PRODUCT BRIEF

Internet of Things
2nd Generation Intel® Xeon®
Scalable Processors

Introducing 2nd Generation Intel® Xeon® Scalable Processors

Your IoT platform for the data-driven era

IoT is reshaping business models and industries based on billions of connected things, the exponential increase of data, much of it video, and the insights driven by AI. According to a study conducted by IDC, 45 percent of all data created by IoT devices will be stored, processed, analyzed, and acted upon close to or at the edge of a network by 2020.[1]

2nd Generation Intel® Xeon® Scalable processors with built-in Intel® Deep Learning Boost deliver advanced AI performance for the next era of data-driven edge platforms and compute. These efficient systems deliver powerful new opportunities and competitive differentiators through workload consolidation and deep learning inference performance that cannot be realized without greater processing power and intelligence at the edge.

Intel Deep Learning Boost extends the instruction set to significantly accelerate inference performance for deep learning workloads, including object detection, image classification, and more. 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.

With the Intel Distribution of OpenVINO toolkit, developers can easily unlock the power of the deep learning capabilities to create AI and IoT applications powered by 2nd Generation Intel Xeon Scalable processors. Developers can create deep learning training models that take advantage of Intel Deep Learning Boost through using Intel-optimized frameworks (such as TensorFlow*, Caffe*, MXNet*, BigDL*, and more). With the Intel Distribution of OpenVINO toolkit, developers can streamline deep learning inference deployment on 2nd Gen Intel Xeon Scalable processors with the performance benefits of Intel Deep Learning Boost. This is in addition to the performance benefits developers have already been seeing with the Intel Distribution of OpenVINO toolkit on other Intel® platforms.

With the capabilities for AI workloads, 2nd Generation Intel Xeon Scalable processors are enabling fast, remote digital security response throughout a city; rapid identification and classification of medical imaging data; personalized retail shopping experiences and automated checkout; and defect detection for manufacturing.

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%
increase 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)

Key benefits for IoT

- **Built-in AI:** Gain key insights and harness the power of data with Intel Deep Learning Boost.
- **Improve performance:** Process more data and improve responsiveness at the edge through higher processing power.
- **Consolidate workloads:** Aggregate various workloads on a single system without degrading performance and helping to ensure deterministic behavior with Intel® Resource Director Technology (Intel® RDT).
- **Enhance security:** Hardware mitigation for side-channel exploits helps protect systems and data by hardening the platform against any malicious attacks.
- **Greater memory capacity and bandwidth:** Speed workloads and time to insight with Intel® Optane™ DC persistent memory, a new, revolutionary memory product for affordable, persistent, and large memory.
- **Enable embedded use cases:** Offers SKUs with 10-year use case reliability to meet stringent conditions for embedded use cases.[5]
- **Gain flexibility:** Rich set of I/Os supports PCIe*, USB, and SATA* on a single platform to connect varied peripherals such as networking (Ethernet), storage (SSDs), and accelerators (FPGAs).
- **Maximize platform investments:** Upgrade existing Intel Xeon Scalable processors with higher performance and new capabilities, without the need for additional infrastructure.

SAMPLE INFERENCE WORKLOADS
Object detection Image recognition Speech recognition Language translation

SAMPLE APPLICATIONS
Medical imaging workflow automation Medical imaging image reconstruction Surveillance and security Personalized shopping Predictive maintenance Machine vision

Intel Deep Learning Boost

Intel Deep Learning Boost provides low-precision integer operations to significantly accelerate performance for deep learning inference applications including image recognition, object detection, speech recognition, language translation, and more.

The dramatic performance improvement and efficiency, which required three separate instructions in previous-generation processors, is now delivered by using a single int8 instruction for deep learning inference applications. Intel Deep Learning Boost delivers up to 14x inference throughput performance[2] (vs. previous-generation Intel Xeon Scalable processors at launch). Software optimizations are achieved via the Intel Distribution of OpenVINO toolkit and Intel software frameworks.

