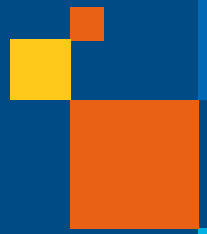
AI on Intel® Xeon® processors

Partner Enablement Package

Addressing customers' AI business challenges with Intel® Xeon® based solutions



Contents

Why Partner With Intel?

At Intel, our goal is to improve lives and outcomes for everyone and every enterprise on this planet

But we aren't doing this alone!

Together with our partners, we are creating real value for our customers by **bringing AI everywhere** and minimizing the risks in AI solution deployment

When you partner with Intel, you partner with a complete AI ecosystem

Our broad portfolio of AI-enabling technologies and collaboration with hardware, software, and solution ecosystem partners delivers real world solutions and differentiated business outcomes for industries, companies, and communities.

Helping you to grow your business.

Intel Leads the Way in AI

More than **300** AI-accelerated ISV features throughout 2024[1]

More than **100M** processors with AI accelerators through 2025[1]

Xeon® install base of **100M+** provisions AI workloads alongside other workloads[2]

Join Us On the Journey to Bring AI Everywhere

[1] https://www.intel.com/content/www/us/en/products/docs/processors/core-ultra/ai-pc-acceleration.html
[2] AI Xeon Sales Card

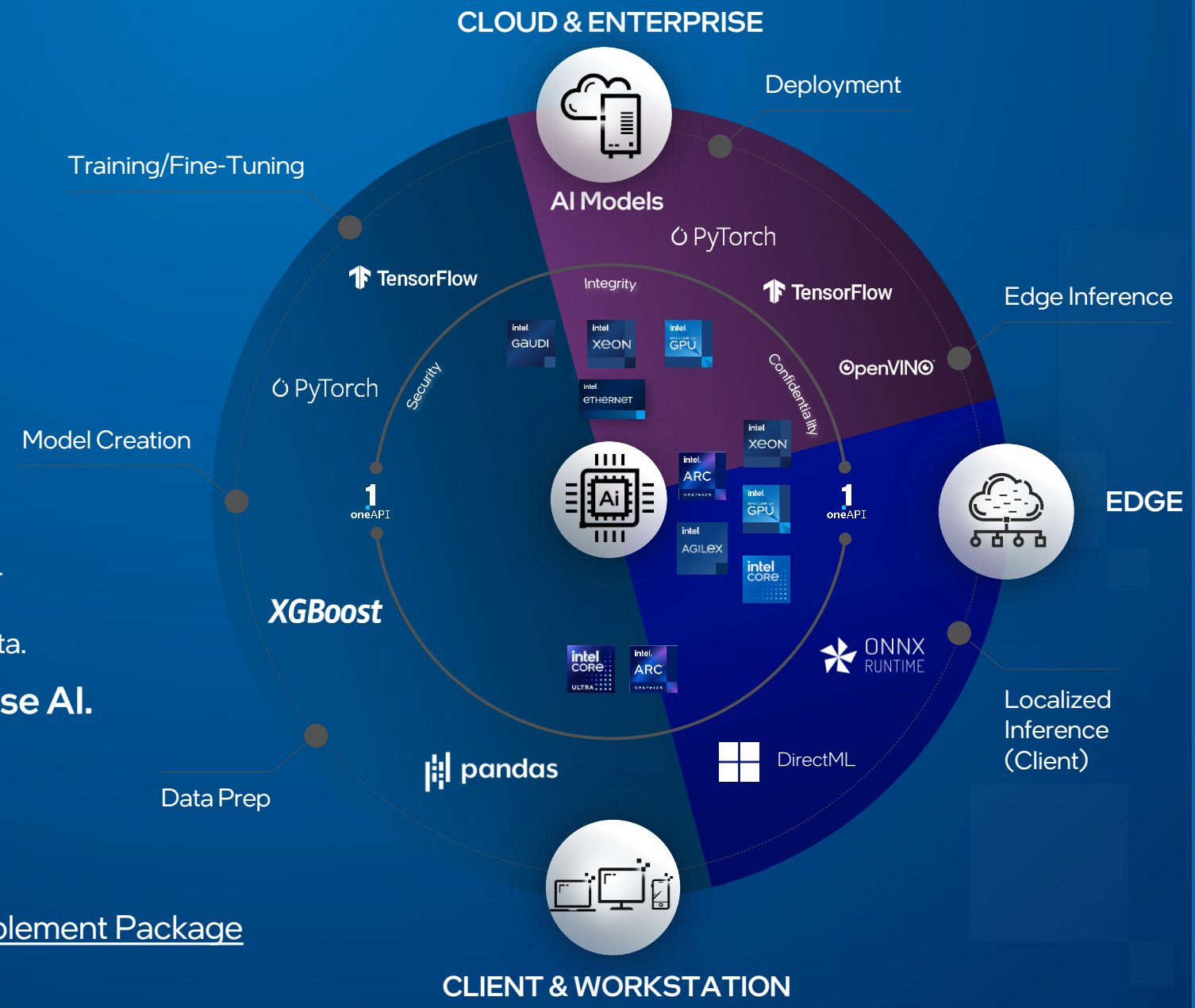
Bringing AI everywhere

In today's hypercompetitive environment, enterprises that embrace AI are pulling ahead.

Intel infrastructure is engineered for enterprise AI, empowering you to maximize your investments and realize your vision at a lower cost. And, with enterprise-ready solutions and open, optimized software, you can go to market fast, even with sensitive and regulated data.

It's time to think differently about enterprise AI.

> Bringing AI Everywhere Infographic



LEARN MORE

- Enterprise AI / Generative AI Partner Enablement Package
- AI Partner Enablement Package

Note: Intel® Core™ Ultra processors integrate NPU low power inference engine from 15th Gen processors onwards.

Why Choose Xeon for Developing an AI Solution

90%

Enterprise Apps will be Infused with AI by 2025[1]

Enterprise AI

100M+

Intel® Xeon Install Base

Intel® Xeon®

100%

of Fortune 500 Global Companies use Enterprise Virtualization[2] Technologies & Services

Virtualization

5th Gen Xeon® with Built-in AI

+

Enterprise ISV Products & Services

Better Together

READ MORE

Bringing AI Everywhere

Business Brief

Deploy AI Everywhere

Business Brief | Video

1 Forbes : https://www.forbes.com/sites/gilpress/2019/11/22/top-artificial-intelligence-ai-predictions-for-2020-from-idc-and-forrester/#4fef9821315a

2 Source: VMware: https://www.vmware.com/files/pdf/VMware-Corporate-Brochure-BR-EN.pdf

Scalable Systems for AI

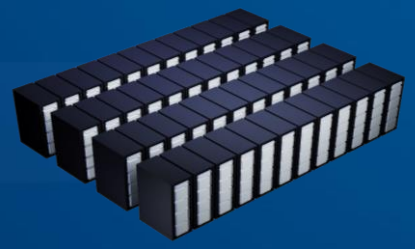
From Cloud & Data Center to the Edge, Intel® Xeon® processors provides optimized performance, scale and efficiency at a cost-effective price

