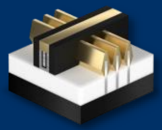
intel xeon Accelerate with Xeon

Accelerating HPC and AI with Intel

Bruno Riva – Intel

intel

We remain confident we will **regain process leadership**

			Angstrom Era	
Intel 7 	Intel 4 	Intel 3 	Intel 20A 	Intel 18A 
Shipping Now	Manufacturing Ready in 2H'22	Manufacturing Ready in 2H'23	Manufacturing Ready in 1H'24	Manufacturing Ready in 2H'24
2022 Milestones	Meteor Lake CPU tile production stepping tape out	Lead server product test wafers running in fab	IP Test Wafers running in Fab	1H'22: Foundry Customer Test Chips 2H'22: First IP shuttle

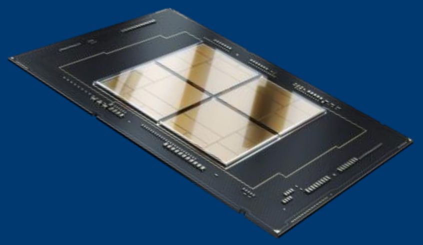
Tick Tock development model enables execution innovation and

5 nodes in 4 years

* Process leadership based on performance per watt

intel.

Expanding the Intel® Xeon® Processor Roadmap



Sapphire Rapids
Intel 7
2022

Emerald Rapids
Intel 7
2023

Granite Rapids
Intel 3
2024

Future Gen
Intel 20A and beyond

P-Core

Perf/core optimized for mainstream & premium cloud and data-center applications

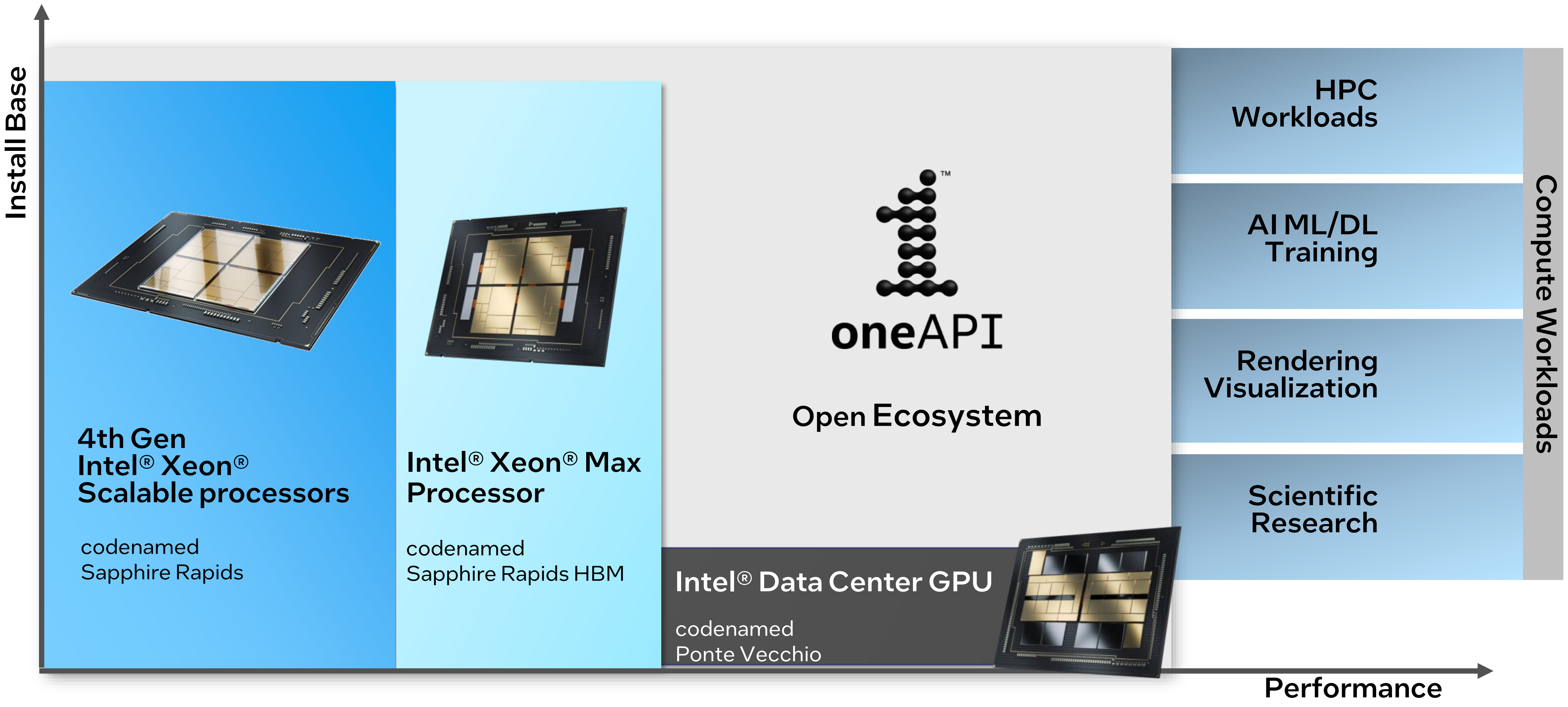
Sierra Forest
Intel 3
2024

Future Gen

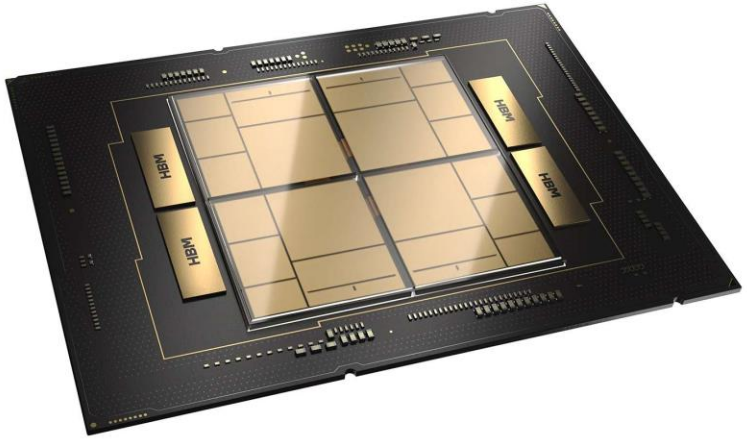
E-Core

Power/perf optimized to support high-density, ultra-efficient compute for the cloud

HPC - AI Super Compute Strategy



Intel® Xeon® CPU Max Series



Designed for HPC, AI, Analytics and other memory bound Workloads

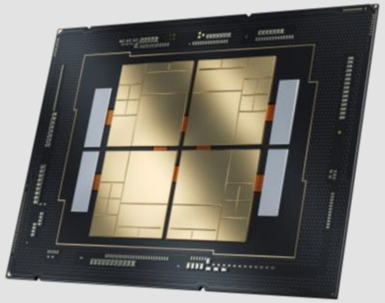
1st x86 CPU to integrate high bandwidth memory and accelerators onto the processor package

Leading performance and efficiency for our customers

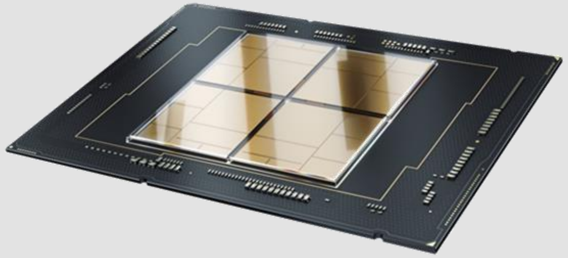
Super Compute Product Portfolio

HPC-AI

CPU
Intel Branded

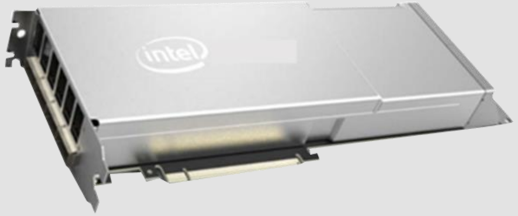


CPU with HBM

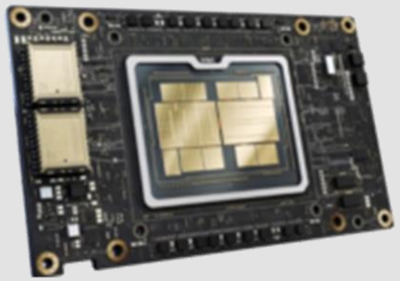


CPUs optimized for HPC

GPU BOARDS
Intel Branded



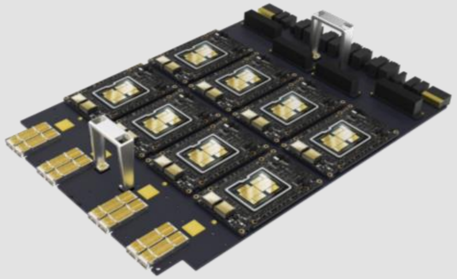
PCIe Add In Cards



OAM Modules

OAM SUBSYSTEMS
Vector Compute Platforms

x4 GPU Subsystem
Vector Compute Platform1.0



x8 GPU Subsystem
Vector Compute Platform1.0

intel®
xeon® Accelerate with Xeon

4th Gen Intel® Xeon® Scalable Processors

intel.

