

Performance
made flexible.

3rd Gen Intel® Xeon® Scalable Platform Technology Preview

Lisa Spelman

Corporate Vice President, Xeon and Memory Group
Data Platforms Group
Intel Corporation

Unmatched Portfolio of Hardware, Software and Solutions

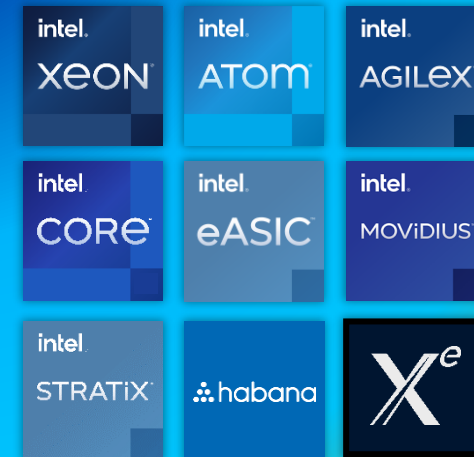
Move Faster



Store More



Process Everything



Optimized Software and System-Level Solutions

Announcing Today – Intel's Latest Data Center Portfolio

Targeting 3rd Gen Intel® Xeon® Scalable processors

Move Faster



Intel® Ethernet E810-2CQDA2

Up to 200GbE per PCIe 4.0 slot for bandwidth-intensive workloads

Store More



Intel® Optane™ SSD P5800X

Fastest SSD on the planet



Intel® Optane™ Persistent Memory 200 series

Up to 6TB memory per socket
+ Native data persistence



Intel® SSD D5-P5316

First PCIe 4.0 144-Layer QLC 3D NAND
Enables up to 1PB storage in 1U

Process Everything



3rd Gen Intel® Xeon® Scalable processor

Intel's highest performing server CPU
with built-in AI and security solutions

Intel® Agilex™ FPGA

Industry leading FPGA logic performance and
performance/watt

Optimized Solutions

intel
SELECT
SOLUTIONS

intel
MARKET
READY

>500
Partner Solutions

World's Most Broadly Deployed Data Center Processor

From the cloud to the intelligent edge

>50M

Intel® Xeon® Scalable
Processors Shipped

>1B Xeon Processor Cores
Deployed in the Cloud Since
2013¹

>800 cloud providers
with Intel® Xeon® processor-
based servers deployed¹

Intel® Xeon® Scalable Processors are
the Foundation for Multi-Cloud Environments

1. Source: Intel internal estimate using IDC Public Cloud Services Tracker and Intel's Internal Cloud Tracker, as of March 1, 2021.

Why Customers Choose Intel

Delivering workload-optimized performance

Platform Features

>2X more TLS requests processed with new Intel Crypto acceleration for fast, predictable response and lower TCO



Comprehensive Portfolio

Realized 2:1 server consolidation while quickly accessing and re-formatting images and video using Intel® Xeon® processors, Intel® Optane™ and Intel FPGAs



AI Capabilities

1.8X performance improvement for Monte Carlo simulation with no accuracy loss using Intel DL Boost and oneAPI-optimized software



Higher Performance Through Optimized Software

3.78X faster inference with lower power and higher utilization after switching from GPU text recognition to Intel® Xeon® processors with optimized PyTorch



Partnership and Co-Engineering

Faster TTM and 20X faster boot time using Intel's early access programs and engineering support to create SMB continuity and recovery solution



Solutions

Quickly scaled from 8-node POC to 68-node full production COVID-19 research cluster using Intel Select Solution for Genomics Analytics



Platform to deliver true flexibility

Performance varies by use, configuration and other factors. Configurations see appendix [45]

Performance made flexible.

3rd Gen Intel® Xeon® Scalable processors

Built specifically for our customers' needs



Optimized for Cloud, Enterprise, HPC, 5G and Edge

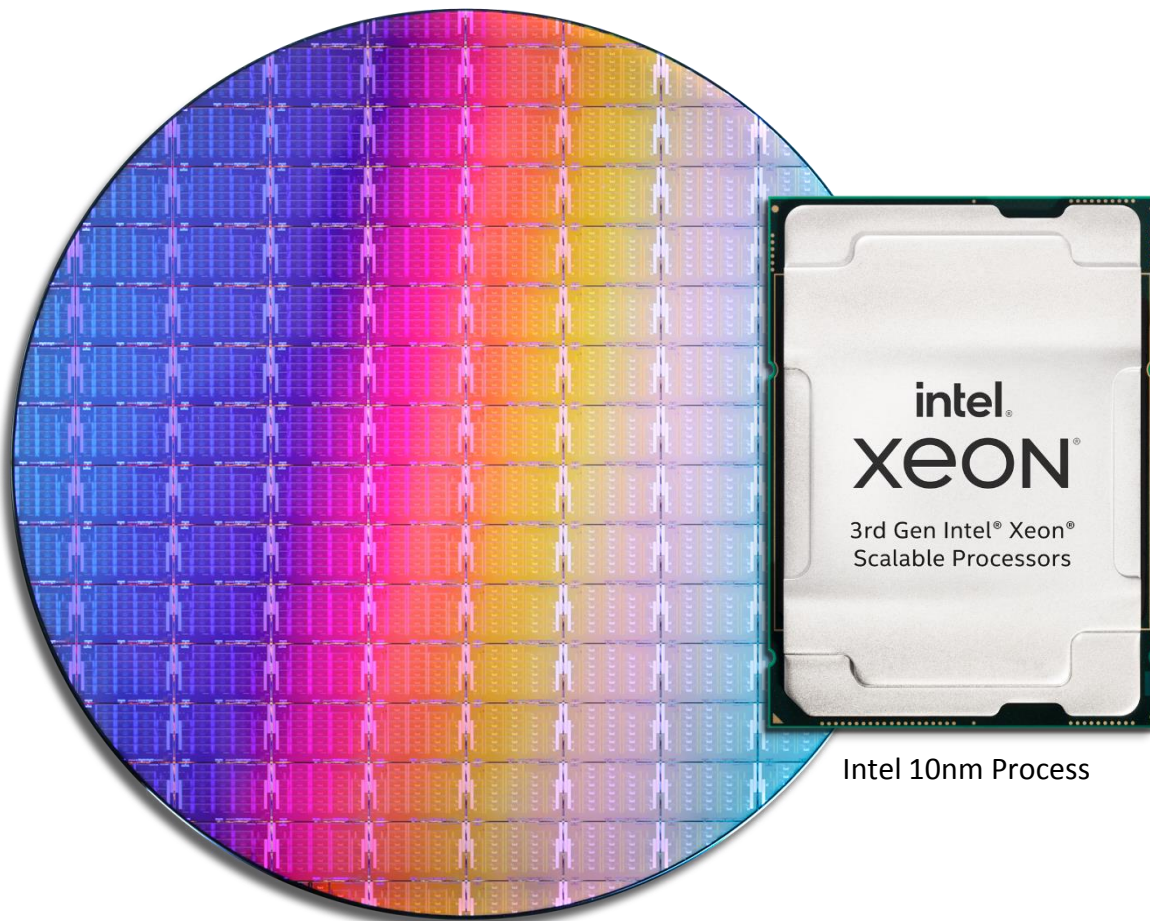
Built-in security with Intel Software Guard Extensions, Platform Firmware Resilience and Total Memory Encryption

Only data center processor with built-in AI (Intel DL Boost)

Built-in crypto acceleration reduces the performance impact of pervasive encryption

3rd Gen Intel® Xeon® Scalable processors

Performance made flexible



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]

3rd Gen Intel® Xeon® Scalable processors

Performance made flexible

Only x86 data center processor with
built-in Artificial Intelligence

Advanced security solutions

Scalable, flexible, customizable



Targeted for 1S-2S systems



Intel Software
Guard Extensions



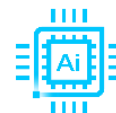
Intel
Crypto
Acceleration



Intel Total Memory
Encryption



Intel Platform
Firmware
Resilience



Intel Deep
Learning Boost



Intel Speed Select
Technology



Intel
AVX-512



Optimized
Software

Next-gen Xeon Scalable Platform

Up to
6TB

System Memory
Capacity
(Per Socket)
DRAM + PMem

Up to
8CH

DDR4-3200
2 DPC
(Per Socket)

Up to
2.6X

Memory Capacity
Increase vs.
2nd Gen Xeon
Scalable

Up to
64

Lanes
PCI Express 4
(per Socket)

Breakthrough Data Performance



Intel®
Optane™
persistent
memory 200
series



Intel®
Optane™ SSD
P5800X series



Intel® SSD
D series

Faster, Flexible, Data Scale



Intel® Ethernet
800 series
network adapters



Intel® Agilex™
FPGA
solutions



Flexible Performance for Most Demanding Workloads

Outstanding gen-on-gen performance from intelligent edge to cloud



Cloud

UP TO
1.5x

Improvement in
Latency Sensitive
Workloads



5G

UP TO
1.62x

Improvement
in Network and
Communications
Workloads



IoT

UP TO
1.56x

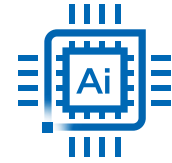
Image Classification
Inference
Improvement



HPC

UP TO
1.57x

Faster Modeling
for Critical Vaccine
Research



Artificial
Intelligence

UP TO
1.74x

Language Processing
Inference
Improvements

Performance varies by use, configuration and other factors. Configurations see appendix [5,7, 17, 19, 52]

Performance made flexible.

3rd Gen Intel® Xeon® Scalable Processors

Broad industry adoption

Cloud Service Providers

- All top CSPs planning to deploy services
- Over 200 ISVs and partners have deployed Intel SGX



OEM/ODMs

- >250 designs with more than 50 unique OxM partners
- Broad software and ecosystem readiness to speed faster time to value



Network Providers

- >15 major TEMs, OEMs and CoSPs readying POCs or initial deployments



HPC Labs and HPCaaS

- Over 20 publicly-declared HPC adopters to date
- HPC solutions coming from all major OEMs
- Growing HPCaaS footprint including:



>200k units shipped in Q1 2021

Only Data Center CPU with Built-in AI Acceleration

Artificial intelligence made flexible



Vs. Competition

- Up to 25x performance on image recognition vs. AMD EPYC 7763
- Xeon leads on average across a broad mix of 20 popular AI & ML workloads:
 - **Up to 1.5x** vs. AMD EPYC 7763
 - **Up to 1.3x** vs. Nvidia A100 GPU

Intel Software Optimizations

- **Up to 10x** performance improvement with TensorFlow for deep learning on ResNet50 vs. default distro
- **Up to 100x** improvement with Scikit-Learn for machine learning on SVC/kNN predict vs. the default distro



Performance

Medical imaging solution exceeds performance requirements using Intel DL Boost powered by OneAPI



Productivity

Drive-thru recommendation engine using Analytics Zoo to unify end-to-end Spark data pipeline on a Xeon-based cluster



Simplicity

Automating underground pipe inspection using a solution developed with Intel AI Builders partner Wipro

150+ Containers and 200+ Turnkey Solutions Accelerate Development and Deployment

Performance varies by use, configuration and other factors. Configurations see appendix [26, 28, 29, 36, 54]

Performance made flexible.

Intel's Most Secure, Compliant and Performant Data Center Platform

Security made flexible



Built-in Intel Software Guard Extensions

- Hundreds of customers currently using Intel SGX to enhance security and enable business transformation through data privacy
- Smallest attack surface of any data center confidential computing technology
- 4000x* increase in protected enclave size to 1TB

Introducing New Security Features

- Intel Total Memory Encryption for basic bulk encryption of the entire memory space to protect against physical attacks
- Intel Platform Firmware Resilience for defending and recovering the underlying firmware layer to protect against permanent denial of service



- Among the largest health insurance providers in Germany
- Intel SGX selected as Trusted Execution Environment for federated digital health record management serving over 26 million people



- Globally \$2T is laundered each year and can be used to fund crime or terrorism
- Intel SGX allows banks to share records without revealing private customer information and reduce false positives 83%

*3rd Gen Intel® Xeon® Scalable processors provides up to 1TB enclave size in a 2S configuration (SKU dependent) vs Xeon E 2100 256MB

AOK and Consilient systems initially deployed on Intel® Xeon® E processors

Built-in Crypto Acceleration for Encryption-Intensive Workloads

Encryption made flexible



Intel Crypto Acceleration

- New instructions and architectural features parallelize execution of encryption functions
- Reduces penalty of implementing pervasive data encryption
- Higher throughput with fast and strong encryption for AVX-512 ISA
- Increases performance of encryption-intensive workloads such as SSL web server, 5G infrastructure and VPN/firewalls

4.2x

Encrypted Web Server

More TLS encrypted web server connections per second;
more content delivered per server

NGINX

1.94x

Vector Packet Processing

More encrypted packets processed per second;
higher network and VPN capacity per node

ipsec 

Performance varies by use, configuration and other factors. Configurations see appendix [17, 51]

The Power of the Intel Portfolio

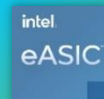
Delivering workload innovation and customer results



5G vRAN

2x

mMIMO throughput in a similar power envelope for a best-in-class 3x100MHz 64T64R configuration



Virtual Desktop (VDI)

1.87x

More VMs per node

300

Compute-intensive VMs per server



Azure Stack HCI

2x

Throughput with lower read and write latency



Performance varies by use, configuration and other factors. Configurations see appendix [39]

Performance made flexible.

Intel® Optane™ Persistent Memory 200 Series

Persistent memory made flexible

Average of
32% higher
memory bandwidth
compared to 100 series



Up to
6 TB total memory
per socket
for faster analysis of the largest
data sets



eADR (Enhanced Asynchronous DRAM Refresh) improves performance of apps that use persistent memory by eliminating “cache flushes”

Persistent memory ecosystem continues to grow



KATANA GRAPH

Computes up to 2X faster graph analytics algorithms used in search, social networks, and fraud detection

vmware®

Lower infrastructure costs by up 25% per VM while delivering the same performance

Performance varies by use, configuration and other factors. Configurations see appendix [53]

Performance made flexible.

Intel® Optane™ SSD P5800X

The world's fastest data center SSD



Up to **26X** more IOPS/GB

Up to **66X** better quality of service

Up to **13X** lower average latency

intel OPTANE SSD

All vs. Intel® SSD D7-P5600 NAND



Optane SSDs accelerate slower bulk storage to increase responsiveness of HCI, VDI, databases and more



Uses a small number of P5800X's per server to absorb writes from a 100GbE network, allowing them to focus on streamlining bulk storage




Achieves a 2.5x bandwidth improvement when moving data to QLC capacity storage

Performance varies by use, configuration and other factors. Configurations see appendix [46]

Performance made flexible.

Intel® SSD D5-P5316

Massive storage capacity made flexible



Up to
7 GB/s

Read-optimized
performance saturates
PCIe 4.0

Up to
48% **better latency**
Compared to previous gen QLC SSD

Up to
5X **higher endurance**
Compared to previous gen QLC SSD

intel
3D NAND SSD

Industry-leading density + no compromise quality and reliability enables customers to confidently deploy at scale in warm storage


Increased per server bandwidth accelerates access to large datasets in CDN, HCI, Big Data, AI, HPC, and Cloud Elastic Storage

Reduce storage footprint by up to 20x

Performance varies by use, configuration and other factors. Configurations see appendix [47]

Intel® Ethernet 800 Series Network Adapters

Connectivity made flexible



Up to **2X** more resources
for greater VM and container density

PCIe 3.0,
PCIe 4.0 and OCP
NIC 3.0

Up to **100** Gbps per port
and **200** Gbps per adapter
with new E810-2CQDA2 adapter

intel.
ETHERNET

Prioritizes application traffic to help deliver the performance required for high-priority, network intensive workloads

Fully programmable pipeline to enable frame classification for advanced and proprietary protocols

Supports RDMA over iWARP or RoCEv2 protocols and NVMe over TCP with ADQ for high throughput, low latency storage, cloud and HPC clusters

New E810-2CQDA2 adapter targets high-performance workloads such as vRAN, NFV forwarding plane, storage, HPC, cloud and content delivery networks

Source: 2x more resources from 100Gbps to 200Gbps

Intel® Agilex™ FPGA + Quartus Prime 20.4 Software

FPGA performance made flexible



~2X Better fabric performance per watt compared to competing 7nm FPGAs

Up to 400 Gbps Ethernet
Industry's highest data rate SerDes transceivers

intel. AGILEX™

Industry-leading FPGA offering customers the adaptability and agility to innovate for diverse workload transitions in 5G, Network, Cloud, and Edge

~2X better fabric performance per watt compared to competing 7nm FPGAs enabling flexible, energy efficient designs for the data center and beyond

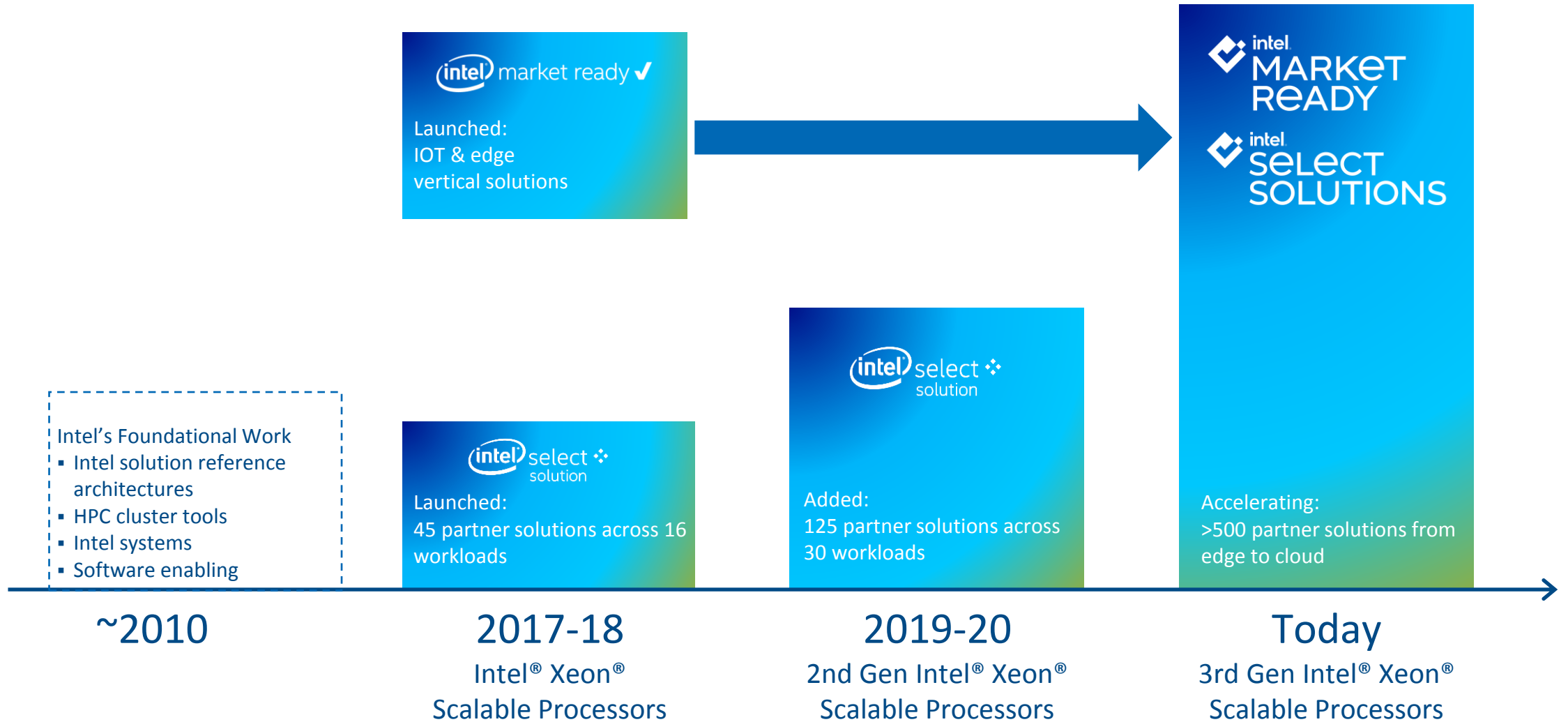
50% faster Video IP performance vs. competing 7nm FPGAs for video processing applications

Up to 49% faster fabric performance for high-speed 5G fronthaul gateway applications critical to enable high speed fiber-based connections

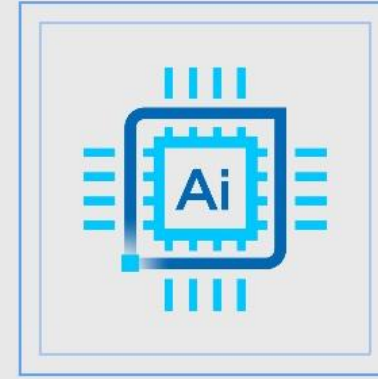
Performance varies by use, configuration and other factors. Configurations see appendix [48]

Decades of Solutions Delivery

Accelerating customer deployments



Technology Demonstration



Performance Made Flexible

Tencent 腾讯 **intel**

Performance made flexible.

