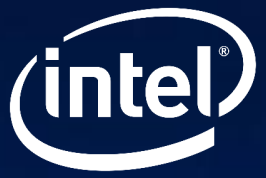


# Start Here

June 2020

# Today

## INDUSTRY INFLECTIONS ARE FUELING THE GROWTH OF DATA

5G Network Transformation, Artificial Intelligence, Intelligent Edge, Cloudification

## AI & ANALYTICS ARE THE DEFINING WORKLOADS OF THE NEXT DECADE

## UNMATCHED PORTFOLIO BREADTH AND ECOSYSTEM SUPPORT

Intel delivers a silicon & software foundation designed for the diverse range of use cases from the cloud to the edge

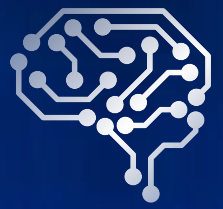
# Intel Purpose

To create world-changing technology
that enriches the lives
of every person on earth

# Industry Inflections

5G Network Transformation

Artificial Intelligence



Intelligent Edge

Cloudification

# Unleashing the Potential of Data

## MOVE FASTER

BAREFOOT NETWORKS | an Intel company

intel ETHERNET

intel SILICON PHOTONICS

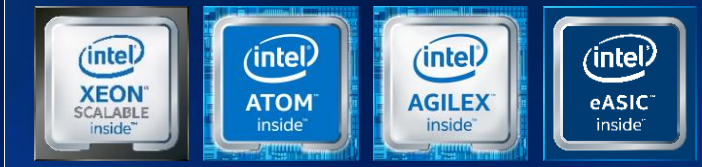
## STORE MORE

intel OPTANE™ PERSISTENT MEMORY

intel OPTANE™ SSD

intel 3D NAND SSD

## PROCESS EVERYTHING





## SOFTWARE & SYSTEM LEVEL OPTIMIZED

intel

# Solving The World's Greatest Challenges

## through new technology-based innovations & approaches

## intel Pandemic Response Technology Initiative

### DIAGNOSE & TREAT

Banner Health | VeeMed

SERVE CRITICALLY ILL COVID PATIENTS THROUGH VIRTUAL CARE

Medical INFORMATICS

RAPIDLY EXPAND REMOTE ICUs TO 100 US HOSPITALS

Montefiore

PREDICT PATIENTS WHO WILL DEVELOP ACUTE RESPIRATORY DISTRESS SYNDROME

### RESEARCH & VACCINES

TACC

ILLUMINATE VIRUS AND DNA REPLICATION TASKS

VERISIMLife

SPEED DISCOVERY OF NEW PHARMACEUTICALS

Lenovo

ACCELERATE GENOMIC ANALYSIS AND DECODE COVID-19

### LOCAL COMMUNITY

International Red Cross

SUPPORT GLOBAL RELIEF EFFORTS

akara

SANITIZE ROOMS AND EQUIPMENT WITH AUTONOMOUS ROBOTS

First Book | CDW | The LEGO Foundation

LEARNING SOLUTION FOR STUDENTS WITHOUT COMPUTERS OR INTERNET

intel

# AI Strategy

| | | | |
|---|---|---|---|
| **Ecosystem** | BUILD THRIVING ECOSYSTEM | DELIVER SOLUTIONS | INNOVATE & INVEST IN THE FUTURE |
| **Software** | OPTIMIZED SOFTWARE | EMPOWER DEVELOPERS | UNIFY APIS |
| **Hardware** | INFUSE AI INTO THE CPU | LEADERSHIP DISCRETE ACCELERATORS | OPTIMIZE AT THE PLATFORM LEVEL |

# Unmatched Silicon & Software Foundation

## for AI & analytics

### SOFTWARE & SOLUTIONS

oneAPI

OpenVINO™

intel select solution

### PROCESS

| 3rd Gen Intel Xeon Scalable | GPU | Intel Stratix 10 NX | Gen 3 Intel Movidius VPU | Habana Gaudi & Goya |
|---|---|---|---|---|
| LAUNCHING | IN DEVELOPMENT | DISCLOSING | EARLY ACCESS | LIMITED SAMPLING |


Xe



habana
An Intel Company

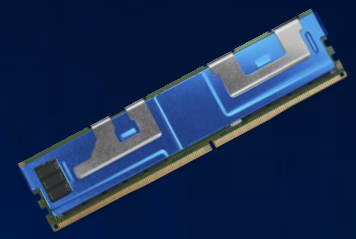
CPU — GPU — FPGA — SPECIALIZED ACCELERATORS

WORKLOAD BREADTH — AI SPECIFIC

### STORE

**Intel Optane Persistent Memory 200 Series**

LAUNCHING



**Intel SSD D7-P5500**
**Intel SSD P5600**

LAUNCHING


intel

LAUNCHING

# 3rd Gen Intel Xeon Scalable Processor

## Built for today's AI-infused, data-intensive services

**1.9X**
AVERAGE PERFORMANCE GAIN
vs 5-YEAR-OLD PLATFORM

### BUILT-IN AI ACCELERATION

Intel Deep Learning Boost
NEW: bfloat16*

**up to 1.98X**
HIGHER DATABASE PERFORMANCE
vs 5-YEAR-OLD PLATFORM

### BREAKTHROUGH MEMORY

Intel Optane Persistent Memory
200 series


TARGETED FOR 4S-8S SYSTEMS

### FLEXIBILITY

Enhanced
Intel Speed Select Technology

Alibaba Cloud · AsiaInfo 亚信科技 5G 原力进化 · Baidu 百度 · FACEBOOK · FUJITSU · GIGABYTE · 紫光集团 H3C 核心企业 数字化解决方案领导者 · 海鑫科金 HISIGN TECHNOLOGY · Hewlett Packard Enterprise · HITACHI · HUAWEI · hyve solutions · inspur 浪潮 · Inventec Inventec Data Center Solutions · Lenovo · Neusoft · Quanta Computer · SAP · sas · SUPERMICRO · Tencent Cloud · wiwynn · ZTE

Performance results are based on testing as of dates in configuration and may not reflect all publicly available security updates. See backup for configuration details.
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

*Available on select 3rd Gen Intel Xeon Scalable processors

intel

# Intel Xeon Scalable Processor

The only mainstream data center CPU with built-in AI acceleration

2017

**1ST GEN**
**Intel Xeon Scalable**

Intel
AVX-512
**FP32**

**24**
OPTIMIZED
TOPOLOGIES
ON XEON

## AI ON 1ST GEN INTEL XEON SCALABLE

AccuRad 盈谷 · datacubes · Descartes Labs · kakao
Kingsoft Cloud KSCLOUD · MicroSeismic · Midea · 南京大学 NANJING UNIVERSITY
SIEMENS Healthineers · SURF SARA · SYNESIS · Taboola

OPTIMIZED LIBRARIES & FRAMEWORKS

Caffe · mxnet · oneAPI · ONNX RUNTIME · OpenVINO
PaddlePaddle · PyTorch · TensorFlow

# Intel Xeon Scalable Processor

## The only mainstream data center CPU with built-in AI acceleration

2019

**2ND GEN**
**Intel Xeon Scalable**

Intel
Deep Learning Boost
INT8

**44**
OPTIMIZED
TOPOLOGIES
ON XEON

## INTEL DL BOOST ADOPTION

CERN
云创大数据 www.cStor.cn
alhua TECHNOLOGY
DARWIN AI
HCL
HLRN
HY 汇医慧影 huiyihuiying.com
智芯原动 ICE TECH.
科大讯飞 iFLYTEK
JD Cloud & AI
kakao
MAX PLANCK INSTITUTE FOR HUMAN COGNITIVE AND BRAIN SCIENCES
Neusoft
QINGCLOUD
RINF TECH
上海交通大學 Shanghai Jiao Tong University
SIEMENS Healthineers
TACC

## OPTIMIZED LIBRARIES & FRAMEWORKS

mxnet
oneAPI
ONNX RUNTIME
OpenVINO
PaddlePaddle
PyTorch
TensorFlow

INTRODUCING

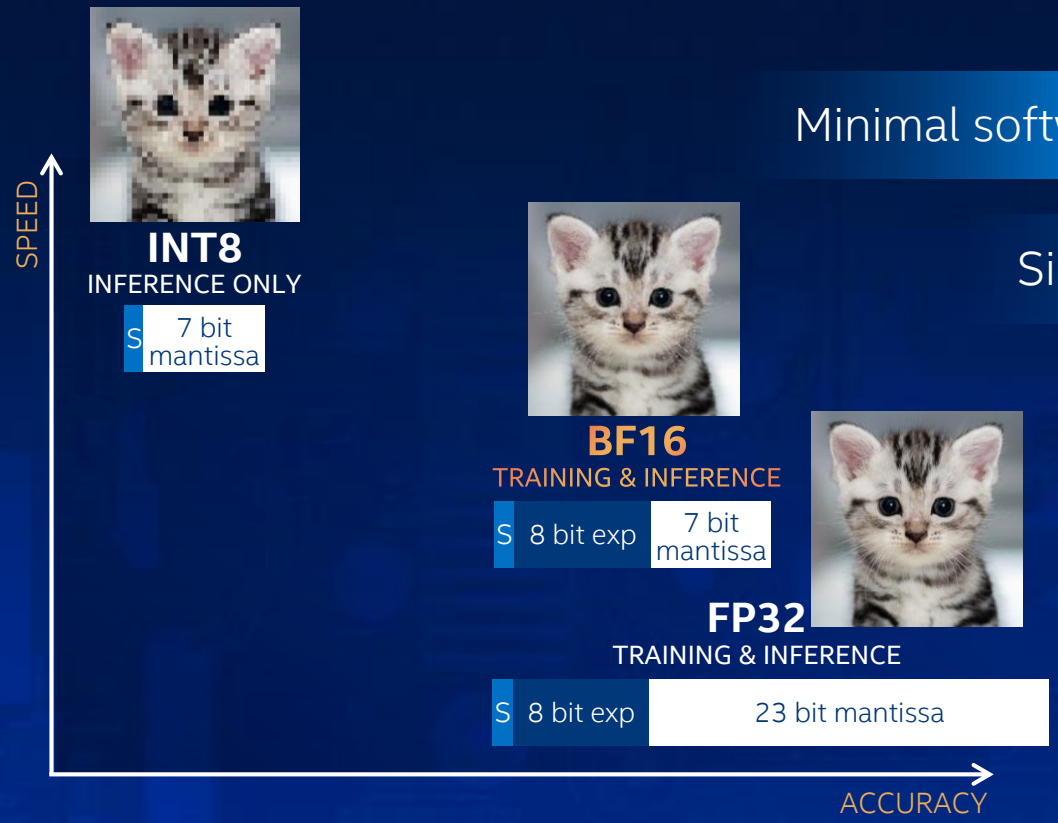
# Intel DL Boost Enhanced With Bfloat16

## The cutting edge of AI innovation

2020

3RD GEN
Intel Xeon Scalable

Intel
Deep Learning Boost
NEW: BF16




Minimal software changes

Similar accuracy to FP32

Performance results are based on testing as of dates in configuration and may not reflect all publicly available security updates. See backup for configuration details.
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

