Intel® AI HPC Workshop

LRZ April 09, 2021

Morning – Machine Learning

9:00 – 9:30	Introduction and Hardware Acceleration for AI OneAPI
9:30 – 10:00	Toolkits Overview, Intel® AI Analytics toolkit and oneContainer
10:00 -10:30	Break
10:30 - 12:30	Deep dive in Machine Learning tools

Quizzes!

LRZ Morning Session Survey

Afternoon – Deep Learning

13:30 – 14:45	Deep dive in Deep Learning tools
14:45 – 14:50	5 min Break
14:50 – 15:20	OpenVINO
15:20 - 15:45	25 min Break
15:45 – 17:00	Distributed training and Federated Learning

Quizzes!

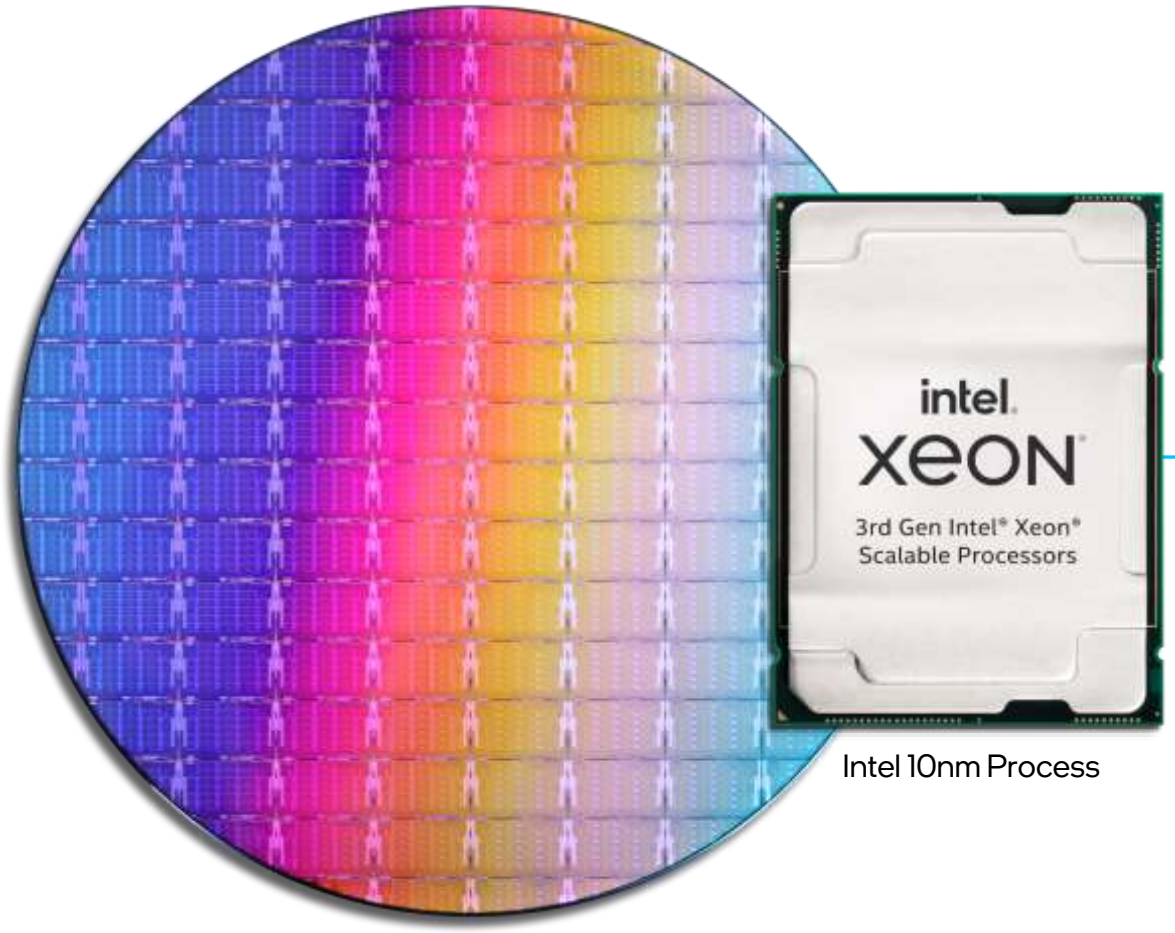
LRZ Afternoon Session Survey


intel.
PENTIUM
SILVER
intel.
PENTIUM
intel.
CORE
i5
intel.
iRIS
Xe
MAX
GRAPHICS
intel.
INSIDE
intel.
CELERON
intel.
CORE
i7
intel.
XEON
SILVER
intel.
XEON
GOLD
intel.
OPTANE
intel.
XEON
PLATINUM
intel.
XEON
intel.
PENTIUM
intel.

INTRODUCING

3rd Gen Intel® Xeon® Scalable processors

Performance made flexible



Intel 10nm Process

Up to 40 cores
per processor

20% IPC improvement
28 core, ISO Freq, ISO compiler

1.46x average performance increase
Geomean of Integer, Floating Point, Stream Triad, LINPACK
8380 vs. 8280

1.74x AI inference increase
8380 vs. 8280 BERT

2.65x average performance increase
vs. 5-year-old system
8380 vs. E5-2699v4

Performance varies by use, configuration and other factors. Configurations see appendix [1,3,5,55]

AI Performance Gains

3rd Gen Intel® Xeon® Scalable Processors with Intel Deep Learning Boost

Machine Learning

SciKit Learn & XGBoost
8380 vs 8280

XGBoost Fit	up to 1.44x
XGBoost Predict	up to 1.30x
Kmeans Fit	up to 1.36x
Kmeans Inference	up to 1.44x
Linear Regression Fit	up to 1.44x
Linear Regression Inference	up to 1.60x

