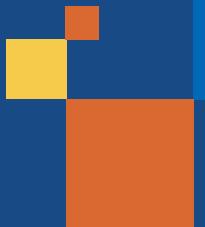


**Intel® Xeon® Scalable Processors: NEX Eagle Stream Platform**

# Xeon Hero Features Overview

Document Number: 784478

August 2023



**intel**<sup>®</sup>

# Notices and Disclaimers

Performance varies by use, configuration and other factors. Learn more on the [Performance Index site](#).

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. Results have been estimated or simulated.

Your costs and results may vary.

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

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.

Code names are used by Intel to identify products, technologies, or services that are in development and not publicly available. These are not "commercial" names and not intended to function as trademarks.

© 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.

# Agenda

- Product Portfolio/Features
- Intel® Advanced Matrix Extensions (Intel® AMX)
- Intel® Data Streaming Accelerator (Intel® DSA)
- Intel® In-memory Analytics Accelerator (Intel® IAA)
- Intel® Dynamic Load Balancer (Intel® DLB)
- Intel® QuickAssist Technology (Intel® QAT)
- Summary

# Product Portfolio/Features

# NEX Eagle Stream Platform

## 4th Gen Intel® Xeon® Scalable Processors

### SAPPHIRE RAPIDS - SP

Enhanced processor and platform performance

Enhanced memory support and I/O

Enhanced AI Acceleration

New Security Features:  
Hardware Enforced Execution Controls

AVAILABLE NOW

## 4th Gen Intel® Xeon® Scalable Processors with Intel® vRAN Boost

### SAPPHIRE RAPIDS - EE

Includes all the features and technologies on Sapphire Rapids.

Exclusive, built-in vRAN hardware acceleration tuned and optimized for optimal power-to-performance.

New Intel® Trust Domain Extensions (Intel® TDX)

AVAILABLE NOW

## Intel® Xeon® Scalable Processors for Edge/IoT

### SAPPHIRE RAPIDS - EE

Includes all the features and technologies on Sapphire Rapids.

Tuned and optimized for optimal power-to-performance for edge usages.

New Intel® Trust Domain Extensions (Intel® TDX) for enhanced confidential computing at the edge

IN DEVELOPMENT

## Emerald Rapids – SP

Includes all the features and technologies on Sapphire Rapids.

Expanded Core Count Options and Memory Value.

VMs and Containers are more secure through Enhanced Data Protection with Trust Domain Extensions

IN DEVELOPMENT

Network & Edge (IoT) solutions of use conditions

Long life

Extended support and availability

Tuned and optimized for specific use conditions

Disclaimer: Not all features are available on all SKUs

# 4th Gen Intel® Xeon® Scalable Processor

Up to  
**15%** Higher Instructions per Cycle vs. prior gen

**Enhanced Processor Performance**  
Compared to Ice Lake-SP<sup>1</sup>  
1 to 8\* socket support

**Intel® Quick Assist Technology**

**PCI Express 5.0 (80 lanes)**  
Up to 2.5x PCIe I/O bandwidth increased

**Compute Express Link (CXL)**  
Four (x16) CXL devices per CPU<sup>3</sup>

**DDR5 Memory**  
8 channels, 16 DIMMs per socket, up to 4800 mt/s  
Up to 50% memory bandwidth compared to DDR4

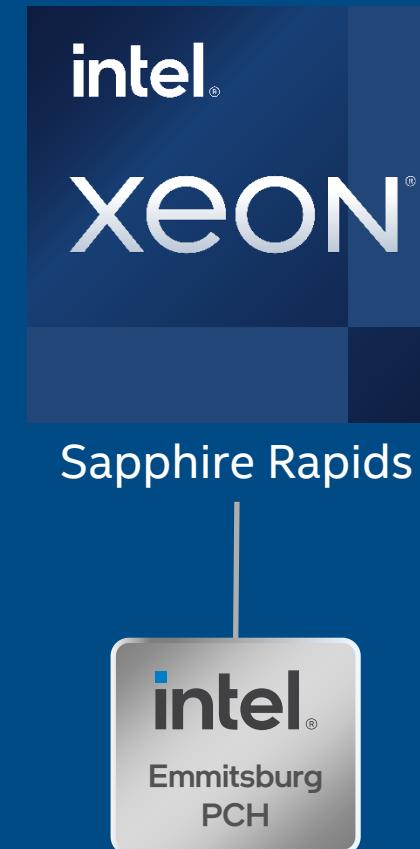


 **Intel Speed Select Technology**

 **Intel Resource Director Technology**

 **Advanced Reliability, Availability, Serviceability**

 **Intel Platform Monitoring Technology**



**New Built-In AI Acceleration**  
Enhanced Intel Deep Learning Boost w/ Advanced Matrix Extensions (AMX)  
Project up to 4x DL inference perf. vs. Ice Lake-SP<sup>2</sup>  
Project up to 2x DL training perf vs. Cooper Lake<sup>2</sup>



**New Integrated Accelerators**  
New Intel Data Streaming Accelerator (DSA)  
New In-Memory Analytics Accelerator (IAA)  
New Intel Dynamic Load Balancer (DLB)



**New Built-In vRAN Acceleration**  
Intel® Advanced Vector Extensions for vRAN



**Intel Ultra Path Interconnect 2.0**  
Up to 4 links per processor (up to 16 GT/s)



**New HW-Enhanced Security**  
New & Enhanced Technologies  
Intel Software Guard Extensions (SGX) w/ integrity  
Intel Platform Firmware Resilience (PFR)

1. Configuration: SPEC CPU2017, DDR5 memory at 4800 MT/s.

2. Intel internal analysis – early performance projections. Performance details referenced above are preliminary and subject to change without notice.

3. All 80 lanes support CXL, but Sapphire Rapids supports only four CXL devices per CPU; x8 (degraded mode) also supported.

\* NEX designs only support 1 socket and 2 sockets.

# Sapphire Rapids Accelerator Engines

## *Purpose-built Workload Features*



### **AMX**

(Intel® Advanced Matrix Extensions)

Built-in AI Acceleration designed for delivering a significant leap in performance for deep learning inference and training.

### **DSA**

(Intel® Data Streaming Accelerator)

Optimizing streaming data movement and transformation operations common in storage, networking and analytics.

### **IAA**

(Intel® In-Memory Analytics Accelerator)

Increases queries per second for analytics workloads.

### **DLB**

(Intel® Dynamic Load Balancer)

Enhances system performance related to handling network data on multi-core Intel Xeon Scalable processors.

### **QAT**

(Intel® QuickAssist Technology)

Accelerate performance for tasks such as security, private key protection, and data compression/decompression.

# Intel® Advanced Matrix Extensions (Intel® AMX)

# Intel® AMX

## New built-in AI acceleration engine

### 2<sup>nd</sup> Gen Intel® Xeon® Scalable Processors

- Intel® Deep Learning Boost (Intro)
- Intel® AVX-512 (VNNI/INT8)

### 3<sup>rd</sup> Gen Intel® Xeon® Scalable Processors

- Intel® Deep Learning Boost
- Intel® AVX-512 (VNNI/INT8 & BF16)

### 4<sup>th</sup> Gen Intel® Xeon® Scalable Processors

- Intel® Deep Learning Boost
- Intel® AMX – INT8 and BFloat16 support
- Intel® AVX-512 (VNNI/INT8)



## VALUE PROP

- Extensive hardware (dedicated silicon/TILEs and set of matrix multiply instructions/TMUL) and software (across market relevant Frameworks, toolkits and libraries) optimizations, to enhance built-in AI acceleration performance on Xeon Scalable
- Intel® AMX/TMUL supports INT8 (Inference) and BFloat16 (Training/Inference) datatypes

## TARGET WORKLOADS/USAGES

- Image recognition
- Recommendation systems
- Machine/Language Translation
- Reinforcement Learning
- Natural Language Processing/NLP
- Media Processing & Delivery
- Media Analytics

## WHAT IS IT?

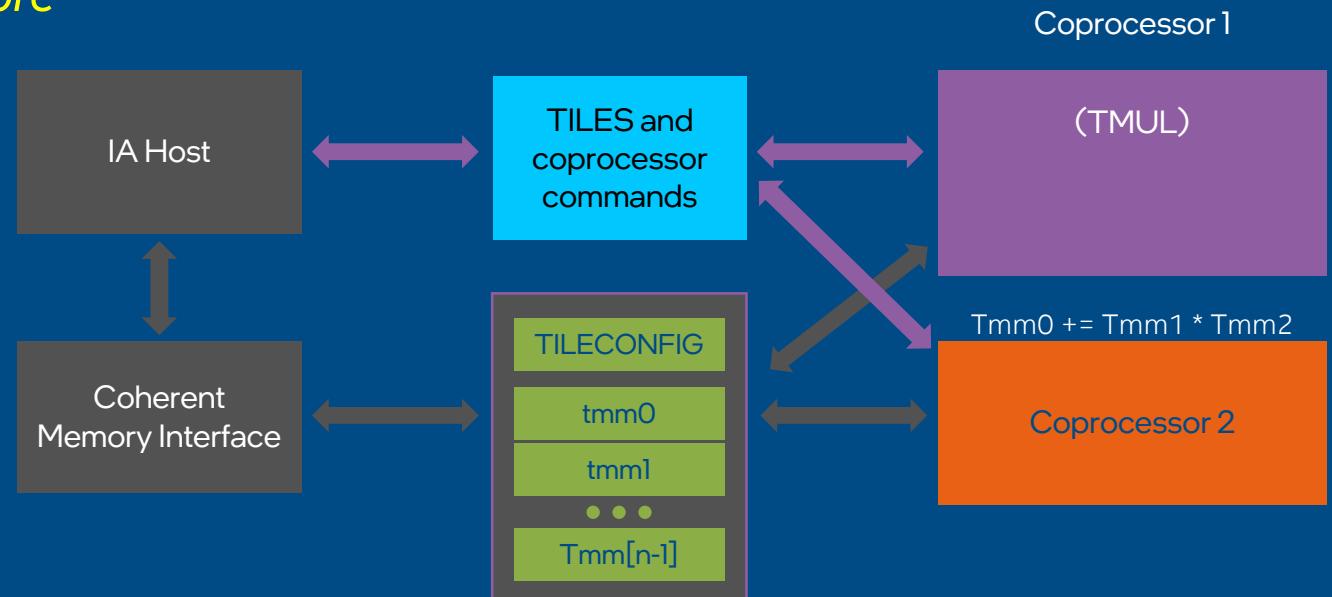
- Significant performance increase for AI/Deep Learning Inference and Training workloads compared to previous generation Xeon Scalable processors

# Intel® AMX

*DL accelerator performance built into every core*

Intel Advanced Matrix Extensions provides a leap in deep learning inference & training performance:

- **For Deep Learning datatypes**
  - int8 (all sign combinations) with int32 accumulation
  - Bfloat16 with IEEE SP accumulation
- **Acceleration at the instruction set level**
  - Full Intel® Architecture (IA) programmability
  - Very low latency if part of normal IA SW flow
  - Available through Intel® oneAPI DNNL and industry-relevant frameworks: TensorFlow, Pytorch, OpenVINO, MXNet, and others



