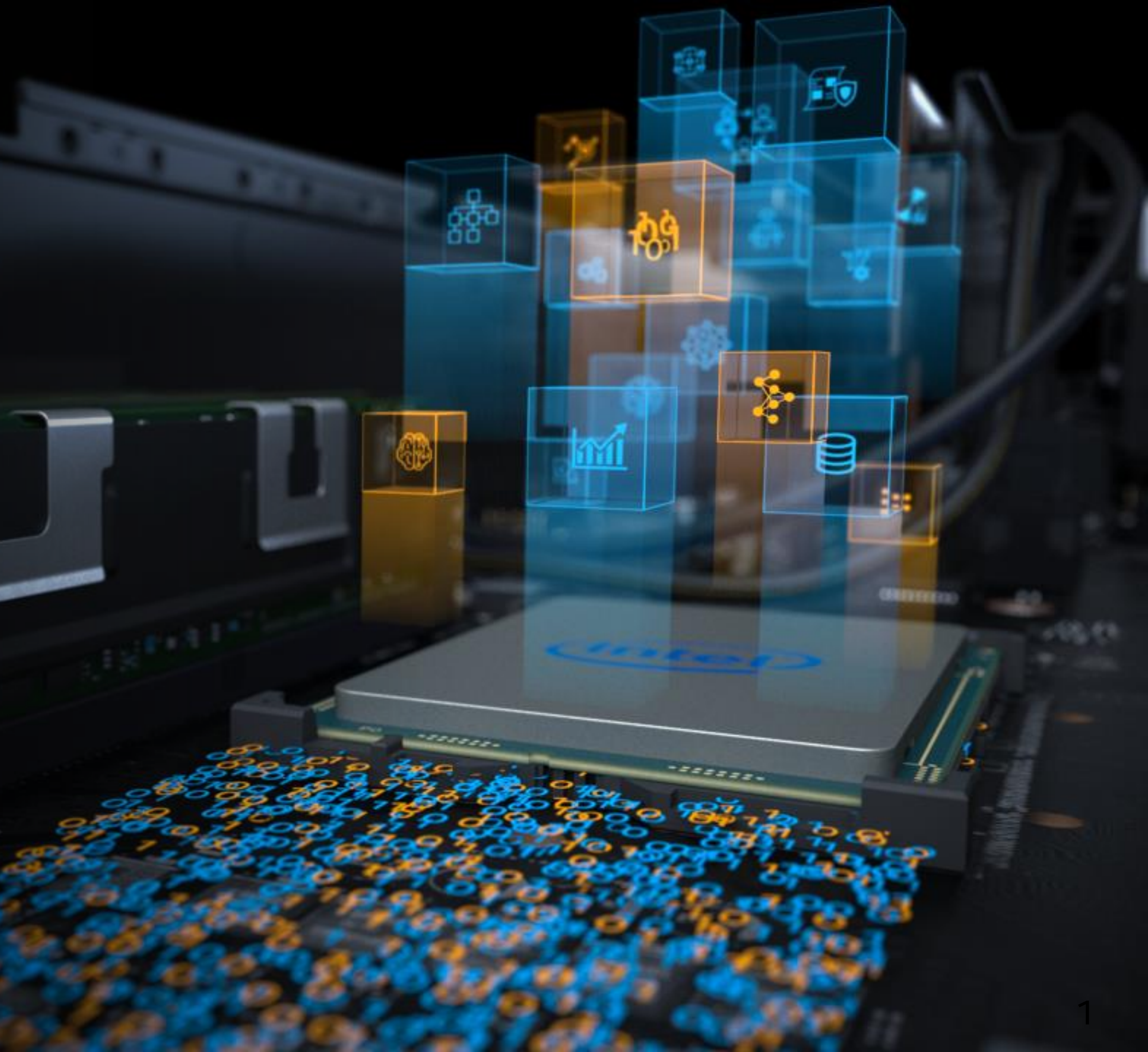
