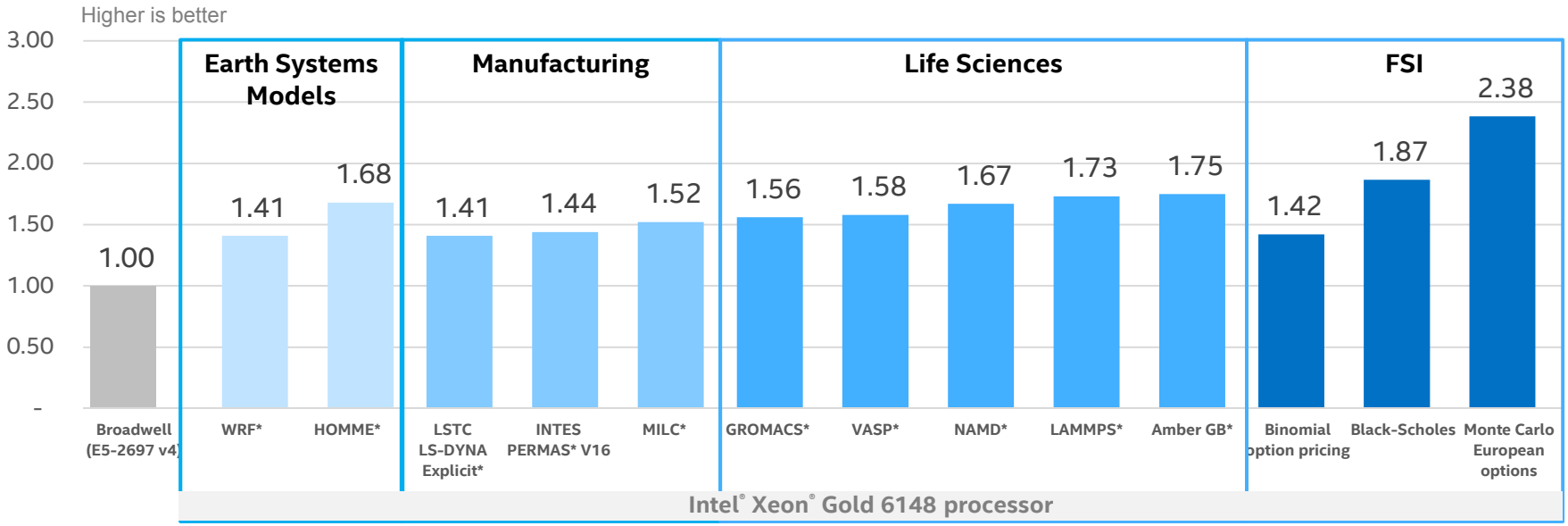
Generational Performance Gains on 2-Socket Servers with Intel® Xeon® Scalable Processor



Geomean based on Normalized Generational Performance (estimated based on Intel internal testing of OLTP Brokerage, SAP SD 2-Tier, HammerDB, Server-side Java, SPEC*int_rate_base2006, SPEC*fp_rate_base2006, Server Virtualization, STREAM* triad, LAMMPS, DPDK L3 Packet Forwarding, Black-Scholes, Intel Distribution for LINPACK.

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/performance 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. Configuration: see page 31, 32

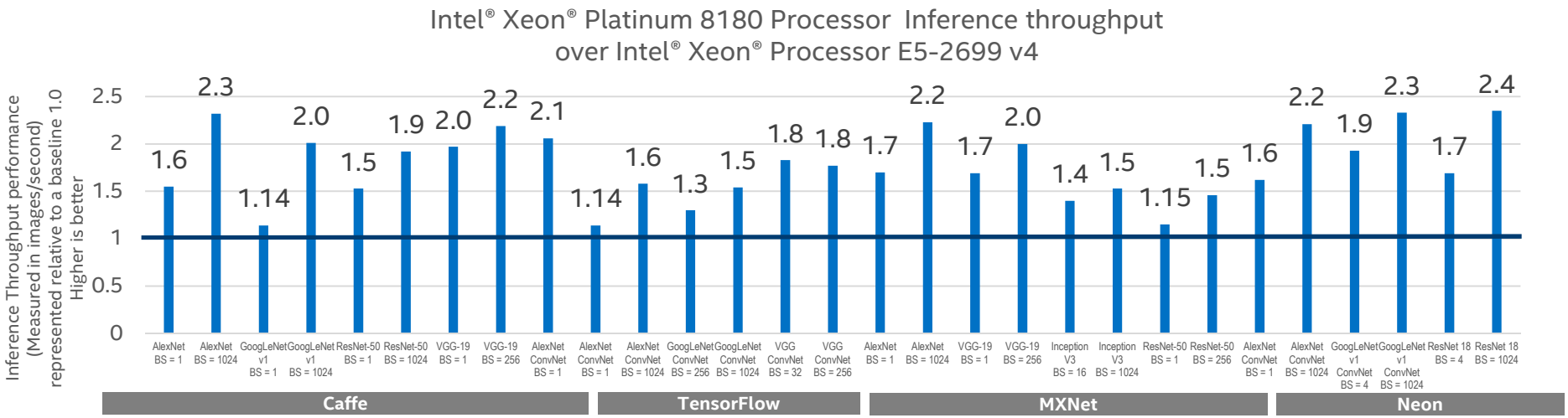
Performance Gains on Technical Compute Workloads



Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Geomean of Weather Research Forecasting - Conus 12Km, HOMME, LSTCLS-DYNA Explicit, INTES PERMAS V16, MILC, GROMACS water 1.5M_pme, VASPSi256, NAMDstmv, LAMMPS, Amber GB Nucleosome, Binomial option pricing, Black-Scholes, Monte Carlo European options. Any difference in system hardware or software design or configuration may affect 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 http://www.intel.com/performance/datacenter.
Configurations: see page 33,34