INTRODUCING

# Intel DL Boost Enhanced With Bfloat16

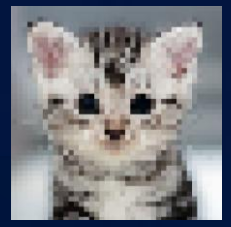
## The cutting edge of AI innovation

2020

3RD GEN
Intel Xeon Scalable

Intel
Deep Learning Boost
NEW: BF16

>100
OPTIMIZED TOPOLOGIES ON XEON



**INT8**
INFERENCE ONLY

S | 7 bit mantissa

SPEED

**BF16**
TRAINING & INFERENCE

S | 8 bit exp | 7 bit mantissa

**FP32**
TRAINING & INFERENCE

S | 8 bit exp | 23 bit mantissa

ACCURACY

Minimal software changes

Similar accuracy to FP32

up to 1.93X
HIGHER TRAINING PERFORMANCE
vs PRIOR GEN FP32

up to 1.9X
HIGHER INFERENCE PERFORMANCE
vs PRIOR GEN FP32

OPTIMIZED LIBRARIES & FRAMEWORKS

oneAPI | OpenVINO | ONNX RUNTIME | PyTorch | TensorFlow

Performance results are based on testing as of dates in configuration and may not reflect all publicly available security updates. See backup for configuration details.
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

intel

INTRODUCING

# Intel DL Boost Enhanced With Bfloat16

## The cutting edge of AI innovation

2020

**3RD GEN**
**Intel Xeon Scalable**

Intel
Deep Learning Boost
NEW: BF16

Alibaba Cloud
**1.58X** HIGHER THROUGHPUT
NLP - TEXTCNN

Alibaba Cloud
**1.83X** FASTER INFERENCE
NLP - BERT

ANT GROUP
**1.72X** FASTER TRAINING
VIDEO ANALYSIS

ANT GROUP
**1.81X** FASTER INFERENCE
VIDEO ANALYSIS

海鑫科金 HISIGN TECHNOLOGY
**1.97X** HIGHER THROUGHPUT
BIOMETRICS

hyve solutions
**1.86X** HIGHER TRAINING PERFORMANCE
IMAGE CLASSIFICATION

Matroid
**1.81X** HIGHER PROCESSING THROUGHPUT
VISUAL MEDIA SEARCH

Neusoft MEDICAL AI RESEARCH
**1.91X** HIGHER PROCESSING THROUGHPUT
MEDICAL IMAGES

Tencent Cloud
**1.68X** HIGHER THROUGHPUT
SEARCH ENGINE

Tencent Cloud
**1.54X** HIGHER THROUGHPUT
TTS - WAVERNN

Tencent Cloud
**1.89X** FASTER INFERENCE
TTS - PARALLEL WAVENET



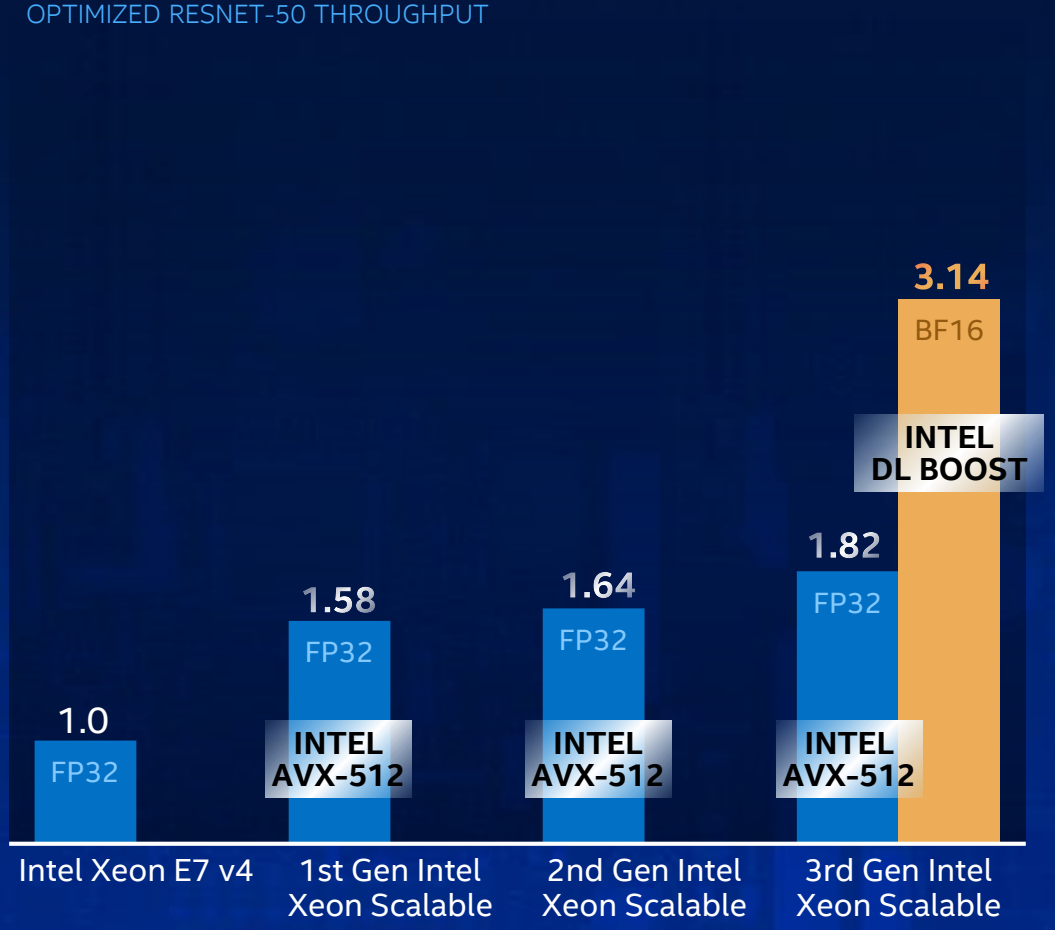
intel

3 Generations of

# Unequaled AI Performance Improvement

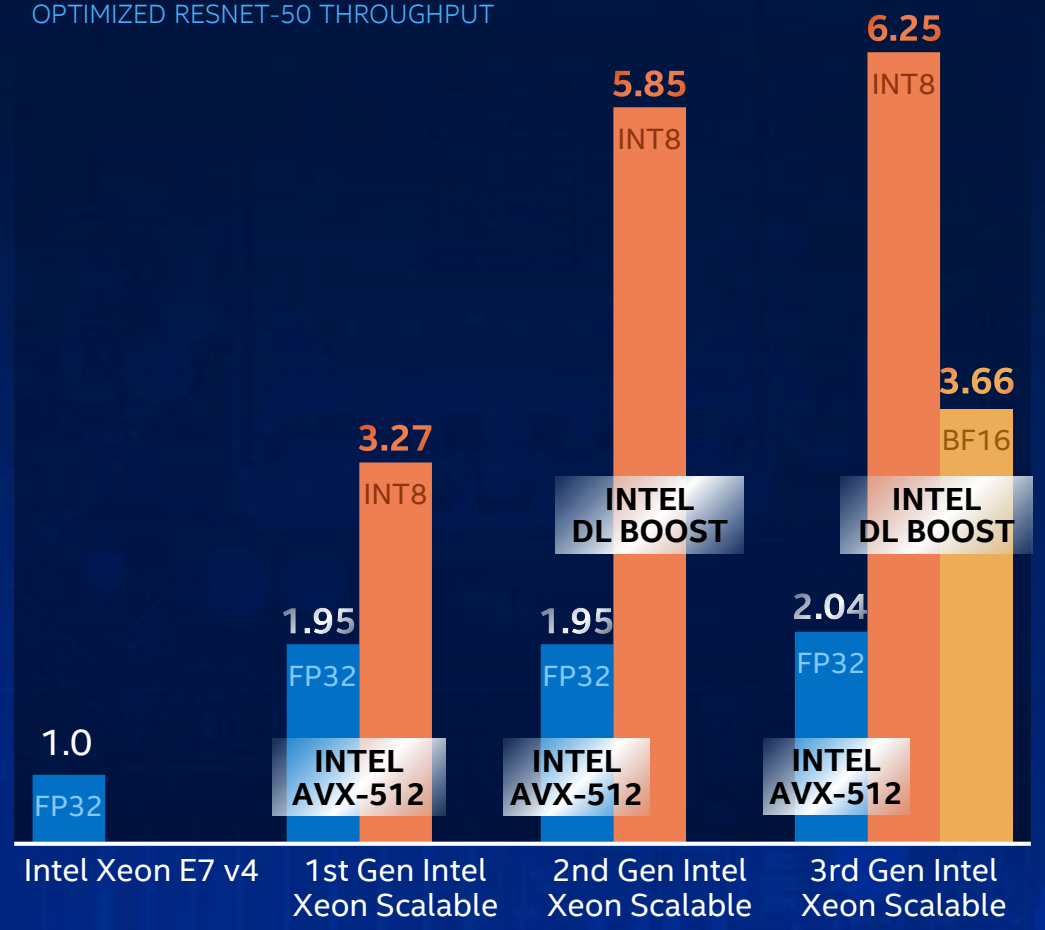
## AI TRAINING PERFORMANCE

OPTIMIZED RESNET-50 THROUGHPUT




## AI INFERENCE PERFORMANCE

OPTIMIZED RESNET-50 THROUGHPUT




Performance results are based on testing as of dates in configuration and may not reflect all publicly available security updates. See backup for configuration details.
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