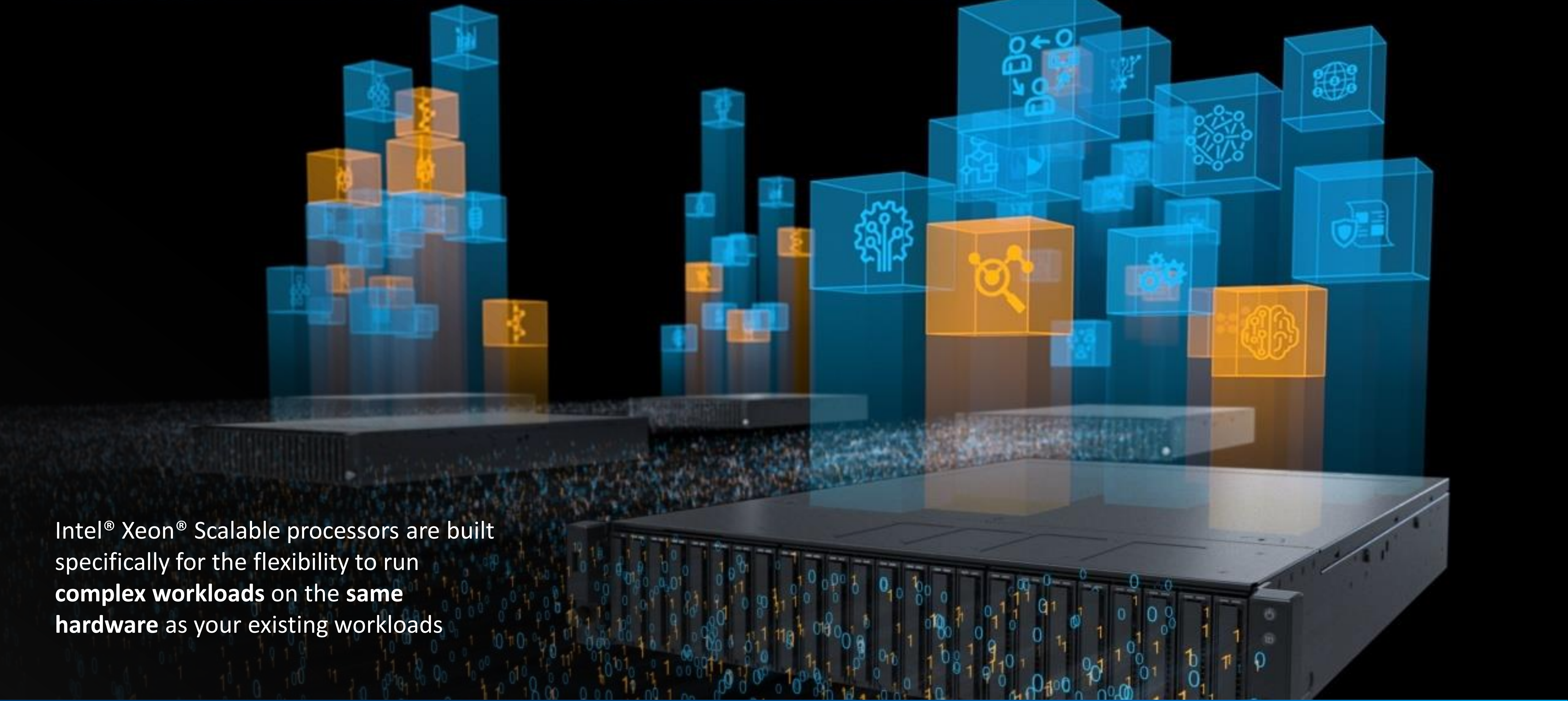
INTEL® DEEP LEARNING BOOST

