FREQUENTLY ASKED QUESTIONS ABOUT 2ND GENERATION INTEL® XEON® SCALABLE PROCESSOR-BASED SERVERS

Why are performance requirements different for data-centric workloads, such as AI, analytics, machine learning, deep learning and others?

Today's data-centric, insight-driven workloads are different than past workloads in two key areas. First, they tend to be compute-intensive and highly-parallel, so they can greatly benefit from highly-scalable processors that deliver outstanding per-core performance.

Second, these workloads are also very data-intensive, which means they really need more memory capacity, faster memory bandwidth, high-speed storage, and fast, reliable network I/O.

Why is there so much discussion about the Intel platform-wide benefits, as well as 2nd Gen Intel® Xeon® Scalable processor benefits?

For many years, system memory and storage technology improvements have simply not kept up with processor improvements.

Each generation of Intel® Xeon® processors has come with significant new advantages in performance and other capabilities, such as more cores, faster cores, new performance-optimized instructions to accelerate critical tasks and workloads, etc.

However, meaningful increases in DRAM capacity and storage performance have come very slowly in comparison—resulting in bottlenecks that have prevented organizations from getting more out of those processor improvements.

Intel recognized that data-centric workloads—such as AI, analytics, machine learning, deep learning and others—are placing greater strains on memory, storage and I/O performance and capacity. Unless these bottlenecks were resolved, processors were going to be forced to wait for data, no matter how fast they could process it.

That's why Intel took an end-to-end approach to platform improvements—with many improvements built into 2nd Gen Intel® Xeon® Scalable processors—while also relieving the DRAM capacity bottleneck with Intel® Optane™ DC persistent memory; improving storage performance with Intel® Optane™ SSDs; increasing affordable storage capacity with Intel® QLC NAND SSDs; and delivering fast, efficient network I/O with Intel® Ethernet adapters.

These platform-wide innovations deliver the scalable, balanced performance required by data-centric workloads, as well as traditional data center workloads.

What are some of the foundational enhancements designed into 2nd Gen Intel® Xeon® Scalable processors and platforms?

2nd Gen Intel® Xeon® Scalable processors include many innovations and enhancements, including:

- **World-class per-core performance** – Boosts compute performance across many workloads
- **Greater memory bandwidth and capacity** – Improves performance and flexibility across data-centric and memory-bound workloads
- **Expanded I/O** – Accelerates throughput for I/O-intensive workloads
- **Intel® Deep Learning Boost with Vector Neural Network Instructions (VNNI)** – New processor instructions that significantly increase AI inference performance
- **Support for Intel® Optane™ DC persistent memory, Intel® Optane® DC SSDs and Intel® QLC 3D NAND SSDs** – Innovations that deliver new levels of performance and capacity across the memory and storage pyramid
- **Intel® Infrastructure Management Technologies** – Enable enhanced monitoring and control of resources to improve data center efficiency and utilization
- **Intel® Security Libraries for Data Center** – Software libraries and components that enable developers to more easily tap into Intel hardware security features to better protect platforms and data

What is Intel® Deep Learning Boost with Vector Neural Network Instructions (VNNI) and why is it valuable?

Deep learning techniques have become transformative across many use cases. And accelerating inference tasks is a critical requirement to maximize the impact of deep learning—enabling high-value use cases such as speech or facial recognition, motion and threat detection, machine vision, predictive analytics and more.

The new Intel® Vector Neural Network Instructions greatly accelerate deep learning inference—a common requirement across deep learning workloads. As a result of this feature, inference performance can be improved up to 30x[1] on an Intel® Xeon® Platinum 9200 processor compared to an Intel® Xeon® Platinum 8180 processor (July 2017).

How do 2nd Gen Intel® Xeon® Scalable processor-based platforms boost memory performance for data-centric workloads?

There are two enormous advantages of 2nd Gen Intel® Xeon® Scalable processor-based servers on the system memory front—and those relate to performance and capacity.

First, the latest Intel® processor-based servers include up to 12 memory channels per processor, to deliver the highest native memory bandwidth of any Intel® Xeon® processor platform.

And second, you can now deploy platforms with support for Intel® Optane™ DC persistent memory, which provides greater memory density per DIMM socket than DRAM—enabling up to 2x memory capacity when combining DRAM and Intel® Optane™ DC Persistent Memory (vs. DRAM only).

These features enable faster memory I/O and more data to be placed in system memory, closer to the processor. That means faster time to insight and support for larger in-memory databases for today's most demanding data-centric workloads.

How do 2nd Gen Intel® Xeon® Scalable processors enhance security?

The key to providing greater security is better protecting both your data and your platforms.

2nd Gen Intel® Xeon® Scalable processors include Intel® Trusted Execution Technology (Intel® TXT), which provides hardware-enhanced capabilities to help ensure that servers are in a known, trusted state[2]. Other Intel hardware features accelerate key and data encryption for data at rest, in-use, and in-flight.

What is Intel® Optane™ DC persistent memory and why is it valuable?

For decades, server platforms have been capacity-constrained in terms of system memory due to limitations on DRAM density.

Intel® Optane™ DC persistent memory is a true technological breakthrough because it has much greater density than DRAM while also being more affordable. Using Intel® Optane™ DC persistent memory in place of DRAM in some of a 2nd Gen Intel® Xeon® Scalable processor-based server's DIMMs, you can achieve 2x memory capacity more cost-efficiently.

Intel® Optane™ DC persistent memory completely rewrites the rules for large in-memory databases, and it's supported only by 2nd Gen Intel® Xeon® Scalable processor-based 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 www.intel.com/benchmarks.

Performance results are based on testing as of dates shown in configuration and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance.

Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice Revision #20110804

The benchmark results may need to be revised as additional testing is conducted. The results depend on the specific platform configurations and workloads utilized in the testing, and may not be applicable to any particular user's components, computer system or workloads. The results are not necessarily representative of other benchmarks and other benchmark results may show greater or lesser impact from mitigations.

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.

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.

1 Up to 30X AI performance with Intel® Deep Learning Boost (Intel DL Boost) compared to Intel® Xeon® Platinum 8180 processor (July 2017). Tested by Intel as of 2/26/2019. Platform: Dragon rock 2 socket Intel® Xeon® Platinum 9282(56 cores per socket), HT ON, turbo ON, Total Memory 768 GB (24 slots/ 32 GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0241.112020180249, Centos* 7 Kernel 3.10.0-957.5.1.el7. x86_64, Deep Learning Framework: Intel® Optimization for Caffe* version: https://github.com/intel/caffe d554cbf1, ICC 2019.2.187, MKL DNN version: v0.17 (commit hash: 830a10059a018cd-2634d94195140cf2d8790a75a), model:https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=64, No datalayer DummyData: 3x224x224, 56 instance/2 socket, Datatype: INT8 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: (https://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs fromhttps://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50),. Intel C++ compiler ver. 17.0.2 20170213, Intel® Math Kernel Library (Intel® MKL) small libraries version 2018.0.20170425. Caffe run with "numactl -l".

2 No computer system can provide absolute security under all conditions. Intel® Trusted Execution Technology (Intel® TXT) requires a computer with Intel® Virtualization Technology, an Intel TXT-enabled processor, chipset, BIOS, Authenticated Code Modules and an Intel TXT-compatible measured launched environment (MLE). Intel TXT also requires the system to contain a TPM v1.s. For more information, visit www.intel.com/technology/security