Target workloads

Vision

AI at the edge is opening up new possibilities in every industry. With 2nd Generation Intel Xeon Scalable processors with Intel Deep Learning Boost, businesses can take advantage of near-real-time insights and make better decisions, faster.

The Intel Distribution of OpenVINO toolkit, in combination with this new platform, significantly accelerates performance for deep learning inference applications at the edge, including traffic pattern monitoring, public safety and emergency response, industrial machine vision, medical imaging, and retail inventory control.

Now, data generated from multiple cameras can be efficiently analyzed and stored by network video recorders (NVRs) or servers at the edge or transmitted to back-end servers and storage—all with improved accuracy and speed.

Health and Life Sciences

As image resolution improves and additional relevant data is embedded or attached, medical image file sizes are growing rapidly, with many images easily at 1 GB or more. When high-end imaging systems (PET, MRI, CT, X-rays) gain the performance boost of the 2nd Generation Intel Xeon Scalable processor, they can better accelerate time-to-image, automate image analysis, provide guidance improving image accuracy, reduce patient exposure to radiation by reducing the number of image captures needed, and enable the personalization of medicine in entirely new ways.

2nd Generation Intel Xeon Scalable processors have access to the entire system memory space, accelerating complex, hybrid workloads, including the larger, memory-intensive models typically found in today's advanced medical imaging, and avoiding the memory constraints inherent to discrete add-in cards.

These processors offer an affordable, flexible platform for AI models, particularly in conjunction with tools like the Intel Distribution of OpenVINO toolkit, which can help accelerate the deployment of high performance computer vision and deep learning inference into medical applications, delivering the right information with the right analysis at the right time to aid the speed and accuracy of time-critical diagnosis and decision-making unique to healthcare.

Education

K–12 and higher education are in transition, as digital platforms and peripherals allow for diverse learning and teaching models—from collaboration and distance learning to highly interactive technologies supporting classrooms and remote venues.

Education deployments include interactive whiteboards that integrate teacher and student PCs with cameras to support in-class interactions and broadcast lessons to remote sites or regions.

Visual Data Devices (VDDs) powered by 2nd Generation Intel Xeon Scalable processors deliver content while performing analytics. Student dashboards with data on student test results, progress, and social interactions during class allow parents and educators to help customize education for each student.

Banking and retail

Industries that depend on highly transactional workloads, such as point of sale (POS), banking, and retail transactions, need greater performance and flexibility to handle the data loads, regulatory requirements, and customer expectations of the digital economy and to enable new services and business models.

Banking industry applications for 2nd Generation Intel Xeon Scalable processors with Intel Deep Learning Boost include both edge and back-end transaction processing and acceleration of complex data analysis.

Retailers can use 2nd Generation Intel Xeon Scalable processor-based solutions to inform and streamline operations. This technology allows retailers to capture a transaction and analyze the data, on premise or in the cloud.

Industrial manufacturing

2nd Generation Intel Xeon Scalable processors with Intel Deep Learning Boost bring the performance and capabilities required to accelerate Industrial IoT (IIoT), helping manufacturers increase performance, use machine vision for defect detection and quality inspection, and consolidate workloads.

The platforms are optimized to deliver the high performance needed for the most demanding computational tasks in IoT and offer options for 10-year use case reliability as well as 15-year production availability.[5]

Accelerate IoT application development

Deep learning revenue is estimated to grow from USD 655 million in 2016 to USD 35 billion by 2025.[6] Integration of computer vision, deep learning, and analytics processing capabilities into applications can help turn data into insights. The opportunity is enormous, with demand growing for intelligent vision applications.

With Intel Distribution of OpenVINO toolkit, software developers and data scientists working on vision solutions can unlock the deep learning capabilities to create AI-based IoT applications powered by 2nd Generation Intel Xeon Scalable processors.