Training and Fine-Tuning

Training | Peak Inference | Mainstream Inference/ Fine-Tuning | Baseline Inference | Endpoint Inference

Inference and Deployment

Cloud Data Center | Edge | Client

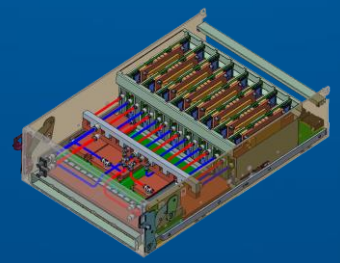


Cluster and Data Center Scale

intel. GAUDI

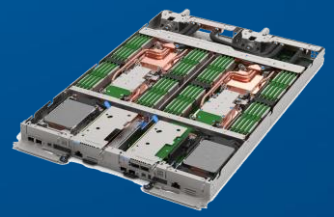
Multi-node Deployment per Rack

intel. GAUDI intel. XEON intel. XEON MAX SERIES



Multi-GPU or Multi-socket CPU

intel. XEON intel. XEON MAX SERIES intel. DATA CENTER GPU FLEX SERIES



Single-Socket CPU or Single GPU

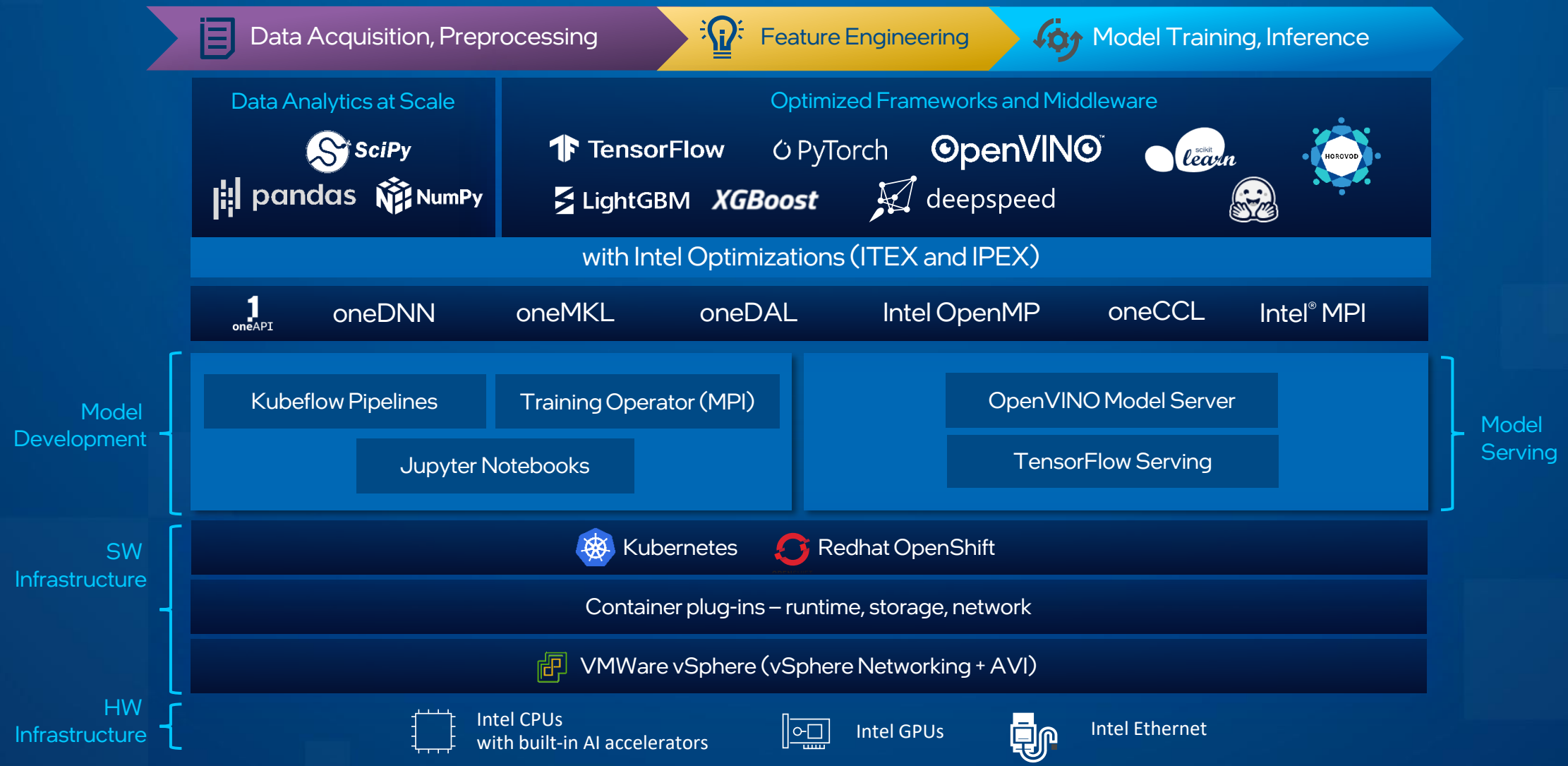
intel. XEON intel. DATA CENTER GPU FLEX SERIES intel CORE ULTRA

Client CPU

intel CORE ULTRA intel. ARC GRAPHICS

intel. ETHERNET

Intel® AI Software is Enterprise Ready



5th Gen Intel® Xeon® Delivers:

Performance

Optimized performance, efficiency, and TCO for the breadth of data center workloads

Leadership CPU AI

Run AI Everywhere with the best CPUs for AI with built-in accelerators, faster memory, and larger last-level cache

Broad Deployments

Designed for efficient performance across all customer deployment models

Intel® Xeon® - The Processor Designed for AI

Efficiently run AI inference

5th Gen Intel® Xeon® processor

The flexibility of Xeon with the built-in DL performance of an AI accelerator

- Up to **29% higher training** and up to **42% higher inference** performance than our previous generation[1]
- Up to **2.69x higher performance than AMD** EPYC 9654 (96C) and 9754 (128C) processors[2]

Build and deploy AI everywhere

Intel AI software suite of optimized open-source frameworks and tools

Enables out of the box AI performance and E2E productivity

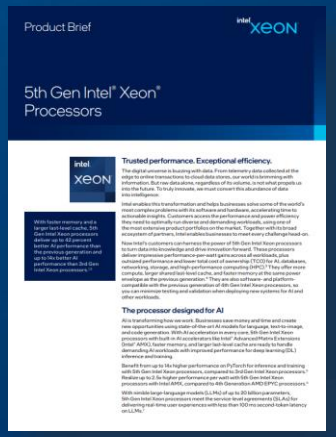
- **5x improvement** on GPT-J in 10 weeks through software optimizations alone[3]
- Optimizing larger models up to **70B parameters** to meet customer SLAs
- Optimized 300+ DL models and 50+ ML and Graph Models