Machine Learning Inference Throughput

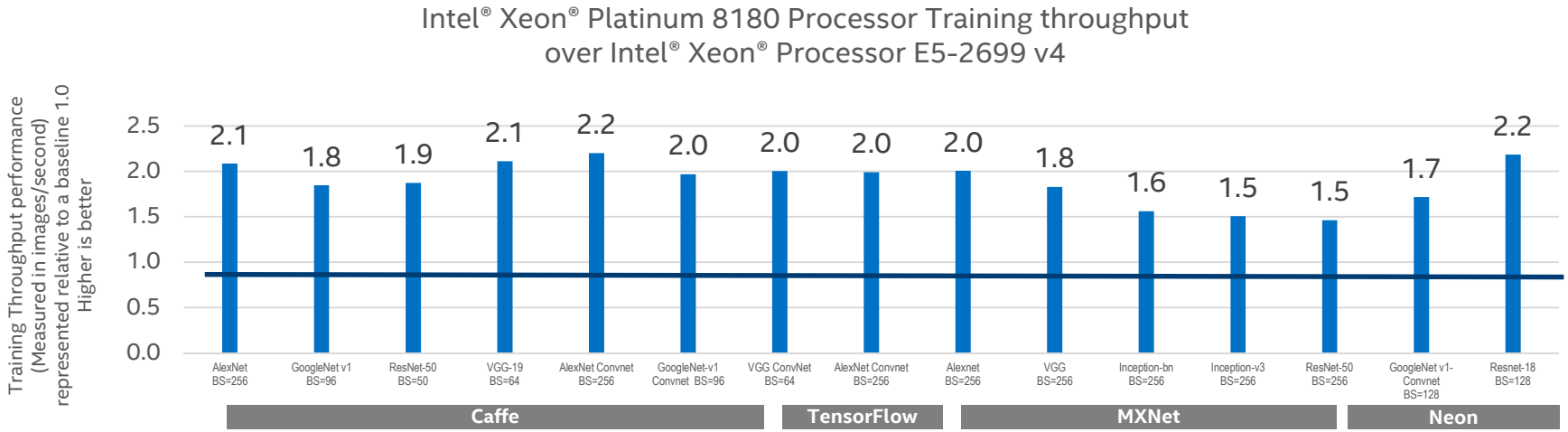
on Intel® Xeon® Platinum 8180 Processor



Intel® Xeon® Platinum Processor delivers up to 2.4x higher Inference throughput 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 complete information visit: http://www.intel.com/performance Source: Intel measured as of June 2017. INFERENCE using FP32. Configuration Details on Slide:35,36

Machine Learning Training Throughput
on Intel® Xeon® Platinum 8180 Processor



Intel® Xeon® Platinum Processor delivers up to 2.2x higher Training throughput 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 complete information visit: http://www.intel.com/performance Source: Intel measured as of June 2017. Configuration Details on Slide:35,36

WRAP UP

Intel® Xeon® Scalable Processor Summary

Architectural innovations to unlock data center performance

- **Up to 60% increase** in compute performance per socket with Intel® AVX-512
- **Improved performance and scalability** with Mesh on-chip interconnect
- L2 and L3 cache hierarchy **optimized for data center workloads**
- Improved memory subsystem with **up to 60% higher memory bandwidth**
- Faster and more efficient Intel® UPI interconnect for **improved scalability**
- Improved integrated IO with **up to 50% higher aggregate IO bandwidth**

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/performance 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. See pages 7, 13, and 16 for supporting performance data and configuration details.

INTEL® XEON® SCALABLE PROCESSOR PLATFORM

The Secure, Agile, Next-Generation Platform for Multi-Cloud Infrastructures

Skylake-SP cores

Intel® AVX-512

Feeds: Intel® UPI, 6x DDR4, 48x PCIe Gen3, Intel® SSDs

Integration: Intel® VMD/ Omni-Path / Intel® QuickAssist / Intel® Ethernet

Intel® AVX-512

Page Protection Key, Mode Based Execution

Intel® QAT w/ Secure Key Management

Intel® Trusted Infrastructure

Intel® Boot Guard

Intel® Volume Management Device Technology

Communication and Storage Acceleration Libraries

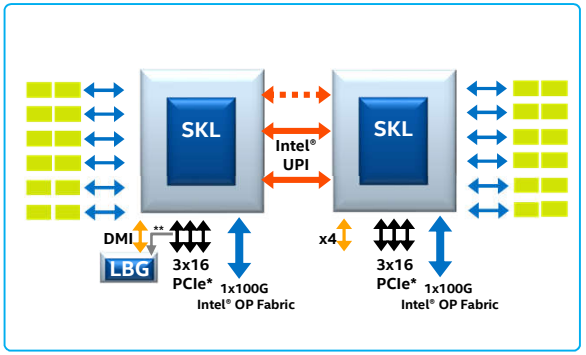
Intel® Run Sure Technology

OpenStack Software Optimizations

ADDITIONAL INFO

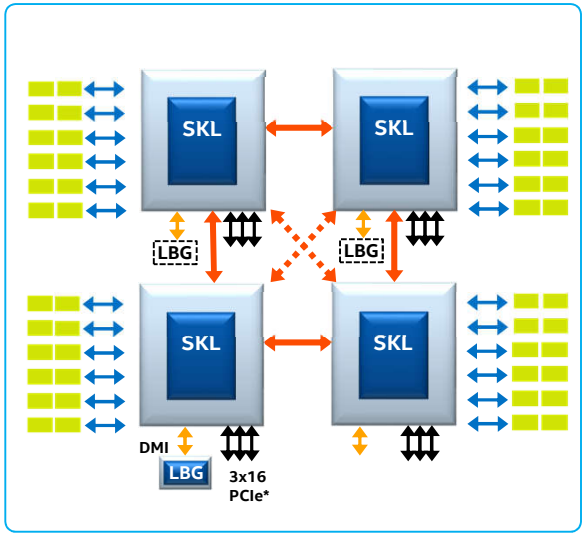
Platform Topologies

2S Configurations



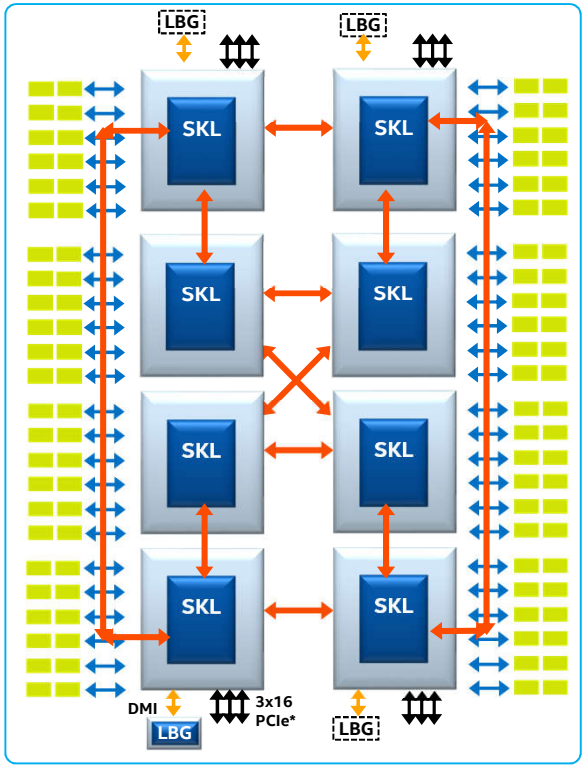
(2S-2UPI & 2S-3UPI shown)

