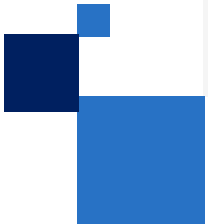May 2024

Data Center Modernization Enablement Package

How our valued partners can address customers' business challenges with Intel based solutions

Presentation Notes

Presentation Name	Data Center Modernization Enablement Package
Abstract	This package has been developed with public Intel content intended to be used by PSAMs and Intel partners to guide in depth conversations with Partner Reps. The focus is centered around Data Center Modernization to align with SSP VX Top Strategic Priorities and should be used for Marketing and Scaling. This package has been customized to enable conversations with partners and the content tailored to a Tech C-Suite and Procurement audience. This is an ongoing project. New versions will be released and the packages will be refreshed on a regular basis to incorporate the latest Intel content. The Intel Focus will be broadened and there will be further customized packages available soon. **Note:** If your partner is not already an IPA member, please include slides 69 and 70 from the backup section to encourage your partner to join IPA
Target Audience	Partner Ecosystem, End Customers, Sales
What's New	• Microsoft Data Center Products Refresh Infographic – slides 5 and 10 • 5th Gen Xeon modernization and benchmarking – slides 20, 21 and 26 • New Cloud TV episodes (Modernization opportunities with Microsoft and 5th Gen Xeon overview) – slide 48
Coming Next	• GEO centric packages • Partner tailored packages

CALL TO ACTION	1. Use this Enablement Package in conversations with your partners 2. Share the public version with your partners – access at www.intel.com/salesenablement 3. If your partner is not a member of Intel Partner Alliance, encourage them to join 4. Provide feedback to Amy Kircos (amy.kircos@intel.com) on any proposed customizations needed for your region or partner to influence upcoming packages

intel.

Microsoft Data Center Products Refresh on 4th Generation Intel® Xeon® Scalable Processors

Windows Server 2012

Microsoft ended support for Windows Server 2012 in October 2023

Why Upgrade Your Hardware Now?

Upgrading to Windows Server 2022 on older, out of warranty hardware will not allow you to use all the features of Windows Server

Upgrading to Intel® 4th Generation Xeon® processors unlocks the full value of Windows Server 2022 for enhanced TCO and Security

LEARN HOW

Support has ended for Microsoft Windows Server 2012 and SQL Server 2012

Transform your customers' infrastructure with 4th Gen Intel® Xeon® Scalable processors + Microsoft Windows 2022 and SQL Server 2022

Enable new features, improve performance, lower cost, and reduce risk

4th Gen Intel® Xeon® Scalable processors

- 1.53x average performance gain[1]
- Up to 80% reduction in server count[2]
- Up to 75% reduction in TCO[2]

- Up to 3.2x backup time improvement[3]
- Up to 22% more NOPM transactions[4]
- Up to 19% faster query response time[4]

when using SQL Server 2022

+ Windows Server 2022
- Additional performance and security enhancements
- Enhanced container support
- Customizable hybrid cloud capabilities with Azure
- Increased scalability (48 TB memory; 2048 logical cores across 64 sockets)

+ SQL Server 2022
- Reduced workload times
- Seamless analytics
- Unified data governance and management
- More secure database
- High availability, business continuity
- Automatic resolution of conflicts

Modernize now to help businesses address cybersecurity risks

88% of corporations see cybersecurity as a business priority[5]

63% of organizations suffered a cyber breach[6]

Intel® Security Engines, built into 4th Gen Intel® Xeon® Scalable processors, deliver enhanced security that allows for even the most sensitive data to be available for new modernization opportunities such as AI analysis, training, or processing—all while remaining private and confidential.
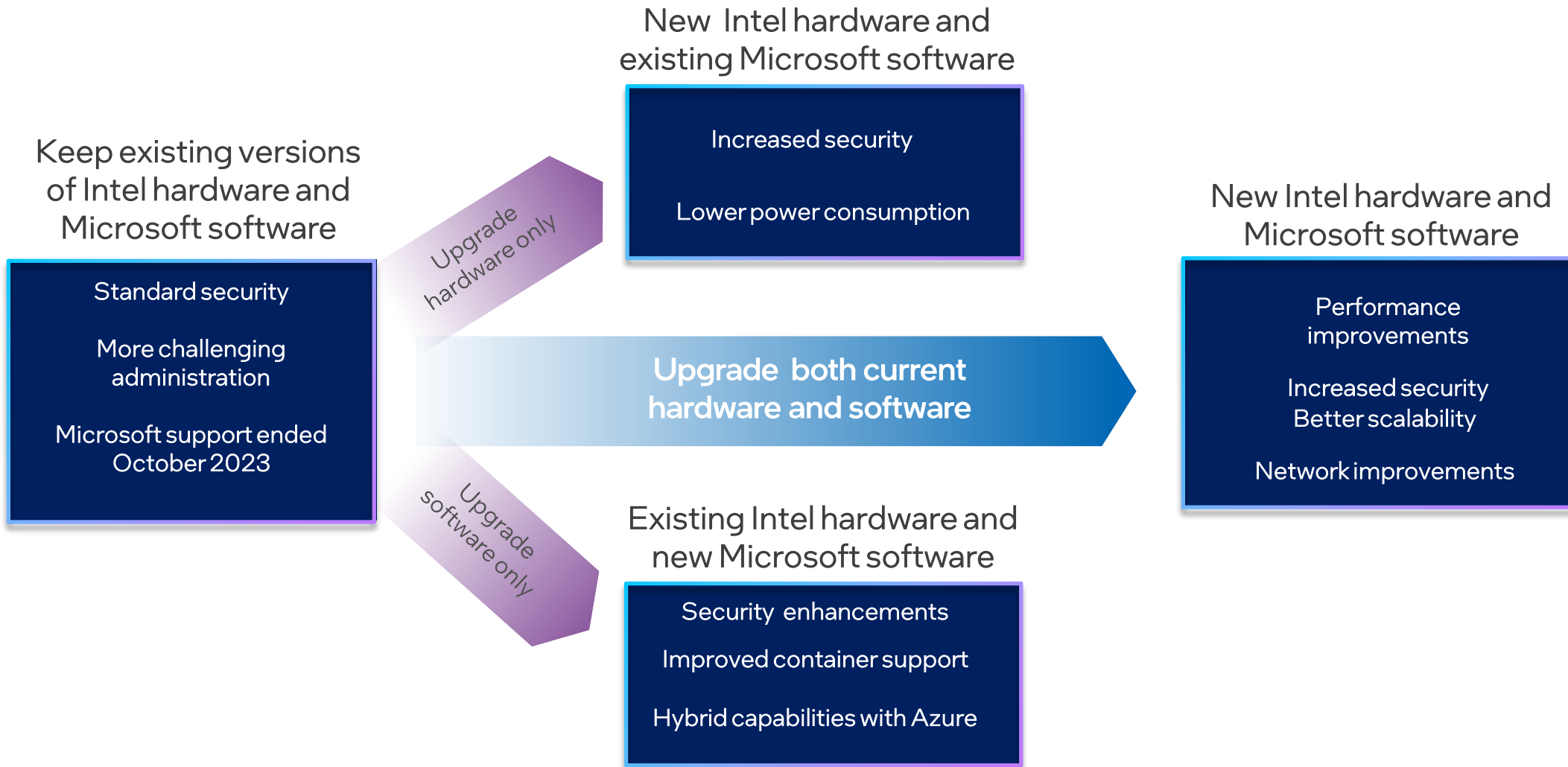
And upgrading to Windows Server 2022 provides the benefit of secured connectivity that adds an additional layer of security during transport—a secured-core server delivers powerful threat protection and multilayer security from chip to cloud.

Find customer solutions and strategies in the Data Center Modernization Enablement package. intel.com/salesenablement

intel.

Data Center Modernization Infographic

intel.

4th Gen Intel® Xeon® Scalable Processor and Windows Server 2022

New Intel hardware and existing Microsoft software

Increased security

Lower power consumption

Keep existing versions of Intel hardware and Microsoft software

Standard security

More challenging administration

Microsoft support ended October 2023

Upgrade hardware only

Upgrade both current hardware and software

New Intel hardware and Microsoft software

Performance improvements

Increased security
Better scalability

Network improvements

Upgrade software only

Existing Intel hardware and new Microsoft software

Security enhancements

Improved container support

Hybrid capabilities with Azure

Better Together: Windows Server 2022 + 4th Gen Intel® Xeon® Scalable Processor

Upgrade Paths / Value Propositions

SW ONLY UPGRADE

1st and 2nd Gen Intel® Xeon® Processors
- ✗ Lesser Platform Security (UEFI, TPM 2.0, Secure Boot)
- ✗ No Shielded VMs (Requires TPM 2.0)
- ✗ No HCI Management
- ✗ No Advanced Flash Support (NVMe, NVDIMM, or Intel® Optane™ SSD)

Windows Server 2022
- ✓ Security Enhancements (VBS, Windows Defender)
- ✓ Performance Improvements (Faster boot, login times)
- ✓ Improved Container support
- ✓ Hybrid Capabilities w/Azure (Arc Enabled; SMB Compression; WAC)
- ✓ Scalability (48TB Memory; 2048 Logical Cores across 64 sockets)

HW + SW UPGRADE

4th Gen Intel® Xeon® Processors
- ✓ Platform Security (UEFI, TPM 2.0, Secure Boot, Secure Core, TME-MK)
- ✓ Support for NVMe, NVDIMM, Intel® Optane™ SSD
- ✓ Up to 60 cores / 120 threads on 2-socket system, 6TB of memory
- ✓ Support for 100Gb NICs
- ✓ Accelerators (AMX, QAT)

Windows Server 2022
- ✓ Security Enhancements (VBS, Windows Defender)
- ✓ Performance Improvements (Faster boot, login times)
- ✓ Improved Container support
- ✓ Hybrid Capabilities w/Azure (Arc Enabled; SMB Compression; WAC)
- ✓ Scalability (48TB Memory; 2048 Logical Cores across 64 sockets)

CURRENT HW & SW – DO NOTHING

1st and 2nd Gen Intel® Xeon® Processors
- ✗ Lesser Platform Security (UEFI, TPM 2.0, Secure Boot)
- ✗ No Shielded VMs (Requires TPM 2.0)
- ✗ No HCI Management
- ✗ No Advanced Flash Support (NVMe, NVDIMM, or Intel® Optane™ SSD)

Windows Server 2012
- ✗ Less secure
- ✗ Non-compliance. Cost of Extended Security Updates
- ✗ Lack of hybrid/multi-cloud integration

HW ONLY UPGRADE

4th Gen Intel® Xeon® Processors
- ✓ Platform Security (UEFI, TPM 2.0, Secure Boot, Secure Core, TME-MK)
- ✓ Support for NVMe, NVDIMM, Intel® Optane™ SSD
- ✓ Up to 60 cores / 120 threads on 2-socket system, 6TB of memory
- ✓ Support for 100Gb NICs
- ✓ Accelerators (AMX, QAT)

Windows Server 2012
- ✗ OS incapable of taking advantage of new HW security features (Secure Core, TME-MK, Shielded VMs)
- ✗ Unable to scale to take advantage of HW innovations (Cores, Memory and VM Support)
- ✗ Lack of hybrid / multi-cloud integration
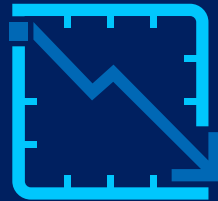- ✗ Non-compliance. Cost of Extended Security Updates

NEW ▲ **Software Modernization**

OLD | **Hardware Modernization** ▶ **NEW**

intel.

4th Gen Intel® Xeon® Scalable Processor
Summary

The 4th Gen Intel® Xeon® Scalable platform improves throughput and efficiency gen-over-gen, enabling server consolidation, improving TCO and helping achieve sustainability goals.

4th Gen Intel® Xeon® Scalable systems run complex AI workloads on the same hardware as your existing workloads, leveraging your data to accelerate insights and innovation.
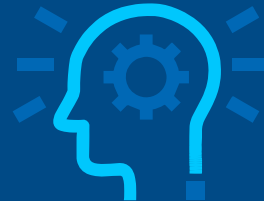
4th Gen Intel® Xeon® Scalable systems help secure and protect your data across environments without compromising performance.

Microsoft SQL Server 2012

Microsoft has ended support for Microsoft SQL 2012

Why Upgrade Your Hardware Now?

Upgrading to Microsoft SQL Server 2022 on older, out of warranty hardware will not allow you to use all the features of SQL 2022

Upgrading to Intel® 4th Generation Xeon® processors unlocks the full value of Windows Server 2022 for enhanced TCO and Security

LEARN HOW

Data Center Modernization Infographic

4th Gen Intel® Xeon® Scalable Processor and Microsoft SQL Server

Flexible, scalable, cloud-connected database for modern workloads

New Intel hardware and existing Microsoft software

Faster database backups

Increased security

Lower power consumption

Keep existing versions of Intel hardware and Microsoft software

New Intel hardware and Microsoft software

Standard security

More challenging administration

Microsoft support no longer available

Upgrade hardware only

Upgrade both current hardware and software

Faster, real-time Insights

Increased security

Significantly easier administration

Upgrade software only

Existing Intel hardware and new Microsoft software

Unified data management

Cloud-based analytics

Access to latest software features

Microsoft® SQL Server

Better Together: Microsoft SQL Server 2022 + 4ᵗʰ Gen Intel® Xeon® Scalable Processor

Upgrade Paths / Value Propositions

SW ONLY UPGRADE

1st and 2nd Gen Intel® Xeon® Processors
- ✗ Limited platform security
- ✗ No Shielded VMs
- ✗ No HCI Management
- ✗ No Advanced Flash Support (NVMe, NVDIMM, or Intel® Optane™ SSD)

SQL Server 2022
- ✓ Reduced workload times with much faster queries
- ✓ Seamless analytics over on-prem, operational data
- ✓ More secure database
- ✓ High availability, business continuity
- ✓ Resolution of conflicts with different data replicas is automatic
- ✓ Unified data governance and management

HW + SW UPGRADE

4th Gen Intel® Xeon® Processors
- ✓ 114% faster database backups with QAT
- ✓ Up to 34% more NOPM transactions compared to 3rd gen
- ✓ Up to 28% faster query response compared to 3rd Gen
- ✓ MSFT Storage Spaces Direct w/NVME, Intel® Optane™
- ✓ Support for NVMe, NVDIMM, Intel® Optane™ SSD
- ✓ Up to 60 cores / 120 threads on 2- socket system, 6TB of memory
- ✓ Support for 100Gb NICs
- ✓ Platform Security (UEFI, TPM 2.0, Secure Boot)

SQL Server 2022
- ✓ Reduced workload times with much faster queries
- ✓ Seamless analytics over on-prem, operational data
- ✓ More secure database
- ✓ High availability, business continuity
- ✓ Resolution of conflicts with different data replicas is automatic
- ✓ Unified data governance and management

CURRENT HW & SW – DO NOTHING

1st and 2nd Gen Intel® Xeon® Processors
- ✗ Limited platform security
- ✗ No Shielded VMs
- ✗ No HCI Management
- ✗ No Advanced Flash Support (NVMe, NVDIMM, or Intel® Optane™ SSD

SQL Server 2012 / 2016 / 2019
- ✗ More difficulty with real-time analysis
- ✗ Lower availability
- ✗ Less encryption and protection
- ✗ Administration can be more difficult
- ✗ Lose opportunity to migrate to Azure or hybrid cloud at same time
- ✗ Resolution of conflicts with different data replicas is manual

HW ONLY UPGRADE

4th Gen Intel® Xeon® Processors
- ✓ 114% faster database backups with QAT
- ✓ Up to 34% more NOPM transactions compared to 3rd gen
- ✓ Up to 28% faster query response compared to 3rd Gen
- ✓ MSFT Storage Spaces Direct w/NVME, Intel® Optane™
- ✓ Support for NVMe, NVDIMM, Intel® Optane™ SSD
- ✓ Up to 60 cores / 120 threads on 2- socket system, 6TB of memory
- ✓ Support for 100Gb NICs
- ✓ Increased platform security (UEFI, TPM 2.0, Secure Boot)

SQL Server 2012 / 2016 / 2019
- ✗ More difficulty with real-time analysis
- ✗ Lower availability
- ✗ Less encryption and protection
- ✗ Administration can be more difficult
- ✗ Lose opportunity to migrate to Azure or hybrid cloud at same time
- ✗ Resolution of conflicts with different data replicas is manual

Software Modernization ▲ **NEW**

OLD **Hardware Modernization** ▶ **NEW**

Summary

Performance across Workloads

Experience up to 34% faster transaction processing[1], 23% faster query response times, and 53%[2] faster backup times for Gen-over-Gen performance[3] on SQL Server 2022 when running on 4th Gen Intel® Xeon® Scalable processor.