Open Ecosystem

Extensive Intel AI products and partnership

Accelerate end customer time to market

READ MORE

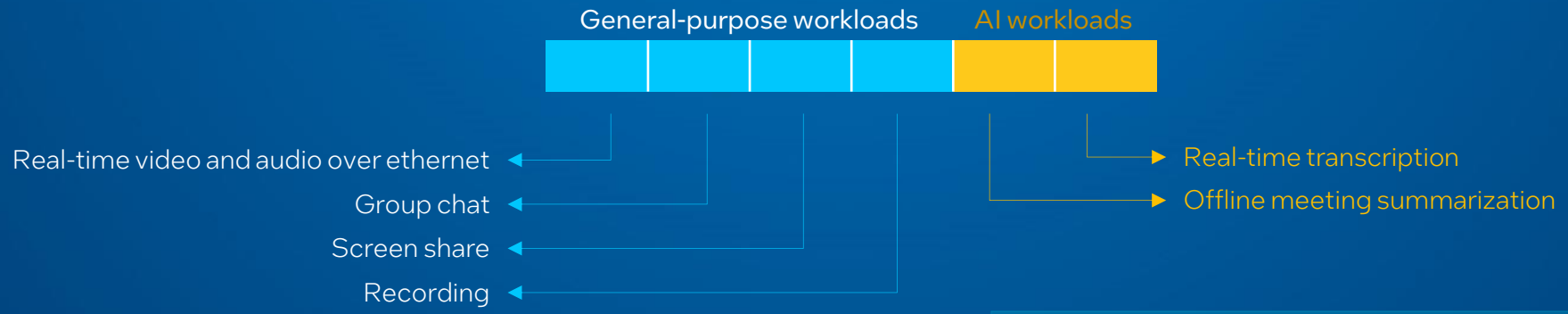


1. Based on performance gains of 1.1x to 1.29x for training (ResNet50v1.5, BERT-Large, SSD-ResNet34, RNN-T, MaskRCNN, and DLRM) and 1.19x to 1.42x for inference (ResNet50v1.5, BERT-Large, SSD-ResNet34, RNN-T (BF16 only), Resnext101 32x16d, MaskRCNN (BF16 only), DistilBERT) compared to 4th Gen Intel® Xeon® processor. See A15-A16 at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary.
2. Based on performance gains of 1.19x to 2.69x with Intel® Advanced Matrix Extensions (Intel® AMX) for inference on GPT-J, LLaMA-2 13B, DLRM, DistilBERT, BERT-Large, and ResNet50v1.5 compared to AMD EYPC 9654 and 9754. See A201, A202, A208-A211 at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary..

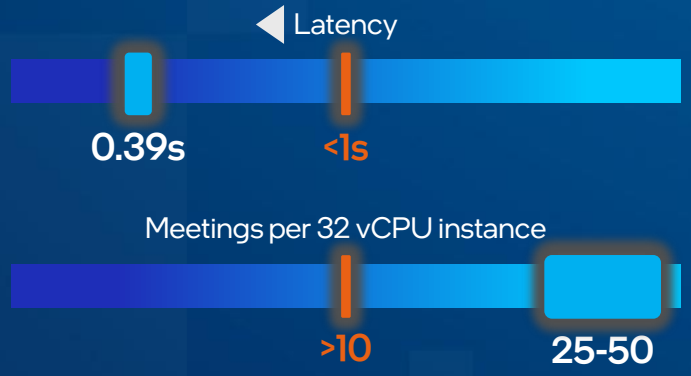
Intel® Xeon® Processor delivers TCO Value for Mixed General-Purpose and AI Workloads

Case Study: Video Conferencing Service

Workloads



Intel performance



Customer SLA

<1 sec latency

>10 Meetings per 32 vCPU instance

Intel engineers collaborating with the customer's technical team were able to beat the latency SLA at **0.39 seconds** and were able to support anywhere between **25-50 meetings on a single 32 vCPU instance**.

See more AI Customer Success Stories on Xeon

Accelerate AI Workloads with Intel® Advanced Matrix Extensions (Intel® AMX)

Intel AMX is a **built-in accelerator** that enables 4th and 5th Gen Intel® Xeon® processors to optimize deep learning (DL) training and inferencing workloads. With Intel® AMX, 4th and 5th Gen Intel® Xeon® processors can quickly pivot between optimizing general computing and AI workloads.

Up to **10x higher performance** and **7x higher performance** per watt for inferencing workloads compared to 3rd Gen Intel Xeon processors[1]

READ MORE

Get Started with Intel® AMX

AI framework optimizations

Tuning Guide

Quick-start Guide

AI reference kits

[1]See linked Solution Brief above for configurations. Results may vary.

USE CASE EXAMPLES

Recommender systems

Deliver a customized end-user experience, whether recommending movies and books or showing targeted ads. Create a DL-based recommender system that accounts for real-time user behavior signals and context features such as time and location.

Natural Language Processing

With a global market projected to reach 80.68 billion USD by 2026,11 NLP applications, including chatbots and sentiment analysis, are critical for businesses to support and scale various functions, including sentiment analysis, chatbots, and machine translation.

Retail e-commerce software solutions

Grow revenue and deliver an exceptional customer experience by minimizing transaction time and effortlessly handling peak demands with DL inference and training, in addition to AI-optimized frameworks like PyTorch and TensorFlow.

Drive Revenue Growth and Improve Customer Experience with Faster, More Effective AI

Leadership performance with the world's best CPU for AI

READ MORE
Leadership Performance

5th Generation Intel® Xeon® processors with Intel® Advanced Matrix Extensions (Intel® AMX) Outperform AMD EPYC[1]

5th Gen Intel Xeon delivers up to

2.2x Higher BERT Large Performance than 4th Gen AMD EPYC[1]

Save up to

$883,000[1]

Reduce TCO Across Your Server Fleet for AI

Up to **41%** Lower TCO than 4th Gen AMD EPYC while running a BERT Large workload[1]

Fewer servers to manage

23 Servers with 5th Gen Intel® Xeon® Scalable processors[1]

vs

50 Servers with 4th Gen AMD EPYC processors[1]

READ MORE
AMD Benchmarks

[1]See linked AMD Benchmarks paper above for configurations. Results may vary.