Accelerating Data Center Growth

Delivering Leading Platforms for our Customers and Partners

Innovating for the Future of the Data Center

Continuing to Advance Products and Services

Focus on Customer Real World Workloads

Artificial Intelligence

Networking 5G

Storage

HPC

Data Analytics

Intel's Differentiated Approach

Workload-First

CPU Cores + Built-In Accelerators Wins

Open Software Ecosystem + oneAPI & AI Tools

Higher Performance	Increased Efficiency	Optimal TCO

INTRODUCING

Intel's Most Feature Rich Server Platform

4th Gen Intel® Xeon® Scalable Processors and Intel® Xeon® CPU Max Series Processors

New Integrated IP Accelerators

Increased I/O Bandwidth with PCIe 5.0

Compute Express Link (CXL) 1.1

4th Gen Intel Xeon Scalable Processor

Intel Xeon CPU Max Series

- Intel® In-Memory Analytics Accelerator (Intel® IAA)
- Intel® Data Streaming Accelerator (Intel® DSA)
- Intel® Dynamic Load Balancer (Intel® DLB)
- Intel® QuickAssist Technology (Intel® QAT)
- PCIe 5.0/4.0 80 Lanes
- 1 to 8 Socket Support
- Intel® In-field Scan **
- Intel® Seamless Firmware Update
- High Bandwidth Memory Support

intel

xeon

8 DMI 4.0 LINKS

Intel® C741 Chipset

- Intel® Advanced Matrix Extensions (AMX)
- Accelerating Interfacing Architecture (AiA)
- 5G Intel® Advanced Vector Extensions for vRAN
- Enhanced Intel® Software Guard Extensions
- 8 Channel DDR5
- Up To 4800 MT/s (1DPC)
- Up To 4400 MT/s (2DPC)
- 16 DIMMS per Socket
- New RAS features (Enhanced ECC, Error Check & Scrub)
- Intel® Ultra Path Interconnect (Up to 4 Links @ 16 GT/s)
- Optimized Power Mode

New Integrated ISA Accelerators

HW-Enhanced Security

Increased Memory Bandwidth with DDR5

Increased Inter-Socket Bandwidth with UPI 2.0

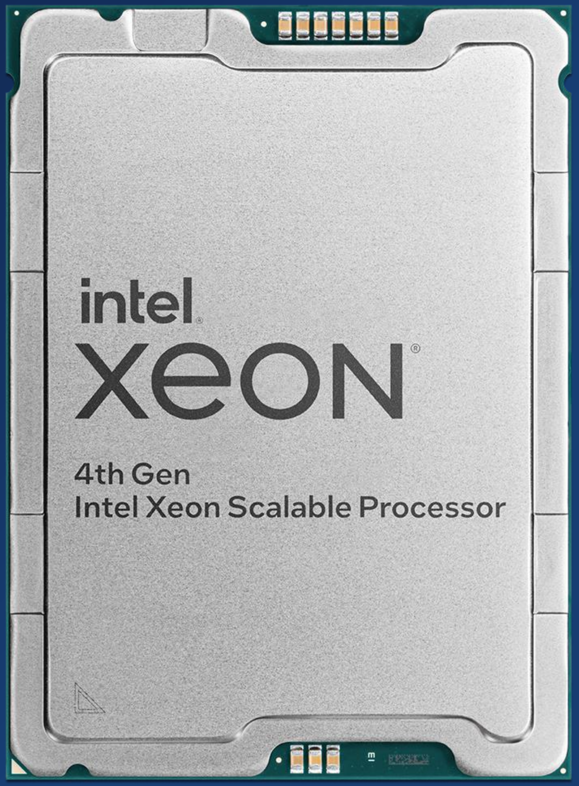
Intel® Optane™ SSDs | **Intel® Agilex™ FPGAs** | **Intel® Data Center GPU Max Series / Intel® Data Center GPU Flex Series** | **Intel® Infrastructure Processing Unit (Intel® IPU)**

intel xeon Accelerate with Xeon

* Enabled for WorkStation + vPro SKUs ONLY

** Intel In-Filed Scan is a new capability available through select providers in 2023

4th Gen Intel® Xeon® Scalable Processors



1 to 8 socket scalability

Up to 60 cores
per processor

Most built-in accelerators of any CPU

Increased memory bandwidth with DDR5

Increased I/O bandwidth with PCIe 5
80 lanes

Increased inter-socket bandwidth with UPI 2.0

Compute Express Link (CXL) 1.1

Hardware enhanced security

Flexibility & Choice for Customers

Most Workload Optimized SKUs on the Market

>56%

Intel® Xeon® Processor Volume

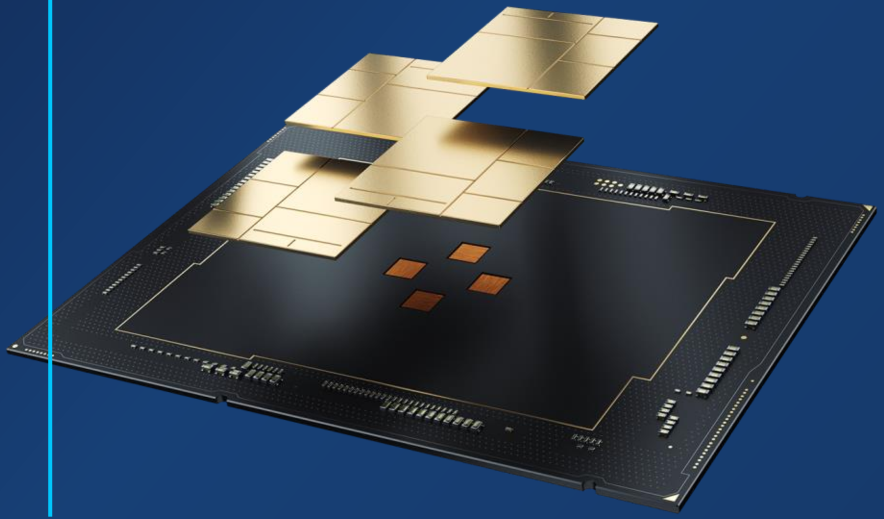
supports customer specific or workload specific demand*

Expanded Options for Workload Optimized SKUs

Cloud (-P, -V, -M)	Network (-N)	Storage (-S)
1-Socket (-U)	Long-Life Use (IOT) (-T)	IMDB Analytics (-H)
HPC (w/HBM)	Liquid Cooled (-Q)	CSP Custom

* Source: Intel Xeon CPU billings on 3rd Gen Intel Xeon Processors, 2022 YTD

4th Gen Intel® Xeon® Scalable Processors



Workload-first approach to innovation, design, and delivery

Most built-in accelerators of any CPU on the market

Leading performance and efficiency for our customers

Industry's most comprehensive Confidential Computing portfolio

Maximize the Effectiveness of Every Core

New Integrated IP Acceleration Engines

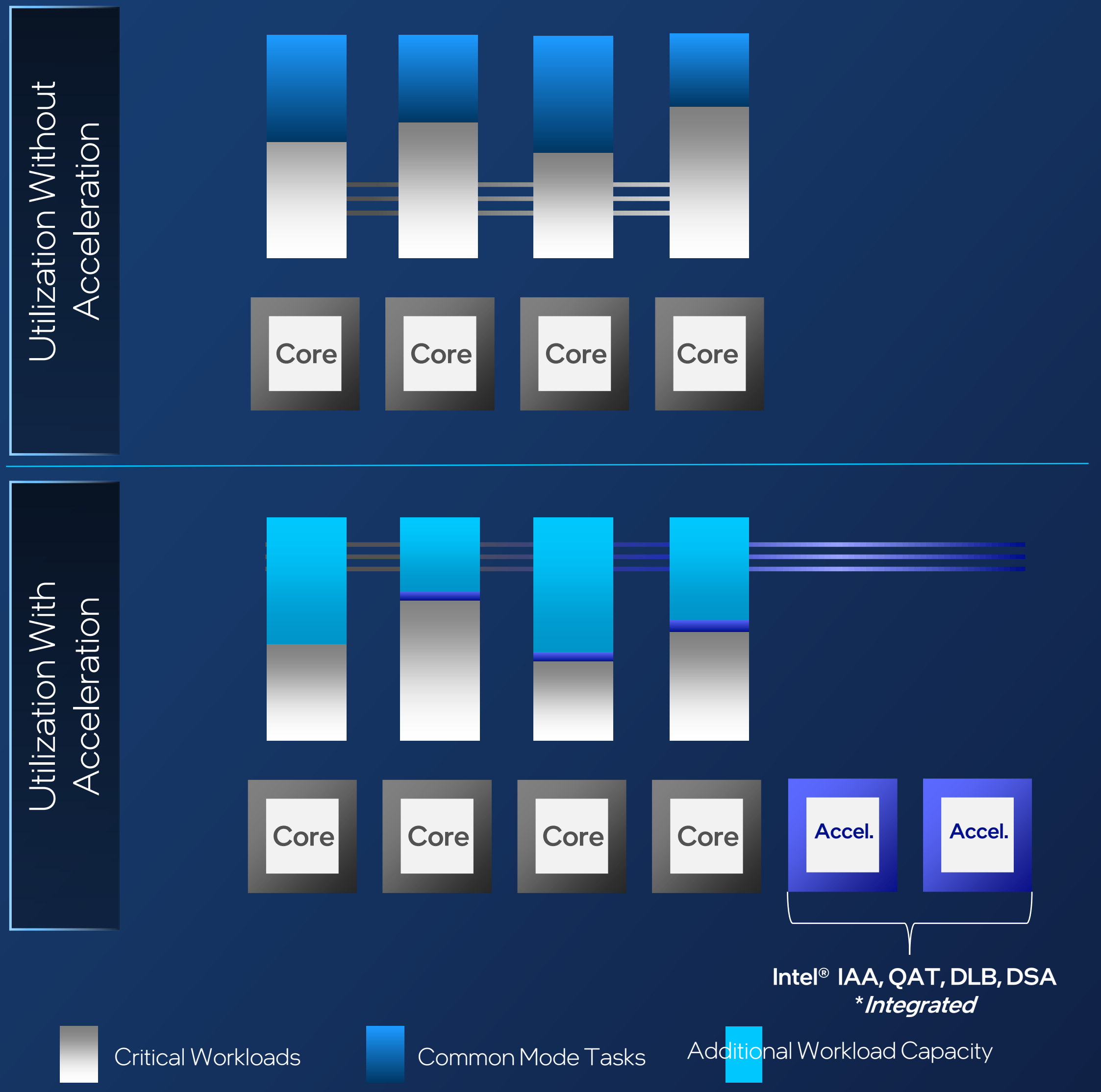
Intel® acceleration engines help free up cores for more general-purpose compute tasks, increasing overall workload performance and power efficiency