Conclusion

Run complex IoT workloads on the same hardware as your existing workloads while taking AI performance to the next level. With 2nd Generation Intel Xeon Scalable processors with Intel Deep Learning Boost, the data-driven era is your opportunity to move, store, and process more.

Try 2nd Generation Intel Xeon Scalable processors and assess the impact on your IoT workloads, applications, and use cases.

Visit the Resource and Design Center ›

Download the Intel Distribution of OpenVINO toolkit and ramp your applications for next-generation computer vision for IoT and AI.

2ND GENERATION INTEL® XEON® SCALABLE PROCESSOR SKUS

PRODUCT	NUMBER OF CORES	BASE NON-AVX CPU FREQUENCY (GHZ)	INTEL® OPTANE™ DC PERSISTENT MEMORY	POWER/ TDP (W)	ROBUST THERMAL PROFILE (HIGH TCASE) AND 10-YEAR RELIABILITY	ADVANCED/ STANDARD RAS[1]
Intel® Xeon® Gold 6238T Processor	22	1.9	√	125	√	A
Intel® Xeon® Gold 6230 Processor	20	2.1	√	125	-	A
Intel® Xeon® Gold 6226 Processor	12	2.7	√	125	-	A
Intel® Xeon® Gold 5218T Processor	16	2.1	√	105	√	A
Intel® Xeon® Gold 5215 Processor	10	2.5	√	85	-	A
Intel® Xeon® Silver 4216 Processor	16	2.1	-	100	-	S
Intel® Xeon® Silver 4215 Processor	8	2.5	√	85	-	S
Intel® Xeon® Silver 4214 Processor	12	2.2	-	85	-	S
Intel® Xeon® Silver 4210 Processor	10	2.2	-	85	-	S
Intel® Xeon® Silver 4209T Processor	8	2.2	-	70	√	S

1. **A** = Advanced RAS; **S** = Standard RAS.

SUPPORTED SOFTWARE

OS TYPE	OPERATING SYSTEM (TARGETED FOR SUPPORT)^	SUPPORT^^	DISTRIBUTION	BIOS
Linux*	Red Hat* Enterprise Linux* 7.5	Red Hat		American Megatrends, Inc. Insyde Software Phoenix Technologies BYOSOFT
	SUSE* Linux Enterprise Server 12 SP4, 15	SUSE, Open Source	SUSE	
	Ubuntu* 18.04 LTS	Canonical, Open Source	Canonical	
	Yocto* Linux v4.19.8	Intel, Open Source	Yocto Project*	
	FreeBSD 11.2	Open Source Community		
	Fedora*	Open Source Community		
	CentOS	Open Source Community		
Windows*	Microsoft Windows Server 2016 Microsoft Windows Server 2019 LTSC Microsoft Windows Server RS3, RS4, RS5 (Core/Nano)	Intel, Microsoft	Microsoft	
VMM	Linux KVM	Open Source Community		
	VMware ESXi* 6.0 u3, 6.5	VMware, Open Source		
	Microsoft Windows* Hyper-V	Microsoft		
	Xen* 4.10, 4.11	Open Source Community		

^ This is the OS list that is tested internally and does NOT reflect the OS vendor support for these exact release versions. Please contact the respective OS vendor(s) for the release version numbers and support. Several software patches will be upstreamed and will be picked up over time. These will be required to enhance platform support.

^^ Intel only provides support for our tools, patches, and utilities on the OS. Actual OS support should come from the OS vendor.

Learn more

Explore 2nd Gen Intel Xeon Scalable processors ›

Find out more about Intel Deep Learning Boost ›

Explore the Intel Distribution of OpenVINO toolkit ›

Explore the Intel Vision Products portfolio ›

Explore Intel solutions for IoT ›

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.
6. https://software.intel.com/en-us/blogs/2018/05/16/openvino-toolkit-accelerates-cv-development-across-intel-platforms.

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.

0419/MW/CMD/PDF 338824-001US