Highly Secured Data Platform

SQL Server is the most secure database[4] and run it on 4th Gen Intel® Xeon® Scalable systems for enabling confidential and protection to the silicon level.

Ready for Modernization

SQL Server 2022 and 4th Gen Intel® Xeon® deliver cloud-connected, flexible platform to leverage digital transformation opportunities.

Microsoft Azure Stack HCI

Why Modernize Your Business with Intel & Microsoft Azure Stack HCI?

Portability
Intel® Virtualization Technology is foundational for seamless movement of data and apps between the data center, ever-growing edge, and Azure public cloud

Flexibility
Tailor performance and capacity to your needs with Intel's broad portfolio of products and accelerators, tools, libraries and frameworks

Increased Hardware-based Security
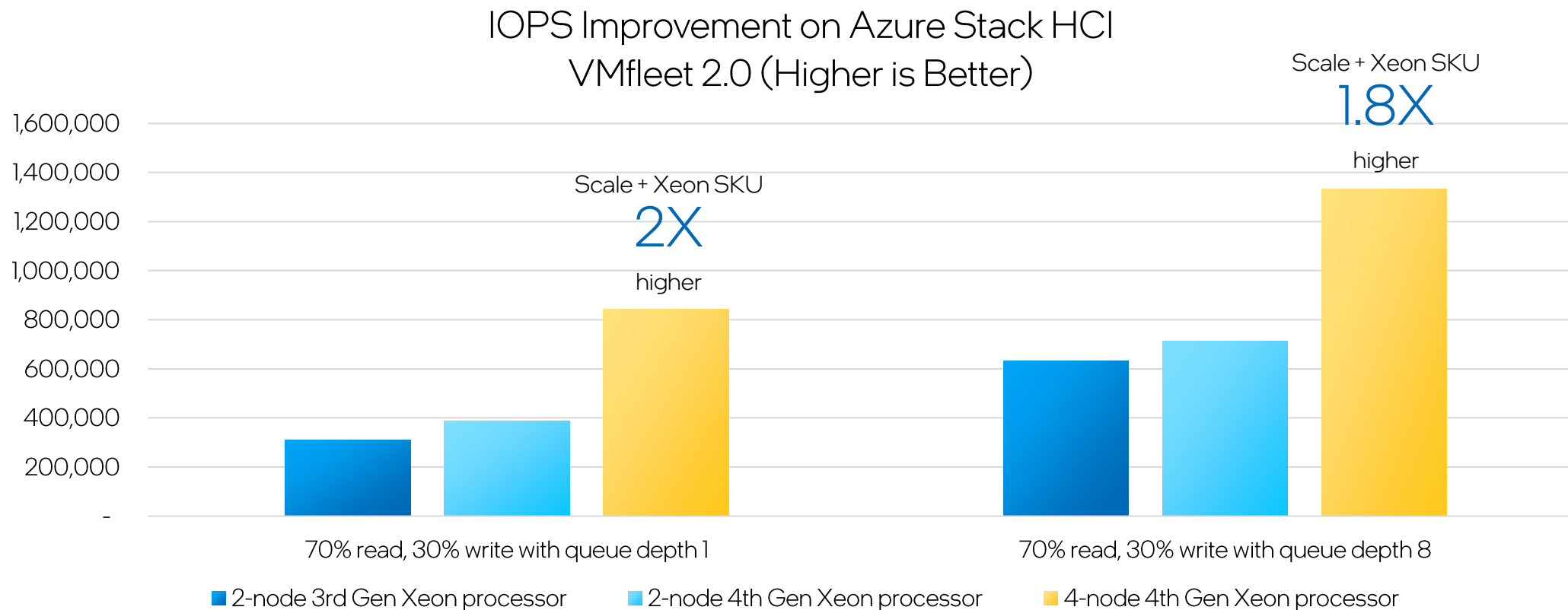with Intel® Total Memory Encryption, Intel® Crypto Acceleration, and Secured Core

Proven Solutions
Tested, verified for modernizing existing IT environments; certified by Microsoft, validated by the server vendor, and with performance verified by Intel

Enhance your infrastructure with hybrid cloud, optimized by trusted partners

Microsoft Azure Stack HCI Using 4ᵗʰ Gen Intel® Xeon® Scalable Processors Delivers Performance for Demanding Workloads

IOPS Improvement on Azure Stack HCI
VMfleet 2.0 (Higher is Better)

Scale + Xeon SKU
1.8X
higher

Scale + Xeon SKU
2X
higher

70% read, 30% write with queue depth 1	70% read, 30% write with queue depth 8

- ■ 2-node 3rd Gen Xeon processor
- ■ 2-node 4th Gen Xeon processor
- ■ 4-node 4th Gen Xeon processor

Y-axis: - , 200,000, 400,000, 600,000, 800,000, 1,000,000, 1,200,000, 1,400,000, 1,600,000

Performance increased gen-to-gen, scales with cluster size, and higher series of Xeon

intel.

Accelerate AI - Image Classification on Microsoft Azure Stack HCI using 4ᵗʰ Gen Intel® Xeon® Scalable Processors with Intel® AMX
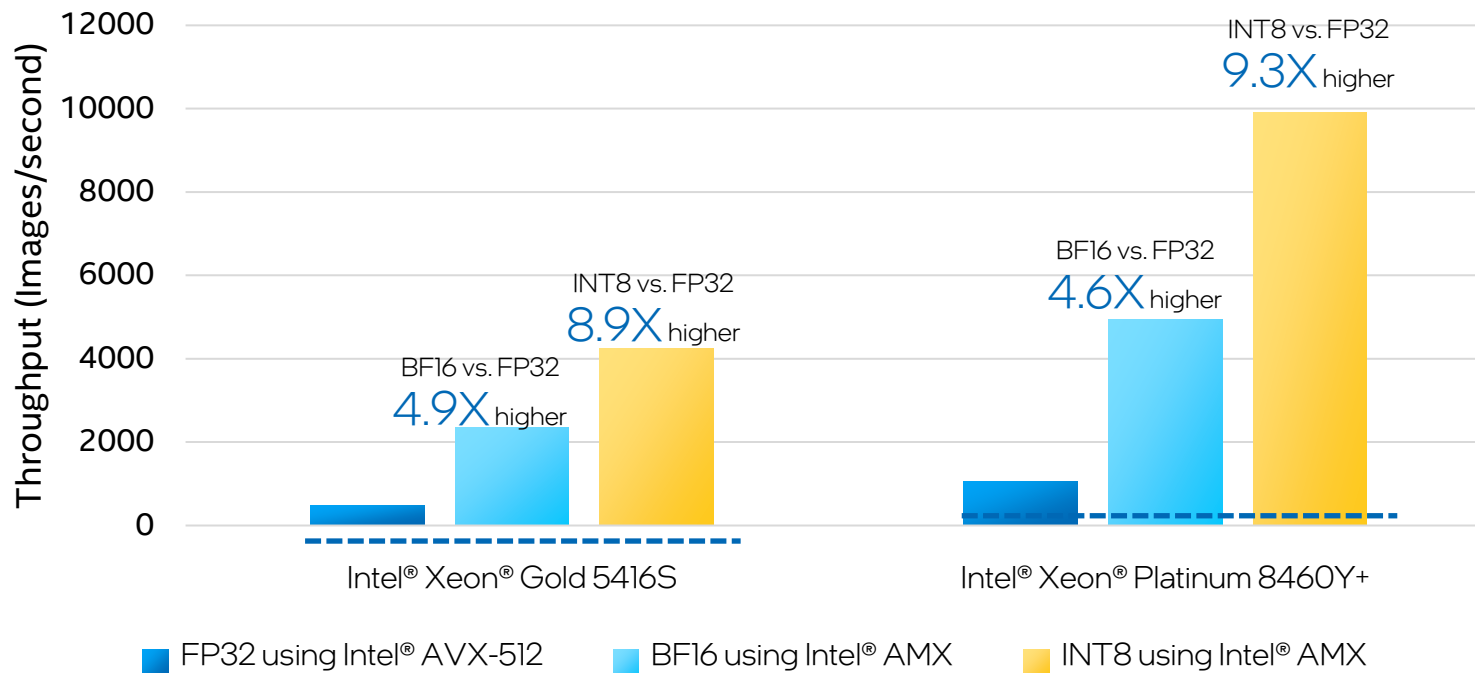

TensorFlow
Higher is better

Image Classification on Tensorflow 2.11 using ResNet50
BS=128, Multi-instance (16x2 and 40x2 instances)

Throughput (Images/second)

INT8 vs. FP32
9.3X higher

BF16 vs. FP32
4.6X higher

INT8 vs. FP32
8.9X higher

BF16 vs. FP32
4.9X higher

Intel® Xeon® Gold 5416S

Intel® Xeon® Platinum 8460Y+

■ FP32 using Intel® AVX-512   ■ BF16 using Intel® AMX   ■ INT8 using Intel® AMX
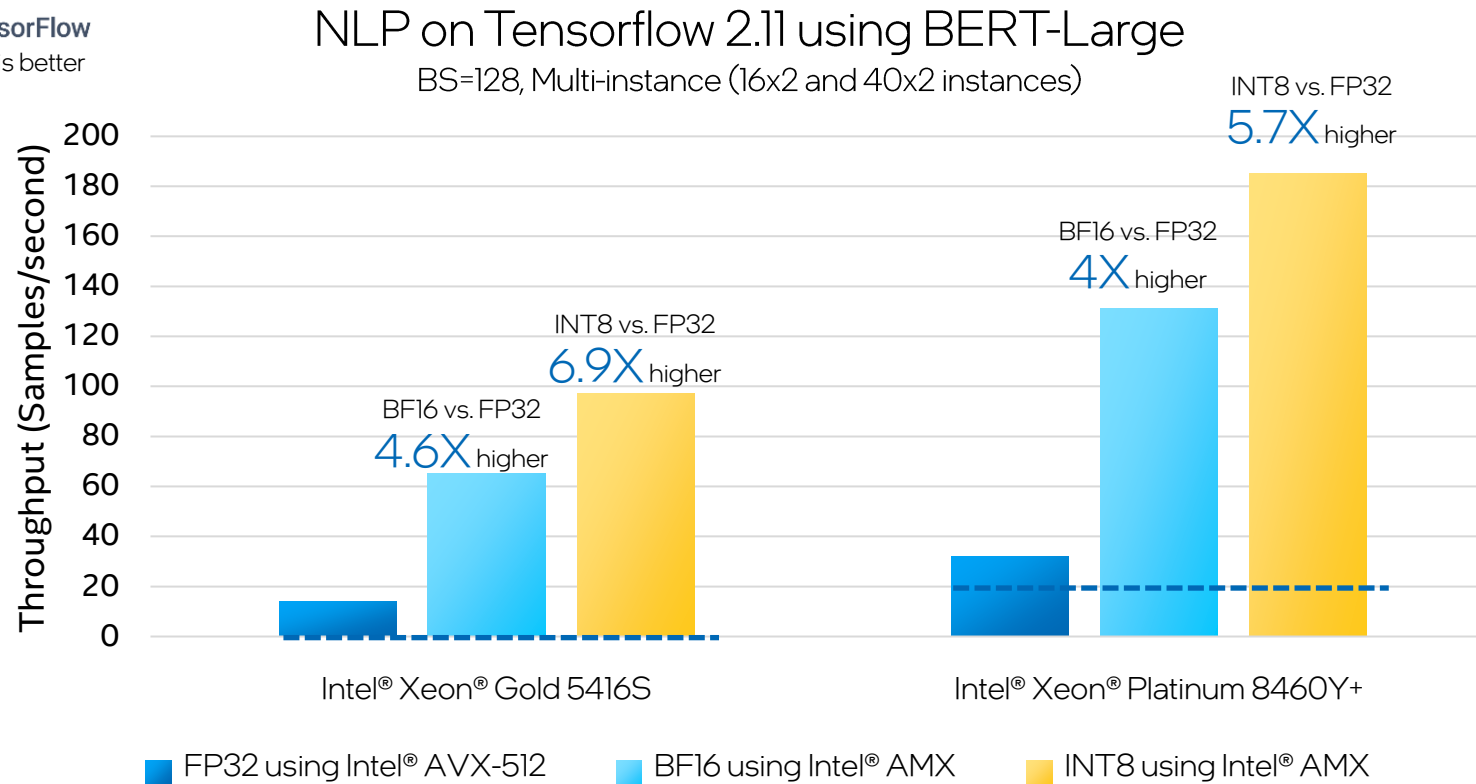
- The ResNet-50 benchmark measures image classification/vision workloads
- FP32 is a standard 32-bit floating point data type used to train deep learning models and for inferencing
- Bfloat16 is a truncated version of 32-bit floating point, used for both training and inference, offering similar accuracy but faster computation
- INT8 offers higher performance and is least computationally demanding for constrained environments, with minimal impact on accuracy
- Many DL workloads are mixed precision and 4ᵗʰ Gen Intel® Xeon® Scalable processors can seamlessly transition between Intel® AMX and Intel® AVX-512 to use the most efficient instruction set

Increase performance with higher series of Intel® Xeon® processor or by changing precision

Accelerate Natural Language Processing (NLP) on Microsoft Azure Stack HCI using 4th Gen Intel® Xeon® Scalable Processors with Intel® AMX

TensorFlow
Higher is better

NLP on Tensorflow 2.11 using BERT-Large
BS=128, Multi-instance (16x2 and 40x2 instances)

INT8 vs. FP32
5.7X higher

BF16 vs. FP32
4X higher

INT8 vs. FP32
6.9X higher

BF16 vs. FP32
4.6X higher

Throughput (Samples/second)

Intel® Xeon® Gold 5416S

Intel® Xeon® Platinum 8460Y+

■ FP32 using Intel® AVX-512   ■ BF16 using Intel® AMX   ■ INT8 using Intel® AMX

- BERT-Large is a pretrained model used for Natural Language Processing

- FP32 is a standard 32-bit floating point data type used to train deep learning models and for inferencing

- Bfloat16 is a truncated version of 32-bit floating point, used for both training and inference, offering similar accuracy but faster computation

- INT8 offers higher performance and is least computationally demanding for constrained environments, with minimal impact on accuracy

- Many DL workloads are mixed precision and 4th Gen Intel® Xeon® Scalable processors can seamlessly transition between Intel AMX and Intel AVX-512 to use the most efficient instruction set

Increase performance with higher series of Intel Xeon processor or by changing precision

Why Modernize Now?
New Applications Demand New Infrastructure

Efficiency

Artificial
Intelligence

Security

5 ways Intel® Xeon® Scalable processors can support your fastest-growing workloads

Significant Performance Leaps

5th Gen Intel® Xeon® CPUs provides generational improvements on CPU and platform upgrades

Efficiency

CPU upgrade

4th Gen Intel® Xeon® CPU vs. 5th Gen Intel Xeon CPU

1.21x
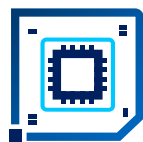average performance gain

Up to **1.42x**
higher inference

Up to **1.4x**
higher HPC performance gain

Up to **1.7x**
higher throughput

General Purpose Compute

AI
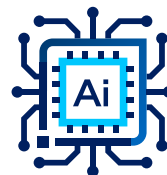
HPC
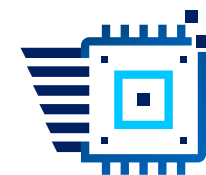
Networking and Storage