## TMUL (Tile Matrix Multiply)

- Set of matrix multiply instructions, the first operators on TILES



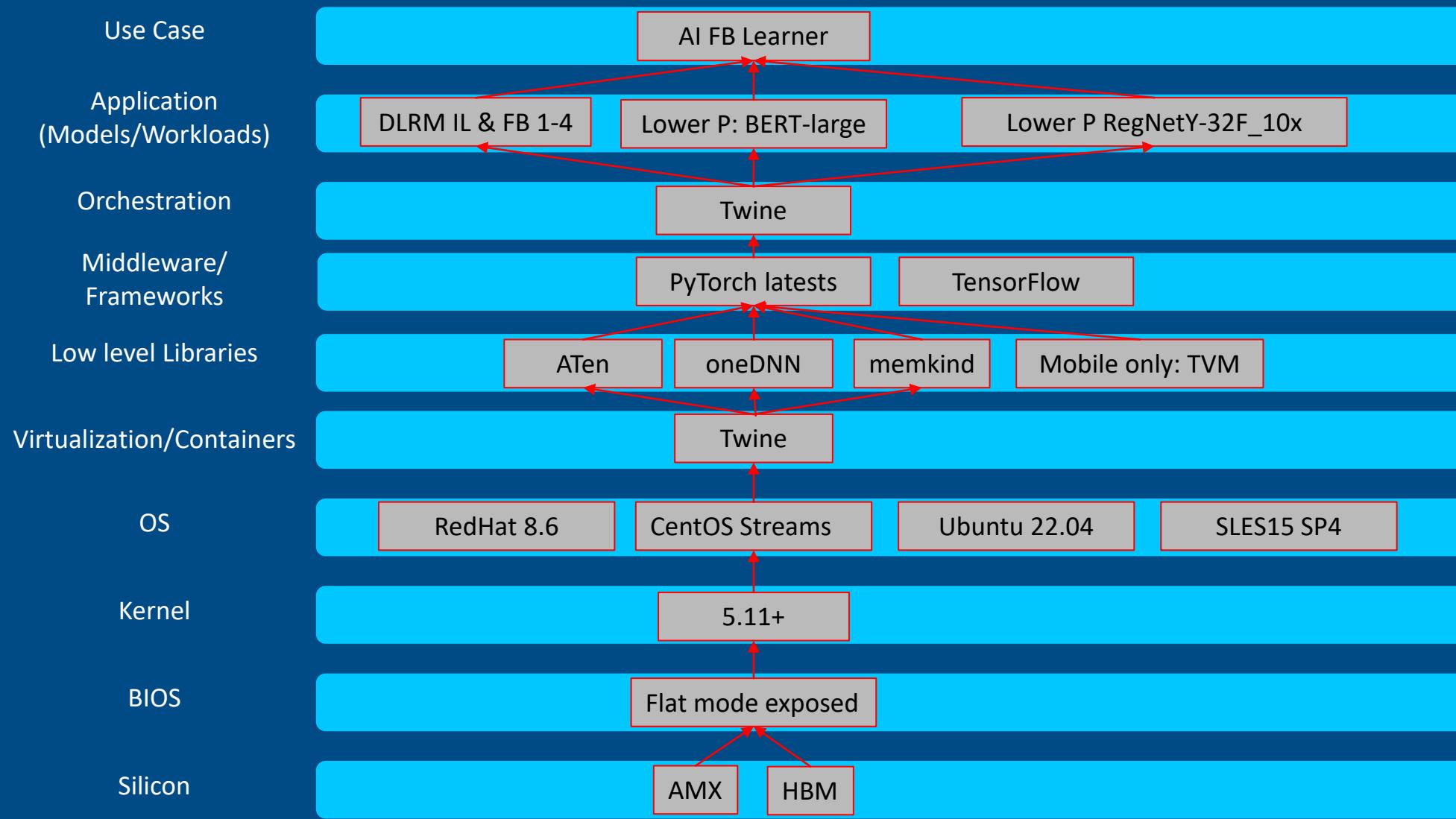
## TILES

- A new expandable 2d register file
- Register file supports basic data operators
- Declares the state and OS-managed by XSAVE\* architecture



# Intel® AMX

Xeon Feature workload + SW stack enabling example view



Disclaimer: Details referenced above are preliminary and subject to change without notice

# Intel® Data Streaming Accelerator (Intel® DSA)

# Intel® DSA

*Improves streaming data movement and transformation operations*

2<sup>nd</sup> Gen Intel® Xeon® Scalable Processors

3<sup>rd</sup> Gen Intel® Xeon® Scalable Processors

4<sup>th</sup> Gen Intel® Xeon® Scalable Processors

Intel Quick Data Technology

(Former codename: *Crystal Beach DMA* or *CBDMA*)

Intel Quick Data Technology

(Former codename: *Crystal Beach DMA* or *CBDMA*)

Data Streaming Accelerator 1.0

## VALUE PROP

- DSA improves performance of applications reliant on data movement. It also eliminates the various CPU compute and latency overheads of typical DMA controllers because it is tightly integrated with the CPU cores, is available in User Mode and works from an application's virtual address space allowing for virtualization or multi-tenancy.
- Cores offload data movement operations to DSA, freeing CPU cycles for higher priority work.

## TARGET WORKLOADS/USAGES

- Virtualization: VM fast-checkpoint analysis
- Network: vSwitch network virtualization
- Storage: Fast replication across non-transparent bridge (NTB), Integrity check offload
- Application usage examples: Messaging, ERP, IMDB, Analytics, AI training

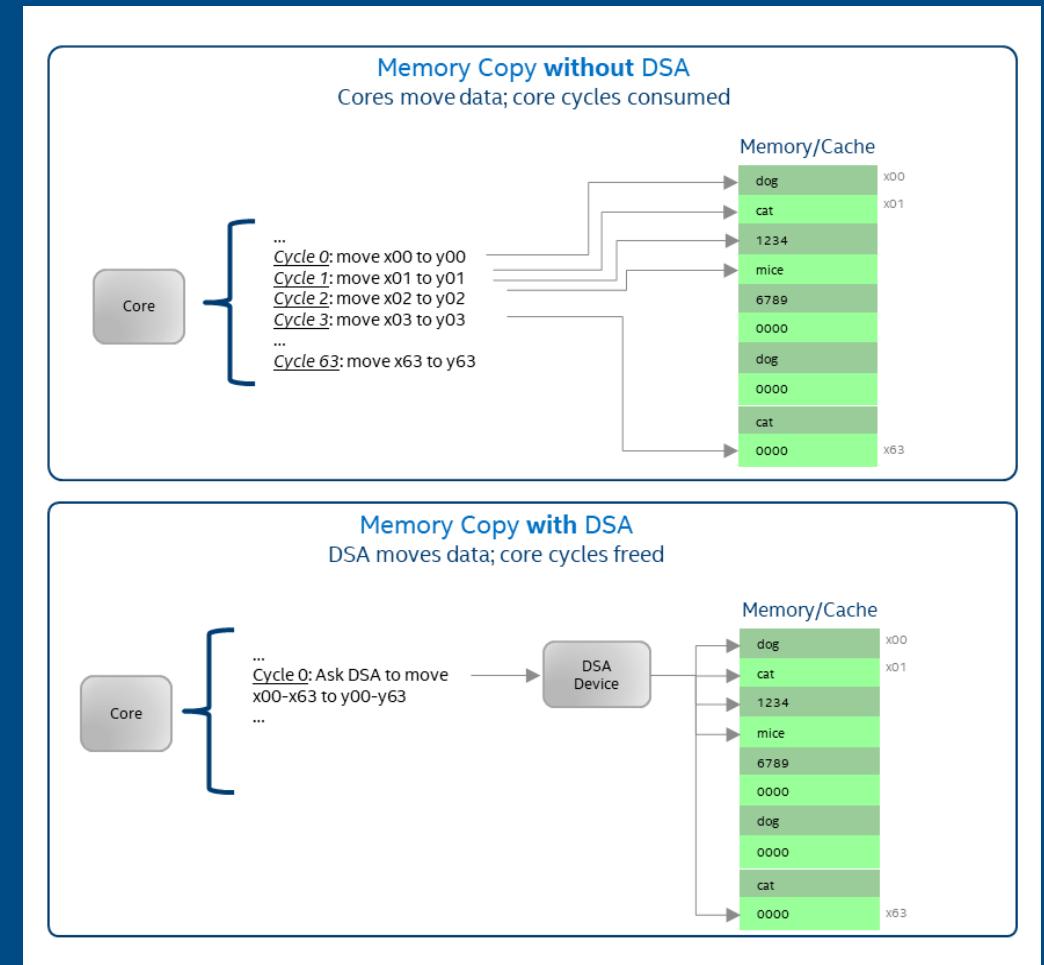
## WHAT IS IT?

- Integrated accelerator IP accelerates common data mover operations
- Up to 4 devices per socket for up to 120GB/s (240GB/s bi-directional) bandwidth
- Advanced offload optimizations such as support for AIA instructions & SVM
- Scalable-IOV support for seamless sharing from application user-process, containers and/or virtual machines.

# Intel® DSA

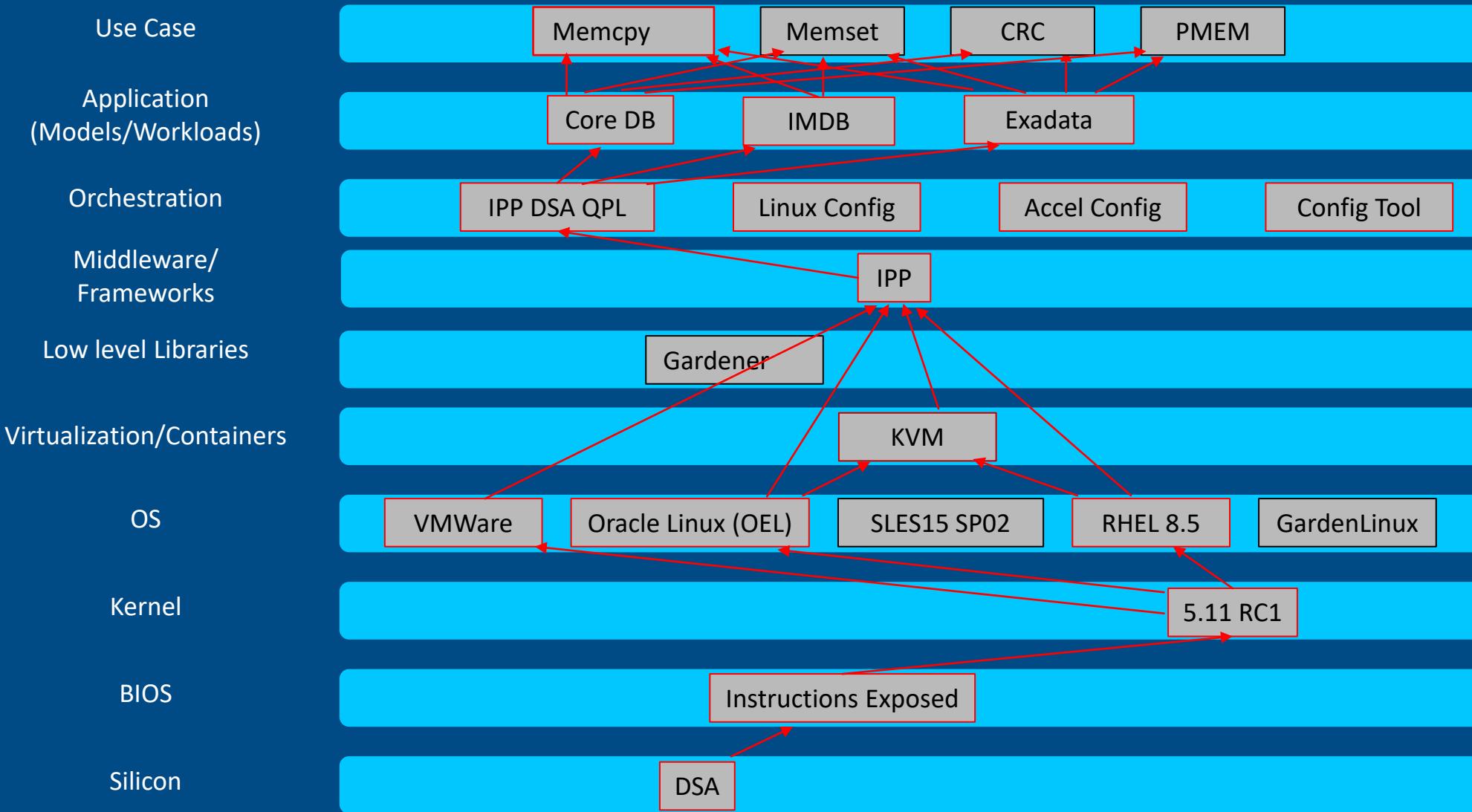
*Improves performance of applications reliant on data movement*

- DSA is an accelerator built into Sapphire Rapids CPU that performs common data mover operations.
- DSA improves the performance of applications reliant on data movement. Cores offload data movement operations to DSA, freeing CPU cycles for higher priority work.
- DSA can be used by/with...
  - **Hypervisors:** DSA aids with VM redundancy and availability via assisting VM migration and check pointing
  - **Storage Apps:** improves performance by offloading data movement between memory and IO
  - **Networking Apps:** improves packet processing throughput via data copy offload
  - **OS:** may improve OS performance by offloading cache flushing, page zeroing, memory moves, etc.
- DSA replaces Intel® Quick Data Technology (prior codename Crystal Beach DMA, or CBDMA), and improves upon it via:
  - Increased bandwidth
  - New functions to assist hypervisors with VM redundancy and availability
  - Support for Scalable I/O Virtualization and Shared Virtual Memory
  - Supports new instructions enabling efficient work dispatch
- DSA supports data movement between CPU caches and/or all EGS platform-compatible attached memory and IO devices. DSA can be shared concurrently across OS/VMM, virtual machines, containers, and application processes.



# Intel® DSA

Xeon Feature workload + SW stack enabling example view



Disclaimer: Details referenced above are preliminary and subject to change without notice

# Intel® In-Memory Analytics Accelerator (Intel® IAA)

# Intel® IAA

*Increases query throughput and decreases memory footprint in Analytics*



## VALUE PROP

- Increases query throughput for in-memory database and analytics workloads
- Decreases memory footprint for analytics workloads

## TARGET WORKLOADS/USAGES

- Commercial in-memory databases
- Open source in-memory database/data stores: RocksDB, Redis, Cassandra, MySQL, PostgreSQL, MongoDB, Memcached, and more
- Columnar Formats for Big Data Analytics: Apache Parquet, Apache ORC

## WHAT IS IT?

- Integrated accelerator IP that accelerates analytics primitives (scan, filter, etc.), CRC calculations, compression, decompression, and more
- Up to 4 Intel® IAA devices per socket
- Advanced offload optimizations such as support for AIA instructions & SVM
- Scalable-IOV support for seamless sharing from application user-process, containers and/or virtual machines

# Intel® IAA

*For Analytics: Intel® IAA increases query throughput and decreases memory footprint*

Analytics WLs typically rely on two operations...