SOLVE COMMON PROBLEMS

Better inform business decisions to drive revenue growth

Reduce repetitive tasks, costs, and time for your business

Improve customer retention and acquisition

Faster analysis for large amounts of data

Enable more responsive smart assistants and chatbots

Improve text prediction speed and accuracy

5th Gen Intel® Xeon® Outperforms Competition Around The Clock

1.70x on HammerDB MySQL OLTP

2.26x on offline batched image classification inference

2.34x on batched recommendation system inference

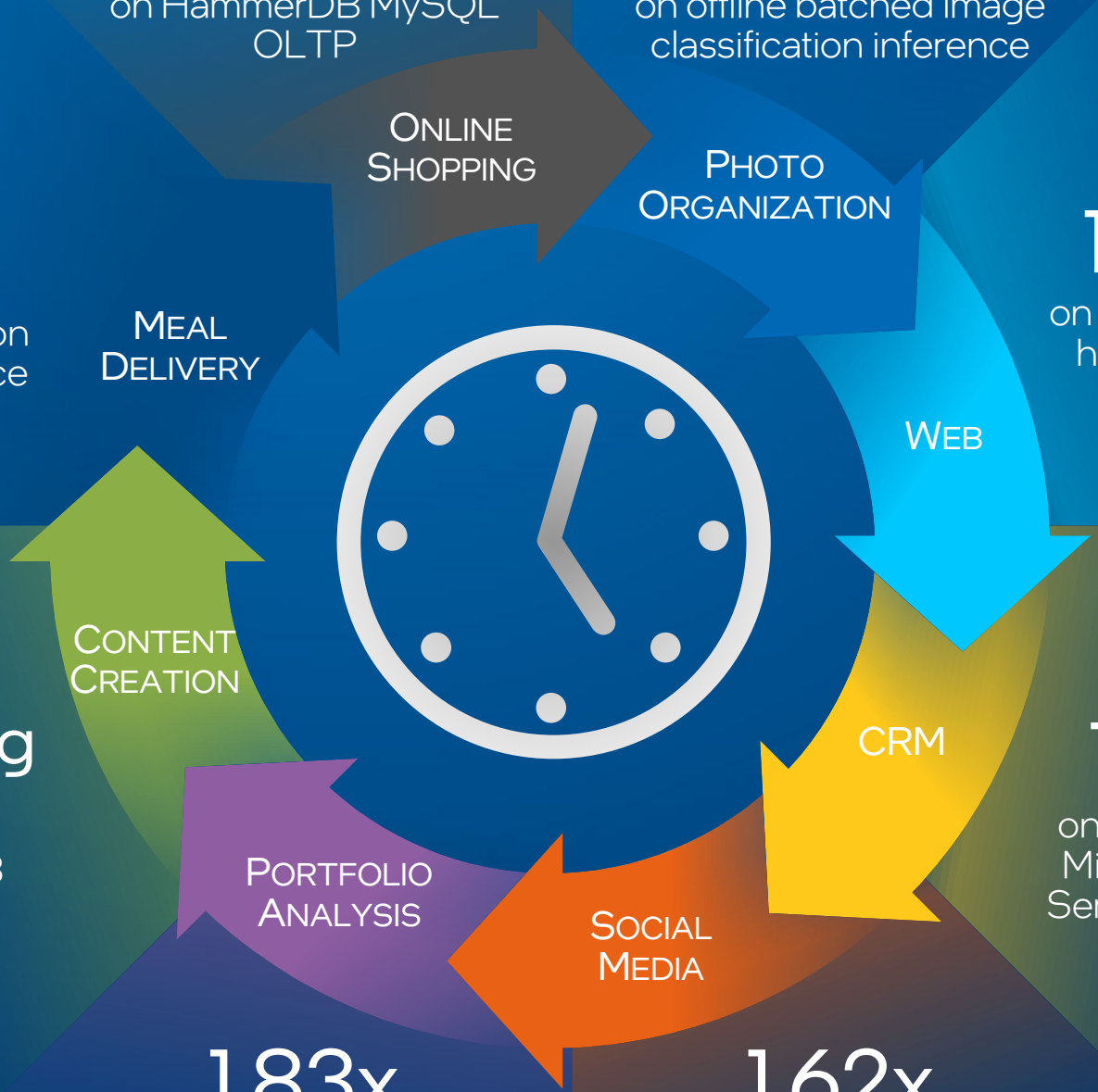
1.66x on NGINX TLS handshakes

Delivering Gen AI on Llama2 13B inferencing

1.93x on HammerDB Microsoft SQL Server + Backup

1.83x on Monte Carlo simulations

1.62x on RocksDB



5th Gen Intel® Xeon® Benchmarks

Compares are relative to 4th Gen EPYC 9654 on AI; relative to 4th Gen EPYC 9554 on all else
See backup for workloads and configurations. Results may vary.

5th Gen Intel® Xeon® TCO advantages over AMD

A comparison against 50 4th Gen AMD EPYC 9554 servers

	Web NGINX TLS	**Data Services** RocksDB	**Data Services** MySQL	**HPC** Monte Carlo	**AI - NLP** DistilBERT
5th Gen Xeon® Servers	31 servers	31 servers	30 servers	28 servers	15 servers
Fleet Energy Saved*	489.7 MWh	1218.1 MWh	684.0 MWh	585.8 MWh	1496.5 MWh
Reduced CO2 Emissions*	207,611 kg	516,402 kg	289,967 kg	248,352 kg	634,428 kg
TCO Savings*	$444K	$471K	$509K	$561K	$1,300K
TCO Delta	21% savings	22% savings	24% savings	27% savings	62% savings

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

AI Case Studies on Intel® Xeon® Processors

Healthcare

Siemens Healthineers is helping practitioners provide **faster, more efficient treatments to patients**, accelerating medical imaging inference time by **35x** while meeting sustainability objectives using Intel® Xeon® CPUs and OpenVINO™ software[1]

SIEMENS Healthineers

READ THE CASE STUDY

Media & Entertainment

Gunpowder accelerated rendering times for **stunning visual effects** while lowering costs with as much as **52% better performance per dollar** compared to previous-gen instances with Intel® Xeon® processors[3]

GUNPOWDER®

READ THE CASE STUDY

Professional Services

Ropers Majeski increased worker productivity by **18.5%**, saving an average of **75 minutes** per user per day by automating email processing, document filing, and report generation with built-in AI acceleration from Intel® Xeon® CPUs[4]

ROPERS MAJESKI

READ THE CASE STUDY

Consulting

BCG and Intel's GenAI solution helps employees **quickly find information relevant to their work and generate business insights**. The solution, powered by Intel® Xeon® CPUs, Intel® Gaudi® accelerators, and hybrid cloud-scale software, delivered a **25%** growth in result relevancy and a **39%** increase in improved work completion rates[2]

BCG

READ MORE

Retail

Meituan uses vision AI services to **improve a wide range of customer experiences**, and achieved **70% cost savings** by migrating from GPUs to Intel® Xeon® CPUs and software for AI inference[5]

美团 Meituan

READ THE CASE STUDY

[1,2,3,4,5] See respective papers and blogs (linked above) for configuration details. Results may vary.

