

# Modernize your data center with HPE Gen 12 Servers and Intel® Xeon® 6

Winnie Chang, Intel Corporation

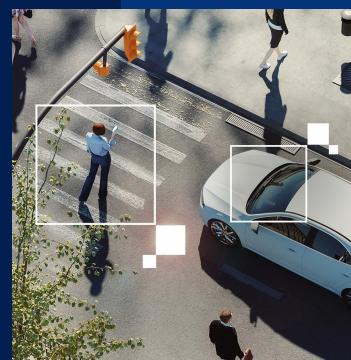
October 16, 2025

# 以 HPE ProLiant Gen12 與 Intel® Xeon® 6 邁向資 料中心現代化

**Winnie Chang – AI Solution Engineer**

**Intel Corporation**

**Oct 16, 2025**



# Data Center Requirements Are Evolving

Varying uses require unique optimization vectors

## AI Everywhere

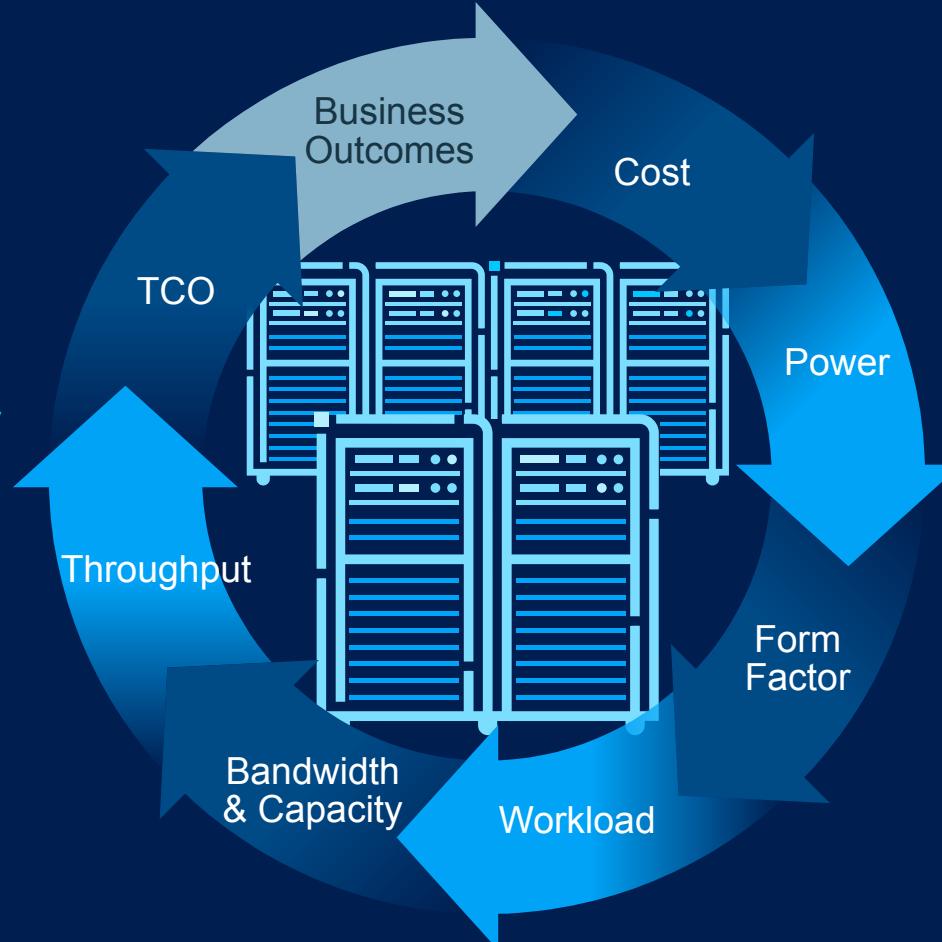
Enable performance at scale with accelerated hardware and open standards-based software

## Throughput & Latency

Maximize system-level performance with best-in-class response time

## Security, Quality & Reliability

Provide security, quality, and reliability for at-scale deployments



## Efficiency & TCO

Increase rack density while meeting power efficiency requirements to improve TCO (total cost of ownership)

## Sustainability

Minimizing carbon emissions through improved energy efficiency and circular product design

## Software Compatibility

ISA consistency for software ecosystem compatibility

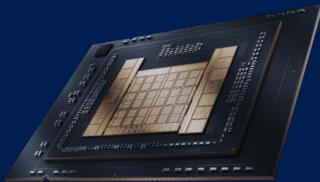
# Addressing Tomorrow's Computing Needs

Across data center, network & edge

## Intel® Xeon® 6 Processors

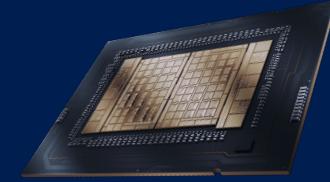
### Maximum Efficiency with E-cores

Performance-  
per-watt for high-density  
compute and scale-out  
workloads



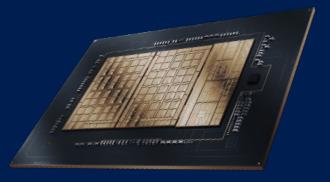
### Peak Performance with P-cores

Per-core-performance  
for compute-intensive  
workloads



### Built for Networking and Edge

More power-efficient,  
servers with Intel® vRAN  
Boost and media  
acceleration, with  
networking built in