Integrated IP

- Intel® QuickAssist Technology (Intel® QAT)
- Intel® Dynamic Load Balancer (Intel® DLB)
- Intel® Data Streaming Accelerator (Intel® DSA)
- Intel® In-Memory Analytics Accelerator (Intel® IAA)

New Instruction Set Architecture (ISA)

- Intel® Advanced Matrix (AMX)
- Intel® Advanced Vector Extensions for vRAN



Intel® Accelerator Engines

Most Built-in Accelerators of any CPU on the market providing customers with increased **performance, costs savings** and **sustainability** advantages for the biggest and fastest-growing workloads

Intel® AI Engines

- Intel® Advanced Matrix Extensions (Intel® AMX)
- Intel® Advanced Vector Extensions 512 (Intel® AVX-512)
- Intel® Deep Learning Boost (Intel® DL Boost)

Intel® Security Engines

- Intel® Control-Flow Enforcement Technology (Intel® CET)
- Intel® Crypto Acceleration
- Intel® Software Guard Extensions (Intel® SGX)
- Intel® Trust Domain Extensions (Intel® TDX)
- Intel® QuickAssist Technology (Intel® QAT)

Intel® HPC Engines

- Intel® Advanced Vector Extensions 512 (Intel® AVX-512)
- Intel® Advanced Matrix Extensions (Intel® AMX)
- Intel® Data Streaming Accelerator (Intel® DSA)
- Intel® QuickAssist Technology (Intel® QAT)

Intel® Network Engines

- Intel® QuickAssist Technology (Intel® QAT)
- Intel® Dynamic Load Balancer (Intel® DLB)
- Intel® Data Streaming Accelerator (Intel® DSA)
- Intel® Advanced Vector Extensions (Intel® AVX) for vRAN
- Intel® Speed Select Technology (Intel® SST)

Intel® Analytics Engines

- Intel® In-memory Analytics Accelerator (Intel® IAA)
- Intel® Data Streaming Accelerator (Intel® DSA)
- Intel® Advanced Vector Extensions 512 (Intel® AVX-512)
- Intel® QuickAssist Technology (Intel® QAT)

Intel® Storage Engines

- Intel® Data Streaming Accelerator (Intel® DSA)
- Intel® QuickAssist Technology (Intel® QAT)
- Intel® In-memory Analytics Accelerator (Intel® IAA)
- Intel® Data Direct I/O (Intel® DDIO)
- Intel® Advanced Vector Extensions 512 (Intel® AVX-512)
- Intel® Crypto Acceleration

Supported in 4th Gen Intel® Xeon® Scalable processors and Intel® Xeon® CPU Max Series processors.
Many features are enabled through Intel® oneAPI Tools and Intel-optimized AI frameworks.

Developer Tools for 4th Gen Intel® Xeon® Scalable Processors

Intel oneAPI, AI tools and optimized AI frameworks help developers maximize application performance by activating advanced capabilities of 4th Gen Intel® Xeon® Scalable processors and Intel® Max Series processors. In multiarchitecture systems with Intel Xeon processors and Intel GPUs, using a single codebase through oneAPI delivers productivity and performance.

Compilers, libraries & analysis tools support built-in accelerators to unleash performance, and fast training and inference for AI workloads.

- **Intel® oneAPI Math Kernel Library**
  for HPC and technical compute
- **Intel® oneAPI Deep Neural Network Library**
  for deep learning training + inference
- **Intel® Query Processing & Intel® Data Mover Library***
  for query processing, compression and data movement
- **Intel® VTune™ Profiler**
  helps locate time-consuming parts of code and identify significant issues affecting application performance

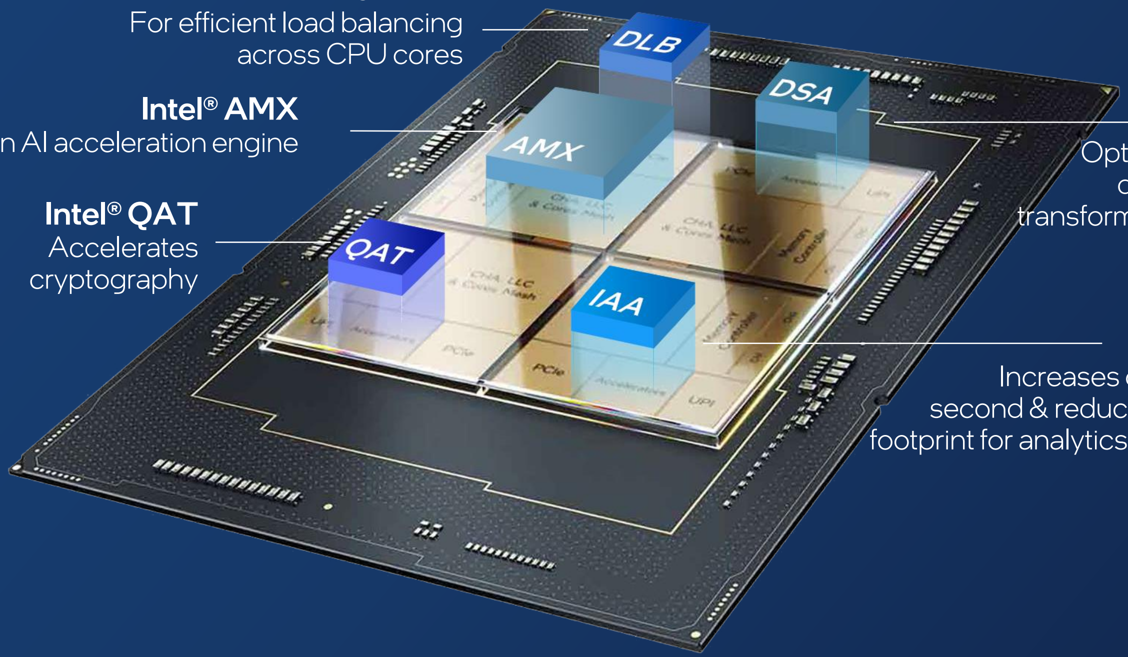
Learn more: Software for 4th Gen Intel Xeon & Max Series Processors

*Intel® QPL is open source. Open source Intel® DML in beta, v1 coming soon

Intel® DLB
For efficient load balancing across CPU cores

Intel® AMX
Built-in AI acceleration engine

Intel® QAT
Accelerates cryptography



Intel® DSA
Optimizes streaming data movement & transformation operations

Intel® IAA
Increases queries per second & reduces memory footprint for analytics workloads

intel 1 oneAPI BASE TOOLKIT
intel 1 oneAPI HPC TOOLKIT
intel 1 oneAPI RENDERING TOOLKIT
intel 1 oneAPI IoT TOOLKIT
intel AI ANALYTICS TOOLKIT
OpenVINO

Powered by oneAPI

Intel® Quick Assist Technology

Acceleration Engine

Function
- Accelerated cryptography and data de/compression

Business Value
- Accelerated compression/decompression offloading leads to greater CPU efficiency
- More encrypted connections and web secure connections between devices with less overhead

Software Support
- Intel® QAT Engine for acceleration of cryptographic operations

Use Cases
- Distributed storage systems, file systems, RocksDB, Data lakes, Apache Spark, Hadoop, NGINX, IPSec

Performance gains vs not using these accelerators

Network Secure Gateway

Up to

84%

fewer cores to achieve same connections/s on NGINX with built-in QAT vs. out-of-the-box software