4S Configurations



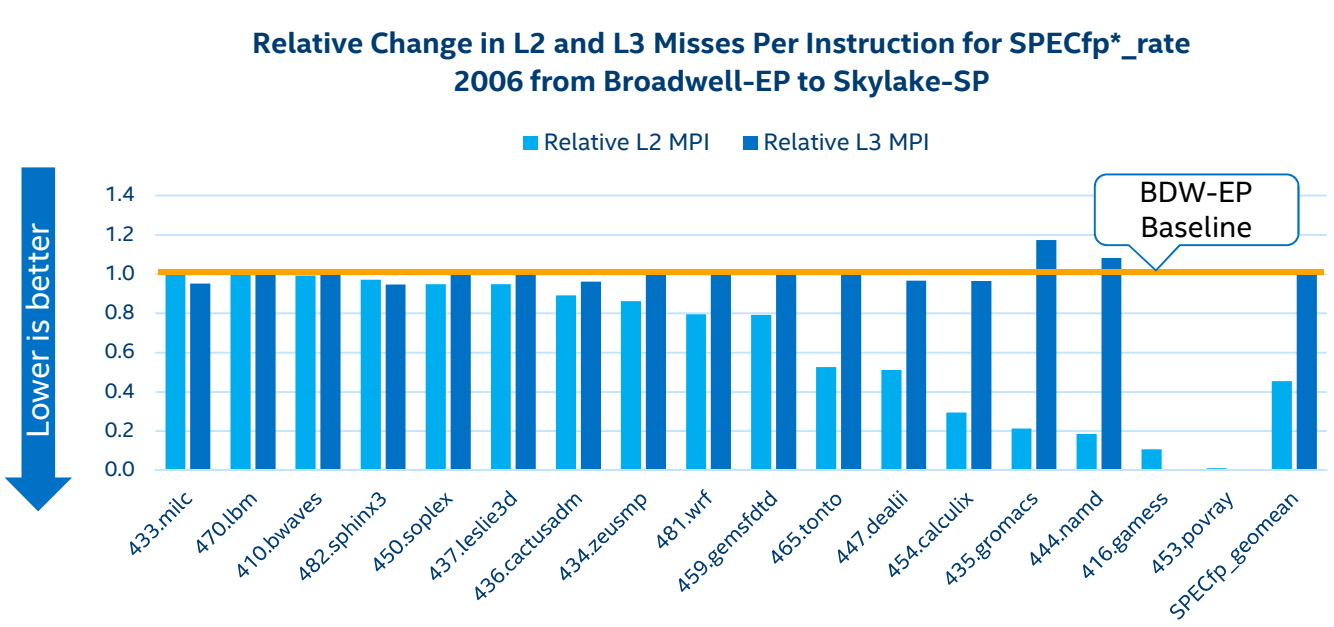
(4S-2UPI & 4S-3UPI shown)

8S Configuration



INTEL® XEON® SCALABLE PROCESSOR SUPPORTS CONFIGURATIONS RANGING FROM 2S-2UPI TO 8S

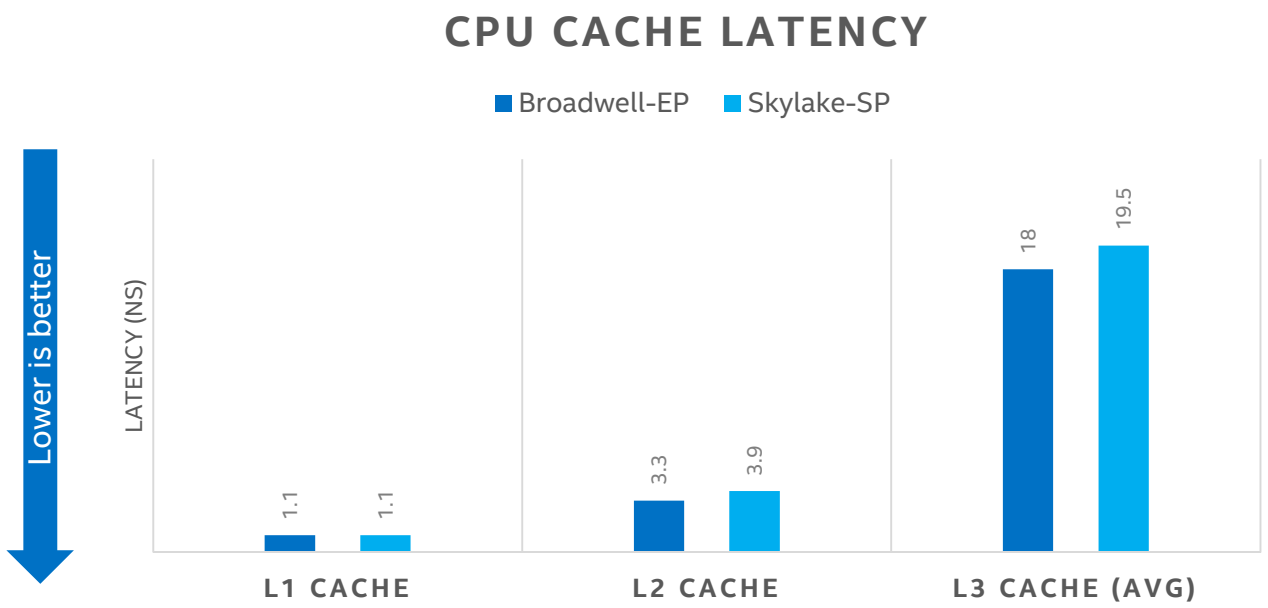
Cache Performance



Skylake-SP cache hierarchy significantly reduces L2 misses without increasing L3 misses compared to Broadwell-EP

Source as of June 2017: Intel internal measurements on platform with Xeon Platinum 8180, Turbo enabled, UPI=10.4, SNC1, 6x32GB DDR4-2666 per CPU, 1 DPC, and platform with E5-2699 v4, Turbo enabled, 4x32GB DDR4-2400, RHEL 7.0. 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.