### Performance Boost for Small Business

Affordable performance  
for business-critical  
services



# Why Intel® Xeon® 6 ?

Addressing the broadest set of enterprise workloads

## Exceptional Performance

with more cores & higher memory bandwidth

## Efficient Compute

improved perf per watt, built-in acceleration, & 16S+ scalability

## Trusted & Secure

most comprehensive confidential compute portfolio

## Foundation for AI

best CPU for AI with Intel® AMX & most deployed AI Accelerated Systems

higher memory bandwidth  
with more cores &

16S+ scalability & acceleration, & built-in security

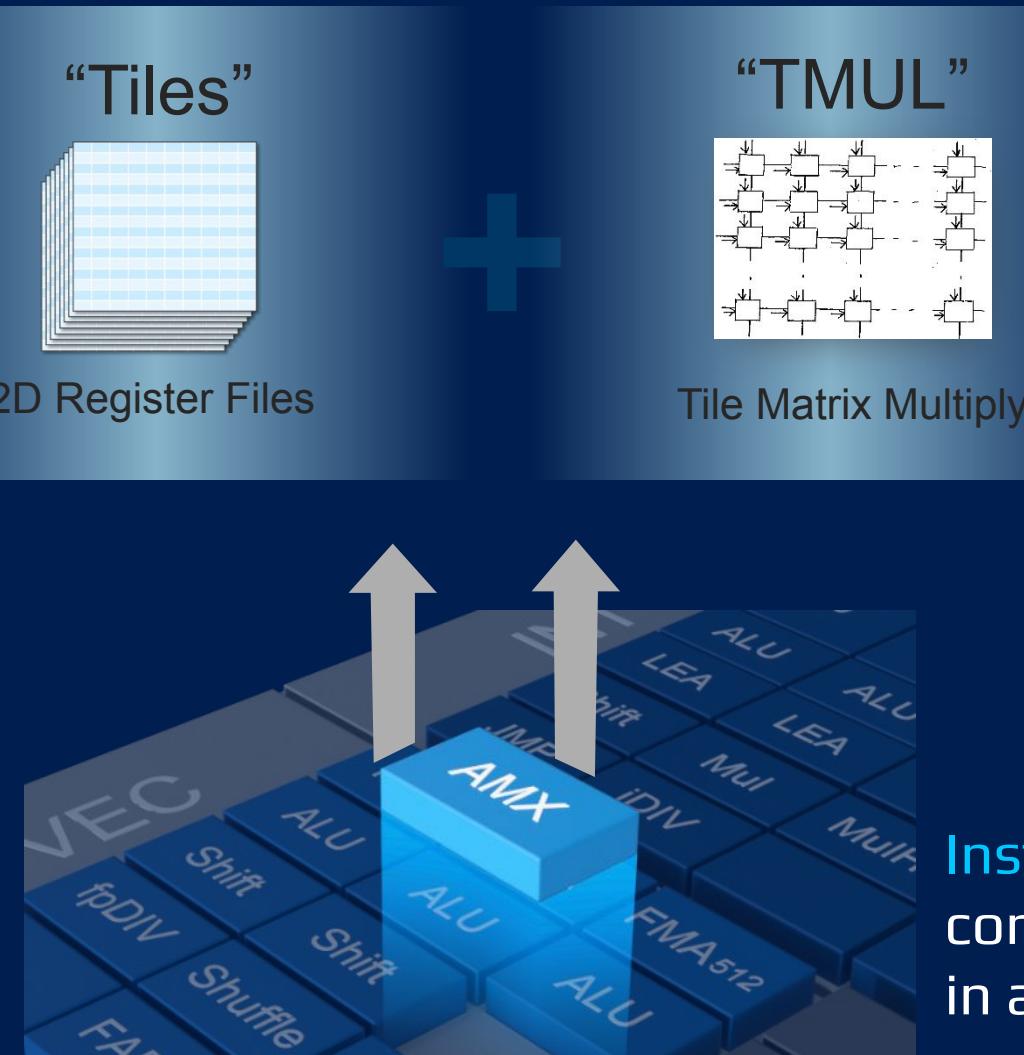
comprehensive confidential compute portfolio

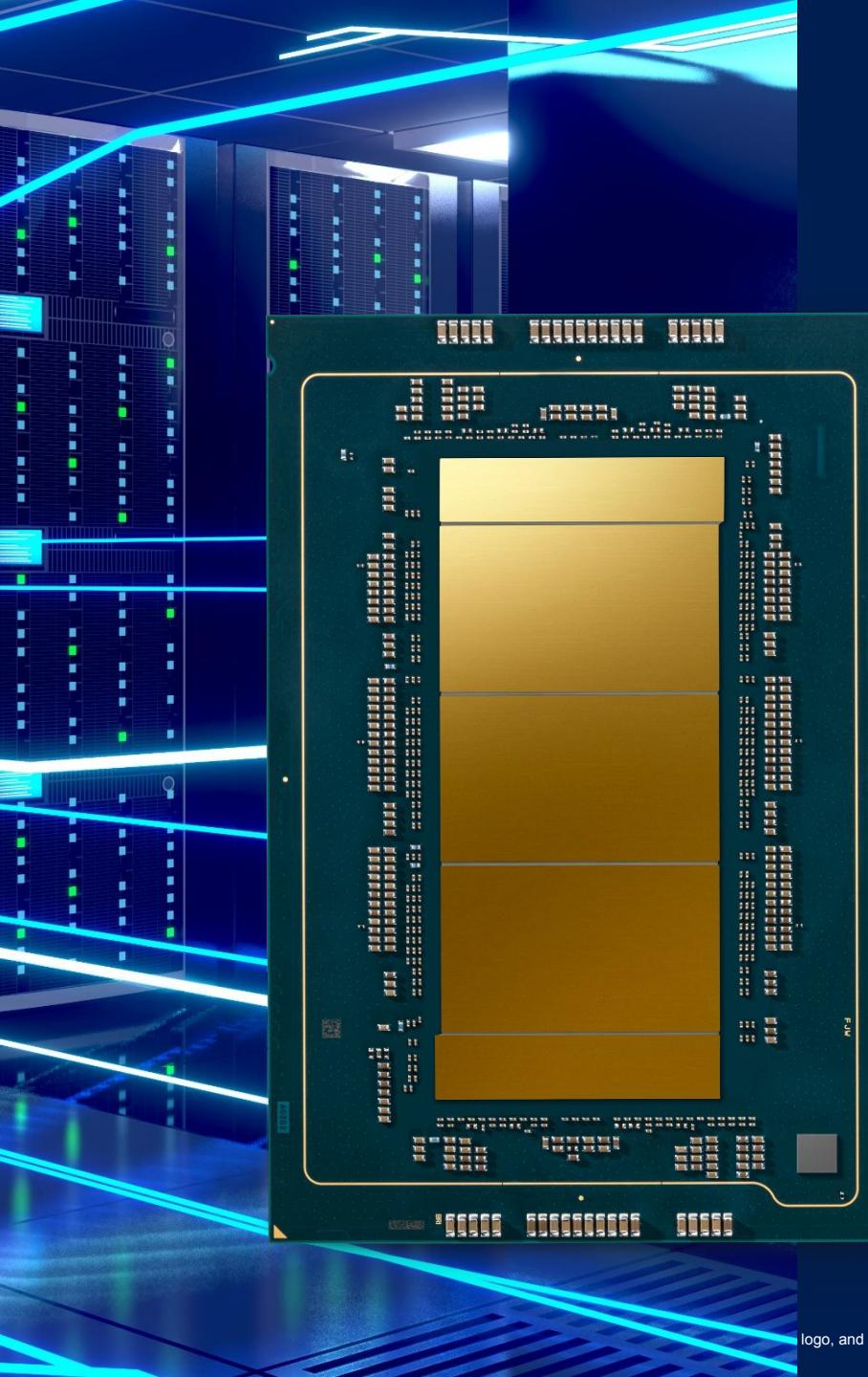
AI accelerated systems  
most deployed AI Accelerated Systems

# AI: Intel® Advanced Matrix Extensions (Intel® AMX) DL Accelerator Performance Built Into Every Core

	4 <sup>th</sup> /5 <sup>th</sup> /6 <sup>th</sup> Gen Intel® Xeon® Scalable processor
AVX512	FP64, FP32
VNNI	INT8
AMX	FP16, BF16, INT8