Coming Soon

Intel® Xeon® 6 Processors

Introducing Intel® Xeon® 6 Processors

P-core

Optimized for **performance** in compute-intensive and AI workloads

Common platform foundation and **shared** software stack

E-core

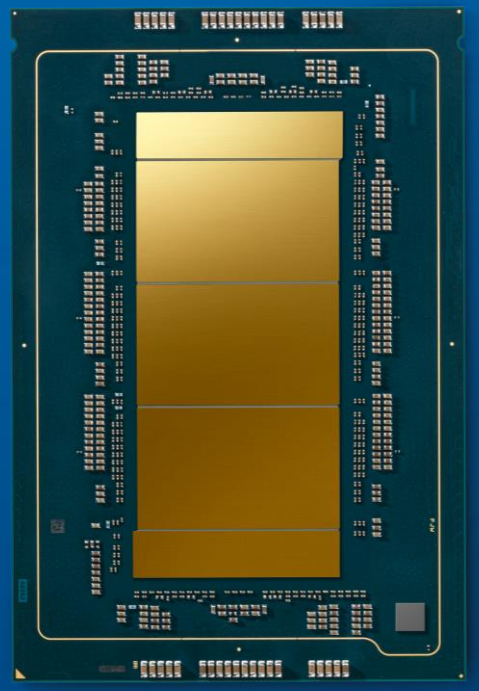
Optimized for **efficiency** in high-density and scale-out workloads

Intel® Xeon® 6 processor with P-cores – Coming Q3 2024

AI | HPC | IaaS | General Compute

Intel® Xeon® 6 processors with P-cores

- Industry-leading Performance-cores (P-cores) are architected for compute-intensive workloads which benefit from multiple data elements being processed in parallel
- Choose from a range of SKUs with up to 128 cores and 12 memory channels for higher overall performance
- Maximize data throughput with the latest DDR5 and Multiplexed Combined Rank (MCR) DIMMs
- Scale AI everywhere with Intel Advanced Matrix Extensions (Intel AMX) to accelerate inferencing for INT8, BF16, and newly supported FP16 datatypes



2x

higher AI inference performance vs. 5th Gen Intel® Xeon® processors[1]

Up to

2.3x

higher HPC performance vs. 5th Gen Intel Xeon processors[1]

2x

higher average performance for general compute vs. 5th Gen Intel Xeon processors[1]

[1]See [9G10, 9H10, 9A10] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.

Call to Action

- Understand your partner / customers' usage model and performance needs FIRST
- When AI is just another workload in a mixed general purpose and AI environment, lead with the Xeon® processors that are already running your customers' business
- For dedicated AI deployments, Xeon® processors paired with Intel® Gaudi® accelerators will deliver the optimal TCO

Access the Intel® Xeon® Processor Advisor Suite to calculate the best route to lower TCO and path to ROI

Additional Resources

Asset Type	Title and Link
Pitch Cards	5th Gen Intel Xeon Pitch Cards PUBLIC SET
Business Brief	Performance. Efficiency. Security. All Across Your Clouds

AI Activation Zones

Digital-first AI workspaces that curate critical resources, tools and benefits - activating partners to build, market, and sell solutions based on Intel technology

Technical Enablement

Sales & Marketing Enablement

Technical Enablement

Sales & Marketing Enablement

Technical Enablement

Sales & Marketing Enablement

Legal Notices and Disclaimers

Notices and Disclaimers.

intel®


intel

AI Use Case Benchmarks on 5th Gen Intel Xeon

Generative AI

GPT-J

- Up to 2.3x performance speedup and 1.81x higher performance/watt with 5th Gen Intel Xeon processor vs 3rd Gen Intel Xeon on GPT-J first token latency (int8)[1]
- Up to 1.64x performance speedup and 1.30x higher performance/watt with 5th Gen Intel Xeon processor vs 3rd Gen Intel Xeon on GPT-J second token latency (int8)[1]

LLaMA

- Up to 2.1x performance speedup and 1.58x higher performance/watt with 5th Gen Intel Xeon processor vs 3rd Gen Intel Xeon on Llama 2 13B first token latency (int8)[2]
- Up to 1.48x performance speedup and 1.11x higher performance/watt with 5th Gen Intel Xeon processor vs 3rd Gen Intel Xeon on Llama 2 13B second token latency (int8)[2]

Recommender Systems

DIEN

- Up to 1.12x higher end-to-end recommendation Speedup with 5th Gen Intel® Xeon® processor compared with prior generation on Deep Interest Evolution Network (DIEN) Recommendation System[5]

DLRM

- Up to 8.7x higher batch Recommendation System inference performance (DLRM) and 6.2x higher performance/watt on 5th Gen Intel Xeon Platinum 8592+ with AMX BF16 vs. 3rd Gen Intel Xeon processor with FP32[6]

DLRM (Competitive)

- Up to 2.34x higher batched Recommendation System inference performance (DLRM) and 2.26x higher performance/watt on 5th Gen Intel Xeon Platinum 8592+ with AMX INT8 vs. AMD EPYC 9654 (Genoa)[7]
- Up to 1.90x higher batched Recommendation System inference performance (DLRM) and 1.72x higher performance/watt on 5th Gen Intel Xeon Platinum 8592+ with AMX INT8 vs. AMD EPYC 9754 (Bergamo)[7]

Natural Language Processing

BERT- Large

- Up to 9.9x higher real-time Natural Language Processing inference (BERT-large) performance and 7.7x higher performance/watt on 5th Gen Intel Xeon Platinum 8592+ with AMX BF16 vs. 3rd gen Intel Xeon processor with FP32[3]
- Up to 10x higher batch Natural Language Processing inference (BERT-large) performance and 7.1x higher performance/watt on 5th Gen Intel Xeon Platinum 8592+ with AMX BF16 vs. 3rd gen Intel Xeon processor with FP32[3]

DistilBERT

- Up to 7x higher real-time Natural Language Processing inference (DistilBERT) performance and 5.6x higher performance/watt on 5th Gen Intel Xeon Platinum 8592+ with AMX BF16 vs. 3rd gen Intel Xeon processor with FP32[4]
- Up to 10x higher batch Natural Language Processing inference (DistilBERT) performance and 8x higher performance/watt on 5th Gen Intel Xeon Platinum 8592+ with AMX BF16 vs. 3rd gen Intel Xeon processor with FP32[4]

Computer Vision

ResNet-50

- Up to 8x higher performance and 6.4x higher performance/watt on real-time image classification(ResNet50) inference (BF16) with 5th Gen Intel Xeon 8592+ (64c) vs. 3rd Gen Intel Xeon 8380 (fp32)[8]
- Up to 8.8x higher performance and 7.4x higher performance/watt on batch image classification(ResNet50) inference (BF16) with 5th Gen Intel Xeon 8592+ (64c) vs. 3rd Gen Intel Xeon 8380 (fp32)[8]

ResNet-34

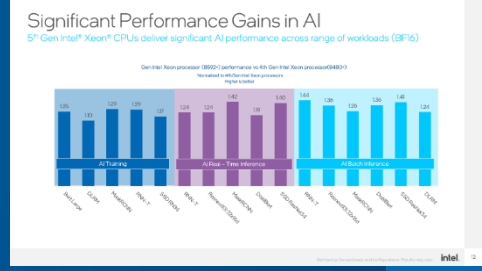
- Up to 14x higher real time object detection inference performance (SSD-ResNet34) and 9.6x higher performance/watt on 5th Gen Intel Xeon Platinum 8592+ with AMX BF16 vs. 3rd Gen Intel Xeon processor with FP32[9]
- Up to 13.8x higher batch object detection inference performance (SSD-ResNet34) and 10x higher performance/watt on 5th Gen Intel Xeon Platinum 8592+ with AMX BF16 vs. 3rd Gen Intel Xeon processor with FP32[9]

Machine Learning

- Up to 1.39x faster end-to-end Census workload performance using 5th Gen Intel Xeon processor compared to prior generation(FP32)[10]

[1,2] See [A1, A2] at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary
[3,4] See [A19, A24] at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary
[5,6,7,] See [A7, A20, A208] at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary
[8,9,] See [A26, A21] at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary
[10,] See [A8] at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary

Configuration details: AI performance



Performance varies by use, configuration and other factors. See backup for configuration details. Results may vary.