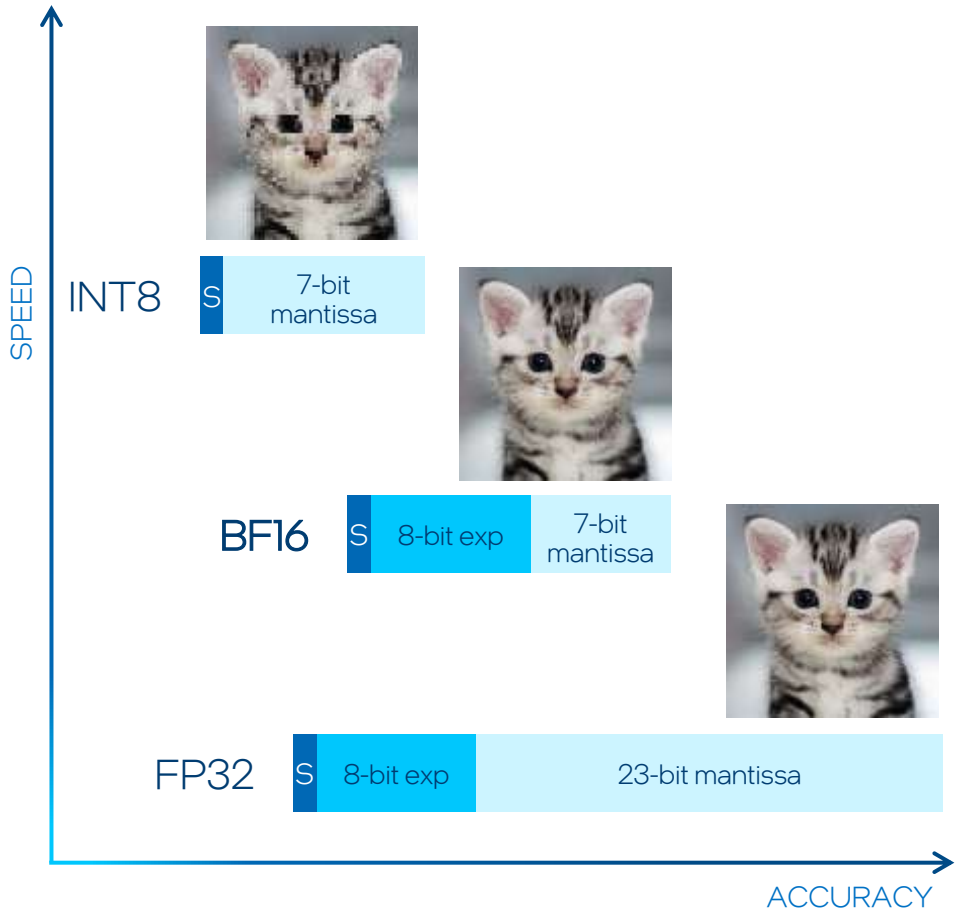
Deep Learning

	Real Time Inference 8380 vs 8280	Batch Inference 8380 vs 8280
Image Recognition Mobilenet-v1	up to 1.59x	up to 1.66x
Image Classification ResNet-50-v1.5	up to 1.52x	up to 1.56x
Language Processing BERT-large	up to 1.45x	up to 1.74x

Performance varies by use, configuration and other factors. Configurations see appendix [5,6,7,25]

Intel Deep Learning Boost, enhanced with bfloat16

The cutting edge of AI innovation with bfloat16 for 4 and 8-socket solutions



Up to

1.93X

HIGHER AI PERFORMANCE WITH INTEL DL BOOST[1]

For image classification vs. Intel Xeon Platinum 8280 processors

Similar accuracy
BF16 vs. FP32

Improved memory utilization
16 bits vs. 32 bits

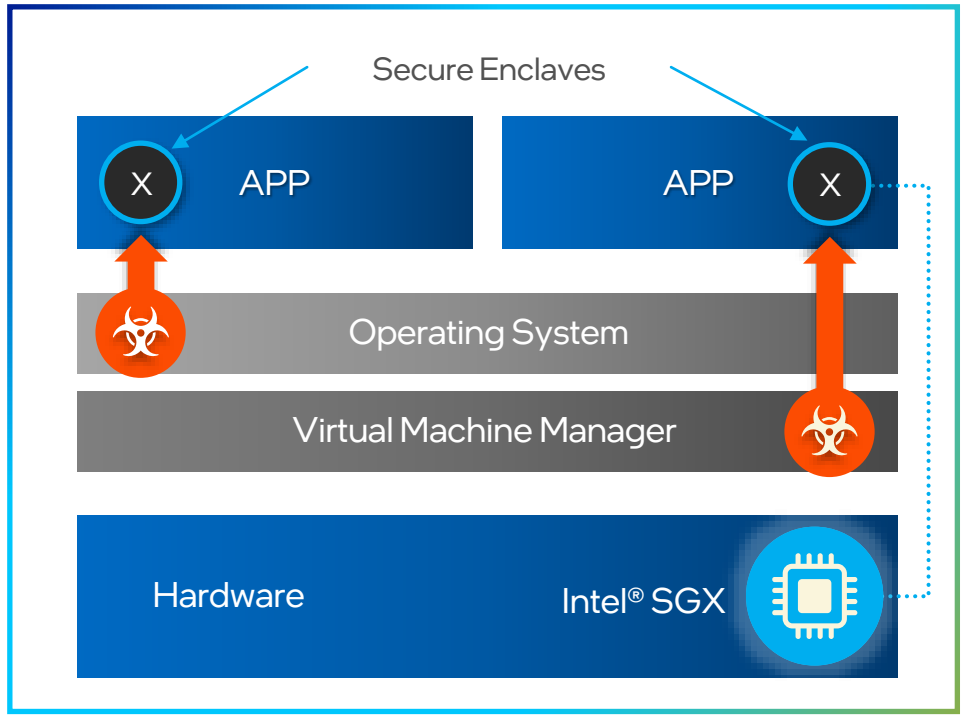
Increased performance
2 BF16 processes / cycle vs. 1 FP32

Optimized libraries & frameworks

oneAPI · OpenVINO · ONNX RUNTIME · PYTORCH · TensorFlow

1 For more complete information about performance and benchmark results, visit www.intel.com/3rd-gen-xeon-configs

Intel Software Guard Extensions (Intel SGX)



Enables **privacy assurances for sensitive data segments** without compromising performance