Cache Performance

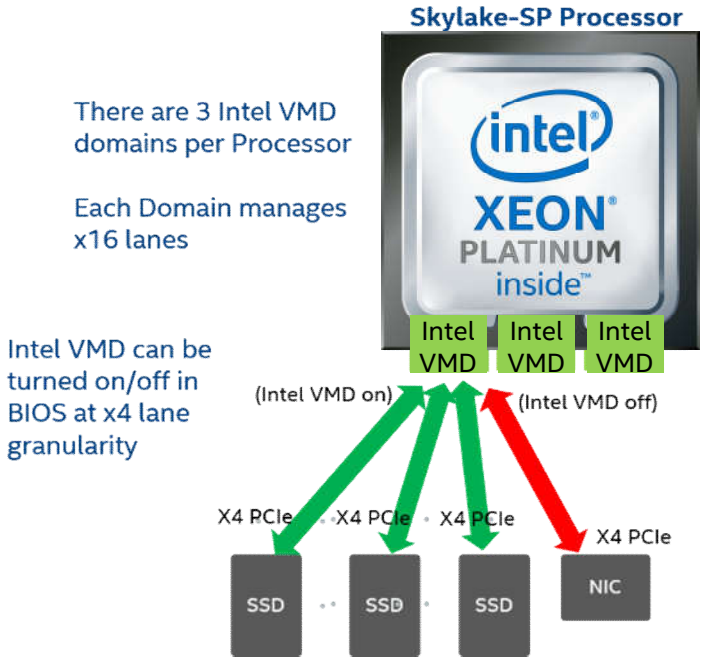


Skylake-SP L2 cache latency has increased by 2 cycles for a 4x larger L2

Skylake-SP achieves good L3 cache latency even with larger core count

Source as of June 2017: Intel internal measurements on platform with Xeon Platinum 8180, Turbo enabled, SNC1, 6x32GB DDR4-2666 per CPU, 1 DPC, and platform with Intel® Xeon® E5-2699 v4, Turbo enabled, without COD, 4x32GB DDR4-2400, RHEL 7.0. Cache latency measurements were done using Intel® Memory Latency Checker (MLC) tool.
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.

Intel® Volume Management Device (Intel® VMD)



Intel® VMD is a CPU-integrated device to aggregate NVMe SSDs into a storage volume and enables other storage services such as RAID

- Intel® VMD is an “integrated end point” that stops OS enumeration of devices under it
- Intel® VMD maps entire PCIe* trees into its own address space (a domain)
- Intel® VMD driver sets up and manages the domain (enumerate, event/error handling), but out of fast IO path

ELIMINATES ADDITIONAL COMPONENTS TO PROVIDE A FULL-FEATURE STORAGE SOLUTION

Intel® Xeon® Scalable Processor with Integrated Fabric

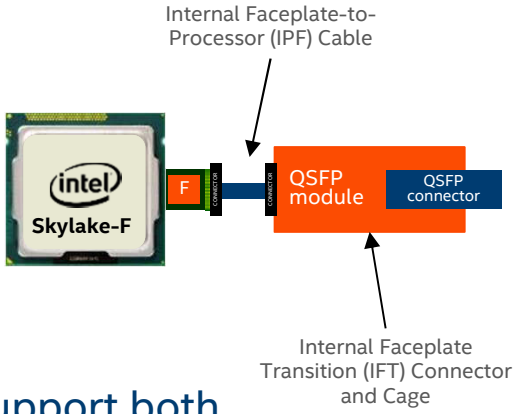
Single on-package Omni-Path Host Fabric Interface (HFI)

Fabric component interfaces to CPU using x16 PCIe* lanes

Fabric PCIe lanes are additional to the 48 PCIe lanes

Single cable from CPU package connector to QSFP module

Same socket for processors with or without Omni-Path fabric



- Intel® Xeon® Scalable Processor Platform can be designed to support both processors
- Platform design requires an expanded keep-out zone and additional board components to accommodate both processors

Configurations: Average Generational Gains on 2S Servers

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/performance 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.