ResNeXT101_32x16d Inference: BS1, BSx

8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 4, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS EGSDCRB1.SYS.0105.D74.2308261933, microcode 0x21000161, 1x Ethernet Controller I225-LM, 1x 3.6T INTEL SSDPE2KX040T8, 1x 931.5G INTEL, CentOS Stream 9, 6.2.0-emr.bkc.6.2.3.6.31.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu, oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1, Test by INTEL as of 09/07/2023.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1.SYS.0102.D37.2305081420, microcode 0x2b0004b1, 1x Ethernet Controller I225-LM, 1x 7.2G Cruzer Blade, 2x 3.7T INTEL SSDPE2KX040T8, 1x 894.3G INTEL SSDSC2KG96, CentOS Stream 8, 5.15.0-spr.bkc.pc.16.4.24.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1. Test by INTEL as of 09/05/2023.

ResNeXT101_32x16d, BS=1: 4cores/instance, BS=x: 1 instance/ numa node; Resnext101: ImageNet

RNN-T Inference : BS1, BSx

8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 4, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS EGSDCRB1.SYS.0105.D74.2308261933, microcode 0x21000161, 1x Ethernet Controller I225-LM, 1x 3.6T INTEL SSDPE2KX040T8, 1x 931.5G INTEL, CentOS Stream 9, 6.2.0-emr.bkc.6.2.3.6.31.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu, oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1, Test by INTEL as of 09/07/2023.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1.SYS.0102.D37.2305081420, microcode 0x2b0004b1, 1x Ethernet Controller I225-LM, 1x 7.2G Cruzer Blade, 2x 3.7T INTEL SSDPE2KX040T8, 1x 894.3G INTEL SSDSC2KG96, CentOS Stream 8, 5.15.0-spr.bkc.pc.16.4.24.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1. Test by INTEL as of 09/05/2023.

RNN-T, BS=1: 4cores/instance, BS=x: 1 instance/ numa node; RNNT: LibriSpeech

DistilBERT Inference : BS1, BSx

8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.0, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic, Test by Intel as of 10/10/23.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 2.0, microcode 0x2b0004d0, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic, Test by Intel as of 10/25/23.

Software configuration: DistilBERT, Intel Model Zoo:https://github.com/IntelAI/models, gcc=12.3, OneDNN3.2, Python 3.9, PyTorch 2.0, IPEX 2.0, Transformer version 4.18.0, physical cores only.

MaskRCNN Inference: : BS1, BSx

8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 4, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS EGSDCRB1.SYS.0105.D74.2308261933, microcode 0x21000161, 1x Ethernet Controller I225-LM, 1x 3.6T INTEL SSDPE2KX040T8, 1x 931.5G INTEL, CentOS Stream 9, 6.2.0-emr.bkc.6.2.3.6.31.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu, oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1, Test by INTEL as of 09/07/2023.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1.SYS.0102.D37.2305081420, microcode 0x2b0004b1, 1x Ethernet Controller I225-LM, 1x 7.2G Cruzer Blade, 2x 3.7T INTEL SSDPE2KX040T8, 1x 894.3G INTEL SSDSC2KG96, CentOS Stream 8, 5.15.0-spr.bkc.pc.16.4.24.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1. Test by INTEL as of 09/05/2023.

MaskRCNN, BS=1: 4cores/instance, BS=x: 1 instance/ numa node; Mask RCNN: COCO 2017

Resources and Configurations

- 5th Gen Xeon Outperforms Competition Around The Clock

- ResNet50v1.5
- Intel Xeon 8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.0, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic. TensorFlow= Intel TF 2.13, OneDNN=3.2, Python 3.8, AI Model=ResNet50v1.5 Large(https://github.com/IntelAI/models/) , Batched Results: best scores achieved using BFloat16, INT8-AMX (BS >1),, Test by INTEL as of 10/10/2023.
- AMD EPYC 9654: 1-node, 2x AMD EPYC 9654 96-Core Processor, 96 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, ZenDNN 4.1 TensorFlow 2.12.1, Python 3.8, AI Model=ResNet50v1.5 Large(https://github.com/IntelAI/models/) , Batched Results: best scores achieved using (BS >1),Test by INTEL as of 09/11/23.
- NGINX TLS
- Intel Xeon 8592+: 1-node, 2x 5th Gen Intel Xeon Scalable processor (64 core) with integrated Intel Quick Assist Technology (Intel QAT), QAT device utilized=4 (1 active socket), HT On, Turbo Off, SNC On, with 1024GB DDR5 memory (16x64 GB 5600), microcode 0x21000161, Ubuntu 22.04.3 LTS, 5.15.0-78-generic, 1x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 1x100GbE, NGINX Async v0.5.1, OpenSSL 3.1.3, IPP Crypto 2021.8, IPsec MB v 1.4, QAT _Engine v 1.4.0, QAT Driver 20.l.1.1..20-00030, TLS 1.3 Webserver: ECDHE-X25519-RSA2K, tested by Intel October 2023.
- AMD EPYC 9554: 1-node, AMD platform with 2x 4th Gen AMD EPYC processor (64 cores), SMT On, Core Performance Boost Off, NPS1, Total Memory 1536GB (24x64GB DDR5 4800), microcode 0xa10113e, Ubuntu 22.04.3 LTS, 5.15.0-78-generic, 1x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 1x100GbE, NGINX Async v0.5.1, OpenSSL 3.1.3, TLS 1.3 Webserver: ECDHE-X25519-RSA2K,tested by Intel October 2023.