Huge enclaves now support demands of mainstream workloads (up to 1TB memory spaces)

the

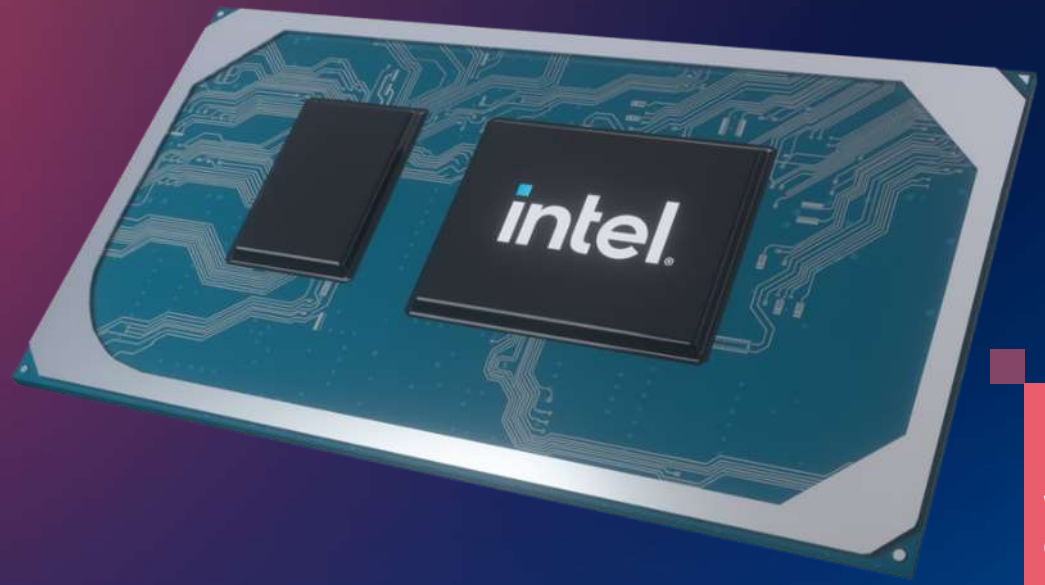
World's Best Processor

for thin and light laptops

new **CPU**

new **GPU**

Industry leading **AI**



Thunderbolt™ 4 integration

best **Wi-Fi 6** experience

deep software **Optimizations**

11th Gen Intel® Core™ Processor

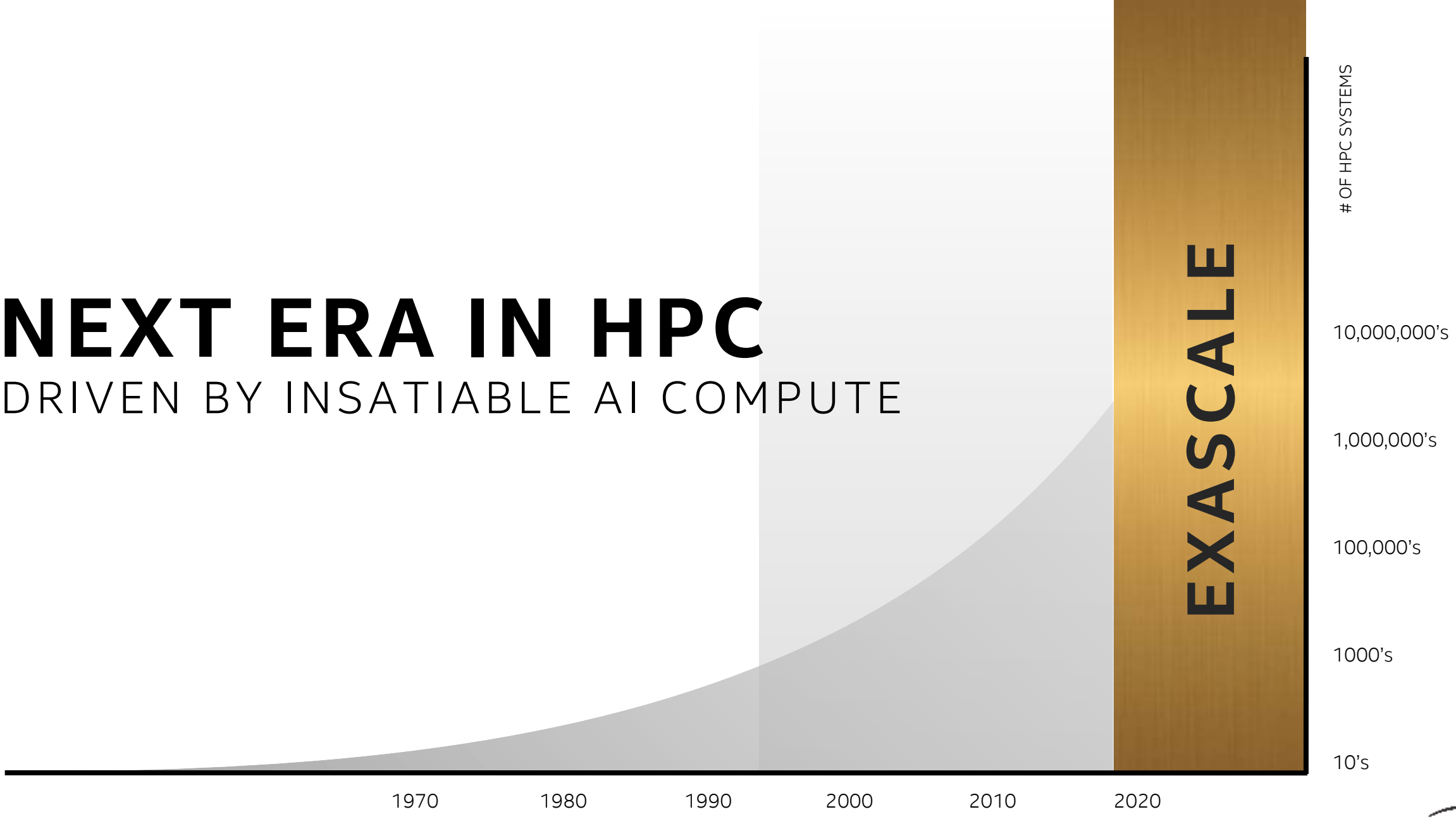
with Iris® Xᵉ Graphics

As measured by Intel® Core™ i7-1185G7 processor's status as the world's best processor for productivity, creation, gaming, collaborating and entertainment on a thin and light laptop. For more complete information about performance and benchmark results, visit www.intel.com/11thgen.