# AI on Intel Xeon Scalable Processors

# Intel Xeon Scalable Roadmap

## 2019

**2ND GEN**
**Intel Xeon Scalable**

1-8
SOCKETS

**Cascade Lake**
PURLEY PLATFORM

## 2020

**3RD GEN**
**Intel Xeon Scalable**

4-8
SOCKETS

**Cooper Lake**
CEDAR ISLAND PLATFORM

**LAUNCHING**

1-2
SOCKETS

**Ice Lake**
WHITLEY PLATFORM

**COMING LATER THIS YEAR**

## 2021

**NEXT GEN**
**Intel Xeon Scalable**

1-8
SOCKETS

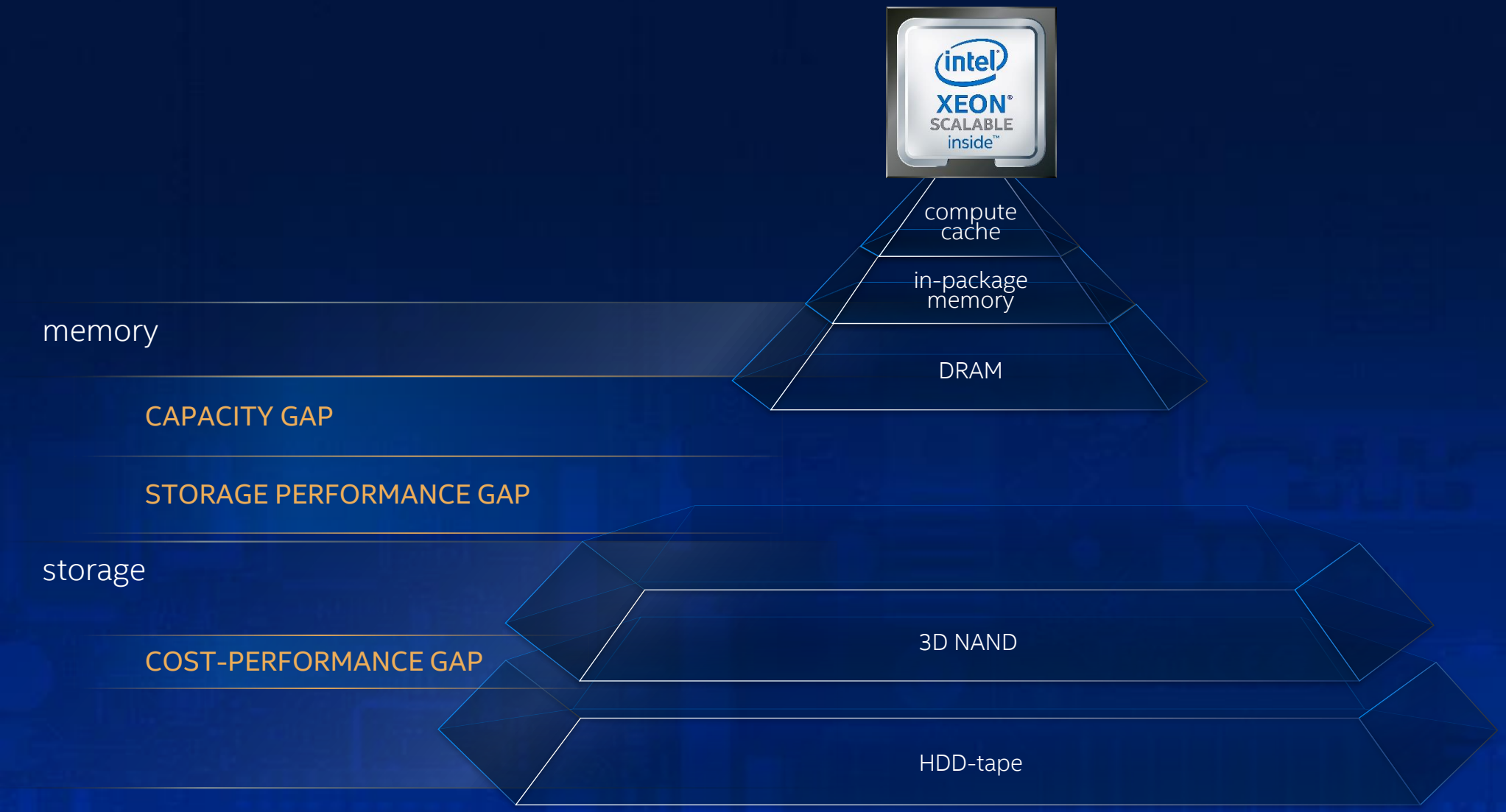
**Sapphire Rapids**
EAGLE STREAM PLATFORM

**NEXT GEN DL BOOST: AMX**

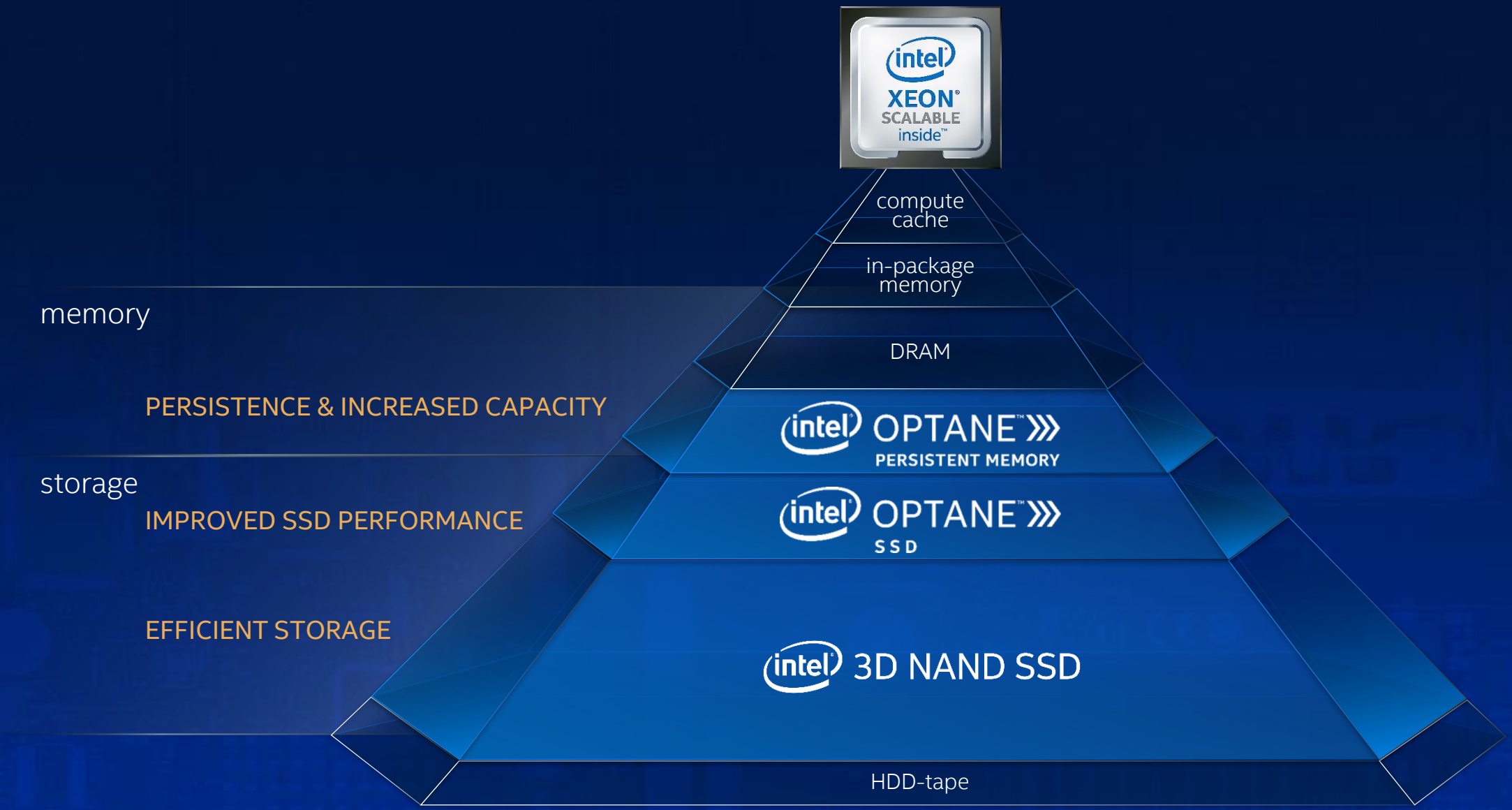
**SILICON POWERED ON**

# Sapphire Rapids Team

# Transforming Memory & Storage



# Transforming Memory & Storage



Intel Optane Persistent Memory

# Delivering Real World Benefits

**CUSTOMER TRACTION SINCE LAUNCH**

**OVER 200** FORTUNE 500

**OVER 85%** POC TO SALE CONVERSION

**OVER 270** PRODUCTION WINS

## TCO SAVINGS

| | |
|---|---|
| Kingsoft Cloud KSCLOUD | **1.3X improvement** IN TCO (REDIS) |
| 快手 KUAISHOU TECHNOLOGY | **30% reduction** IN RECOMMENDATION SYSTEM & REDIS SERVICE |
| MAX-PLANCK-GESELLSCHAFT | **15X reduction** IN MEMORY FOR IMAGE PROCESSING |
| 平安云 | **22.5%-48% improvement** IN TCO (REDIS) |
| SoftBank | **41% reduction** ON INFRASTRUCTURE COST |

## INCREASED THROUGHPUT

| | |
|---|---|
| CDW PEOPLE WHO GET IT StudioCloud | **~3X improvement** IN JOBS PER PHYSICAL HOST RATIO |
| GPORTAL | **2.78X increase** IN GAMES HOSTED ON A SINGLE SERVER |
| NITRADO SERVER HOSTING | **1.1X increase** IN CPU UTILIZATION VS. DRAM-ONLY |
| SK telecom | **~2X** VM INSTANTIATION FOR 5G MULTI-ACCESS EDGE (REDIS) |
| ZTO 中通快递 ZTO EXPRESS | **~40% more** VMS & CONTAINERS WITHIN SAME BUDGET (REDIS) |

## FASTER TIME TO INSIGHTS

| | |
|---|---|
| epcc | **8X faster** SOLVER RUN COMPARED TO LUSTRE FILESYSTEM |
| phoenixNAP GLOBAL IT SERVICES | **80%** LATENCY REDUCTION & **3X** ACCELERATED INDEXING (ELASTICSEARCH) |
| SIEMENS | **15X faster** DATABASE DATA LOAD STARTUP (SAP HANA) |
| T-Systems | **13.7X accelerated** DATABASE STARTUP (SAP HANA) |
| UC San Diego | **up to 17X faster** STORAGE APPLICATIONS (ROCKSDB, MONGODB, MYSQL) |

Performance results are based on testing as of dates in configuration and may not reflect all publicly available security updates. See backup for configuration details.
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