Performance gains vs prior generation products

Enterprise Storage and Data Analytics

Up to

95%

fewer cores and

2x

higher level 1 compression throughput leveraging integrated QAT vs. prior generation

See [N15,N16] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

Intel® Data Streaming Accelerator

Acceleration Engine

Function
- Optimizing streaming data movement and transformation operations

Business Value
- Accelerated data protection for NVMe/TCP improving efficiency for data storage applications via CPU offload

Software Support
- Intel® Data Mover Library

Use Cases
- Virtualization, fast replication across non-transparent bridge, ERP, In-Memory Databases

Performance gains vs not using these accelerators

Data Integrity (Throughput)

Up to

1.7x

higher IOPs for large packet sequential reads with built-in Intel® DSA vs. ISA-L software

Performance gains vs prior generation products

Data Integrity (Throughput & Latency)

Up to

1.6x

higher IOPs and

37%

latency reduction for large packet sequential reads with built-in Intel® DSA vs. prior generation

See [N18] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

Intel® Dynamic Load Balancer

Acceleration Engine

Function
- Dynamic redistribution of data load across cores when static NIC distribution causes a load-imbalance

Business Value
- Improves system performance related to handling network data on multi-core Intel® Xeon® Scalable processors
- Improved performance for distributed processing, dynamic load balancing and dynamic network processing reordering

Software Support
- Intel® Data Mover Library

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

Performance gains vs not using these accelerators

Microservices

Up to

96%

lower latency at the same throughput with built-in Intel® DLB vs. software for Istio ingress gateway

Performance gains vs prior generation products

Microservices

Up to

89%

lower latency and

57%

lower CPU utilization at same core count with built-in Intel® DLB vs. prior generation

See [W6] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

Intel® Advanced Matrix Extensions

Acceleration Engine

Function
- Provides extensive hardware and software optimizations to enhance AI acceleration

Business Value
- Significant performance increases for AI/Deep Learning inference and training workloads
- Delivers common applications faster through hardware acceleration

Software Support
- Market relevant frameworks, toolkits and libraries (PyTorch, TensorFlow), Intel® oneAPI Deep Neural Network Library (oneDNN)

Use Cases
- Image recognition, recommendation systems, machine/ language translation, NLP, media processing, and delivery

Performance gains vs prior generation products

Speech Recognition Inference

Up to

8.6x

higher speech recognition inference performance with built-in Intel® Advanced Matrix Extensions (Intel® AMX) (BF16) vs. the prior generation (FP32)

Performance gains vs prior generation products

PyTorch Training and Inference

Up to

10x

higher PyTorch for both real-time inference and training performance with built-in Intel® Advanced Matrix Extensions (Intel® AMX) (BF16) vs. the prior generation (FP32)

https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/advanced-matrix-extensions/ai-solution-brief.html

See [A26, A16] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

Intel® In-Memory Advanced Analytics Accelerator

Acceleration Engine

Function

- Integrated accelerator IP accelerating analytics primitives, CRC calculations, compression, and decompression

Business Value

- Increases query throughput for in-memory databases and analytics workloads
- Decreases memory and bandwidth footprint for analytics workloads, freeing up space on CPU

Software Support

- Intel® Query Processing Library, Intel® Data Mover Library

Use Cases

- Commercial in-memory databases, open-source in-memory databases (RocksDB, Redis, Cassandra, MySQL, MongoDB), columnar formats for big data analytics

Performance gains vs not using these accelerators

Embedded Databases

Up to

2.1x

Higher RocksDB performance with built-in Intel® IAA vs. Zstd software

Performance gains vs prior generation products

Embedded Databases

Up to

3x

higher RocksDB performance with

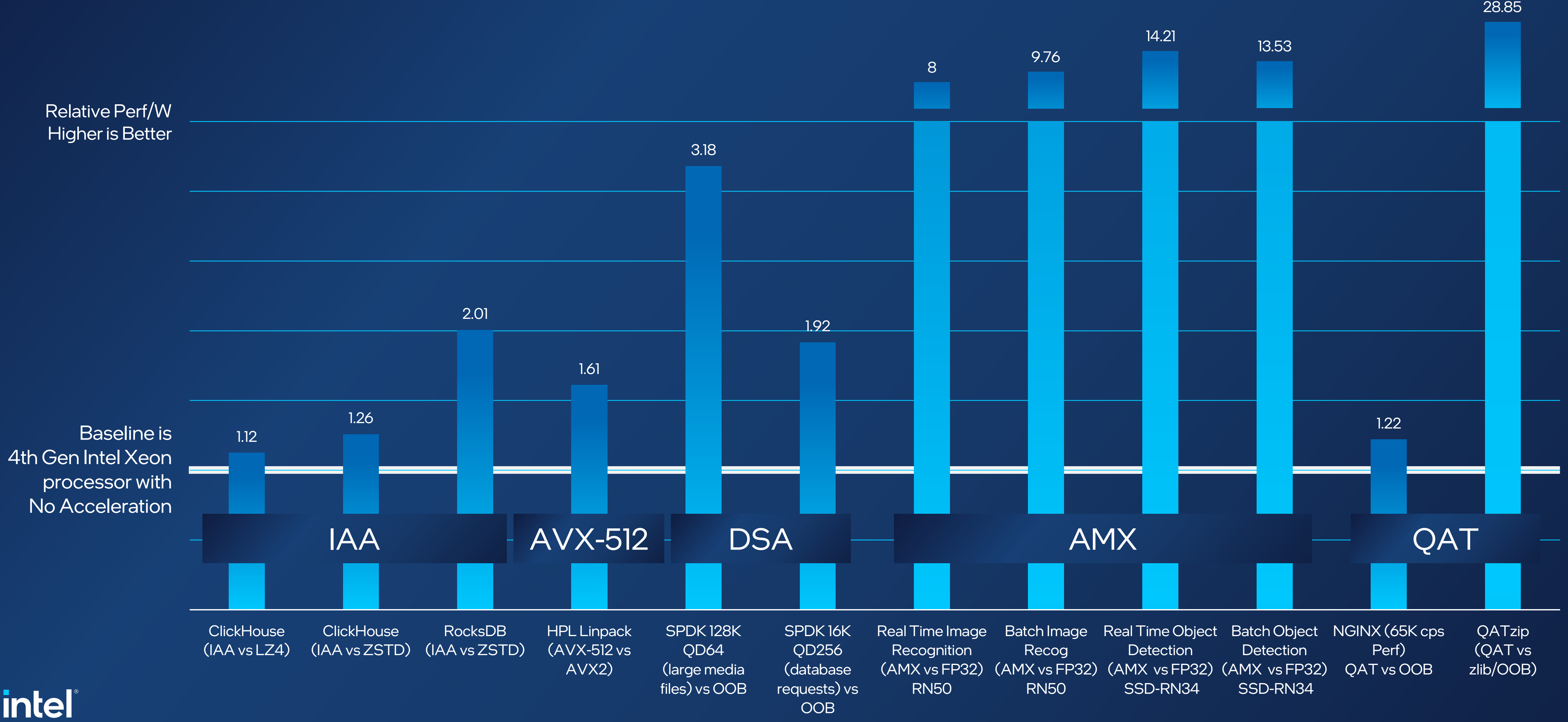
66%

latency reduction with built-in Intel® IAA vs. prior generation

See [D1] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

A More Energy Efficient Server Architecture

Intel® Accelerator Engines Raise Performance Per Watt Ceilings



intel XEON Accelerate with Xeon

See backup for workloads and configurations. Results may vary.

A More Cost-Efficient Server Architecture

Benefits of Workload Optimized Products

When considering new purchases for the data center, deploy fewer 4th Gen Intel® Xeon® processor-based servers or Intel® Xeon® CPU Max processor-based servers to meet the same performance requirement

Comparisons to deploying 50 servers with 3rd Gen Intel Xeon processor	Artificial Intelligence (Real time Inferencing, RSN50 w/ Intel® AMX)	Database (Rocks DB w/Intel® IAA)	Large Media File Requests (SPDK w/Intel® DSA)	HPC (OpenFOAM)
Number of Intel Xeon processor-based servers	17 servers with 4th Gen Intel® Xeon processors	18 servers with 4th Gen Intel® Xeon processors	15 servers With 4th gen Intel® Xeon processors	16 servers with Intel® Xeon® CPU Max Series
Lower Fleet Power (kilowatts)	22.1 kW	15.4 kW	8.6 kW	25.7 kW
Reduced CO2 emissions (kg)*	524,000 kg	366,000 kg	206,577 kg	611,000 kg
TCO savings ($)*	$1.3M	$1.2M	1.4M	$1.5M
	55% Lower TCO	52% Lower TCO	60% Lower TCO	66% Lower TCO

* Estimated over 4 years
See backup for workloads and configurations. Results may vary.