Server platform upgrade

3rd Gen Intel® Xeon® CPU vs. 5th Gen Intel Xeon CPU

1.84x
average performance gain

Up to **14x**
higher inference and training performance

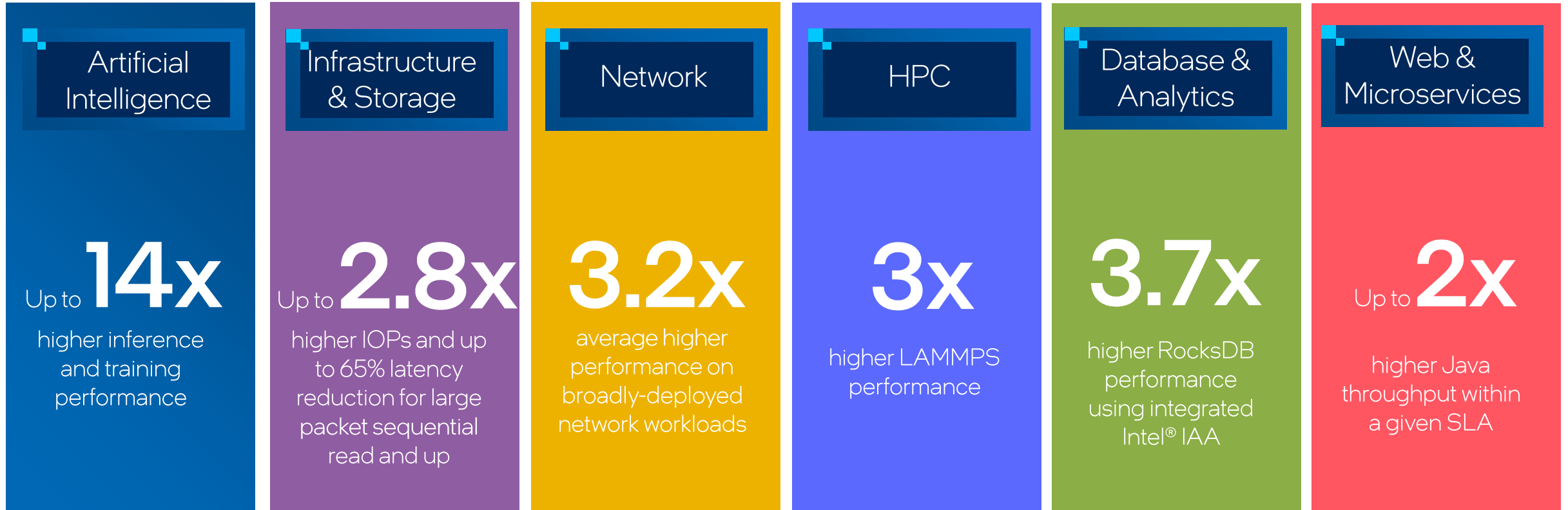Up to **2.1x**
average performance gain

3.6x
higher throughput

intel

Significant Performance Gains
5th Gen Intel® Xeon® Scalable processors vs. 3rd Gen Intel® Xeon® processors

Efficiency

Artificial Intelligence	Infrastructure & Storage	Network	HPC	Database & Analytics	Web & Microservices
Up to **14x** higher inference and training performance	Up to **2.8x** higher IOPs and up to 65% latency reduction for large packet sequential read and up	**3.2x** average higher performance on broadly-deployed network workloads	**3x** higher LAMMPS performance	**3.7x** higher RocksDB performance using integrated Intel® IAA	Up to **2x** higher Java throughput within a given SLA

See [A17, N16, N11, H10, D1, W3] at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary.

intel. 21

The Sustainable Data Center

Efficiency

-20% Reduce power consumption, even in demanding workloads

2.9x Improve efficiency and performance across targeted workloads

Up to 20% Maximize energy efficiency on lower-utilization workloads

Watch
Cloud TV: Sustainability and the Cloud
Discover 5 ways you can help your customers practice sustainability with Intel technologies

More Sustainability Guidance

Infographic
The Sustainable Data Center

Video
Sustainability with Intel Technologies

Study
The Sustainable CTO: The Road to Tech Positive

intel

22

The Sustainable Data Center

How Intel® Xeon® processor–powered servers, compares to AMD EPYC processor–powered servers

96% Less electricity used in AI workloads

51% Less electricity used in Network Infrastructure

39% Less electricity used in HPC workloads
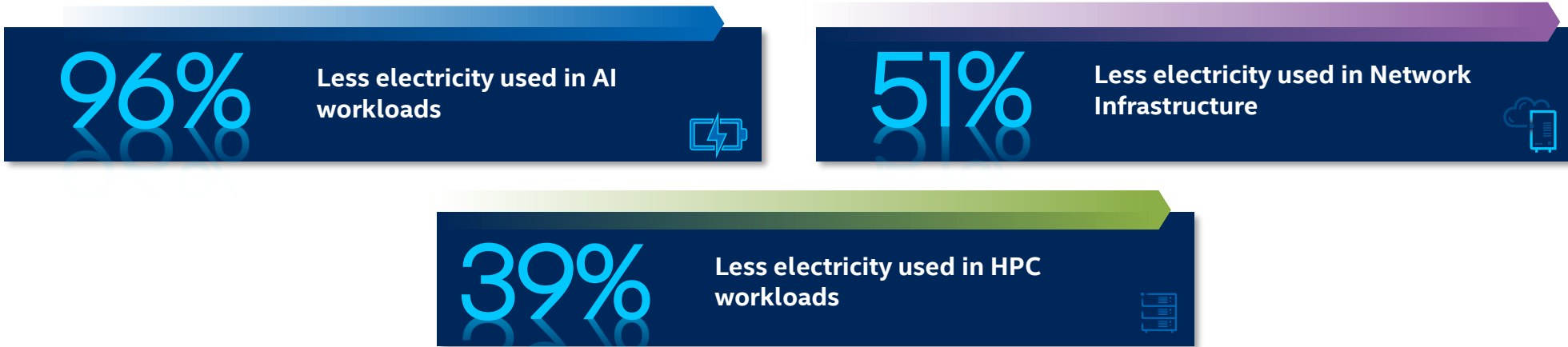
5 Reasons Why Processor Selection Makes a Difference Infographic

Liquid Cooling
Improving TCO through Energy Efficiency and Water Reduction

Liquid Cooling Benefits

Power / Performance

- Reduce PUE[1] (Power Usage Effectiveness) from 1.3+ to as low as 1.03[3]

- Lower power consumption by up to 30%[3]

- Extends cooling range for higher system thermals

Density

- More compute volume within same rack footprint[4]

- Less real estate needed per compute output[4]

Water

- Liquid cooling can significantly reduce the billions of gallons of water used in air cooled data centers[2]
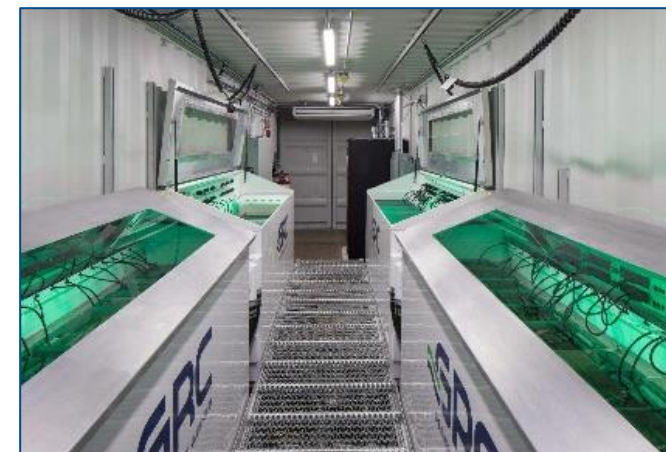
Liquid Cooling Solution Benefits

Energy

Up to 40%[2]
reduction in TCO

~1.03 Enables PUEs to 1.03[1]
PUE: Power Usage Effectiveness

40% Reduction in cooling CAPEX[1]

95% Reduction in cooling OPEX[1]

~30% Reduction in power consumption[1]

Water

Up to 100%
reduction in water use[1]

Use Heat for District Heating

Use Heat for Urban Farming

If evaporative cooling is used, water reduction can still be significant over conventional rack air cooling

Density

10x
Increase in compute density[2]

Enables compute in dense edge environments

Increases compute density per sqm

Eliminates physical components (fans, chillers)

Forecasted Growth for Liquid Cooling (2022-2027): 36.3%[3]

PUE: Power Usage Effectiveness = amount of power entering data center/amount of power needed to run IT equipment

[1]Source: Green Revolution Cooling
Water reduction dependent on system and heat reuse

[2]Source: Submer

[3]BIS Research

intel

25

Why Choose 5th Gen Intel® Xeon® processors for Server Refresh?

Efficiency

Lower Total Cost of Ownership (TCO)

Intel's portfolio of hardware, software, systems, and tools can help advance your data center's overall efficiency, creating energy savings and reducing your carbon footprint, without sacrificing performance, while giving you the TCO and flexibility you need.

Up to
77%
reduction in TCO[1]

Efficiency

Intel® Accelerator Engines boost CPU utilization, reduce electricity consumption resulting in lower impact on the environment.

10x
better efficiency (perf/watt) with built-in accelerators[3]

Optimized workload performance

By delivering more performance per core with built-in accelerators, 5th Gen Intel® Xeon® processors help you meet requirements for even the most demanding workloads.

84%
Performance gain[2]

Modernization

Intel® Xeon® processors deliver the low-latency, high-bandwidth capabilities required by modern and AI-infused workloads. Replacing aging infrastructure with these speedy and energy efficient processors will help you keep pace with rapidly evolving market needs.

Up to
16:1
server consolidation[1]

Confidential Computing

intel **XEON**

With Intel, you can choose from the most deployed confidential computing options in data centers on the market today—now including application or VM-level isolation.
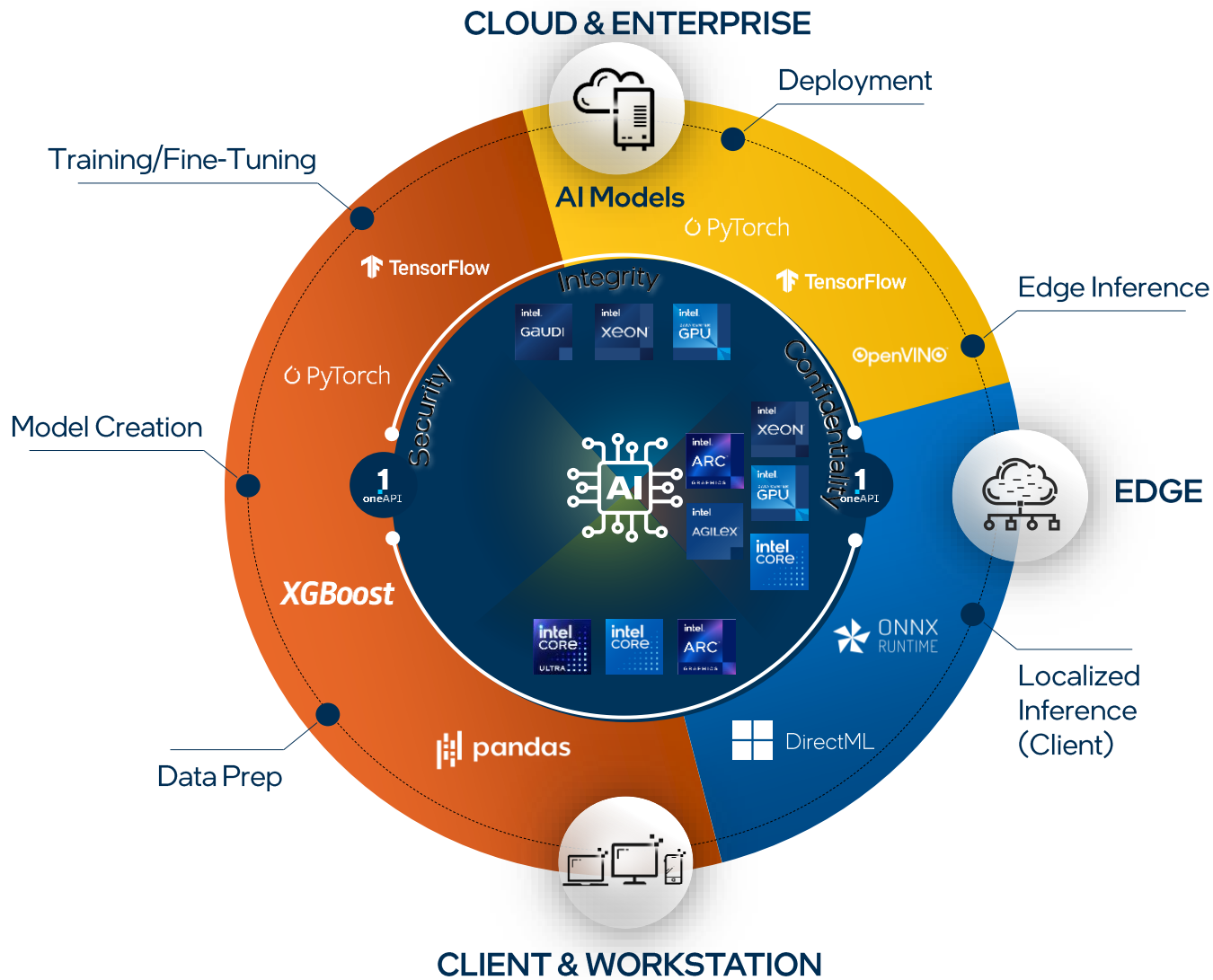
[What's the right transition for your customer?]
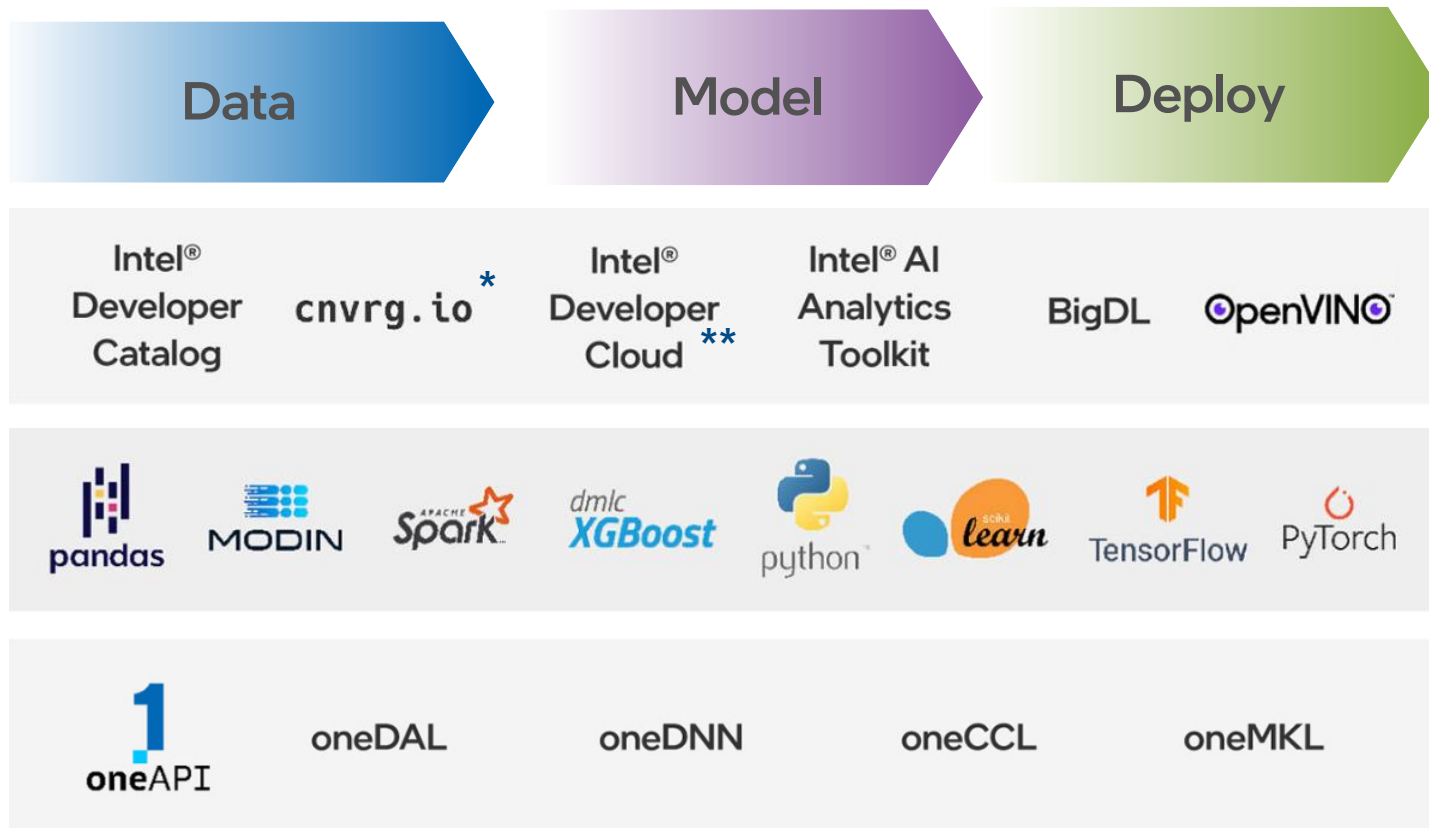
AI Continuum
Bringing AI Everywhere

Note: Intel® Core Ultra integrates NPU low power inference engine from Meteor Lake onwards.