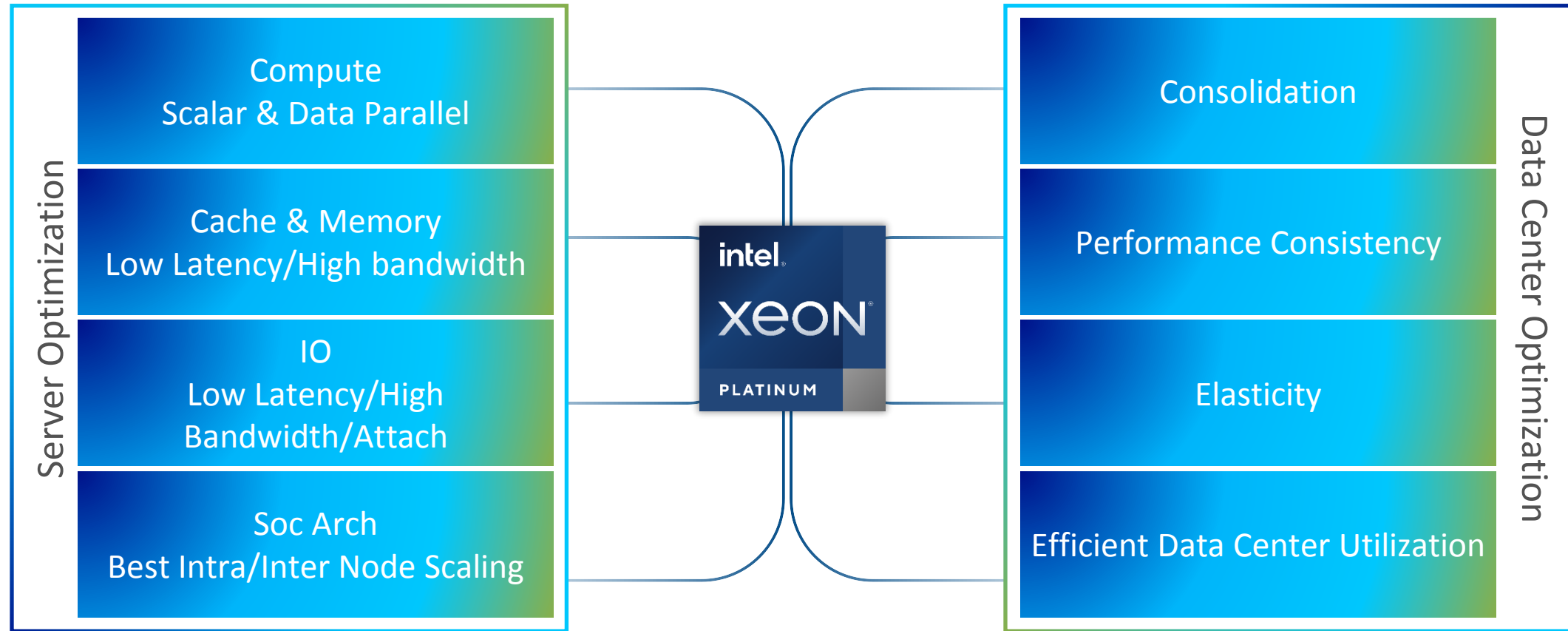
3rd Gen Intel® Xeon® Scalable Platform Architectural Deep Dive

Sailesh Kottapalli

Senior Fellow, Chief Architect
Data Center Processor Architecture
Intel Corporation

Xeon® Architecture

Industry's best latency and throughput



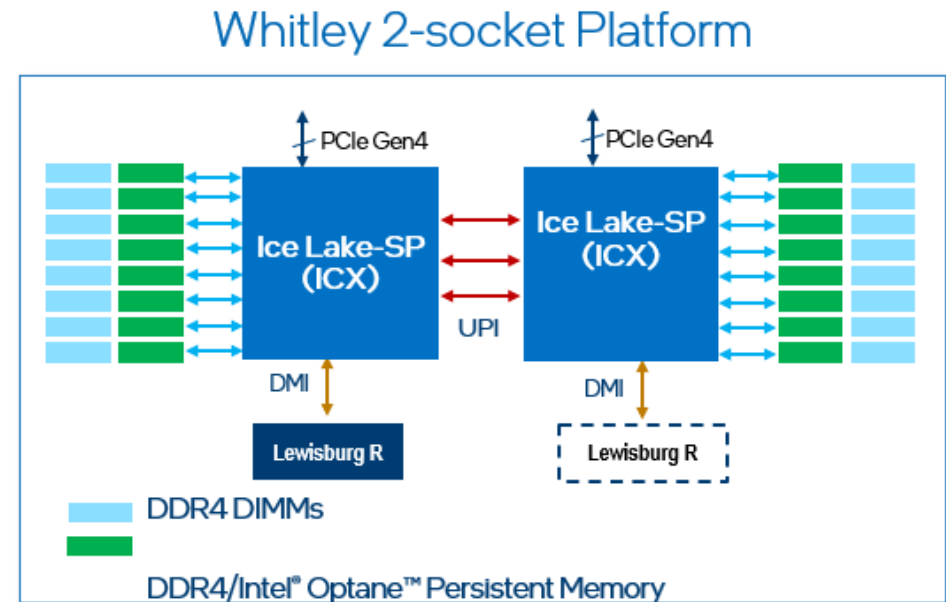
Perf @SLA, Perf @TCO, Perf @Scale

Ice Lake-SP

3rd Gen Intel® Xeon® Scalable Processor

Scalable and balanced architecture delivered through advancements in all technology pillars

- **Compute:** Next generation core with significant IPC improvements and new ISA instructions
- **Memory:** Significant improvement in DDR and Intel® Optane™ Persistent Memory performance
- **IO:** 64 lanes PCIe Gen4 and 3 high-speed UPI links
- **Intra/Inter Node Scaling:** Consistent low latencies to cache, memory and inter socket



Compute Microarchitecture

Sunny Cove Core

- Improved Front-end: higher capacity and improved branch predictor
- Wider and deeper machine: wider allocation , larger structures and execution resources
- Enhancements in TLBs, single thread execution, prefetching
- Datacenter optimized capabilities – larger Mid-level Cache (L2) + higher Vector throughput

	Cascade Lake (per core)	Ice Lake (per core)
Out-of-order Window	224	352
In-flight Loads + Stores	72 + 56	128 + 72
Scheduler Entries	97	160
Register Files – Integer + FP	180 + 168	280 + 224
Allocation Queue	64/thread	70/thread; 140/1 thread
L1D Cache (KB)	32	48
L2 Unified TLB (STLB)	1.5K	2K
STLB-IG Page support	16	1024 (shared w/4K)
STLB-IG Page support	16	1024 (shared w/4K)
Mid-level Cache (MB)	1	1.25

Compute Architecture

New instructions

Cryptography

- Big-Number Arithmetic (AVX-512 Integer IFMA)
- Vector AES and Vector Carry-less Multiply Instructions
- Galois Field New Instructions (GFNI)
- SHA-NI

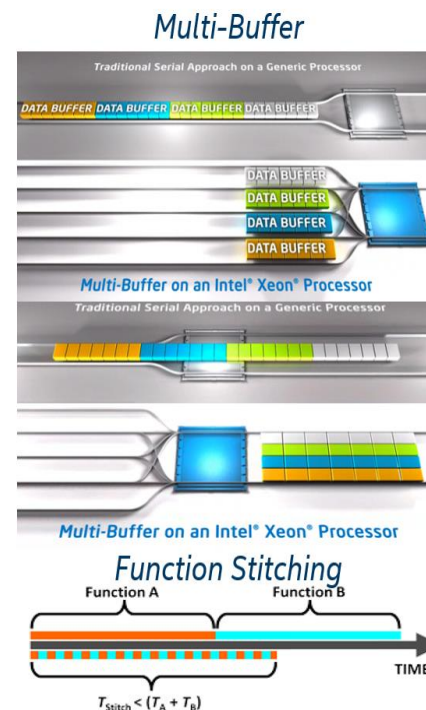
Compression/Decompression and Special SIMD

- Bit Algebra
- VBMI – Vector Bit Manipulation Instruction

New SIMD ISA Utilizing AVX512 on ICX



Software / Algorithms



Ice Lake vs. Cascade Lake Per Core Performance

ECDHE x25519	4.12X
RSA Sign 2048	5.63X
ECDHE p256	2.73X
AES-CTR	3.84X
AES-CMAC	3.78X
AES-XTS	3.5X
AES-GCM	3.34x
ECDSA Sign p256	1.9X
CRC	2.3X
ZUC	1.5X

Cache, Memory & IO

Latency and coherency optimizations

Cache

- Shared LLC: >1.5x increase over CLX
- Hemisphere mode
- Innovations to improve latency, BW and SOC scaling

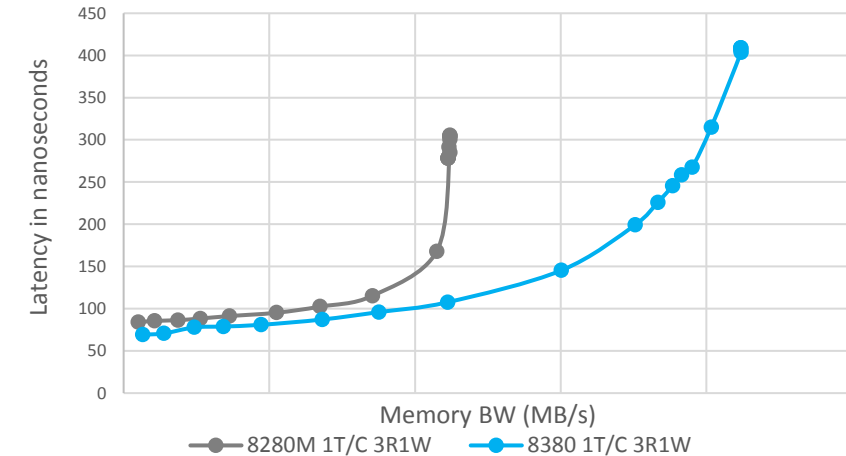
Memory

- 8 channels of DDR4 3200
- Memory scheduler improvements for lower effective latency and higher BW

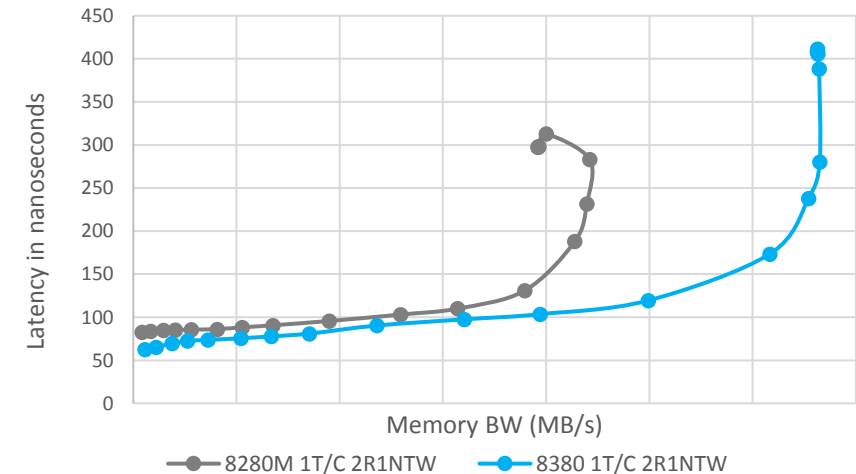
IO

- 64 lanes PCIe Gen4
- 3 UPI links at 11.2 GT/s
- IO latency improvements

3R1W traffic with RFO

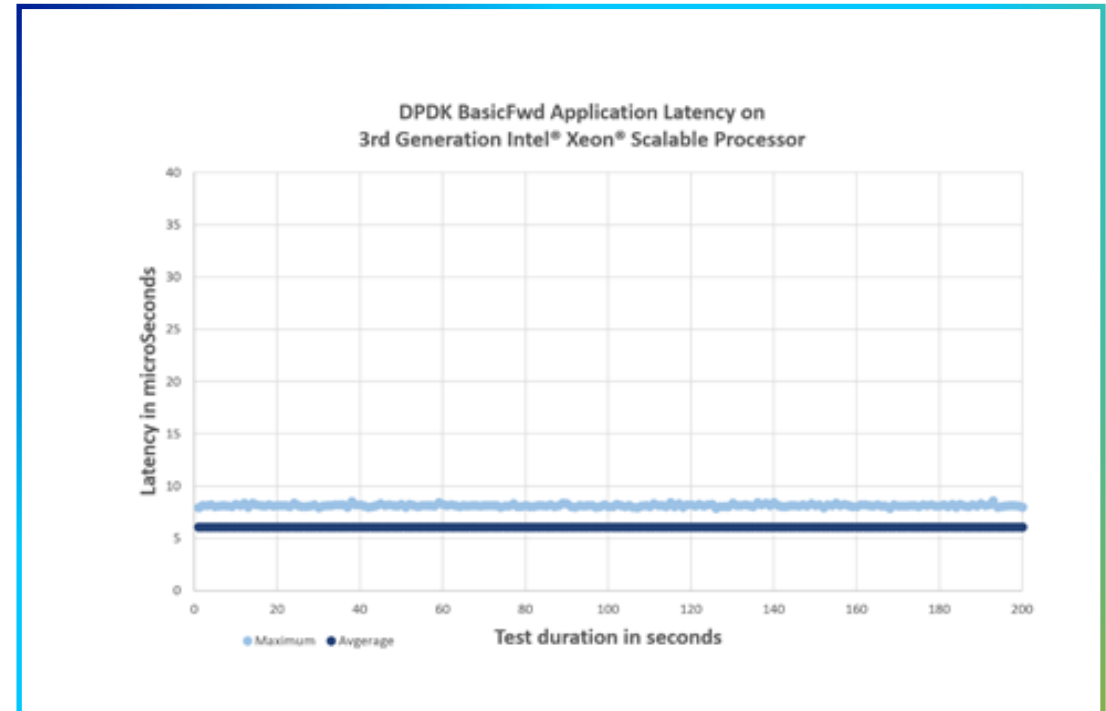
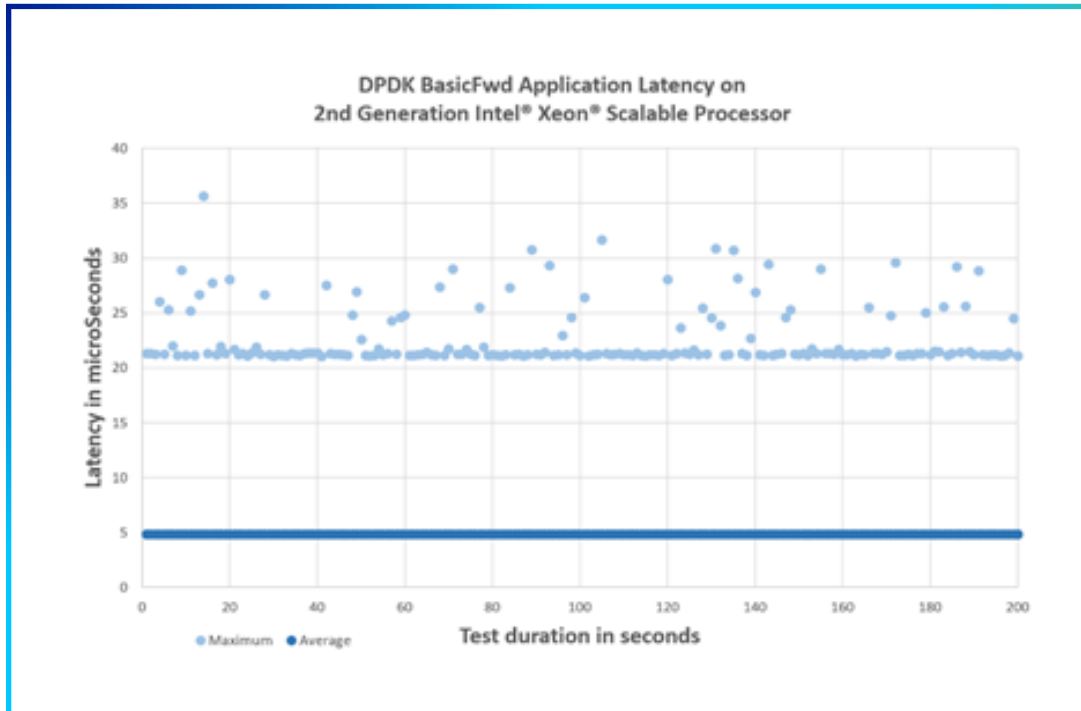


2R1NTW (non-temporal Write) traffic



SoC Architecture Improvements

Latency and Responsiveness



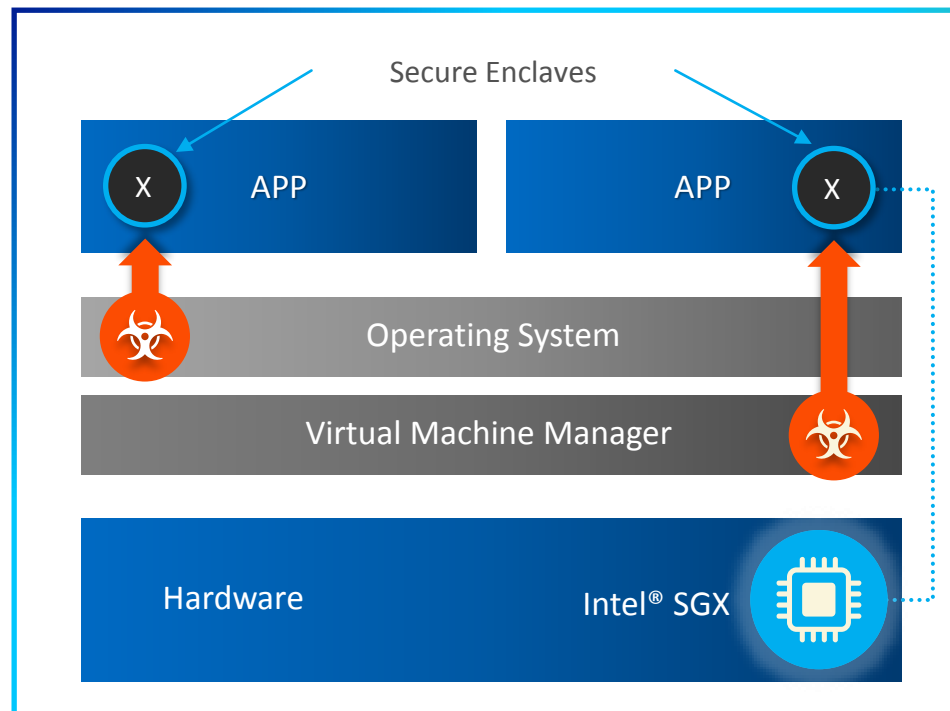
- Seamless power management architecture
- Architecture changes to Improve performance consistency
- Minimize frequency impact on AVX 512 operations

Results have been estimated on pre-production parts as of July 2020. Results may vary.