This offering is not approved or endorsed by OpenCFD Limited, producer and distributor of the OpenFOAM software via www.openfoam.com, and owner of the OPENFOAM® and OpenCFD® trademark

CPU + Accelerators: Differentiated Performance On Real Workloads

4th Gen Intel® Xeon® Scalable processors					Intel® Xeon® CPU Max Series
General Purpose Compute	Artificial Intelligence	Network 5G vRAN	Networking & Storage	Data Analytics	HPC
53%	Up to 10x	Up to 2x	Up to 2x	Up to 3x	Up to 3.7x
average performance gain*	higher inference and training performance*	capacity for vRAN workloads at same power envelope*	higher data compression with 95% fewer cores*	higher performance*	on memory-bound workloads**

See [G1, A17, N10, N16, D1] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary

*4th Gen Intel Scalable Processor vs. 3rd Gen Intel Xeon Scalable processors

** Intel Xeon CPU Max Series vs. Intel Xeon 8380

Architected to Accelerate Real World Workloads

Cloud

Up to 89% performance increase with Intel® QAT vs. prior gen.[11]

inspur

"We were pleased to observe a 20% increase in performance over the current generation C2 VMs from Google Cloud in testing with one of our key workloads."[12]

Google Cloud

Security

Intel® SGX performs up to 4.6x higher vs. prior gen.[13]

Fortanix®

AI

Up to 2.48x performance improvement with Intel® AMX vs. prior gen.[14]

Alibaba

Up to 4x performance gain with Intel® AMX vs. prior gen.[15]

BRAINPOOL.AI FUJITSU

"Intel's [4th Gen Xeon processor] provides unprecedented levels of performance for critical graph intelligence tasks."

KATANA GRAPH

5G

"It is not just a software, it is not interfaces, it is not only radio. It is how we can build all the pieces in our architecture."

Telefónica

HPC

Up to 4.3x performance improvement with Intel AMX® on Intel Xeon Max Series vs. prior gen.[16]

CERN openlab SURF

Up to 8.57x performance improvement on Intel Xeon Max Series vs. Intel E5V4.[17]

Los Alamos NATIONAL LABORATORY Hewlett Packard Enterprise

"The reason we use the 4th Gen Intel® Xeon® processor as the building block for immersion born systems is really because of its unrivaled power and efficiency."

Hypertec

Intel® Accelerator Engines in Action

Intel® AMX

Solution Brief

intel xeon

Deep Learning Models
4th Gen Intel Xeon Scalable processors

Intel® Advanced Matrix Extensions (Intel® AMX) Enhances AI Inference Performance for Alibaba Cloud Address Purification

4th Gen Intel Xeon Scalable processors with Intel AMX boost end-to-end inferencing performance 2.48x as compared to the previous generation.[1]

Alibaba Cloud

As an important technique of artificial intelligence (AI), deep learning (DL) has been widely implemented in many areas, such as computer vision (CV), natural language processing (NLP) and recommendation systems. However, with the explosive growth of data and the increasing complexity of DL models, using inferencing in production can be challenging. Users expect to optimize hardware, software, and algorithms to improve performance and reduce overall cost. Optimizing DL inferencing helps users adopt more complex DL models to improve accuracy while maintaining the same latency.

To improve the performance of address-purification services, Alibaba Cloud's machine learning platform (PAI) and the Alibaba Academy for Discovery, Adventure, Momentum and Outlook (DAMO Academy) NLP team collaborated with Intel. 4th Gen Intel® Xeon® Scalable processors, with Intel AMX, along with optimization tools, improved end-to-end inferencing by up to 2.48 times, compared to using a previous-generation platform.[1]

Alibaba Cloud Address Purification

Address purification is the automated process of standardizing, correcting, and validating postal address. It is used in many industries including logistics, e-commerce, retail, and finance. Alibaba Cloud Address Purification is an efficient standard address algorithm as a service (AaaS) developed by the NLP team of Alibaba DAMO Academy based on Alibaba Cloud's enormous address collection.[2] Faster end-to-end performance translates to better business results for Alibaba Cloud's customers. This AaaS is a one-stop, closed-loop service platform for address data processing. It uses the NLP algorithm to correct, complete, normalize, structurize, and label the address data registered in business systems. It supplies more than 20 types of address services[3] and can be deployed on public, private, or hybrid clouds. Alibaba Cloud objectives are:

- Accelerate one-stop performance of the platform with an overall consideration in multiple workloads such as data cleaning and model inference
- Use existing hardware resources efficiently and make full use of customers' server resources in public, private, and hybrid clouds to reduce hardware costs

Solution Brief

intel xeon

Machine Learning
4th Gen Intel® Xeon® Scalable processors

Optimizing Machine Learning (ML) Models with Intel® Advanced Matrix Extensions (Intel® AMX)

Bidirectional Encoder Representations from Transformers (BERT) model throughput shows 2x-3x performance gains with 4th Gen Intel® Xeon® Scalable processors and Intel AMX versus the previous generation[1,2]

In this solution brief, standard BERT models of 12 layers, 768 hidden size, 12 heads, and 128 sequence length (token size) are used as the proxy model for introduction of the fusion optimization methodology.

Overview

Bidirectional Encoder Representations from Transformers (BERT) is a widely used ML model and technique for natural language processing (NLP). BERT has been used to refresh countless records in NLP tasks since its inception. It has also performed extremely well in practical core-bound applications.

For search, machine translation, man-machine interaction, and other NLP tasks, BERT has been widely adopted across multiple user scenarios. Because BERT performance directly affects the user experience with applications and increases the queries per second (QPS) throughput rate, engineers have considered a wide variety of ways to optimize the model to improve its performance.

Tencent StarLake Lab personnel explore advanced cloud computing, artificial intelligence (AI), security, storage, and network technologies to deliver solutions that improve data center performance and reduce the total cost of ownership (TCO) of data centers. The Tencent Machine Learning Platform Department (MLPD) is the heart of the Tencent AI platform, constantly working to drive innovations across Tencent's internet and technology businesses. The MLPD engages in R&D covering a broad range of fields, including computer vision, voice recognition, graph computation, and NLP. Solutions created by the MLPD have been broadly applied to major scenarios in social media, personalized advertising, gaming AI, and content recommendation and search. BERT plays a key role in applications across all these tech sectors.

Intel has closely collaborated with Tencent MLPD and Tencent StarLake laboratory on BERT inference optimization using Intel® AMX, a built-in accelerator for 4th Gen Intel® Xeon® Scalable processors. The teams demonstrated that BERT model throughput [INT8] could increase 2x and BERT model throughput [BF16] could increase 3x when running on systems powered by 4th Gen Intel Xeon Scalable processors using Intel AMX.[1,2] By combining Intel AMX and software optimizations into a powerful unified solution, Tencent aims to evolve its capabilities to deliver a consistent service experience and to optimize TCO.

Alibaba Cloud's machine learning platform (PAI) used 4th Gen Intel® Xeon® Scalable processors, featuring Intel® AMX and optimization tools to improve end-to-end inferencing over a previous generation platform.

Using Intel® AMX, Intel and Tencent demonstrated BERT model throughput gains compared to the previous generation. Now, Tencent can use the optimized BERT model to deliver better service experiences and to help reduce TCO.

Intel's Most Sustainable Data Center Processor Ever

Perf/watt improvements
from the most built-in accelerators ever offered in an Intel processor

New Optimized Power Mode
delivers up to 20 percent power savings with negligible performance impact on select workloads

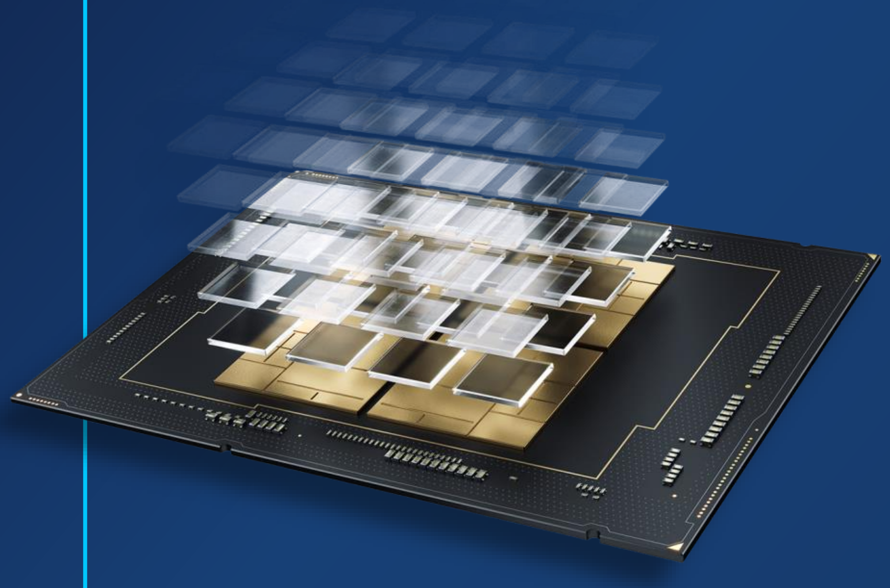
Built-in advanced telemetry
enables monitoring and control of electricity consumption and carbon emissions

