INSIGHTS. NOW ON DEMAND WITH INTEL® DEEP LEARNING BOOST.

intel

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*

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**

IDEAL FOR

- Image recognition
- Object detection
- Recommendation systems
- Speech recognition

WHY UPGRADE

ACCELERATED INSIGHTS

- Up to 2x faster inference with new int8 instructions vs. previous generations[1]
- Up to 30x improved deep learning performance vs. previous generations[1]

OPTIMIZED FRAMEWORKS & LIBRARIES

Caffe

PaddlePaddle

mxnet

TensorFlow

INTEL® MKL-DNN

*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."

"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?"

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[1]

2ND GENERATION INTEL XEON PLATINUM PROCESSOR 8200 SERIES

with Intel DL Boost

Up to 14X better inference throughput

compared to previous-generation technology[3]

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, 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.

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

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 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 14X AI Performance Improvement with Intel® DL Boost compared to Intel® Xeon® Platinum 8180 Processor" (July 2017). 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, DummyData, 4 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, 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). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".