intel

NEXT ERA IN HPC

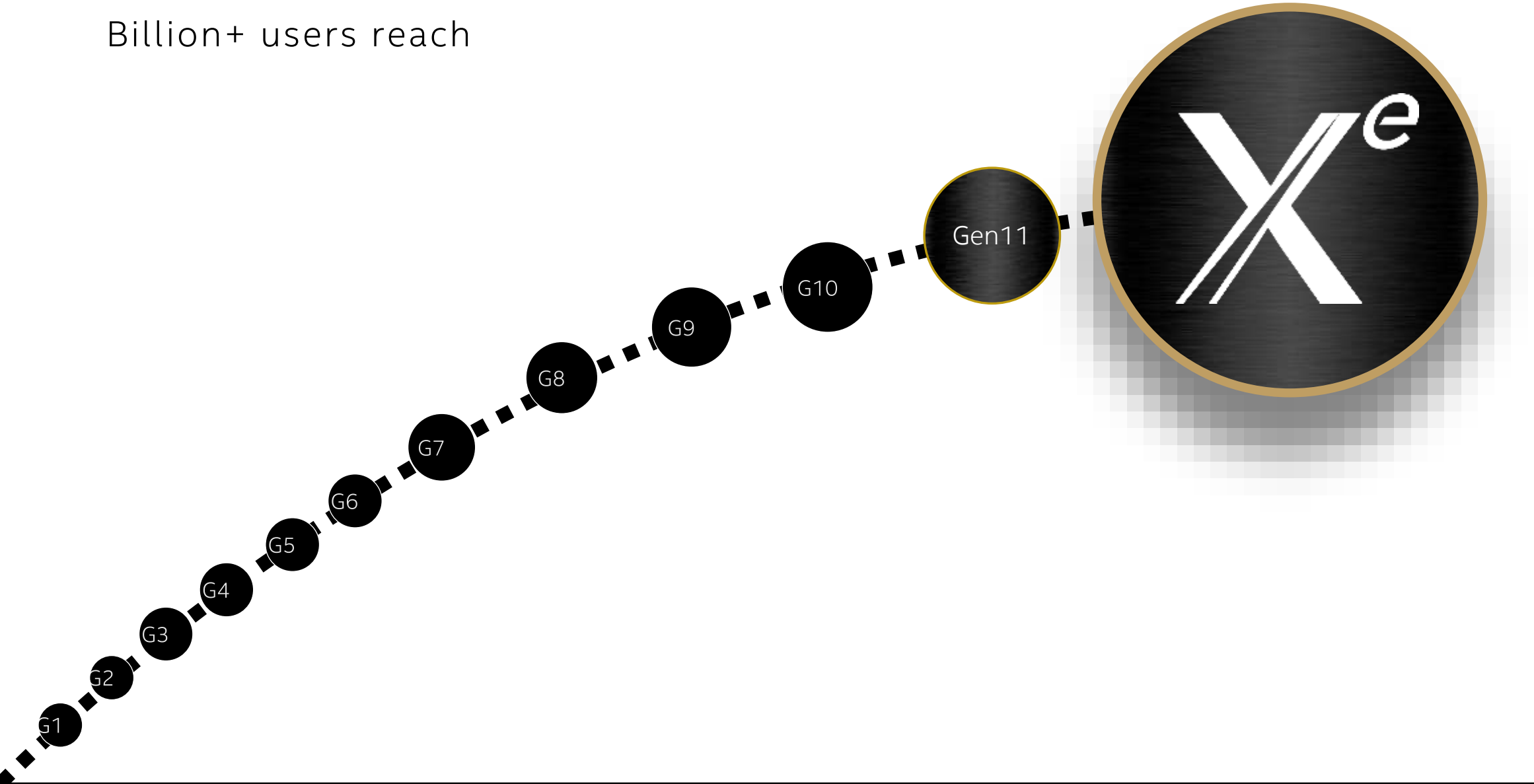
DRIVEN BY INSATIABLE AI COMPUTE



intel

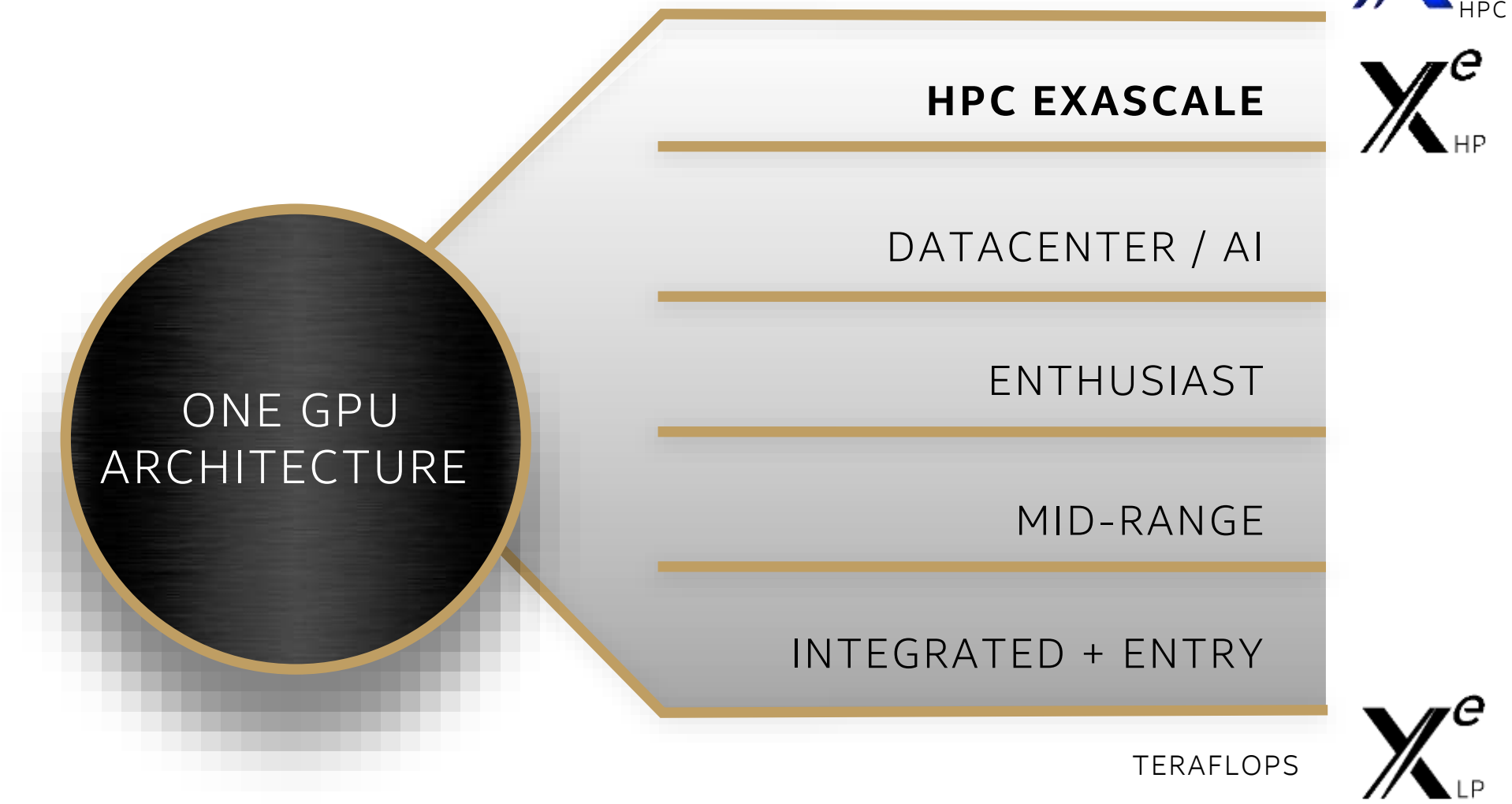
INTEL GPU IMPACT

Billion+ users reach



Xe ARCHITECTURE

FROM TERA FLOPS TO EXASCALE



intel

HPC FEATURES

- COMPUTE
  - SCALABILITY
  - AI PERFORMANCE
  - HPC PERFORMANCE
- MEMORY
  - SCALABILITY
  - BANDWIDTH
  - UNIFIED MEMORY