Built-in acceleration for training and inference workloads



RUN COMPLEX WORKLOADS ON THE SAME PLATFORM



Intel® Xeon® Scalable processors are built specifically for the flexibility to run **complex workloads** on the **same hardware** as your existing workloads

INTEL AVX-512

INTEL DEEP LEARNING BOOST

INTEL VNNI, BFLOAT16

Intel AVX-512

1st, 2nd & 3rd Generation Intel Xeon Scalable Processors

Ultra-wide 512-bit vector operations capabilities with up to two fused-multiply add units and other optimizations accelerate performance for demanding computational tasks.

Intel VNNI

2nd & 3rd Generation Intel Xeon Scalable Processors

Based on Intel Advanced Vector Extensions 512 (Intel AVX-512), the Intel DL Boost Vector Neural Network Instructions (VNNI) delivers a significant performance improvement by combining three instructions into one—thereby maximizing the use of compute resources, utilizing the cache better, and avoiding potential bandwidth bottlenecks.

bfloat16

3rd Generation Intel Xeon Scalable Processors on 4S+ Platform

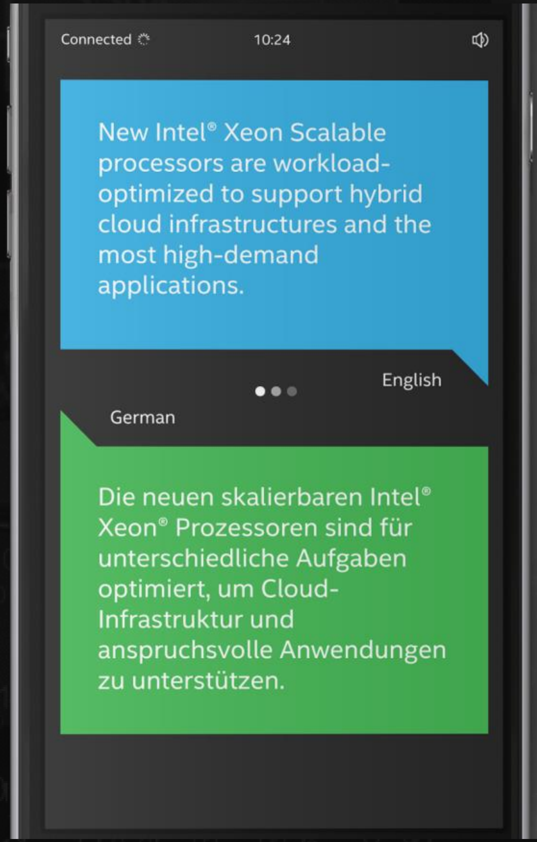
Brain floating-point format (bfloat16 or BF16) is a number encoding format occupying 16 bits representing a floating-point number. It is a more efficient numeric format for workloads that have high compute intensity but lower need for precision.