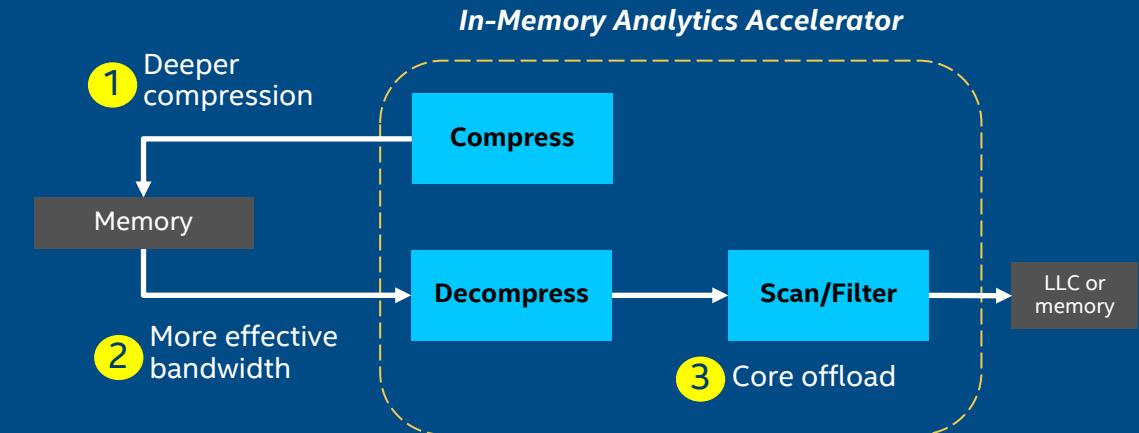
- Scan/filter: identify relevant data in large data set
- Compress/decompress: reduces memory consumption

**IAA increases query throughput and decreases memory footprint via:**

- ① Deeper compression than software-only techniques
- ② More effective bandwidth, as deeply compressed data consumes less bandwidth
- ③ Core offload, as Intel® IAA performs computationally demanding scan and filter operations in place of cores

IAA is an accelerator built into Sapphire Rapids CPU that accelerates analytics primitives (scan, filter, etc.), CRC calculations, compression, decompression, and more.

- Works with all EGS compatible memory
- Negligible power footprint



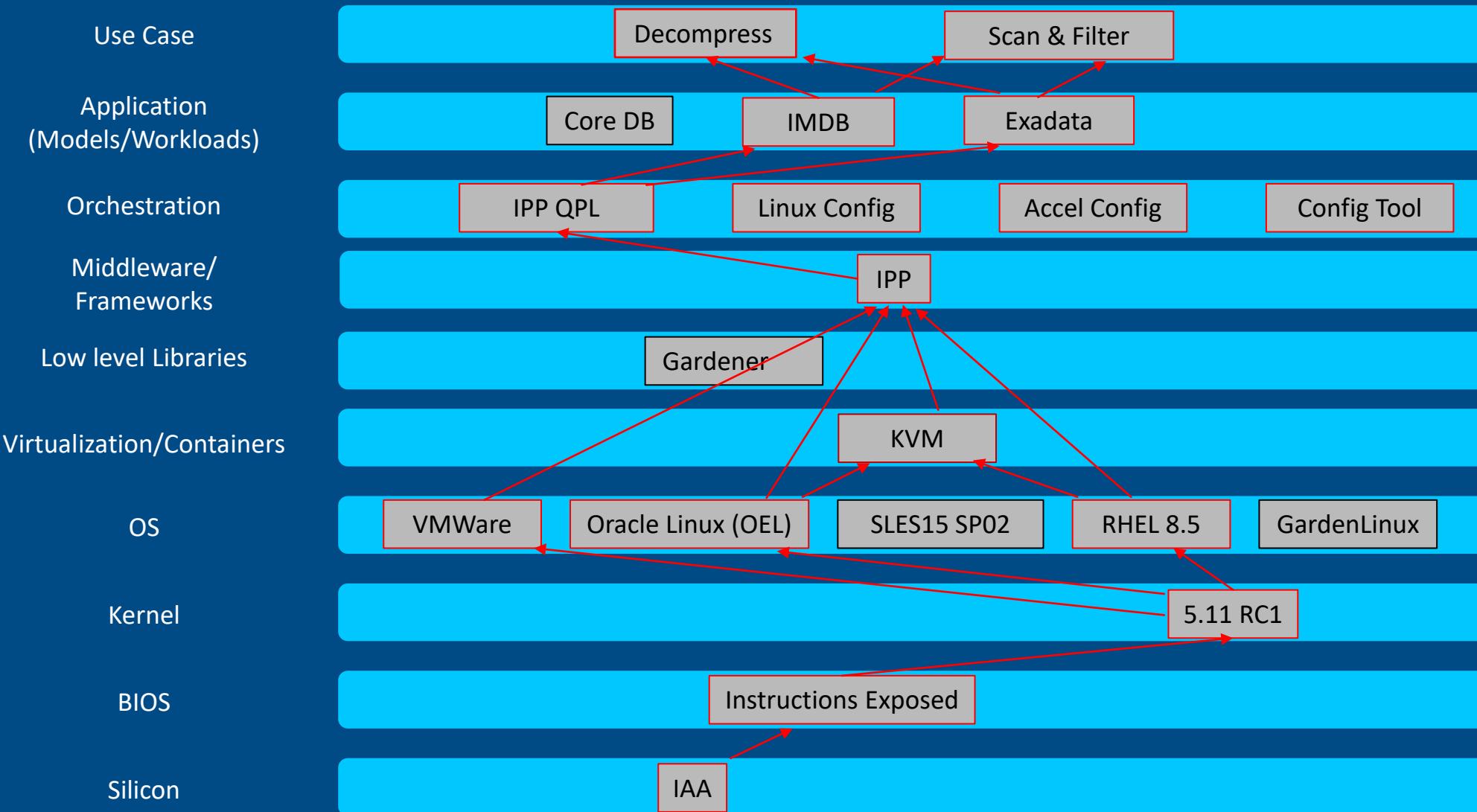
③ While profiling their PrestoDB analytics solution, Facebook observed...

***"Nearly 60 percent of our global Presto CPU time is attributed to table scan..."***

*Aria Presto: Making table scan more efficient, Facebook Engineering ([link](#))*

# Intel® IAA

Xeon Feature workload + SW stack enabling example view



Disclaimer: Details referenced above are preliminary and subject to change without notice

# Intel® Dynamic Load Balancer (Intel® DLB)

# Intel® DLB

*Scalable. Dynamic. Atomic*

2<sup>nd</sup> Gen Intel® Xeon® Scalable Processors

3<sup>rd</sup> Gen Intel® Xeon® Scalable Processors

4<sup>th</sup> Gen Intel® Xeon® Scalable Processors

Intel DLB 2.0

## VALUE PROP

- Improves the system performance related to handling network data on multi-core Intel® Xeon Scalable processors.
- Improved performance for Distributed Processing, Dynamic Load Balancing, Dynamic Network Processing Reordering

## TARGET WORKLOADS/USAGES

- IPSec security gateway
- VPP router
- UPF
- vSwitch
- Streaming data processing
- Elephant flow handling

## WHAT IS IT?

- Improved performance over software load balancing (packet reordering, priority scheduling, load-balancing)
- Improved dynamic redistribution of load across cores when static NIC distribution (RSS/FDIR) causes load-imbalance

# Intel® DLB

Intel® DLB improves the system performance related to handling network data on multi-core Intel® Xeon Scalable processors:

- **Distributed Processing**

- Intel® DLB enables the efficient distribution of network processing across multiple CPU cores/threads;

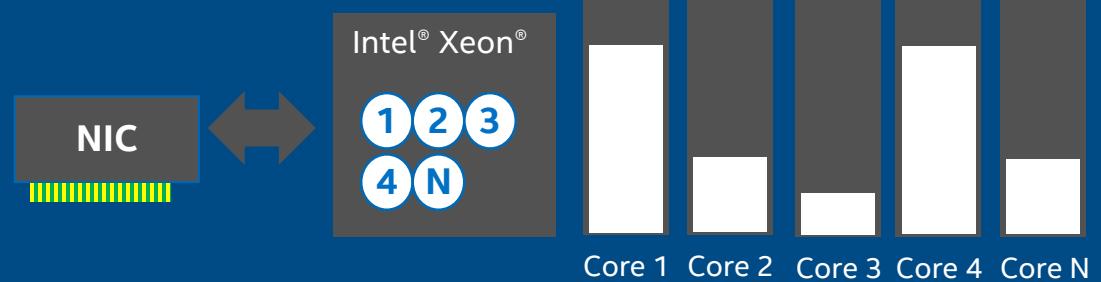
- **Dynamic Load Balancing**

- Intel® DLB dynamically distributes network data across multiple CPU cores for processing as the system load varies;

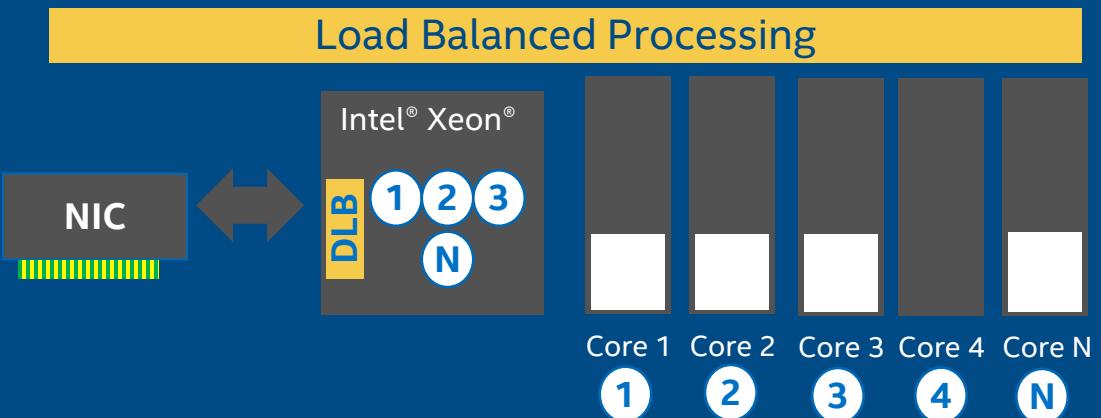
- **Dynamic Network Processing Reordering**

- Intel® DLB restores the order of networking data packets processed simultaneously on CPU cores;