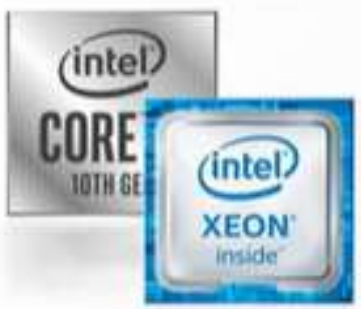
EXASCALE GPU

Xe HPC

intel

Intel AI Portfolio

GENERAL PURPOSE — PURPOSE BUILT



BF16 - AMX | BF16 - XMX | STRATIX 10 NX | MYRIAD | GOYA

DL BOOST – VNNI – DP4A | AGILEX | MYRIAD X | GAUDI | KEEM BAY

intel | TECHNOLOGY PILLARS

Under embargo until August 13th, 2020 at 6:00 a.m. Pacific Time.

Architecture Day 2020


intel

Notices and Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Intel contributes to the development of benchmarks by participating in, sponsoring, and/or contributing technical support to various benchmarking groups, including the BenchmarkXPRT Development Community administered by Principled Technologies.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Some results may have been estimated or simulated.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

All product plans and roadmaps are subject to change without notice.

Statements in this document that refer to future plans or expectations are forward-looking statements. These statements are based on current expectations and involve many risks and uncertainties that could cause actual results to differ materially from those expressed or implied in such statements. For more information on the factors that could cause actual results to differ materially, see our most recent earnings release and SEC filings at www.intc.com.

Appendix

1. 1.46x average performance gain - Ice Lake vs Cascade Lake: Geomean of 1.5x SPECrate2017_int_base (est), 1.52x SPECrate2017_fp_base (est), 1.47x Stream Triad, 1.38x Intel distribution of LINPACK. Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 processor on Coyote Pass with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0x261, HT on (SPECcpu2017), off (others), Turbo on, Ubuntu 20.04, 5.4.0-66-generic, 1x S4610 SSD 960G, SPECcpu2017 v1.1.0, Stream Triad, Linpack, ic19.1u2, MPI: Version 2019u9; MKL:2020.4.17, test by Intel on 3/15/2021. Platinum 8280: 1-node, 2x Intel® Xeon® Platinum 8280 processor on Wolf Pass with 384 GB (12 slots/ 32GB/ 2933) total DDR4 memory, ucode 0x5003003, HT on (SPECcpu2017), off (others), Turbo on, Ubuntu 20.04, 5.4.0-62-generic, 1x S3520 SSD 480G, SPECcpu2017 v1.1.0, Stream Triad, Intel distribution of LINPACK, ic19.1u2, MPI: Version 2019u9; MKL:2020.4.17, test by Intel on 2/4/2021.

2. **1.54x average performance gain - Ice Lake vs Skylake**: Geomean of 1.6x SPECrate2017_int_base (est), 1.62x SPECrate2017_fp_base (est), 1.52x Stream Triad, 1.44x Intel distribution of LINPACK. 3rd Gen Intel® Xeon® Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 processor on Coyote Pass with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0x261, HT on (SPECcpu2017), off (others), Turbo on, Ubuntu 20.04, 5.4.0-66-generic, 1x S4610 SSD 960G, SPECcpu2017 v1.1.0, Stream Triad, Linpack, ic19.1u2, MPI: Version 2019u9; MKL:2020.4.17, test by Intel on 3/15/2021. Intel® Xeon® Platinum 8180: 1-node, 2x Intel® Xeon® Platinum 8180 processor on Wolf Pass with 192 GB (12 slots/ 16GB/ 2933[2666]) total DDR4 memory, ucode 0x2006a08, HT on (SPECcpu2017), off (others), Turbo on, Ubuntu 20.04, 5.4.0-62-generic, SPECcpu2017 v1.1.0, Stream Triad, Intel distribution of LINPACK, ic19.1u2, MPI: Version 2019 Update 9 Build 20200923; MKL: psxe_runtime_2020.4.17, test by Intel on 1/27/21.

3. **2.65x average performance gain - Ice Lake vs Broadwell**: Geomean of 2.34x SPECrate2017_int_base (est), 2.6x SPECrate2017_fp_base (est), 2.55x Stream Triad, 3.18x Intel distribution of LINPACK. 3rd Gen Intel® Xeon® Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 processor on Coyote Pass with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0x261, HT on (SPECcpu2017), off (others), Turbo on, Ubuntu 20.04, 5.4.0-66-generic, 1x S4610 SSD 960G, SPECcpu2017 v1.1.0, Stream Triad, Linpack, ic19.1u2, MPI: Version 2019u9; MKL:2020.4.17, test by Intel on 3/15/2021. Intel® Xeon® E5-2699v4: 1-node, 2x Intel® Xeon® processor E5-2699v4 on Wildcat Pass with 256 GB (8 slots/ 32GB/ 2400) total DDR4 memory, ucode 0x038, HT on (SPECcpu2017), off (others), Turbo on, Ubuntu 20.04, 5.4.0-62-generic, 1x S3700 400GB SSD, SPECcpu2017 v1.1.0, Stream Triad, Intel distribution of LINPACK, ic19.1u2, MPI: Version 2019 Update 9 Build 20200923; MKL: psxe_runtime_2020.4.17, test by Intel on 1/17/21.