COMMON TRAINING AND INFERENCE WORKLOADS

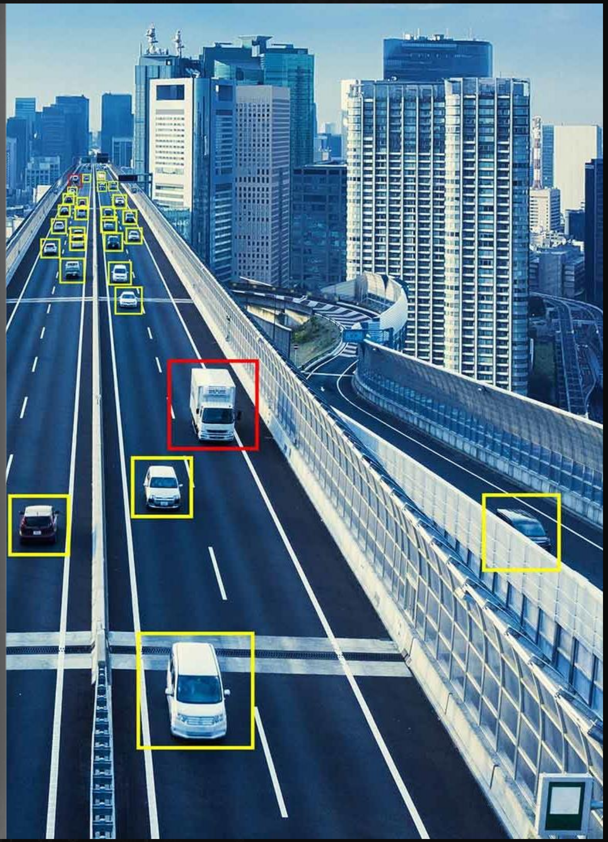
IMAGE CLASSIFICATION

SPEECH RECOGNITION

LANGUAGE TRANSLATION

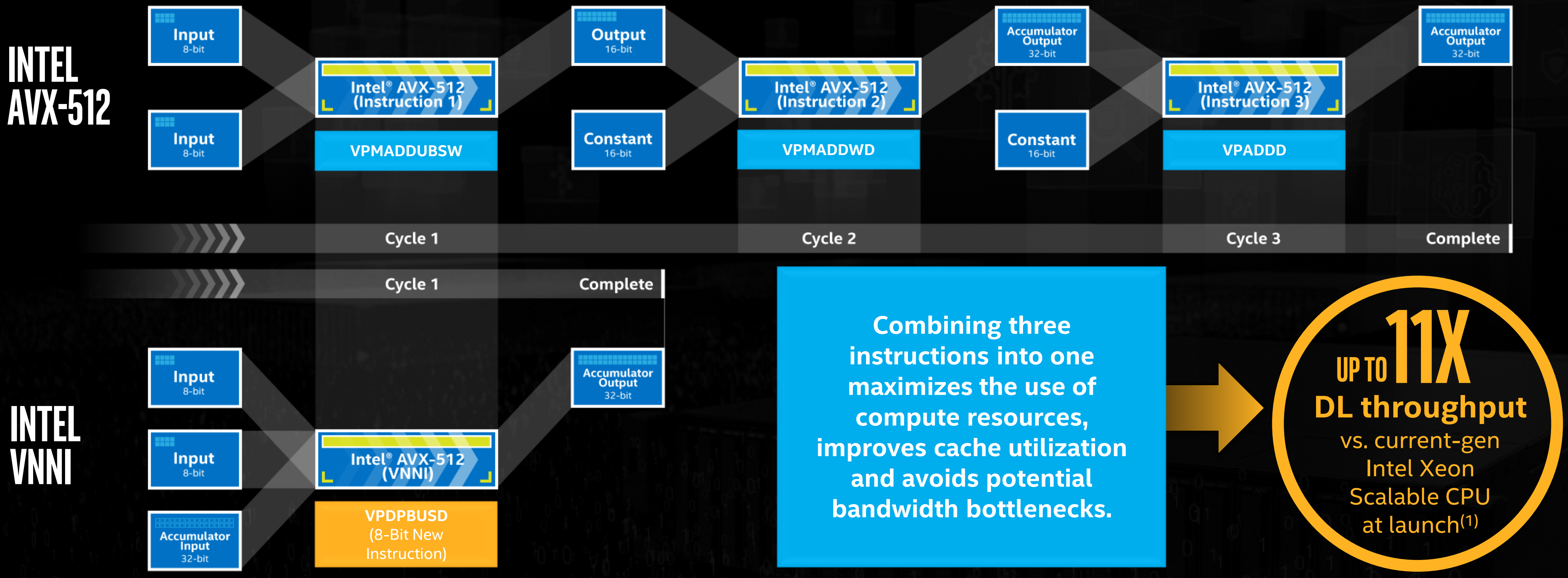


OBJECT DETECTION



INTEL DEEP LEARNING BOOST

A VECTOR NEURAL NETWORK INSTRUCTION (VNNI) EXTENDS INTEL AVX-512 TO ACCELERATE AI/DL INFERENCE



Future Intel Xeon Scalable processor (codename Cascade Lake) results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance vs Tested by Intel as of July 11th 20.17. For more complete information about performance and benchmark results visit www.intel.com/benchmarks.

INTEL DEEP LEARNING BOOST

A VECTOR NEURAL NETWORK INSTRUCTION (VNNI) EXTENDS INTEL AVX-512 TO ACCELERATE AI/DL INFERENCE

PROBLEMS SOLVED

Intel AVX-512

VPMADDUBSW
VPMADDWD
VPADDD

1^{st} Gen Intel Xeon scalable processors

Intel DL Boost (VNNI)

VPDPBUSD
(8-bit new instruction)

More efficient Inference acceleration

2^{nd} Gen Intel Xeon scalable processors

Low Precision Integer Operations

END CUSTOMER VALUE

Designed to accelerate AI/Deep Learning use cases (image classification, object detection, speech recognition, language translation and more)

Animation & whitepaper: **https://ai.intel.com/intel-deep-learning-boost**

INTRODUCING BRAIN FLOATING-POINT FORMAT WITH 16 BITS (BFLOAT16)

- Floating Point 32 (FP32) provides high precision based on the number of bits used to represent a number
- Many AI functions do not require the level of accuracy provided by FP32
- Bfloat16 supports the same range of numbers based on the same exponent field but with lower precision
- Conversion between bfloat16 and FP32 is simpler than FP16
- Twice the throughput per cycle can be achieved with bfloat16 when comparing FP32

Example:
Number: 0.56580972671508789062596
As FP32: 0.565809726715087890625

32 Bits

FP32: 0 | 0 1 1 1 1 1 1 0 | 0 0 1 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 0 1 0 0 0

Simple conversion from BF16 to FP32

bfloat16: 0 | 1 1 1 1 1 1 1 0 | 0 0 1 0 0 0 0

Bfloat16 has the same number of exponent bits in can represent numbers as large as FP32, but with less precision – fewer bits to store actual number