## Without DLB: CPU Utilization Per Core

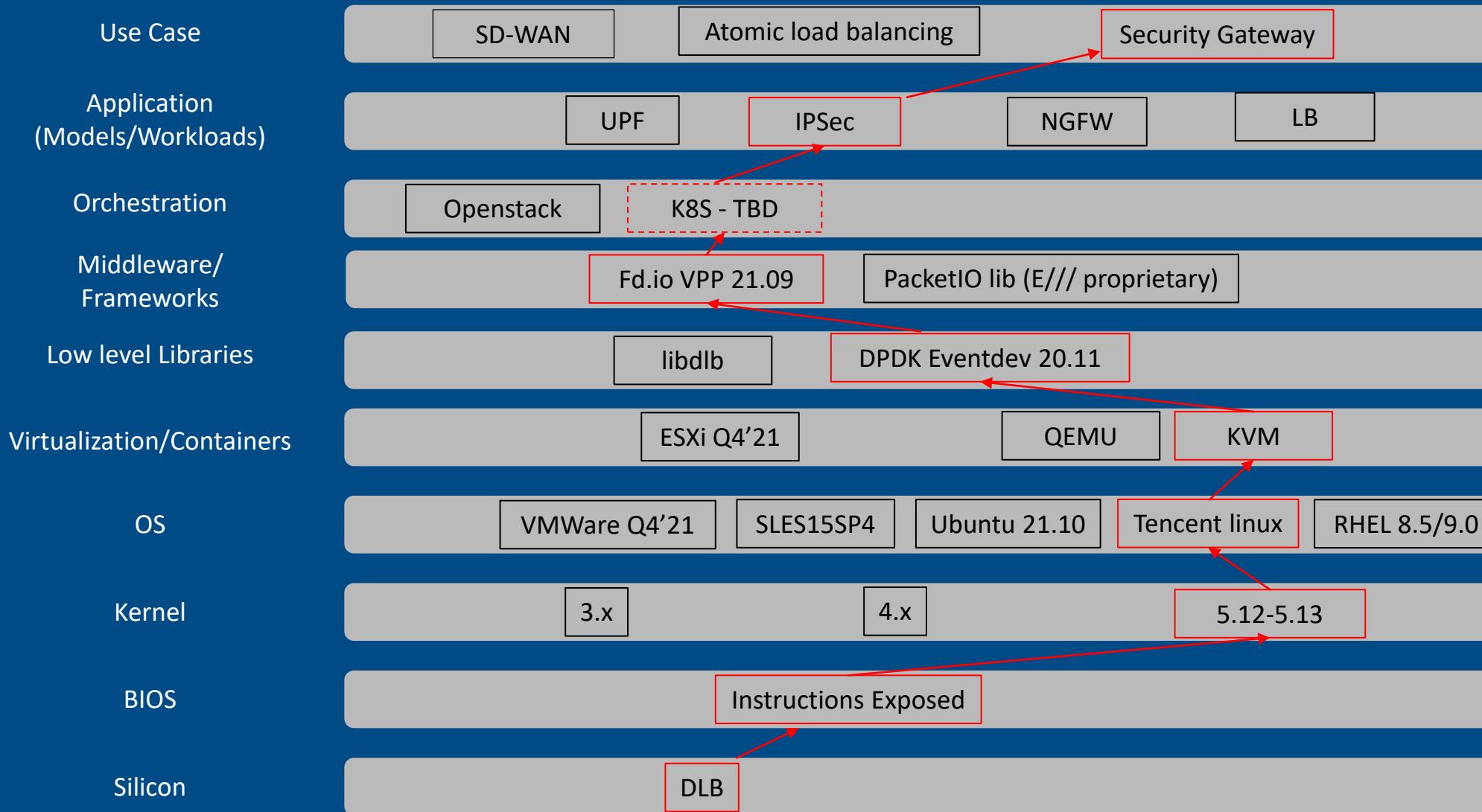


## With DLB: CPU Utilization Per Core



# Intel® DLB

Xeon Feature workload + SW stack enabling example view

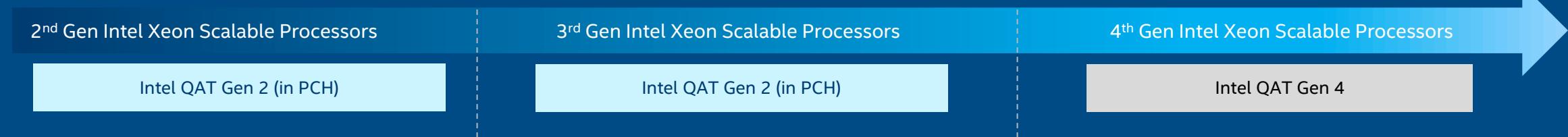


Disclaimer: Details referenced above are preliminary and subject to change without notice

# Intel® QuickAssist Technology (Intel® QAT)

# Intel® QAT

*Hardware native accelerator in 4th Gen Intel Xeon Scalable Processors*



## VALUE PROP

- Offers outstanding capabilities: 400/200Gbs Crypto, 160Gbs verified compression, 100kops PFS ECDHE & RSA 2K Decrypt

## TARGET WORKLOADS/USAGES

- Proxies and Content Delivery (NGINX & HAProxy)
- Distributed storage systems (Ceph)
- Databases/Data lakes
- Service Mesh (ENVOY/BoringSSL)
- http compression
- Memory infrastructure optimization
- Packet Processing (VPP/DPDK)
- Data Deduplication

## WHAT IS IT?

- Scalable hardware accelerators exposed to IA.
- High performance Security, Private Key Protection, and Compression/Decompression
- New usages for high compression ratio in the data path - Storage, HPC & Big Data, File Systems, Database, NextGen Firewall, Application Delivery Controller, Wireless Core & Edge, Content Delivery Controllers.

# Intel® QAT

*Hardware Native in 4th Gen Intel® Xeon® Scalable Processors*

- Scalable hardware accelerators exposed to IA. High performance Security, Private Key Protection, and Compression/Decompression
  - Server: Content Delivery Networks, secure browsing, email, search, secure multi-tenancy, Security Middle Box, IPsec, SSL/TLS, OpenSSL
  - Networking: firewall, IDS/IPS, VPN, secure routing, Web proxy, WAN optimization (IP Comp), 3G/4G authentication
  - Storage: real-time data compression, secure storage.
  - Big Data & HPC: Secure Storage & Low Latency Compression/Decompression for efficient storage and data movement.
- Intel® QAT on Sapphire Rapids offers outstanding capabilities: 200Gbs Crypto, 160Gbs verified compression, 100kops PFS ECDHE & RSA 2K Decrypt



Cloud

More Secure Gateway  
SDWAN  
Compute  
Content Delivery



Networking

More Secure Routing  
Load Balance  
Web Proxy  
WAN Optimization  
Infrastructure Gateway



Big Data / HPC

Analytics (Hadoop\*)  
More Secure Data Transfer  
Lossless Compression

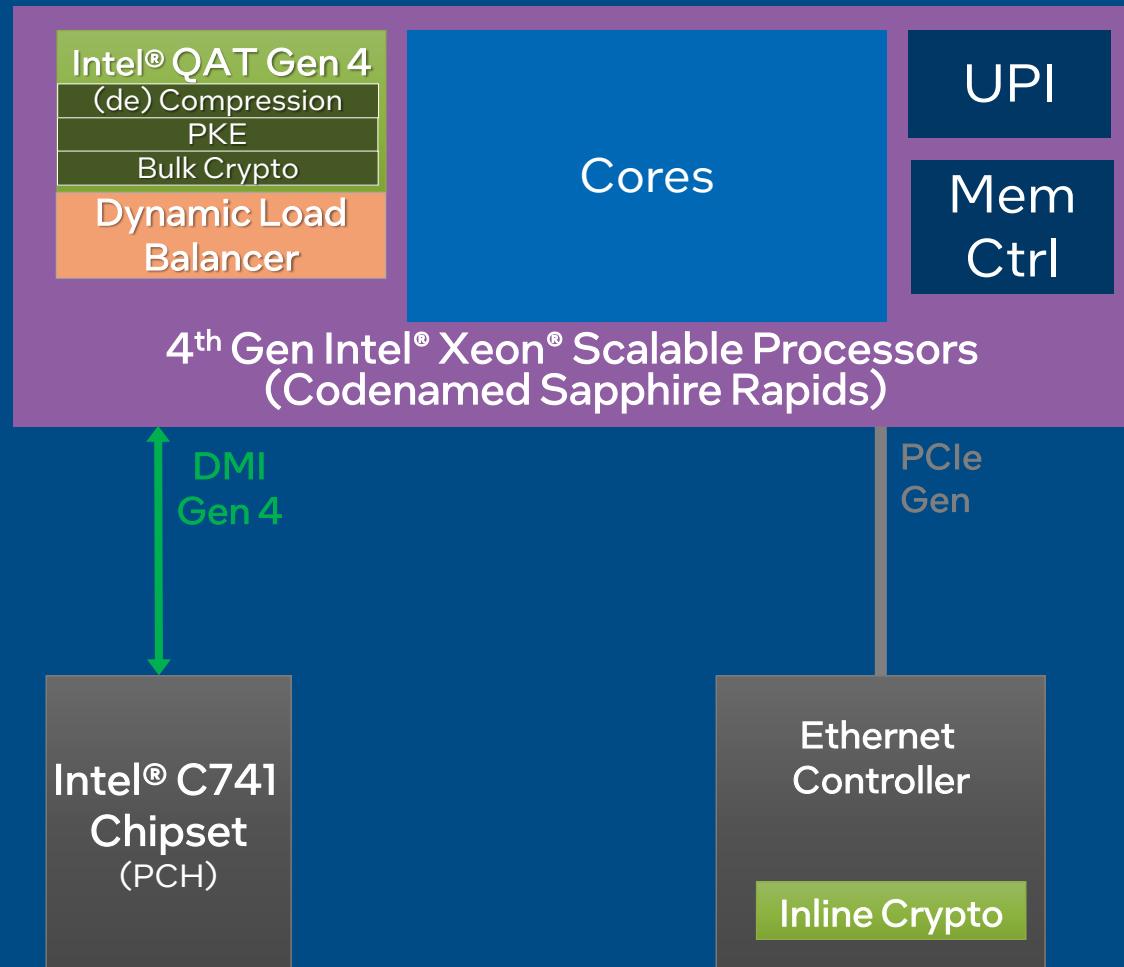


Storage

Data Compression  
More Secure Storage  
Hyper-Convergence

[www.intel.com/quickassist](http://www.intel.com/quickassist)

# Intel® QAT



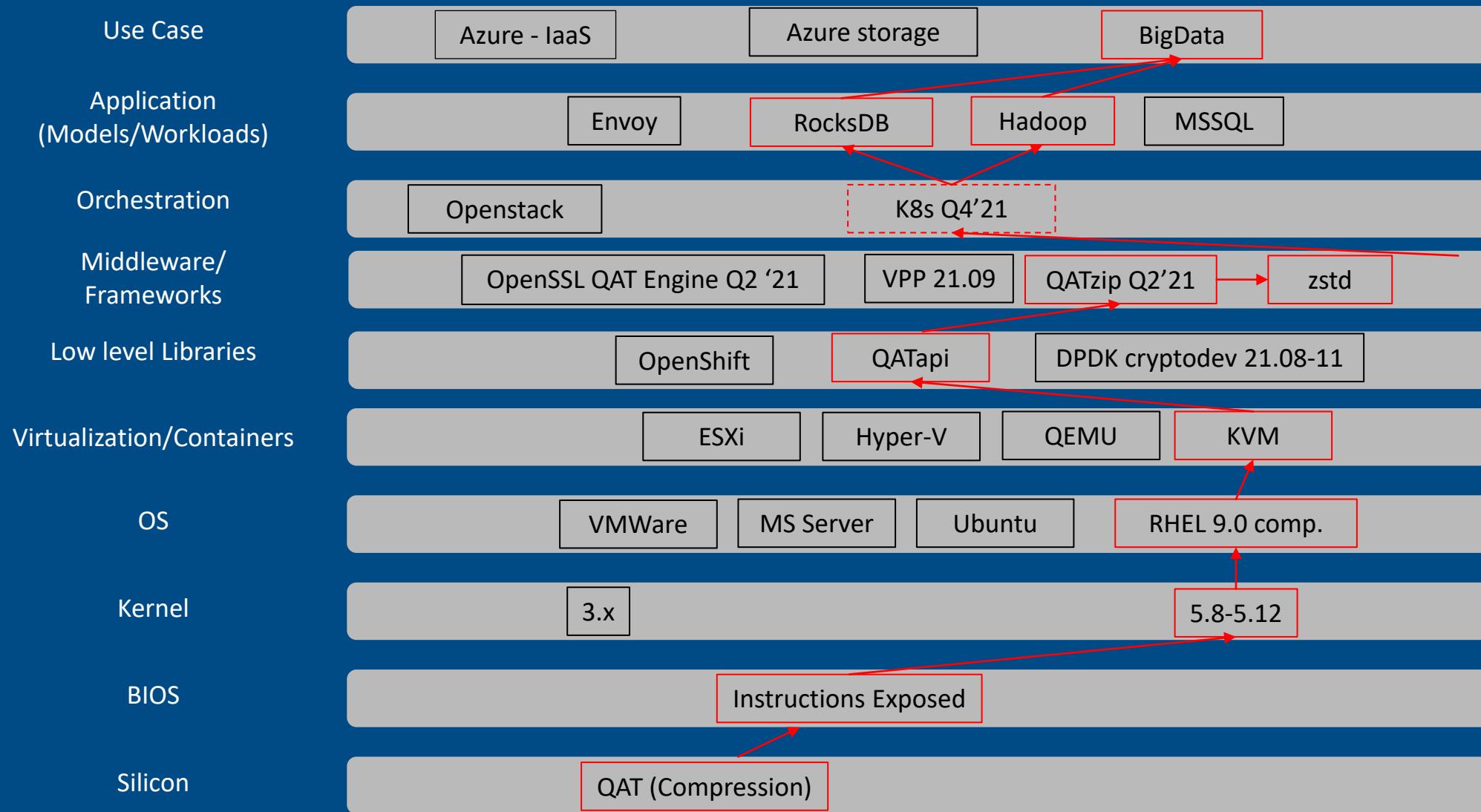
## Maximum Capacities\*

SKUs	Sapphire Rapids	
	XCC	MCC
QAT Compression (Gb/s)	160	80
QAT Crypto AEAD (Gb/s)	400	200
QAT Crypto Other modes (Gb/s)	200	100
QAT PKE (KOp/s)	100	150
DLB (Md/s)	400	200

\*Device BW shown are targeted max device BW; actual BW may depend on operation and data