4. **1.65x Average Performance:** Geomean based on Normalized Generational Performance (estimated based on Intel internal testing of OLTP Brokerage, SAP SD 2-Tier, HammerDB, Server-side Java, SPEC*int_rate_base2006, SPEC*fp_rate_base2006, Server Virtualization, STREAM* triad, LAMMPS, DPDK L3 Packet Forwarding, Black-Scholes, Intel Distribution for LINPACK.
   a) **Up to 1.33x on TPC*-E**: 1-Node, 2 x Intel® Xeon® Processor E5-2699 v4 on Lenovo Group Limited with 512 GB Total Memory on Windows Server* 2012 Standard using SQL Server 2016 Enterprise Edition. Data Source:http://www.tpc.org/tpce/results/tpce_result_detail.asp?id=116032402 , Benchmark: TPC Benchmark* E (TPC-E), Score: 4938.14 vs. 1-Node, 2 x Intel® Xeon® Platinum 8180 processor on Lenovo Group Limited with 1536 GB Total Memory on Windows Server* 2016 Standard using SQL Server 2017 Enterprise Edition. Data Source: http://www.tpc.org/tpce/results/tpce_result_detail.asp?id=117062701, Benchmark: TPC Benchmark* E (TPC-E), Score: 6598.36. Higher is better
   b) **Up to 1.40x on SPECvirt_sc* 2013**: Claim based on best-published 2-soclet SPECvirt_sc* 2013 result submitted to/published at http://www.spec.org/virt_sc2013/results/res2016q3/virt_sc2013-20160823-00060-perf.html as of 11 July 2017, Score: 2360 @ 137 VMs vs. 1-Node, 2 x Intel® Xeon® Platinum 8180 Processor with 768 GB (24 x 32 GB, 2R x4 PC4-2666 DDR4 2666MHz RDIMM) Total Memory on SUSE Linux Enterprise Server 12 SP2. Data Source: http://www.spec.org, Benchmark: SPECvirt_sc* 2013, Score: 3323 @ 189 VMs Higher is better
   c) **Up to 1.44x on 2-Tier SAP* SD** : Claim based on best-published two-socket SAP SD 2-Tier on Linux* result published at http://global.sap.com/solutions/benchmark/sd2tier.epx as of 11 July 2017. New configuration: 2-tier, 2 x Intel® Xeon® Platinum 8180 Processor (56 cores/112 threads) on DellEMC PowerEdge* R740xd with 768 GB total memory on Red Hat Enterprise Linux* 7.3 using SAP Enhancement Package 5 for SAP ERP 6.0, SAP NetWeaver 7.22 pl221, and Sybase ASE 16.0. Source: Certification #: 2017017: www.sap.com/benchmark, SAP* SD 2-Tier enhancement package 5 for SAP ERP 6.0 score: 32,085 benchmark users.
   d) **Up to 1.53x on SPECint*_rate_base2006** : Claim based on best-published two-socket SPECint*_rate_base2006 result submitted to/published at http://www.spec.org/cpu2006/results/ as of 11 July 2017. New configuration: 1-Node, 2 x Intel® Xeon® Platinum 8180 Processor on Huawei 2288H V5 with 384 GB total memory on SUSE Linux Enterprise Server 12 SP2 (x86_64) Kernel 4.4.21-69-default, using C/C++: Version 17.0.1.132 of Intel C/C++ Compiler for Linux. Source: submitted to www.spec.org, SPECint*_rate_base2006 Score: 2800. Results are pending SPEC approval; they are considered estimates until SPEC approves
   e) **Up to 1.58x on SPECjbb*2015 MultiJVM critical-jOPS**: Claim based on best-published two-socket SPECjbb*2015 MultiJVM critical-jOPS results published at http://www.spec.org/jbb2015/results/jbb2015multijvm.html as of 11 July 2017. New configuration: 1-Node, 2 x Intel® Xeon® Platinum 8180 Processor on Cisco* Systems UCS C240 M5 with 1536 GB total memory on Red Hat Enterprise Linux* 7.3 (Maipo) using Java* HotSpot 64-bit Server VM, version 1.8.0_131. Source: submitted to http://www.spec.org, SPECjbb2015* - MultiJVM scores: 141,360 max-jOPS and 118,551 critical-jOPS
   f) **Up to 1.65x on est SPECfp*_rate_base2006** :Claim based on best-published two-socket SPECfp*_rate_base2006 result submitted to/published at http://www.spec.org/cpu2006/results/ as of 11 July 2017. New configuration: 1-Node, 2 x Intel® Xeon® Platinum 8180 Processor on Huawei 2288H V5 with 384 GB total memory on SUSE Linux Enterprise Server 12 SP2 (x86_64) Kernel 4.4.21-69-default, using C/C++ and Fortran: Version 17.0.0.098 of Intel C/C++ and Intel Fortran Compiler for Linux. Source: submitted to www.spec.org, SPECfp*_rate_base2006 Score: 1850. Results are pending SPEC approval; they are considered estimates until SPEC approves
   g) **Up to 1.65x on est STREAM - triad**: 1-Node, 2 x Intel® Xeon® Processor E5-2699 v4 on Grantley-EP (Wellsburg) with 256 GB Total Memory on Red Hat Enterprise Linux* 6.5 kernel 2.6.32-431 using Stream NTW avx2 measurements. Data Source: Request Number: 1709, Benchmark: STREAM - Triad, Score: 127.7 Higher is better vs. 1-Node, 2 x Intel® Xeon® Platinum 8180 Processor on Neon City with 384 GB Total Memory on Red Hat Enterprise Linux* 7.2-kernel 3.10.0-327 using STREAM AVX 512 Binaries. Data Source: Request Number: 2500, Benchmark: STREAM - Triad, Score: 199 Higher is better

Configurations: Average Generational Gains on 2S Servers

h) **Up to 1.73x on HammerDB**:1-Node, 2 x Intel® Xeon® Processor E5-2699 v4 on Grantley-EP (Wellsburg) with 384 GB Total Memory on Red Hat Enterprise Linux* 7.1 kernel 3.10.0-229 using Oracle 12.1.0.2.0 (including database and grid) with 800 warehouses, HammerDB 2.18. Data Source: Request Number: 1645, Benchmark: HammerDB, Score: 4.13568e+006 Higher is better vs. 1-Node, 2 x Intel® Xeon® Platinum 8180 Processor on Purley-EP (Lewisburg) with 768 GB Total Memory on Oracle Linux* 7.2 using Oracle 12.1.0.2.0, HammerDB 2.18. Data Source: Request Number: 2510, Benchmark: HammerDB, Score: 7.18049e+006 Higher is better

i) **Up to 1.73x on LAMMPS**: LAMMPS is a classical molecular dynamics code, and an acronym for Large-scale Atomic/Molecular Massively Parallel Simulator. It is used to simulate the movement of atoms to develop better therapeutics, improve alternative energy devices, develop new materials, and more. E5-2697 v4: 2S Intel® Xeon® processor E5-2697 v4, 2.3GHz, 36 cores, Intel® Turbo Boost Technology and Intel® Hyperthreading Technology on, BIOS 86B0271.R00, 8x16GB 2400MHz DDR4, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327. Gold 6148: 2S Intel® Xeon® Gold 6148 processor, 2.4GHz, 40 cores, Intel® Turbo Boost Technology and Intel® Hyperthreading Technology on, BIOS 86B.01.00.0412.R00, 12x16GB 2666MHz DDR4, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327.

j) **Up to 1.77x on DPDK L3 Packet Forwarding**: E5-2658 v4: 5 x Intel® XL710-QDA2, DPDK 16.04. Benchmark: DPDK l3fwd sample application Score: 158 Gbits/s packet forwarding at 256B packet using cores. Gold 6152: Estimates based on Intel internal testing on Intel Xeon 6152 2.1 GHz, 2x Intel®, FM10420(RRC) Gen Dual Port 100GbE Ethernet controller (100Gbit/card) 2x Intel® XXV710 PCI Express Gen Dual Port 25GbE Ethernet controller (2x25G/card), DPDK 17.02. Score: 281 Gbits/s packet forwarding at 256B packet using cores, IO and memory on a single socket

k) **Up to 1.87x on Black-Scholes:** which is a popular mathematical model used in finance for European option valuation. This is a double precision version. E5-2697 v4: 2S Intel® Xeon® processor CPU E5-2697 v4 , 2.3GHz, 36 cores, turbo and HT on, BIOS 86B0271.R00, 128GB total memory, 8 x16GB 2400 MHz DDR4 RDIMM, 1 x 1TB SATA, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327. Gold 6148: Intel® Xeon® Gold processor 6148@ 2.4GHz, H0QS, 40 cores 150W. QMS1, turbo and HT on, BIOS SE5C620.86B.01.00.0412.020920172159, 192GB total memory, 12 x 16 GB 2666 MHz DDR4 RDIMM, 1 x 800GB INTEL SSD SC2BA80, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327

l) **Up to 2.27x on LINPACK***: 1-Node, 2 x Intel® Xeon® Processor E5-2699 v4 on Grantley-EP (Wellsburg) with 64 GB Total Memory on Red Hat Enterprise Linux* 7.0 kernel 3.10.0-123 using MP_LINPACK 11.3.1 (Composer XE 2016 U1). Data Source: Request Number: 1636, Benchmark: Intel® Distribution of LINPACK, Score: 1446.4 Higher is better vs. 1-Node, 2 x Intel® Xeon® Platinum 8180 Processor on Wolf Pass SKX with 384 GB Total Memory on Red Hat Enterprise Linux* 7.3 using mp_linpack_2017.1.013. Data Source: Request Number: 3753, Benchmark: Intel® Distribution of LINPACK, Score: 3295.57 Higher is better