Resources and Configurations

- 5th Gen Xeon Outperforms Competition Around The Clock

- HammerDB Microsoft SQL Server + Backup
- Intel Xeon 8592+: 1-node, 2x 5th Gen Intel Xeon Scalable processor 8592+ (64 cores) with integrated Intel Quick Assist Technology (Intel QAT), Number of IAA device utilized=8(2 sockets active), HT On, Turbo On, SNC off, Total Memory 1024GB (16x64GB DDR5 5600), microcode 0x21000161, 2x Ethernet Controller 10-Gigabit X540-AT2, 7x 3.5T INTEL SSDPE2KE032T807, QATZip 2.0.W.1.9.0-0008, Microsoft Windows Server Datacenter 2022, Microsoft SQL Server 2022, SQL Server Management Studio 19.0.1, HammerDB 4.5, tested by Intel October 2023.
- AMD EPYC 9554: 1-node, AMD platform with 2x 4th Gen AMD EPYC processor (64 cores), SMT On, Core Performance Boost On, NPS1, Total Memory 1536GB (24x64GB DDR5 4800), microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 7x 3.5T INTEL SSDPE2KE032T807, Microsoft Windows Server Datacenter 2022, Microsoft SQL Server 2022, SQL Server Management Studio 19.0.1, HammerDB 4.5, tested by Intel October 2023.
- RocksDB
- Intel Xeon 8592+: 1-node, 2x 5th Gen Intel Xeon Scalable processor 8592+ (64 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), Number of IAA device utilized=8(2 sockets active), HT On, Turbo On, SNC off, Total Memory 1024GB (16x64GB DDR5 5600), microcode 0x21000161, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 6.5.0-060500-generic, QPL v1.2.0, accel-config-v4.0, iaa_compressor plugin v0.3.0, ZSTD v1.5.5, gcc 10.4.0, RocksDB v8.3.0 trunk (commit 62fc15f) (db_bench), 4 threads per instance, 64 RocksDB instances, tested by Intel October 2023.
- AMD EPYC 9554: 1-node, AMD platform with 2x 4th Gen AMD EPYC processor (64 cores), SMT On, Core Performance Boost On, NPS1, Total Memory 1024GB (16x64GB DDR5 4800), microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 6.5.0-060500-generic, ZSTD v1.5.5, gcc 10.4.0, RocksDB v8.3.0 trunk (commit 62fc15f) (db_bench), 4 threads per instance, 28 RocksDB instances, tested by Intel October 2023.
- Monte Carlo
- Intel Xeon 8592+: 1-node 2x Intel Xeon 8592+, HT On, Turbo On, SNC2, 1024 GB DDR5-5600, ucode 0x21000161, Red Hat Enterprise Linux 8.7, 4.18.0-425.10.1.el8_7.x86_64, Monte Carlo v1.2, cmkl:2023.2.0 icc:2023.2.0 tbb:2021.10.0. Test by Intel as of October 2023.
- AMD EPYC 9554: 1-node, 2x AMD EPYC 9554, SMT On, Turbo On, CTDP=360W, NPS=4, 1536GB DDR5-4800, ucode= 0xa101111, Red Hat Enterprise Linux 8.7, Kernel 4.18,Monte Carlo v1.2, cmkl:2023.2.0 icc:2023.2.0 tbb:2021.10.0. Test by Intel as of March 2023

Resources and Configurations

- 5th Gen Xeon Outperforms Competition Around The Clock
  - DLRM
  - Intel Xeon 8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS2.0, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic. Framework=Pytorch 2.1, IPEX=2.1, Python 3.8, AI Model= DLRM(https://github.com/IntelAI/models/) , Batched Results: best scores achieved using BS>1, Precision=INT8-AMX, Test by INTEL as of 10/10/2023.
  - AMD EPYC 9654: 1-node, 2x AMD EPYC 9654 96-Core Processor, 96 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, Framework=Pytorch 2.1, IPEX=2.1, Python 3.8, AI Model= DLRM(https://github.com/IntelAI/models/) , Batched Results: best scores achieved using BS>1, Precision=INT8. Test by INTEL as of 09/11/23.
  - HammerDB MySQL
  - Intel Xeon 8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 8 [0], DSA 8 [0], IAX 8 [0], QAT 8 [0], Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.0, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, 2x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, Ubuntu 22.04.3 LTS, 5.15.0-84-generic, HammerDB Mv4.4, MySQL 8.0.33. Test by Intel as of 10/04/23.
  - AMD EPYC 9554: 1-node, 2x AMD EPYC 9554 64-Core Processor, 64 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 0 [0], DSA 0 [0], IAX 0 [0], QAT 0 [0], Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, 2x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, HammerDB v4.4, MySQL 8.0.33. Test by Intel as of 10/05/23.

Configurations: 5th Gen Xeon TCO Advantages (1 of 6)

Claim: 5th Gen Intel Xeon delivers up to 24% lower TCO than the 4th Gen AMD Epyc while running a HammerDB MySQL OLTP database workload.

- Based on 5th Gen Intel Xeon delivers up to a 1.70x faster than the 4th Gen AMD Epyc while running a HammerDB MySQL OLTP workload. This performance drives a fleet reduction from 50 to 30
- servers which, over 4 years, saves: 684.0 MWH of energy, 289,967 kgCo2 emissions, and $508.9k of cost.
- 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 8 [0], DSA 8 [0], IAX 8 [0], QAT 8 [0], Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.0, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, 2x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, Ubuntu 22.04.3 LTS, 5.15.0-84-generic, HammerDB Mv4.4, MySQL 8.0.33. Test by Intel as of 10/04/23.
- 1-node, 2x AMD EPYC 9554 64-Core Processor, 64 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 0 [0], DSA 0 [0], IAX 0 [0], QAT 0 [0], Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, 2x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, HammerDB v4.4, MySQL 8.0.33. Test by Intel as of 10/05/23.

- For a 50 server fleet of AMD EPYC 9554, estimated as of October 2023:
- CapEx costs: $1.40M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $755.1K
- Energy use in kWh (4 year, per server): 47654, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

- For a 30 server fleet of 5th Gen Xeon 8592+ as of October 2023
- CapEx costs: $1.17M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $480.0K
- Energy use in kWh (4 year, per server): 58625, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394
- Costs based on Intel estimates andinformation from thinkmate.com as of October 2023.

Configurations: 5th Gen Xeon TCO Advantages (2 of 6)

Claim: 5th Gen Intel Xeon delivers up to 22% lower TCO than the 4th Gen AMD Epyc while running a RocksDB database workload.

- Based on 5th Gen Intel Xeon delivers up to a 1.62x faster than the 4th Gen AMD Epyc while running a RocksDB database workload. This performance drives a fleet a reduction from 50 to 31 servers which, over 4 years, saves: 1,218 MWH of energy, 516,402 kgCo2 emissions, and $471.8k of cost.
- 1-node, 2x 5th Gen Intel Xeon Scalable processor 8592+ (64 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), Number of IAA device utilized=8(2 sockets active), HT On, Turbo On, SNC off, Total Memory 1024GB (16x64GB DDR5 5600), microcode 0x21000161, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 6.5.0-060500-generic, QPL v1.2.0, accel-config-v4.0, iaa_compressor plugin v0.3.0, ZSTD v1.5.5, gcc 10.4.0, RocksDB v8.3.0 trunk (commit 62fc15f) (db_bench), 4 threads per instance, 64 RocksDB instances, tested by Intel October 2023.
- 1-node, AMD platform with 2x 4th Gen AMD EPYC processor (64 cores), SMT On, Core Performance Boost On, NPS1, Total Memory 1024GB (16x64GB DDR5 4800), microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 6.5.0-060500-generic, ZSTD v1.5.5, gcc 10.4.0, RocksDB v8.3.0 trunk (commit 62fc15f) (db_bench), 4 threads per instance, 28 RocksDB instances, tested by Intel October 2023.

- For a 50 server fleet of AMD EPYC 9554, estimated as of October 2023:
- CapEx costs: $1.36M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $809.5K
- Energy use in kWh (4 year, per server): 58531, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

- For a 31 server fleet of 5th Gen Xeon 8592+ as of October 2023
- CapEx costs: $1.21M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $491.3K
- Energy use in kWh (4 year, per server): 55111, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394
- Costs based on Intel estimates and information from thinkmate.com as of October 2023

Configurations: 5th Gen Xeon TCO Advantages (3 of 6)

Claim: 5th Gen Intel Xeon delivers up to 21% lower TCO than the 4th Gen AMD Epyc while running a NGINX TLS Handshake workload.