Trusted Execution

Intel Software Guard Extensions (SGX)

Helps provide enhanced security protections for application data independent of operating system or hardware configuration



- Helps protect against SW attacks even if OS/drivers/BIOS/VMM/SMM are compromised
- Helps increase protections for secrets (data/keys/etc) even when attacker has full control of platform
- Helps prevent attacks such as memory bus snooping, memory tampering, and “cold boot” attacks against memory contents in RAM
- Provides an option for hardware-based attestation capabilities to measure and verify valid code and data signatures

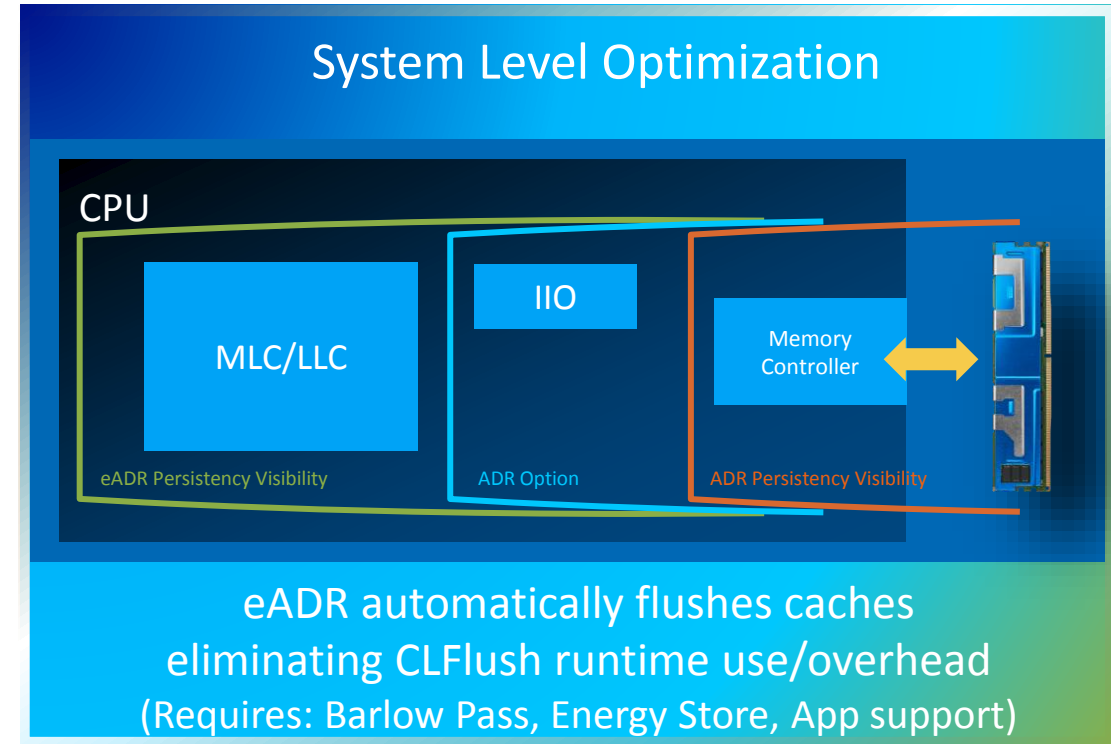
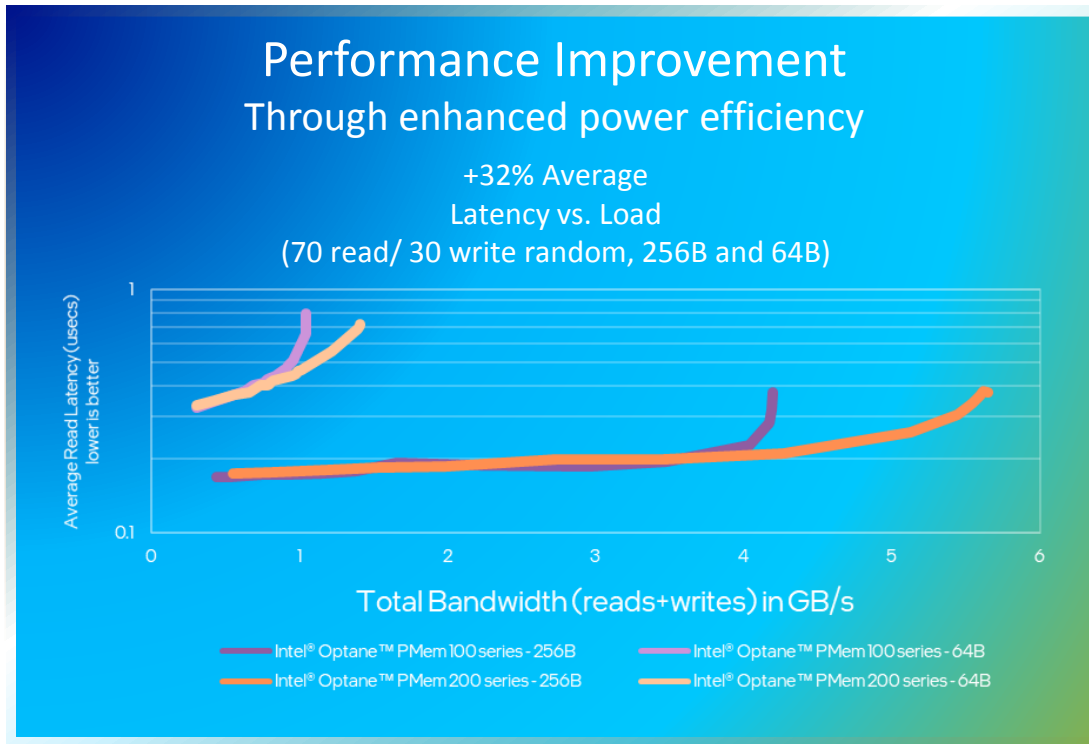
Minimally-sized Trusted Compute Base (TCB)

Other technologies allow some privileged SW in their trust boundary

Helps enhance protections for hard-to-protect spaces

Helps increase transparency and accountability

Intel® Optane™ Persistent Memory 200 Series (Barlow Pass)



Intel® Optane™ Persistent Memory 200 Series

- Consistent developer target: capacity, latency...
- Increased bandwidth, better power efficiency
- Cross system innovation for increased application performance



Intel® Optane™ SSD P5800X

Storage Acceleration with the world's fastest data center SSD

P5800X gains vs Intel® SSD D7-P5600 NAND (both Gen4 PCIe)

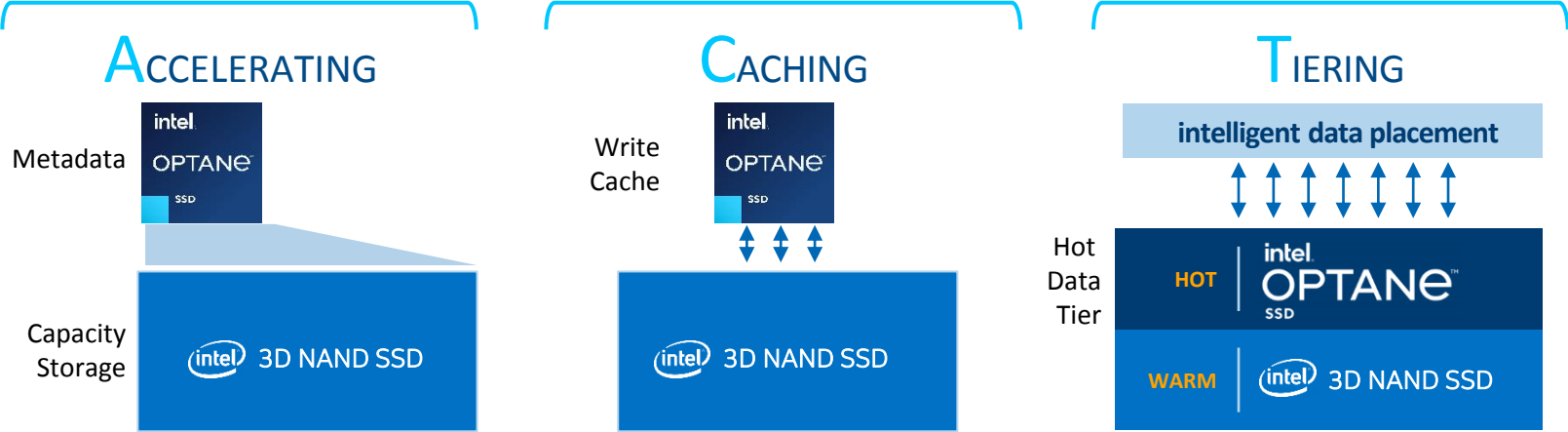
up to **13x** Average Lower Latency at QD=1

up to **66x** Better QoS (4K Read, 70/30 Mixed Random, QD=1, 5 9's)

up to **26x** Greater IOPS/GB 70/30 Random Read

More than **33x** Higher endurance (Drive Writes Per Day)

Accelerate Slower Capacity Storage



Gen3 PCIe Intel® Optane™ SSD DC P4800X improvements vs PCIe Gen 3 NAND

ceph
70% Lower 4 9's latency vs all-TLC NAND at ~ same cost
Also: MS-SQL, MySQL

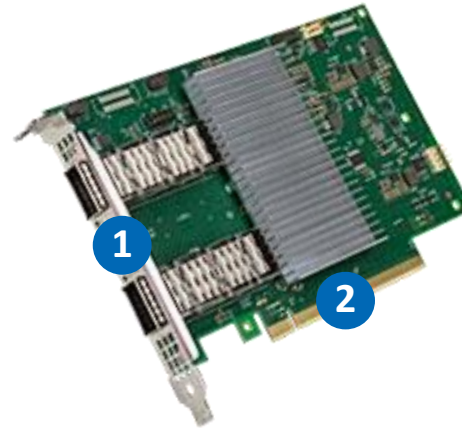
vmware
60% Greater per-node VM density
Also: azure Stack HCI, Cisco Hyperflex, Nutanix

DELL EMC
50% Better response times
Also: Pure Storage FlashArray//X, Vast Data

Performance varies by use, configuration and other factors. Configurations see appendix [46]

Intel® Ethernet Network Adapter

E810-2CQDA2



Hardware Features

Up to 200Gbps of bandwidth in a single PCIe 4.0 x16 slot

Two QSFP28 ports enable active/active configuration with up to 100Gbps on each port

Uses PCIe slot bifurcation to enable the functionality of two 100Gbps adapters in one slot

Key Use Cases

- Bandwidth-intensive Comms workloads, including 5G UPF, vRAN, and CDN
- Cloud workloads, including edge services, database applications, caching servers
- High-bandwidth Storage targets
- HPC/AI fabrics

Intel® Ethernet 800 Series Features

- Flexible port configurations: 2x100Gb, 2x50Gb
- Application Device Queues: improved application response time predictability
- Dynamic Device Personalization: enhanced packet classification capabilities improve throughput
- RDMA iWARP and RoCEv2: high-speed, low latency connectivity

Up to 200Gbps in a single PCIe slot for bandwidth-intensive workloads

Intel® Agilex™ FPGA

Architecture Innovations

Features and On-Going Investments¹

Agilex Family Variants	<ul style="list-style-type: none"> F-Series: Flexible FPGA fabric with up to 58Gbps PAM4 transceivers, PCIe Gen4x16, DDR4 3200 Mbps, and quad-core Arm* Cortex A53 I-Series: Adds enhanced connectivity, up to 116Gbps PAM4 transceivers, PCIe Gen5x16, and CXL M-Series: Adds enhanced memory interface capabilities including DDR5, HBM2e, and Intel® Optane™ Persistent Memory support
------------------------	--

Agilex FPGA Fabric Innovations
Delivering Performance/Watt Gains

- Intel 10nm SuperFin Technology, metal layer stackup and transistor types optimized for FPGAs
- Second-generation Hyperflex™ fabric architecture, re-timing and pipelining registers in the interconnect routing redesigned to reduce register delay and area
- Re-architected interconnect routing to reduce loading and delay, long high-fanout lines replaced with multiple shorter segments, direct logic block outputs added, improved point-to-point routing using shorter wire types
- Floorplan discontinuities removed, resulting in consistent, predictable, and improved timing performance

Quartus Software	<ul style="list-style-type: none"> Retiming-aware synthesis, place and route, and global retiming algorithms optimized to extract maximum performance from fabric innovations and floorplan enhancements Concurrent setup and hold timing optimizations through fine grained register re-timing and clock skew scheduling with signoff-quality timing analysis
------------------	--

Agilex Target Applications

- Embedded, Edge and IOT: Video, Vision, Medical Imaging, Test, Aerospace, Radar, Industry 4.0
- Network: 5G fronthaul, 5G baseband, 5G radio head, NFVi, vRAN, O-RAN, OTN
- Cloud and Enterprise: SmartNIC, Computational Storage, Application Acceleration

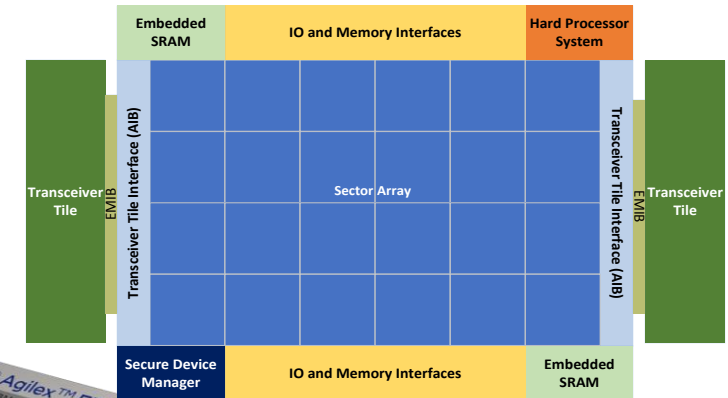


Fig1: Agilex FPGA floorplan (not drawn to scale)

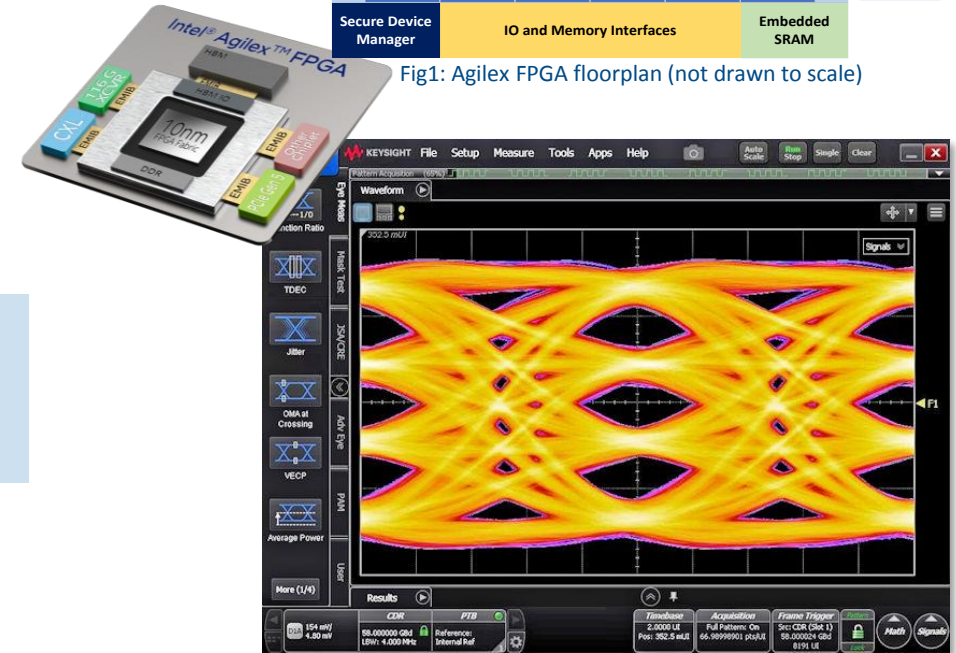


Fig2: Transceiver transmit eye diagram, 116Gbps PAM4

¹ Feature availability occurs over time with each release of Quartus software, preliminary and subject to change

New 3rd Gen Intel® Xeon® Scalable Processor

Optimized for the cloud



Scalable architecture and power and performance optimizations achieve both high throughput and per core performance

New Sunny Cove architecture improves per core performance

Increased DDR4 memory channels feed memory-bound workloads

PCIe Gen4 lanes provide consistent IO latency and performance to scale services on distributed infrastructures

Intel 10nm process is ready to support a fast cloud ramp

Results have been estimated based on pre-production tests as of July 2020. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

3rd Gen Intel® Xeon® Scalable Platform Technical & Performance Overview

Ram Ramakrishnan

Data Center Performance Director
Data Platforms Group
Intel Corporation

3rd Gen Intel® Xeon® Scalable Platform

Feature	2nd Gen Intel® Xeon® Scalable Processor (Cascade Lake)	3rd Gen Intel® Xeon® Scalable Processors (Ice Lake)	Notes
Cores per Socket	4-28	8-40	New Sunny Cove architecture
L1/L2/L3 cache per core	32KB/1MB/1.375MB	48KB/1.25MB/1.5MB	Larger caches to enable fast access to data
Memory Channels and DIMM Speed	6 Up to 2933	8 Up to 3200	Huge boost in memory bandwidth & support for Intel® Optane™ PMem 200
Processor Interconnect: UPI links, speed	2 or 3, 10.4 GT/s	2 or 3, 11.2 GT/s	Improved bandwidth between processors
PCIe lanes per socket	PCIe 3.0, 48 Lanes (x16, x8, x4)	PCIe 4.0, 64 lanes (x16, x8,x4)	2x bandwidth and more PCIe lanes to support new Gen 4 SSD, Ethernet and other adjacencies
Workload Acceleration Instructions	AVX-512 VNNI DDIO	AVX-512, VNNI, DDIO vAES, vPCLMULQDQ, VPMADD52, VBMI, PFR, Crypto, SHA extensions, TME, SGX	Enable new capabilities and speedup performance
Platform Adjacencies		Intel® Optane™ PMem 200 series, Intel® Optane™ P5800X SSD, Intel DC P5510 SSD, Intel E810-C ethernet adapter	