Store bigger  
chunks of **data**





# Intel® Xeon® 6900 Processor

## with Performance-cores (P-cores)

Up to 6400 MT/s DDR5

8800 MT/s MRDIMM memory

Up to 128 performance cores

6 UPI 2.0 links, up to 24 GT/s

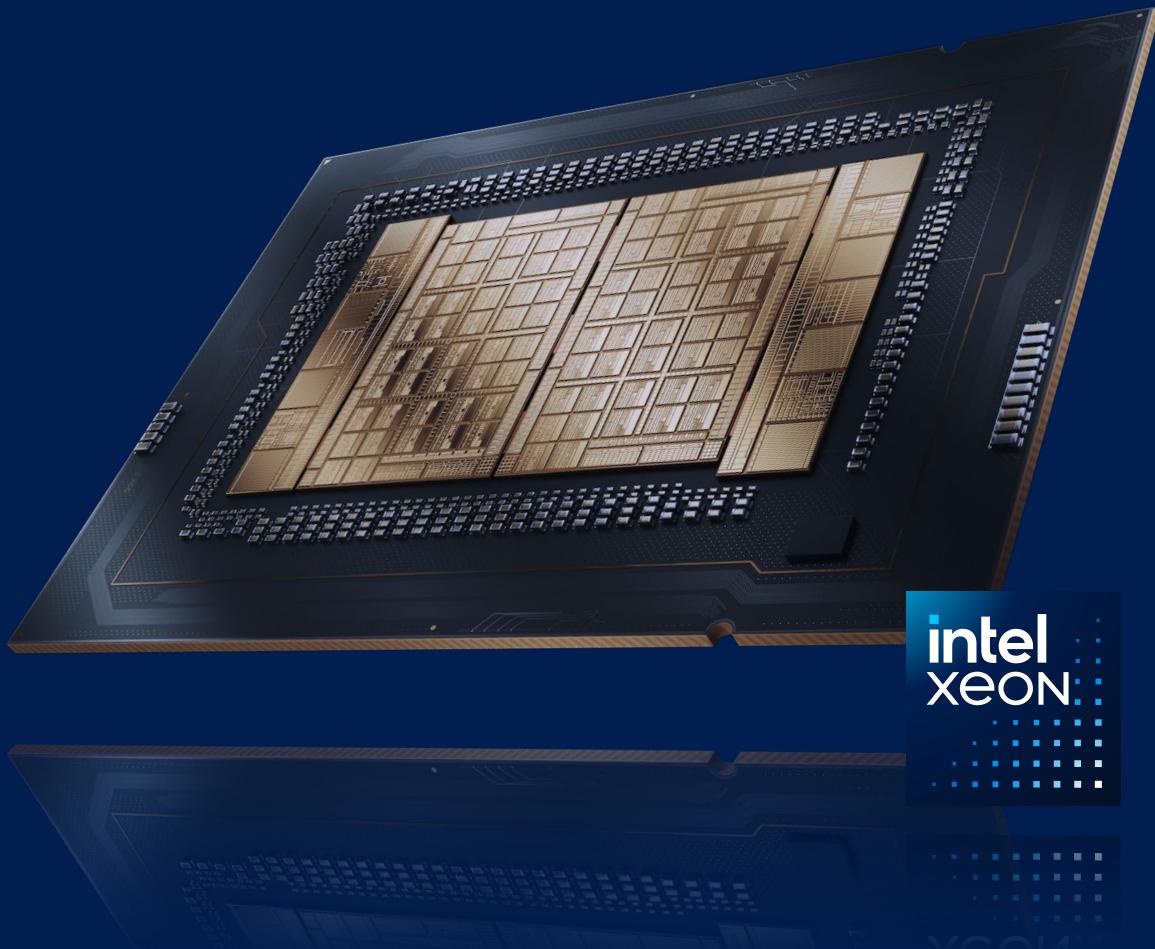
Up to 96 lanes PCIe 5.0/CXL® 2.0

L3 cache as large as 504 MB

Intel Advanced Matrix Extensions (Intel AMX) with FP16 support

# Intel® Xeon® 6700P & 6500P Processors

The perfect balance of power, performance, & efficiency



## Run the broadest range of enterprise workloads

meeting the compute and virtualization demands in today's challenging data centers

## Optimize your data center

with server consolidation, delivering better efficiency, lower power costs and improved total cost of ownership

## Scale AI everywhere

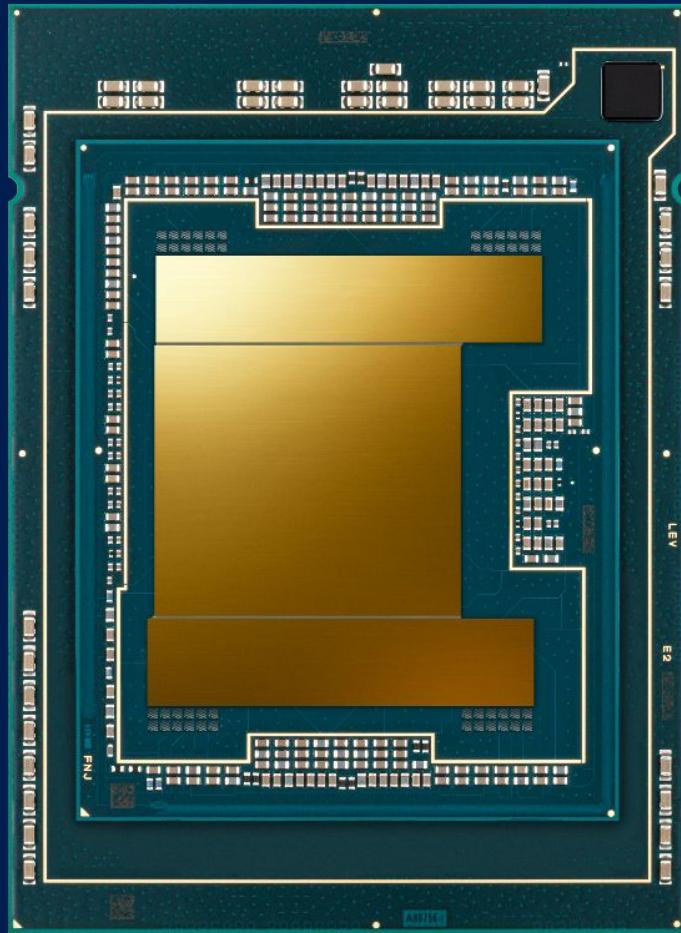
with the best CPU for AI inferencing and the most deployed host CPU for accelerated AI systems

## Protect your data

hardware- based security, confidential computing, and trust services

# Intel® Xeon® 6700E processors with Efficient-cores

Optimize and Scale Your Infrastructure with Cloud-Native Agility



Delivers distinct advantages for cloud-scale workloads



Data services,  
networking, media,  
and microservices

Unmatched core density for scale-out capacity

Up to  
**144**

Cores  
per socket

Highest task parallel performance per watt

Up to  
**3X**

Higher performance per watt as compared to 2<sup>nd</sup> Gen Intel® Xeon® processor

Increase rack utilization for better efficiency and total cost of ownership

**3:1**

Consolidate servers using 2<sup>nd</sup> Gen Intel Xeon processors to Intel Xeon 6 with E-cores