Configurations: Technical Compute Workloads

Up to 1.63x Gains based on Geomean of Weather Research Forecasting - Conus 12Km, HOMME, LSTCLS-DYNA Explicit, INTES PERMAS V16, MILC, GROMACS water 1.5M_pme, VASPSi256, NAMDstmv, LAMMPS, Amber GB Nucleosome, Binomial option pricing, Black-Scholes, Monte Carlo European options. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect 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 http://www.intel.com/performance/datacenter.

1. **PERMAS by INTES** is an advanced Finite Element software system that offers a complete range of physical models at high performance, quality, and reliability. It plays a mission-critical role in the design process at customers from automotive, ship design, aerospace, and more. E5-2697 v4: 2S Intel® Xeon® processor E5-2697v4, 2.3GHz, 18 cores, turbo on, HT off, NUMA on, BIOS 338.R00, 256 GB total memory (8x 32GB w/ 2400 MT/s, DDR4 LRDIMM), 4x Intel® SSD DC P3600 2 TB in RAID 0 (stripe size 64k). CentOS Linux* release 7.2, kernel 3.10.0-327.13.1.el7.x86_64. Intel® Composer 2015.5.223. INTES PERMAS V16.00. Gold 6148: Intel® Xeon® Gold 6148 processor, 2.4 GHz, 20 cores, turbo on, HT off, NUMA on, BIOS SE5C620.86B.01.00.0412.020920172159, 384 GB total memory (12x 32GB w/ 2400 MT/s, DDR4 LRDIMM), 3x Intel® SSD DC P3600 2 TB in RAID 0 (stripe size 64k), CentOS* Linux* release 7.3, kernel 3.10.0-514.10.2.el7.x86_64. Intel® Composer 2015.7.235. INTES PERMAS V16.00.
2. **LS-DYNA** is the leading product in the crash simulation market. It is used by the automobile, aerospace, construction, military, manufacturing, and bioengineering industries in worldwide. Workload: 2M elements Car2car model with 120ms simulation time. LS-DYNA explicit standard benchmarks tested by Intel, March 2017. E5-2697 V4: 2S Intel® Xeon® processor E5-2697 v4, 2.3GHz, 18 cores, turbo and HT on, BIOS SE5C610.86B.01.01.0016.033120161139, 128GB total memory, 8 memory channels / 8x16GB / 2400 MT/s / DDR4, Red Hat Enterprise Linux* 7.3 kernel 3.10.0-229.20.1.el6.x86_64.knl2. GOLD 6148: 2S Intel® Xeon® Gold 6148 processor, 2.4GHz, 20 cores, turbo and HT on, BIOS version 412, 192GB total memory, 12 memory channels / 12x16GB / 2400 MT/s / DDR4, Red Hat Enterprise Linux* 7.3 kernel 3.10.0-514.el7.x86_64.
3. **Binomial option pricing** is a lattice-based approach that uses a discrete-time model of the varying price over time of the underlying financial instrument. This is compute bound, double precision workload. FSI Binomial workload. OS: Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327. Testing by Intel March 2017. E5-2697 v4: 2S Intel® Xeon® processor CPU E5-2697 v4 , 2.3GHz, 36 cores, turbo and HT on, BIOS 86B0271.R00, 128GB total memory, 8 slots / 16GB / 2400 MT/s / DDR4 RDIMM, 1 x 1TB SATA, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327. Gold 6148: Intel® Xeon® Gold 6148 processor, 2.4GHz, 40 cores, turbo and HT on, BIOS 86B.01.00.0412, 192GB total memory, 12 slots / 16 GB / 2666 MT/s / DDR4 RDIMM, 1 x 800GB INTEL SSD SC2BA80, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327.
4. **Monte Carlo** is a numerical method that uses statistical sampling techniques to approximate solutions to quantitative problems. In finance, Monte Carlo algorithms are used to evaluate complex instruments, portfolios, and investments. This is compute bound, double precision workload. FSI Monte Carlo workload. OS: Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327. Testing by Intel March 2017. E5-2697 v4: 2S Intel® Xeon® processor CPU E5-2697 v4 , 2.3GHz, 36 cores, turbo and HT on, BIOS 86B0271.R00, 128GB total memory, 8 x16GB 2400 MHz DDR4 RDIMM, 1 x 1TB SATA, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327. Gold 6148: Intel® Xeon® Gold 6148 processor@ 2.4GHz, H0QS, 40 cores 150W. QMS1, turbo and HT on, BIOS SE5C620.86B.01.00.0412.020920172159, 192GB total memory, 12 x 16 GB 2666 MHz DDR4 RDIMM, 1 x 800GB INTEL SSD SC2BA80, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327
5. **Black-Scholes** is a popular mathematical model used in finance for European option valuation. This is a double precision version. E5-2697 v4: 2S Intel® Xeon® processor CPU E5-2697 v4 , 2.3GHz, 36 cores, turbo and HT on, BIOS 86B0271.R00, 128GB total memory, 8 x16GB 2400 MHz DDR4 RDIMM, 1 x 1TB SATA, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327. Gold 6148: Intel® Xeon® Gold 6148 processor@ 2.4GHz, H0QS, 40 cores 150W. QMS1, turbo and HT on, BIOS SE5C620.86B.01.00.0412.020920172159, 192GB total memory, 12 x 16 GB 2666 MHz DDR4 RDIMM, 1 x 800GB INTEL SSD SC2BA80, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327
6. **Amber*** is a suite of programs for classical molecular dynamics and statistical analysis. The main MD program is PMEMD (Particle Mesh Ewald Molecular Dynamics) employs two separate algorithms for implicit- and explicit-solvent dynamics. Here performance for explicit solvent (PME) is presented. Amber: Version 16 with all patches applied at December, 2016. Workloads: PME Cellulose NVE(408K atoms), PME stmv(1M atoms), GB Nucleosome (25K), GB Rubisco (75K). No cut-off was used for GB workloads. Compiled with -mic2_spdp –intelmpi - openmp, –DMIC2 * defined. Tests performed on March 2017. E5-2697 v4: Executed with 36 MPI, 2 OpenMP. 2S Intel® Xeon® processor E5-2697 v4, 2.3GHz, 36 cores, turbo and HT on, BIOS 86B0271.R00, 8x16GB 2400MHz DDR4, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327. Gold 6148: Executed with 40 MPI and 2 OpenMP. 2S Intel® Xeon® Gold 6148 processor, 2.4GHz, 40 cores, turbo on, HT on, BIOS 86B.01.00.0412.R00, 12x16GB 2666MHz DDR, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327.