# Intel® QAT

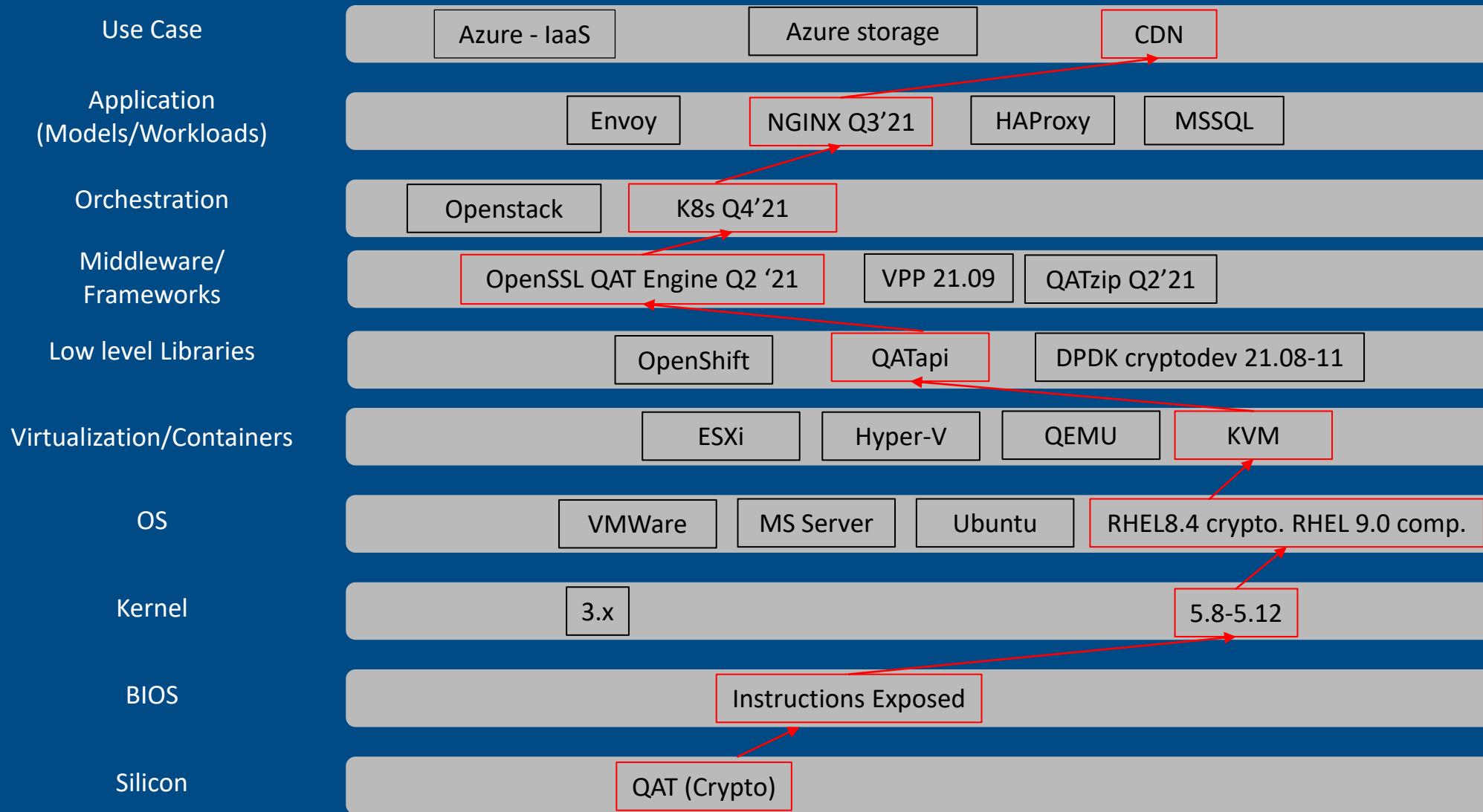
Xeon Feature workload + SW stack enabling example view



Disclaimer: Details referenced above are preliminary and subject to change without notice

# Intel® QAT

Xeon Feature workload + SW stack enabling example view

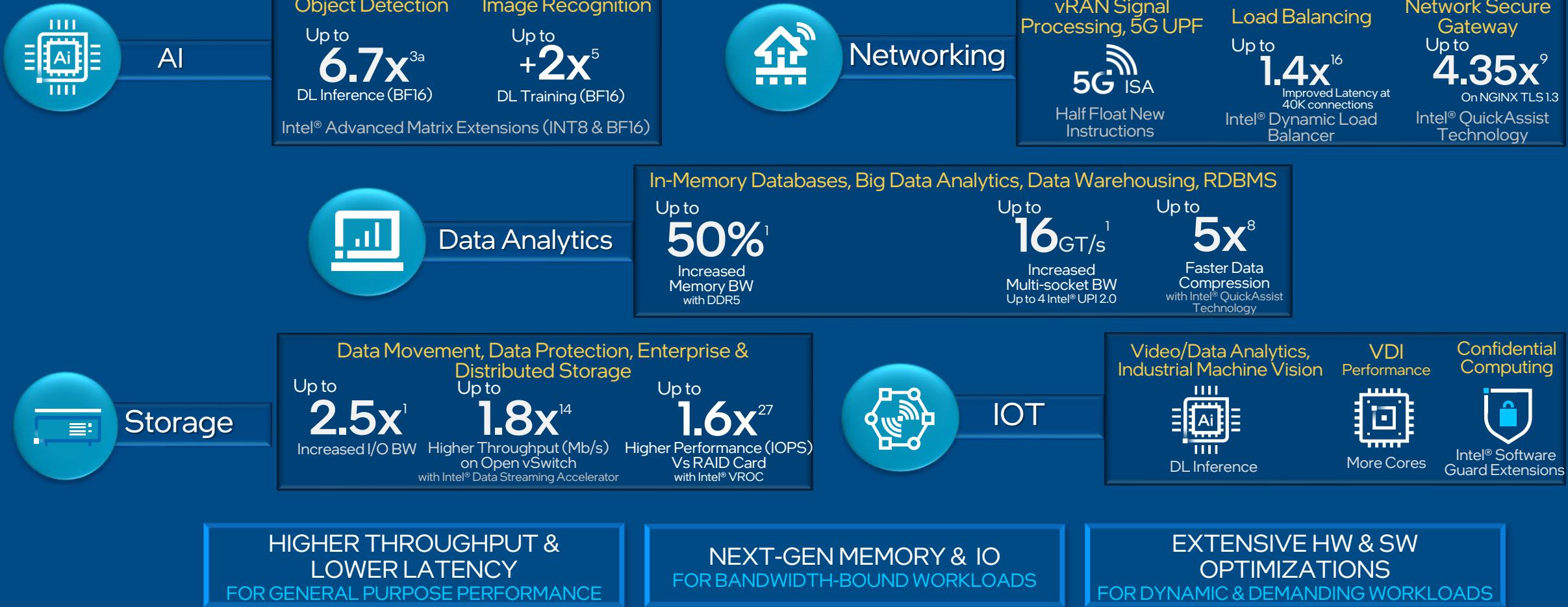


Disclaimer: Details referenced above are preliminary and subject to change without notice

# Summary

## 4TH GEN INTEL XEON SCALABLE PROCESSORS

### Accelerating Customer Usages with Unique Workload Optimizations



new  
enhanced

Intel® AMX, Intel® QAT, Intel® DLB, Intel® DSA, HFNI (5G ISA)  
Intel® SGX

<sup>1</sup>Compared to 3rd Gen Intel Xeon Scalable processors (Ice Lake)

Network and Edge Group



The Intel Xeon logo consists of the word "intel" in its signature white font with a blue outline, followed by "xeon" in a bold, blue, sans-serif font.

Accelerate with Xeon



A small version of the Intel logo, featuring the word "intel" in white on a blue square background.

# Appendix

# Sapphire Rapids Performance – Configuration/Disclaimers

## SECTION: SEGMENT USE CASES & MARKET TRENDS

1. **5G CORE/UPF: Baseline Configuration:** 1-node, 2x Intel(R) Xeon(R) Gold 6338N CPU @ 2.20GHz, 32 cores, HT On, Turbo Off, Total memory 512GB (16x32GB DDR4 3200 MT/s [2666 MT/s]), BI/OS 1.4, microcode 0xd000375, 4x Intel E810-CQDA2 (CVL, Tacoma Rapids), 1x 745.2G INTEL SSDSC2BA800G3, Ubuntu 22.04 LTS, 5.15.0-27-generic, GCC 7.5.0, DPDK 20.11, FlexCore 5G UPF (April 2021), VPP 20.09, Test by Intel as of 10/18/22. **New Configuration:** 1-node, 2x Intel(R) Xeon(R) Gold 8470N CPU, 32 cores, HT On, Turbo On, Total Memory 512GB (16x32GB DDR5 4800 MT/s [4800 MT/s]), BI/OS EGSDCRB1.SYS.0090.D03.2210040200, microcode 0x2b0000c0, 3x Intel E810-2CQDA2 (CVL, Chapman Beach, Total – 6x100G ports), 1x 223.6G INTEL SSDSC2KB240G8, 1x 745.2G INTEL SSDSC2BA800G3, Ubuntu 22.04 LTS, 5.15.0-27-generic, GCC 7.5.0, DPDK 20.11, FlexCore 5G UPF (April 2021), VPP 20.09, Test by Intel as of 10/14/22.
2. **VPP FIB: Baseline Configuration:** 1-node, 2(1 used)x Intel Xeon Gold 6338N on Wilson City with 256 GB (16 slots/ 16GB/ DDR4-3200[2666]) total memory, ucode 0xd000375, HT on, Turbo off, Ubuntu 22.04 LTS, 5.15.35, 1x INTEL SSDSC2KB240G8, x 4x Gen4 x16 PCIe NICs Intel Ethernet Controller E810-2CQDA2 (total 800Gb/s), VPP 22.06-release FIB ipv4 router , GCC 9.4, Dataset size 64B / 512B, IxNetwork 9.00.1900.17, test by Intel on 10/5/2022. **New Configuration:** 1-node, pre-production platform with 2(1 used)x Intel® Xeon® Platinum 8470N on Archer City with 512 GB (16 slots/ 32GB/ DDR5-4800) total memory, ucode 0xab0000c0, HT on, Turbo off, Ubuntu 22.04 LTS, 5.15.35, 1x INTEL SSDSC2KB240G8, x 5x Gen4 x16 PCIe NICs Intel Ethernet Controller E810-2CQDA2 (total 1000Gb/s), VPP 22.06-release FIB ipv4 router , GCC 9.4, Dataset size 64B / 512B, IxNetwork 9.00.1900.17, test by Intel on 9/30/2022.
3. **VPP IPSEC: Baseline Configuration:** 1-node, 2(1 used)x Intel Xeon Gold 6338N on Wilson City with 256 GB (16 slots/ 16GB/ DDR4-3200[2666]) total memory, ucode 0xd000375, HT on, Turbo off, Ubuntu 22.04 LTS, 5.15.35, 1x INTEL SSDSC2KB240G8, x 4x Gen4 x16 PCIe NICs Intel Ethernet Controller E810-2CQDA2 (total 800Gb/s), VPP 22.06-release 1 tunnel per VF port, GCC 9.4, Intel-ipsec-mb libraries v1.1, Dataset size 1420B, IxNetwork 9.00.1900.17, test by Intel on 10/5/2022. **New Configuration:** 1-node, pre-production platform with 2(1 used)x Intel® Xeon® Platinum 8470N on Archer City with 512 GB (16 slots/ 32GB/ DDR5-4800) total memory, ucode 0xab0000c0, HT on, Turbo off, Ubuntu 22.04 LTS, 5.15.35, 1x INTEL SSDSC2KB240G8, x 5x Gen4 x16 PCIe NICs Intel Ethernet Controller E810-2CQDA2 (total 1000Gb/s), VPP 22.06-release 1 tunnel per VF port, GCC 9.4, Intel-ipsec-mb libraries v1.1, Dataset size 1420B, IxNetwork 9.00.1900.17, test by Intel on 9/30/2022.
4. **Istio Envoy Ingress with QAT:** 8480+: 1-node, pre-production platform with 2x Intel(R) Xeon(R) Platinum 8480+ with Intel QAT on Intel ArcherCity with GB (16 slots/ 32GB/ DDR5 4800) total memory, ucode 0x2b0000a1, HT on, Turbo off, Ubuntu 22.04.1 LTS, 5.17.0-051700-generic, 1x 54.9G INTEL SSDPEK1A058GA, 1x Ethernet Controller I225-LM, 4x Ethernet Controller E810-C for QSFP, 2x Ethernet Controller XXV710 for 25GbE SFP28, Nighthawk, gcc version 11.2.0, Docker 20.10.17, Kubernetes v1.22.3, Calico 3.21.4, Istio 1.13.4. DLB SW v 7.8, qatlib is 22.07.1, Nighthawk POD's with response size: 25 PODs each with 1kB/10kB/1MB/mixed size, test by Intel on 10/27/2022.
5. **AI Inference up to 10x:** 5.7x to 10x higher PyTorch real-time inference performance on 4th Gen Intel Xeon Scalable processor with built in Intel AMX (BF16) vs. prior generation (FP32); 5.7-10x & 7x: PyTorch geomean of ResNet50, Bert-Large, MaskRCNN, SSD-ResNet34, RNN-T, Resnext101; 5.8x to 9.6x higher PyTorch batch inference performance on 4th Gen Intel Xeon Scalable processor with built in Intel AMX (BF16) vs. prior generation (FP32); 5.8-9.6x & 7x: PyTorch geomean of ResNet50, Bert-Large, MaskRCNN, SSD-ResNet34, RNN-T, Resnext101, DLRM
6. **5G AVX ISA “up to 2x capacity gains”.** The is projected claim (theoretical) as of 2/10/2022 based on Sapphire Rapids architecture improvements vs 3<sup>rd</sup> Gen Intel Xeon Scalable processors at similar core counts on a test scenario using FlexRAN software; disclosed at Intel Investor Day (February 17, 2022).
7. **CDN: Baseline Configuration:** Test by Intel as of 09/28/22. 1-node, 1x Intel® Xeon® Platinum 8380 Processor, 40 cores, HT On, Turbo On, Total Memory 256 GB (8 slots/ 32 GB/ 3200 MT/s), 8x Intel® P5800X, 2x Intel® E810-2CQDA2, BI/OS 1.4 (ucode 0xd000375), Ubuntu 22.04, kernel 5.15.0-48-generic, gcc (Ubuntu 11.2.0-19ubuntu1) 11.2.0, OpenSSL 3.0.2 15 Mar 2022 (Library: OpenSSL 3.0.2 15 Mar 2022), NGINX 1.22.0, wrk master 02/07/2021 (keep alive OR connection: close, 400 OR 4000 OR 20000 total connections) Throughput measured with 100% Transport Layer Security (TLS) traffic with 93.3% cache hit ratio. **New Configuration:** Test by Intel as of 09/28/22. 1-node, 1x Intel® Xeon® Platinum 8480+ Processor, 56 cores, HT On, Turbo On, Total Memory 256 GB (8 slots/ 32 GB/ 4800 MT/s), 8x Intel® P5800X, 2x Intel® E810-2CQDA2, BI/OS EGSDCRB1.SYS.0087.D13.2208261709 (ucode 0x2b000070), Ubuntu 22.04, kernel 5.15.0-48-generic, gcc (Ubuntu 11.2.0-19ubuntu1) 11.2.0, OpenSSL 3.0.2 15 Mar 2022 (Library: OpenSSL 3.0.2 15 Mar 2022), NGINX 1.22.0, wrk master 02/07/2021 (keep alive OR connection: close, 400 OR 4000 OR 20000 total connections) Throughput measured with 100% Transport Layer Security (TLS) traffic with 93.3% cache hit ratio.