# Intel® Xeon® 6 Processors

	Intel Xeon 6700/6500 Series Processors with P-cores	Intel Xeon 6900 Series Processors with P-cores	Intel Xeon 6 Processors with E-cores		
Cores	up to <b>86 cores</b>	up to <b>128 cores</b>	up to <b>144 cores</b> (6700E series)   <b>288 cores</b> (6900E series)		
Sockets	1S, 2S, <b>4S, 8S</b>	1S, 2S	1S, 2S		
Max TDP	150 to 350W	400 to 500W	205 to 500W		
Memory	up to <b>8 channels</b> DDR5   MRDIMM	up to <b>12 channels</b> DDR5   MRDIMM	up to <b>12 channels</b> DDR5		
Max Memory Speed	6400 (1DPC) DDR5   5200 (2 DPC) DDR5   8000 MRDIMM (1DPC)	6400 (1DPC) DDR5   8800 MRDIMM (1DPC)	6400 (1DPC) DDR5   5200 (2 DPC) DDR5 (6700E)   6400 (1DPC) DDR5 (6900E)		
Intel® UPI	up to <b>4 UPI 2.0</b> at up to 24 GT/s per lane	up to <b>6 UPI 2.0</b> at up to 24 GT/s per lane	up to <b>6 UPI 2.0</b> at up to 24 GT/s per lane		
PCI Express 5.0	up to <b>88 lanes</b> up to <b>136 lanes</b> for single socket designs	up to <b>96 lanes</b>	up to <b>96 lanes</b>		
Compute Express Link	up to <b>64 lanes CXL 2.0</b>	up to <b>64 lanes CXL 2.0</b>	up to <b>64 lanes CXL 2.0</b>		
AI Acceleration Intel® Deep Learning Boost	<b>Intel AMX</b> (INT8, BF16, FP16)	<b>Intel AVX 512</b> (VNNI/INT8)	<b>Intel AMX</b> (INT8, BF16, FP16)	<b>Intel AVX 512</b> (VNNI/INT8)	<b>Intel AVX 2</b> (VNNI/INT8)
Security	Intel Software Guard Extensions, Intel Trust Domain Extensions				
Crypto	Vector AES, SHA2-256 extensions, VPMADD52				
Integrated Accelerators	Intel QuickAssist Technology, Intel Dynamic Load Balancer, Intel Data Streaming Accelerator, Intel In-memory Analytics Accelerator				

# Driving Platform Enhancements with Intel® Xeon® 6 Processors



6900 Series

6700/6500 Series

DDR5

Up to **2.3X**

higher memory bandwidth  
(w/MRDIMM memory in P-core)  
vs. 5<sup>th</sup> Gen Intel® Xeon® processors

Up to **1.4X**

higher memory bandwidth  
(w/MRDIMM memory in P-core)  
vs. 5<sup>th</sup> Gen Intel Xeon processors

PCIe5

Up to **1.2X**

increased I/O Bandwidth  
vs. 5<sup>th</sup> Gen Intel Xeon processors

Up to **1.1X**

increased I/O Bandwidth  
vs. 5<sup>th</sup> Gen Intel Xeon processors

UPI 2.0

Up to **1.8X**

increased inter-socket bandwidth  
vs. 5<sup>th</sup> Gen Intel Xeon processors

Up to **1.2X**

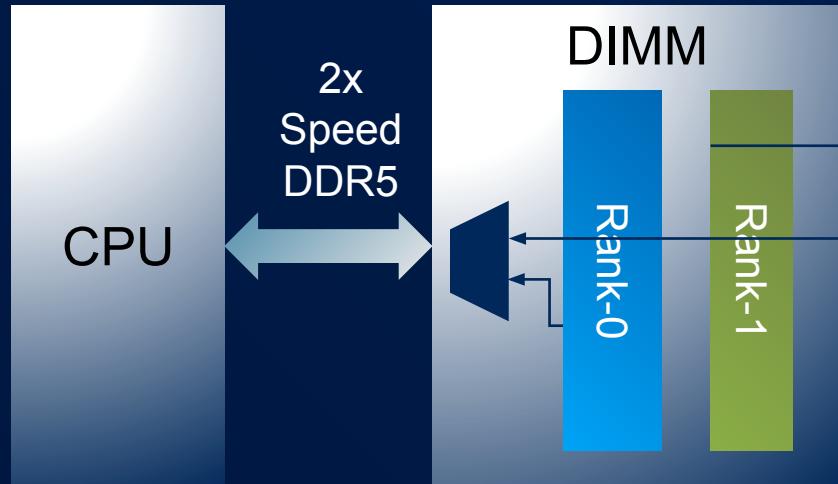
increased inter-socket bandwidth  
vs. 5<sup>th</sup> Gen Intel Xeon processors