4. **3.14x average performance gain - Ice Lake vs Haswell**: Geomean of 2.85x SPECrate2017_int_base (est), 3.08x SPECrate2017_fp_base (est), 2.8x Stream Triad, 3.97x Intel distribution of LINPACK. 3rd Gen Intel® Xeon® Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 processor on Coyote Pass with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0x261, HT on (SPECcpu2017), off (others), Turbo on, Ubuntu 20.04, 5.4.0-66-generic, 1x S4610 SSD 960G, SPECcpu2017 v1.1.0, Stream Triad, Linpack, ic19.1u2, MPI: Version 2019u9; MKL:2020.4.17, test by Intel on 3/15/2021. Intel® Xeon® E5-2699v3: 1-node, 2x Intel® Xeon® processor E5-2699v3 on Wildcat Pass with 256 GB (8 slots/ 32GB/ 2666[2133]) total DDR4 memory, ucode 0x44, HT on (SPECcpu2017), off (others), Turbo on, Ubuntu 20.04, 5.4.0-62-generic, 1x S3700 400GB SSD, SPECcpu2017 v1.1.0, Stream Triad, Intel distribution of LINPACK, ic19.1u2, MPI: Version 2019 Update 9 Build 20200923; MKL: psxe_runtime_2020.4.17, test by Intel on 2/3/21.

5. **BERT-Large SQuAD : 1.45x higher INT8 real-time inference throughput & 1.74x higher INT8 batch inference throughput & 1.22x performance/core :** Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 processor on Coyote Pass with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0x261 , HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Intel_SSDSC2KG96, Intel® SSDPE2KX010T8, BERT - Large SQuAD, gcc-9.3.0, oneDNN 1.6.4, BS=1,128 INT8, TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/, test by Intel on 3/12/2021. Platinum 8280: 1-node, 2x Intel® Xeon® Platinum 8280 processor on Wolf Pass with 384 GB (12 slots/ 32GB/ 2933) total DDR4 memory, ucode 0x5003003, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-48-generic, 1x Samsung_SSD_860, Intel® SSDPE2KX040T8, BERT - Large SQuAD, gcc-9.3.0, oneDNN 1.6.4, BS=1,128 INT8, TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/, test by Intel on 2/17/2021.

6. **MobileNet-v1: 1.59x higher INT8 real-time inference throughput & 1.66x higher INT8 batch inference & 1.16x performance/core throughput: Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 processor on Coyote Pass with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0x261, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Intel_SSDSC2KG96, Intel® SSDPE2KX010T8, MobileNet-v1, gcc-9.3.0, oneDNN 1.6.4, BS=1,56 INT8, TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/, test by Intel on 3/12/2021. Platinum 8280: 1-node, 2x Intel® Xeon® Platinum 8280 processor on Wolf Pass with 384 GB (12 slots/ 32GB/ 2933) total DDR4 memory, ucode 0x5003003, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-48-generic, 1x Samsung_SSD_860, Intel® SSDPE2KX040T8,, MobileNet-v1, gcc-9.3.0, oneDNN 1.6.4, BS=1,56 INT8, TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/, test by Intel on 2/17/2021.**

7. **ResNet-50 v1.5 : 1.52x higher INT8 real-time inference throughput & 1.56x higher INT8 batch inference throughput on Ice Lake vs. prior generation Cascade Lake Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 processor on Coyote Pass with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0x261 , HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Intel_SSDSC2KG96, Intel® SSDPE2KX010T8, ResNet-50 v1.5, gcc-9.3.0, oneDNN 1.6.4, BS=1,128 INT8, TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/, test by Intel on 3/12/2021. Platinum 8280: 1-node, 2x Intel® Xeon® Platinum 8280 processor on Wolf Pass with 384 GB (12 slots/ 32GB/ 2933) total DDR4 memory, ucode 0x5003003, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-48-generic, 1x Samsung_SSD_860, Intel® SSDPE2KX040T8, ResNet-50 v1.5, gcc-9.3.0, oneDNN 1.6.4, BS=1,128 INT8, TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/, test by Intel on 2/17/2021.**

Appendix

20. **3.34x higher IPSec AES-GCM performance,3.78x higher IPSec AES-CMAC performance,3.84x higher IPSec AES-CTR performance,1.5x higher IPSec ZUC performance:** 8380: 1-node, 2x Intel(R) Xeon(R) Platinum 8380 CPU on M50CYP2SB2U with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0x8d055260, HT On, Turbo Off , Ubuntu 20.04.2 LTS, 5.4.0-66-generic, 1x Intel 1.8TB SSD OS Drive, intel-ipsec-mb v0.55, gcc 9.3.0, Glibc 2.31, test by Intel on 3/17/2021. 8280M: 1-node, 2x Intel(R) Xeon(R) Platinum 8280M CPU on S2600WFT with 384 GB (12 slots/ 32GB/ 2933) total DDR4 memory, ucode 0x4003003, HT On, Turbo Off , Ubuntu 20.04.2 LTS, 5.4.0-66-generic, 1x Intel 1.8TB SSD OS Drive, intel-ipsec-mb v0.55, gcc 9.3.0, Glibc 2.31, test by Intel on 3/8/2021.

21. **3.5x higher ISA-L AES-XTS performance, 2.30x higher ISA-L CRC performance: ISA-L**: 8380: 1-node, 2x Intel® Xeon® Platinum 8380 Processor, 40 cores HT On Turbo OFF Total Memory 512 GB (16 slots/ 32GB/ 3200 MHz), Data protection (Reed Solomon EC (10+4)), Data integrity (CRC64), Hashing (Multibuffer MD5),Data encryption (AES-XTS 128 Expanded Key), Data Compression (Level 3 Compression (Calgary Corpus)), BIOS: SE5C6200.86B.3021.D40.2103160200 (ucode: 0x8d05a260), Ubuntu 20.04.2, 5.4.0-67-generic, gcc 9.3.0 compiler, yasm 1.3.0, nasm 2.14.02, isal 2.30, isal_crypto 2.23, OpenSSL 1.1.1.i, zlib 1.2.11, Test by Intel as of 03/19/2021. 8280: 1-node, 2x Intel® Xeon® Platinum 8280 Processor, 28 cores HT On Turbo OFF Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.02.01.0013.121520200651 (ucode:0x4003003), Ubuntu 20.04.2, 5.4.0-67-generic,, gcc 9.3.0 compiler, yasm 1.3.0, nasm 2.14.02, isal 2.30, isal_crypto 2.23, OpenSSL 1.1.1.i, zlib 1.2.11 Test by Intel as of 2/9/2021. Performance measured on single core.