# Sapphire Rapids Performance – Configuration/Disclaimers

8. **VoD: Baseline Configuration:** Test by Intel as of 09/28/22. 1-node, 1x Intel® Xeon® Platinum 8380 Processor, 40 cores, HT On, Turbo On, Total Memory 256 GB (8 slots/ 32 GB/ 3200 MT/s), 8x Intel® P5510, 2x Intel® E810-2CQDA2, BIOS 1.4 (ucode 0xd000375), Ubuntu 22.04, kernel 5.15.0-48-generic, gcc (Ubuntu 11.2.0-19ubuntu1) 11.2.0, OpenSSL 3.0.2 15 Mar 2022 (Library: OpenSSL 3.0.2 15 Mar 2022), NGINX 1.22.0, wrk master 02/07/2021 (keep alive OR connection: close, 400 OR 4000 OR 20000 total connections) Throughput measured with 100% Transport Layer Security (TLS) traffic with 100% cache hit ratio.**New Configuration:** Test by Intel as of 09/28/22. 1-node, 1x Intel® Xeon® Platinum 8480+ Processor, 56 cores, HT On, Turbo On, Total Memory 256 GB (8 slots/ 32 GB/ 4800 MT/s), 8x Intel® P5510, 2x Intel® E810-2CQDA2, BIOS EGSDCRBI.SYS.0087.D13.2208261709 (ucode 0x2b0000070), Ubuntu 22.04, kernel 5.15.0-48-generic, gcc (Ubuntu 11.2.0-19ubuntu1) 11.2.0, OpenSSL 3.0.2 15 Mar 2022 (Library: OpenSSL 3.0.2 15 Mar 2022), NGINX 1.22.0, wrk master 02/07/2021 (keep alive OR connection: close, 400 OR 4000 OR 20000 total connections) Throughput measured with 100% Transport Layer Security (TLS) traffic with 100% cache hit ratio.

## SECTION: PLATFORM OVERVIEW

9. **AMX (BF16):** Estimated performance on 2S Pre-production 4<sup>th</sup> Gen Intel Xeon Scalable processor (Sapphire Rapids) 56C, 350W TDP, with 1TB (8 channels/ 64GB/ 4800) total DDR5 memory, using BKC 46, using AMX/int8 and BF16, CentOS Stream 8, oneDNN optimized AMX kernels compared to 3<sup>rd</sup> Gen Intel Xeon Scalable processor (Cooper Lake), 28C, 250W (8380H); Due to linear scaling across sockets on inference performance, 2S socket data obtained by applying 0.25x factor on 8S measurement; Configuration: 1-node, 8x 3rd Gen Intel® Xeon® Platinum 8380H processor (28C, 250W) on Intel Reference Platform (Cooper City) with 384 GB (48 slots / 64GB / 2933) total memory, ucode 0x7002302, HT on, Turbo on, with Ubuntu 20.04 LTS, Linux 5.4.0-29-generic, Intel SSD 800GB OS Drive; measurements may vary. Test by Intel on 1/27/2022.

9a Object Detection (RT): With SSD-RN34, BS=1, 56, BF16; internal Intel optimized TensorFlow 2.8

9b : Recommendation System: With DLRM, BS=128, BF16; internal Intel optimized PyTorch v1.10

9c : Natural Language Processing: With BERT-Large, BS=1, 16, BF16; internal Intel optimized TensorFlow 2.8, Squad 1.1 dataset.

9d : Image Recognition: With ResNet-50 v1.5, BS=1, 116, BF16; internal Intel optimized TensorFlow 2.8

10. **IAA:** Estimated performance comparing 4th Gen Intel® Xeon® Scalable processor configuration with Intel® IAA enabled, versus same processor running software on CPU cores without IAA offload. Configuration: 1-node, 2 sockets (1 socket tested) 4th Gen Intel Xeon Scalable processor (56-cores, 4x IAA devices) pre-production platform with 512GB (16x32GB 4800MT/s [4800MT/s]) total memory, HT on, Turbo on, internal pre-production BIOS 0x8e000260, CentOS Linux 8.4.2105, 5.15.0-spr.bkc.pc.3.21.0.x86\_64, internal RocksDB v7.1 with pluggable compression support (db\_bench, read only). Software Configuration GCC 8.5.0, ZSTD v1.4.4, p99 latency used. Results depend on block size and database entry size. Read-Only results (Relative ops/s vs data size – 16kB block, 32B value) and Read-Write results (Relative ops/s vs data size – 16kB block, 32B value). Tradeoff up to 16% compressed data size. Test by Intel as of 4/25/2022.

11. **QAT :** Estimated performance comparing 4th Gen Intel Xeon Scalable processor configuration with Intel® QAT enabled, versus same processor without QAT offload, with software optimizations. Configuration: 1-node, 2-sockets (1 socket tested, SPR-E3) 4th Gen Intel Xeon Scalable processor, 52C, 300W TDP with 1, 2 and 4 Intel QAT active devices (with 4 cores/8 threads). Pre-production platform with 512GB (16x32GB 4800MT/s [4800MT/s]) total memory, HT on, Turbo off, internal pre-production BIOS 0x890000a0, Ubuntu 22.04 LTS, 5.15.0-27-generic; Workload: Async NGINX 0.4.7; NGINX TLS 1.3 Webserver with ECDHE-X448-RSA4k algorithm. Software Configuration: GCC 11.2.0, libraries: OpenSSL 1.1.1o, QAT engine v0.6.12, Intel IPsec MB v1.2, IPP Crypto ipccp\_2021.5. Test by Intel as of 6/30/2022.

12. **DSA:** Estimated performance on pre-production configuration: 1-node, 2x 4th Gen Intel Xeon Scalable processor, XCC 56C with 4 DSA devices – tests done on 1 device (up to 2.2x) and 4 devices (up to 8.9x), Archer City pre-production system; 1024 GB (16x64GB DDR5-4800MT/S) total DDR5 memory, HT on, Turbo on, ucode 0x8e000140, Red Hat Enterprise Linux 8.5 (Ootpa), 5.12.0-512MPS-intel-next.sprd0po.05242021, workload: DSA\_microbenchmarks, gcc version 8.5.0 20210514, 1x DSA memory to memory 32K size: 31 GB/s, 4x DSA memory to memory 32K size: 125 GB/s, 2x DSA memory to Gen5 x16 MMIO 32K size: 36 GB/s, 1x DSA memory to Gen4 x16 MMIO 32K size: 22 GB/s, test by Intel on 2/14/2022; versus Baseline platform 1-node, 2x 3rd Gen Intel Xeon processor Platinum 8380 on Wilson City with 512 GB (16x32GB 3200MT/s) total DDR4 memory, ucode 0xd000331, HT on, Turbo off, Red Hat Enterprise Linux 8.1 (Ootpa), 5.4.2-Perf, DMA tool, gcc version 8.3.1 20190507, memory to memory 32K size: 14 GB/s, memory to Gen4 x16 MMIO 32K size: 14 GB/s, memory to Gen4 x16 MMIO 32K size: 14 GB/s, test by Intel on 2/8/2022.