FP16: 0 | 0 1 1 1 0 | 0 0 1 0 0 0 0 1 1 0

16 Bits

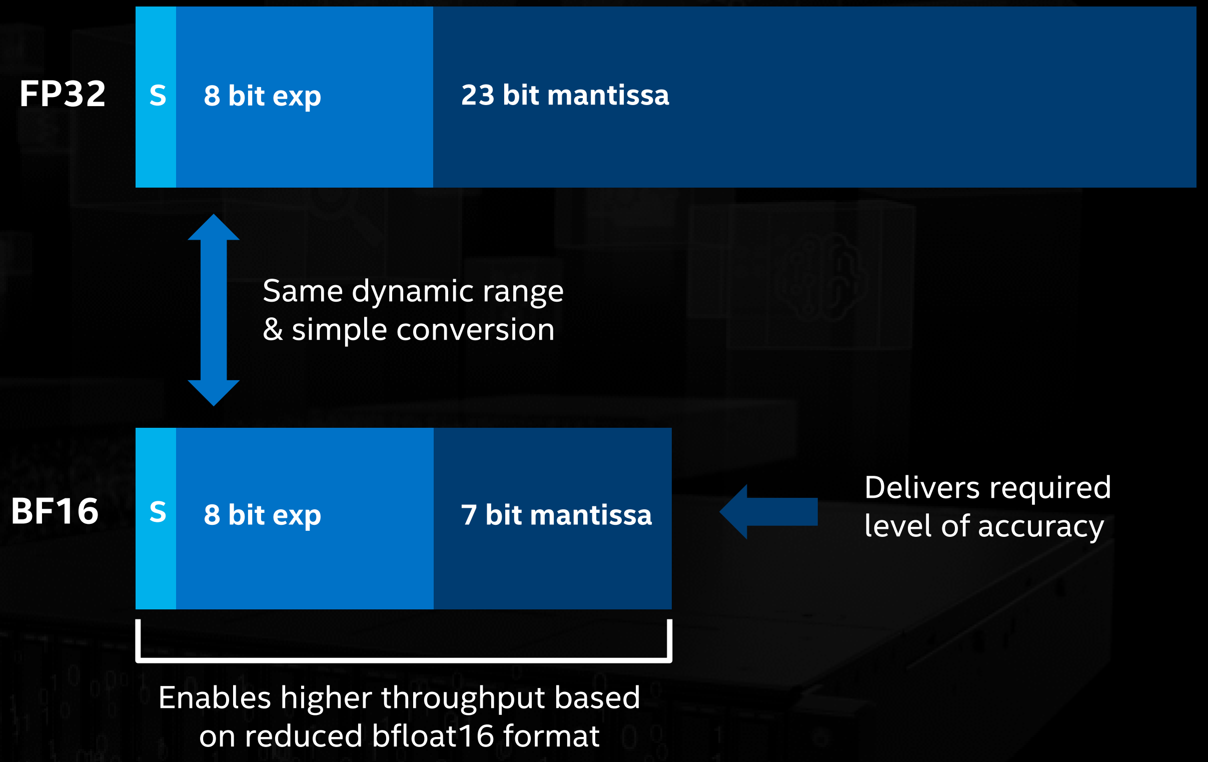
FP16 can provide a more accuracy than bfloat16 but cannot support the same range of numbers due to fewer bits for exponent detail.

- Sign – Indicates positive or negative number
- Exponent – Indicates the position of the decimal point in the fraction/mantissa bits
- Fraction/Mantissa – Bits used to store the "number"