Designed to Move Faster, Store More, Process Everything

New 3rd Gen Intel® Xeon® Scalable Platform Delivers Amazing Performance vs. Previous Xeon Generations

3rd Gen Intel® Xeon® Platinum 8380 (Ice Lake)	Integer SPECrate2017 _int_base (est)	Floating Point SPECrate2017_ fp_base (est)	Memory Bandwidth Stream Triad	LINPACK	Geomean
Vs 2nd Gen Intel® Xeon® Platinum 8280 (Cascade Lake)	up to 1.5x	up to 1.52x	up to 1.47x	up to 1.38x	up to 1.46x
Vs Intel® Xeon® Platinum 8180 (Skylake)	up to 1.6x	up to 1.62x	up to 1.52x	up to 1.44x	up to 1.54x
Vs Intel® Xeon® E5-2699 v4 (Broadwell)	up to 2.34x	up to 2.6x	up to 2.55x	up to 3.18x	up to 2.65x
Vs Intel® Xeon® E5-2699 v3 (Haswell)	up to 2.85x	up to 3.08x	up to 2.8x	up to 3.97x	up to 3.1x

Performance varies by use, configuration and other factors. Configurations see appendix [1,2,3,4]

3rd Gen Intel® Xeon® Scalable Processors

Built-in acceleration

DDIO
enabled vs disabled

Up to
1.3x
higher

DPDK L3 Packet Forwarding
6338N

Intel AVX-512
vs AVX2

Up to
1.62x
higher

LINPACK
8380

Intel Crypto Acceleration
enabled vs disabled

Up to
4.2x
higher

NGINX ECDHE-X25519-RSA2K
6338N v 6252N

Intel DL Boost (VNNI)
int8 vs fp32

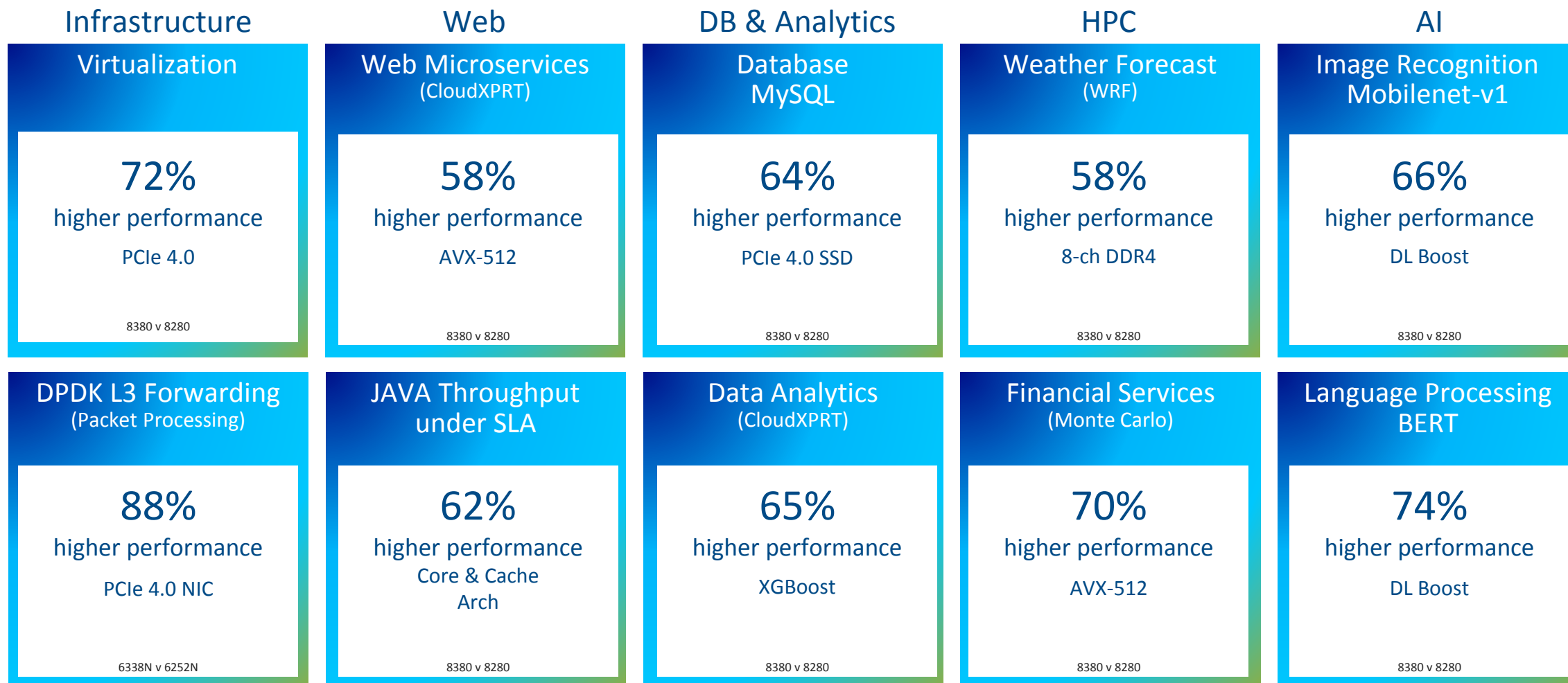
Up to
4.3x
higher

ResNet50-v1.5
8380

Performance varies by use, configuration and other factors. Configurations see appendix [14,15,16,51]

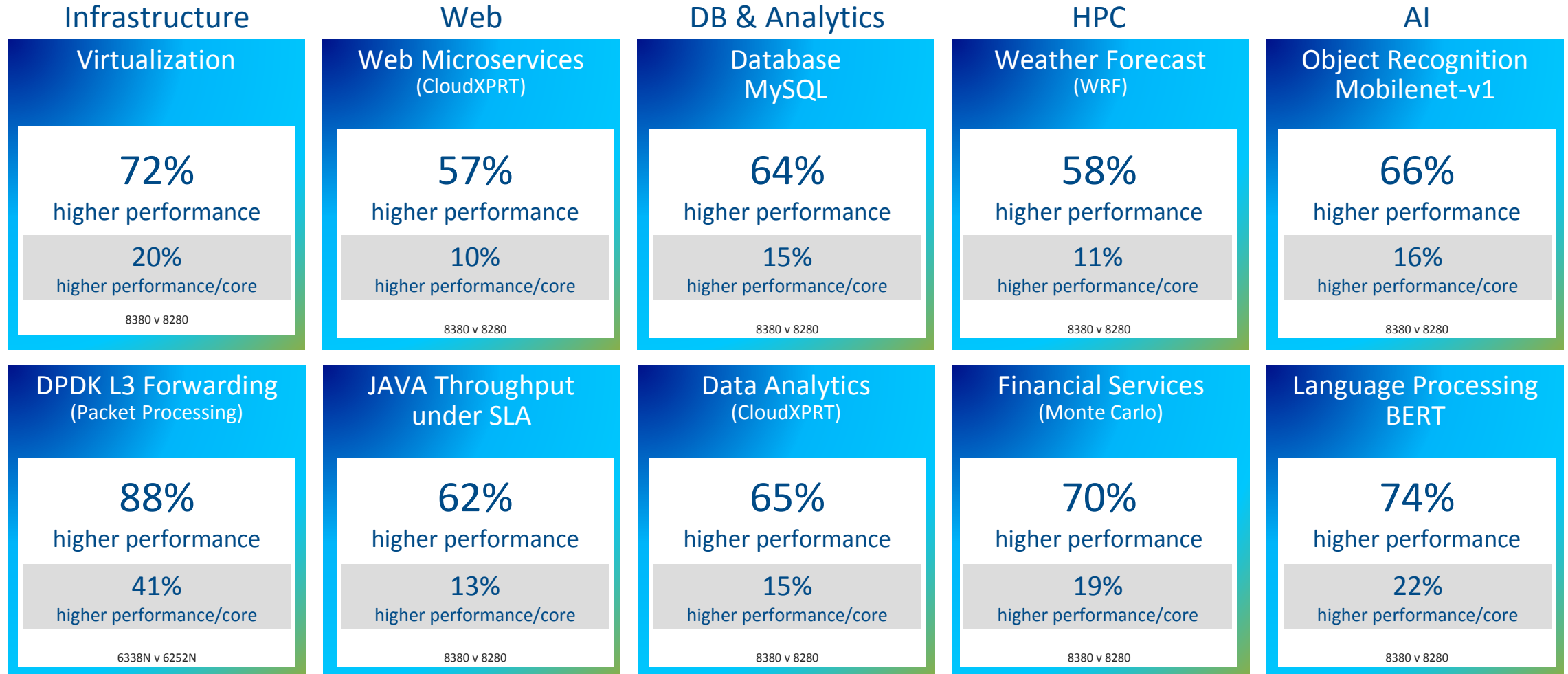
Performance made flexible.

3rd Gen Intel® Xeon® Scalable Platform offer Excellent Real-World Performance Gains on Popular & QoS Sensitive Datacenter Workloads



Performance varies by use, configuration and other factors. Configurations see appendix [5,6,8,9,10,12,13,17,19]
QoS: Quality of Service

Improved Performance per Core Across Various Workloads Delivers Significant Customer Value in the Cloud & On-Premise

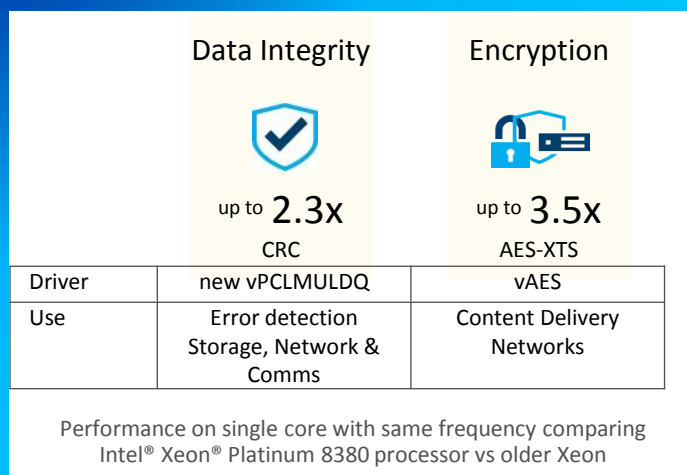


Performance varies by use, configuration and other factors. Configurations see appendix [5,6,8,9,10,12,13,17,19]

3rd Gen Intel® Xeon® Scalable Platform

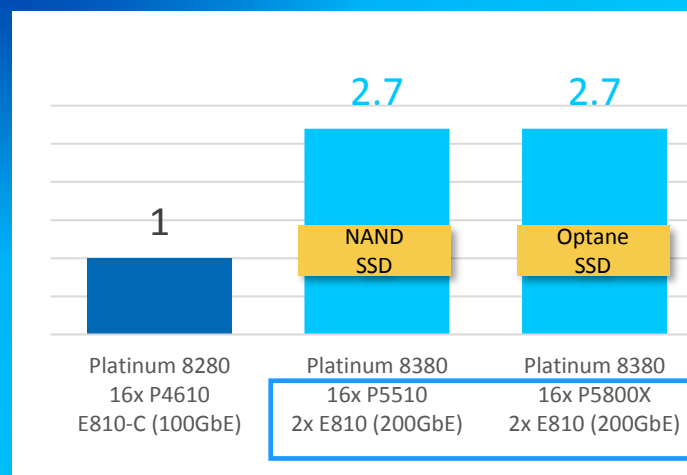
New capabilities deliver big improvements on storage performance acceleration

Platinum 8380 vs prior gen (Platinum 8280)



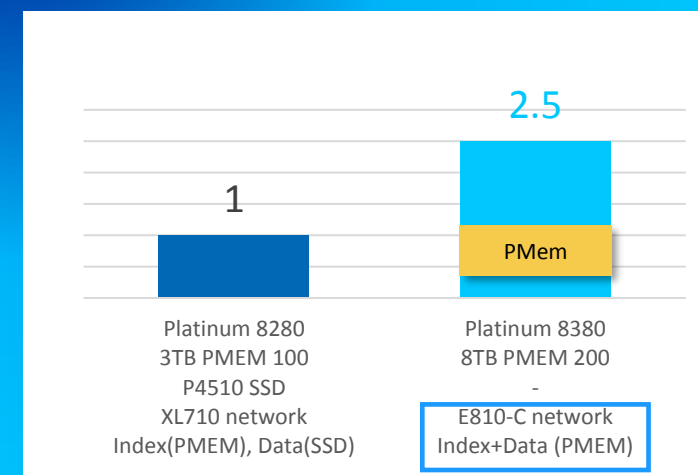
New instructions and architecture improvements on 3rd Gen Intel® Xeon® Scalable processors enable huge improvements in Intel® Intelligent Storage Acceleration Library Performance

NVMe-over-TCP 4kB Random 70R/30W IOPS



NVMe-over-TCP allows customers to provision scalable storage without having to change their ethernet network and providing latencies similar to direct-attached storage

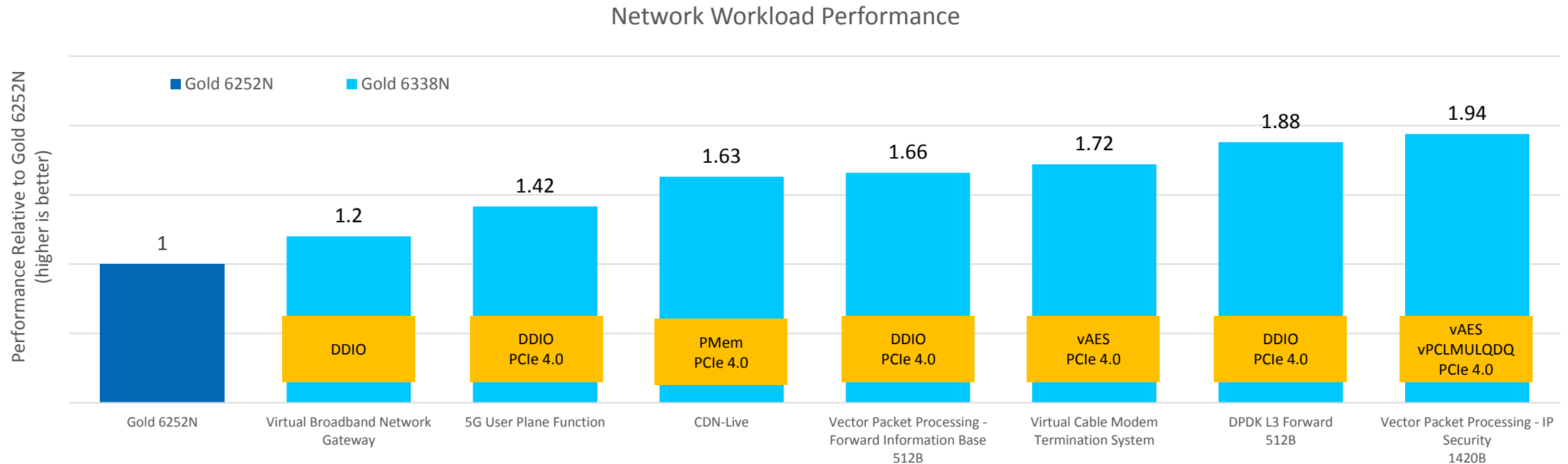
Aerospike Database Transactions Per Second



By moving data and index to Intel® Optane™ PMem 200, customers can see up to 2.5x higher transactions on the new 3rd Gen Intel® Xeon® Scalable platform

Network Workload Performance for Comms Infrastructure

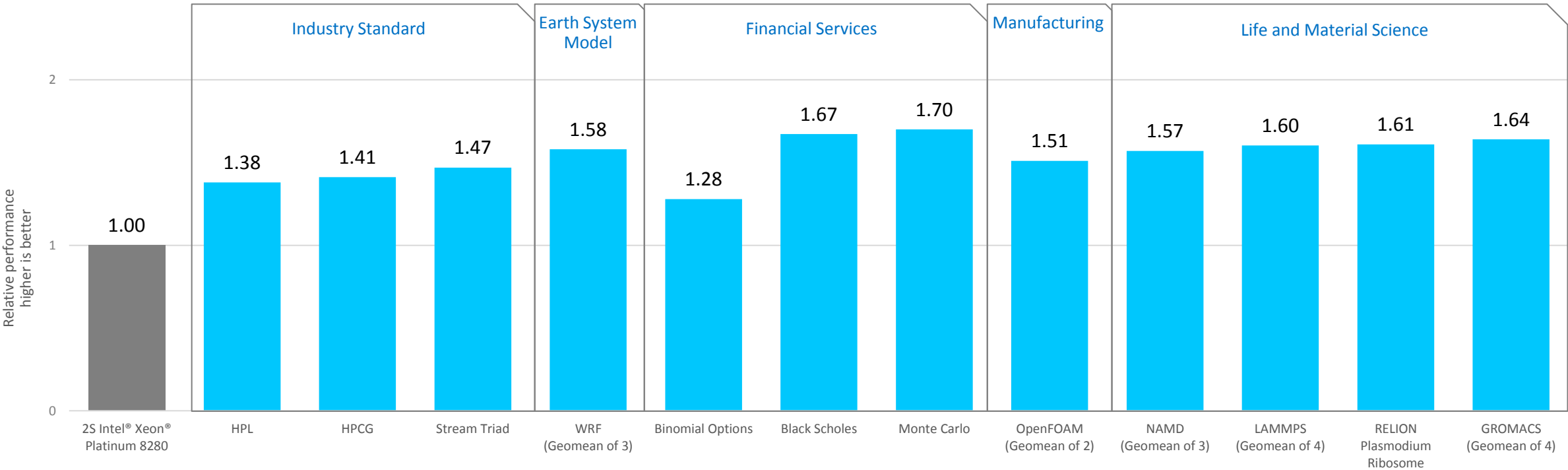
Featuring new Intel Xeon Gold 6338N processor



- New Ice Lake core with vAES, vPCLMULQDQ instructions, PCIe 4.0, DDIO deliver exciting improvements in 3rd Gen Intel® Xeon® Scalable platform for the networking market
- 1.62x average performance gain on a range of broadly-deployed network workloads vs prior generation

Performance varies by use, configuration and other factors. Configurations see appendix [17]

3rd Gen Intel® Xeon® Scalable Processor with Intel AVX-512 and 8-channels Delivers Big Boost for HPC Customers