CXL® 2.0

Type 1, Type 2, and Type 3

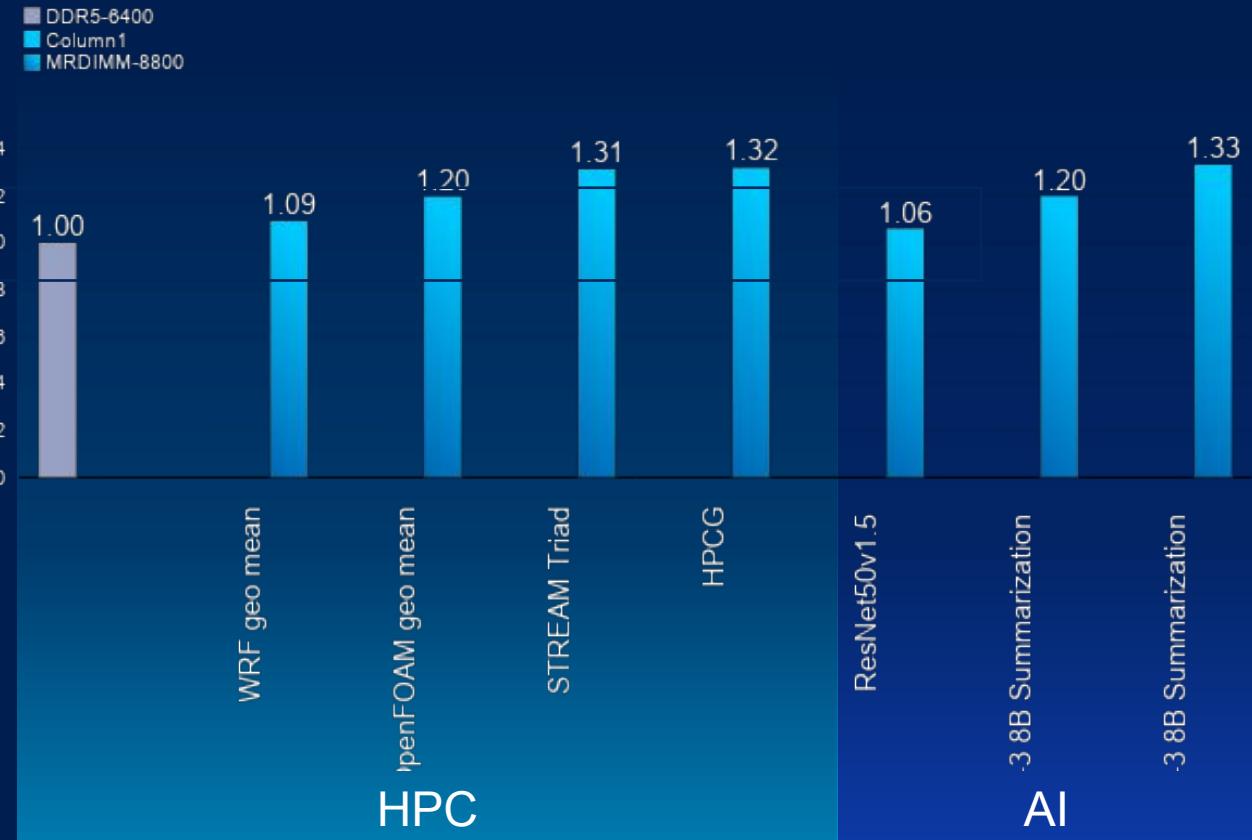
# Multiplexed Rank DIMMs

First to market on Intel® Xeon® 6 processors with P-core



MRDIMMs deliver up to 8800 MT/s data rate on Intel Xeon 6 with P-cores

Intel® Xeon® 6 with P-cores (128c)  
MRDIMM-8800 Performance Gains Over DDR5-6400  
Higher is better



See backup for workload and configurations [13]. Results may vary. \* 6972P (96c) used.  
This offering is not approved or endorsed by OpenCFD Limited, producer and distributor of the OpenFOAM software via [www.openfoam.com](http://www.openfoam.com), and owner of the OPENFOAM® and OpenCFD® trademark

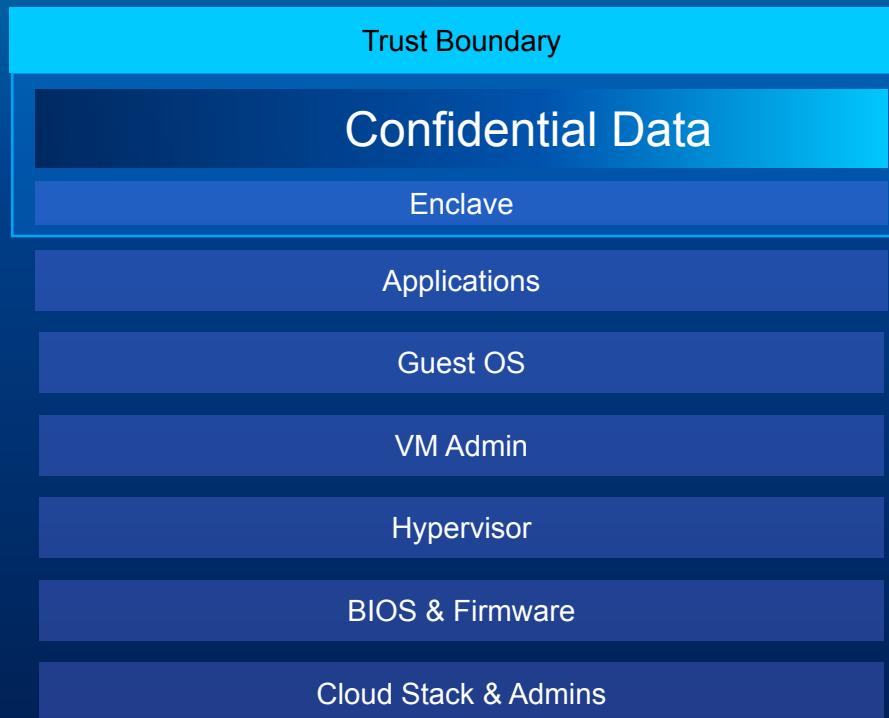
# The Most Comprehensive Portfolio

for confidential computing

## App Isolation

Intel® SGX

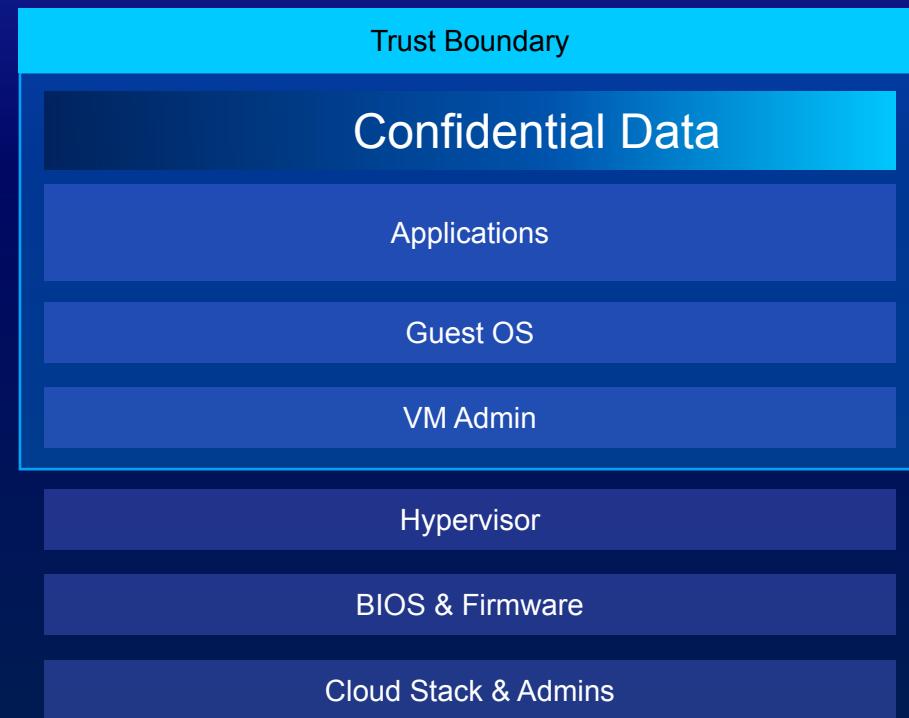
Smallest trust boundary for greatest data protection & code integrity



## VM Isolation

Intel® TDX

Most straightforward path to greater security, compliance & control for legacy apps



Intel® Software Guard Extensions (Intel® SGX), Intel® Trust Domain Extensions (Intel® TDX)

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

# The Next Milestone in Confidential AI with Intel® TDX Connect

Provides a high-performance encrypted connection between the CPU and PCIe devices with direct memory access and lower overhead





Storage



Networking



Content Delivery



Host CPU for AI

## Intel® Xeon® 6 | 1-socket platform with 136 PCIe lanes

Delivers enhanced performance and lower TCO for today and tomorrow's data center demands