Intel AI Software Enables AI Everywhere Faster

Accelerating development with optimizations of the most popular industry libraries and open source tools, the Intel® AI software suite unleashes the performance of Intel® Xeon® Scalable processors without code changes

Data ⟶ **Model** ⟶ **Deploy**

Intel Productivity Tools & Kits

Intel® Developer Catalog | cnvrg.io * | Intel® Developer Cloud ** | Intel® AI Analytics Toolkit | BigDL | OpenVINO™

Optimized Frameworks

pandas | MODIN | Spark™ | dmlc XGBoost | python™ | learn | TensorFlow | PyTorch

oneAPI Libraries

oneAPI | oneDAL | oneDNN | oneCCL | oneMKL

The Intel® AI software suite has been validated on over 400 AI models and use cases to help ensure that you achieve out-of-the-box application performance

* Now known as: Intel® Tiber™ AI Studio
** Now known as: Intel® Tiber™ Developer Cloud

Accelerate AI Development with Reference Kits

Optimized AI reference kits help developers and data scientists innovate faster

Why it matters

Built on the oneAPI open, standards-based, heterogeneous programming model and components of Intel's end-to-end AI software portfolio, such as Intel® AI Analytics Toolkit and the Intel® Distribution of OpenVINO™ toolkit, the reference kits enable AI developers to streamline the process of introducing AI into their applications, enhancing existing intelligent solutions and accelerating deployment.

The result is proven performance improvements with a shorter, more productive workflow versus a traditional model development workflow

Using the **AI reference kit** designed to set up interactions with an enterprise conversational AI chatbot, users can experience inferencing in batch mode up to 45% faster with oneAPI optimizations

The **AI reference kit** designed to automate visual quality control inspections for life sciences demonstrated training up to 20% faster and inferencing 55% faster for visual defect detection with oneAPI optimizations.
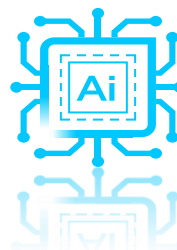
To enable developers to predict utility asset health and deliver higher service reliability, there is an **AI reference kit** that provides up to a 25% increase in prediction accuracy.

Press Release: Intel Accelerates AI Development with Reference Kits
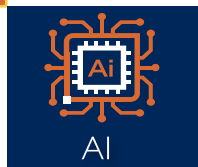
intel.

4th Gen Intel® Xeon® Scalable Processors
AI Accelerators

Intel® Advanced Matrix Extensions (Intel® AMX)

Intel® Advanced Matrix Extensions (Intel® AMX) accelerates deep learning fine-tuning and inference on Intel® Xeon® Scalable processors. Intel AMX is built into every core on 4th and 5th Gen Xeon® processors, accelerating bfloat16 (BF16) and INT8 data types.

More Information
Website
Solution Brief
Video
User Guide

Get started with Intel® AMX

Intel AMX can deliver up to **10x generational performance gains**[1] for AI workloads. It is enabled in **Intel® 4th Gen Xeon® Scalable processors** available through OEMs, partners, or hosted on cloud service providers, such as:

GCP-C3	C7i M7i R7i	VPC	{GCR} 8i Q4'22	{GCM} 6	Red Hat Enterprise Linux	SUSE Linux Enterprise Server 15 SP4	Ubuntu 20.04 & 21.10	Oracle Linux

Intel AI Optimizations Quick Start Guide

4th Gen Intel® Xeon® Scalable Processors
AI Accelerators

Intel® Advanced Vector Extensions 512 (Intel® AVX-512)

Significantly accelerates deep learning training and inference, ideal for workloads like natural language processing, recommendation systems and image recognition
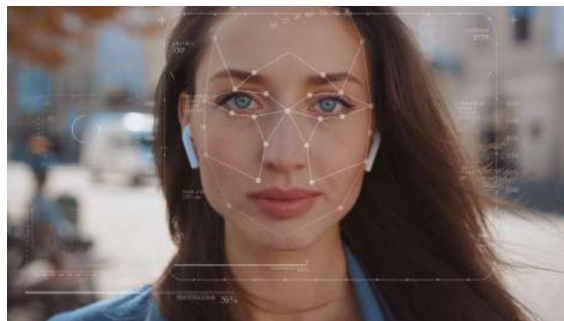
[Website](#)
[Solution Brief](#)
[Video](#)
[User Guide and Downloads](#)
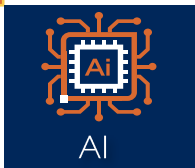
4th Gen Intel® Xeon® Scalable Processor AI Inference Performance Results with Customers & Partners

Tencent Cloud

2-3x AI throughput for BERT models used by Tencent Search application with Intel® AMX vs previous generation

Tencent can use the optimized BERT model to deliver better service experiences and to help reduce TCO

[Case Study](#)

美团 Meituan

3.4x AI throughput with Bfloat16 optimizations for Meituan's Compute Vision Platform with Intel AMX vs. without AMX optimizations

Meituan increased the overall efficiency of its online resources by over 3x and saved 70% on service costs

[Case Study](#)

vmware®

5.7x Natural Language Processing (NLP) on vSphere/vSAN 8.0 Using a 4th Gen Intel® Xeon® Scalable Processor with Intel® AMX

Intel offers a broad range of open and free-to-use tools, optimized libraries, and industry frameworks to deliver the best out-of-the-box performance and end to end productivity

[Case Study](#)   [Article & Demo](#)

See case study links for workloads and configurations. Results may vary.
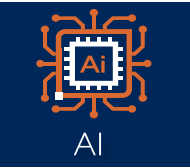
intel®

4th Gen Intel® Xeon® Scalable Processor AI Case Studies

FOX SPORTS

"We've shaved weeks off of setup time"

"For us, Intel Xeon processors are a cornerstone of how we deploy technology. We run only on Intel Xeon CPUs, and that gives us the ability to run everywhere: in VMs, in dedicated on-premises bare metal, in the cloud."

Case Study

SIEMENS

35x speedup in AI inference time for auto contouring algorithms compared to previous gen[1]

20% reduction in energy consumption compared to previous gen[2]

Case Study Video

Intel Offers the Most Comprehensive Portfolio

Security

Intel® Software Guard Extensions (Intel® SGX)

Application isolation

Intel® Trust Domain Extensions (Intel® TDX)

Virtual machine isolation

Intel® Tiber™ Trust Services formerly Intel® Trust Authority

Independent trust verification services for multi-cloud & hybrid cloud

Software Solutions, Cloud, OEM and System Integrator Ecosystem

Intel Security-First Development & Lifecycle Support

*Intel® TDX available through select cloud providers

intel

Intel Trusted Execution Environments

Application-level isolation: Intel® SGX

VM-level isolation: Intel® TDX

Advantages

- Separation from cloud provider and other tenants
- Smaller trust boundary and potential attack surface
- More amenable to code inspection and monitoring
- Deployable in VMs, cloud-native containers and bare-metal

Considerations

- Apps may require specific development or tailoring
- Frequent calls outside the enclave may impact performance

Advantages

- Separation from cloud provider and other tenants
- Lowest porting effort for existing applications
- More amenable to enterprise-wide deployment mandates
- Can be a simple instance configurator setting

Considerations

- Larger trust boundary (guest OS, all apps, VM admins)
- Possible re-validation with updated guest OS & hypervisor
- Less granular attestation

Intel® TDX Availability

Security

Intel® TDX is available on 4th Gen Intel® Xeon® Scalable instances in public preview through three leading cloud providers

Click on the logos below for more information on each cloud provider's offering

Microsoft Azure   **Alibaba Cloud**   **Google Cloud**

IBM Cloud
*public preview TBA

Intel® TDX is enabled on the following guest OS vendors

Red Hat   **SUSE**   **Canonical Ubuntu**

How to Get Started

Security

Intel® Software Guard Extensions (Intel® SGX)

More information   Get Started

Cloud Service Providers
Click on logos for more info

Azure   CORE   OVHcloud   phoenixNAP GLOBAL IT SERVICES

swisscom   IBM Cloud   CloudSigma

OEMs
Click on logos for more info

Hewlett Packard Enterprise   SUPERMICRO

Training & Documentation

Training Videos   Technical Library   Solution Brief

Intel® Trust Domain Extensions (Intel® TDX)

More information

Documentation

Trust Domain Security Guidance for Developers

Get Started

Intel® Trust Domain Extension (Intel® TDX) Module Download

Intel® Trust Domain Extension (Intel® TDX) Loader

Intel® Tiber™ Trust Services

formerly Intel® Trust Authority

Put Zero Trust Within Reach and Get Public Cloud Flexibility with Private Cloud Security

Intel® Tiber™ Trust Services is a new portfolio of software and services that brings enhanced security and assurance to Confidential Computing with Zero Trust principles
In its first generation, it offers an independent attestation service that attests to Trusted Execution Environments (TEEs) that are based on (Intel® SGX) and (Intel® TDX)

Implement the tenets of Zero Trust without incurring the cost and complexity of building your own attestation service

Independent ➤ Scalable ➤ Easy to Deploy

Learn More

Product Brief

What That Means Video

Noname Case Study

Thales Case Study

Zscaler Case Study

n noname

THALES

zscaler

VMware vSphere 8.0 on
4th Gen Intel® Xeon® Scalable Processors

VMware vSphere 8

The Enterprise Workload Platform

intel.
Xeon®
4th Gen
Intel Xeon Scalable Processor

Scalability

VMware vSAN

Distributed
Service Engine

Intel® Advanced Matrix
Extensions
(Intel® AMX)

Intel® Advanced Vector
Extensions(Intel® AVX)

Security

Virtual NUMA

Intel® Dynamic Load
Balancer (Intel® DLB)

Intel® QuickAssist
Technology (Intel® QAT)

Tanzu
Kubernetes

Intel® SGX

Lifecycle
Management

AVX-512,
AMX vMotion
Supported

DRS and
vMotion

New Virtual
Hardware

Unlocking the Value of Accelerators with Software
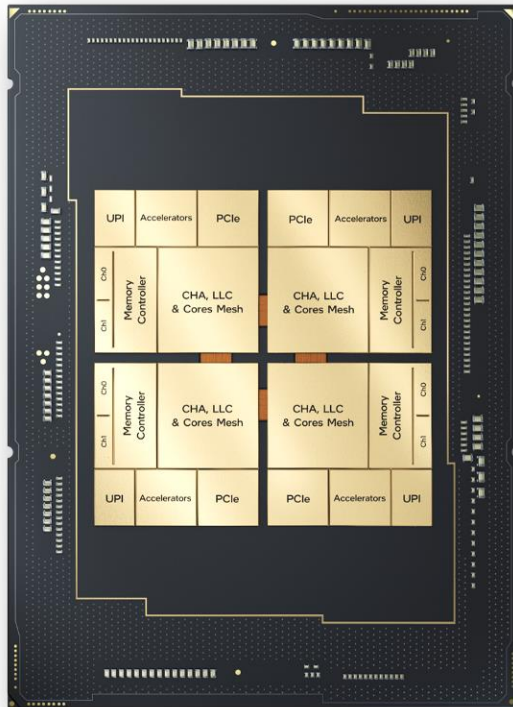
Intel® Advanced Matrix Extensions (Intel® AMX)

- TensorFlow
- PyTorch
- ONNX Runtime
- OpenVINO
- oneDNN (Intel oneAPI)

Intel® Advanced Vector Extensions (Intel® AVX) for vRAN

- FlexRAN
- Data Plane dev Kit (DPDK)*

Intel® In-memory Analytics Accelerator (Intel® IAA)

- Intel Query Processing Library

Intel® Data Streaming Accelerator (Intel® DSA)

- Storage Perf Dev Kit (SPDK)*
- Data Plane Dev Kit (DPDK)*

Intel® QuickAssist Technology (Intel® QAT)

- QATzip* (Intel lib)
- OpenSSL**
- Boring SSL

Intel® Dynamic Load Balancer (Intel® DLB)

- VPP IPsec
- Data Plane Dev Kit (DPDK)*

*Intel open-source library (not part of stock SW).
**Difference between Intel version and stock version.
***Intel® QPL and *Intel® DML in open-source beta, v1.0.0 coming shortly.

intel®

Accelerate AI - Image Classification on vSphere / vSAN 8.0 using 4th Gen Intel® Xeon® Scalable Processor with Intel® AMX

Image Classification on Tensorflow 2.11 using ResNet50
BS=128, Multi-instance (4 cores per instance)

Higher is better
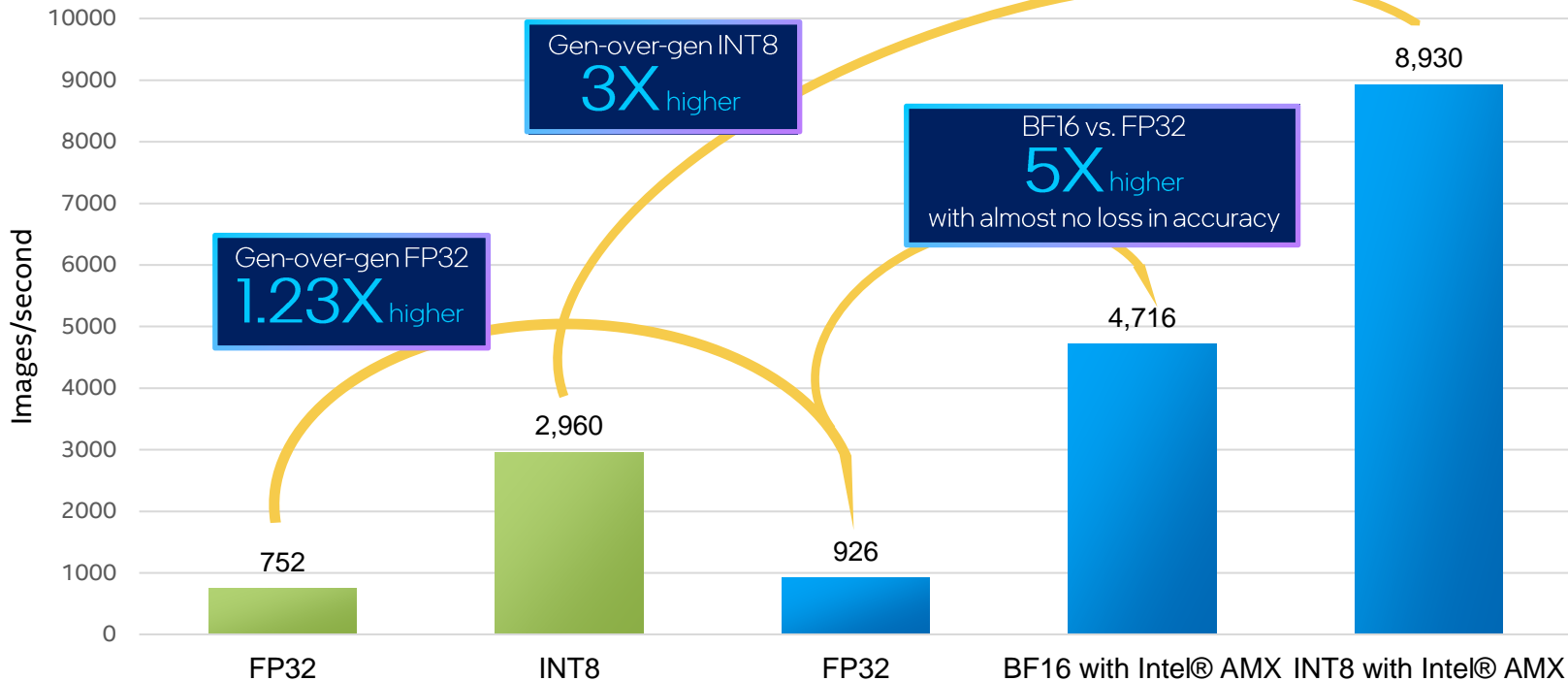
Images/second (Y-axis, 0 to 10000)

