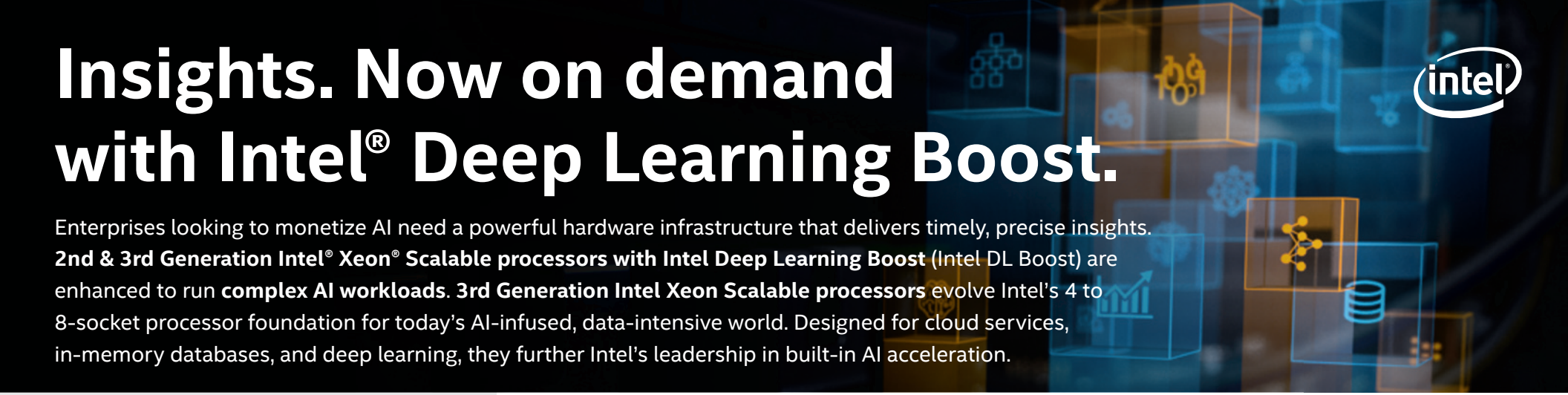
Insights. Now on demand with Intel® Deep Learning Boost.



Enterprises looking to monetize AI need a powerful hardware infrastructure that delivers timely, precise insights. **2nd & 3rd Generation Intel® Xeon® Scalable processors with Intel Deep Learning Boost** (Intel DL Boost) are enhanced to run **complex AI workloads**. **3rd Generation Intel Xeon Scalable processors** evolve Intel's 4 to 8-socket processor foundation for today's AI-infused, data-intensive world. Designed for cloud services, in-memory databases, and deep learning, they further Intel's leadership in built-in AI acceleration.

When to recommend

Talk about Intel Xeon Scalable processors with Intel DL Boost to customers who want exceptional AI performance with lower memory requirements, enabling their hardware footprint to do more.

Customer pain points

- Data-center bottlenecks are obstructing **real-time intelligence**
- Computational capacity is an issue, like in **convolutional neural networks** (CNNs) and **deep neural networks** (DNNs)
- Low-latency hardware solutions are needed to drive inference at scale

Ideal for

- Image classification
- Object detection
- Language translation
- Speech recognition

Why upgrade

The evolution of **Intel DL Boost** in 3rd Generation Intel Xeon processors makes it the first general-purpose server CPU to offer **built-in bfloat16 instructions** to accelerate deep learning training and inference—without greater complexity or loss of accuracy. This addition makes Intel DL Boost even more broadly useful for a range of AI applications. Bfloat16 extends the Intel DL Boost feature set, joining today's INT8-based Vector Neural Network Instructions (VNNI).

Accelerated insights

- Up to 2x single-precision FLOPs vs. FP32 datatype (theoretical max)
- Up to 30x improved deep learning performance vs. previous generations[1]

Optimized frameworks & libraries

Caffe

PaddlePaddle

mxnet

TensorFlow

Deep Neural Network Library (oneDNN)

*Other names and brands may be claimed as the property of others.