INCREASE TRAINING AND INFERENCE THROUGHPUT USING BFLOAT16

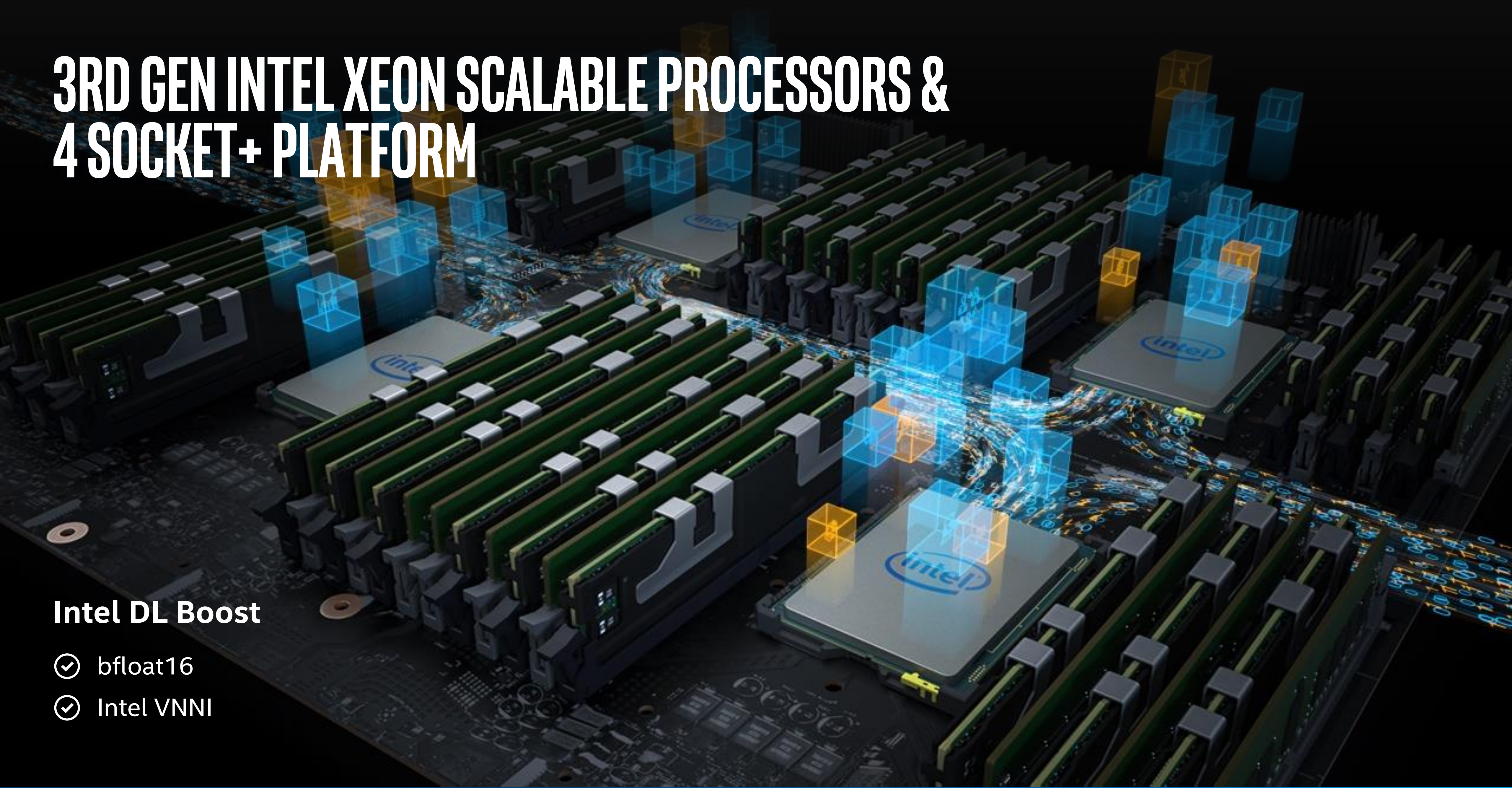
AVAILABLE ON 3RD GEN INTEL XEON SCALABLE PROCESSORS ON 4S+ PLATFORM

- Training & Inference Acceleration
- Native support for bfloat16 datatype
- 2x bfloat16 peak throughput/cycle vs. fp32
- Improved throughput and efficiencies
- Seamless integration with popular AI frameworks



New Built-in AI-acceleration capabilities in select 3rd Generation Intel® Xeon® Scalable Processors targets higher training and inference performance with the required level of accuracy