# Sapphire Rapids Performance – Configuration/Disclaimers

13. **HBM:** 2.7x Manufacturing Performance (OpenFOAM 28M\_cell\_motorbike @250 iterations) – Sapphire Rapids + High Bandwidth Memory vs. Ice Lake: New: Sapphire Rapids: 1-node, 2x 56c Intel Xeon 512 GB (16 slots/ 32GB/ 4800) total DDR5, ucode 0x8f000050, Red Hat Enterprise Linux 8.4, Kernel 4.18. App Version: v8; Build notes: Tools: Intel FORTRAN Compiler 2022.0, Intel C Compiler 2022.0, Intel MPI 2021.5; threads/core: 1; Turbo: used; Build knobs: -O3 -ip -xCORE-AVX512. OpenFOAM® v1912.; Build notes: Motorbike 28M @ 250 iterations; Build notes: Tools: Intel Parallel Studio 2020u4, Build knobs: -O3 -ip -xCORE-AVX512. Baseline: 1-node, 2x Intel Xeon Platinum 8380 512 GB (16 slots/ 32GB/ 3200) total DDR4, ucode 0xd000270, Rocky Linux 8.5 Kernel 4.18. App Version: v8; Build notes: Tools: Intel FORTRAN Compiler 2021.2, Intel C Compiler 2021.2, Intel MPI 2021.2; threads/core: 1; Turbo: used; Build knobs: -O3 -ip -xCORE-AVX512. Data collected: 2/2022.
14. **AMX (BF16)** : Projected performance on 1S Pre-Production 4<sup>th</sup> Gen Intel Xeon Scalable processor (Sapphire Rapids) compared to 3rd Gen Intel Xeon Scalable processor (Cooper Lake) on ResNet-50 v1.5 for Deep Learning Training
15. **QAT:** Estimated performance comparing 4<sup>th</sup> Gen Intel® Xeon® Scalable processor (60C, 350W TDP) with 4 Intel® QAT devices enabled (level 9), versus same processor running Zstd (SW compression level 7) on CPU cores without QAT offload. The approximate SW levels tested on will get to the same compression ratio approximately. Configuration: 1-node, 2-sockets, 4<sup>th</sup> Gen intel Xeon Scalable processor (56-cores, 4xQAT devices) pre-production platform with 256GB (16x16GB 4800MT/s [4800MT/s]) total memory, HT on, Turbo on, internal pre-production BIOS 0x8e0001e0, CentOS Stream 8, 5.15.0-spr.bkc.pc.2.10.0.x86\_64, ZSTD-QAT\_REL\_0.1.1\_001; Software Configuration GCC 8.4.1, QAT driver: QAT20.L.0.8.0-00071; Workload: Industry standard Silesia corpus file; Test by Intel as of 3/8/2022.
16. **DSA:** Estimated performance on pre-production configuration: 1-node, 2x 4th Gen Intel Xeon Scalable processor (XCC 56C with 4 DSA devices) Archer City pre-production system; CPU: SPR EO; MEMORY: 256GB (8x32GB 4800MT/s) - Dual Rank total DDR5 memory; HT on, Turbo off, ucode 0x8e0001c0, Ubuntu 20.04.3 LTS, 5.15.0-18-generic; Workload: Open vSwitch ver2.16.90 (+patches to enable DSA, internal branch: spaig\_dsa\_22\_03 - e281f85ef); gcc version (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0; DPDK ver: 22.03-rc0 (+patches to enable DSA, internal branch: spaig\_vhost\_dsa\_for\_dpdk\_22\_03 - 4508f526be); VM OS (Kernel): Ubuntu 20.04.3 LTS (5.4.0-99-generic); NIC: 2x Chaman Beach CVL cards (total 400 GbE set up) Each Chapman Beach has 2x100 GbE link; DPDK testpmd application is used in each VM for packet processing. Test cases: VM with Checksum and VM with MAC processing. Test by Intel on 2/16/22.
17. **DLB (Average Latency at 40K connections):** Estimated performance comparing 4th Gen Intel Xeon Scalable processor configuration with Intel® DLB enabled versus same processor running the workloads on cores without DLB offload for Webserver use cases. Configuration: 1-node, 2-sockets (1-socket tested) 4th Gen Intel Xeon Scalable processor, 56C, 350W TDP with 4 DLB devices. Pre-production platform with 128GB (4x32GB 4800MT/s [4800MT/s]) DDR5 total memory, HT on, Turbo on, internal pre-production BIOS 0x8d0003f0, CentOS Linux 8, 5.12.0-0507.intel\_next.10\_26\_po.49.x86\_64+server; Workload: NGINX Load Balancer (NGINX 1.16.1); Software Configuration: GCC: 8.5.0 20210514 (Red Hat 8.5.0-4), BKC #47, DLB driver: RELEASE\_VER\_7.4.0-V1. Test by Intel as of April 2022.
18. **Integer & Floating-Point Throughput, HP LINPACK and STREAM Triad (pg-39):** 4<sup>th</sup> Gen Intel® Xeon® Scalable processor (Sapphire Rapids) Performance optimized SKUs – 56C 350W and 32C, 300W versus 3<sup>rd</sup> Gen Intel® Xeon® Scalable processor (Ice Lake) – 40C, 270W (8380) and 32C, 265W TDP respectively. Performance projections for Integer & Floating-Point throughput data on ICC 18 u1 (pre silicon) with SPR 56C (E2) & 32C (E2), 1-node, 1-socket pre-production projections versus 3<sup>rd</sup> Gen Intel® Xeon® Scalable processor (ICX). STREAM data : 4<sup>th</sup> Gen Intel Xeon Scalable processor, 56C (E2): 1-node, 2x pre-production Sapphire Rapids on Archer City with 1TB GB (16 slots/ 64GB/ 4800 MT/s) total DDR5 memory, ucode 0xf0002e1, HT OFF, Turbo ON, Ubuntu 22.04 Jammy Jellyfish (development branch), 5.15.0-23-generic, 1x 1TB Crucial MX500 SSD, STREAM v5.10, ICC 2022.1, test by Intel on 4/4/2022. Intel Xeon Platinum 8380: 1-node, 2x Intel(R) Xeon(R) Platinum 8380 CPU @ 2.30GHz on Coyote Pass (Whitley) with 1TB GB (16 slots/ 64GB/ 3200 MT/s) total DDR4 memory, ucode 0xd0002b1, HT OFF, Turbo ON, Ubuntu 22.04 Jammy Jellyfish (development branch), 5.15.0-22-generic, 1x 1.92TB Intel S4610 SSD, STREAM v5.10, ICC 2022.1, test by Intel on 3/25/2022.

# Sapphire Rapids Performance – Configuration/Disclaimers

## SECTION : Network Optimized N skus

19. a. **Secure Gateway Baseline Configuration:** 1-node, 2(1 used)x Intel Xeon Gold 6338N on Wilson City with 256 GB (16 slots/ 16GB/ DDR4-3200[2666]) total memory, ucode 0xd000375, HT on, Turbo off, Ubuntu 22.04 LTS, 5.15.35, 1x INTEL SSDSC2KB240G8, x 4x Gen4 x16 PCIe NICs Intel Ethernet Controller E810-2CQDA2 (total 800Gb/s), VPP 22.06-release 1 tunnel per VF port, GCC 9.4, Intel-ipsec-mb libraries v1.1, Dataset size 1420B, lxFNetwork 9.00.1900.17, test by Intel on 10/5/2022. **New Configuration:** 1-node, pre-production platform with 2(1 used)x Intel® Xeon® Platinum 8470N on Archer City with 512 GB (16 slots/ 32GB/ DDR5-4800) total memory, ucode 0xab0000c0, HT on, Turbo off, Ubuntu 22.04 LTS, 5.15.35, 1x INTEL SSDSC2KB240G8, x 5x Gen4 x16 PCIe NICs Intel Ethernet Controller E810-2CQDA2 (total 1000Gb/s), VPP 22.06-release 1 tunnel per VF port, GCC 9.4, Intel-ipsec-mb libraries v1.1, Dataset size 1420B, lxFNetwork 9.00.1900.17, test by Intel on 9/30/2022.
- b. **Next Generation Firewall Baseline Configuration:** 3rd Gen Gold 6338N: Test by Intel as of 11/2022. 1-node, 2x Intel(R) Xeon(R) Gold 6338N CPU @ 2.20GHz on Supermicro X12DPG-QT6, 32 cores, HT On, Turbo Off/On, Total Memory 512GB (16x32GB DDR4 3200 MT/s [2666 MT/s]), BIOS 1.4, microcode 0xd000375, 1x Intel Ethernet Controller E810-CQDA2, 1x 223.6G INTEL SSDSC2BW240H6, 1x 240M Disk, Ubuntu 22.04 LTS, 5.15.35, GCC 9.4, NGFW22.09-1, VPP : v22.06.0-16, Snort:3.1.36.0, DAQ: 3.0.9, LuaJIT: 2.1.0-beta3, OpenSSL: 1.1.1f 31 Mar 2020, Libpcap: 1.10.1(with TPACKET\_V3), PCRE:8.45 2021-06-15, ZLIB: 1.2.11, Hyperscan: 5.4.0 2021-01-26, LZMA: 5.2.5 **New Configuration:** 4th Gen Platinum 8470N: Test by Intel as of 11/2022. 1-node, pre-production platform with 2(1 used)x Intel(R) Xeon(R) Platinum 8470N on Intel Corporation M50FCP, 52 cores, HT On, Turbo Off/On, Total Memory 512GB (16x32GB DDR5 4800 MT/s [4800 MT/s]), microcode 0x2b000310, 1x Intel Ethernet Controller E810-CQDA2, 1x 223.6G INTEL SSDSC2KB240G8, 1x 240M Disk, Ubuntu 22.04 LTS, 5.15.0-27-generic, GCC 11.3, NGFW22.09-1, VPP : v22.06.0-16, Snort:3.1.36.0, DAQ: 3.0.9, LuaJIT: 2.1.0-beta3, OpenSSL: 1.1.1f 31 Mar 2020, Libpcap: 1.10.1(with TPACKET\_V3), PCRE:8.45 2021-06-15, ZLIB: 1.2.11, Hyperscan: 5.4.0 2021-01-26, LZMA: 5.2.5
20. **Web Servicing Connections for Intel® QuickAssist Technology. NGINX Webserver Handshake Only TLS 1.3 ECDHE-X25519-RSA2K (24C48T) Baseline Configuration:** 1-node, 2(1 used)x Intel(R) Xeon(R) Gold 6338N CPU @ 2.20GHz, 32 (24C48T used) cores on Supermicro SYS-740GP-TNRT, HT On, Turbo Off, Total Memory 256GB (16x16GB DDR4 3200 MT/s [2666 MT/s]), BI/OS 1.4, microcode 0xd000375, 4x Ethernet Controller E810-C for QSFP, 2x Ethernet Controller 10G X550T, 1x 223.6G INTEL SSDSC2KB240G8, Ubuntu 22.04 LTS, 5.15.0-27-generic, gcc (Ubuntu 11.2.0-19ubuntul) 11.2.0, LBG 62X Chipset (3 QAT), NGINX (async mode nginx 0.4.7), GCC 11.2.0, openssl 1.1.1m, qatengine v0.6.14 (Optimized SW and QAT HW), IPsecmb v1.2 (Optimized SW), IPP-Crypto ipp-crypto\_2021\_5 (Optimized SW), QAT Driver (CPM 1.7): QAT.L.4.18.1-00001, test by Intel on 09/19/2022. **New Configuration:** Test by Intel as of Sep 19 10:18:38. 1-node, pre-production platform 2(1 used)x Intel(R) Xeon(R) Gold 6428N, 32 (24C48T used) cores, HT On, Turbo Off, Total Memory 512GB (16x32GB 4800 MT/s [4000 MT/s]), BI/OS EGSDCRB1.86B.8612.P03.2208120625, microcode 0xab000060, 1x Ethernet Controller I225-LM, 6x Ethernet Controller E810-C for QSFP, 1x Ethernet interface, 1x 223.6G INTEL SSDSC2BB240G4 Ubuntu 22.04 LTS, 5.15.0-27-generic, gcc (Ubuntu 11.2.0-19ubuntul) 11.2.0. NGINX (async mode nginx 0.4.7), GCC 11.2.0, openssl 1.1.1m, qatengine v0.6.14 (Optimized SW and QAT HW), IPsecmb v1.2 (Optimized SW), IPP-Crypto ipp-crypto\_2021\_5 (Optimized SW), QAT driver QAT.20.L.0.9.5 (2 QAT HW)
21. **CDN: Baseline Configuration:** Test by Intel as of 09/28/22. 1-node, 1x Intel® Xeon® Platinum 8380 Processor, 40 cores, HT On, Turbo On, Total Memory 256 GB (8 slots/ 32 GB/ 3200 MT/s), 8x Intel® P5800X, 2x Intel® E810-2CQDA2, BI/OS 1.4 (ucode 0xd000375), Ubuntu 22.04, kernel 5.15.0-48-generic, gcc (Ubuntu 11.2.0-19ubuntul) 11.2.0, OpenSSL 3.0.2 15 Mar 2022 (Library: OpenSSL 3.0.2 15 Mar 2022), NGINX 1.22.0, wrk master 02/07/2021 (keep alive OR connection: close, 400 OR 4000 OR 20000 total connections) Throughput measured with 100% Transport Layer Security (TLS) traffic with 93.3% cache hit ratio. **New Configuration:** Test by Intel as of 09/28/22. 1-node, pre-production platform with 1x Intel® Xeon® Platinum 8480+ Processor, 56 cores, HT On, Turbo On, Total Memory 256 GB (8 slots/ 32 GB/ 4800 MT/s), 8x Intel® P5800X, 2x Intel® E810-2CQDA2, BI/OS EGSDCRB1.SYS.0087.D13.2208261709 (ucode 0x2b000070), Ubuntu 22.04, kernel 5.15.0-48-generic, gcc (Ubuntu 11.2.0-19ubuntul) 11.2.0, OpenSSL 3.0.2 15 Mar 2022 (Library: OpenSSL 3.0.2 15 Mar 2022), NGINX 1.22.0, wrk master 02/07/2021 (keep alive OR connection: close, 400 OR 4000 OR 20000 total connections) Throughput measured with 100% Transport Layer Security (TLS) traffic with 93.3% cache hit ratio.
22. **5G CORE/UPF: Baseline Configuration:** 6338N: Test by Intel as of 10/18/22. 1-node, 2(1 used)x Intel(R) Xeon(R) Gold 6338N CPU @ 2.20GHz, 32 cores on Supermicro X12DPG-QT6, HT On, Turbo Off, Total memory 512GB (16x32GB DDR4 3200 MT/s [2666 MT/s]), BI/OS 1.4, microcode 0xd000375, 4x Intel E810-CQDA2 (CVL, Tacoma Rapids), 1x 745.2G INTEL SSDSC2BA800G3, Ubuntu 22.04 LTS, 5.15.0-27-generic, GCC 7.5.0, DPDK 20.11, FlexCore 5G UPF (April 2021), VPP 20.09 **New Configuration:** 8470N: Test by Intel as of 10/14/22. 1-node, pre-production platform 2(1 used)x Intel(R) Xeon(R) Gold 8470N CPU, 32 cores, HT On, Turbo Off/On, Total Memory 512GB (16x32GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1.SYS.0090.D03.2210040200, microcode 0x2b0000c0, 3x Intel E810-2CQDA2 (CVL, Chapman Beach, Total – 6x100G ports), 1x 223.6G INTEL SSDSC2KB240G8, 1x 745.2G INTEL SSDSC2BA800G3, Ubuntu 22.04 LTS, 5.15.0-27-generic, GCC 7.5.0, DPDK 20.11, FlexCore 5G UPF (April 2021), VPP 20.09