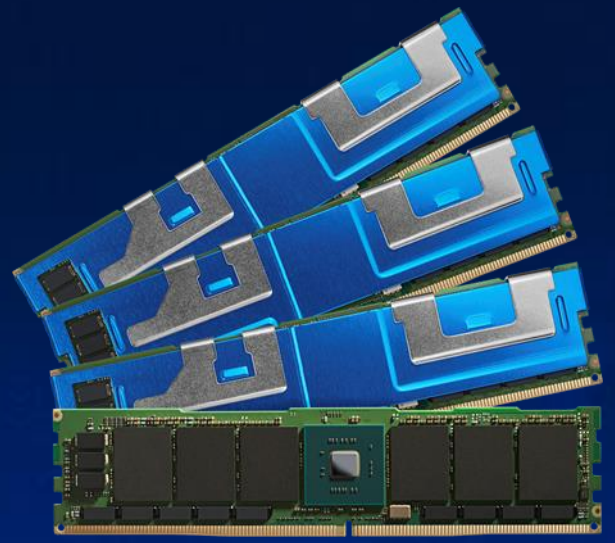
intel

LAUNCHING

# Intel Optane Persistent Memory 200 Series

## Making real-time big data analytics possible

up to **4.5TB**
TOTAL MEMORY
PER SOCKET



average **25%**
HIGHER MEMORY BANDWIDTH
vs PRIOR GEN

REDUCE I/O BOTTLENECKS TO ANALYZE DATA FASTER

over **225X**
FASTER ACCESS TO DATA
vs MAINSTREAM NAND SSD

BOOST APPLICATION PERFORMANCE

Performance results are based on testing as of dates in configuration and may not reflect all publicly available security updates. See backup for configuration details.
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

intel

LAUNCHING

# Intel 3D NAND SSD D7-P5500 & P5600

## Ready for the intense IO requirements of AI & analytics

up to **40%** LOWER LATENCY vs PRIOR GEN

intel 3D NAND SSD

DELIVERING AN OPTIMAL BALANCE OF PERFORMANCE AND CAPACITY

up to **33%** MORE PERFORMANCE vs PRIOR GEN

**Accelerates** all-flash arrays

**Advanced** IT efficiency & data security features

**Most advanced** TLC 3D NAND


Performance results are based on testing as of dates in configuration and may not reflect all publicly available security updates. See backup for configuration details.
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

# Speeding The Time To Deployment

## Intel Select Solutions for AI & Analytics

### ARTIFICIAL INTELLIGENCE

- AI INFERENCING
- BIGDL ON APACHE SPARK

### ANALYTICS

- update — MICROSOFT SQL SERVER (WINDOWS SERVER, LINUX)
- soon — SAP HANA
- PINGCAP TIDB
- GBASE
- update — MEDIA ANALYTICS
- new — TRANSWARP ARGO DB

### HPC

- SIMULATION & MODELING
- SIMULATION & VISUALIZATION
- update — GENOMICS ANALYTICS
- HPC & AI CONVERGED CLUSTERS (MAGPIE, UNIVA)

### HCI / STORAGE

- VMWARE VSAN
- new — VMWARE HORIZON VDI ON vSAN
- MICROSOFT AZURE STACK HCI
- new — NUTANIX HCI
- XSKY

intel select solution

ADVANTECH · ASUS · BOSTON Servers | Storage | Solutions · DataON · INSILICOGEN www.insilicogen.com · inspur 浪潮 · Lenovo

NUTANIX · PENGUIN COMPUTING · TRANSWARP · VAST · vmware · World Wide Technology

intel

# David Moore

Corporate Vice President
General Manager, Programmable Solutions Group
Data Platforms Group

# FPGAs Deliver Hardware Customization

## with integrated AI

# Enabling FPGA Developers With AI

**Ecosystem**

**Software**

mxnet · oneAPI · ONNX RUNTIME · OpenVINO · TensorFlow · BigDL · Spark · APACHE hadoop

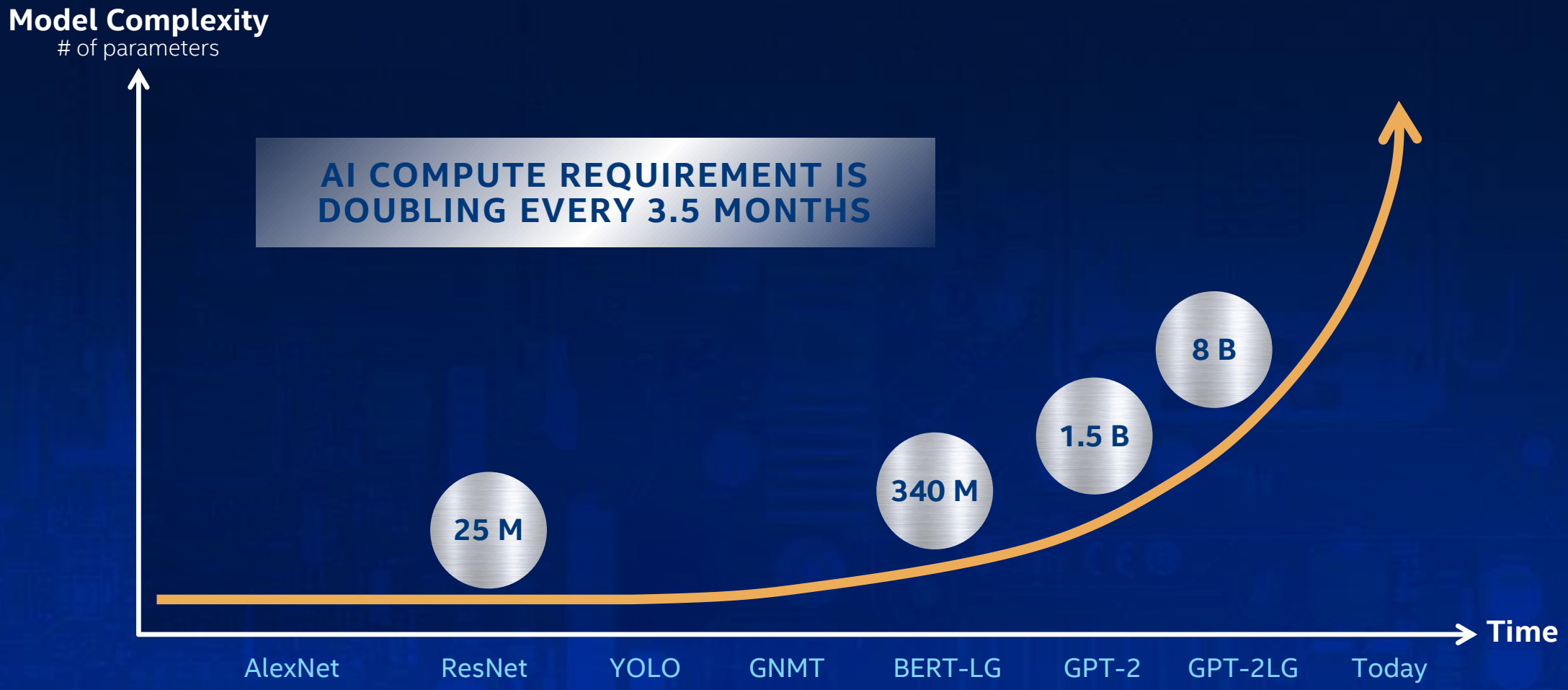
Intel® MKL

Intel® FPGA
Deep Learning Accelerator Suite


**Hardware**


intel

# Increasing AI Model Complexity
## requires innovation




Source: https://openai.com/blog/ai-and-compute/

DISCLOSING

# Intel Stratix 10 NX FPGA

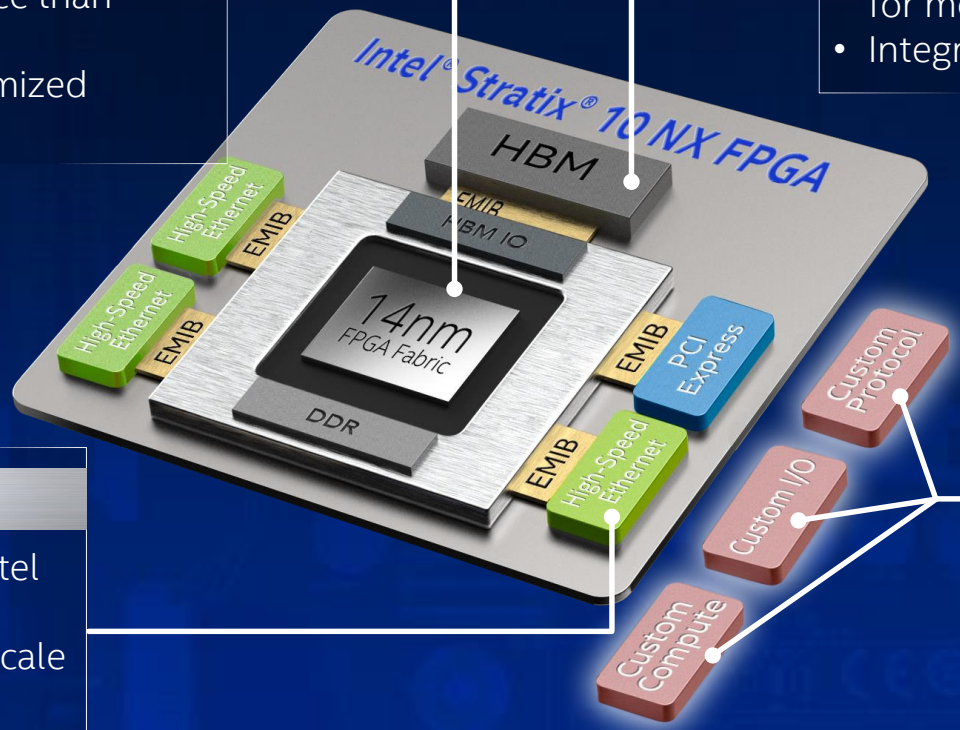
## Intel's first AI-optimized FPGA

### HIGH PERFORMANCE AI TENSOR BLOCKS

- Up to 15X more INT8 compute performance than today's Stratix 10 MX for AI workloads
- Hardware programmable for AI with customized workloads

### ABUNDANT NEAR-COMPUTE MEMORY

- Embedded and customizable memory hierarchy for model persistence
- Integrated HBM for high memory bandwidth




### HIGH BANDWIDTH NETWORKING

- Up to 57.8G PAM4 transceivers and hard Intel Ethernet blocks for high efficiency
- Flexible and customizable interconnect to scale across multiple nodes

### EXTENSIBLE

- Chiplets enable easier interface customization and ASIC extensions

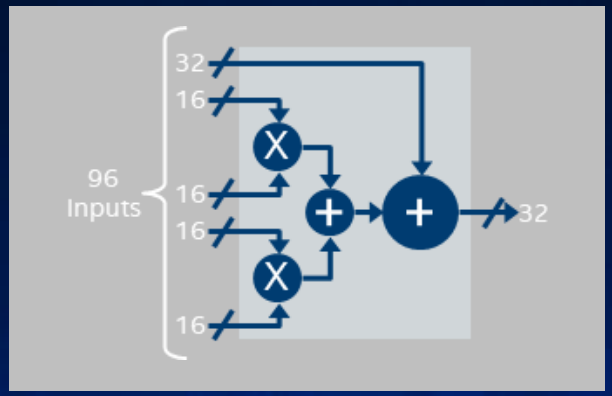
**TENSOR COMPUTE, NEAR MEMORY, AND NETWORKING DELIVERS HIGH PERFORMANCE HARDWARE OPTIMIZED FOR AI**