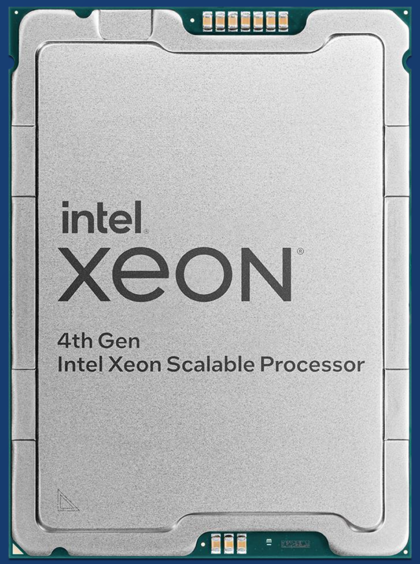
Available immersion cooling warranty rider for Intel® Xeon® processors

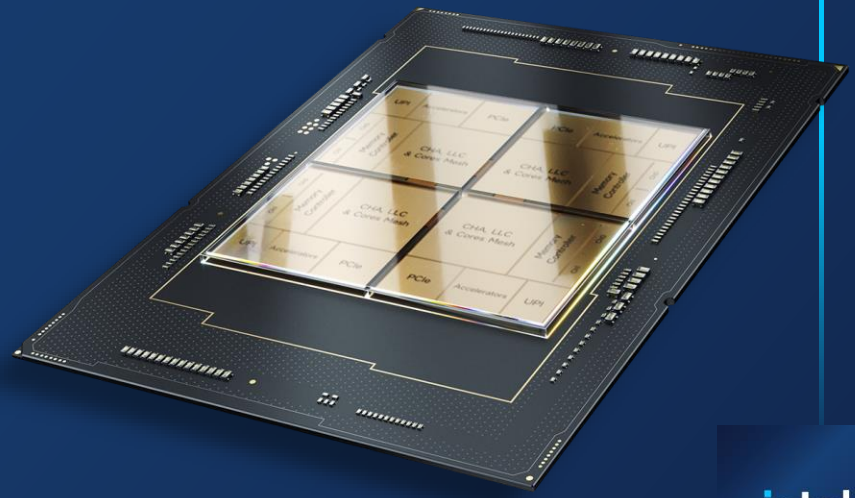
Scope 3 GHG emissions benefits
due to manufacturing with 90-100 percent renewable electricity

Manufactured at sites with state-of-the-art water reclaim
facilities that in 2021 recycled 2.8 billion gallons of water

Thank you!







Learn more

- Intel® Xeon® Scalable Processors
- 4th Gen Intel® Xeon® Scalable Processors
- 4th Gen Intel® Xeon® Scalable Processor product brief
- Intel® Accelerator Engines
- Software for 4th gen Intel Xeon Scalable and Intel® Xeon® Max Series

intel®

intel®

CPU + Accelerators: Groundbreaking Efficiency

Higher Performance per Watt

2.9x

average improvement of perf/watt with built-in accelerators*

Lower Power Bills

up to 70W

power savings per CPU with Optimized Power Mode

Lower TCO More Sustainable

55%

lower TCO and power consumption while reducing 524K kg of CO_2 emissions*

AI Real Time Inferencing workload, ResNet50

See [E1, E6, E7] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary

*for selected workloads.

*4th Gen Intel® Xeon® Scalable Processor vs. 3rd Gen Intel Xeon Scalable processors.

Acceleration Delivers TCO Value

AI real time inferencing	Database	High Performance Computing
55%	52%	66%
vs. prior gen	vs. prior gen	vs. prior gen
lower TCO	lower TCO	lower TCO

intel xeon Accelerate with Xeon

See [E7, E8, E9] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

Sustainable Compute: Optimized Power Mode

up to 20%

lower processor power savings

<5%

performance impact

for select workloads

Lower Power Bills

up to 70W

power savings per CPU with Optimized Power Mode

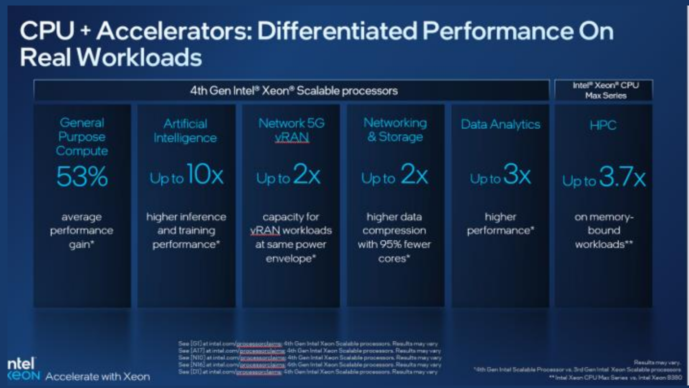
intel. xeon

See [E6] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

Accelerate with Xeon

Resources & Configurations

Resources and Configurations



Architecting to Accelerate Customer Workloads

Leading Performance with the most built – in accelerators

- **Up to 3.7x on memory-bound workloads** - Intel® Xeon® 8380: Test by Intel as of 10/7/2022. 1-node, 2x Intel® Xeon® 8380 CPU, HT On, Turbo On, Total Memory 256 GB (16x16GB 3200MT/s DDR4), BIOS Version SE5C620.86B.01.01.0006.2207150335, ucode revision=0xd000375, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, Stream v5.10; Intel® Xeon® CPU Max Series: Test by Intel as of 9/2/2022. 1-node, 2x Intel® Xeon® CPU Max Series, HT On, Turbo On, SNC4, Total Memory 128 GB (8x16GB HBM2 3200MT/s), BIOS Version SE5C7411.86B.8424.D03.2208100444, ucode revision=0x2c000020, CentOS Stream 8, Linux version 5.19.0-rc6.0712.intel_next.1.x86_64+server, Stream v5.10

Resources and Configurations

Bringing the Architecture to Life (1 of 3)

- Get up to 53% faster results for life and material sciences for more effective research and Meet tight timelines with up to 45% faster results for options pricing
  - DeePMD (Multi-Instance Training)
    8480+: Test by Intel as of 10/12/2022. 1-node, 2x Intel Xeon Platinum 8480+, Total Memory 512 GB, kernel 4.18.0-365.el8_3x86_64, compiler gcc (GCC) 8.5.0 20210514 (Red Hat 8.5.0-10), https://github.com/deepmodeling/deepmd-kit, Tensorflow 2.9, Horovod 0.24.0, oneCCL-2021.5.2, Python 3.9
    8380: Test by Intel as of 10/20/2022. 1-node, 2x Intel Xeon Platinum 8380 processor, Total Memory 256 GB, kernel 4.18.0-372.26.1.el8_6.crt1.x86_64, compiler gcc (GCC) 8.5.0 20210514 (Red Hat 8.5.0-10), https://github.com/deepmodeling/deepmd-kit, Tensorflow 2.9, Horovod 0.24.0, oneCCL-2021.5.2, Python 3.9
  - LAMMPS
    8480+: Test by Intel as of 9/29/2022. 1-node, 2x Intel Xeon Platinum 8480+, HT On, Turbo On, SNC4, Total Memory 512 GB (16x32GB 4800MT/s, DDR5), BIOS Version SE5C7411.86B.8713.D03.2209091345, ucode revision=0x2b000070, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, LAMMPS v2021-09-29 cmkl:2022.1.0, icc:2021.6.0, impi:2021.6.0, tbb:2021.6.0; threads/core:; Turbo:off; BuildKnobs:-O3 -ip -xCORE-AVX512 -g -debug inline-debug-info -qopt-zmm-usage=high;
    8380: Test by Intel as of 10/11/2022. 1-node, 2x Intel Xeon Platinum 8380 CPU, HT On, Turbo On, NUMA configuration SNC2, Total Memory 256 GB (16x16GB 3200MT/s, Dual-Rank), BIOS Version SE5C620.86B.01.01.0006.2207150335, ucode revision=0xd000375, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, LAMMPS v2021-09-29 cmkl:2022.1.0, icc:2021.6.0, impi:2021.6.0, tbb:2021.6.0; threads/core:; Turbo:on; BuildKnobs:-O3 -ip -xCORE-AVX512 -g -debug inline-debug-info -qopt-zmm-usage=high;
    LAMMPS (Atomic Fluid, Copper, DPD, Liquid_crystal, Polyethylene, Protein, Stillinger-Weber, Tersoff, Water)
  - Quantum Espresso (AUSURF112, Water_EXX)
    8480+: Test by Intel as of 9/2/2022. 1-node, 2x Intel Xeon Platinum 8480+, HT On, Turbo On, Total Memory 512 GB (16x32GB 4800MT/s, Dual-Rank), ucode revision= 0x90000c0, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, Quantum Espresso 7.0, AUSURF112, Water_EXX
    8380: Test by Intel as of 9/30/2022. 1-node, 2x Intel Xeon Platinum 8380 CPU, HT On, Turbo On, Total Memory 256 GB (16x16GB 3200MT/s, Dual-Rank), ucode revision=0xd000375, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, Quantum Espresso 7.0, AUSURF112, Water_EXX
  - VASP(Geomean: CuC, Si, PdO4, PdO4_k221)
    8480+: Test by Intel as of 10/7/2022. 1-node, 2x 4th Gen Intel® Xeon® Platinum 8480+, HT On, Turbo On, SNC4, Total Memory 512 GB (16x32GB 4800MT/s, DDR5), BIOS Version SE5C7411.86B.8713.D03.2209091345, ucode revision=0x2b000070, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, VASP6.3.2
    8380: Test by Intel as of 10/7/2022. 1-node, 2x Intel® Xeon® 8380 CPU, HT On, Turbo On, NUMA configuration SNC2, Total Memory 256 GB (16x16GB 3200MT/s, Dual-Rank), BIOS Version SE5C620.86B.01.01.0006.2207150335, ucode revision=0xd000375, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, VASP6.3.2
  - GROMACS (geomean: benchMEM, benchPEP, benchPEP-h, benchRIB, hecbiosim-3m, hecbiosim-465k, hecbiosim-61k, ion_channel_pme_large, lignocellulose_rf_large, rnase_cubic, stmv, water1.5M_pme_large, water1.5M_rf_large)
    8480+: Test by Intel as of 10/7/2022. 1-node, 2x 4th Gen Intel® Xeon® Scalable Processor, HT On, Turbo On, SNC4, Total Memory 512 GB (16x32GB 4800MT/s, DDR5), BIOS Version SE5C7411.86B.8713.D03.2209091345, ucode revision=0x2b000070, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, GROMACS v2021.4_SP
    8380: Test by Intel as of 10/7/2022. 1-node, 2x Intel® Xeon® 8380 CPU, HT On, Turbo On, NUMA configuration SNC2, Total Memory 256 GB (16x16GB 3200MT/s, Dual-Rank), BIOS Version SE5C620.86B.01.01.0006.2207150335, ucode revision=0xd000375, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, Converge GROMACS v2021.4_SP