# Sapphire Rapids Performance – Configuration/Disclaimers

## SECTION : Network Optimized N skus

23. **Packet Processing: Baseline Configuration:** 1-node, 2(1 used)x Intel Xeon Gold 6338N on Wilson City with 256 GB (16 slots/ 16GB/ DDR4-3200[2666]) total memory, ucode 0xd000375, HT on, Turbo off, Ubuntu 22.04 LTS, 5.15.35, 1x INTEL SSDSC2KB240G8, x 4x Gen4 x16 PCIe NICs Intel Ethernet Controller E810-2CQDA2 (total 800Gb/s), VPP 22.06-release FIB ipv4 router , GCC 9.4, Dataset size 64B / 512B, IxNetwork 9.00.1900.17, test by Intel on 10/5/2022: **New Configuration:** 1-node, pre-production platform with 2(1 used)x Intel® Xeon® Platinum 8470N on Archer City with 512 GB (16 slots/ 32GB/ DDR5-4800) total memory, ucode Oxab0000c0, HT on, Turbo off, Ubuntu 22.04 LTS, 5.15.35, 1x INTEL SSDSC2KB240G8, x 5x Gen4 x16 PCIe NICs Intel Ethernet Controller E810-2CQDA2 (total 1000Gb/s), VPP 22.06-release FIB ipv4 router , GCC 9.4, Dataset size 64B / 512B, IxNetwork 9.00.1900.17, test by Intel on 9/30/2022.
24. **VCMTS: Baseline Configuration:** 1-node, 2(1 used)x Intel Xeon Gold 6338N on SuperMicro X12DPG-QT6 with 512 GB (16 slots/ 32GB/ DDR4-3200[2666]) total memory, ucode 0xd000375, HT on, Turbo on, Ubuntu 22.04 LTS, 5.15.0-27-generic, 1x INTEL SSDSC2KB240G8, 1x Ethernet Controller E810-C for QSFP, vCMTS 22.10 beta, DPDK 22.03, GCC 11.2.0, DPDK 22.03, Collectd 5.12.0, Grafana 8.5.3, Prometheus 2.0.0, test by Intel on 10/10/2022. **New Configuration:** 1-node, pre-production platform with 2(1 used)x Intel® Xeon® Platinum 8470N on Archer City with 512 GB (16 slots/ 32GB/ DDR5-4800[4800]) total memory, ucode Oxab000080, HT on, Turbo on, Ubuntu 22.04 LTS, 5.15.0-27-generic, 1x INTEL SSDSC2KB240G8, 1x Ethernet Controller E810-C for QSFP, vCMTS 22.10 beta, DPDK 22.03, GCC 11.2.0, DPDK 22.03, Collectd 5.12.0, Grafana 8.5.3, Prometheus 2.0.0, test by Intel on 9/20/2022.
25. **Malware Detection: Baseline Configuration:** Test by Intel as of 11/25/2022. 1-node, 2x Intel® Xeon® Platinum 8380, 40 cores, HT Off, Turbo On, Total Memory 512 GB DDR4 (16 slots/ 32 GB/ 3200 MHz [run @ 3200 MHz]), ucode 0xd000375, OS CentOS Stream 8, kernel 5.16.0-rc1-intel-next-00543-g5867b0a2a125, gcc (GCC) 8.5.0 20210514 (Red Hat 8.5.0-13); Python 3.9.13, Network AI Malware Detection Throughput: Intel-optimized-MalConv-for-TF: input size=1 MB, embed.=4, filter = 32, kernel = 2500, stride = 600, Tensorflow 2.10.0, oneDNN 2.6.1, VNNI INT8, 28 instances. **New Configuration:** Test by Intel as of 11/25/2022. 1-node, pre-production platform with 2x Intel® Xeon® Platinum 8481C, 56 cores, HT Off, Turbo On, Total Memory 512 GB DDR5 (16 slots/ 32 GB/ 4800 MHz [run @ 4800 MHz]), ucode 0x2b000041, OS CentOS Stream 8, kernel 5.16.0-rc1-intel-next-00543-g5867b0a2a125, gcc (GCC) 8.5.0 20210514 (Red Hat 8.5.0-13); Python 3.9.13, Network AI Malware Detection Throughput: Intel-optimized-MalConv-for-TF: input size=1 MB, embed.=4, filter = 32, kernel = 2500, stride = 600, Tensorflow 2.10.0, oneDNN 2.6.1, AMX INT8, Neural Network optimized and quantized (INT8) by Intel based on Ember MalConv open-source model, 56 instances

## SECTION : Edge (IoT) SKUs

26. **General Purpose Compute:** "New: 1-node, 2x Intel(R) Xeon(R) Gold 6448Y on ArcherCity with 512 GB (16 slots/ 32GB/ 4800) total DDR5 memory, ucode 0x2b000111, HT ON, Turbo ON, CentOS Stream 8, 5.15.0-spr.bkc.pc.12.7.15.x86\_64, 1x Samsung SSD 870 1TB, SpecCPU 2017 (n-copy) (est.), ic2022.1, test by Intel on Nov 25 2022.Baseline: 1-node, 2x Intel(R) Xeon(R) Gold 6348 CPU on WilsonCity with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xd000375, HT ON, Turbo ON, CentOS Stream 8, 5.15.0-spr.bkc.pc.12.7.15.x86\_64, 1x Crucial MX500 1TB, SpecCPU 2017 (n-copy) (est.), ic2022.1, test by Intel on Nov 23 2022."
27. **Video Analytics: Baseline Configuration:** Tested by Intel as of 11/30/22. 2-socket Intel® Xeon® Gold 6348 SRKHP, Up to 3.5GHz, 235 watts TDP, 28 cores, SMT On, Turbo On, Total Memory 512GB (16x32GB DDR4 3200 MT/s [3200 MT/s]), BI/OS WLYDCRB1.SYS.0029.P30.2209011945, microcode 0xd00037b, 1x I210 Gigabit Network Connection, 2x Ethernet Controller X710 for 10GBASE-T, 2x 931.5G ST1000NX0423, 2x 931.5G ST1000LM048-2E71, 1x 931.5G INTEL SSDPELKX010T8, Red Hat Enterprise Linux 8.6, 4.18.0-372.9.1.el8.x86\_64, 5 threads given to Detection, 2 threads given to each Classification, SW Used: OpenVINO 2022.2.0, OpenCV 4.6.0, gStreamer 1.21.1.1, DL Streamer 1.7.0.0, FFMPEG 5.1-1, Intel OneAPI 2021.7.0, libx264 1.20.2, NNs Used: OpenVINO 2022.2.0 resnet-50-tf, yolov3-tf, MobileNet\_V2, INT8, BS=1 **New Configuration:** Tested by Intel as of 11/26/22. pre-production platform with 2-socket Intel® Xeon® Gold 6448Y Q27M, Up to 4.1GHz, 225W TDP, 32 cores, SMT On, Turbo On, Total Memory 512GB (16x32GB 4800 MT/s [4800 MT/s]), BI/OS EGSDCRB1.SYS.9207.P03.2211041115, microcode 0x2b000111, 1x Ethernet Controller I225-LM, 2x Ethernet Controller X710 for 10GbE SFP+, 1x 114.6G SanDisk 3.2Gen1, 1x 931.5G Samsung SSD 870, 3x 931.5G WD Green 2.5 100, 1x 465.8G Samsung SSD 970 EVO Plus 500GB, Red Hat Enterprise Linux 8.6, 4.18.0-372.9.1.el8.x86\_64, 2 threads given to Detection, 1 thread given to each Classification, SW Used: Red Hat Enterprise Linux 8.6, kernel 4.18.0-372.9.1.el8.x86\_64, OpenVINO 2022.2.0, OpenCV 4.6.0, gStreamer 1.21.1.1, DL Streamer 1.7.0.0, FFMPEG 5.1-1, Intel OneAPI 2021.7.0, libx264 1.20.2. NNs Used: OpenVINO 2022.2.0 resnet-50-tf, yolov3-tf, MobileNet\_V2, INT8, BS=1

# Sapphire Rapids Performance – Configuration/Disclaimers

## SECTION: Edge (IoT) SKUs

28. **Object Detection & Image Classification with OpenVINO:** **Baseline Configuration:** 1-node, 2x Intel(R) Xeon(R) Gold 6348 CPU on WilsonCity with 512 GB (16 slots/ 32GB/ 3200 MT/s) total DDR4 memory, ucode 0xd000375, HT ON, Turbo ON, CentOS Stream 8, 5.15.0-spr.bkc.pc.12.7.15.x86\_64, 1x Crucial MX500 1TB, OpenVINO 2022.3, library Build: 2022.3.0-8831-4f0b846dla5, SSD-ResNet 34-1200 (INT8), Resnet-50-TF, test by Intel on Fri Dec 2, 2022. **New Configuration:** 1-node, pre-production platform 2x Intel(R) Xeon(R) Gold 6448Y ArcherCity with 512 GB (16 slots/ 32GB/ 4800) total DDR4 memory, ucode 0x2b000111, HT ON, Turbo ON, CentOS Stream 8, 5.15.0-spr.bkc.pc.12.7.15.x86\_64, 1x Samsung SSD 870 1TB, OpenVINO 2022.3, library Build: 2022.3.0-8831-4f0b846dla5, SSD-ResNet 34-1200 (INT8), Resnet-50-TF, test by Intel on Mon Dec 5, 2022

29. **Self Checkout: Baseline Configuration:** Tested by Intel as of 11/30/2022. 2-socket Intel® Xeon® Gold 6348 SRKHP, Up to 3.5GHz, 235 watts TDP, 28 cores, SMT On, Turbo On, Total Memory 512GB (16x32GB DDR4 3200 MT/s [3200 MT/s]), BIOS WLYDCRB1.SYS.0029.P30.2209011945, microcode 0xd00037b, 1x I210 Gigabit Network Connection, 2x Ethernet Controller X710 for 10GBASE-T, 2x 931.5G ST1000NX0423, 2x 931.5G ST1000LM048-2E71, 1x 931.5G INTEL SSDPELKKX010T8, Red Hat Enterprise Linux 8.6, 4.18.0-372.9.1.el8.x86\_64, 5 threads given to Detection, 2 threads given to each Classification. **New Configuration:** Tested by Intel as of 11/26/2022. 2-socket Intel® Xeon® Gold 6448Y Q27M, Up to 4.1GHz, 225W TDP, 32 cores, SMT On, Turbo On, Total Memory 512GB (16x32GB 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1.SYS.9207.P03.2211041115, microcode 0x2b000111, 1x Ethernet Controller I225-LM, 2x Ethernet Controller X710 for 10GbE SFP+, 1x 114.6G SanDisk 3.2Gen1, 1x 931.5G Samsung SSD 870, 3x 931.5G WD Green 2.5 100, 1x 465.8G Samsung SSD 970 EVO Plus 500GB, Red Hat Enterprise Linux 8.6, 4.18.0-372.9.1.el8.x86\_64, 2 threads given to Detection, 1 thread given to each Classification. **SW used:** Red Hat Enterprise Linux 8.6, kernel 4.18.0-372.9.1.el8.x86\_64, OpenVINO 2022.2.0, OpenCV 4.6.0, gStreamer 1.21.1.1, DL Streamer 1.7.0.0, FFMPEG 5.1-1, Intel OneAPI 2021.7.0, libx264 1.20.2. **NNs Used:** OpenVINO 2022.2.0 resnet-50-tf, yolov3-tf, MobileNet\_V2, INT8, BS=1

## SECTION: Eagle Stream platform overview – energy efficiency

30. **Up to 14.21x and 13.53x higher performance/W** using 4th Gen Intel Xeon Scalable w/Advanced Matrix Extensions using AMX vs VNNI instructions on SSD-ResNet34 on Object Detection -- 1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable processor (60 core) with Intel® Advanced Matrix Extensions (Intel AMX), Intel platform with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000a1, HT On, Turbo On, SNC Off, CentOS Stream 8, 5.19.16-301.fc37.x86\_64, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Intel TF 2.10, AI Model=SSD-ResNet34, best scores achieved: BS1 FP32 60 cores/instance (max. 100ms SLA), BS1 INT8 4 cores/instance (max. 100ms SLA), BS1 AMX 4 core/instance (max. 100ms SLA), BS8 FP32 8 cores/instance, BS2 INT8 1 cores/instance, BS2 AMX 1 cores/instance, using physical cores, tested by Intel November 2022

## SECTION: Intel Infrastructure Power Manager

31. **Up to 30% IPM Power Savings & 93% Perf/Watt with IPM (New):** Tested by Intel as of 01/26/23. 1-node, 2x Intel(R) Xeon(R) Gold 6438N CPU, 32 cores, HT On, Turbo Off, Total Memory 512GB (16x32GB DDR5 4800 MT/s [4000 MT/s]), BIOS EGSDCRB1.SYS.0090.D03.2210040200, microcode 0x2b0000c0, 2x Intel E810-2CQDA2 (CVL, Chapman Beach, Total – 4x100G ports), 1x 223.6G INTEL SSDSC2KB240G8, 1x 745.2G INTEL SSDSC2BA800G3, Ubuntu 22.04 LTS, 5.15.0-27-generic, GCC 7.5.0, DPDK 22.11