Performance results are based on Intel estimates. See backup for configuration details.
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

INTRODUCING

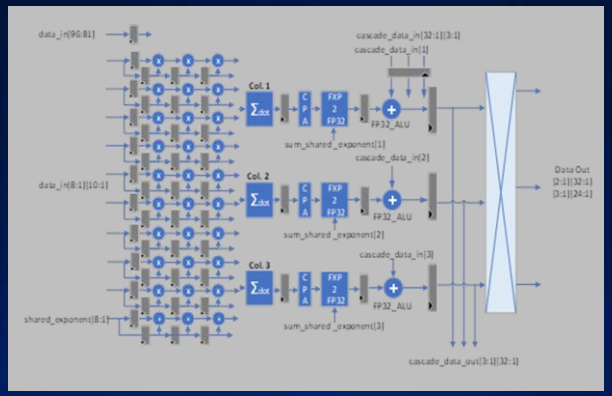
# AI Tensor Block

## INTEL STRATIX 10 MX
### DSP Block




2 MULTIPLIERS
2 ACCUMULATORS

## INTEL STRATIX 10 NX
### AI Tensor Block



30 MULTIPLIERS
30 ACCUMULATORS
INT4, INT8, BLOCK FP12, BLOCK FP16

**UP TO 15X MORE INT8 COMPUTE**
THAN STRATIX 10 MX

Performance results are based on Intel estimates. See backup for configuration details.
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

# AI Workload Examples

## for Intel Stratix 10 NX FPGA

### NATURAL LANGUAGE PROCESSING


- Custom FPGA pooling for large systems
- BERT batch 1 performance up to **2.3X** faster than Nvidia V100

### FRAUD DETECTION


- Direct network ingest for low latency data movement
- LSTM batch 1 performance up to **9.5X** faster than Nvidia V100

### SMART CITY


- Integrated video ingestion, data transformation, and AI for low and deterministic latency
- ResNet50 batch 1 performance up to **3.8X** faster than Nvidia V100

Performance results are based on Intel estimates and Nvidia's published specs. See backup for configuration details.
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

intel

DISCLOSING

# Intel Stratix 10 NX FPGA

## Silicon available later this year


FPGAs deliver hardware customization with integrated AI

AI Tensor block delivers up to 15X more INT8 compute performance than today's Stratix 10 MX for AI workloads

Intel Stratix 10 NX FPGA adds a new innovative capability to Intel's broad portfolio of silicon and software for AI

Microsoft

"As Microsoft designs our real-time multi-node AI solutions, we need flexible processing devices that deliver ASIC-level tensor performance, high memory and connectivity bandwidth, and extremely low latency. Intel Stratix 10 NX FPGAs meet Microsoft's high bar for these requirements, and we are partnering with Intel to develop next-generation solutions to meet our hyperscale AI needs."

**-Doug Burger**
Technical Fellow, Distinguished Engineer
Cloud & AI

**INTEL STRATIX 10 NX IS INTEL'S FIRST AI-OPTIMIZED FPGA**
FOR HIGH-BANDWIDTH, LOW-LATENCY AI ACCELERATION

Performance results are based on Intel estimates. See backup for configuration details.
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

intel

# Unleashing the Potential of Data

## MOVE FASTER

BAREFOOT NETWORKS | an Intel company

intel ETHERNET

intel SILICON PHOTONICS

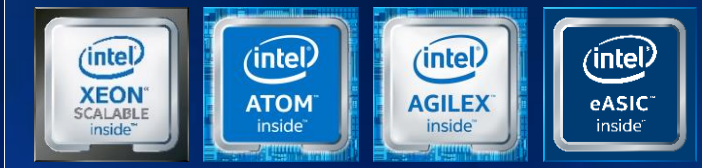
## STORE MORE

intel OPTANE™ PERSISTENT MEMORY

intel OPTANE™ SSD

intel 3D NAND SSD

## PROCESS EVERYTHING





## SOFTWARE & SYSTEM LEVEL OPTIMIZED

intel

# Notices and Disclaimers

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

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.


intel

### 3rd Gen Intel Xeon Scalable Processor: Built for today's AI-infused, data-intensive services

**Up to 1.98X higher database performance:** New: 1-node, 4x 3rd Gen Intel® Xeon® Platinum 8380H processor (pre-production 28C, 250W) on Intel Reference Platform (Cooper City) with 768 GB (24 slots / 32 GB / 3200) total memory, microcode 0x700001b, HT on, Turbo on, with Redhat 8.1, 4.18.0-147.3.1.el8_1.x86_64, 1x Intel 240GB SSD OS Drive, 2x6.4T P4610 for DATA, 2x3.2T P4610 for REDO, 1Gbps NIC, HammerDB 3.2, Popular Commercial Database, test by Intel on 5/13/2020. Baseline: 1-node, 4x Intel® Xeon® processor E7-8890 v3 on Intel Reference Platform (Brickland) with 1024 GB (64 slots / 16GB / 1600) total memory, microcode 0x16, HT on, Turbo on, with Redhat 8.1, 4.18.0-147.3.1.el8_1.x86_64, 1x Intel 800GB SSD OS Drive, 1x1.6T P3700 for DATA, 1x1.6T P3700 for REDO, 1Gbps NIC, HammerDB 3.2, Popular Commercial Database, test by Intel on 4/20/2020.

**Up to 1.9X average performance gain:** Average performance based on Geomean of est SPECrate®2017_int_base 1-copy, est SPECrate®2017_fp_base 1-copy, est SPECrate®2017_int_base, est SPECrate®2017_fp_base, Stream Triad, Intel distribution of LINPACK, Virtualization and OLTP Database workloads.

**SPECcpu_2017, Stream, LINPACK Performance:** New: 1-node, 4x 3rd Gen Intel® Xeon® Platinum 8380H processor (pre-production 28C, 250W) on Intel Reference Platform (Cooper City) with 768 GB (24 slots / 32 GB / 3200) total memory, microcode 0x87000016, HT on for SPECcpu, off for Stream, LINPACK), Turbo on, with Ubuntu 19.10, 5.3.0-48-generic, 1x Intel 240GB SSD OS Drive, est SPECcpu_2017, Stream Triad, Intel distribution of LINPACK, test by Intel on 5/15/2020. Baseline: 1-node, 4x Intel® Xeon® processor E7-8890 v3 on Intel Reference Platform (Brickland) with 512 GB (32 slots / 16 GB / 2133 (@1600)) total memory, microcode 0x16, HT on for SPECcpu, off for Stream, LINPACK), Turbo on, with Ubuntu 20.04 LTS, 5.4.0-29-generic, 1x Intel 480GB SSD OS Drive, est SPECcpu_2017, Stream Triad, Intel distribution of LINPACK, test by Intel on 5/15/2020.

**HammerDB OLTP Database Performance:** New: 1-node, 4x 3rd Gen Intel® Xeon® Platinum 8380H processor (pre-production 28C, 250W) on Intel Reference Platform (Cooper City) with 768 GB (24 slots / 32 GB / 3200) total memory, microcode 0x700001b, HT on, Turbo on, with Redhat 8.1, 4.18.0-147.3.1.el8_1.x86_64, 1x Intel 240GB SSD OS Drive, 2x6.4T P4610 for DATA, 2x3.2T P4610 for REDO, 1Gbps NIC, HammerDB 3.2, Popular Commercial Database, test by Intel on 5/13/2020. Baseline: 1-node, 4x Intel® Xeon® processor E7-8890 v3 on Intel Reference Platform (Brickland) with 1024 GB (64 slots / 16GB / 1600) total memory, microcode 0x16, HT on, Turbo on, with Redhat 8.1, 4.18.0-147.3.1.el8_1.x86_64, 1x Intel 800GB SSD OS Drive, 1x1.6T P3700 for DATA, 1x1.6T P3700 for REDO, 1Gbps NIC, HammerDB 3.2, Popular Commercial Database, test by Intel on 4/20/2020.

**Virtualization Performance:** New: 1-node, 4x 3rd Gen Intel® Xeon® Platinum 8380H processor (pre-production 28C, 250W) on Intel Reference Platform (Cooper City) with 1536 GB (48 slots / 32 GB / 3200 (@2933)) total memory, microcode 0x700001b, HT on, Turbo on, with RHEL-8.1 GA, 4.18.0-147.3.1.el8_1.x86_64, 1x Intel 240GB SSD OS Drive, 4x P4610 3.2TB PCIe NVME, 4 x 40 GbE x710 dual port, Virtualization workload, Qemu-kvm 2.12 (inbox), WebSphere 8.5.5, DB2 v9.7, Nginx 1.14.1, test by Intel on 5/20/2020. Baseline:1-node, 4x Intel® Xeon® processor E7-8890 v3 on Intel Reference Platform (Brickland) with 1024 GB (64 slots / 16GB / 1600) total memory, microcode 0x0000016, HT on, Turbo on, with RHEL-8.1 GA, 4.18.0-147.3.1.el8_1.x86_64, 1x Intel 240GB SSD OS Drive, 4x P3700 2TB PCIe NVME, 4 x 40 GbE x710 dual port, Virtualization workload, Qemu-kvm 2.12 (inbox), WebSphere 8.5.5, DB2 v9.7, Nginx 1.14.1, test by Intel on 5/20/2020.

### Intel DLBoost Enhanced with Bfloat16: The cutting edge of AI innovation

**Up to 1.93x Training Performance Improvement:** New: 1-node, 4x 3rd Gen Intel® Xeon® Platinum 8380H processor (pre-production 28C, 250W) on Intel Reference Platform (Cooper City) with 384 GB (24 slots / 16GB / 3200) total memory, ucode 0x700001b, HT on, Turbo on, with Ubuntu 20.04 LTS, Linux 5.4.0-26,28,29-generic, Intel 800GB SSD OS Drive, ResNet-50 v 1.5 Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit#828738642760358b388d8f615ded0c213f10c99a, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneDNN 1.4, BF16, BS=512, test by Intel on 5/18/2020. Baseline: 1-node, 4x Intel® Xeon® Platinum 8280 processor on Intel Reference Platform (Lightening Ridge) with 768 GB (24 slots / 32 GB / 2933 ) total memory, ucode 0x4002f00, HT on, Turbo on, with Ubuntu 20.04 LTS, Linux 5.4.0-26,28,29-generic, Intel 800GB SSD OS Drive, ResNet-50 v 1.5 Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit#828738642760358b388d8f615ded0c213f10c99a, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneDNN 1.4, FP32, BS=512, test by Intel on 5/18/2020.