Resources and Configurations

Bringing the Architecture to Life (2 of 3)

- Meet tight timelines with up to 45% faster results for options pricing

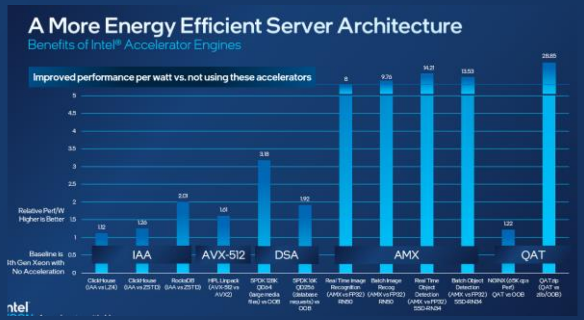
- Binomial Options, Black Scholes, Monte Carlo
  8480+: Test by Intel as of 10/7/2022. 1-node, 2x Intel Xeon Platinum 8480+, HT On, Turbo On, SNC4, Total Memory 512 GB (16x32GB 4800MT/s, DDR5), BIOS Version SE5C7411.86B.8713.D03.2209091345, ucode revision=0x2b000070, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, Binomial Options v1.1, Black Scholes v1.4, Monte Carlo v1.2
  8380: Test by Intel as of 10/7/2022. 1-node, 2x Intel Xeon Platinum 8380 CPU, HT On, Turbo On, Total Memory 256 GB (16x16GB 3200MT/s DDR4), BIOS Version SE5C620.86B.01.01.0006.2207150335, ucode revision=0xd000375, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, Binomial Options v1.1, Black Scholes v1.4, Monte Carlo v1.2

Resources and Configurations

Bringing the Architecture to Life (3 of 3)

- Run social network microservices up to 88% faster for better user experiences.
  - 8480+:4 (1master, 3worker)-node, each-node, pre-production platform with 2x Intel(R) Xeon(R) Platinum 8480+ on QuantaGrid D54Q-2U with GB (16 slots/ 64GB/ DDR5 4800) total memory, ucode 0x2b000081 , HT on, Turbo on, CentOS Linux release 8.4.2105, 6.0.6, 1x 2.9T INTEL SSDPE2KE032T7, 1x 893.8G AVAGO JBOD, 2x Ethernet Controller X710 for 10GBASE-T, 2x Ethernet Controller E810-C for QSFP, DeathStarBench Social Network, wrk2 - load generator, ICE driver (CVL): 6.0.6, Cilium CNI - 1.11.4, Kubernetes - 1.21.14, ContainerD - 1.4.12, deathstarbench/social-network-microservices:0.0.8, nginx-thrift: yg397/openresty-thrift:xenial, memcached:1.6.7, mongo:4.4.6, redis 7.0.5, dataset: DeathStarBench/socialNetwork/datasets/social-graph/socfb-Reed98/ , test by Intel on 11/2/2022. \
  - 8360Y:4 (1master, 3worker)-node, each-node, 2x Intel(R) Xeon(R) Platinum 8360Y on Intel Whitley with GB (16 slots/ 32GB/ DDR4 3200) total memory, ucode 0xd000375, HT on, Turbo on, CentOS Linux release 8.4.2105, 6.0.6, 1x 894.3G INTEL SSDSC2KG96, 2x Ethernet Controller X710 for 10GBASE-T, 1x Ethernet Controller E810-C for QSFP, DeathStarBench Social Network, wrk2 - load generator, ICE driver (CVL): 6.0.6, Cilium CNI - 1.11.4, Kubernetes - 1.21.14, ContainerD - 1.4.12, deathstarbench/social-network-microservices:0.0.8, nginx-thrift: yg397/openresty-thrift:xenial, memcached:1.6.7, mongo:4.4.6, redis 7.0.5, dataset: DeathStarBench/socialNetwork/datasets/social-graph/socfb-Reed98/ , test by Intel on 11/2/2022.
  - https://github.com/delimitrou/DeathStarBench#publications
- Offer personalized product recommendations up to 6.3x faster for smoother e-commerce.
  - 8480+: 1-node, pre-production platform with 2x Intel Xeon Platinum 8480+ on Archer City with 1024 GB (16 slots/ 64GB/ DDR5-4800) total memory, ucode 0x2b0000a1, HT on, Turbo on, CentOS Stream 8, 5.15.0, 1x INTEL SSDSC2KW256G8 (PT)/Samsung SSD 860 EVO 1TB (TF), DLRM, Inf: bs=n [1socket/instance], Inference: bs: fp32=128, amx bf16=128, amx int8=128, Training bs:fp32/amx bf16=32k [1 instance, 1socket], Criteo Terabyte Dataset, Framework: https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66; Modelzoo: https://github.com/IntelAI/models/tree/spr-launch-public, PT:1.13, IPEX: 1.13, OneDNN: v2.7, test by Intel on 10/24/2022.
  - 8380: 1-node, 2x Intel Xeon Platinum 8380 on M50CYP2SBSTD with 1024 GB (16 slots/ 64GB/ DDR4-3200) total memory, ucode 0xd000375, HT on, Turbo on, Ubuntu 22.04 LTS, 5.15.0-27-generic, 1x INTEL SSDSC2KG960G8, DLRM, Inf: bs=n [1socket/instance], Inference: bs: fp32=128, int8=128, Training bs:fp32=32k [1 instance, 1socket], Criteo Terabyte Dataset, Framework: https://github.com/intel-innersource/frameworks.ai.pytorch.private-cpu/tree/d7607bdd983093396a70713344828a989b766a66; Modelzoo: https://github.com/IntelAI/models/tree/spr-launch-public, PT:1.13, IPEX: 1.13, OneDNN: v2.7, test by Intel on 10/24/2022.

Resources and Configurations



A More Energy Efficient Server Architecture

Up to 1.12x and 1.26x higher performance/W using 4th Gen Xeon Scalable w/Intel Analytics Accelerator vs LZ4 and Zstd on ClickHouse
1-node, 2x pre-production 4th Gen Intel Xeon Scalable processor (60 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), Number of IAA device utilized=8(2 sockets active), on pre-production Intel platform and software, HT On, Turbo On, SNC off, Total Memory 1024GB (16x64GB DDR5 4800), microcode 0x2b0000a1, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generic, QPL v0.1.21, accel-config-v3.4.6.4, gcc 11.2, Clickhouse 21.12, Star Schema Benchmark, tested by Intel November 2022.

Up to 2.01x higher performance/W using 4th Gen Xeon Scalable w/Intel Analytics Accelerator vs Zstd on RocksDB
1-node, 2x pre-production 4th Gen Intel Xeon Scalable Processor (60 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), on pre-production Intel platform and software, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 4800), microcode 0x2b0000a1, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generic, QPL v0.2.1,accel-config-v3.4.6.4, ZSTD v1.5.2, RocksDB v6.4.6 (db_bench), tested by Intel November 2022.

Up to 1.61 higher performance/W using 4th Gen Xeon Scalable w/AVX-512 vs AVX2 on Linpack
1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable processor (60 core), on pre-production Supermicro SYS-221H-TNR and software with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000c0, HT On, Turbo On, SNC 4, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, One API BaseKit 2022.2.0.262, One API HPC 2022.2.0.191, Linpack ver 2.3, tested by Intel November 2022.