### Increased PCIe lanes

Up to 136 lanes for greater lane volume for peripherals & devices

### Improved I/O performance

Removal of power and latency penalties in a single socket



### Enhanced performance and power efficiency

Greater core density, MRDIMMs and Intel® Accelerators

### Cost optimization

Opportunity for server consolidation for improved total cost of ownership

# Only x86 Solution Meeting the Market's Need for Scalability

Intel® Xeon® 6 processor 4S/8S and beyond configurations provide more processing power, I/O bandwidth, and memory capacity

Generational Specs			
	2 <sup>nd</sup> Gen Intel Xeon	4 <sup>th</sup> Gen Intel Xeon	Intel Xeon 6700 with P-cores
Max Cores / socket	28 cores	60 cores	86 cores
DDR Mem capacity / socket	Up to 3TB	Up to 4TB	Up to 4TB
Memory speed	Up to 2933 MT/s (DDR4)	Up to 4800 MT/s (DDR5)	Up to 6400 MT/s (DDR5)

### Topology Support

4 socket Ring (2 UPI)

4 socket Fully connected (3 UPI)

8 socket Optimized (4 UPI)

# Scalability and Flexibility with Modular SoC Architecture

Intel® Xeon® 6 processor architecture

Module-die Fabric	Enables flexible construction offering customers a breadth of compute choices
Multi-die Architecture	I/O die with UPI, PCIe, CXL and Intel® Accelerator Engines Compute die with cores, cache, and memory controllers
Embedded Multi-die Interconnect Bridge	In-package high-density interconnect enabling high bandwidth, low power, and low latency



# Modular I/O Die Architecture

Universal and common I/O stacks across Intel® Xeon® 6 processors on Intel 7 process

## Universal I/O Stacks

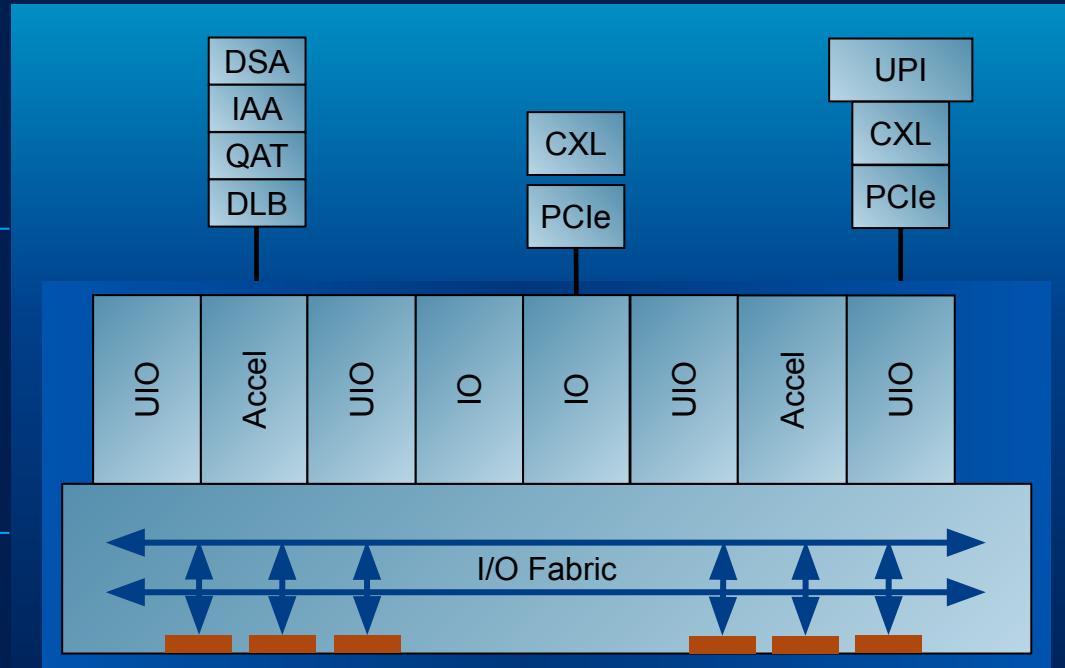
UPI, PCIe, CXL and Intel® Accelerator Engines

## New Capabilities

Full CXL support, extends Intel® Resource Director Technology (RDT),  
secure interconnect

## Enhanced I/O Performance

UPI @ 24GT/s w/6-links, UPI affinity, distributes traffic across all  
mesh columns



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

# Higher Performance Efficiency Across Server Utilization

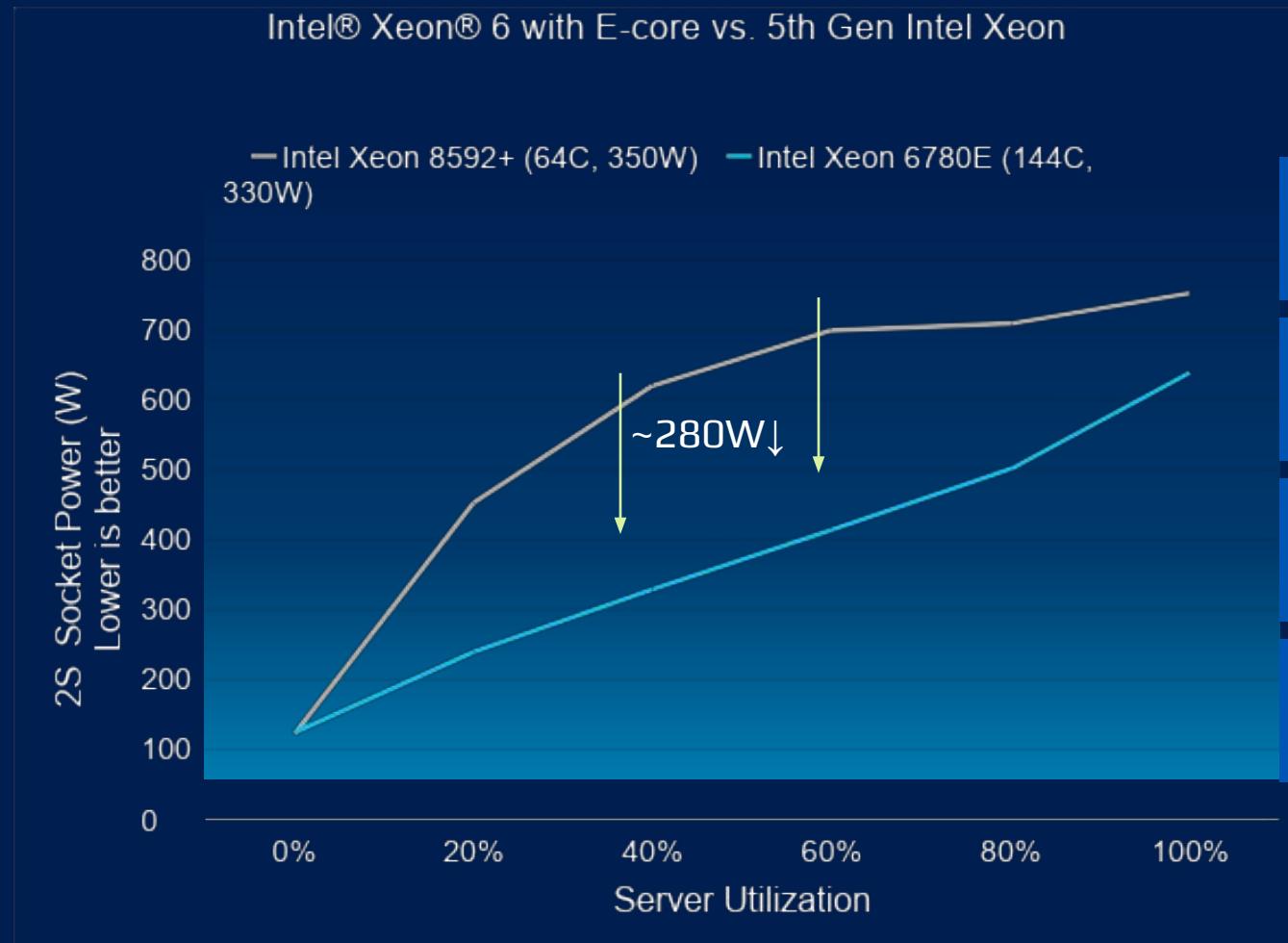
Intel® Xeon® 6 processor with P-cores delivers significant advantages in performance per watt at typical 40% server utilization