- Gen-over-gen INT8 **3X** higher
- Gen-over-gen FP32 **1.23X** higher
- BF16 vs. FP32 **5X** higher with almost no loss in accuracy

Bar values:
- FP32: 752
- INT8: 2,960
- FP32: 926
- BF16 with Intel® AMX: 4,716
- INT8 with Intel® AMX: 8,930

Legend:
- Intel® Xeon® Gold 6348, 2.6GHz, 28c Intel® AVX-512 + Intel® DL Boost for FP32
- Intel® Xeon® Gold 6448Y, 2.1GHz, 32c Intel® AVX-512 + Intel® DL Boost for FP32; Intel® AMX for BF16, INT8

- The ResNet-50 benchmark measures image classification/vision workloads

- FP32 is a standard 32-bit floating point data type used to train deep learning models and for inferencing – more computationally demanding, but typically achieves higher accuracies

- Bfloat16 is a truncated version of 32-bit floating point, used for both training and inference, offering similar accuracy but faster computation

- INT8 offers higher performance and is least computationally demanding for constrained environments, with minimal impact on accuracy

- Many DL workloads are mixed precision and 4th Gen Xeon® Scalable processors can seamlessly transition between AMX and AVX-512 to use the most efficient instruction set
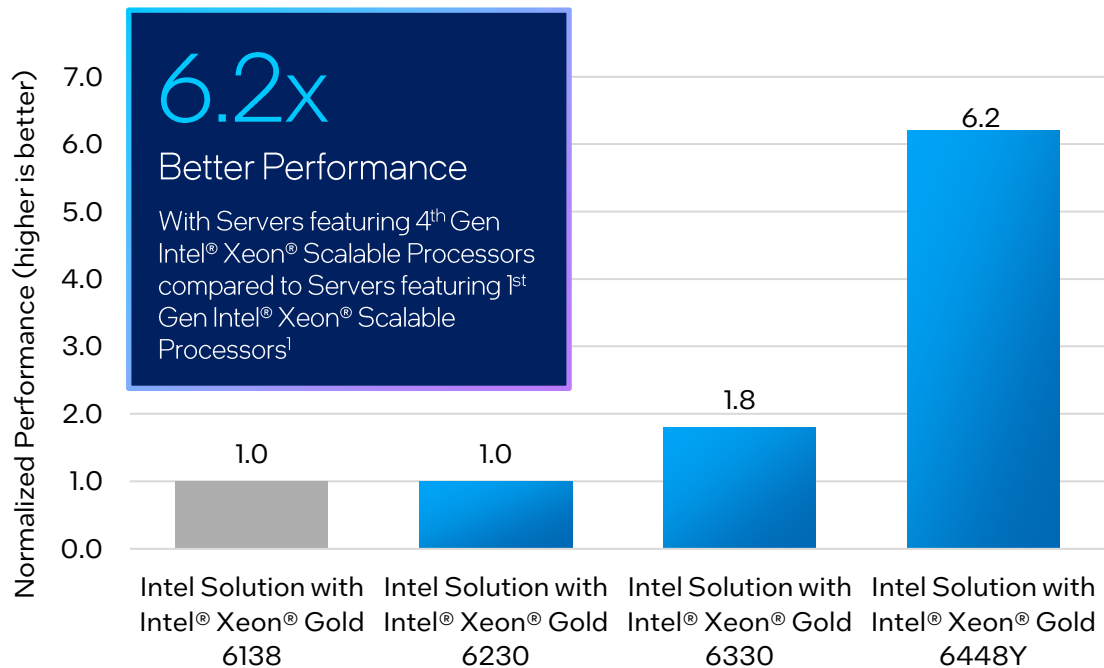
Partner Benchmarking

vmware®

NLP on Tensorflow 2.11 using BERT-Large
BS=128, 28x2 and 32x2 instances
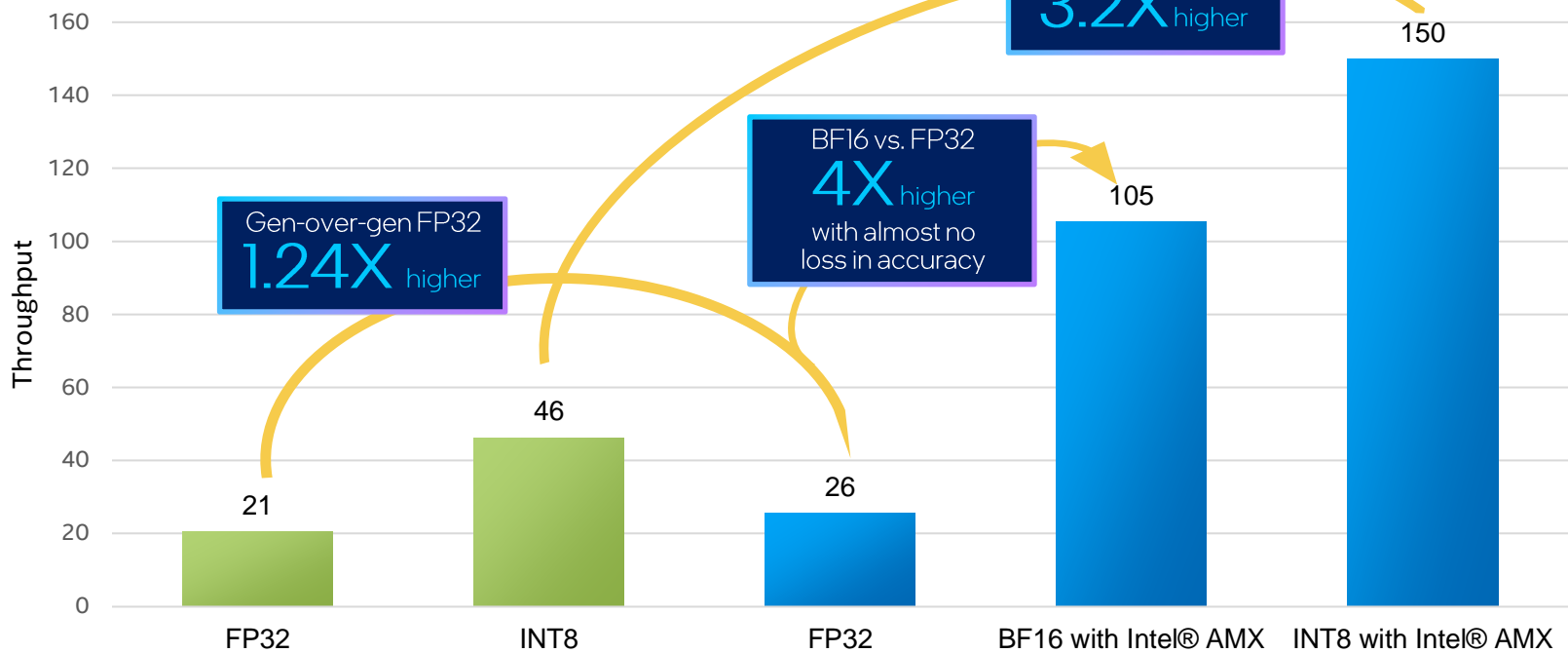
TensorFlow

Higher is better

Throughput

Gen-over-gen FP32
1.24X higher

BF16 vs. FP32
4X higher
with almost no loss in accuracy

Gen-over-gen INT8
3.2X higher

- FP32: 21
- INT8: 46
- FP32: 26
- BF16 with Intel® AMX: 105
- INT8 with Intel® AMX: 150

■ Intel® Xeon® Gold 6348, 2.6GHz, 28c
Intel® AVX-512 + Intel® DL Boost for FP32

■ Intel® Xeon® Gold 6448Y, 2.1GHz, 32c
Intel® AVX-512 + Intel® DL Boost for FP32; Intel® AMX for BF16, INT8

- BERT-Large is a pretrained model used for Natural Language Processing

- FP32 is a standard 32-bit floating point data type used to train deep learning models and for inferencing – more computationally demanding, but typically achieves higher accuracies

- Bfloat16 is a truncated version of 32-bit floating point, used for both training and inference, offering similar accuracy but faster computation

- INT8 offers higher performance and is least computationally demanding for constrained environments, with minimal impact on accuracy

- Many DL workloads are mixed precision and 4th Gen Xeon® Scalable processors can seamlessly transition between AMX and AVX-512 to use the most efficient instruction set

VMware vSAN 8.0 Select Modernization Test Results

Partner Benchmarking

vmware

Performance and Latency Improvements Due to HW and SW Advancements

Generational performance and latency improvements are due to the newest Intel® technologies combined with the new Express Storage Architecture (ESA) introduced in VMware vSphere 8.0. ESA is an optional alternative architecture in vSAN that is designed to process and store data with all-new levels of efficiency, scalability, and performance.

Performance of HCI Bench
Higher is better

6.2x

Better Performance

With Servers featuring 4th Gen Intel® Xeon® Scalable Processors compared to Servers featuring 1st Gen Intel® Xeon® Scalable Processors[1]

Normalized Performance (higher is better)

Value	Solution
1.0	Intel Solution with Intel® Xeon® Gold 6138
1.0	Intel Solution with Intel® Xeon® Gold 6230
1.8	Intel Solution with Intel® Xeon® Gold 6330
6.2	Intel Solution with Intel® Xeon® Gold 6448Y

Scenario 8k block size 70% reads 100% random

Latency of HCI Bench
Lower is better

7.1x

Lower Latency

With Servers featuring 4th Gen Intel® Xeon® Scalable Processors.[1]

Normalized latency (lower is better)

Value	Solution
7.1	Intel Solution with Intel® Xeon® Gold 6138
7.1	Intel Solution with Intel® Xeon® Gold 6230
6.1	Intel Solution with Intel® Xeon® Gold 6330
1.0	Intel Solution with Intel® Xeon® Gold 6448Y
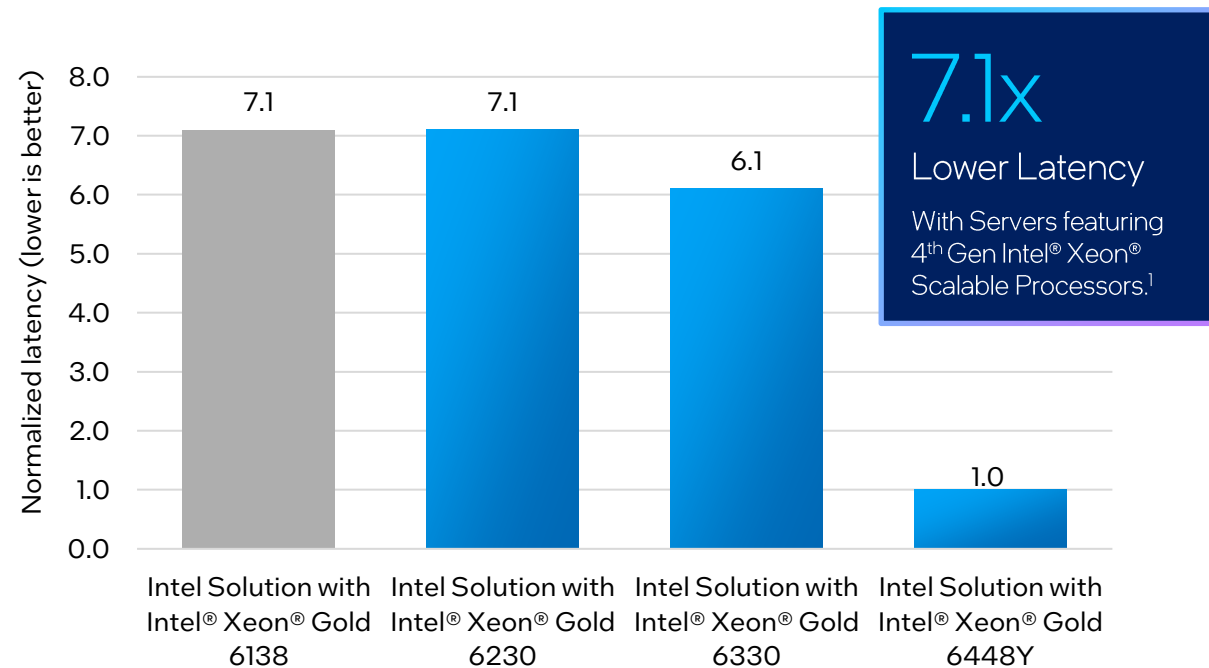
Scenario 8k block size 70% reads 100% random

1. See backup for workloads and configurations. Results may vary.

4th Gen Intel® Xeon® Scalable Processor Support Red Hat Q3'23

Feature	RHEL – BareMetal	RHEL – Virtualized	OpenShift (K8S)	Accelerator Getting Started Guide
4th Gen Xeon®	8.6, 9.0	8.6, 9.0	4.11	N/A
DSA (Data Streaming Accelerator)	8.6 / 9.0	TBD	4.13 (Q4'23/Q1'24)	DSA Guide
IAA (In-memory Analytics Accel.)	8.6 / 9.0	TBD	4.13 (Q4'23/Q1'24)	IAA Guide
QAT (Quick Assist Technology)	8.6 & 9.0	TBD - OOT* until Q2'24	4.12	QAT Guide
AMX (Adv. Matrix eXtensions)	8.6 & 9.0	8.6, 9.0	4.11	AMX Guide
AVX (Adv. Vector eXtensions)	8.6, 9.0	8.6, 9.0	4.11	Not Available
DLB (Dynamic Load Balancer)	OOT* until Q2'24	OOT* until Q2'24	TBD	TBD
SGX (SW Guard eXtensions)	8.6, 9.0	8.6, 9.0	4.11	SGX Guide
TDX (Trust Domain eXtensions)	TBD	8.8, 9.2 (VM Guest & TBD on Host)	TBD	TDX Guides
SIOV (Scalable I/O Virtualization)	9.2 (target)	-	-	Not Available
SST (Speed Select Technology)	8.6 / 9.0	N/A	Power Operator (Q1'23)	SST Guide
Intel On-Demand	8.7 / 9.1	N/A	N/A	N/A

* OOT = Out-of-Tree

Accelerators monitoring tool: pcm-accel (https://github.com/intel/pcm)

intel

Boosting AI Performance with Red Hat® OpenShift® 4.12 on 4th Gen Intel® Xeon® Scalable Processors
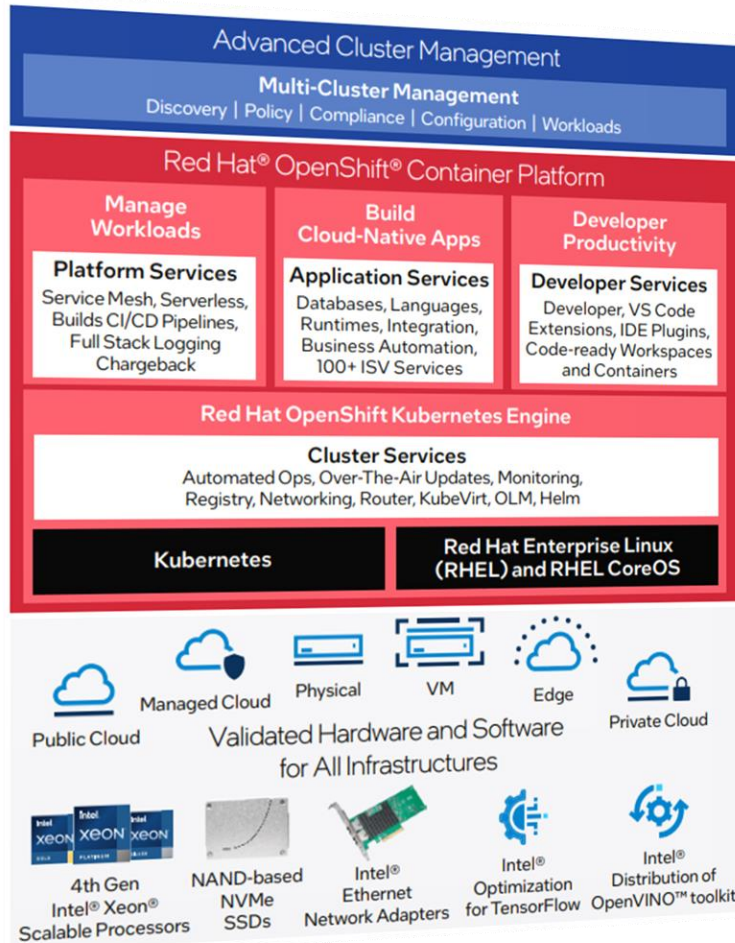
Natural Language Processing:
Smoother Experiences with Faster Responses