Up to 3.18x and 1.92x higher performance/W using 4th Gen Xeon Scalable w/Data Streaming Accelerator vs out-of-box OS software on SPDK NVMe TCP
1-node, 2x pre-production 4th Gen Intel Xeon Scalable processor (60 core) with integrated Intel Data Streaming Accelerator (Intel DSA), DSA device utilized=1(1 active socket), on pre-production Intel platform and software with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000a1, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x 1.92TB Intel® SSDSC2KG01, 4x 1.92TB Samsung PM1733, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 2x100GbE, FIO v3.30, SPDK 22.05, tested by Intel November 2022.

Up to 8x and 9.76x higher performance/W using 4th Gen Xeon Scalable w/Advanced Matrix Extensions using AMX vs VNNI instructions on ResNet50 Image Processing
1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable processor (60 core) with Intel® Advanced Matrix Extensions (Intel AMX), on pre-production Supermicro SYS-221H-TNR with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000c0, 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=Resnet 50 v1_5, best scores achieved: BS1 FP32 8 cores/instance (max. 15ms SLA), BS1 INT8 2 cores/instance (max. 15ms SLA), BS1 AMX 1 core/instance (max. 15ms SLA), BS16 FP32 5 cores/instance, BS16 INT8 5 cores/instance, BS16 AMX 5 cores/instance, using physical cores, tested by Intel November 2022.

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.

Up to 1.22x higher performance/W using 4th Gen Intel Xeon Scalable w/QuickAssist Accelerator vs out-of-box software on NGINX TLS Handshake.
QAT Accelerator: 1-node, 2x pre-production 4th Gen Intel Xeon Scalable Processor (60 cores) with integrated Intel QuickAssist Accelerator (Intel QAT), Number of QAT device utilized=4(1 socket active), on pre-production Intel platform and software with DDR5 memory total 1024GB (16x64 GB), microcode 0x2b0000a1, HT On, Turbo Off, SNC Off, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 1x100GbE, QAT engine v0.6.14, QAT v20.l.0.9.1, NGINX 1.20.1, OpenSSL 1.1.1l, IPP crypto v2021_5, IPSec v1.1 , TLS 1.3 AES_128_GCM_SHA256, ECDHE-X25519-RSA2K, 65K CPS target SLA, tested by Intel November 2022. Out of box configuration: 1-node, 2x pre-production 4th Gen Intel Xeon Scalable Processor (60 cores) with integrated Intel QuickAssist Accelerator (Intel QAT), Number of QAT device utilized=0, on pre-production Intel platform and software with DDR5 memory total 1024GB (16x64 GB), microcode 0x2b0000a1, HT On, Turbo Off, SNC Off, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 1x100GbE, NGINX 1.20.1, OpenSSL 1.1.1l, TLS 1.3 AES_128_GCM_SHA256, ECDHE-X25519-RSA2K, 65K CPS target SLA, tested by Intel November 2022.

Up to 28.85x higher performance/W using 4th Gen Intel Xeon Scalable w/QuickAssist Accelerator vs out-of-box zlib on QATzip compression
1-node, 2x pre-production 4th Gen Intel® Xeon Scalable Processor (60 core) with integrated Intel QuickAssist Accelerator (Intel QAT), QAT device utilized=8(2 sockets active), on pre-production Intel platform and software with DDR5 memory Total 1024GB (16x64 GB), microcode 0x2b0000a1, HT On, Turbo Off, SNC Off, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, QAT v20.l.0.9.1 , QATzip v1.0.9, tested by Intel November 2022.

Resources and Configurations

A More Cost-Efficient Server Architecture

ResNet50 Image Classification

New Configuration: 1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable 8490H processor (60 core) with Intel® Advanced Matrix Extensions (Intel AMX), on pre-production SuperMicro SYS-221H-TNR with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000c0, 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=Resnet 50 v1_5, best scores achieved: BS1 AMX 1 core/instance (max. 15ms SLA), using physical cores, tested by Intel November 2022. Baseline: 1-node, 2x production 3rd Gen Intel Xeon Scalable 8380 Processor (40 cores) on SuperMicro SYS-220U-TNR , DDR4 memory total 1024GB (16x64 GB), microcode 0xd000375, 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=Resnet 50 v1_5, best scores achieved: BS1 INT8 2 cores/instance (max. 15ms SLA), using physical cores, tested by Intel November 2022.

For a 50 server fleet of 3rd Gen Xeon 8380 (RN50 w/DLBoost), estimated as of November 2022:
CapEx costs: $1.64M
OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $739.9K
Energy use in kWh (4 year, per server): 44627, PUE 1.6
Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

For a 17 server fleet of 4th Gen Xeon 8490H (RN50 w/AMX), estimated as of November 2022:
CapEx costs: $799.4K
OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $275.3K
Energy use in kWh (4 year, per server): 58581, PUE 1.6
Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

RocksDB
New Configuration: 1-node, 2x pre-production 4th Gen Intel Xeon Scalable 8490H Processor (60 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), on pre-production Intel platform and software, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 4800), microcode 0x2b0000a1, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generic, QPL v0.2.1,accel-config-v3.4.6.4, ZSTD v1.5.2, RocksDB v6.4.6 (db_bench), tested by Intel November 2022. Baseline: 1-node, 2x production 3rd Gen Intel Xeon Scalable 8380 Processor (40 cores) on SuperMicro SYS-220U-TNR , HT On, Turbo On, SNC Off, Total Memory 1024GB (16x64GB DDR4 3200), microcode 0xd000375, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generic, ZSTD v1.5.2, RocksDB v6.4.6 (db_bench), tested by Intel November 2022.

For a 50 server fleet of 3rd Gen Xeon 8380 (RocksDB), estimated as of November 2022:
CapEx costs: $1.64M
OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $677.7K
Energy use in kWh (4 year, per server): 32181, PUE 1.6
Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

For a 18 server fleet of 4th Gen Xeon 8490H (RockDB w/IAA), estimated as of November 2022:
CapEx costs: $846.4K
OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $260.6K
Energy use in kWh (4 year, per server): 41444, PUE 1.6
Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

OpenFOAM
New Configuration: 1-node, 2x pre-production 4th Gen Intel Xeon CPU Max Series (56 cores) on pre-production Intel platform and software, HT On, Turbo On, SNC4 mode, Total Memory 128 GB (8x16GB HBM2 3200MT/s), microcode 0x2c000020, 1x3.5TB INTEL SSDPF2KX038TZ NVMe, CentOS Stream 8, 5.19.0-rc6.0712.intel_next.1.x86_64+server, OpenFOAM 8, Motorbike 20M @ 250 iterations, Motorbike 42M @ 250 iterations, Tools: ifort:2021.6.0, icc:2021.6.0, impi:2021.6.0, tested by Intel December 2022. Baseline: 1-node, 2x production 3rd Gen Intel Xeon Scalable 8380 Processor (40 cores) on SuperMicro SYS-220U-TNR, HT On, Turbo On, 512GB (16x32GB DDR4 3200 MT/s), microcode 0xd000375, 1x2.9TB INTEL SSDPE2KE032T8 NVMe, CentOS Stream 8, 4.18.0-408.el8.x86_64, OpenFOAM 8, Motorbike 20M @ 250 iterations, Motorbike 42M @ 250 iterations, Tools: ifort:2021.6.0, icc:2021.6.0, impi:2021.6.0, tested by Intel December 2022

For a 50 server fleet of 3rd Gen Xeon 8380 (OpenFOAM), estimated as of December 2022:
CapEx costs: $1.50M
OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $780.3K
Energy use in kWh (4 year, per server): 52700, PUE 1.6
Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

For a 16 server fleet of Intel Xeon CPU Max Series 56 core, estimated as of December 2022:
CapEx costs: $507.2K
OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $274.9K
Energy use in kWh (4 year, per server): 74621, PUE 1.6
Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

Resources and Configurations

A More Cost-Efficient Server Architecture

SPDK

New Configuration: 1-node, 2x pre-production 4th Gen Intel Xeon Scalable processor (60 core) with integrated Intel Data Streaming Accelerator (Intel DSA), DSA device utilized=1(1 active socket), on pre-production Intel platform and software with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000a1, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x 1.92TB Intel® SSDSC2KG01, 4x 1.92TB Samsung PM1733, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 2x100GbE, FIO v3.30, SPDK 22.05, tested by Intel November 2022. Baseline: 1-node, 2x production 3rd Gen Intel Xeon Scalable Processors(40 cores) on Supermicro SYS-220U-TNR , DDR4 memory total 1024GB (16x64 GB), HT On, Turbo On, SNC Off, microcode 0xd000375, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x 1.92TB Intel SSDSC2KG01, 4x 1.92TB Samsung PM1733, 1x Intel Ethernet Network Adapter E810-2CQDA2, 2x100GbE, FIO v3.30, SPDK 22.05, tested by Intel November 2022.

For a 50 server fleet of 3rd Gen Xeon 8380 (SPDK), estimated as of November 2022:
CapEx costs: $1.77M
OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $630.6K
Energy use in kWh (4 year, per server): 22762, PUE 1.6
Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

For a 15 server fleet of 4th Gen Xeon 8490H (SPDK w/DSA), estimated as of November 2022:
CapEx costs: $743.8K
OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $220.1K
Energy use in kWh (4 year, per server): 43387, PUE 1.6
Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394