**Up to 1.9x inference performance improvement:** New: 1-node, 4x 3rd Gen Intel® Xeon® Platinum 8380H processor (pre-production 28C, 250W) on Intel Reference Platform (Cooper City) with 384 GB (24 slots / 16GB / 3200) total memory, ucode 0x700001b, HT on, Turbo on, with Ubuntu 20.04 LTS, Linux 5.4.0-26,28,29-generic, Intel 800GB SSD OS Drive, BERT-Large (QA) Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit#828738642760358b388d8f615ded0c213f10c99a, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Squad 1.1 dataset, oneDNN 1.4, BF16, BS=32, 4 instances, 28-cores/instance, test by Intel on 5/18/2020. Baseline: 1-node, 4x Intel® Xeon® Platinum 8280 processor on Intel Reference Platform (Lightening Ridge) with 768 GB (24 slots / 32 GB / 2933 ) total memory, ucode 0x4002f00, HT on, Turbo on, with Ubuntu 20.04 LTS, Linux 5.4.0-26,28,29-generic, Intel 800GB SSD OS Drive, BERT-Large (QA) Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit#828738642760358b388d8f615ded0c213f10c99a, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Squad 1.1 dataset, oneDNN 1.4, FP32, BS=32, 4 instances, 28-cores/instance, test by Intel on 5/18/2020.

## Intel DLBoost Enhanced with Bfloat16: The cutting edge of AI innovation (customer examples)

**Hisign* Facial Recognition Throughput Performance on 3rd Gen Intel® Xeon® Scalable Processor:**

NEW: Tested by Intel as of 5/15/2020. 1-node, 4x Intel® Xeon® Platinum 8380H (pre-production) Processor on Intel Reference Platform, 28 cores HT On Turbo ON Total Memory 768 GB (24 slots/ 32GB/ 3200 MHz), BIOS: WLYDCRB1.SYS.0015.D19.2002140555 (microcode: 0x87000016), NIC: Intel X550T; Storage: 1x Intel 800GB SSD, OS: RedHat 8.1, 4.18.0-147.8.1.el8_1.x86_64, Framework: Internal Tensorflow 2.1 Branch: UTB, Commit id: 4c711446a4d42fa1ef8759602345fb75f50154ee, ssh://git@gitlab.devtools.intel.com:29418/TensorFlow/Direct-Optimization/private-tensorflow.git, Topology/ML Algorithm: customized FaceResNet, Compiler: GCC 8.3.1, MKL DNN, Dataset: Customer provided 4906 images, 128x128x3, Precision: BF16
BASELINE: Tested by Intel as of 5/15/2020. 1-node, 4x Intel® Xeon® Platinum 8280L Processor on Inspur NF8260M5, 28 cores HT On Turbo ON Total Memory 768 GB (24 slots/ 32GB/ 2933 MHz), BIOS: Inspur 4.1.10 (microcode: 0x400002C), NIC: Intel X550T; Storage: 1x Intel 800GB SSD, OS: RedHat 8.1, 4.18.0-147.8.1.el8_1.x86_64, Framework: Internal Tensorflow 2.1 Branch: UTB, Commit id: 4c711446a4d42fa1ef8759602345fb75f50154ee, ssh://git@gitlab.devtools.intel.com:29418/TensorFlow/Direct-Optimization/private-tensorflow.git, Topology/ML Algorithm: customized FaceResNet, Compiler: GCC 8.3.1, MKL DNN, Dataset: Customer provided 4906 images, 128x128x3, Precision: FP32

**TensorFlow on Neusoft Pathology Inference Throughput Performance on 3rd Gen Intel® Xeon® Scalable Processor:**

NEW: Tested by Intel as of 5/15/2020. 1-node, 4x Intel® Xeon® Platinum 8380H (pre-production) Processor on Intel Reference Platform, 28 cores HT On Turbo ON Total Memory 768 GB (24 slots/ 32GB/ 3200 MHz), BIOS: WLYDCRB1.SYS.0015.D19.2002140555 (microcode: 0x87000016), NIC: Intel X550T; Storage: 1x Intel 800GB SSD, OS: RedHat 8.1, 4.18.0-147.8.1.el8_1.x86_64, Framework: Internal Tensorflow 2.1 Branch: UTB, Commit id: 4c711446a4d42fa1ef8759602345fb75f50154ee, ssh://git@gitlab.devtools.intel.com:29418/TensorFlow/Direct-Optimization/private-tensorflow.git, Topology/ML Algorithm: customized DNN topology, Compiler: GCC 8.3.1, MKL DNN, Dataset: Customer provided 1728 images, 32x32x3, Precision: BF16
BASELINE: Tested by Intel as of 5/15/2020. 1-node, 4x Intel® Xeon® Platinum 8280L Processor on Inspur NF8260M5, 28 cores HT On Turbo ON Total Memory 768 GB (24 slots/ 32GB/ 2933 MHz), BIOS: Inspur 4.1.10 (microcode: 0x400002C), NIC: Intel X550T; Storage: 1x Intel 800GB SSD, OS: RedHat 8.1, 4.18.0-147.8.1.el8_1.x86_64, Framework: Internal Tensorflow 2.1 Branch: UTB, Commit id: 4c711446a4d42fa1ef8759602345fb75f50154ee, ssh://git@gitlab.devtools.intel.com:29418/TensorFlow/Direct-Optimization/private-tensorflow.git, Topology/ML Algorithm: customized DNN topology, Compiler: GCC 8.3.1, MKL DNN, Dataset: Customer provided 1728 images, 32x32x3, Precision: FP32

**AliCloud PAI Customized TextCNN on TF1.14 Run Time Performance on 3rd Gen Intel® Xeon® Scalable Processor:**

New: Tested by Intel as of 4/23/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family (Ali Customized SKU) Processor using Intel Reference Platform, 24 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x7000017), Storage: Intel SSDPE2KX010T7, NIC: 2x Intel Ethernet Controller 10G X550T, OS: CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized TextCNN(Confidential), BS=32, Dummy data, 4 instances/4 socket, Datatype: BF16
Baseline: Tested by Intel as of 4/23/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family (Ali Customized SKU) Processor, using Intel Reference Platform 24 cores HT On Turbo ON Total Memory 384 GB (24 slots / 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x7000017), Storage: Intel SSDPE2KX010T7, NIC: 2x Intel Ethernet Controller 10G X550T, OS: CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, MKL version: 2020.1.217, Customized TextCNN(Confidential), BS=32, Dummy data, 4 instances/4 socket, Datatype: FP32

**AliCloud PAI Customized BERT on TF1.14 Latency Performance on 3rd Gen Intel® Xeon® Scalable Processor:**

New: Tested by Intel as of 4/23/2020. 4 socket Intel® Xeon® Platinum 83xxH (Ali Customized SKU) Processor using Intel Reference Platform, 24 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x7000017), Storage: Intel SSDPE2KX010T7, NIC: 2x Intel ethernet Controller 10G x550T, OS: CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized BERT(Confidential), BS=1, MRPC data, 12 instance/4 socket, Datatype: BF16
Baseline: Tested by Intel as of 4/23/2020. 4 socket Intel® Xeon® Platinum 83xxH (Ali Customized SKU) Processor using Intel Reference Platform, 24 cores HT On Turbo ON Total Memory 384 GB (24 slots / 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x7000017), Storage: Intel SSDPE2KX010T7, NIC: 2x Intel ethernet Controller 10G x550T, OS:CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, MKL version: 2020.1.217, Customized BERT(Confidential), BS=1, MRPC data, 12 instance/4 socket, Datatype: FP32

## Intel DLBoost Enhanced with Bfloat16: The cutting edge of AI innovation (customer examples CONT)

**Alibaba Ant Financial Inference and Training on 3rd Gen Intel® Xeon® Scalable Processor:**

Tested by Intel as of 4/20/2020. 4 socket 3rd Gen Intel® Xeon® Scalable processor (18-core, 170W, pre-production) Processor using Intel Reference Platform, 18 cores HT OFF, Turbo ON Total Memory 768 GB (24 slots / 32GB / 2666), BIOS Version: 166.08 (6BC51780-BFDE-1000-03E6-000000000000) Microcode: 0x8600000b, CentOS 7.7.1908, 3.10.0-957.el7.x86_64, Deep Learning Framework: Pytorch Intel optimized Pytorch-1.0.0a0+3ca7205 https://gitlab.devtools.intel.com/cce-ai/pytorch, dnnl (mkldnn) commit id:7b53785 https://github.com/oneapi-src/oneDNN, Model: 3d CNN I3D, Compiler: gcc 7.3.1, Libraries: dnnl (mk-dnn), Dataset: UCF101 (size: 13320 shape: 3x64x224x224, Baseline Training: BS=24*4, FP32, New Training: BS=24*4, BF16; Baseline Inference: BS=32, 4 instances/4sockets, FP32, New Inference: BS=32, 4 instances/ 4 sockets, BF16.

**Tencent Search Engine Customized NLP model on TF1.14 Throughput Performance on 3rd Generation Intel® Xeon® Processor Scalable Family:**

New: Tested by Intel as of 4/28/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family(CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x86000017), CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, OneDNN version: DNNLv1.3, Customized NLP model(Confidential), BS=1, MRPC data, 8 instances/4 socket, Datatype: BF16
Baseline: Tested by Intel as of 4/28/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family (CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots / 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x86000017),CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, OneDNN version: DNNLv1.3, Customized NLP model(Confidential), BS=1, MRPC data, 8 instances/4 socket, Datatype: FP32

**Tencent Cloud Xiaowei Customized WaveRNN on MXNetv1.7 Throughput Performance on 3rd Generation Intel® Xeon® Processor Scalable Family:**

Opt. BF16 Solution: Tested by Intel as of 4/28/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family(CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x86000017), CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: MXNet1.7 https://github.com/apache/incubator-mxnet/tree/v1.7.x, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized WaveRNN(Confidential), BS=1 , Customer Provided data, 104 Instances/4 socket, Datatype: BF16
BASELINE(Opt. FP32 Solution): Tested by Intel as of 4/28/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family(CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots / 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x86000017),CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: MXNet1.7 https://github.com/apache/incubator-mxnet/tree/v1.7.x, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized WaveRNN(Confidential), BS=1, Customer Provided data, 104 Instances/4 socket, Datatype: FP32

**Tencent Cloud Xiaowei TTS P_Wavenet on TF1.14 Run Time Performance on 3rd Generation Intel® Xeon® Processor Scalable Family:**

New: Tested by Intel as of 5/11/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family(CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x86000017), CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized TTS Pwavenet(Confidential), BS=1, Customer Provided data, 4 instances/4 Socket, Datatype: BF16
Baseline: Tested by Intel as of 5/11/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family (CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots / 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x86000017),CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized TTS Pwavenet(Confidential), BS=1, Customer Provided data, 4 instances/4 Socket, Datatype: Datatype: FP32