Configurations: Technical Compute Workloads

7. **VASP** CONFIGURATION: The Vienna Ab initio Simulation Package (VASP) is a computer program for atomic scale materials modeling and performs electronic structure calculations and quantum-mechanical molecular dynamics from first principles. VASP provides scientists with fast and precise calculation of materials properties covering wide range of MD methods from DFT, DFT-HF to Random-Phase approximation (GW, ACDFT). Beta VASP, a release candidate for v6.0. Developer branch provided as "Package" included with download: https://github.com/vasp-dev/vasp-knl. AVX512: Intel® Compiler 17.0.1.132, Intel® MPI 2017u1, ELPA 2016.05.004. Optimization Flags: "-O3 -xCORE-AVX512". AVX2: Intel® Compiler 17.0.1.132, Intel® MPI 2017u1, ELPA 2016.05.004. Optimization Flags: "-O3 -xCORE-AVX2". E5-2697 v4: 2S Intel® Xeon® processor E5-2697 v4 2.3 GHz , 18 Cores/Socket, 36 Cores, 72 Threads, HT on, turbo off, BIOS 86B0271.R00, 128GB total memory, 2400 MT/s DDR4 RDIMM, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327. Gold 6148: Dual Socket Intel® Xeon® processor Gold 6148 2.4 GHz , 20 Cores/Socket, 40 Cores, 80 Threads, HT on, turbo off, BIOS 86B.01.00.0412, 192GB total memory, 2666 MT/s / DDR4 RDIMM, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327.
8. **NAMD**: NAMD, recipient of a 2002 Gordon Bell Award, is a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems. Based on Charm++ parallel objects, NAMD scales to hundreds of cores for typical simulations and beyond 200,000 cores for the largest simulations. Version 2.12 Dec2016. Workloads: apoa1(92K atoms), stmv(1M atoms). Compiled with –DNAMD_KNL* define. Tests performed on March 2017. E5-2697 v4: 2S Intel® Xeon® processor E5-2697 v4, 2.3GHz, 36 cores, turbo and HT on, BIOS 86B0271.R00, 8x16GB 2400MHz DDR4, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327. Compiler option "-xCORE-AVX2". Gold 6148: 2S Intel® Xeon® Gold 6148 processor, 2.4GHz, 40 cores, turbo on, HT on, BIOS 86B.01.00.0412.R00, 12x16GB 2666MHz DDR, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327. Compiler option "-xCORE-AVX512".
9. **LAMMPS**: LAMMPS is a classical molecular dynamics code, and an acronym for Large-scale Atomic/Molecular Massively Parallel Simulator. It is used to simulate the movement of atoms to develop better therapeutics, improve alternative energy devices, develop new materials, and more. E5-2697 v4: 2S Intel® Xeon® processor E5-2697 v4, 2.3GHz, 36 cores, Intel® Turbo Boost Technology and Intel® Hyperthreading Technology on, BIOS 86B0271.R00, 8x16GB 2400MHz DDR4, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327. Gold 6148: 2S Intel® Xeon® Gold 6148 processor, 2.4GHz, 40 cores, Intel® Turbo Boost Technology and Intel® Hyperthreading Technology on, BIOS 86B.01.00.0412.R00, 12x16GB 2666MHz DDR4, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327.
10. **GROMACS** is a versatile package to perform classical Molecular Dynamics simulations. Heavily optimized for most modern platforms and provides extremely high performance. GROMACS AVX2 CONFIGURATION: Version 2016.3: ftp://ftp.gromacs.org/pub/gromacs/gromacs-2016.3.tar.gz , Intel® Compiler 17.0.1.132, Intel® MPI 2017u1. Optimization Flags: "-O3 -xCORE-AVX2". Cmake options: "-DGMX_FFT_LIBRARY=mkl -DGMX_SIMD=AVX2_256". GROMACS AVX512 CONFIGURATION: Version 2016.3: ftp://ftp.gromacs.org/pub/gromacs/gromacs-2016.3.tar.gz , Intel® Compiler 17.0.1.132, Intel® MPI 2017u1. Optimization Flags: "-O3 -xCORE-AVX512". Cmake options: "-DGMX_FFT_LIBRARY=mkl -DGMX_SIMD=AVX_512". E5-2697 V4: GROMACS AVX2 binary, Dual Socket Intel® Xeon® processor E5-2697 v4 2.3 GHz, 18 Cores/Socket, 36 Cores, 72 Threads (HT on, Turbo on), DDR4 128GB, 2400 MHz, Red Hat 7.2. Gold 6148: GROMACS AVX512 binary, Dual Socket Intel® Xeon® processor Gold 6148 2.4 GHz , 20 Cores/Socket, 40 Cores, 80 Threads (HT on, Turbo on), DDR4 192GB, 2666 MT/s DDR4 RDIMMs, Red Hat 7.2.
11. **Weather Research and Forecasting** (WRF) Model is a next-generation mesoscale numerical weather prediction system designed for both atmospheric research and operational forecasting needs. It features two dynamical cores, a data assimilation system, and a software architecture facilitating parallel computation and system extensibility. The model serves a wide range of meteorological applications across scales from tens of meters to thousands of kilometers. 2S Intel® Xeon® processor CPU E5-2697 v4 , 2.3GHz, 36 cores, turbo and HT on, BIOS 86B0271.R00, 128GB total memory, 8 slots / 16GB / 2400 MT/s / DDR4 RDIMM, 1 x 1TB SATA HDD, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327. Software: WRF version 3.6.1 Compiled using Intel config option with "-O3 -fp-model fast=1 -xCORE-AVX2". Executed with 36 MPI ranks and OMP_NUM_THREADS=1. Intel® Xeon® Gold 6148 processor, 2.4GHz, 40 cores, turbo and HT on, BIOS 86B.01.00.0412, 192GB total memory, 12 slots / 16 GB / 2666 MT/s / DDR4 RDIMM, 1 x 800GB INTEL SSD SC2BA80, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327. Software: WRF version 3.6.1 Compiled using Intel config option with "-O3 -fp-model fast=1 -xCORE-AVX512". Executed with 40 MPI ranks and OMP_NUM_THREADS=1.
12. **HOMME** is the spectral element dynamical core that solves the equations of motion in the CAM-SE atmospheric model, part of the NSF Community Earth System Model (CESM) as well as the related DOE ACME model. CESM is a widely-used Earth system model and an important source of simulations used by the Intergovernmental Panel on Climate Change. HOMME version: https://svn-homme-model.cgd.ucar.edu/branch_tags/dungeon_tags/dungeon016. Compiled with "-O3 -fp-model fast -xCORE-AVX2". Running "perfTestWACCM" benchmark from code repository with size NE=8. Executed with 64 MPI ranks and OMP_NUM_THREADS=1. E5-2697 v4: 2S Intel® Xeon® processor CPU E5-2697 v4 , 2.3GHz, 36 cores, turbo and HT on, BIOS 86B0271.R00, 128GB total memory, 8 slots / 16GB / 2400 MT/s / DDR4 RDIMM, 1 x 1TB SATA, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327. Gold 6148: Intel® Xeon® Gold 6148 processor, 2.4GHz, 40 cores, turbo and HT on, BIOS 86B.01.00.0412, 192GB total memory, 12 slots / 16 GB / 2666 MT/s / DDR4 RDIMM, 1 x 800GB INTEL SSD SC2BA80, Red Hat Enterprise Linux* 7.2 kernel 3.10.0-327.