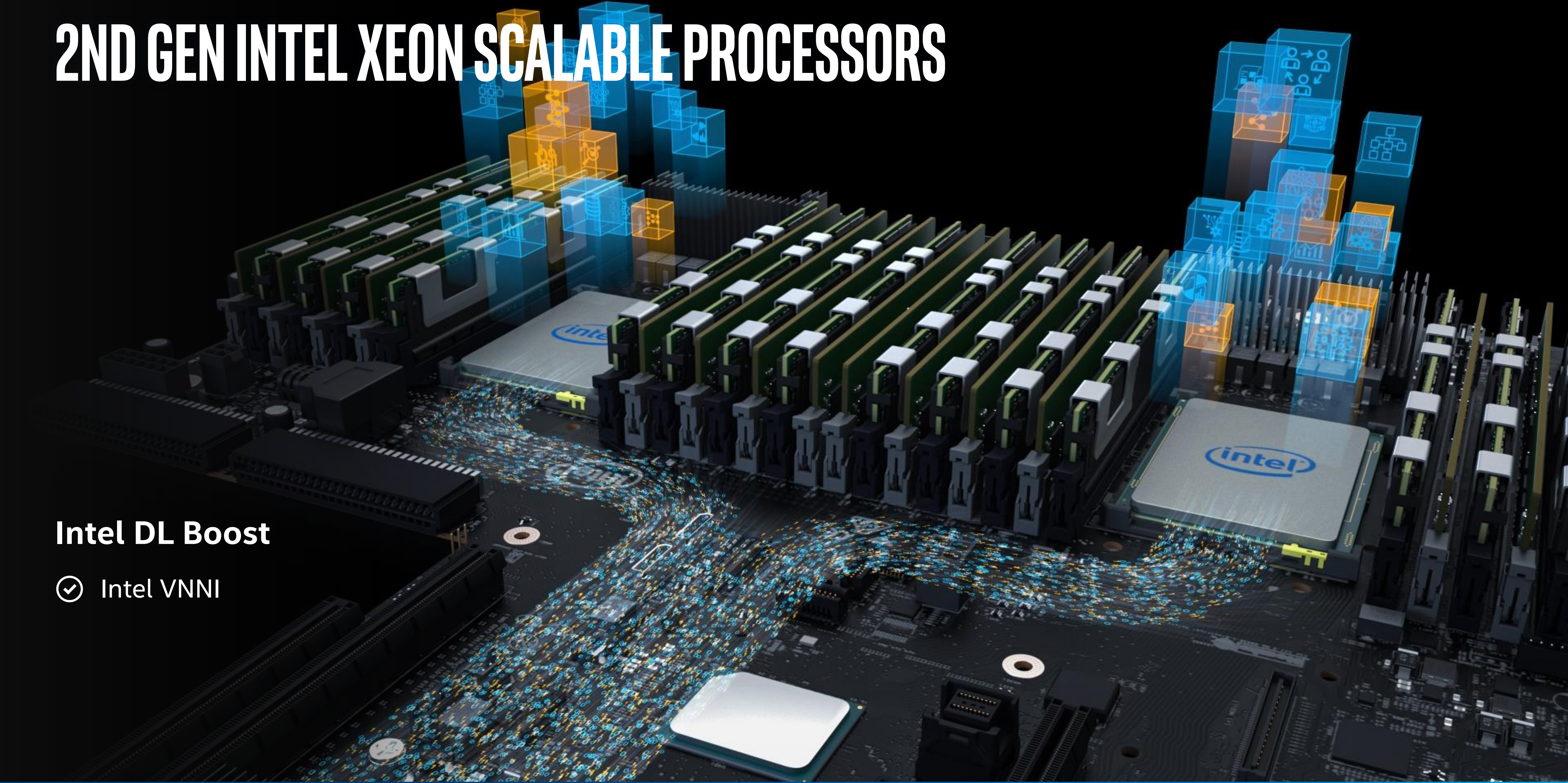
3RD GEN INTEL XEON SCALABLE PROCESSORS & 4 SOCKET+ PLATFORM



Intel DL Boost

- ✓ bfloat16
- ✓ Intel VNNI

2ND GEN INTEL XEON SCALABLE PROCESSORS



Intel DL Boost

- Intel VNNI

SOLUTION: CARDIAC MRI EXAM POC
SIEMENS HEALTHINEERS

intel XEON PLATINUM inside

RESULT:
5.5X FASTER
COMPARING INT8 WITH DL BOOST TO FP32[1]

- 2nd Gen Intel Xeon Scalable Processors
- Intel Deep Learning Boost
- Intel Distribution of OpenVINO™ toolkit

Client: Siemens Healthineers is a pioneer in the use of AI for medical applications. They are working with Intel to develop medical imaging use cases that don't require the added cost or complexity of accelerators.

Challenge: 1/3 of all deaths worldwide are due to cardiovascular disease.[2] Cardiac magnetic resonance imaging (MRI) exams are used to evaluate heart function, heart chamber volumes, and myocardial tissue.

This is a flood of data for radiology departments, resulting in potentially long turn-around-time (TAT)—even when the scan is considered stat.

Solution: Siemens Healthineers is developing AI-based technologies for the analysis of cardiac MRI exams.

They are working with Intel to optimize their heart chamber detection and quantification model for 2nd Gen Intel Xeon Scalable processors.

1. This Siemens Healthineers' feature is currently under development and not available for sale. 5.5x speedup: based on Siemens Healthineers and Intel analysis on 2nd Gen Intel Xeon Platinum 8280 Processor (28 Cores) with 192GB, DDR4-2933, using Intel OpenVino 2019 R1. HT ON, Turbo ON. CentOS Linux release 7.6.1810, kernel 4.19.5-1.el7.elrepo.x86_64.Custom topology and dataset (image resolution 288x288). Comparing FP32 vs Int8 with Intel DL Boost performance on the system. 2. Journal of the American College of Cardiology, 2017. Performance results are based on testing as of February 2018, and may not reflect all publicly available security updates. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

SOLUTION: VIDEO SURVEILLANCE

RINF TECH
Code-written Business Stories

RESULTS

UP TO **7.4X INCREASE**

Inference performance over baseline using OpenVINO R5 on 2nd generation Intel® Xeon® Scalable Processor and Intel DL Boost

intel XEON GOLD inside | OpenVINO | INTEL® AI BUILDERS MEMBER

Customer: RINF Tech specializes in cross-platform integration for checkout systems in retail, automotive, video surveillance and business intelligence.