Up To 5.7x higher
End-to-End Real-Time Inference Performance Speedup[2]

Up To 6.2x higher
Real-Time NLP Inference Performance[3]

Recommendation Systems:
Recommendations in Real Time

Up To 6.3x higher
Batch Recommendation System Inference Performance[4]

Up To 4x higher
Recommendation System Training Performance[4]

UPDATE: 5th Gen Xeon®

Accelerate Red Hat OpenShift AI Workflows using 5th Gen Xeon® Features with Red Hat Validated Patterns

4th Gen Intel® Xeon® Scalable Processor & Red Hat Enterprise Linux

QAT

Intel® QuickAssistTechnology (Intel® QAT) accelerated throughput (Gb/s)

Software throughput (Gb/s)

Ratio of hardware to software throughput
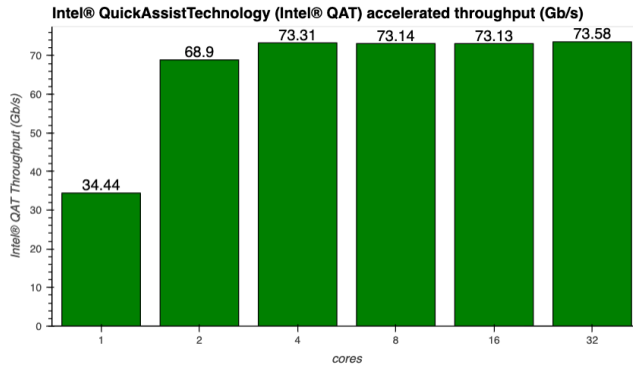
AMX
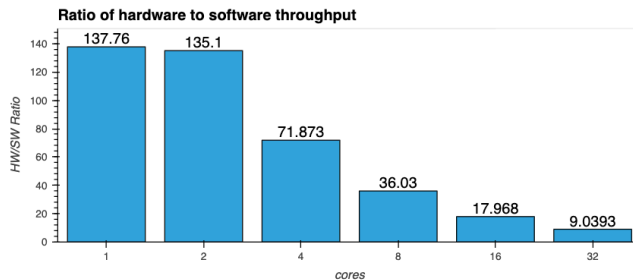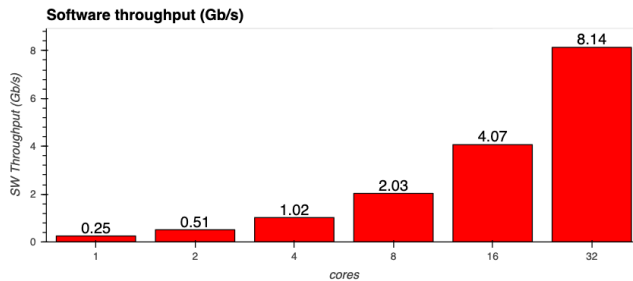
2P Sapphire Rapids Phoronix-Test-Suite speedup factors vs 4P Cooper Lake
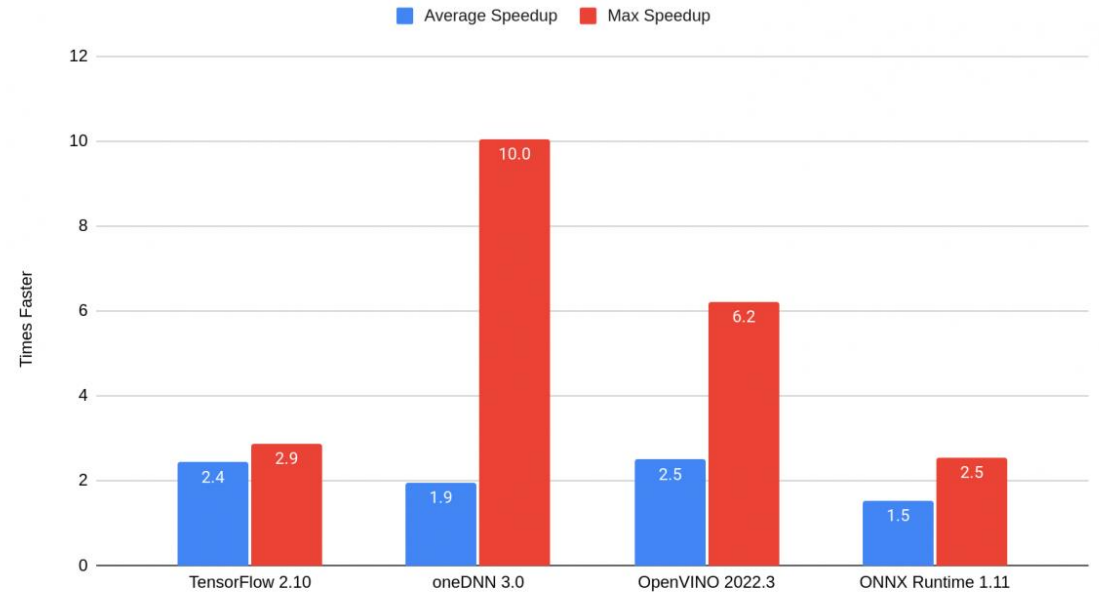
Call to Action

Education

Understand the value of modernizing your customers' data centers on 4th and 5th Gen Intel® Xeon® Scalable processors and how it will reduce operating costs and increase your AI & Security capabilities over older infrastructure

Engagement

Connect with your Intel representative to understand how to leverage Intel's technology portfolio to modernize your customers' data centers

Cloud TV

Intel® Cloud TV explores cloud computing news, trends, and strategies to drive your success

Sapphire Rapids in the Cloud

Sustainability and the Cloud

Modernizing the Hybrid Data Center

5th Gen Intel® Xeon® Scalable Processors Overview

Modernization Opportunities with Microsoft

Intel® Xeon® Processor Advisor Tool Suite

**New: Updated advisors for 4th Generation Intel® Xeon®
Scalable processors are now available!**

Sign Up!
Then choose your deployment environment to begin

On-Prem Advisors

Find the best solutions for your workload whether it's refreshing existing infrastructure or building a new data center. Get instant recommendations and optimize based on TCO, Sustainability and Power.

Cloud Advisors

Discover the best solutions for refreshing or migrating your workload to the cloud. Get instant recommendations optimized for performance and TCO. Find Intel based CSP Instances and pricing.

Modernization
Information and Resources

Asset Type	Title and Link
Infographic	Intel Sustainability Use Case - Server Consolidation
Sales Brief	Intel Sustainability Use Case - AI
Solution Brief	Advance Your Energy Initiatives
Whitepaper	Worker Experiences Redefined with 4th Gen Intel® Xeon® Scalable Processors and New Accelerators Innovate Faster with Integrated AI
Video	Sustainability with Intel technologies
Case Study	Gunpowder Cuts Digital Rendering Time and Cost on New Google Cloud Instances
Performance Index	4th Generation Intel® Xeon® Scalable Processors
Live Webinar	Cloud Solution Architect (CSA) Tech Talk: Reduce TCO and Improve Efficiency with 4th Gen Intel® Xeon® Scalable Processors
Recorded Webinar	Cloud Solution Architect (CSA) Tech Talk: Building Sustainability Practices in the Data Center and Cloud
Recorded Webinar	Cloud Solution Architect (CSA) Tech Talk: Accelerating Critical Workloads with 4th Generation Intel® Xeon® Scalable Processors
Intel® Optimization Hub	Optimizations as Code
Training	In-deck links to Online Tutorials

Resources

Microsoft Data Center Products Refresh
Information and Resources

Asset Type	Title and Link
Microsoft SQL Server 2022	
Tuning Guide	Tuning SQL Server for OLTP
Tuning Guide	Tuning SQL Server for OLAP
Solution Brief	Optimizing Microsoft SQL Server 2022 on Lenovo ThinkSystem SR650 V3
Solution Snapshot	Microsoft SQL Server 2022 on 4th Gen Intel® Xeon® Scalable Processors
Solution Design Brief	Microsoft SQL Server 2022 on Intel® Technologies
Whitepaper	Intel QAT Performance on 4th Gen Intel® Xeon® Processors
Microsoft Azure Stack HCI	
Whitepaper	Unify Operations Across Hybrid and Multi-Cloud Environments
Solution Design Brief	Microsoft Azure Stack HCI on 4th Gen Intel® Xeon® Scalable Processors
Article	MSFT Azure HCI & Arc Wall Street Journal - The Path to Greener IT in a Hybrid Cloud World
Article	MSFT Azure HCI & Arc Wall Street Journal - Driving Sustainability for IT Infrastructure
Case Study	Franz Morat Group Gears Up for the Future
Security Assets	Infographic White paper Video animation
Windows Server 2022	
Report	Deploying Windows Server 2022 on Dell PowerEdge Servers

Modernization Opportunities with Microsoft

Modernization Opportunties

Helping customers migrate to:

MICROSOFT WINDOWS SERVER 2022

MICROSOFT SQL SERVER 2022

intel cloud tv

Learn how to capitalize on two critical modernization opportunities for you and your customers with the end of support for Microsoft Windows Server and SQL Server 2012

How to Access Intel® Partner Alliance Customer Support

Intel Virtual Assistant
This Chat Bot, located in the bottom-right corner of each Partner Alliance webpage, provides self-help to most questions or a quick link to a live support agent.

Get Help "Blade"
Submit an online support request.
This link is found on the footer of most pages within the Partner Alliance website.

Get Help

✉ **Request Support**

Contact us anytime to create a support request.
Submit request ›

Partner Alliance "Get Help" page
The Get Help page provides detailed self-help guides on most of the tools and benefits available to Partner Alliance members.

intel.   PRODUCTS   SUPPORT   SOLUTIONS   DEVELOPERS   PARTNERS

Intel® Partner Alliance ⌄ / Get Help

Intel® Partner Alliance Help   Dashboard →

Account & Membership   Technical Tools   Training & Events   Sales & Marketing
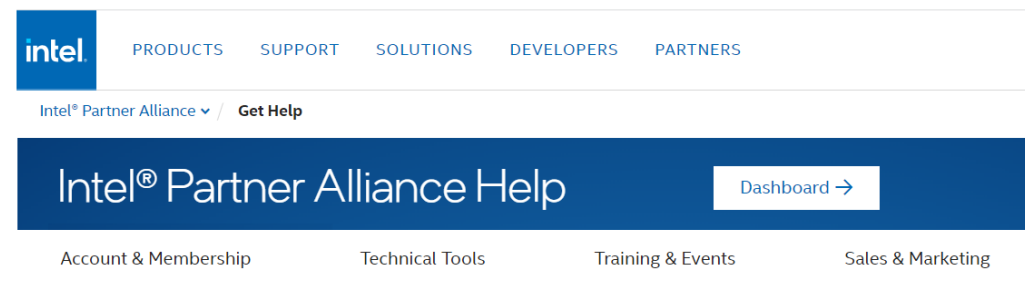
Training

Topic -- Audience
Data Center Sustainability with Intel Data Center Manager DevOps / Cloud Architects
Corporate Sustainability: A Blueprint for Reducing Carbon Emissions C-Suite
One Intel: Introduction to Intel Sustainability Initiative ALL
How to Reduce Data Center Power Cost with Sustainability Regulations C-Suite / Procurement
Intel's Contribution to Cloud Native DevOps / Cloud Architects
Application Architecture and Development in the Cloud DevOps

Topic -- Audience
AWS: Intel Instances and Affinity Cloud Architects
Azure: Intel Instances and Affinity Cloud Architects
Google Cloud Platform: Intel Instances and Affinity Cloud Architects
Workload Placement Cloud Architects

Backup

Deploy a Cloud-Native Core with Confidence

Intel is at the forefront of virtualizing and cloudifying the 5G core

We work in unison with service providers ensuring our products, like the 4th Gen Intel® Xeon® Scalable processors, maximize performance while unlocking agility and scale

90% Core network virtualization will be reached by 2023

1Tbps Breakthrough 5G UPF performance on a single dual socket server

93% average performance-per-watt improvement using Intel® Infrastructure Power Manager with no impact to performance

InfoBrief

Cloud-Native Architectures

Breakthrough Performance and Power Savings

Using the Intel® Infrastructure Power Manager for 5G Core reference software delivers an average power savings of 30% while maintaining key telco performance metrics by dynamically matching run-time CPU power consumption to traffic

Intel uses 90% renewable electricity[3] for a lower carbon footprint

Architecture and built-in accelerators increased perf/watt on workloads that matter

Integrated AI for 4x perf/watt boost[4] for AI inference workloads - TensorFlow

10x increase in energy efficiency by 2030 for client and server CPUs[3]

Built-in telemetry and power management control per-core frequency and power consumption

Built with circular economy strategies for waste; 5% of total waste to landfill[3]

Solution Brief

Dynamically Tune Intel® CPUs to Maximize Network Energy Efficiency

3rd Gen Intel® Xeon® Scalable Processor with SQL Server 2019 vs 4th Gen Intel ® Xeon® Scalable Processor with SQL Server 2022 on Standard Edition

Average New Orders Per Minute (NOPM)

Higher is better

34.2 %

1140000 (3rd Gen)
1530000 (4th Gen)
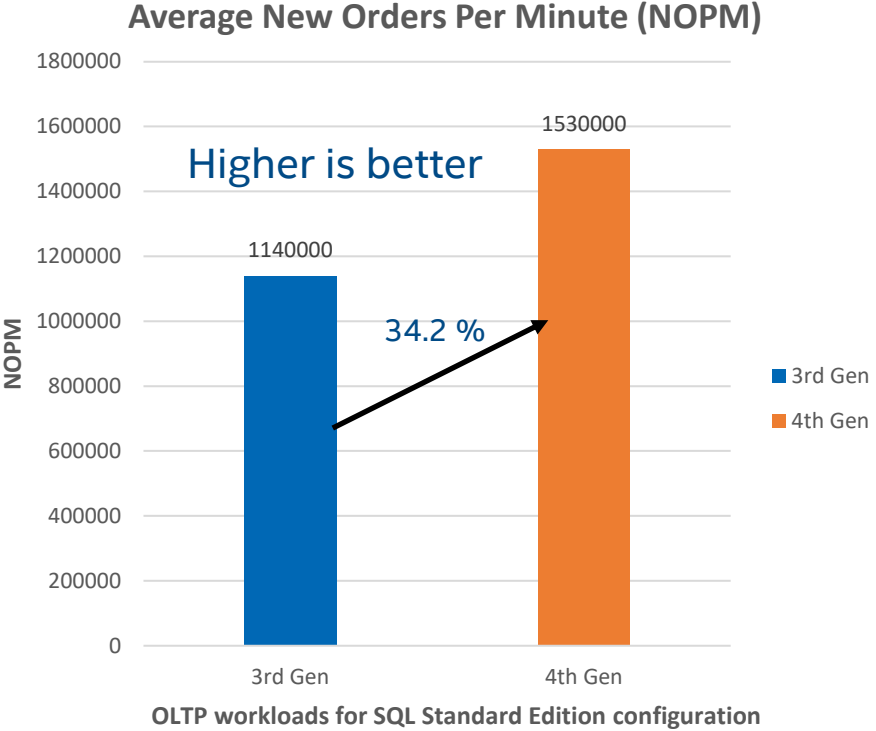
- ■ 3rd Gen
- ■ 4th Gen

OLTP workloads for SQL Standard Edition configuration

Up to 34% more NOPM transactions with 4th Gen Xeon processors over 3rd Gen Xeon processors

Average Query Response Time (mins)

105.24 (3rd Gen)
28.08 %
75.68 (4th Gen)