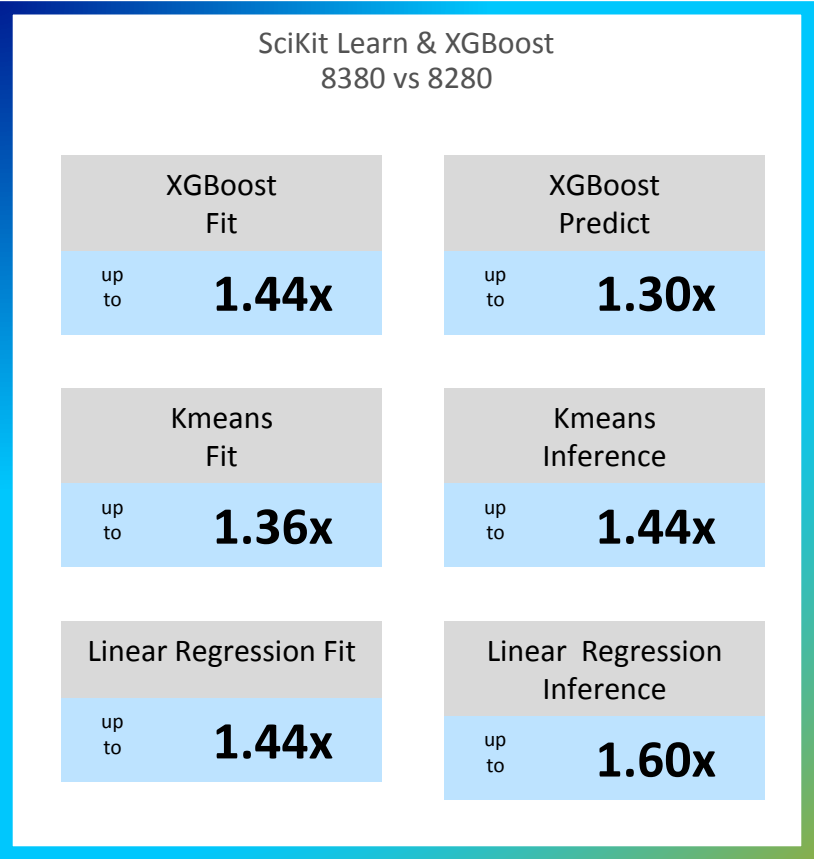
1.53x higher average HPC Performance with Intel® Xeon® Platinum 8380 vs. prior generation featuring Intel® AVX-512 and 8 channels of DDR4-3200 across a broad set of HPC codes

Performance varies by use, configuration and other factors. Configurations see appendix [19]
OpenFOAM Disclaimer: This offering is not approved or endorsed by OpenCFD Limited, producer and distributor of the OpenFOAM software via www.openfoam.com, and owner of the OPENFOAM® and OpenCFD® trademark

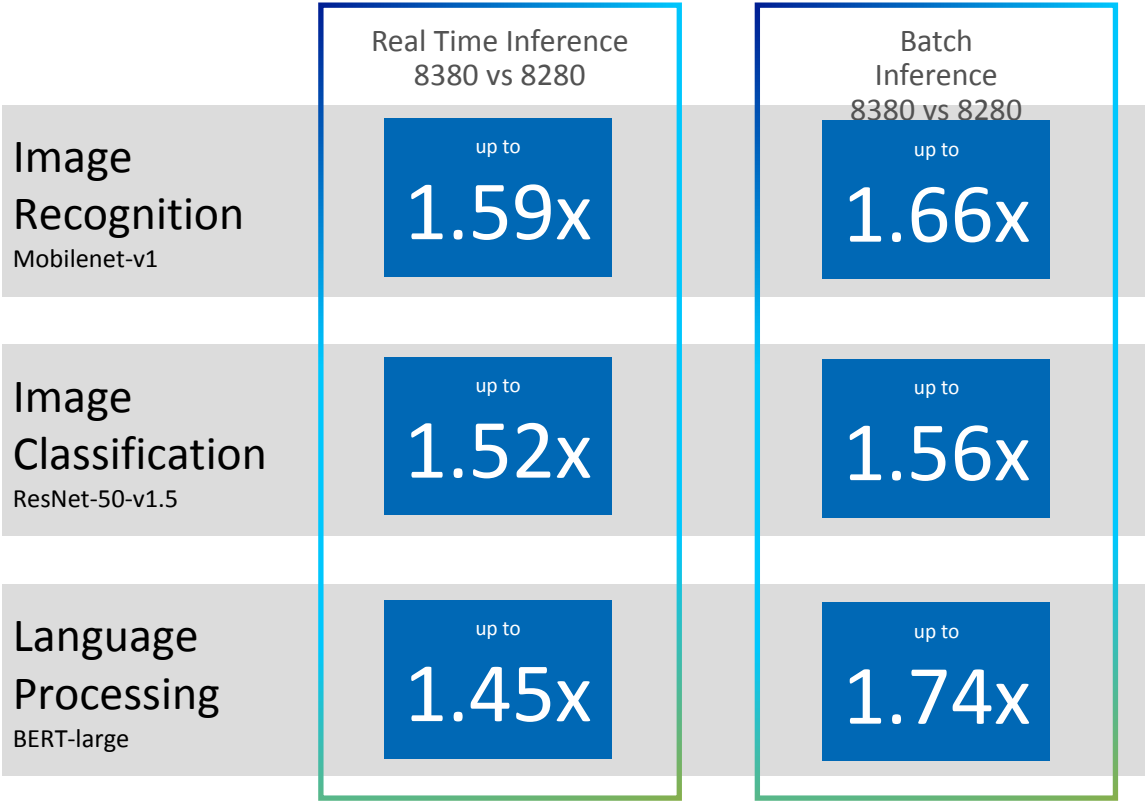
AI Performance Gains

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

Machine Learning



Deep Learning



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

Performance Summary

- 3rd Gen Intel® Xeon® Scalable Platform delivers amazing performance and performance/core across broad range of workloads
- Built-in acceleration and new instructions offer big performance boost in AI, HPC, Networking and Cloud
- New platform capabilities such as Intel® Pmem 200 series, Intel® SSDs & Intel® Ethernet supporting PCIe Gen4 and software optimizations and tools like OneAPI, allow for enhanced performance and scaling

3rd Gen Intel® Xeon® Scalable Platform Technical & Performance Overview

Dave Hill

Senior Data Center Performance Director
Data Platforms Group
Intel Corporation

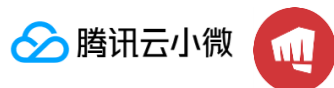
Why Customers Choose Intel

Portfolio Performance



- Up to 2.5x higher transactions on Aerospike database
- Built-in and discrete workload accelerators
- Performance through innovative memory, storage and connectivity products

Predictable Results



- Customers choose Xeon for critical workloads that require fast response times with low and consistent latencies to meet their SLAs
- Tencent: Consistent speech synthesis results across 100s of instances

Optimizations & Domain Expertise



- Extend your ROI by utilizing ongoing Intel SW optimizations (Ex: AI, HPC)
- >700 world leading ISVs and open source projects

Virtualization Compute Pool Consistency

- Reduced deployment and management complexity with ability to live VM migrate 5 generations of Xeon® processors to/from 3rd Generation Xeon
- Consistent memory latency for a wide range applications

Mature Global Supply Chain

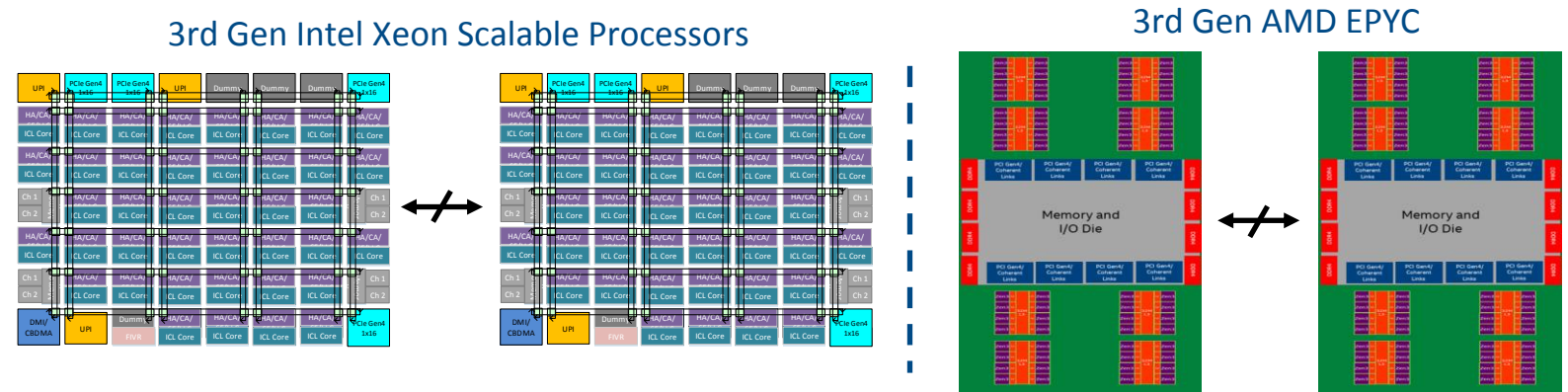


- IDM advantage ensures we control our supply chain
- 25% increase in capacity in 2020

Performance varies by use, configuration and other factors. Configurations see appendix [23]

3rd Gen Intel Xeon Scalable Processors

Processor architecture, cache latencies



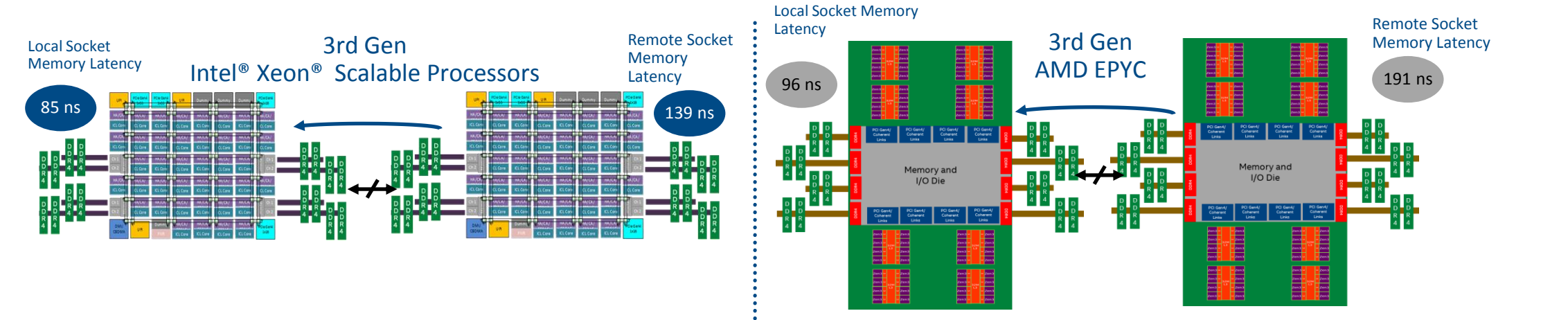
Latency	Intel Xeon Platinum 8380 Processor (Ice Lake)	AMD EPYC 7763 Processor (Milan)	Intel Xeon Platinum 8280 Processor (Cascade Lake)
L1 hit cache, cycles	5	4	4
L2 hit cache, cycles	14	12	14
L3 hit cache (same socket), ns	21.7	13.4 (< 32MB) local die 112 (> 32MB) remote die	20.2
L3 hit cache (remote socket), ns	118	209	180

3rd Gen Intel Xeon Scalable processor have consistent and low latencies to local cache and the second socket

Performance varies by use, configuration and other factors. Configurations see appendix [50]

3rd Gen Intel® Xeon® Scalable Processors

Memory controller, memory latency and capacity



	Intel® Xeon® Platinum 8380 Processor (Ice Lake)	AMD EPYC 7763 Processor (Milan)	Intel® Xeon® Platinum 8280 Processor (Cascade Lake)
Memory Controller	On die – 8 ch	Multi chip module – 8 ch	On Die – 6ch
Max DIMM Capability	2 DPC 3200/2933/2666 (SKU dependent) <small>New: PMem runs at memory channel speed</small>	1 DPC 3200 2 DPC 2933/2666	1 DPC 2933/2 DPC 2666 (SKU dependent)
DRAM read latency local socket, ns	85	96	81
DRAM read latency (remote socket), ns	139	191	138
Max Memory Capacity per Socket	6TB (DDR+PMem). 4TB (DDR)	4TB (DDR)	4.5 TB (DDR+PMem). 3TB (DDR)

3rd Gen Xeon Scalable provides lower latencies to DRAM and supports larger memory capacity

Performance varies by use, configuration and other factors. Configurations see appendix [50]
Results may vary

3rd Gen Intel® Xeon® Scalable Processors

Instructions providing workload acceleration

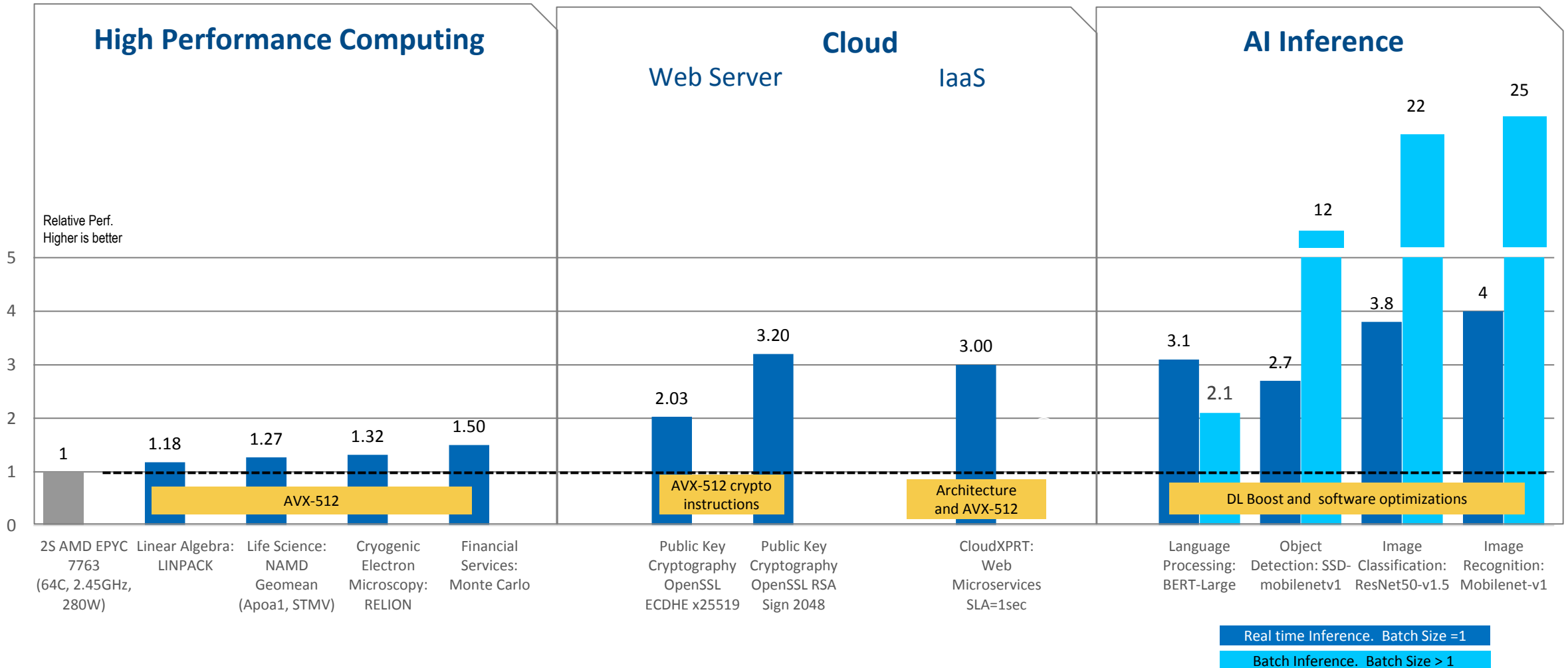
	Application	Usage/Workloads	3rd Gen Intel® Xeon® Scalable Processors (Ice Lake)	2nd Gen Intel® Xeon® Scalable Processors (Cascade Lake)	3rd Gen AMD EPYC (Milan)
VNNI	Accelerated 8-bit integer processing	AI workload performance acceleration	Yes	Yes	No
AVX-512	Vector acceleration for computational demanding tasks	HPC and other compute intensive applications	Yes	Yes	No AVX2
Vector AES (Advanced Encryption Standard) Vector CLMUL	Encryption (CTR, CBC, XTS) and authenticated encryption (AES-GCM)	Data move to/from disk or across AES. Ex. Database encryption, cloud encryption (Hadoop, etc.)	Yes AVX-512 (light)	No	Yes AVX2
VPMADD52	Public key crypto generation (RSA & DH)	SSL Front End Web Server connections (NGINX, HA-Proxy, WordPress)	Yes	No	No
SHA (Secure Hash Algorithm) Extensions	SHA-1 & SHA-256 Hashing Algorithms	Hashing, SSL, TLS, IPSec, Data Dedup, Blockchain, Bitcoin	Yes	No	Yes
VBMI AVX 512 (Vector Byte Manipulation Instructions)	In line compression while feeding small chunks of data to algorithms for immediate operation	In-Memory Database (IMDB) workloads; improves decompression/compression performance	Yes	No	No
VBMI AVX 512 (Vector Byte Manipulation Instructions)	In line compression while feeding small chunks of data to algorithms for immediate operation	In-Memory Database (IMDB) workloads; improves decompression/compression performance	Yes	No	No

3rd Gen Xeon® provides differentiated instructions that enable outstanding performance on leading edge workloads

Source: Intel instructions at <https://software.intel.com/content/www/us/en/develop/articles/intel-sdm.html#combined>. AMD instructions at <https://www.amd.com/system/files/TechDocs/24594.pdf>

HPC, Cloud, and AI Performance Comparisons

2S Intel® Xeon® Platinum 8380 (40C) vs. 2S AMD EPYC 7763 (64C)



Performance varies by use, configuration and other factors. Configurations see appendix [27, 30-38]

Intel Focus on Continuing Software Optimizations to Improve Customer Performance



30x Improved
AI Inference
Performance



10x Lower Latency
Database Read
Performance on X8M



8.2x Improved
Medical Imaging AI Inference
Performance



5.5x Improved
Hadoop Data Compression
Performance



2.5x Improved
Cloud-native Container OpenSSL
Performance



1.8x Improved
Molecular Dynamics
Performance

Source:

30x <https://blog.roblox.com/2020/05/scaled-bert-serve-1-billion-daily-requests-cpus/>

10x Oracle. Compared with the X8 and its InfiniBand fabric, the X8M will offer 100Gb RDMA over Converged Ethernet (RoCE) as the internal fabric to deliver latency of under 19 microseconds (10X improvement over the X8). With 1.5TB of persistent memory (PMem) per storage server and up to 21.5TB of PMem per standard full rack, organizations can achieve up to 16 million OLTP 8K read IOPS, 2.5X the X8.

