Enterprises looking to monetize AI need a powerful hardware infrastructure that delivers timely, precise insights. 2nd Generation Intel® Xeon® Scalable processors with new Intel® Deep Learning Boost (Intel® DL Boost) are enhanced specifically to run performance-hungry AI applications alongside existing cloud and data center workloads.

Keywords: Inference, deep learning, image recognition, object detection, recommendation systems, speech recognition, deep neural network, convolutional neural networks

"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."

"Intel DL Boost works by extending the Intel AVX-512 instruction set to do with one instruction what took three instructions in previous-generation processors. How would your organization benefit from those dramatic increases in efficiency?"

**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
- When computational capacity is an issue, like in convolutional neural networks (CNNs) and deep neural networks (DNNs)
- Where low-latency, low-power hardware solutions are needed to drive inference at scale

**WHY UPGRADE**

**ACCELERATED INSIGHTS**
- Up to 2x faster inference with new int8 instructions vs. previous generations
- Up to 30x improved deep learning performance vs. previous generations

**OPTIMIZED FRAMEWORKS & LIBRARIES**
- Caffe
- PaddlePaddle
- mxnet
- TensorFlow
- INTEL® MKL-DNN

**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."

"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."

"Intel DL Boost works by extending the Intel AVX-512 instruction set to do with one instruction what took three instructions in previous-generation processors. How would your organization benefit from those dramatic increases in efficiency?"

*Other names and brands may be claimed as the property of others.*
HARDWARE AND STORAGE INNOVATION

ACCELERATE INFERENCE PERFORMANCE

2ND GENERATION INTEL XEON PLATINUM PROCESSOR
9200 SERIES with Intel DL Boost

Up to 30X better inference performance on image classification compared to competing processors

2ND GENERATION INTEL XEON PLATINUM PROCESSOR
8200 SERIES with Intel DL Boost

Up to 14X better inference throughput compared to previous-generation technology

ADDED VALUE FOR DEEP LEARNING WITH INTEL® OPTANE™ TECHNOLOGY

Together with the Intel Xeon Scalable processor with Intel DL Boost, Intel® Optane™ technology can enhance both training and inference in deep learning.

INTEL OPTANE DC 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.

BENEFITS

FOR TRAINING

Larger datasets and optimized batch training mean AI solutions can get smarter, faster.

FOR INFERENCE

Larger datasets enable real-time and batch expansion of inference workloads.

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.

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, operating systems, 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.

1. Configurations for (1) “Up to 2x more inference throughput improvement on Intel® Xeon® Platinum 9282 processor with Intel® DL Boost” (2) “Up to 30X AI performance improvement with Intel® Xeon® Platinum 8180 processor” (July 2017). Tested by Intel as of 2/26/2019. Platform: Dragon rock 2 socket Intel® Xeon® Platinum 9282 (36 cores per socket), HT ON, turbo ON, Total Memory 768 GB (32 slots/ 32 GB/ 2933 MHz), BIOS: 0004.001.01, 2 01.02.10.0124T, 11.02.2018.0405, 0.0.19, Open-Computing: 1.1.0, OpenCL: 1.2.0, OpenMP: 3.1, OpenFabrics: 0.0.17 (commit hash: 830a059a1b0b26349e41914025f42b70a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/intel_optimized_models/intel害羞50_int8_full Conv.prototxt, BS=64. No dataloader synthetic. Dataset: Synthetic Data of 12244242, 56 instances/2 socket, Dataset and RTB vs. Intel of as of July 11th 2017. 25 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 7.3.1511, (Core), Linux kernel 3.10.0-162.16.17.x86_64 64 64 Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC). Performance measured with: Environment variables: KMF AFFINITY=granularity-fine, compact, OMP_NUM_THREADS=56, CPU freq set with cpupower freq-set --set -d 2.5G. g-model performance. Caffe: http://github.com/intel/caffe, revision: 9967659762dbf281356905a0a0a7f555b6f2156b05c. Inference measured with “caffe time –forward_only” command, training measured with “caffe time” command. Form “ConvNet” topologies, synthetic dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology species from https://github.com/intel/caffe/tree/master/master_models/intel_optimized_models (phin-Net-50). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.2170425. Caffe ran with “numactl –f”.


3. Configurations for (1) “Up to 14X AI Performance Improvement with Intel® DL Boost compared to Intel® Xeon® Platinum 8180 Processor” (July 2017). Tested by Intel as of 2/26/2019. 2 socket Intel® Xeon® Platinum 8280 Processor, 28 cores. HT ON, Turbo ON. Total Memory 384 GB (24 slots/ 12GB/ 2933 MHz), BIOS: 0004.001.01, 2 01.02.10.0124T, 11.02.2018.0405, 0.0.19, Open-Computing: 1.1.0, OpenCL: 1.2.0, OpenMP: 3.1, OpenFabrics: 0.0.17 (commit hash: 830a059a1b0b26349e41914025f42b70a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/intel_optimized_models/intel害羞50_int8_full Conv.prototxt, BS=64. No dataloader synthetic. Dataset: Synthetic Data of 12244242, 56 instances/2 socket, Dataset and RTB vs. Intel of as of July 11th 2017. 25 Intel® Xeon® Platinum 8280 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 7.3.1511, (Core), Linux kernel 3.10.0-162.16.17.x86_64 64 64 Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC). Performance measured with: Environment variables: KMF AFFINITY=granularity-fine, compact, OMP_NUM_THREADS=56, CPU freq set with cpupower freq-set --set -d 2.5G. -g-perfomance. Caffe: http://github.com/intel/caffe, revision: 9967659762dbf281356905a0a0a7f555b6f2156b05c. Inference measured with “caffe time –forward_only” command, training measured with “caffe time” command. Form “ConvNet” topologies, synthetic dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology species from https://github.com/intel/caffe/tree/master/master_models/intel_optimized_models (phin-Net-50). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.2170425. Caffe ran with “numactl –f”.

Copyright © 2019 Intel Corporation. Intel, the Intel logo, the Intel Inside logo, Intel Optane, and Xeon are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries. *Other names and brands may be claimed as the property of others.