Lower is better
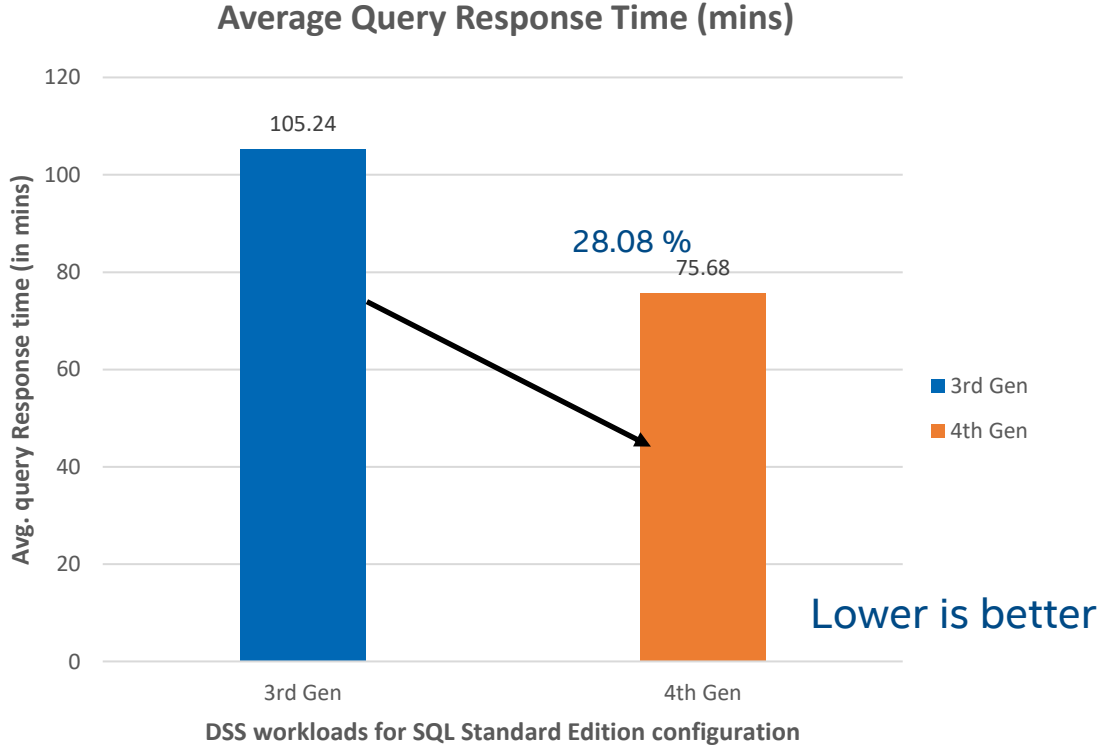
- ■ 3rd Gen
- ■ 4th Gen

DSS workloads for SQL Standard Edition configuration

Up to 28 % faster query response time with 4th Gen Xeon processors over 3rd Gen Xeon processors

3rd Gen Intel® Xeon® Scalable Processor vs 4th Gen Intel® Xeon® Scalable Processor with SQL Server 2022 on Plus Enterprise Configuration

Average New Orders Per Minute (NOPM)
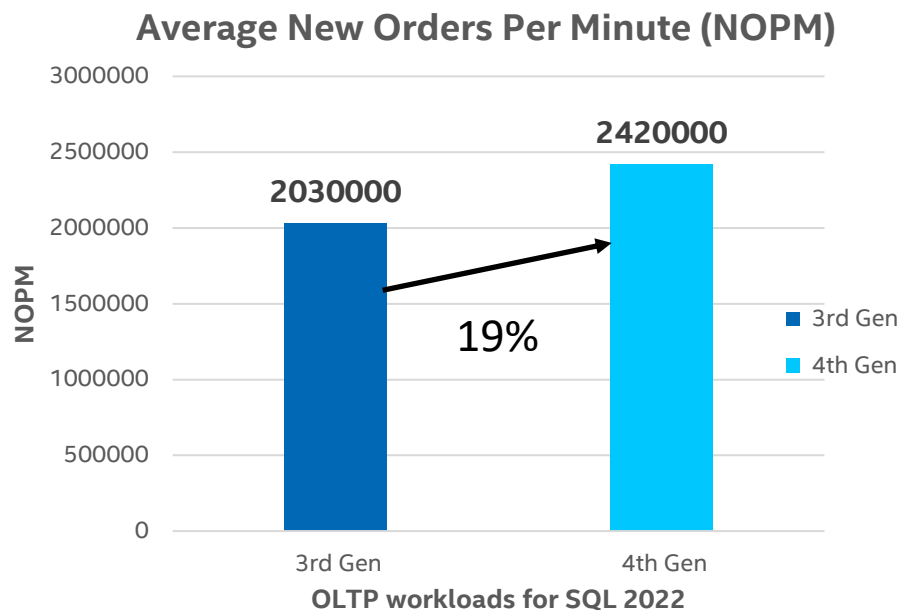
Up to 19% more NOPM transactions with 4th Gen Intel® Xeon® processors over 3rd Gen Intel® Xeon® processors

Average Query Response Time (mins)

Up to 19% faster query response time with 4th Gen Intel® Xeon® processors over 3rd Gen Xeon processors

3rd Gen Intel® Xeon® Scalable Processor with SQL server 2019 vs 4th Gen Intel® Xeon® Scalable Processor with SQL Server 2022 on Base Enterprise Edition

Average New Orders Per Minute (NOPM)

OLTP workloads for SQL Server Base Enterprise Configuration

Average Query Response Time (mins)

DSS workloads for SQL Server Base Enterprise Configuration
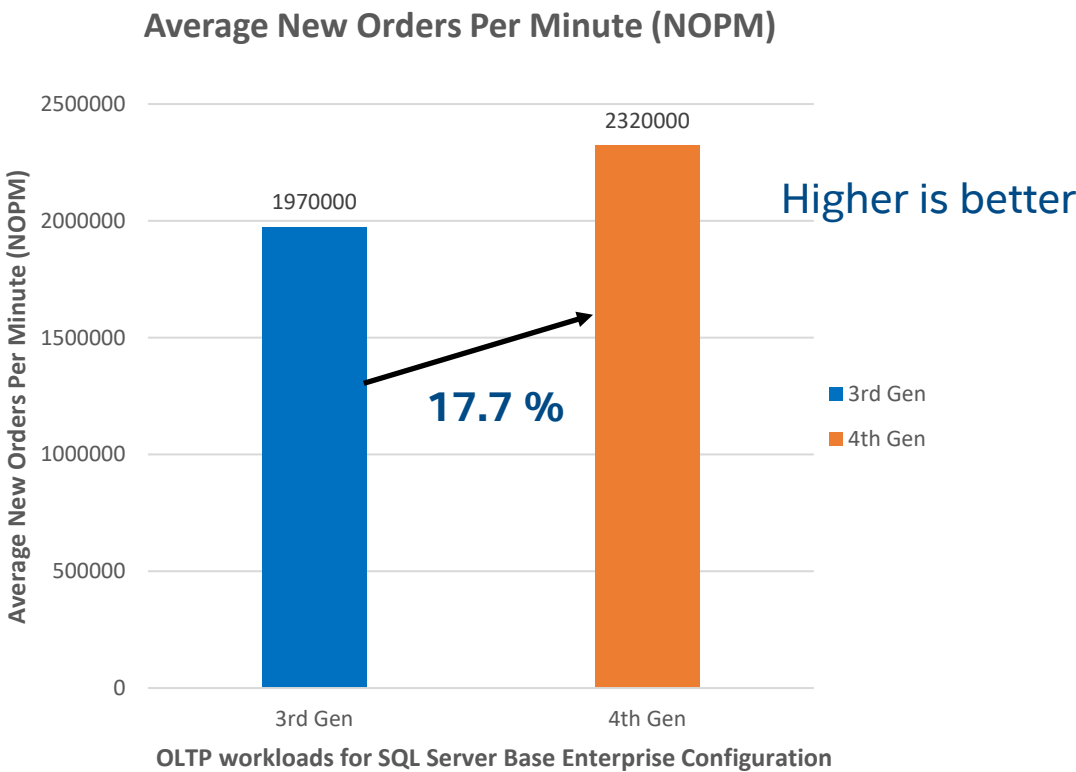
Up to 17.7% more NOPM transactions with 4th Gen Xeon processors over 3rd Gen Xeon processors
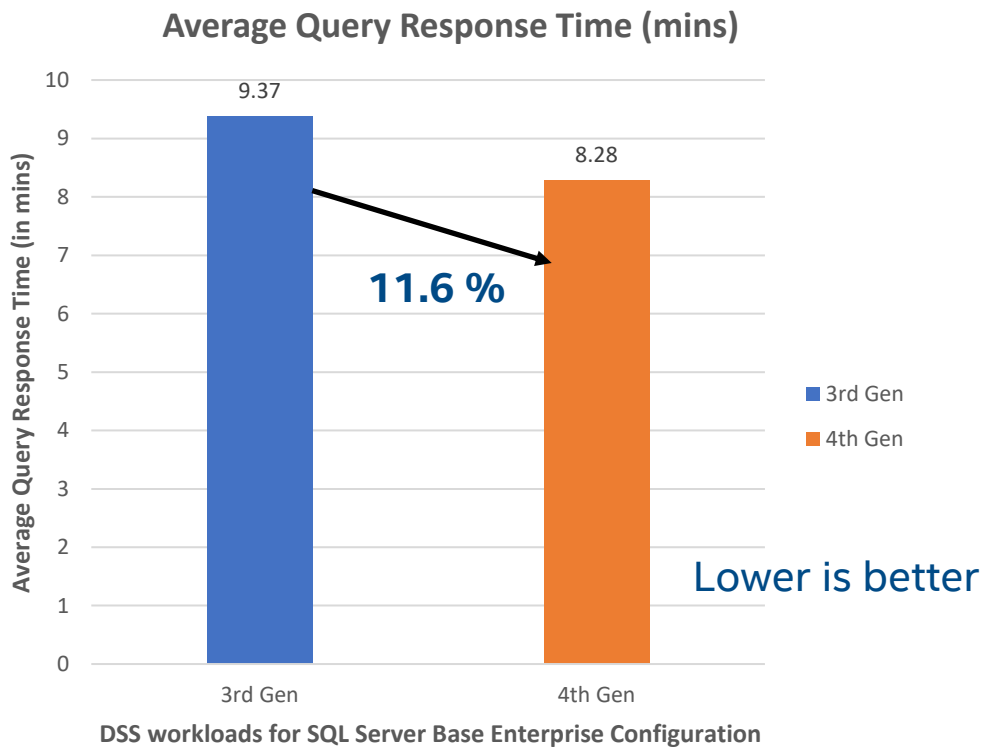
Up to 11.6% faster query response time with 4th Gen Xeon processors over 3rd Gen Xeon processors
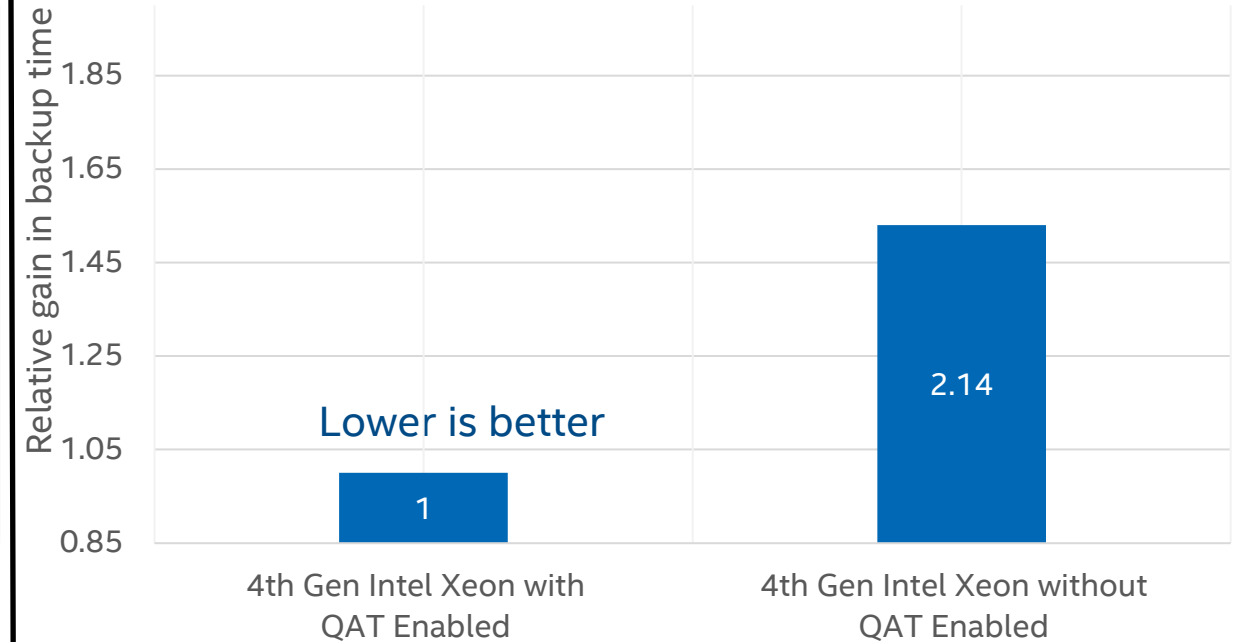
SQL Server 2022 with QAT Performance Comparisons

Intel® 4th Generation Xeon® Scalable processor with QAT vs Intel® 4th Generation Xeon® Scalable processor without QAT enabled
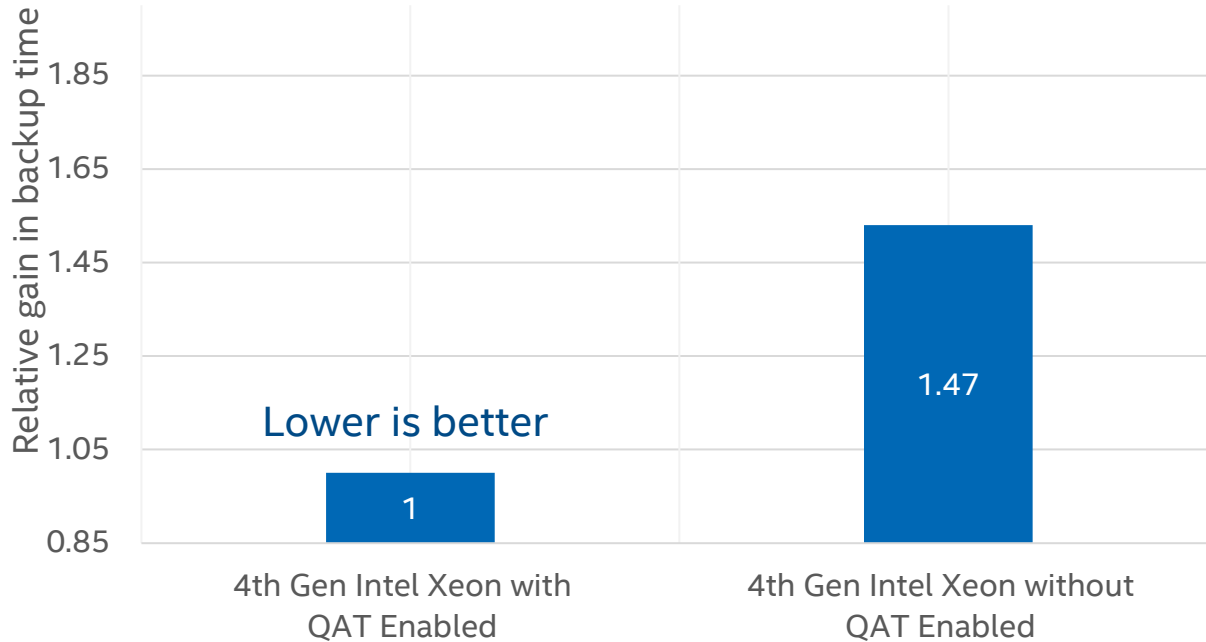
Up to 47% faster backup with Intel 4th Gen Intel® Xeon® Processor and QAT enabled in idle state

Up to 114% faster backup with Intel 4th Generation Intel® Xeon® Processor and QAT under peak load

Tested by Intel as of 12/12/2022. 1-node, 2x Intel® Xeon® Gold 8460Y+ Processor, 32 cores, HT On, Turbo On, Total Memory 512 GB (16 slots/ 32GB/ 4800 MHz [run @ 4800MHz]) DDR4 memory, one QAT device enabled, ucode 0x2B000081, Windows 2022 Standard Edition 21H2, 10.0.20348, SQL Server 2022, 16.0.1000.6 (X64), database backup without QAT using Xpress software compression.