**1.86x ResNet-50 Training Throughput Performance Improvement on Catalina platform with BF16:**

1-node, 8x 3rd Gen Intel® Xeon® Platinum 8380H processor( 28C) on Catalina with 768 GB (48 slots / 16GB / 3200) total memory, microcode 0x86000017, HT on, Turbo on, Ubuntu 20.04 LTS(Host | Ubuntu 18.04 (Docker) Kernel 5.4.0-28-generic (Host), 1x INTEL_SSDSC2BX01, 8x Intel E810-C, ResNet-50 v 1.5 Throughput, Intel optimized TensorFlow 2.2, https://github.com/Intel-tensorflow/tensorflow/commits/bf16/base, https://github.com/IntelAI/models/blob/v1.6.1/models/image_recognition/tensorflow/ResNet50v1_5/training/mlperf_resnet/resnet_model.py, gcc version 7.5.0 (docker) , ImageNet Challenge 2012 Dataset, oneDNN v1.4, FP32 and BF16,, test by Intel on 05/24/2020, *16-node projected performance

## Intel DLBoost Enhanced with Bfloat16: The cutting edge of AI innovation (customer examples CONT)

**Matroid 1.81X higher processing throughput:** TensorFlow on Matroid Inference Throughput Performance on 3rd Gen Intel® Xeon® Scalable Processor:
NEW: Tested by Intel as of 6/8/2020. 1-node, 4x Intel® Xeon® Platinum 8380H (pre-production) Processor on Intel Reference Platform, 28 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 3200 MHz), BIOS: WLYDCRB1.SYS.0015.P96.2005070242 (microcode: 0x700001b), NIC: Intel I210; Storage: 1x INTEL SSDSC2KG96 800GB SSD, OS: RedHat 8.0, 4.18.0-80.el8.x86_64, Framework: TensorFlow 2.2.0 (custom tensorflow-mkl), Topology/ML Algorithm: Custom CNN; Neural Architecture Search, Compiler: GCC 7.3.0, MKL DNN 2020.1, Python 3.7.0, Dataset: Customer provided images - 320x294x3, Precision: BF16
BASELINE: Tested by Intel as of 6/8/2020. 1-node, 4x Intel® Xeon® Platinum 8380H (pre-production) Processor on Intel Reference Platform, 28 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 3200 MHz), BIOS: WLYDCRB1.SYS.0015.P96.2005070242 (microcode: 0x700001b), NIC: Intel I210; Storage: 1x INTEL SSDSC2KG96 800GB SSD, OS: RedHat 8.0, 4.18.0-80.el8.x86_64, Framework: TensorFlow 2.2.0 (Eigen), Topology/ML Algorithm: Custom CNN; Neural Architecture Search, Compiler: GCC 7.3.0, Python 3.7.0, Dataset: Customer provided images - 320x294x3, Precision: FP32

## 3 Generations of Unequaled AI Performance Improvement

**ResNet-50 Inference Throughput Performance multi-gen Improvement**

3rd Gen Intel Xeon Scalable Processor (Cooper Lake): 1-node, 4x 3rd Gen Intel® Xeon® Platinum 8380H processor (pre-production 28C, 250W) on Intel Reference Platform (Cooper City) with 384 GB (24 slots / 16GB / 3200) total memory, ucode 0x700001b, HT on, Turbo on, with Ubuntu 20.04 LTS, Linux 5.4.0-29-generic, Intel SSD 800GB OS Drive, Inference: ResNet-50 v 1.5 Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit#6ef2116e6a09, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneDNN 1.4, FP32, INT8-VNNI, BF16, BS=128, 4 instances, 28-cores/instance, test by Intel on 06/01/2020.

2nd Gen Intel Xeon Scalable Processor (Cascade Lake): 1-node, 4x Intel® Xeon® Platinum 8280 processor on Intel Reference Platform (Lightning Ridge) with 768 GB (24 slots / 32 GB / 2933 ) total memory, ucode 0x4002f00, HT on, Turbo on, with Ubuntu 20.04 LTS, Linux 5.4.0-29-generic, Intel SSD 800GB OS Drive, Inference: ResNet-50 v 1.5 Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit# 6ef2116e6a09, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneDNN 1.4, FP32, INT8-VNNI, BS=128, 4 instances, 28-cores/instance, test by Intel on 06/01/2020.

Intel Xeon Scalable Processor (Skylake): 1-node, 4x Intel® Xeon® Platinum 8180 processor on Intel Reference Platform (Lightning Ridge) with 768 GB (24 slots / 32 GB / 2966 ) total memory, ucode 0x2000069, HT on, Turbo on, with Ubuntu 20.04 LTS, 5.4.0-26-generic, Intel SSD 800GB OS Drive, Inference: RN50-v1.5 Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit#6ef2116e6a09, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneDNN 1.4, FP32, INT8, BS=128, 4 instances, 28-cores/instance, test by Intel on 6/02/2020.

Intel Xeon processor E7 v4 (Broadwell): 1-node, 4x Intel® E7-8890 v4processor on Intel Reference Platform (Brickland) with 512 GB (32 slots /16GB/ 1600) total memory, ucode 0xb000038, HT on, Turbo on, with Ubuntu 20.04 LTS, Linux 5.4.0-29-generic, Intel SSD 800GB OS Drive, Inference: RN50-v1.5 Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit#6ef2116e6a09, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneDNN 1.4, FP32, BS=128, 4 instances, 24-cores/instance, test by Intel on 6/08/2020.

**ResNet-50 Training Performance multi-gen Improvement**

3rd Gen Intel Xeon Scalable Processor: 1-node, 4x 3rd Gen Intel® Xeon® Platinum 8380H processor (pre-production 28C, 250W) on Intel Reference Platform (Cooper City) with 384 GB (24 slots / 16GB / 3200) total memory, ucode 0x700001b, HT on, Turbo on, with Ubuntu 20.04 LTS, Linux 5.4.0-29-generic, Intel SSD 800GB OS Drive, ResNet-50 v 1.5 Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit#6ef2116e6a09, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneDNN 1.4, FP32, BF16, global BS=1024, 4 instances, 28-cores/instance, test by Intel on 06/01/2020.

2nd Gen Intel Xeon Scalable Processor: 1-node, 4x Intel® Xeon® Platinum 8280 processor on Intel Reference Platform (Lightning Ridge) with 768 GB (24 slots / 32 GB / 2933 ) total memory, ucode 0x4002f00, HT on, Turbo on, with Ubuntu 20.04 LTS, Linux 5.4.0-29-generic, Intel SSD 800GB OS Drive, ResNet-50 v 1.5 Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit# 6ef2116e6a09, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneDNN 1.4, FP32, global BS=1024, 4 instances, 28-cores/instance, test by Intel on 06/01/2020.

Intel Xeon Scalable Processor: 1-node, 4x Intel® Xeon® Platinum 8180 processor on Intel Reference Platform (Lightning Ridge) with 768 GB (24 slots / 32 GB / 2966 ) total memory, ucode 0x2000069, HT on, Turbo on, with Ubuntu 20.04 LTS, 5.4.0-26-generic, Intel SSD 800GB OS Drive, Training: RN50-v1.5,Inference: RN50-v1.5 Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit#6ef2116e6a09, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneDNN 1.4, FP32, global BS=1024, 4 instances, 28-cores/instance, test by Intel on 6/02/2020.

Intel Xeon processor E7 v4: 1-node, 4x Intel® E7-8890 v4processor on Intel Reference Platform (Brickland) with 512 GB (32 slots /16GB/ 1600) total memory, ucode 0xb000038, HT on, Turbo on, with Ubuntu 20.04 LTS, Linux 5.4.0-29-generic, Intel SSD 800GB OS Drive, Training: RN50-v1.5,Inference: RN50-v1.5 Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit#6ef2116e6a09, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneDNN 1.4, FP32, global BS=1024 , 4 instances, 24-cores/instance, test by Intel on 6/08/2020.

## Intel Optane Persistent Memory: Delivering Real World Benefits

**Kingsoft Cloud REDIS service*** (self-defined workload); OS: Red Hat Enterprise Linux* 7.5 4.18.8-x86_64. Testing by Intel and Kingsoft Cloud completed on Jan 10, 2019. Security Mitigations for Variants 1, 2, 3 and L1TF in place. BASELINE: 2nd Gen Intel® Xeon® Platinum 8260 processor, 2.3 GHz, 24 cores, turbo, and HT on, BIOS 1.018, 1536GB total memory, 12 slots / 64GB / 2666 MT/s / DDR4 LRDIMM, 1 x 480GB / Intel® SSD DC S4500 + 1 x 1TB / Intel® SSD DC P4500. NEW: 2nd Gen Intel® Xeon® Platinum 8260 processor, 2.3GHz, 24 cores, turbo and HT on, BIOS 1.018, 1536GB total memory, 12 slots / 16GB / 2933 MT/s / DDR4 LRDIMM and 12 slo/ 128 GB / Intel® Optane™ DC persistent memory, 1 x 480GB / Intel® SSD DC S4500 + 1 x 1TB / Intel® SSD DC P4500. For more complete information about performance and benchmark results, visit: https://www.intel.com/content/www/us/en/processors/xeon/scalable/software-solutions/kingsoft-cloud-redis-service.html.

**Kuaishou Technolgy:** Test results are based on Kuaishou's internal tests and evaluation. For more details, please contact Kuaishou https://www.intel.cn/content/www/cn/zh/architecture-and-technology/kuaishou-recommendation-system-and-redis-services-storage-upgrade.html.

**Max Planck**: 2x 6248 CPUs with 2-2-2 128GB Apache Pass modules configured in memory mode. 32GBx12 DDR4 2666MHz RAM, CentOS* 7.6. For more complete information about performance and benchmark results, visit https://www.intel.com/content/www/us/en/customer-spotlight/stories/max-planck-institute-customer-story.html.

**Ping An Cloud** Total Cost of Ownership (TCO): This cost reduction data is derived from the joint calculation by of Ping An Cloud and Intel. For more complete information about performance and benchmark results, visit https://www.intel.com/content/www/us/en/customer-spotlight/stories/ping-an-cloud-customer-story.html.

**SoftBank** Results of the validation at SoftBank: Intel® Xeon® Silver 4114 processor: 40 cores with Intel® Hyper-Threading Technology enabled, 512 GB 1VM resource at 30VM capacity: 1.3 cores, 17.0 GB. Intel® Xeon® Gold 6222V processor: 80 cores with Intel® Hyper-Threading Technology enabled, 1536 GB 1VM resource at 60VM capacity: 1.3 cores, 25.6 GB. The information was described as of 13th December 2019. For more complete information about performance and benchmark results, visit https://www.intel.com/content/www/us/en/customer-spotlight/stories/ping-an-cloud-customer-story.html.