Configuration Details: Machine Learning

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/), 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 (GoogLeNet, AlexNet, and ResNet-50), 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). 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), commit id 207203253b6f8ea5e938a512798429f91d5b4e7e. Performance numbers were obtained for three convnet benchmarks: alexnet, googlenetv1, 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 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/), revision 5efd91a71f36fea483e882b0358c8d46b5a7aa20. 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 2018.0.20170425.

- Neon: ZP/MKL_CHWN branch commit id:52bd02acb947a2adabb8a227166a7da5d9123b6d. 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.

Configuration Details: Machine Learning

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

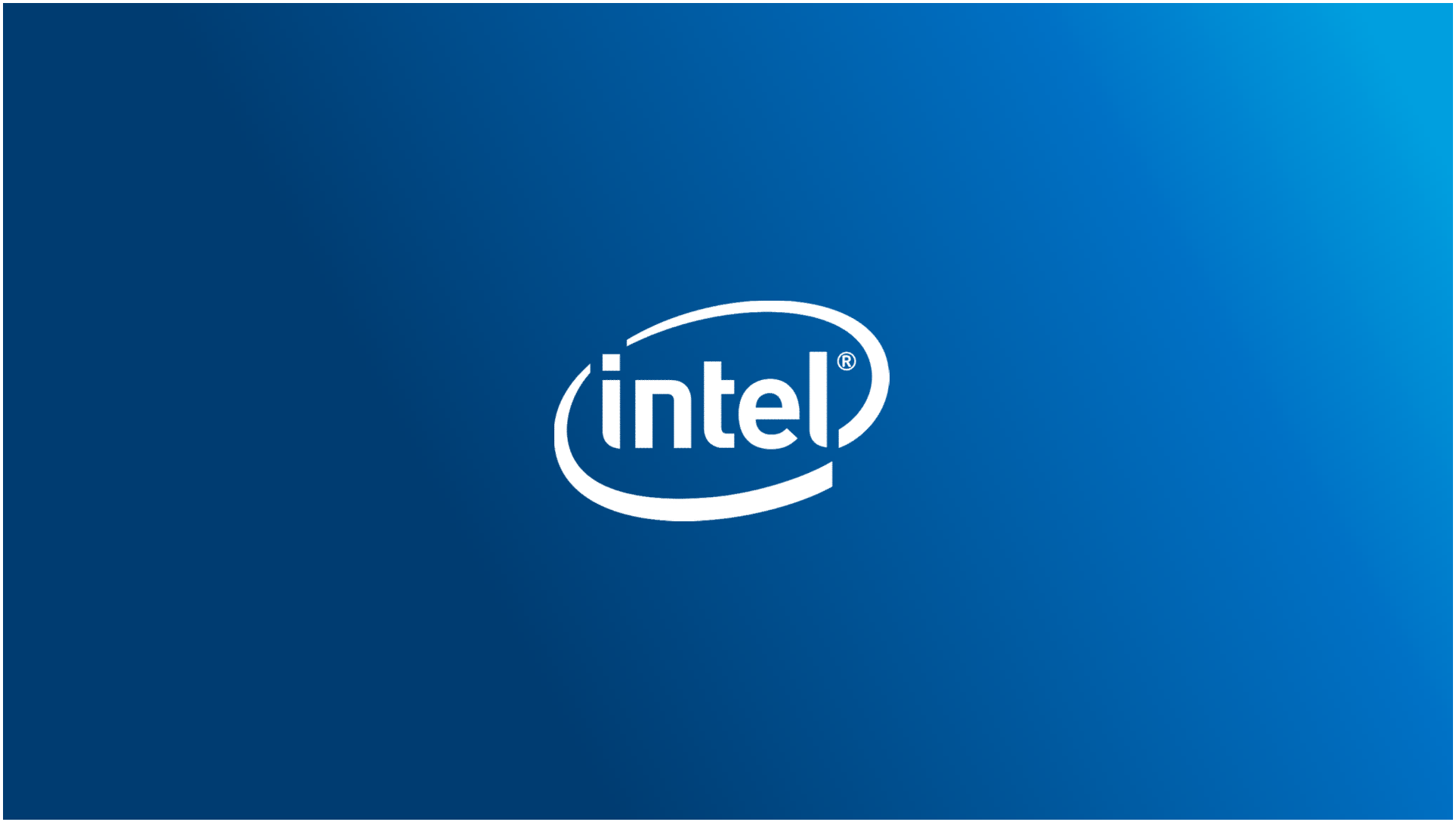
Deep Learning Frameworks:

- 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 (GoogLeNet, AlexNet, and ResNet-50), 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, googlenetv1, 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 googlenet 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 e9f281a27584cdb78db8ce6b66e648b3dbc10d37. 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:52bd02acb947a2adabb8a227166a7da5d9123b6d. 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.


intel®