22. **NVMe-over-TCP IOPS Throughput** : Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 Processor, 40 cores HT On Turbo ON Total Memory 1024 GB (16 slots/ 64GB/ 3200), BIOS:SE5C6200.86B.2021.D40.2103100308 (ucode:0x261), Fedora 30, Linux Kernel 5.7.12, gcc 9.3.1 compiler, fio 3.20, SPDK 21.01, Storage: 16x Intel® SSD D7-P5510 7.68 TB or 16x Intel® Optane™ SSD 400GB P5800X, Network: 2x 100GbE Intel E810-C, Test by Intel as of 3/17/2021. Platinum 8280: 1-node, 2x Intel® Xeon® Platinum 8280 Processor, 28 cores HT On Turbo ON Total Memory 768 GB (24 slots/ 32GB/ 2666), BIOS: SE5C620.86B.02.01.0013.121520200651 (ucode:0x4003003), Fedora 30, Linux Kernel 5.7.12, gcc 9.3.1 compiler, fio 3.20, SPDK 21.01, Storage: 16x Intel® SSD DC P4610 1.6TB, Network: 1x 100GbE Intel E810-C, Test by Intel as of 2/10/2021.

23. **2.5x higher transactions on Aerospike Database**: Platinum 8368: 1-node, 2x Intel® Xeon® Platinum 8368 processor on Coyote Pass with 256 GB (16 slots/ 16GB/ 3200) total DDR4 memory, 8192 GB (16 slots/ 512 GB/ 3200) total PMem, ucode x261, HT on, Turbo on, CentOS 8.3.2011, 4.18.0-193.el8.x86_64, 1x Intel 960GB SSD, 7x P5510 3.84TB, 2x Intel E810-C 100Gb/s, Aerospike Enterprise Edition 5.5.0.2; Aerospike C Client 5.1.0 Benchmark Tool; 70R/30W. Dataset size: 1.1TB, 9.3 billion 64B records, PMDK libPMem, Index (PMem)+data (SSD) and Index+data (PMem), test by Intel on 3/16/2021. Platinum 8280: 1-node, 2x Intel® Xeon® Platinum 8280L processor on Wolf Pass with 768 GB (12 slots/ 64GB/ 2666) total DDR4 memory, 3072 GB (12 slots/ 256 GB/ 2666) total PMem, ucode 0x5003003, HT on, Turbo on, CentOS 8.3.2011, 4.18.0-193.el8.x86_64, 7x P4510 1.8TB PCIe 3. 1, 2x Intel XL710 40Gb/s, Aerospike Enterprise Edition 5.5.0.2; Aerospike C Client 5.1.0 Benchmark Tool; 70R/30W. Dataset size: 1.1TB, 9.3 billion 64B records, PMDK libPMem, Index (PMem)+data (SSD), test by Intel on 3/16/2021.

24. **5.63x higher OpenSSL RSA Sign 2048 performance,1.90x higher OpenSSL ECDSA Sign p256 performance,4.12x higher OpenSSL ECDHE x25519 performance,2.73x higher OpenSSL ECDHE p256 performance**, 8280M:1-node, 2x Intel(R) Xeon(R) Platinum 8280M CPU on S2600WFT with 384 GB (12 slots/ 32GB/ 2933) total DDR4 memory, ucode 0x5003003, HT On, Turbo Off, Ubuntu 20.04.1 LTS, 5.4.0-65-generic, 1x INTEL_SSDSC2KG01 , OpenSSL 1.1.1j, GCC 9.3.0, test by Intel on 3/5/2021. 8380: 1-node, 2x Intel(R) Xeon(R) Platinum 8380 CPU on M50CYP2SB2U with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xd000270, HT On, Turbo Off, Ubuntu 20.04.1 LTS, 5.4.0-65-generic, 1x INTEL_SSDSC2KG01 , OpenSSL 1.1.1j, GCC 9.3.0, QAT Engine v0.6.4, test by Intel on 3/24/2021. 8380: 1-node, 2x Intel(R) Xeon(R) Platinum 8380 CPU on M50CYP2SB2U with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xd000270, HT On, Turbo Off, Ubuntu 20.04.1 LTS, 5.4.0-65-generic, 1x INTEL_SSDSC2KG01 , OpenSSL 1.1.1j, GCC 9.3.0, QAT Engine v0.6.5, test by Intel on 3/24/2021.

25. **1.44x XGBoost fit, 1.30x XGBoost predict, 1.36x Kmeans fit, 1.44x Kmeans inference, 1.44x Linear Regression fit, 1.60x Linear Regression inference:** 8380: 1-node, 2x Intel® Xeon® Platinum 8380 processor on Coyote Pass with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0x261, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-64-generic, 1x Intel® SSDSC2KG960G7, 1x Intel® SSDSC2KG960G7, Python 3.7.9, Sklearn 0.24.1(https://github.com/IntelPython/scikit-learn_bench), Daal4py 2021.2, XGBoost 1.3.3, test by Intel on 3/19/2021. 8280: 1-node, 2x Intel® Xeon® Platinum 8280L processor on S2600WFT with 384 GB (12 slots/ 32GB/ 2933) total DDR4 memory, ucode 0x5003003, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Intel® SSDSC2BB800G7, 1x Intel® SSDSC2BB800G7, Python 3.7.9, Sklearn 0.24.1(https://github.com/IntelPython/scikit-learn_bench), Daal4py 2021.2, XGBoost 1.3.3, test by Intel on 2/5/2021.

26. 10x higher batch AI inference performance with Intel-optimized Tensor Flow vs. stock Cascade Lake FP32 configuration8380: 1-node, 2x Intel Xeon Platinum 8380 processor on Coyote Pass with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode X261 , HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Intel_SSDSC2KG96, Intel SSDPE2KX010T8, ResNet-50 v1.5, gcc-9.3.0, oneDNN 1.6.4, BS=128 FP32,INT8, TensorFlow 2.4.1 with Intel optimizations for 3rd Gen Intel Xeon Scalable processor, upstreamed to TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/, Unoptimized model : TensorFlow- 2.4.1, Modelzoo:https://github.com/IntelAI/models –b master, test by Intel on 3/12/2021.8280: 1-node, 2x Intel Xeon Platinum 8280 processor on Wolf Pass with 384 GB (12 slots/ 32GB/ 2933) total DDR4 memory, ucode 0x5003003, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-48-generic, 1x Samsung_SSD_860, Intel SSDPE2KX040T8, ResNet-50 v1.5, gcc-9.3.0, oneDNN 1.6.4, BS=128 FP32,INT8, Optimized model : TensorFlow 2.4.1 with Intel optimizations for 3rd Gen Intel Xeon Scalable processor, upstreamed to TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/, Unoptimized model : TensorFlow- 2.4.1, Modelzoo:https://github.com/IntelAI/models –b master, test by Intel on 2/17/2021.