**CDW Canada StudioCloud**. Test results are based on StudioCloud internal tests and evaluation as of December 2019. For more complete information about performance and benchmark results, visit https://www.cdw.ca/content/cdwca/en/industries/studiocloud.html.

**GPORTAL**: 1 Baseline Configuration: Dell EMC PowerEdge R640 server; 2x Intel® Xeon® Gold 6154 processor @ 3.0 GHz (18 cores/36 threads); 768 GB DDR4; BIOS = 2.3.10; OS = Linux Results: 180 Minecraft game instances DUT Configuration: Dell EMC PowerEdge R640 server; 2x Intel® Xeon® Platinum 8268 processor @ 2.90 GHz (24 cores/48 threads); 12 x 32 GB DDR4 + 12 x 128 GB Intel® Optane™ DC persistent memory modules; BIOS = 2.3.10; OS = Linux Results: 500 Minecraft game instances. Testing by GPORTAL as of 5 December 2019.

**Nitrado**: Testing by Nitrado as of February 7, 2019. All-DRAM configuration: Dual-socket Intel® Xeon® Gold 6148 processor (8x 64 GB DDR4-2666 DRAM), total memory installed = 512 GB. System memory available = 512 GB. Number of Minecraft* instances: 182. CPU utilization: 40%. DRAM + Intel® Optane™ DC persistent memory configuration: Dual-socket Intel® Xeon® Gold 6252 processor (12x 128 GB (1.5TB) Intel® Optane™ DC persistent memory plus 12x 16 GB (192 GB) DDR4-2600 DRAM), total memory installed = 1,692 GB. System memory available = 1,536 GB. Number of Minecraft instances: 500. CPU utilization: 85%. Final results were extrapolated from Nitrado's testing data. For more complete information about performance and benchmark results, visit https://www.intel.com/content/www/us/en/customer-spotlight/stories/nitrado-online-gaming-customer-story.html.

**SK Telecom:** Testing conducted by SKT and Intel as of June 7, 2019. For more complete information about performance and benchmark results, visit https://builders.intel.com/docs/networkbuilders/case-study-of-scaled-up-skt-5g-mec-reference-architecture.pdf.

**ZTO Express**: For more complete information about performance and benchmark results, visit https://www.intel.com/content/www/us/en/customer-spotlight/stories/zto-express-improves-infrastructure-video.html.

**EPCC**: Performance results provided by EPCC and may not reflect all released security updates. No product can provide absolute security. For more complete information about performance and benchmark results, visit https://www.intel.com/content/www/us/en/customer-spotlight/stories/edinburgh-parallel-computing-center-customer-story.html.

**phoenixNAP and Panzura configurations**: Up to 3x indexing and 80% cache latency decrease – based on phoenixNAP and Panzura testing as of March 2019 on Elasticsearch: Intel® Xeon® Gold 6230 processor, Total Memory 256 GB RAM, 1.5TB of Intel® Optane™ DC persistent memory, HyperThreading: Enabled, Turbo: Enabled, ucode: 0x043, OS: ('centos-release-7-5.1804.el7.centos.x86_64'), Kernel: (3.10.0-862) vs. AWS i3xlarge (Intel) Instance, Elasticsearch, Memory: 30.5GB, Hypervisor: KVM, Storage Type: EBS Optimized, Disk Volume: 160GB, Total Storage: 960GB, Elasticsearch version: 6.3. For more complete information about performance and benchmark results, visit https://www.intel.com/content/www/us/en/customer-spotlight/stories/phoenixnap-panzura-customer-story.html.

**Siemens**: 15X faster database data load at startup performance results are based Siemens testing in April 2019 and may not reflect all released security updates. No product can provide absolute security. For more complete information about performance and benchmark results, visit https://www.intel.com/content/www/us/en/customer-spotlight/stories/siemens-in-memory-processing-customer-story.html.

## Intel Optane Persistent Memory: Delivering Real World Benefits (CONT)

**T-Systems:** Testing by T-Systems as of March 18, 2019. Baseline Configuration Hardware: HPE Superdome Flex* server with 4x CPU sockets (Intel® Xeon® Platinum processor Beta 8276M 2.20 GHz ; Memory = 4x6 256 GB Intel® Optane™ DC persistent memory (6 TB) - DEACTIVATED and 4x6 64 GB DDR4 Memory (1.5 TB) for a total memory configuration of 1.5 TB Software: Database: 4 TB SAP S/4HANA* database in App Direct Mode; OS: Standard SUSE Linux Enterprise Server* 12 Service Pack 4 microcode = 0xb00002e; kernel = Linux 4.12.14-95.16, standard NetApp cDot*-based storage used for persistence; SAP HANA 2.0 SPS4 rev. 40 installation with BW-Benchmark workload Re-start time: 10, 248 seconds (approximately 2.85 hours) Proof of Concept Configuration Hardware: HPE Superdome Flex* server with 4x CPU sockets (Intel® Xeon® Platinum processor Beta 8276M 2.20 GHz ; Memory = 4x6 256 GB Intel® Optane™ DC persistent memory (6 TB) and 4x6 64 GB DDR4 Memory (1.5 TB) for a total memory configuration of 7.5 TB Software: Database: 4 TB SAP S/4HANA* database in App Direct Mode; OS: Standard SUSE Linux Enterprise Server* 12 Service Pack 4 microcode = 0xb00002e; kernel = Linux 4.12.14-95.16, standard NetApp cDot*-based storage used for persistence; SAP HANA 2.0 SPS4 rev. 40 installation with BW-Benchmark workload Re-start Time: 748 seconds (approximately 12.47 minutes). For more complete information about performance and benchmark results, visit https://www.intel.com/content/www/us/en/customer-spotlight/stories/t-systems-in-memory-database-customer-story.html.

**UC San Diego**: Testing conducted by UC San Diego as of August 8, 2019. For more complete information about performance and benchmark results, visit https://arxiv.org/pdf/1903.05714.pdf.

## Intel Optane Persistent Memory 200 Series: Making real-time big data analytics possible

**>225X faster access to data:** Intel® Optane persistent memory idle read latency of 340 nanoseconds. Intel® SSD DC P4610 Series TLC NAND solid state drive idle read latency of 77 microseconds.

**Average of 25% higher memory bandwidth vs prior gen:** Baseline: 1-node, 1x Intel® Xeon® 8280L 28C @ 2.7GHz processor on Neon City with Single PMem module config (6x32GB DRAM; 1x{128GB,256GB,512GB} Intel Optane PMem 100 Series module at 15W) ucode Rev: 04002F00 running Fedora 29 kernel 5.1.18-200.fc29.x86_64, and MLC ver 3.8 with App-Direct. Source: 2020ww18_CPX_BPS_DI. Tested by Intel, on 27 Apr 2020. New configuration: 1-node, 1x Intel® Xeon® pre-production CPX6 28C @ 2.9GHz processor on Cooper City with Single PMem module config (6x32GB DRAM; 1x{128GB,256GB,512GB} Intel Optane PMem 200 Series module at 15W), ucode pre-production running Fedora 29 kernel 5.1.18-200.fc29.x86_64, and MLC ver 3.8 with App-Direct. Source: 2020ww18_CPX_BPS_BG. Tested by Intel, on 31 Mar 2020.

## Intel 3D NAND SSD D7-P5500 & P5600

**Up to 40% lower latency:** Source – Intel. Comparing datasheet figures for 4KB Random Write QD1 latency between the Intel® SSD D7-P5500 Series 7.68TB and Intel® SSD DC P4510 Series 8TB with both drives running on PCIe 3.1. Measured latency was 15µs and 25µs for the D7-P5500 and DC P4510, respectively. Performance for both drives measured using FIO Linux CentOS 7.2 kernel 4.8.6 with 4KB (4096 bytes) of transfer size with Queue Depth 1 (1 worker). Measurements are performed on a full Logical Block Address (LBA) span of the drive once the workload has reached steady state but including all background activities required for normal operation and data reliability. Power mode set at PM0. Any differences in your system hardware, software or configuration may affect your actual performance. Intel expects to see certain level of variation in data measurement across multiple drives.

**Up to 33% more performance:** Source – Intel. Comparing datasheet figures for 4KB Random Read QD256 performance between the Intel® SSD D7-P5500 Series 7.68TB and Intel® SSD DC P4510 Series 8TB with both drives running on PCIe 3.1. Measured performance was 854K IOPS and 641.8K IOPS for the D7-P5500 and DC P4510, respectively. Performance for both drives measured using FIO Linux CentOS 7.2 kernel 4.8.6 with 4KB (4,096 bytes) of transfer size with Queue Depth 64 (4 workers). Measurements are performed on a full Logical Block Address (LBA) span of the drive once the workload has reached steady state but including all background activities required for normal operation and data reliability. Power mode set at PM0. Any differences in your system hardware, software or configuration may affect your actual performance. Intel expects to see certain level of variation in data measurement across multiple drives.

## Intel Stratix 10 NX FPGA

**Up to 15X more INT8 compute performance than today's Stratix 10 MX for AI workloads:** When implementing INT8 computations using the standard Stratix 10 DSP Block, there are 2 multipliers and 2 accumulators used. On the other hand, when using the AI Tensor Block, you have 30 multipliers and 30 accumulators. Therefore 60/4 provides up to 15X more INT8 compute performance when comparing the AI Tensor Block with the standard Stratix 10 DSP block.

**BERT 2.3X faster, LSTM 9.5X faster, ResNet50 3.8X faster**: BERT batch 1 performance 2.3X faster than Nvidia V100 (DGX-1 server w/ 1x NVIDIA V100-SXM2-16GB | TensorRT 7.0 | Batch Size = 1 | 20.03-py3 | Precision: Mixed | Dataset: Sample Text); LSTM batch 1 performance 9.5X faster than Nvidia V100 (Internal server w/Intel® Xeon® CPU E5-2683 v3 and 1x NVIDIA V100-PCIE-16GB | TensorRT 7.0 | Batch Size = 1 | 20.01-py3 | Precision: FP16 | Dataset: Synthetic); ResNet50 batch 1 performance 3.8X faster than Nvidia V100 (DGX-1 server w/ 1x NVIDIA V100-SXM2-16GB | TensorRT 7.0 | Batch Size = 1 | 20.03-py3 | Precision: INT8 | Dataset: Synthetic). Estimated on Stratix 10 NX FPGA using -1 speed grade, tested in May 2020. Each end-to-end AI model includes all layers and computation as described in Nvidia's published claims as of May 2020. Result is then compared against Nvidia's published claims. Link for Nvidia: https://developer.nvidia.com/deep-learning-performance-training-inference. Results have been estimated or simulated using internal Intel analysis, architecture simulation, and modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.