- Intel Xeon 8592+
- Intel Xeon 6760P Latency Optimized Mode
- Intel Xeon 6760P Out Of Box (OPM)

# Lower Power Across Server Utilization

Intel® Xeon® 6 processor with E-cores delivers improved energy efficiency across the load line



Power increases linearly with load on Intel Xeon 6 with E-cores

Save up to 280W power when operating at sweet spot 40-60% server utilization

18% improved performance on Intel Xeon 6780E vs. Intel Xeon Platinum 8592+

Lower power and cooling costs in your datacenter with default out-of-box settings

\*Out-of-Box mode: Assumes default energy-efficient BIOS and OS settings.  
Socket power is power consumed by CPUs  
See [773] at [intel.com/processorclaims](http://intel.com/processorclaims): Intel® Xeon® 6. Results may vary

# Save Power and Money on New Server Purchases

Performance advantage and TCO savings vs AMD EPYC 9005 servers

## Intel Xeon 6900P Processor-based Servers

Recommendation  
System  
DLRM

**1.87x** Perf / Server



Computational Fluid Dynamics  
OpenFOAM

**1.43x** Perf / Server



## Intel Xeon 6700P Processor-based Servers

Web Services  
NGINX TLS (1S) on 6760P

**1.55x** Perf / Server



Image Construction  
Vision Transformer on 6760P

**2.09x** Perf / Server



\*Estimated over 4 years. See [9T223, 9T222, 7T223, 7T221] intel.com/processorclaims: Intel Xeon 6. 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.

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

# Intel® Xeon® 6 Delivers Performance Advantage

## Across Diverse Data Center Workloads

Intel Xeon 6900P vs AMD EPYC 9005

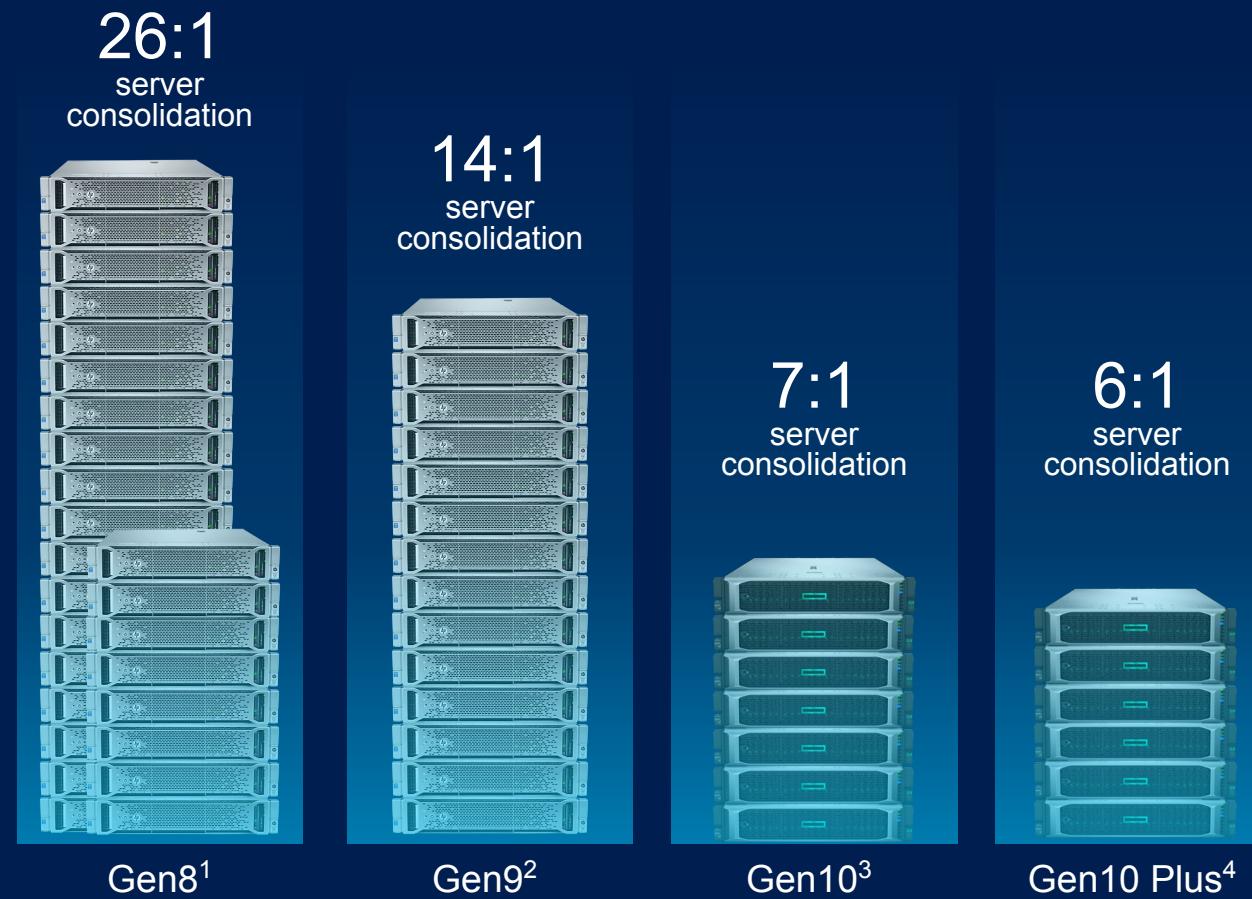


Performance comparison at ~similar core counts.  
See [9D220 - 9D222, 9W220, 9H221 - 9H224, 9A221 - 9A224] intel.com/processorclaims: Intel Xeon 6. Results may vary