8.2x <https://www.intel.com/content/www/us/en/customer-spotlight/stories/hyhy-customer-story.html>

5.5x: Customer software example of Kuaishou big data storage platform compression optimizations of Kuaishou Short Video solution. Optimizations included use of Intel 01.org open source Intelligent Storage Acceleration Library (ISA-L) compression algorithm DEFLATE. Since Hadoop is implemented in Java, JNI used to call ISA-L (C interface) interface. Because the data structure of Java and C are inconsistent, used JNIEnv's GetPrimitiveArrayCritical method, reducing the number of memory copies and improved JNI packaging. Test used Kuaishou Intel SKX6148 machine, the Hadoop compression performance using zlib1.2.11 and zlib1.2.22 resulting in a 5.51x improvement. Measured by Intel Oct 2020.





2.5x Red Hat OpenSSL: <https://www.intel.com/content/www/us/en/big-data/partners/redhat/red-hat-openshift-hybrid-cloud-reference-architecture.html>, page 25, table 17

1.8x <https://www.hpcwire.com/2020/08/12/intel-speeds-namd-by-1-8x-saves-Xeon-processor-users-millions-of-compute-hours/>

3rd Gen Intel® Xeon® Scalable Processors

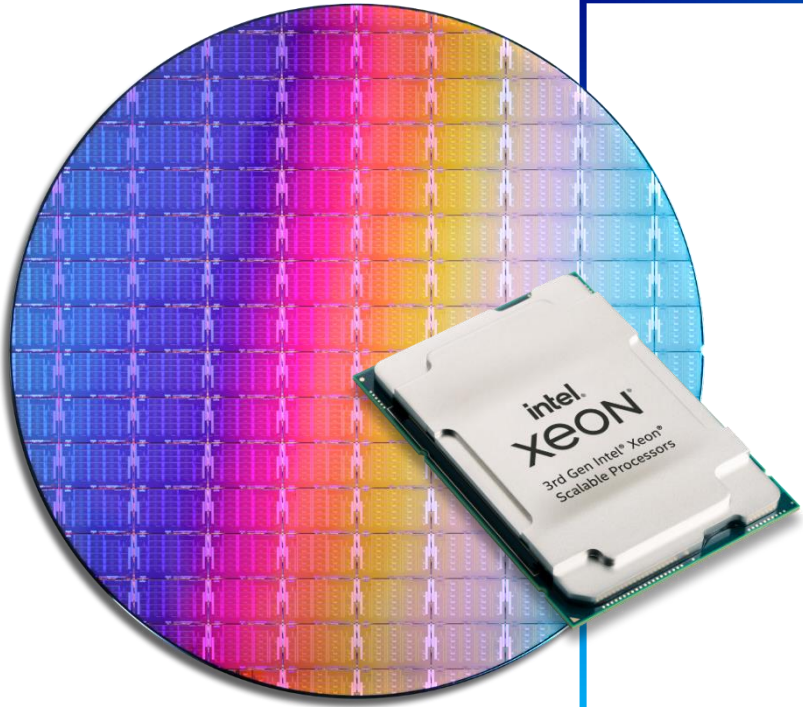
Competitive differentiators

Improvements compared to
2nd Gen Intel® Xeon® Scalable Processors

	Feature	Customer Value
Processor 	Intel Deep Learning Boost, built-in AI Acceleration	Gain new insights for fast decisions through increased AI performance and the flexibility to run other cloud and data center workloads
	Intel AVX-512 Vector Instructions	Accelerate performance for a wide range of compute-intensive workloads (AI, HPC, security, imaging, networking)
	Network Acceleration with Intel Direct Data I/O (DDIO)	Improve application response time and increase transaction rates by implementing low latency and fast data transfer from IO devices to Xeon processor cache, bypassing memory
	As Low as 105W TDP SKUs	Energy efficient alternatives for power constrained environments such as edge and 5G networks
	Workload and Customer Optimized SKUs	A tailored portfolio of Xeon processor SKUs to meet demanding workload and customer requirements – targeted to cloud, network, IoT/ NEBS, and liquid cooled.
	Enhanced Intel Speed Select Technology (Intel SST)	Customize Xeon to your workload through granular control over processor frequency, core count and power
	Advanced Security Technologies: Intel Software Guard Extensions (Intel SGX), Intel Crypto Acceleration, Intel Total Memory Encryption (Intel TME)	Encrypt everything. Helps protect customer data at-rest, in-flight, and in-use
	Vector Bit Manipulation Instructions (VBMI)	Accelerate applications such as in-memory databases by improving compression/decompression performance
Portfolio 	Intel® Optane™ Persistent Memory 200 series, supports up to 6TB memory per socket	World leading memory per socket enables support for large memory workloads, future business growth, data persistence, and lower memory costs/TCO in some configurations.
	Unmatched Portfolio: Connectivity, Storage, Processors, FPGA	Enhance workload optimized solutions designed to move, store and process everything
	Additional Crypto/Compression Acceleration with Intel QuickAssist Technology	Improves performance of security and networking infrastructure workloads by offloading crypto acceleration and compression capabilities to hardware thereby reserving processor cycles for other tasks.
	Storage specific features: Intel VMD, eDPC, Intel VROC, Intel RSTe, Intel CAS, Intel ISA-L	Allows for managing, reducing cost, and improving performance for storage workloads.
Ecosystem  	Broad Software and Tools Ecosystem	Great out-of-box experience on applications from Intel open source and ISV software enabling. Offers the flexibility and choice from solutions across a wide range system hardware vendors and deployment options including public cloud and on-premise.
	Broad Hardware Ecosystem	
	System Solutions with Intel Select Solutions and Intel Market Ready Solutions	Enables solutions that are workload optimized, easy to deploy and verified by Intel.
	VMware vMotion Live Migration Capable within Intel® Xeon® Infrastructure	Ease of management and reduced deployment complexity with seamless VM portability within last five generations of Intel® Xeon® Processor based infrastructure.
	SAP HANA Certification: Intel® Xeon® Processors are the only x86 CPU supported	Solutions offering real time data insights on a mission critical in-memory database environment with deployment choice across public cloud and on-premise.

Performance made flexible.

Summary



- Intel's highest performing data center processor with built-in security and AI and crypto acceleration
- Unmatched portfolio of hardware and software solutions to move, store and process data
- Broadest ecosystem and decades of experience to ease customer deployments

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.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

Q&A



Performance
made flexible.

Appendix

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

8. 1.72x higher virtualization performance & 1.2x performance/core: Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 processor on Coyote Pass with 2048 GB (32 slots/ 64GB/ 3200) total DDR4 memory, ucode 0x261, HT on, Turbo on, RedHat 8.3, 4.18.0-240.el8.x86_64, 1x S4610 SSD 960G, 4x P5510 3.84TB NVME, 2x Intel E810, Virtualization workload, Qemu-kvm 4.2.0-34 (inbox), WebSphere 8.5.5, DB2 v9.7, Nginx 1.14.1, test by Intel on 3/14/2021. Platinum 8280: 1-node, 2x Intel® Xeon® Platinum 8280 processor on Wolf Pass with 1536 GB (24 slots/ 64GB/ 2933[2666]) total DDR4 memory, ucode 0x5003005, HT on, Turbo on, RedHat 8.1 (Note: selected higher of RedHat 8.1 and 8.3 scores for baseline), 4.18.0-147.el8.x86_64, 1x S4510 SSD 240G, 4x P4610 3.2TB NVME, 2x Intel XL710, Virtualization workload, Qemu-kvm 4.2.0-34 (inbox), WebSphere 8.5.5, DB2 v9.7, Nginx 1.14.1, test by Intel on 12/22/2020.
9. 1.62x higher throughput under SLA for Server Side Java & 1.13x 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 20.04.1 LTS, 5.4.0-64-generic, 1x SSDSC2BA40, Java workload, JDK 1.15.0.1, 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, Turbo on, Ubuntu 20.04.1 LTS, 5.4.0-64-generic, 1x INTEL_SSDSC2KG01, Java workload, JDK 1.15.0.1, test by Intel on 2/18/2021.
10. 1.64x HammerDB MySQL & 1.15x 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, Redhat 8.3, 4.18.0-240.el8.x86_64 x86_64, 1x Intel® SSD 960GB OS Drive, 1x Intel P5800 1.6T, onboard 1G/s, HammerDB 4.0, MySQL 8.0.22, test by Intel on 3/11/2021. 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, Redhat 8.3, 4.18.0-240.el8.x86_64 x86_64, 1x Intel 240GB SSD OS Drive, 1x Intel 6.4T P4610, onboard 1G/s, HammerDB 4.0, MySQL 8.0.22, test by Intel on 2/5/2021.
11. **1.48x higher responses on WordPress with HTTPS: 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, 5.4.0-65-generic, 1x Intel 895GB SSDSC2KG96, 1x XL710-Q2, WordPress 4.2 with HTTPS, gcc 9.3.0, GLIBC 2.31-0ubuntu9.1, mysqld Ver 10.3.25-MariaDB-0ubuntu0.20.04.1, PHP 7.4.9-dev (fpm-fcgi), Zend Engine v3.4.0, 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, Turbo on, Ubuntu 20.04, 5.4.0-65-generic, 1x Intel 1.8T SSDSC2KG01, 1x Intel X722, test by Intel on 2/5/2021.
12. **1.65x higher responses with CloudXPRT Data Analytics & 1.15x 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, 5.4.0-65-generic, 1x S4610 SSD 960G, CloudXPRT v1.0, Data Analytics (Analytics per minute @ p.95 <= 90s), 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, 5.4.0-65-generic, 1x S3520 SSD 480G, CloudXPRT v1.0, test by Intel on 2/4/2021. 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.
13. **1.58x higher responses with CloudXPRT Web Microservices:** 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, 5.4.0-65-generic, 1x S4610 SSD 960G, CloudXPRT v1.0, Web Microservices (Requests per minute @ p.95 latency <= 3s), 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, 5.4.0-54-generic, 1x S3520 SSD 480G, CloudXPRT v1.0, test by Intel on 2/4/2021. 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.
14. **1.3x DDIO on/off with DPDK L3 Packet Forwarding:** 1-node, 2(1 socket used)x Intel® Xeon® Gold 6338N on Intel® Whitley with 128 GB (8 slots/ 16GB/ 2666) total DDR4 memory, ucode 0x261, HT on, Turbo off, Ubuntu 20.04 LTS (Focal Fossa), 5.4.0-40-generic, 1x INTEL® 240G SSD, 1x E810-2CQDA2 (Chapman Beach), v20.08.0, Gcc 9.3.0, DPDKL3FWD (4c8t), PCIe Writes = Non Allocating for DDIO off case, test by Intel on 3/24/2021
15. **1.62x AVX-512 vs AVX2 on Linpack:** 1-node, 2x Intel® Xeon® Platinum 8380 on Coyote Pass with 512 GB (12 slots/ 32GB/ 3200) total DDR4 memory, ucode 0x270, HT off, Turbo on, Ubuntu 20.04.2 LTS, 5.4.0-67-generic, 1x INTEL_SSDSC2BB80, 1x X710, Intel distribution of Linpack with AVX-512, AVX2, test by Intel on 3/24/2021
16. 4.3x Intel DL Boost FP32 to INT8: 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, ResNet-50 v1.5, gcc-9.3.0, oneDNN 1.6.4, BS=128, FP32/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 3/19/2021.

Appendix

17. **1.62x average network performance gains:** geomean of Virtual Broadband Network Gateway, 5G User Plane Function, Virtual Cable Modem Termination System, Vector Packet Processing - Forward Information Base 512B, DPDK L3 Forward 512B, CDN-Live, Vector Packet Processing - IP Security 512B.
- a. **1.2x Virtual Broadband Network Gateway:** Gold 6338N: 1-node, 2(1 socket used)x Intel® Xeon® Gold 6338N on Intel® Whitley with 256 GB (16 slots/ 16GB/ 2666) total DDR4 memory, ucode 0x261, HT on, Turbo off, Ubuntu 20.04 LTS (Focal Fossa), 5.4.0-40-generic, 1x INTEL® 240G SSD , 3x E810-CQDA2 (Tacoma Rapids), vBNG 20.07, Gcc 9.3.0, test by Intel on 3/11/2021. Gold 6252N: 1-node, 2(1 socket used)x Intel® Xeon® Gold 6252N on SuperMicro® X11DPG-QT with 192 GB (12 slots/ 16GB/ 2933) total DDR4 memory, ucode 0x5002f01, HT on, Turbo off, Ubuntu 20.04 LTS (Focal Fossa), 5.4.0-40-generic, 1x INTEL® 240G SSD , 3x E810-CQDA2 (Tacoma Rapids), vBNG 20.07, Gcc 9.3.0, test by Intel on 2/2/2021.
 - b. **1.42x 5G User Plane Function:** 1-node, 2(1 socket used)x Intel® Xeon® Gold 6338N on Whitley Coyote Pass 2U with 128 GB (8 slots/ 16GB/ 2666) total DDR4 memory, ucode 0x261, HT on, Turbo off, Ubuntu 18.04.5 LTS , 4.15.0-134-generic , 1x Intel 810 (Columbiaville), FlexCore 5G UPF, Jan' 2021 MD5 checksum: c4ad7f8422298ceb69d01e67419ff1c1, GCC 7.5.0, 5G UPF228 Gbps / 294 Gbps, test by Intel on 3/16/2021. 1-node, 2(1 socket used)x Intel® Xeon® Gold 6252N on SuperMicro® X11DPG-QT with 96 GB (6 slots/ 16GB/ 2933) total DDR4 memory, ucode 0x5003003, HT on, Turbo off, Ubuntu 18.04.5 LTS , 4.15.0-132-generic , 1x Intel 810 (Columbiaville), FlexCore 5G UPF, Jan' 2021 MD5 checksum: c4ad7f8422298ceb69d01e67419ff1c1, GCC 7.5.0, 5G UPF161 Gbps / 213 Gbps , test by Intel on 2/12/2021.
 - c. **1.63x CDN Live:** 1 node, 2x Intel® Xeon® Gold 6338N Processor, 32 core HT ON Turbo ON, Total DRAM 256GB (16 slots/16GB/2666MT/s), Total Optane Persistent Memory 200 Series 2048GB (16 slots/128GB/2666MT/s), BIOS SE5C6200.86B.2021.D40.2103100308 (ucode: 0x261), 4x Intel® E810, Ubuntu 20.04, kernel 5.4.0-65-generic, gcc 9.3.0 compiler, openssl 1.1.1h, varnish-plus 6.0.7r2. 2 clients, Test by Intel as of 3/11/2021. Gold 6252N: 2x Intel® Xeon® Gold 6252N Processor, 24 core HT ON Turbo ON, Total DRAM 192GB (12 slots/16GB/2666MT/s), Total Optane Persistent Memory 100 Series 1536GB(12 slots/128GB/2666MT/s), 1x Mellanox MCX516A-CCAT, BIOS: SE5C620.86B.02.01.0013.121520200651 (ucode: 0x5003003), Ubuntu 20.04, kernel 5.4.0-65-generic, wrk master 4/17/2019. Test by Intel as of 2/15/2021. Throughput measured with 100% Transport Layer Security (TLS) traffic with 93.3% cache hit ratio and keep alive on, 512 total connections.
 - d. **1.66 Vector Packet Processing - Forward Information Base 512B:**1-node, 2(1 socket used)x Intel® Xeon® Gold 6338N on Intel® Whitley with 128 GB (8 slots/ 16GB/ 2666) total DDR4 memory, ucode 0x261, HT on, Turbo off, Ubuntu 20.04 LTS (Focal Fossa), 5.4.0-40-generic, 1x INTEL® 240G SSD , 1x E810-2CQDA2 (Chapman Beach), v20.05.1-release, Gcc 9.3.0, VPPFIB(24c24t), test by Intel on 3/17/2021. 1-node, 2(1 socket used)x Intel® Xeon® Gold 6252N on SuperMicro® X11DPG-QT with 96 GB (6 slots/ 16GB/ 2933) total DDR4 memory, ucode 0x5002f01, HT off, Turbo off, Ubuntu 20.04 LTS (Focal Fossa), 5.4.0-40-generic, 1x INTEL® 240G SSD , 1x E810-CQDA2 (Tacoma Rapids), v20.05.1-release, Gcc 9.3.0, VPPFIB (18c18t), test by Intel on 2/2/2021.
 - e. **1.72x Virtual Cable Modem Termination System:** Gold 6338N: 1-node, 2(1 socket used)x Intel® Xeon® Gold 6338N on Coyote Pass with 256 GB (16 slots/ 16GB/ 2666) total DDR4 memory, ucode 0x261, HT on, Turbo off(no SST-BF)/on(SST-BF), Ubuntu 20.04 LTS (Focal Fossa), 5.4.0-40-generic, 1x INTEL® 240G SSD , 3x E810-CQDA2 (Tacoma Rapids), vCMTS 20.10, Gcc 9.3.0, SST-BF (2.4 Ghz,1.9 Ghz frequencies for the priority cores and the other cores respectively), test by Intel on 3/11/2021. Gold 6252N: 1-node, 2(1 socket used)x Intel® Xeon® Gold 6252N on SuperMicro® X11DPG-QT with 192 GB (12 slots/ 16GB/ 2933) total DDR4 memory, ucode 0x5002f01, HT on, Turbo off, Ubuntu 20.04 LTS (Focal Fossa), 5.4.0-40-generic, 1x INTEL® 240G SSD , 2x E810-CQDA2 (Tacoma Rapids), vCMTS 20.10, Gcc 9.3.0, vCMTS90 (14 instances), test by Intel on 2/2/2021.
 - f. **1.88x DPDK L3 Forward 512B & 1.41x performance/core:**1-node, 2(1 socket used)x Intel® Xeon® Gold 6338N on Intel® Whitley with 128 GB (8 slots/ 16GB/ 2666) total DDR4 memory, ucode 0x261, HT on, Turbo off, Ubuntu 20.04 LTS (Focal Fossa), 5.4.0-40-generic, 1x INTEL® 240G SSD , 1x E810-2CQDA2 (Chapman Beach), v20.08.0, Gcc 9.3.0, DPDKL3FWD (24c24t), test by Intel on 3/17/2021, 2(1 socket used)x Intel® Xeon® Gold 6252N on SuperMicro® X11DPG-QT with 96 GB (6 slots/ 16GB/ 2933) total DDR4 memory, ucode 0x5002f01, HT off, Turbo off, Ubuntu 20.04 LTS (Focal Fossa), 5.4.0-40-generic, 1x INTEL® 240G SSD , 1x E810-CQDA2 (Tacoma Rapids), v20.08.0, Gcc 9.3.0, DPDKL3FWD (12c12t), test by Intel on 2/2/2021.
 - g. **1.94x Vector Packet Processing - IP Security 1420B:** 1-node, 2(1 socket used)x Intel® Xeon® Gold 6338N on Intel® Whitley with 128 GB (8 slots/ 16GB/ 2666) total DDR4 memory, ucode 0x261, HT on, Turbo off, Ubuntu 20.04 LTS (Focal Fossa), 5.4.0-40-generic, 1x INTEL® 240G SSD , 1x E810-2CQDA2 (Chapman Beach), v21.01-release, Gcc 9.3.0, VPPISEC(24c24t) test by Intel on 3/17/2021. 1-node, 2(1 socket used)x Intel® Xeon® Gold 6252N on SuperMicro® X11DPG-QT with 96 GB (6 slots/ 16GB/ 2933) total DDR4 memory, ucode 0x5002f01, HT off, Turbo off, Ubuntu 20.04 LTS (Focal Fossa), 5.4.0-40-generic, 1x INTEL® 240G SSD , 1x E810-CQDA2 (Tacoma Rapids), v21.01-release, Gcc 9.3.0, VPPISEC(18c18t) test by Intel on 2/2/2021.
18. 2x MIMO Throughput: Results have been estimated or simulated. Based on 2x estimated throughput from 32Tx32R (5Gbps) on 2nd Gen Intel® Xeon® Gold 6212U processor to 64Tx64R (10Gbps) on 3rd Gen Intel® Xeon® Gold 6338N processor at similar power ~185W.