Say this to your customer

"The Intel Xeon Scalable platform introduces a common platform for AI with high throughput for both inference and training, so you can **do both without purchasing a GPU**."[2]

"Intel is partnering with developers to continue optimizing popular frameworks and libraries to further accelerate inference performance."

"Intel DL Boost unlocks insights by optimizing systems for **impactful automation**. Imagine the efficiency you can deliver to your business by **no longer having to purchase dedicated hardware** to uncover the data you need."

"Based on Intel Advanced Vector Extensions 512 (Intel AVX-512), VNNI delivers significant performance improvement by **combining three instructions into one**. How would your organization benefit from this **dramatic increase in efficiency**?"

Hardware and storage innovation

Added value for deep learning

Intel Xeon Scalable processors with Intel DL Boost have integrated features and technologies designed to enhance both training and inference in deep learning.

Brain floating-point Format (bfloat16)

This number-encoding format, available on 3rd Generation Intel Xeon Scalable processors, delivers efficiency for workloads that have **high compute intensity** but **lower need for precision**, making mainstream AI training widely deployable on more CPUs.

Intel® Optane™ persistent memory

Lower latency and more memory closer to the CPU enable larger in-memory working datasets and persistence across power cycles.

Intel Optane Solid State Drives

Bigger, more affordable datasets and application acceleration help enterprises take advantage of next-level insights.

UP TO 2.1X

TRAINING PERFORMANCE IMPROVEMENT
over baseline on 2nd Gen Intel Xeon Gold 6248 processor.[3]

UP TO 60%

PERFORMANCE INCREASE FOR TRAINING COMPARED TO FP-32[4]

Help businesses deliver AI readiness across the data center with Intel Xeon Scalable processors featuring Intel Deep Learning Boost. Contact your Intel Authorized Distributor or visit **ai.intel.com**

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Performance varies depending on system configuration.

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. 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 intel.com/benchmarks.

Intel Advanced Vector Extensions (Intel AVX) provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at intel.com/go/turbo.

1. Configurations for "Up to 30X AI performance with 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: 830a10059a018cd2634d94195140cf2d8790a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=64, No datalayer syntheticData: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: (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".

2. "Product Brief: 2nd Generation Intel® Xeon® Scalable Processors." Intel, 2019. https://www.intel.com/content/www/us/en/products/docs/processors/xeon/2nd-gen-xeon-scalable-processors-brief.html

3. Configurations for "Up to 2.1x improvement in training performance over baseline on 2nd Gen Intel Xeon Gold 6248 processor" (Sept. 2019). New: Tested by Intel as of Sept. 2019. 2x Intel Xeon Gold 6248 processor (2 socket), 20 cores HT On, Turbo On. Total Memory: 376GB (6 slots / 32GB / 2666 MTs / DDR4 DIMM), BIOS: SE5C620.86B.02.01.0008.031920191559, Ubuntu 18.04.2 LTS, Deep Learning Framework/Libraries: Python 3.6.8, TensorFlow 1.12.3, Intel MKL 2019.4 for Intel 64 architecture applications, GCC 7.4, Numpy (1.16.4), Topology (Custom Graph Convolutions). Baseline: Tested by Intel as of Sept. 2019. 2x Intel Xeon Gold 6248 processor (2 socket), 20 cores HT On, Turbo On. Total Memory: 376 GB (6 slots / 32GB / 2666 MTs / DDR4 DIMM), BIOS: SE5C620.86B.02.01.0008.031920191559, Ubuntu 18.04.2 LTS, Deep Learning Framework/Libraries: Python 3.6.8, TensorFlow 1.12.3, Numpy (1.16.4), Topology (Custom Graph Convolutions), customer dataset.

4. Configurations for "1x inference throughput improvement in 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, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50),and https://github.com/soumith/convnet-benchmarks/tree/master/-caffe/imagenet_winners (ConvNet benchmarks; files were updated to use newer Caffe prototxt format but are functionally equivalent). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".