- Based on 5th Gen Intel Xeon delivers up to a 1.66x faster than the 4th Gen AMD Epyc while running a NGINX TLS Handshake workload. This performance drives a fleet a reduction from 50 to 31 servers which, over 4 years, saves: 489.7 MWH of energy, 207,611 kgCo2 emissions, and $443.5k of cost.
- 1-node, 2x 5th Gen Intel Xeon Scalable processor (64 core) with integrated Intel Quick Assist Technology (Intel QAT), Integrated Accelerators Available [used]: DLB 2 [0], DSA 2 [0], IAA 2 [0], QAT 2 [0], HT On, Turbo Off, SNC On, with 1024GB DDR5 memory (16x64 GB 5600), microcode 0x21000161, Ubuntu 22.04.3 LTS, 5.15.0-78-generic, 1x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 1x100GbE, NGINX Async v0.5.1, OpenSSL 3.1.3, IPP Crypto 2021.8, IPsec MB v 1.4, QAT _Engine v 1.4.0, QAT Driver 20.l.1.1..20-00030, TLS 1.3 Webserver: ECDHE-X25519-RSA2K, tested by Intel October 2023.
- 1-node, 9554: 1-node, AMD platform with 2x 4th Gen AMD EPYC processor (64 cores), SMT On, Core Performance Boost Off, NPS1, Total Memory 1536GB (24x64GB DDR5 4800), microcode 0xa10113e, Ubuntu 22.04.3 LTS, 5.15.0-78-generic, 1x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 1x100GbE, NGINX Async v0.5.1, OpenSSL 3.1.3, TLS 1.3 Webserver: ECDHE-X25519-RSA2K,tested by Intel October 2023.

- For a 50 server fleet of AMD EPYC 9554, estimated as of October 2023:
- CapEx costs: $1.41M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $698.7K
- Energy use in kWh (4 year, per server): 36386, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

- For a 31 server fleet of 5th Gen Xeon 8592+ as of October 2023
- CapEx costs: $1.21M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $453.4K
- Energy use in kWh (4 year, per server): 42889, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394
- Cost based on Intel estimates and information from thinkmate.com as of October 2023

Configurations: 5th Gen Xeon TCO Advantages (4 of 6)

Claim: 5th Gen Intel Xeon delivers up to 27% lower TCO than the 4th Gen AMD Epyc while running Monte Carlo workload.

- Based on 5th Gen Intel Xeon delivers up to a 1.83x faster than the 4th Gen AMD Epyc while running a Monte Carlo workload. This performance drives a fleet a reduction from 50 to 28 servers which, over 4 years, saves: 585.8 MWH of energy, 248,352 kgCo2 emissions, and $561.0k of cost.
- 1-node 2x Intel Xeon 8592+, HT On, Turbo On, SNC2, 1024 GB DDR5-5600, ucode 0x21000161, Red Hat Enterprise Linux 8.7, 4.18.0-425.10.1.el8_7.x86_64, Monte Carlo v1.2, cmkl:2023.2.0 icc:2023.2.0 tbb:2021.10.0. Test by Intel as of October 2023.
- 1-node, 2x AMD EPYC 9554, SMT On, Turbo On, CTDP=360W, NPS=4, 1536GB DDR5-4800, ucode= 0xa101111, Red Hat Enterprise Linux 8.7, Kernel 4.18,Monte Carlo v1.2, cmkl:2023.2.0 icc:2023.2.0 tbb:2021.10.0. Test by Intel as of March 2023

- For a 50 server fleet of AMD EPYC 9554, estimated as of October 2023:
- CapEx costs: $1.36M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $739.3K
- Energy use in kWh (4 year, per server): 44505, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

- For a 28 server fleet of 5th Gen Xeon 8592+ as of October 2023
- CapEx costs: $1.09M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $453.3K
- Energy use in kWh (4 year, per server): 58550, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394
- Costs based on Intel estimates and information from thinkmate.com as of October 2023.

Configurations: 5th Gen Xeon TCO Advantages (5 of 6)

Claim: 5th Gen Intel Xeon delivers up to 46% lower TCO than the 4th Gen AMD Epyc while running a real-time Natural Language Processing inference (BERT-Large) workload.

- Based on 5th Gen Intel Xeon delivers up to a 2.44x faster than the 4th Gen AMD Epyc while running a real-time Natural Language Processing inference (BERT-Large) workload. This performance drives a fleet a reduction from 50 to 21 servers which, over 4 years, saves: 1231.2 MWH of energy, 521,941 kgCo2 emissions, and $982.9k of cost.
- Intel Xeon 8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.0, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic. Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=BERT Large(https://github.com/IntelAI/models/) , INT8-AMX , Real Time (BS=1) results while maintaining 130ms latency SLA, Test by INTEL as of 10/10/2023.
- 9554: 1-node, 2x AMD EPYC 9554 64-Core Processor, 64 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=BERT Large(https://github.com/IntelAI/models/) , INT8, Real Time (BS=1) results while maintaining 130ms latency SLA. Test by INTEL as of 09/11/23.

- For a 50 server fleet of AMD EPYC 9554, estimated as of October 2023:
- CapEx costs: $1.36
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $766.9K
- Energy use in kWh (4 year, per server): 50021, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

- For a 21 server fleet of 5th Gen Xeon 8592+ as of October 2023
- CapEx costs: $801K
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $344.0K
- Energy use in kWh (4 year, per server): 60472, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394
- Costs based on Intel estimates and information from thinkmate.com as of October 2023.

Configurations: 5th Gen Xeon TCO Advantages (6 of 6)

Claim: 5th Gen Intel Xeon delivers up to 62% lower TCO than the 4th Gen AMD Epyc while running real-time Natural Language Processing inference (DistilBERT) workload.

- Based on 5th Gen Intel Xeon delivers up to a 3.49x faster than the 4th Gen AMD Epyc while running a real-time Natural Language Processing inference (DistilBERT) workload. This performance drives a fleet a reduction from 50 to 15 servers which, over 4 years, saves: 1496.5 MWH of energy, 634,428 kgCo2 emissions, and $1,300k of cost.
- 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.0, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic. Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=DistilBERT (https://github.com/IntelAI/models/) , INT8-AMX, Real Time (BS=1) results while maintaining 5ms latency SLA, Test by INTEL as of 10/10/2023.
- 1-node, 2x AMD EPYC 9554 64-Core Processor, 64 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=DistilBERT Large(https://github.com/IntelAI/models/) , INT8, Real Time (BS=1) results while maintaining 5ms latency SLA. Test by INTEL as of 09/11/23.

- For a 50 server fleet of AMD EPYC 9554, estimated as of October 2023:
- CapEx costs: $1.36
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $749.7K
- Energy use in kWh (4 year, per server): 46573, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

- For a 15 server fleet of 5th Gen Xeon 8592+ as of October 2023
- CapEx costs: $572K
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $238.3K
- Energy use in kWh (4 year, per server): 55475, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394
- Costs based on Intel estimates and information from thinkmate.com as of October 2023.