Appendix

19. 1.53x higher HPC performance (geomean HPL, HPCG, Stream Triad, WRF, Binomial Options, Black Scholes, Monte Carlo, OpenFOAM, GROMACS, LAMMPS, NAMD, RELION)
- a. 1.53x higher FSI Kernel performance (geomean Binomial Options, Black Scholes, Monte Carlo)
 - b. 1.60x higher Life and Material Science performance (geomean GROMACS, LAMMPS, NAMD, RELION)
 - c. Platform setup: 8380: 1-node, 2x Intel® Xeon® Platinum 8380 (40C/2.3GHz, 270W TDP) processor on Intel Software Development Platform with 256 GB (16 slots/ 16GB/ 3200) total DDR4 memory, ucode 0x055260, HT on, Turbo on, CentOS Linux 8.3.2011, 4.18.0-240.1.1.el8_3.crt1.x86_64, 1x Intel_SSDSC2KG96. 8280: 1-node, 2x Intel® Xeon® Platinum 8280 (28C/2.7GHz, 205W TDP) processor on Intel Software Development Platform with 192GB (12 slots/ 16GB/ 2933) total DDR4 memory, ucode 0x4002f01, HT on, Turbo on, CentOS Linux 8.3.2011, 4.18.0-240.1.1.el8_3.crt1.x86_64, 1x Intel_SSDSC2KG48. tested by Intel and results as of March 2021
 - d. 1.41x higher HPCG performance: App Version: 2019u5 MKL; Build notes: Tools: Intel MKL 2020u4, Intel C Compiler 2020u4, Intel MPI 2019u8; threads/core: 1; Turbo: used; Build knobs: -O3 -ip -xCORE-AVX-512
 - e. 1.38x higher HPL performance App Version: The Intel Distribution for LINPACK Benchmark; Build notes: Turbo: used; BIOS settings: HT=off Turbo=On SNC=Off
 - f. 1.47x higher Stream Triad Performance: App Version: McCalpin_STREAM_OMP-version; Build notes: Turbo: used; BIOS settings: HT=off Turbo=On SNC=On
 - g. 1.58x higher WRF performance & 1.11x performance/core: (geomean Conus-12km, Conus-2.5km, NWSC-3-NA-3km) App Version: 4.2.2; Build notes: Intel Fortran Compiler 2020u4, Intel MPI 2020u4; threads/core: 1; Turbo: used; Build knobs: -ip -w -O3 -xCORE-AVX2 -vec-threshold0 -ftz -align array64byte -qno-opt-dynamic-align -fno-alias \$(FORMAT_FREE) \$(BYTESWAPIO) -fp-model fast=2 -fimf-use-svml=true -inline-max-size=12000 -inline-max-total-size=30000
 - h. 1.28x higher Binomial Options performance: App Version: v1.0; Build notes: Tools: Intel C Compiler 2020u4, Intel Threading Building Blocks ; threads/core: 2; Turbo: used; Build knobs: -O3 -xCORE-AVX-512 -qopt-zmm-usage=high -fimf-domain-exclusion=31 -fimf-accuracy-bits=11 -no-prec-div -no-prec-sqrt
 - i. 1.67x higher Black Scholes performance: App Version: v1.3; Build notes: Tools: Intel MKL , Intel C Compiler 2020u4, Intel Threading Building Blocks 2020u4; threads/core: 1; Turbo: used; Build knobs: -O3 -xCORE-AVX-512 -qopt-zmm-usage=high -fimf-precision=low -fimf-domain-exclusion=31 -no-prec-div -no-prec-sqrt -fimf-domain-exclusion=31
 - j. 1.70x higher Monte Carlo performance & 1.19x performance/core: App Version: v1.1; Build notes: Tools: Intel MKL 2020u4, Intel C Compiler 2020u4, Intel Threading Building Blocks 2020u4; threads/core: 1; Turbo: used; Build knobs: -O3 -xCORE-AVX-512 -qopt-zmm-usage=high -fimf-precision=low -fimf-domain-exclusion=31 -no-prec-div -no-prec-sqrt
 - k. 1.52x higher OpenFOAM performance: (geomean 20M_cell_motorbike, 42M_cell_motorbike) App Version: v8; Build notes: Tools: Intel FORTRAN Compiler 2020u4, Intel C Compiler 2020u4, Intel MPI 2019u8; threads/core: 1; Turbo: used; Build knobs: -O3 -ip -xCORE-AVX-512. OpenFOAM Disclaimer: This offering is not approved or endorsed by OpenCFD Limited, producer and distributor of the OpenFOAM software via www.openfoam.com, and owner of the OPENFOAM® and OpenCFD® trademark
 - l. 1.64x higher GROMACS performance: (geomean ion_channel_pme, lignocellulose_rf, water_pme, water_rf) App Version: v2020.5_SP; Build notes: Tools: Intel MKL 2020u4, Intel C Compiler 2020u4, Intel MPI 2019u8; threads/core: 2; Turbo: used; Build knobs: -O3 -ip -xCORE-AVX-512
 - m. 1.60x higher LAMMPS performance: (geomean Polyethylene, Stillinger-Weber, Tersoff, Water) App Version: v2020-10-29; Build notes: Tools: Intel MKL 2020u4, Intel C Compiler 2020u4, Intel Threading Building Blocks 2020u4, Intel MPI 2019u8; threads/core: 2; Turbo: used; Build knobs: -O3 -ip -xCORE-AVX-512 -qopt-zmm-usage=high
 - n. 1.57x higher NAMD performance: (geomean Apoa1, f1atpase, STMV) App Version: 2.15-Alpha1 (includes AVX tiles algorithm); Build notes: Tools: Intel MKL , Intel C Compiler 2020u4, Intel MPI 2019u8, Intel Threading Building Blocks 2020u4; threads/core: 2; Turbo: used; Build knobs: -ip -fp-model fast=2 -no-prec-div -qoverride-limits -qopenmp-simd -O3 -xCORE-AVX-512 -qopt-zmm-usage=high
 - o. 1.61x higher RELION Plasmodium Ribosome performance: App Version: 3_1_1; Build notes: Tools: Intel C Compiler 2020u4, Intel MPI 2019u9; threads/core: 2; Turbo: used; Build knobs: -O3 -ip -g -debug inline-debug-info -xCOMMON-AVX-512 -qopt-report=5 --restrict

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 configuration** 8380: 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

27. **3.00x higher CloudXPRT Web Microservices with SLA < 1 sec.** Ice Lake: 2-socket Intel® Xeon® 8380 (40C/2.3GHz, 270W TDP) on Intel Software Development, HT on, Turbo on, SNC off, 512GB (16x32GB DDR4-3200), ucode x270, Ubuntu 20.04 LTS, 5.8.0-40-generic, CloudXPRT version 1.1, Tested by Intel and results as of February 2021. Milan: 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92, SMT on, Boost on, Power deterministic mode, NPS=1, 512 GB (16 x32GB DDR4-3200), ucode 0xa001114, Ubuntu 20.04 LTS, 5.8.0-40-generic, CloudXPRT version 1.1. Tested by Intel and results as of March 2021. 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.
28. 1.5x higher AI performance with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost vs. FP32 AMD EPYC 7763 (64C Milan): (geomean of 20 workloads including logistic regression inference, logistic regression fit, ridge regression inference, ridge regression fit, linear regression inference, linear regression fit, elastic net inference, XGBoost Fit, XGBoost predict, SSD-ResNet34 inference, Resnet50-v1.5 inference, Resnet50-v1.5 training, BERT Large SQuAD inference, kmeans inference, kmeans fit, brute_knn inference, SVC inference, SVC fit, dbscan fit, traintestsplint)
- 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, 5.4.0-65-generic/5.4.0-64-generic, 1x Intel_SSDSC2K96, Intel SSDPE2KX010T8, tested by Intel, and results as of March 2021. 7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Samsung_MZ7LH3T8/INTEL SSDSC2K96019T8, tested by Intel, and results as of March 2021.
- ResNet50-v1.5 Intel : gcc-9.3.0, oneDNN 1.6.4, BS=128, 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> , ResNet50-v1.5 AMD : gcc-9.3.0, oneDNN 1.6.4, BS=128, FP32, TensorFlow- 2.4.1, Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/image_recognition/tensorflow/resnet50v1_5
- ResNet50-v1.5 Training Intel : gcc-9.3.0, oneDNN 1.6.4, BS=256, FP32, 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> , ResNet50-v1.5 Training AMD : gcc-9.3.0, oneDNN 1.6.4, BS=256, FP32, TensorFlow- 2.4.1, Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/image_recognition/tensorflow/resnet50v1_5
- SSD-ResNet34 Intel : gcc-9.3.0, oneDNN 1.6.4, BS=1, 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/>, AMD : SSD-ResNet34, gcc-9.3.0, oneDNN 1.6.4, BS=1, FP32, TensorFlow- 2.4, Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/object_detection/tensorflow/ssd-resnet34
- BERT-Large SQuAD Intel : gcc-9.3.0, oneDNN 1.6.4, BS=1, 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/>, AMD : BERT-Large SQuAD, gcc-9.3.0, oneDNN 1.6.4, BS=1, FP32, TensorFlow- 2.4.1, Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/language_modeling/tensorflow/bert_large
- Python : Python 3.7.9, SciKit-Learn : Sklearn 0.24.1, oneDAL : Daal4py 2021.2, XGBoost : XGBoost 1.3.3 : Benchmarks: https://github.com/IntelPython/scikit-learn_bench
29. 1.3x higher AI performance with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost vs. NVIDIA A100 GPU: (geomean of 20 workloads including logistic regression inference, logistic regression fit, ridge regression inference, ridge regression fit, linear regression inference, linear regression fit, elastic net inference, XGBoost Fit, XGBoost predict, SSD-ResNet34 inference, Resnet50-v1.5 inference, Resnet50-v1.5 training, BERT Large SQuAD inference, kmeans inference, kmeans fit, brute_knn inference, SVC inference, SVC fit, dbscan fit, traintestsplint)
- 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, 5.4.0-65-generic, 1x Intel_SSDSC2K96, Intel SSDPE2KX010T8, tested by Intel, and results as of March 2021.
- DL Measurements on A100: 1-node, 2-socket AMD EPYC 7742 (64C) with 256GB (8 slots/ 32GB/ 3200) total DDR4 memory, ucode 0x8301038, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-42-generic, INTEL SSDSC2KB01, NVIDIA A100-PCIe-40GB, HBM2-40GB, Accelerator per node =1, tested by Intel, and results as of March 2021. ML Measurements on A100 : 1-node, 2-socket AMD EPYC 7742 (64C) with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0x8301034, HT on, Turbo on, Ubuntu 18.04.5 LTS, 5.4.0-42-generic, NVIDIA A100 (DGX-1) , 1.92TB M.2 NVMe, 1.92TB M.2 NVMe RAID tested by Intel, and results as of March 2021.
- ResNet50-v1.5 Intel : gcc-9.3.0, oneDNN 1.6.4, BS=1, 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>
- ResNet50-v1.5 NVIDIA :A100 (7 instance/GPU), BS=1,TensorFlow - 1.5.5 (NGC: tensorflow:21.02-tf1-py3), <https://github.com/NVIDIA/DeepLearningExamples/tree/master/TensorFlow/Classification/ConvNets/resnet50v1.5>, TF AMP (FP16+TF32); ResNet50-v1.5 Training Intel : gcc-9.3.0, oneDNN 1.6.4, BS=256, FP32, 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> , ResNet50-v1.5 Training NVIDIA :A100, BS=256,TensorFlow - 1.5.5 (NGC: tensorflow:21.02-tf1-py3), <https://github.com/NVIDIA/DeepLearningExamples/tree/master/TensorFlow/Classification/ConvNets/resnet50v1.5>, TF32;
- BERT-Large SQuAD Intel : gcc-9.3.0, oneDNN 1.6.4, BS=1, 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/>
- A100 : BERT-Large SQuAD, BS=1, A100 (7 instance/GPU), TensorFlow - 1.5.5 (NGC: tensorflow:20.11-tf1-py3), <https://github.com/NVIDIA/DeepLearningExamples/tree/master/TensorFlow/LanguageModeling/BERT>, TF AMP (FP16+TF32) ; SSD-ResNet34 Intel : gcc-9.3.0, oneDNN 1.6.4, BS=1, 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/>, SSD-ResNet34 NVIDIA :A100 (7 instance/GPU), BS=1,Pytorch – 1.8.0a0 (NGC Container, latest supported): A100 : SSD-ResNet34 (NGC: pytorch:20.11-py3), <https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/Detection/SSD>, AMP (FP16 +TF32) ;
- Python : Intel: Python 3.7.9, Scikit-Learn : Sklearn 0.24.1, OneDAL : Daal4py 2021.2, XGBoost : XGBoost 1.3.3 Python : NVIDIA A100 : Python 3.7.9, Scikit-Learn : Sklearn 0.24.1, CuML 0.17, XGBoost 1.3.0dev.rapidsai0.17, Nvidia RAPIDS : RAPIDS 0.17, CUDA Toolkit : CUDA 11.0.221 Benchmarks: https://github.com/IntelPython/scikit-learn_bench

Appendix

- 30. LINPACK.** Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 (40C/2.3GHz, 270W TDP) processor on Intel Software Development Platform with 256 GB (16 slots/ 16GB/ 3200) total DDR4 memory, ucode 0x261, HT on, Turbo on, CentOS Linux 8.3.2011, 4.18.0-240.1.1.el8_3.crt1.x86_64, 1x Intel_SSDDSC2KG96, App Version: The Intel Distribution for LINPACK Benchmark; Build notes: Tools: Intel MPI 2019u7; threads/core: 1; Turbo: used; Build: build script from Intel Distribution for LINPACK package; 1 rank per NUMA node: 1 rank per socket, tested by Intel and results as of March 2021 EPYC 7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, NPS=4, Red Hat Enterprise Linux 8.3, 4.18, 1x Samsung_MZ7LH3T8, App Version: AMD official HPL 2.3 MT version with BLIS 2.1; Build notes: Tools: hpc-x 2.7.0; threads/core: 1; Turbo: used; Build: pre-built binary (gcc built) from <https://developer.amd.com/amd-aocl/blas-library/>; 1 rank per L3 cache, 4 threads per rank, tested by Intel and results as of March 2021
- 31. Monte Carlo FSI Kernel.** Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 (40C/2.3GHz, 270W TDP) processor on Intel Software Development Platform with 256 GB (16 slots/ 16GB/ 3200) total DDR4 memory, ucode 0x261, HT on, Turbo on, CentOS Linux 8.3.2011, 4.18.0-240.1.1.el8_3.crt1.x86_64, 1x Intel_SSDDSC2KG96, App Version: v1.1; Build notes: Tools: Intel MKL 2020u4, Intel C Compiler 2020u4, Intel Threading Building Blocks 2020u4; threads/core: 1; Turbo: used; Build knobs: -O3 -xCORE-AVX-512 -qopt-zmm-usage=high -fimf-precision=low -fimf-domain-exclusion=31 -no-prec-div -no-prec-sqrt tested by Intel and results as of March 2021 EPYC 7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, NPS=4, Red Hat Enterprise Linux 8.3, 4.18, 1x Samsung_MZ7LH3T8, App Version: v1.1; Build notes: Tools: Intel MKL 2020u4, Intel C Compiler 2020u4, Intel Threading Building Blocks 2020u4; threads/core: 2; Turbo: used; Build knobs: -O3 -march=core-avx2 -fimf-precision=low -fimf-domain-exclusion=31 -no-prec-div -no-prec-sqrt tested by Intel and results as of March 2021
- 32. NAMD Geomean of ApoA1, STMV.** Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 (40C/2.3GHz, 270W TDP) processor on Intel Software Development Platform with 256 GB (16 slots/ 16GB/ 3200) total DDR4 memory, ucode 0x261, HT on, Turbo on, CentOS Linux 8.3.2011, 4.18.0-240.1.1.el8_3.crt1.x86_64, 1x Intel_SSDDSC2KG96, App Version: 2.15-Alpha1 (includes AVX tiles algorithm); Build notes: Tools: Intel MKL, Intel C Compiler 2020u4, Intel MPI 2019u8, Intel Threading Building Blocks 2020u4; threads/core: 2; Turbo: used; Build knobs: -ip -fp-model fast=2 -no-prec-div -qoverride-limits -qopenmp-simd -O3 -xCORE-AVX-512 -qopt-zmm-usage=high, tested by Intel and results as of March 2021 EPYC 7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, NPS=4, Red Hat Enterprise Linux 8.3, 4.18, 1x Samsung_MZ7LH3T8, App Version: 2.15-Alpha1 (includes AVX tiles algorithm); Build notes: Tools: Intel MKL, AOCO 2.2.0, gcc 9.3.0, Intel MPI 2019u8; threads/core: 2; Turbo: used; Build knobs: -O3 -fomit-frame-pointer -march=znver1 -ffast-math, tested by Intel and results as of March 2021
- 33. RELION Plasmodium Ribosome.** Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 (40C/2.3GHz, 270W TDP) processor on Intel Software Development Platform with 256 GB (16 slots/ 16GB/ 3200) total DDR4 memory, ucode 0x261, HT on, Turbo on, CentOS Linux 8.3.2011, 4.18.0-240.1.1.el8_3.crt1.x86_64, 1x Intel_SSDDSC2KG96, App Version: 3_1_1; Build notes: Tools: Intel C Compiler 2020u4, Intel MPI 2019u9; threads/core: 2; Turbo: used; Build knobs: -O3 -ip -g -debug inline-debug-info -xCORE-AVX-512 -qopt-report=5 -restrict, tested by Intel and results as of March 2021 EPYC 7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, NPS=4, Red Hat Enterprise Linux 8.3, 4.18, 1x Samsung_MZ7LH3T8, App Version: 3_1_1; Build notes: Tools: Intel C Compiler 2020u4, Intel MPI 2019u9; threads/core: 2; Turbo: used; Build knobs: -O3 -ip -g -debug inline-debug-info -march=core-avx2 -qopt-report=5 -restrict tested by Intel and results as of March 2021
- 34. 3.88x higher INT8 real-time inference throughput & 22.09x higher INT8 batch inference throughput on ResNet-50 with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost vs. FP32 AMD EPYC Milan**
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, 5.4.0-65-generic, 1x Intel_SSDDSC2KG96, Intel SSDPE2KX010T8, ResNet50-v1.5, gcc-9.3.0, oneDNN 1.6.4, BS=1,128, 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/>, tested by Intel, and results as of March 2021.
7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, NPS=1, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Samsung_MZ7LH3T8, ResNet50-v1.5, gcc-9.3.0, oneDNN 1.6.4, BS=1,128, FP32, TensorFlow- 2.4.1, Model : https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/image_recognition/tensorflow/resnet50v1_5, tested by Intel, and results as of March 2021.
- 35. 2.79x higher INT8 real-time inference throughput & 12x higher INT8 batch inference throughput on SSD-MobileNet-v1 with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost vs. FP32 AMD EPYC Milan**
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, 5.4.0-65-generic, 1x Intel_SSDDSC2KG96, Intel SSDPE2KX010T8, SSD-MobileNet-v1, gcc-9.3.0, oneDNN 1.6.4, BS=1,448, 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 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: <https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/>, tested by Intel, and results as of March 2021.
7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, NPS=1, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Samsung_MZ7LH3T8, SSD-MobileNet-v1, gcc-9.3.0, oneDNN 1.6.4, BS=1,448, FP32, TensorFlow- 2.4.1, Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/object_detection/tensorflow/ssd-mobilenet, tested by Intel, and results as of March 2021.

