2ND GENERATION
INTEL® XEON® SCALABLE PROCESSORS
Accelerate IoT and artificial intelligence workloads

2nd Generation Intel® Xeon® Scalable processors with built-in Intel® Deep Learning Boost deliver artificial intelligence (AI) performance for the next era of data-driven edge platforms and compute.

PERFORMANCE AND EDGE INTELLIGENCE TO TRANSFORM YOUR BUSINESS

By 2020, **45%** of all data created by IoT devices will be stored, processed, analyzed, and acted upon close to or at the edge of a network.[1]

UP TO 14X inference throughput performance[2] (vs. previous-generation Intel® Xeon® Scalable processors at launch)

UP TO 1.33X average[3] generational gains on Intel® Xeon® Gold processor mainstream SKUs (vs. previous generation)

UP TO 37% in integer throughput performance[4] (vs. previous generation)

UP TO 36% increase in floating point throughput performance[4] (vs. previous generation)

UP TO 34% increase in LINPACK* performance[4] (vs. previous generation)

Now you can get improved deep learning capabilities, speed deployment, and lower TCO—simply with one integrated CPU with AI workloads optimized by the Intel® Distribution of OpenVINO™ toolkit.

Built-in AI acceleration with Intel Deep Learning Boost for inference-based applications, including image recognition, object detection, speech recognition, language translation, and more.

Speed IoT and AI application development with Intel Distribution of OpenVINO toolkit.

Consolidate workloads on a single system without performance degradation.

Enhance security with hardware mitigation for side-channel exploits.

Increase memory capacity and bandwidth with Intel® Optane™ DC persistent memory for affordable, persistent, and large memory.

Enable embedded use cases with 10-year use case reliability to meet stringent requirements.[5]

Maximize platform investments: Upgrade existing Intel Xeon Scalable processors without the need for additional infrastructure.

SMART CITIES
- Public safety
- Traffic pattern monitoring
- Smart parking

HEALTH AND LIFE SCIENCES
- Accelerate high-end imaging systems (MRI, CT scans)
- Improve accuracy
- Reduce patient exposure to radiation by reducing image captures

INDUSTRIAL MANUFACTURING
- Accelerate Industrial IoT (IIoT)
- Industrial machine vision
- Defect detection and quality inspection
- Consolidate workloads

EDUCATION
- Deploy interactive whiteboards
- Integrate Visual Data Devices (VDDs)
- Support in-class interactions
- Broadcast lessons to remote sites

BANKING
- Enable edge and back-end transaction processing
- Accelerate complex data analysis

RETAIL
- Inform and streamline operations
- Personalize shopping
- Capture data, such as store traffic patterns
- Gender and age classification for targeted retailing
- Improve customer service

MOVE YOUR IOT AND AI WORKLOADS FORWARD

Try **2nd Generation Intel Xeon Scalable processors ›**

Visit the **Resource and Design Center ›**

Find out more about **Intel Deep Learning Boost ›**

Download the **Intel Distribution of OpenVINO toolkit ›**

Explore the **Intel® Vision Products portfolio ›**

1. https://innovationatwork.ieee.org/how-the-edge-computing-layer-helps-with-latency/.

2. **1x inference throughput improvement on Intel® Xeon® Platinum 8180 processor (July 2017) baseline:** 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, synthetic 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". **14x inference throughput improvement on Intel® Xeon® Platinum 8280 processor with Intel® DL Boost:** Tested by Intel as of 2/20/2019. 2 socket Intel® Xeon® Platinum 8280 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0271.120720180605 (ucode: 0x200004d), Ubuntu 18.04.1 LTS, kernel 4.15.0-45-generic, SSD 1x sda INTEL SSDSC2BA80 SSD 745.2GB, nvme1n1 INTEL SSDPE2KX040T7 SSD 3.7TB, Deep Learning Framework: Intel® Optimization for Caffe version: 1.1.3 (commit hash: 7010334f159da247db3fe3a9d96a3116ca06b09a), ICC version 18.0.1, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a, model https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=64, syntheticData, 4 instance/2 socket, Datatype: INT8 vs. Tested by Intel as of July 11, 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, synthetic 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".

3. Geomean of est. SPECrate2017_int_base, est. SPECrate2017_fp_base, Stream Triad, Intel Distribution of Linpack, server side Java. Gold 5218 vs Gold 5118: 1-node, 2x Intel® Xeon® Gold 5218 CPU on Wolf Pass with 384 GB (12x 32GB 2933 [2666]) total memory, ucode 0x4000013 on RHEL7.6, 3.10.0-957.el7.x86_65, IC18u2, AVX2, HT on all (off Stream, Linpack), Turbo on, result: est. int throughput=162, est. fp throughput=172, Stream Triad=185, Linpack=1088, server side java=98333, test by Intel on 12/7/2018. 1-node, 2x Intel® Xeon® Gold 5118 CPU on Wolf Pass with 384 GB (12x 32GB 2666 [2400]) total memory, ucode 0x200004D on RHEL7.6, 3.10.0-957.el7.x86_65, IC18u2, AVX2, HT on all (off Stream, Linpack), Turbo on, result: est. int throughput=119, est. fp throughput=134, Stream Triad=148.6, Linpack=822, server side java=67434, test by Intel on 11/12/2018.

4. 1.27x average generational gains on Silver 4210 vs. Silver 4110: 1-node, 2x Intel® Xeon® Silver 4210 CPU on Wolf Pass with 384 GB (12x 32GB 2933 [2666]) total memory, ucode 0x400000A on RHEL7.6, 3.10.0-957.el7.x86_65, IC18u2, AVX2, HT on all (off Stream, Linpack), Turbo on, score: est. int throughput=98, est. fp throughput=114, Stream Triad=86, Linpack=678.35, server side java=55537, test by Intel on 1/28/2019. 1-node, 2x Intel® Xeon® Silver 4110 CPU on Wolf Pass with 384 GB (12x 32GB 2666 [2400]) total memory, ucode 0x200004D on RHEL7.6, 3.10.0-957.el7.x86_65, IC18u2, AVX2, HT on all (off Stream, Linpack), Turbo on, score: est. int throughput=71.6, est. fp throughput=84, Stream Triad=81, Linpack=506.925, server side java=44338, test by Intel on 11/19/2018.

5. IOTG product availability: 15-year availability spans from Intel Corporation introduction of component family (initial Intel launch date) to last shipments. Component family introduction dates are available at https://ark.intel.com.

Performance results are based on testing as of February 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

Software and workloads used in performance tests may have been optimized for performance only on Intel microplatforms. 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 intel.com/benchmarks.

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.

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 computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com/iot.

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.