Appendix

53. Intel® Optane™ Persistent Memory 200 Series
    a. Average 32% more memory bandwidth:
    Based on testing by Intel as of April 27, 2020 (Baseline) and March 23, 2021 (New). Baseline configuration: 1-node, 1 x Intel® Xeon® Platinum 8280L processor (28 cores at 2.7 GHz) on Neon City with a single Intel® Optane™ PMem module configuration (6 x 32 GB DRAM; 1 x {128 GB, 256 GB, 512 GB} Intel® Optane™ PMem module), ucode rev: 04002F00 running Fedora 29 kernel 5.1.18-200.fc29.x86_64 and Intel Memory Latency Checker (Intel MLC) version 3.8 with App Direct Mode.
    New Configuration: 1-node, 1 x pre-production 3rd Gen Intel® Xeon® Scalable processor (38 cores at 2.0 GHz) on Wilson City with a single Intel® Optane™ PMem module configuration (8 x 32 GB DRAM; 1 x {128 GB, 256 GB, 512 GB} Intel® Optane™ PMem module), ucode rev: 8d000270 running RHEL 8.1 kernel 4.18.0-147.el8.x86_64 and Intel MLC version 3.9 with App Direct Mode.
    b. Katana 2X faster graph analytics computations:
    Baseline: Test by Intel as of 3/11/2021. 1-node, 2x Intel® Xeon® Platinum 8260 Processor, 24 cores HT On Turbo ON Total Memory 768GB (12 slots/ 64GB/ 2666 MHz), Total PMEM 6TB (12 slots/512GB/2666MHZ), BIOS: SE5C620.86B.0X.02.0001.051420190324(ucode:0x4003003), UBUNTU 24.04,5.4.0-65-generic, gcc 9.3.0 compiler, Galois (https://github.com/IntelligentSoftwareSystems/Galois), HT-OFF, page_alloc.shuffle=1
    New Config: Test by Intel as of 3/11/2021. 1-node, 2x Intel® Xeon® Platinum 8368 Processor, 38 cores HT Off Turbo ON Total Memory 1 TB (16 slots/ 64GB/ 3200 MHz), Total PMEM 8TB (12 slots/512GB/3200MHZ), BIOS:SE5C6200.86B.SE5C6200.86B.2021.D40.2103100308 (ucode: 0x8d055260), Ubuntu 24.04,5.4.0-66-generic, gcc 9.3.0 compiler, Galois (https://github.com/IntelligentSoftwareSystems/Galois), HT-OFF, page_alloc.shuffle=1
    c. VMware 25% lower cost per VM:
    Based on testing by Intel as of March, 23 2021. Baseline Configuration: 2x Intel® Xeon® Platinum 8380 processor @ 2.3 GHz (Microcode: 0x8d055260), 1x Intel® Server Platform M50CYP, 2.0TB DDR4, 32 slots/64 GB/3200 MT/s. BIOS: SE5C6200.86B.0020.P16.2101262103, Hyper Threading: Enabled, Turbo: Enabled, NVM Performance Setting: Balanced Performance Uncore Power management -> Performance P-limit Enabled, 1 NUMA Nodes per Socket, Data Storage: 4x 4.0 TB Intel® SSD P4510+1x 8.0 TB Intel® SSD P4510, Network: 1x Intel® Ethernet X540-T2.
    New Configuration: 2x Intel® Xeon® Platinum 8380 processor @ 2.3 GHz (Microcode: 0x8d055260), 1x Intel® Server Platform M50CYP, 2.0TB DDR4, 32 slots/64 GB/3200 MT/s. 2.0 TB, 16 x128 GB Intel® Optane™ PMem 200 series/3200 MT/s and 512 GB DDR4, 16 x 32 GB/3200 MT/s (PMem Firmware Version: 02.02.00.1540). BIOS: SE5C6200.86B.0020.P16.2101262103, Hyper Threading: Enabled, Turbo: Enabled, NVM Performance Setting: Balanced Performance Uncore Power management -> Performance P-limit Enabled, 1 NUMA Nodes per Socket, Data Storage: 4x 4.0 TB Intel® SSD P4510+1x 8.0 TB Intel® SSD P4510, Network: 1x Intel® Ethernet X540-T2.
    OS/Software: VMware ESXi 7.0.2 (VMware_bootbank_cpu-microcode_7.0.2-0.0.17473468), Workload: VMmark 3.1 benchmark with modifications to VMmark tile to consume a larger amount of memory without increasing the CPU requirements.
    More information available: Intel® Optane™ Persistent Memory "Memory Mode" Virtualized Performance Study (vmware.com)
54. Upto 100x gains due to software improvement on SciKit learn workloads : linear regression fit, SVC inference, kdtree_knn inference and elastic-net fit on Ice Lake with Daal4py optimizations compared with stock Scikit-learnUpto 100x gains due to software improvement on SciKit learn workloads : linear regression fit, SVC inference, kdtree_knn inference and elastic-net fit on Ice Lake with Daal4py optimizations compared with stock Scikit-learn 8380: 1-node, 2x Intel Xeon Platinum 8380 (40C/2.3GHz, 270W TDP) processor on Intel Software Development Platform with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode X55260, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-64-generic, 2x Intel_SSDSC2KG96, Unoptimized : Python : Python 3.7.9, SciKit-Learn : Sklearn 0.24.1, Optimized : oneDAL : Daal4py 2021.2, Benchmarks: https://github.com/IntelPython/scikit-learn_bench, tested by Intel, and results as of March 2021
55. 20% IPC improvement: 3rd Gen Xeon Scalable processor: 1-node, 2x 28-core 3rd Gen Intel Xeon Scalable processor, Wilson City platform, 512GB (16 slots / 32GB / 3200) total DDR4 memory, HT on, ucode=x270, RHEL 8.0, Kernel Version4.18.0-80.el8.x86_64, test by Intel on 3/30/2021. 2nd Gen Intel Xeon Scalable processor: 1-node, 2x 28-core 2nd Gen Intel Xeon Scalable processor, Neon City platform, 384GB (12 slots / 32GB / 2933) total DDR4 memory, HT on, ucode=x2f00, RHEL 8.0, Kernel Version4.18.0-80.el8.x86_64, test by Intel on 3/30/2021. SPECrate2017_int_base (est). Tests at equal frequency, equal uncore frequency, equal compiler.