Results may vary.

Configuration Details

4-Node 4th Gen Intel® Xeon® Scalable cluster with Azure Stack HCI: Tested by Intel as of 04/23/2023. 4 Node, 2x Intel® Xeon® Gold 8460Y+, 1x Intel® Server Board M50CYP, Total Memory: 512GB (16 x 32 GB 4800MHz DDR5 DIMM), Intel® Hyper-Threading Technology: Enabled, Turbo: Enabled, Storage (boot): 1 x Solidigm DC S4610 , 960 GB , Storage: 6x Solidigm® D7 P5510 Series (3.84TB) (NVMe), Network devices: 1 x 25 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: 1 GbE, OS/Software: Microsoft Azure Stack HCI build 20385 with SQL Server 2019 Standard Edition

DiskSpd (QD=8,30w:70r): 1,334,067 IOPS @6.51ms(r), @11.22ms(w)

Throughput for OLTP workloads : 30890658 NOPM

2-Node 4th Gen Intel® Xeon® Scalable cluster with Azure Stack HCI: 2 Node, 2x Intel® Xeon® Gold 5416S CPU, 1x Intel® Server Board M50CYP, Total Memory: 512GB (16 x 32 GB 4800MHz DDR5 DIMM), Intel® Hyper-Threading Technology: Enabled, Turbo: Enabled, Storage (boot): 1 x Solidigm DC S4610 , 960 GB , Storage: 4x Solidigm® D7 P5510 Series (3.84TB) (NVMe), Network devices: 1 x 25 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: 1 GbE, OS/Software: Microsoft Azure Stack HCI build 20385 with SQL Server 2019 Standard Edition

DiskSpd (QD=8,30w:70r): 634656 IOPS @2.08ms(r), @5.32ms(w)

Throughput for OLTP workloads : 1634126 NOPM

2-Node 3rd Gen Intel® Xeon® Scalable cluster with Azure Stack HCI: Tested by Intel as of 12/01/2022. 2 Node, 2x Intel® Xeon® Gold 5416S CPU, 1x Intel® Server Board M50CYP, Total Memory: 512GB (16 x 32 GB 4800MHz DDR5 DIMM), Intel® Hyper-Threading Technology: Enabled, Turbo: Enabled, Storage (boot): 1 x Solidigm DC S4610 , 960 GB , Storage: 4x Solidigm® D7 P5510 Series (3.84TB) (NVMe), Network devices: 1 x 25 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: 1 GbE, OS/Software: Microsoft Azure Stack HCI build 20385 with SQL Server 2019 Standard Edition

DiskSpd (QD=8,30w:70r): 713K IOPS @2.55ms(r), @2.97ms(w)

Throughput for OLTP workloads : 1.63M NOPM

Configuration Details

2-Node Azurestack : Tested by Intel as of 28/04/2023. 2 Node, 2x Intel® Xeon® Gold 5416S CPU, 1x Intel® Server Board M50CYP, Total Memory: 512GB (16 x 32 GB 4800MHz DDR5 DIMM), Intel® Hyper-Threading Technology: Enabled, Turbo: Enabled, Storage (boot): 1 x Solidigm DC S4610 , 960 GB , Storage: 4x Solidigm® D7 P5510 Series (3.84TB) (NVMe), Network devices: 1 x 25 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: 1 GbE, OS/Software: Microsoft Azure Stack HCI build 20385 with Ubuntu Server 2022

4-Node Azurestack : Tested by Intel as of 28/04/2023. 4 Node, 2x Intel® Xeon® Platinum 8460Y+ CPU, 1x Intel® Server Board M50CYP, Total Memory: 512GB (16 x 32 GB 4800MHz DDR5 DIMM), Intel® Hyper-Threading Technology: Enabled, Turbo: Enabled, Storage (boot): 1 x Solidigm DC S4610 , 960 GB , Storage: 6x Solidigm® D7 P5510 Series (3.84TB) (NVMe), Network devices: 1 x 25 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: 1 GbE, OS/Software: Microsoft Azure Stack HCI build 20385 with Ubuntu Server 2022

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.

Leadership Performance with 4th Gen Intel® Xeon® processors
Disclaimers

53% average performance gain over the prior generation1 See [G1] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

Up to 10x higher PyTorch real-time inference performance with built-in Intel® Advanced Matrix Extensions (Intel® AMX) (BF16) vs. the prior generation (FP32)4 See [A17] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

Up to 10x higher PyTorch training performance with built-in Intel® Advanced Matrix Extensions (Intel® AMX) (BF16) vs. the prior generation (FP32)5 See [A16] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary

Up to 5:1 consolidation and 75% TCO savings with 4th Gen Intel Xeon processors: Calculations as of March 28, 2023 based on the Intel® Node TCO & Power Calculator using default cost, power and TCO assumptions over a 5-year TCO horizon comparing replacing 50 older servers with Intel Xeon 4110 processors with new servers using new Intel Xeon 5420+ processors. Results may vary. Performance measurements based on published SPECrate®2017_int_base on spec.org as of March 28, 2023
[4110: https://www.spec.org/cpu2017/results/res2020q4/cpu2017-20201015-24218.html 5420+: https://www.spec.org/cpu2017/results/res2023q1/cpu2017-20230130-33925.html]

2.9x average performance per watt efficiency improvement for targeted workloads utilizing built-in accelerators compared to the previous generation2 See [E1] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

Intel SGX is the most researched, updated, and deployed confidential computing technology in data centers on the market today. With Intel® Security Engines, 4th Gen Intel Xeon Scalable processors help bring a zero-trust security strategy to life while unlocking new opportunities for business collaboration and insights—even with sensitive or regulated data. Intel® Software Guard Extensions (Intel® SGX) is designed to enhance data protection at rest, in motion, and in use. Intel SGX is the most researched, updated, and deployed confidential computing technology in data centers on the market today. Intel SGX provides the smallest trust boundary of any confidential computing technology in the data center today.

Built-in accelerators for encryption help keep data protected while preserving performance. Intel® Crypto Acceleration reduces the impact of implementing pervasive data encryption and increases the performance of encryption-sensitive workloads, such as for Secure Sockets Layer (SSL) web servers, 5G infrastructure, and VPNs/firewalls.. Networking Encryption: Up to 47% fewer cores to achieve the same connections/second using integrated Intel® QuickAssist Technology (Intel® QAT) vs. the prior generation on NGINX key handshake.4 See [N15] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

Most deployed platform, backed by extensive testing and validation: With more deployments than any other data center CPU in the market, Intel® Xeon® Scalable processors are widely trusted to run critical workloads at scale. From next-gen memory and I/O to software optimizations, 4th Gen Intel Xeon Scalable processors have been extensively tested and validated to deliver the high performance and reliability organizations demand.

Businesses can speed up time to deployment with the largest ecosystem of partners they know and use—hardware and software vendors and solution integrators around the world build their products on Intel® Xeon® Scalable processors, offering maximum choice and interoperability with the reassurance of thousands of real-world implementations

Refresh and consolidate Intel® Xeon® processor-based servers Disclaimers

Up to 5:1 consolidation with 75% TCO reduction with 4th Gen Intel Xeon processors

Calculations as of March 28, 2023 based on the Intel® Node TCO & Power Calculator using default cost, power and TCO assumptions over a 5-year TCO horizon comparing replacing 50 older servers with Intel Xeon 4110 processors with new servers using new Intel Xeon 5420+ processors. Results may vary. Performance measurements based on published SPECrate®2017_int_base on spec.org as of March 28, 2023
4110: https://www.spec.org/cpu2017/results/res2020q4/cpu2017-20201015-24218.html
5420+: https://www.spec.org/cpu2017/results/res2023q1/cpu2017-20230130-33925.html]

4th Gen Intel® Xeon® processors can significantly lower your total cost of ownership

Calculations as of March 28, 2023 based on the Intel® Node TCO & Power Calculator using default cost, power and TCO assumptions over a 5-year TCO horizon comparing replacing 50 older servers with Intel Xeon 4110 processors with new servers using new Intel Xeon 5420+ processors. Results may vary. Performance measurements based on published SPECrate®2017_int_base on spec.org as of March 28, 2023

8160   https://www.spec.org/cpu2017/results/res2018q4/cpu2017-20181112-09655.html

8460Y   https://www.spec.org/cpu2017/results/res2023q1/cpu2017-20221223-33229.html

6130   https://www.spec.org/cpu2017/results/res2019q2/cpu2017-20190506-13570.html

6430   https://www.spec.org/cpu2017/results/res2023q1/cpu2017-20221223-33187.html

5120   https://www.spec.org/cpu2017/results/res2018q4/cpu2017-20181015-09160.html

5420+   https://www.spec.org/cpu2017/results/res2023q1/cpu2017-20230130-33925.html

4110   https://www.spec.org/cpu2017/results/res2020q4/cpu2017-20201015-24218.html

Vmware Configuration

BASELINE: Intel Xeon Gold 6348 (3rd gen Intel® Xeon): 4-node cluster, Each node: 2x Intel® Xeon® Gold 6348 Processor, 1x Server Board M50CYP2UR, Total Memory 512 GB (16x 32GB DDR4 3200MHz), HyperThreading: Enable, Turbo: Enabled, NUMA noSNC, Intel VMD: Enabled, BIOS:SE5C620.86B.01.01.0006.2207150335 (ucode:0xd000375), Storage (boot): 2x 80 GB Intel SSD P1600X, Storage (cache): 2x 400 GB Intel® Optane™ DC SSD P5800X Series, Storage (capacity): 6x 3.84 TB Intel SSD DC P5510 Series PCIe NVMe , Network devices: 1x Intel Ethernet E810CQDA2 E810-CQDA2, fw 4.0, at 100 GbE RoCE, Network speed: 100 GbE, OS/Software: VMware/vSAN 8.0, 20513097, Test by Intel as of 03/08/2023 using Ubuntu Server 22.04 VM (vHW=20, vmxnet3), vSAN default policy (RAID-1, 2DG), Kernel 5.15, intel-optimized-tensorflow:2.11.0, ResNet50v1.5, Batch size=128, VM=56vCPU+64GBRAM, Multi instance scenario (4 cores per instance), BERT-Large, SQuAD 1.1, Batch size=128, VM=56vCPU+64GBRAM

New Gen: **Intel Xeon Gold 6448Y (4th gen Intel® Xeon):** 4-node cluster, Each node: 2x Intel® Xeon® Gold 6448Y Processor, 1x Server Board M50FCP2SBSTD, Total Memory 512 GB (16x DDR5 32GB 4800MHz), HyperThreading: Enable, Turbo: Enabled, NUMA noSNC, Intel VMD: Enabled, BIOS: SE5C741.86B.01.01.0002.2212220608 (ucode:0x2b000161), Storage (boot): 2x240GB S4520, Storage (data): 6x 3.84 TB Intel SSD DC P5510 Series PCIe NVMe , Network devices: 1x Intel Ethernet E810CQDA2 E810-CQDA2, fw 4.0, at 100 GbE RoCE, Network speed: 100 GbE, OS/Software: VMware/vSAN 8.0, 20513097, Test by Intel as of 03/13/2023 using Ubuntu Server 22.04 VM (vHW=20, vmxnet3), vSAN ESA – Optimal default policy (RAID-5, flat), Kernel 5.15, intel-optimized-tensorflow:2.11.0, ResNet50v1.5, Batch size=128, VM=64vCPU+64GBRAM, Multi instance scenario (4 cores per instance), BERT-Large, SQuAD 1.1, Batch size=128, VM=64vCPU+64GBRAM,

Intel's Vibrant Ecosystem
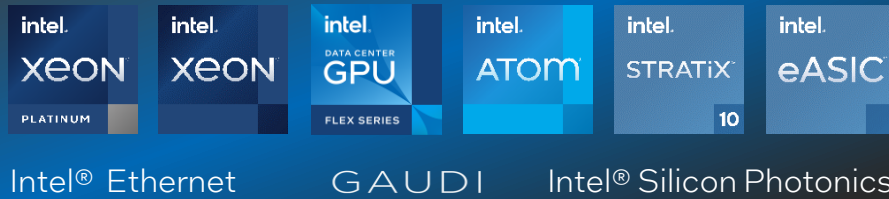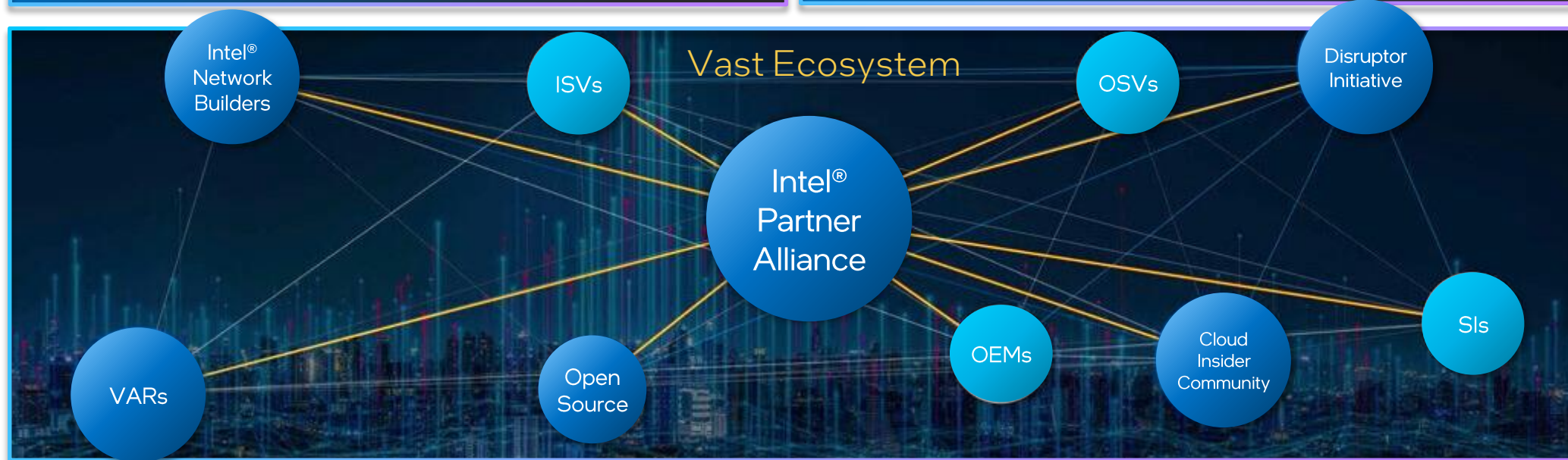
Get Started with Intel® Partner Alliance

Intel Partner Alliance membership gives you exclusive business-building opportunities, like entry to our global marketplace, advanced training, and promotional support – all tailored to your needs

Training and Competencies

Admission to Intel® Partner University provides you with specialized training on advanced technologies, competency programs and rewards for learning

Marketing Resources

Entry to the Intel® Solutions Marketplace and the Intel® Marketing Studio helps you create more demand for your products and services

Valuable Rewards

Earn points for your qualifying activities, advance your membership status and get access to additional resources to build your business

If you're not already a Member
Join Now

Benefits of a Membership

Earn Points

One of the most popular and differentiated benefits within Intel® Partner Alliance are points we award partners to recognize their business results with Intel and their engagement in high priority activities.

There are over 1,000 ways to earn points within Intel Partner Alliance, and 100's of redemption opportunities.

Cloud Insider Community

Intel® Cloud Insider Community offers continuously refreshed, world-class cloud content and tools. Members have the opportunity to connect with peers and the ecosystem to take innovative, joint cloud solutions to market

Learn More

Industry Insights

Gold and Titanium members can access specifically curated quarterly industry insights to help fuel their growth

Learn More

Financial Incentives

Membership unlocks powerful marketing development funds and incentive programs to accelerate your product marketing success

Speak to your Intel Representative to learn about Intel Partner Alliance Accelerator Initiatives and more Financial Incentives