Appendix

36. Upto 25x higher AI performance with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost vs. FP32 AMD EPYC 7763 (64C Milan) ,4.01x higher INT8 real-time inference throughput & 25.05x higher INT8 batch inference throughput on MobileNet-v1 with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost vs. FP32 AMD EPYC Milan : 1-node, 2x Intel Xeon Platinum 8380 processor on Coyote Pass with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode X55260 , 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 2.4.1 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/>, test by Intel on March 2021. 1-node, 2x AMD Epyc 7763 on GigaByte with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xa001114, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Samsung_MZ7LH3T8, MobileNet-v1, gcc-9.3.0, oneDNN 1.6.4, BS=1,56, FP32, TensorFlow- 2.4.1, Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/image_recognition/tensorflow/mobilenet_v1, tested by Intel and results as of March 2021.
37. **3.18x higher INT8 real-time inference throughput & 2.17x higher INT8 batch inference throughput on BERT Large SQuAD with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost vs. FP32 AMD EPYC Milan**
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, 5.4.0-65-generic, 1x Intel_SSDSC2KG96, Intel SSDPE2KX010T8, BERT Large SQuAD , gcc-9.3.0, oneDNN 1.6.4, BS=1,128, 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/>, tested by Intel, and results as of March 2021.
7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, NPS=1, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Samsung_MZ7LH3T8, BERT Large SQuAD , gcc-9.3.0, oneDNN 1.6.4, BS=1,128, FP32, TensorFlow- 2.4.1, Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/language_modeling/tensorflow/bert_large, tested by Intel, and results as of March 2021.
38. **3.20x higher OpenSSL RSA Sign 2048 performance, 2.03x higher OpenSSL ECDHE x25519 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 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, Tested by Intel and results as of March 2021. 7763 : 1-node, 2x AMD EPYC 7763 64-Core Processor on R282-Z92-00 with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xa001114, HT On, Turbo Off, Ubuntu 20.04.1 LTS, 5.4.0-65-generic, 1x SAMSUNG_MZ7LH3T8, OpenSSL 1.1.1j, GCC 9.3.0, Tested by Intel and results as of March 2021.
39. 1) 5G vRAN: Results have been estimated or simulated: Based on 2x throughput from 32Tx32R (5Gbps) on 2nd Gen Intel® Xeon® Gold 6212U processor to 64Tx64R (10Gbps) on 3rd Gen Intel® Xeon® Gold 6338N processor at ~185W
2) VDI: <https://www.principledtechnologies.com/VMware/VMware-HCI-Intel-Optane-VDI-0420.pdf>
3) Azure stack HCI – Ice Lake Configuration:
4 Node, 2x Intel® Xeon® Gold 6330 CPU, 1x Intel® Server Board M50CYP, Total Memory: 256GB (16 x 16 GB 3200MHz DDR4 RDIMM, HyperThreading: Enable, Turbo: Enable ,Storage (boot): 1 x Intel® SSD D3-S4510 Series (480GB, 2.5in SATA 6Gb/s, 3D2, TLC), Storage: 4x Intel® SSD DC P4610 Series (3.2TB) (NVMe), Network devices: 1 x 100 GbE Intel(R) Ethernet Network Adapter E810-C-Q2, Network speed: 25 GbE, 1 x 10 GbE Intel(R) Ethernet Converged Network Adapter X550-T2, Network Speed: 10 GbE, OS/Software: Microsoft Azure Stack HCI build 17763, Benchmarks: DiskSpd (QD=8,30w:70r): 1.396M IOPS @4.03ms(r), @6.95ms(w) for 90% requests, Tested by Intel as of 12-Mar-2021.
Azure stack HCI – Cascade Lake Configuration:
4 Node, 2x Intel® Xeon® Gold 6230, 1x Intel® Server Board S2600WFT, Total Memory: 512 GB Intel® Optane™™ DC persistent Memory, 4 slots/128 GB/2666 MT/s and 192 GB, 12 slots/16 GB/2666 MT/s, HyperThreading: Enable, Turbo: Enable, Storage (boot): 1x 480 GB Intel® SSD 3520 Series M.2 SATA, Storage (cache): 2x 375 GB Intel® Optane™ DC SSD P4800X, Storage (capacity): 4x 4 TB Intel® SSD DC P4510 PCIe NVMe, Network devices: 1x 25 Gbps Chelsio* Network Adapter, Network speed: 25GbE, OS/Software: Windows Server* 2019 Datacenter Edition build 17763, Benchmarks: DiskSpd (QD=8,30w:70r): 588K IOPS @4.99ms(r), @19.54ms(w) for 90% requests, Tested by Intel as of 22-Feb-2019.

Appendix

45. Why Customers Choose Intel Delivering workload-optimized performance:
- a. Kingsoft Cloud: <https://www.intel.com/content/www/us/en/customer-spotlight/stories/kingsoft-cloud-cdn-customer-story.html>
 - b. VK: <https://www.intel.com/content/www/us/en/customer-spotlight/stories/vk-storage-customer-story.html>
 - c. CERN: <https://www.nextplatform.com/2021/02/01/cern-uses-dlboost-oneapi-to-juice-inference-without-accuracy-loss/>
 - d. Naver: <https://www.intel.com/content/www/us/en/customer-spotlight/stories/naver-ocr-customer-story.html>
 - e. Datto: <https://www.intel.com/content/www/us/en/customer-spotlight/stories/datto-customer-story.html>
 - f. BIH: <https://www.intel.com/content/www/us/en/customer-spotlight/stories/berlin-institute-health-customer-story.html>
46. P5800X IOPS Test and System Configs & Specifications
- a. World's Fastest Data Center SSD : Intel. As compared to generally available PCIe Gen4 x4 (4 lanes) Enterprise and Data Center industry SSDs. Results may vary.
 - b. Alibaba. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Results may vary.
 - c. Excelero, December 16, 2020. <https://www.excelero.com/blog/a-breakthrough-technology-helps-ai-ml-and-database-storage/>. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Results may vary.
 - d. P5800X gains vs Intel® SSD D7-P5600 NAND (both Gen4 PCIe) :Intel. Date tested – March 18, 2021. Workload – FIO rev 3.5, based on random 512B transfer size with total queue depth of 64 (QD=8, workers/jobs=8) workload, 4KB transfer size with total queue depth of 32 (QD=4, workers/jobs=8) workload, 8KB transfer size with total queue depth of 16 (QD=4, workers/jobs=4) workload in most case, except where specified.
LATENCY System configuration: Intel Optane SSD P5800X: 1.6TB: CPU: Intel® Xeon® Platinum 8380 2.30GHz 270W 40 cores per socket, CPU Sockets: 2, BIOS: SE5C6200.86B.3021.D40.2103160200, UCODE: 0X8D05A260, RAM: 32GB @3200 MT/s DDR4, DIMM Slots Populated: 16 slots, PCIe Attach: CPU (not PCH lane attach), OS: Ubuntu 20.04.2 LTS, Kernel: 5.4.0-67-generic, FIO version: 3.16; NVMe Driver: Inbox, C-states: Disabled, Hyper Threading: Disabled, CPU Governor (through OS): Performance Mode. Intel Turbo Mode, and P-states = Disabled; IRQ Balancing Services (OS) = Off; SMP Affinity, set in the OS; FIO with ioengine=io_uring. Test by Intel on 3/18/2021 vs Intel® SSD D7-P5600: see <https://www.intel.com/content/www/us/en/products/docs/memory-storage/solid-state-drives/data-center-ssds/d7-p5600-p5500-series-brief.html>
QoS, IOPS/GB system configuration: Intel Optane SSD P5800X: CPU: Intel® Xeon® Gold 6254 3.10GHz 30MB 160W 18 cores per socket, CPU Sockets: 2, BIOS: SE5C620.86B.02.01.0009.092820190230, RAM Capacity: 32G, RAM Model: DDR4, RAM Stuffing: NA, DIMM Slots Populated: 4 slots, PCIe Attach: CPU (not PCH lane attach), Chipset: Intel C610 chipset, Switch/ReTimer Model/Vendor: Intel G4SAC switch (PCIe Gen4), OS: CentOS 7.5.1804, Kernel: 4.14.74, FIO version: 3.5; NVMe Driver: Inbox, C-states: Disabled, Hyper Threading: Disabled, CPU Governor (through OS): Performance Mode. Measurements are performed on a full Logical Block Address (LBA) span of the drive. Power mode set at PM0. Test by Intel Nov 2020. vs Intel® SSD D7-P5600: see <https://www.intel.com/content/www/us/en/products/docs/memory-storage/solid-state-drives/data-center-ssds/d7-p5600-p5500-series-brief.html>
 - e. CEPH : Intel® Optane™ SSD DC P4800X + Intel® SSD P4510: Tested by Intel on 2/20/2019, 5-nodes, 2x Intel® Xeon Gold 6252 on WolfPass with 12 x 16GB 2666MHz DDR4 (total 384GB), NIC: 25x2 GbE Mellanox Connect-4 Lx CX4121A, Storage: Intel® SSD DC S3610 1.5TB, Application drive: 1x Intel® Optane™ SSD DC P4800X (375GB) + 6x Intel® SSD DC P4510 (4TB), Bios: SE5C620.86B.0D.01.0250.112320180145, ucode: 0x4000010 (HT=ON, Turbo=ON), OS: RedHat 7.6, Kernel: 3.10.0-957.el7.x86_64, Benchmark: Ceph 13.2.4 Mimic, QD= 64, Results: 4KB read = 1313300 IOPS & Latency (99.99th Percentile)= 320.13 ms, 4KB write = 291006.67 IOPS & (99.99th Percentile)= 499.68 ms, 4KB read/write (70/30) = 656517.67 IOPS & (99.99th Percentile)= 519.43 ms. Intel® SSD DC P4510 test results captured on 4TB model, while cost calculations are based on 8TB model pricing. Drive capacity is not material to test results in this benchmarking scenario.Baseline: Tested by Intel on 2/20/2019, 5-nodes, 2x Intel® Xeon Gold 6152 on WolfPass with 12 x 16GB 2666MHz DDR4 (total 384GB), NIC: 25x2 GbE Mellanox Connect-4 Lx CX4121A, Storage: Intel® SSD DC S3610 1.5TB, Application drive: 1x Intel® SSD DC P4600 (2TB) + 6x Intel® SSD DC P4500 (4TB), Bios: SE5C620.86B.0D.01.0250.112320180145, ucode: 0x4000010 (HT=ON, Turbo=ON), OS: RedHat 7.6, Kernel: 3.10.0-957.el7.x86_64, Benchmark: Ceph 13.2.4 Mimic, QD= 64, Results: 4KB read = 1149766.67 IOPS & (99.99th Percentile)= 381.58 ms, 4KB write = 230116.67 IOPS & (99.99th Percentile)= 556.35 ms, 4KB read/write (70/30) = 536652.33 IOPS & (99.99th Percentile)= 574.75 ms. Intel. Test Date Feb 20, 2019. For further test configuration and test setup documentation refer to the Ceph Benchmarking Best-Known Methods (BKMs): Installation Guide @ <https://www.intel.com/content/www/us/en/partner/cloud-insider/content-library.html?grouping=rdc%20Content%20Types&sort=title:asc>
 - f. VMware : Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Results may vary. <https://www.evaluatorgroup.com/document/lab-insight-latest-intel-technologies-power-new-performance-levels-vmware-vsan-2018-update/>
 - g. DELL EMC : Intel. <https://www.intel.com/content/dam/www/public/us/en/documents/solution-briefs/dell-emc-powermax-Optane-ssd-brief.pdf>. Results may vary.

Appendix

47. Intel® SSD D5-P5316 Massive storage capacity made flexible
- Up to 7GB/s higher sequential read - 128K sequential read bandwidth between Intel® SSD D5-P5316 15.36TB (7.0 GB/s)
 - Up to 48% better latency performance at 99.999%: Source-Intel product specification. Comparing measured performance for 4KB Random Read, QD1 latency performance at 99.999% between Intel® SSD D5-P5316 15.36TB with Intel® SSD D5-P4326 15.36TB. Measured performance are 600 us and 1150 us for Intel® SSD D5-P5316 and Intel® SSD D5-P4326 respectively.
 - Up to 5x higher endurance gen over gen - Comparing endurance (64K random write) between Intel® SSD D5-P5316 30.72TB (22,930 TBW) and Intel® SSD D5-P4326 15.36TB (4,400 TBW).
 - Up to 20x reduction of warm storage footprint. With 4TB HDD drive, it takes 10 (2U) of rack space to fill up 1PB of storage. With Intel® SSD D5-P5316 30.72TB E1.L or U.2, it takes 1U of rack space to fill up 1PB of storage.
48. Intel® Agilex™ FPGA + Quartus Prime 20.4 Software FPGA performance made flexible
- ~2x Better Fabric Performance per Watt vs. Versal: The Agilex and Versal devices (part number/speed grade) used in the perf/watt comparison are as follows: Agilex: AGF014-2, Versal: Equivalent density to AGF014-2 in 2M speed grade,ted March 2021 by Intel. Design profile used for the comparison: Base Stratix 10 frequency: 450MHz, Agilex Fmax = 450 * 1.59 = 716Mhz, Versal Fmax = 450 * 1.19 = 536Mhz, Resource usage: 60% of AGF014 resource (logic, M20K memory, DSP), power at the respective Fmax; Version of all the tools used for this data :Agilex: Quartus 20.4/PTC 21.1 b149, Versal :Vivado 2020.2/XPE: 2020.2
 - 50% faster Video IP performance: Derived from a set of five video IP designs comparing Fmax of each design achieved in Xilinx Versal ACAP devices with the Fmax achieved in Intel® Agilex™ devices, using Intel® Quartus® Prime Software (version 20.4) and Xilinx Vivado Software (version 2020.2). On geomean, designs running in the mid speed grade of Intel® Agilex™ FPGAs achieve a 50% higher in Fmax compared to the same designs running in the mid speed grade of Xilinx Versal devices (-2M speed grade), and 42% higher in Fmax compared to the same designs running in the fast speed grade of Xilinx Versal devices (-2H speed grade) and 24% higher in Fmax compared to the same designs running in the mid speed grade of Xilinx 16nm VUP devices (-2 speed grade) , tested January 2021.
 - Up to 49% faster fabric performance compared to prior generation FPGA for high-speed 5G fronthaul gateway applications –Derived from comparing the Fmax result of Agilex FPGA and Stratix 10 FPGA in a fronthaul gateway reference example using Quartus Prime 20.4 software, tested in February, 2021.
 - Software Configurations: Tests were done by running internal builds of Intel® Quartus® Prime Pro Design Software on a wide variety of internal benchmarks. The computer systems used for the evaluations were Intel® Skylake CPU @ 3.3GHz 256G Memory class machines running SUSE Linux Enterprise Server 12 operating system. The performance results represent average improvements across a wide variety of internal benchmarks, and results may vary for each testcase. Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates.
50. Cache and memory latency comparisons
- Cascade Lake: 2-socket Intel® Xeon® 8280 (2.7GHz, 28C, 205W TDP) on Intel Software Development Platform, HT on, Turbo on, SNC off, prefetchers disabled, 384 GB (12x32GB DDR4-2933), ucode 0x5000024, Ubuntu 19.04, 5.3.0-rc3-custom , kernel patched for security mitigations, Latencies using Intel Memory Latency Checker version 3.7, <https://software.intel.com/en-us/articles/intelr-memory-latency-checker>. Tested by Intel as of and results as of October 2019.
 - Ice Lake: 2-socket Intel® Xeon® 8380 (40C/2.3GHz, 270W TDP) on Intel Software Development, HT on, Turbo on, SNC off, prefetchers disabled, 512GB (16x32GB DDR4-3200), ucode 0xd0001e0, Ubuntu 20.04 LTS, 5.4.0-65-generic, Latencies using Intel Memory Latency Checker version 3.9, <https://software.intel.com/en-us/articles/intelr-memory-latency-checker>. Tested by Intel and results as of March 2021.
 - Milan: 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92, SMT on, Boost on, Power deterministic mode, NPS=1, prefetchers disabled, 512 GB (16 x32GB DDR4-3200), ucode 0xa001114, Ubuntu 20.04 LTS, 5.4.0-65-generic, Latencies using Intel Memory Latency Checker version 3.9, <https://software.intel.com/en-us/articles/intelr-memory-latency-checker>. Tested by Intel and results as of March 2021
51. **4.2x NGINX (TLS 1.2 Handshake) web server connections/sec with ECDHE-X25519-RSA2K Multi-buffer: 6338N: 1-node, 2x Intel® Xeon® Gold 6338N processor on Coyote Pass with 256 GB (16 slots/ 16GB/ 2666) total DDR4 memory, ucode x261, HT on, Turbo off, Ubuntu 20.04.1 LTS, 5.4.0-65-generic, x 3 x Quad Ethernet Controller E810-C for SFP 25 GBE, Async NGINX v0.4.3, OpenSSL 1.1.1h, QAT Engine v0.6.4, Crypto MB-ippcp_2020u3, GCC 9.3.0, GLIBC 2.31 , test by Intel on 3/22/2021. 6252N: 1-node, 2x Intel® Xeon® Gold 6252N processor on Supermicro X11DPG-QT with 192 GB (12 slots/ 16GB/ 2933) total DDR4 memory, ucode 0x5003003, HT on, Turbo off, Ubuntu 20.04.1 LTS, 5.4.0-65-generic, x 2 x Quad Ethernet Controller XXV710 for 25GbE SFP28, 1 x Dual Ethernet Controller XXV710 for 25GbE SFP28, Async NGINX v0.4.3 , OpenSSL 1.1.1h, GCC 9.3.0, GLIBC 2.31 , test by Intel on 1/17/2021.**
52. **50% higher performance on MySQL, Redis, and Nginx:** Alibaba 7th Gen ECS Cloud Server claim, <https://developer.aliyun.com/article/781850?spm=a2c6h.17735062.0.0.5bca49f2vpLXaT>

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: SE5C6200.86B.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](https://www.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-learn 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-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.