Challenge: Analysing and understanding images faster and improving accuracy is the key to better decision making. The challenge is to provide rapid and accurate assessment of imagery to support daily operations efficiently, while providing critical information in near real time and in a cost effective manner

Solution: This challenge was resolved through the combination of RINF Tech's camera at the edge and 2nd generation Intel® Xeon® Scalable processors delivering competitive computing capacities. Additionally, higher Inference throughput was achieved using Intel® Distribution of OpenVINO® Toolkit

Configuration : NEW: Tested by Intel as of 03/18/2019. 2 socket Intel® Xeon® Gold 6252 Processor @ 2.10 GHZ, 24 cores per socket, , HT On, Turbo On, OS Linux,Deep Learning Framework: Caffe; tool : OpenVINO R5
Baseline : Tested by Intel as of 03/18/2019. 2 socket Intel® Xeon® Gold 6252 Processor @ 2.10 GHZ, 24 cores per socket,, HT On, Turbo On, OS Linux, Deep Learning Framework: Caffe
*Other names 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 http://www.intel.com/performance. Performance results are based on testing as of August 2018 and may not reflect all publicly available security updates.
.

PUBLIC

SOLUTION: FACE RECOGNITION
CLOUDWALK

RESULTS

UP TO **3.3X** INCREASE

Inference performance over baseline (Quantized from FP32 to INT8 processing) on 2S Intel® Xeon® CLX 8260 processors

Customer: CloudWalk is one of the Top 3 computer vision solution providers in PRC, delivering services to the public security and finance sectors.

Challenge: Deploying facial recognition solutions in bank, security government or police station face two bottlenecks - network bandwidth and computing capabilities. These negatively impact deep learning inference throughput and latency, thereby resulting in less than optimal user experiences.

Solution: This challenge was resolved through the combination of CloudWalk's camera at the edge and 2nd Gen Intel® Xeon® Scalable processors that addressed the computing bottleneck, as well as optimization for image processing and inferencing using Intel® Caffe and Intel® MKL-DNN. Result was a significant reduction in inference latency, while maintaining SLAs for accuracy.

Configuration: Cloudwalk Facial Recognition* (self-defined workload); OS: CentOS* 7.5 Kernel 3.10.0-957.1.3.el7.x86_64. Testing by Intel and Cloudwalk completed on Dec 18, 2018. Security Mitigations for Variants 1, 2, 3 and L1TF in place. TEST SYSTEM CONFIG: 2nd Gen Intel® Xeon® Platinum processor 8260L, 2.3 GHz, 24 cores, turbo and HT on, BIOS 1.0180, 192GB total memory, 12 slots / 16GB / 2666 MT/s / DDR4 LRDIMM, 1 x 480GB / Intel® SSD Data Center (Intel® SSD DC) S4500 + 1 x 1TB / Intel® SSD DC P4500; Intel® Optimization for Caffe*.Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Testing done on ____ 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 http://www.intel.com/performance. Performance results are based on testing as of August 2018 and may not reflect all publicly available security updates.

GET MAXIMUM UTILIZATION USING INTEL XEON SCALABLE PROCESSORS

running data center and AI workloads side-by-side

Improve Inference Throughput

UP TO **14X** **better inference throughput**

compared to previous-generation technology[1]

Accelerate Insights

UP TO **30X** **improved deep learning performance**

compared to previous-generation technology[2]

1. Configurations for "Up to 14X Al 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.OD.01 .0271.120720180605 (ucode: Ox200004d), Ubuntu 18.04.1 LTS, kernel 4.15.0-45-generic, SSD 1x sda INTEL SSDSC2BA80 SSD 745.2GB, nvme1 n1 INTEL SSDPE2KX040T7 SSD 3.7TB, Deep Learning Framework: Intel" Optimization for Caffe version: 1.1.3 (commit hash: 7010334f159da247db3fe3a9d96a3116ca0Gb09a), ICC version 18.0.1, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140d2d8790a75a, mode https://github.com/intel/caffe/blob/master/models/imel_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.SOGHz (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.Sin 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.SG -u 3.86 -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 -1".

2. Configurations for (1) "Up to 2x more inference throughput improvement on Intel® Xeon® Platinum 9282 processor with Intel® DL Boost" + (2) "Up to 30X Al 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.OD.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, 1cc 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_irnt8_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.eJ7.x86_64. SSD: Intel0 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 f96b759f71 b2281835f690af267158b82b150b5c. 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/imel_optimized_models (ResNet-50). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -1".

NOTICES & DISCLAIMERS

- Intel technologies may require enabled hardware, software or service activation. Your costs and results may vary.
- No product or component can be absolutely secure.
- Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks .
- 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 http://www.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 http://www.intel.com/go/turbo.
- 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.
- 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.
- Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.