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

# The most compelling case for server refresh yet

41% higher performance per watt versus Gen10 <sup>6</sup>
Free up data center capacity with up to 26:1 consolidation <sup>2</sup>
Reduce energy consumption and achieve up to 84% power savings <sup>1</sup>
Mitigating risk and technical debt from dated operating system and software



Power savings per year (up to) 84% 61% 65% 50% >

SPEC and the names SPECrate are registered trademarks of the Standard Performance Evaluation Corporation (SPEC). The stated results [SPECrate2017\_int\_base: #36693 (1), #36691 (2), #20893 (3), #37007 (4)] are published as of 01-01-2025, see spec.org, and compared against a 48-core estimated Gen12 system. All rights reserved. Power savings based on the Thermal Design Power of the systems. (6)The performance per watt advantages are based on internal power and performance measurements on similar configured high energy efficient servers and compared against an estimated 86-core Gen12 system. Source: HPE. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

# Intel® Xeon® 6 Processors for AI

World's Best  
CPU for AI

Up to 128 P-cores  
on 6900-series  
up to 86 P-cores on 6500/6700-series

More bandwidth & cache  
MRDIMM memory support  
Up to 504MB low latency LLC

AI accelerators built-in  
Intel® AMX, Intel® AVX-512,  
and Intel® AVX-2

Comprehensive SW suite  
AI development across classical ML  
and small GenAI models

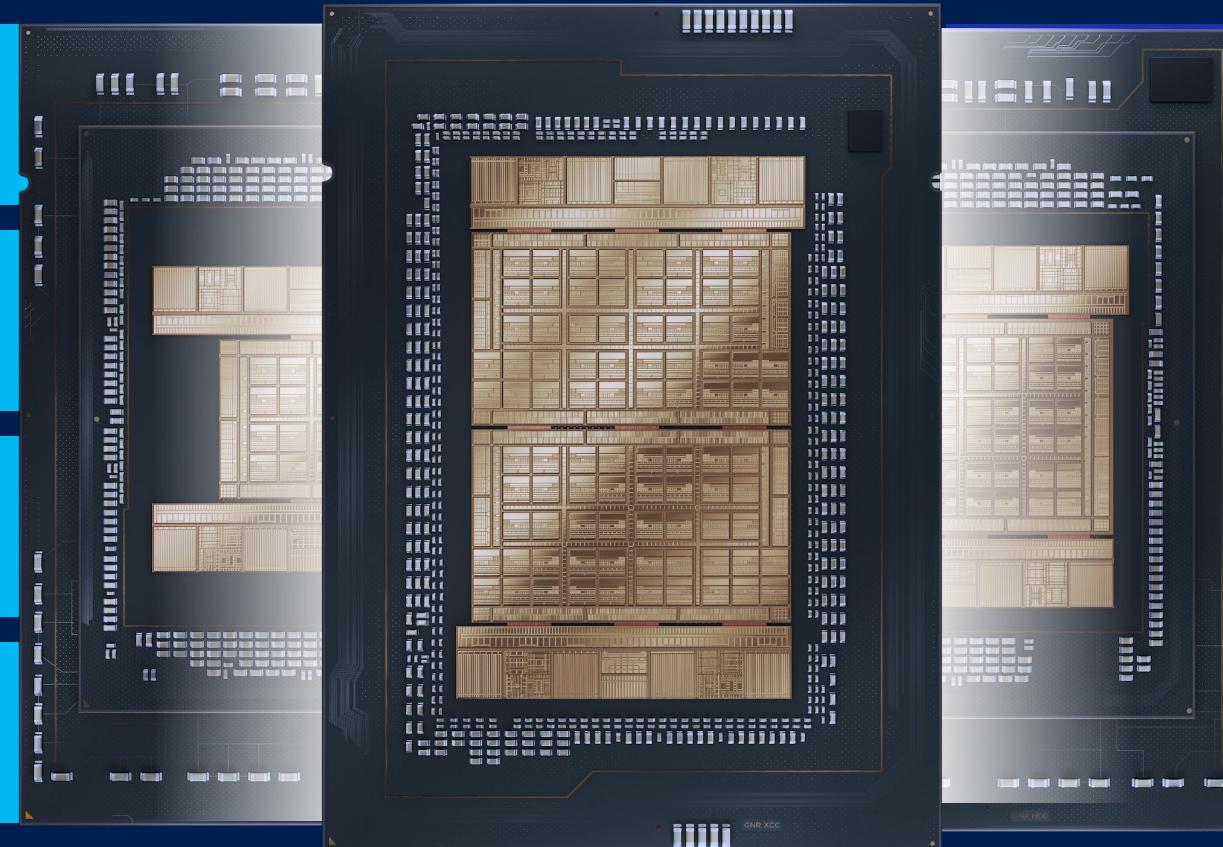
The Most  
Deployed Host CPU

Superior I/O performance  
up to 192 PCIe 5.0 lanes

High Single Threaded  
Performance  
With Intel's latest generation P-core

Top Tier Memory Support  
30% higher memory B/W with MRDIMMs  
Expandability with CXL 2.0

Ready for Deployment  
DC-MHS & NVIDIA MGX™  
form factors supported



Best CPU: See ISC 2024 section of [intel.com/performanceindex](http://intel.com/performanceindex) for workloads and configurations. Your results may vary. Intel technologies may require enabled hardware, software or service activation.

Most Deployed Source: IDC Server Tracker report, based on 1H'24 system volume.

\*NVIDIA logo and MGX are trademarks of NVIDIA and/or its subsidiaries

# Elevating Xeon for AI - A two prong approach

## The best CPU for General AI

ML workloads, Small & Mid Model Inference (<20B) & Small Model Fine-tuning

P-core perf/thread & core count

Built-in AI Accelerators (AMX)

High Bandwidth & Cache

SW Enablement

1.38x  
Perf Llama2-7B  
chat-hf

1.53x  
Perf on  
ResNet-50

1.27x  
Perf on  
BERT Large

1.21x  
Perf on  
DLRM

## Most deployed Host CPU for Large Scale AI

Foundational LLM Training, Model Fine-tuning, Large Model Inference

Superior I/O Performance

Single Thread Performance

Core Clock Frequency

Top Tier Memory Support

2S AMD EPYC 9755 (128 cores) vs. 2S Intel Xeon 6787P (86 cores)  
Llama2-7B is 2S AMD EPYC 9965 (192 cores) vs. 2S Intel Xeon 6980P (128 cores)

See [7A220 – 7A224] intel.com/processorclaims: Intel Xeon 6. Results may vary. \* Xeon 6: 96 PCIe gen5, 5<sup>th</sup> Gen Xeon: 80 PCIe gen5. ^See configuration [13d] in backup

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

## DeepSeek-R1-8B (BF16) on Intel Xeon 6

AMX - 12.4 tokens/sec

OpenVINO DeepSeek-R1-Distill-Llama-8B Chatbot

Chatbot

Solve the equation:  $2x + 5 = 15$ .

Submit 

Clear

AVX512 - 5.3 tokens/sec

OpenVINO DeepSeek-R1-Distill-Llama-8B Chatbot

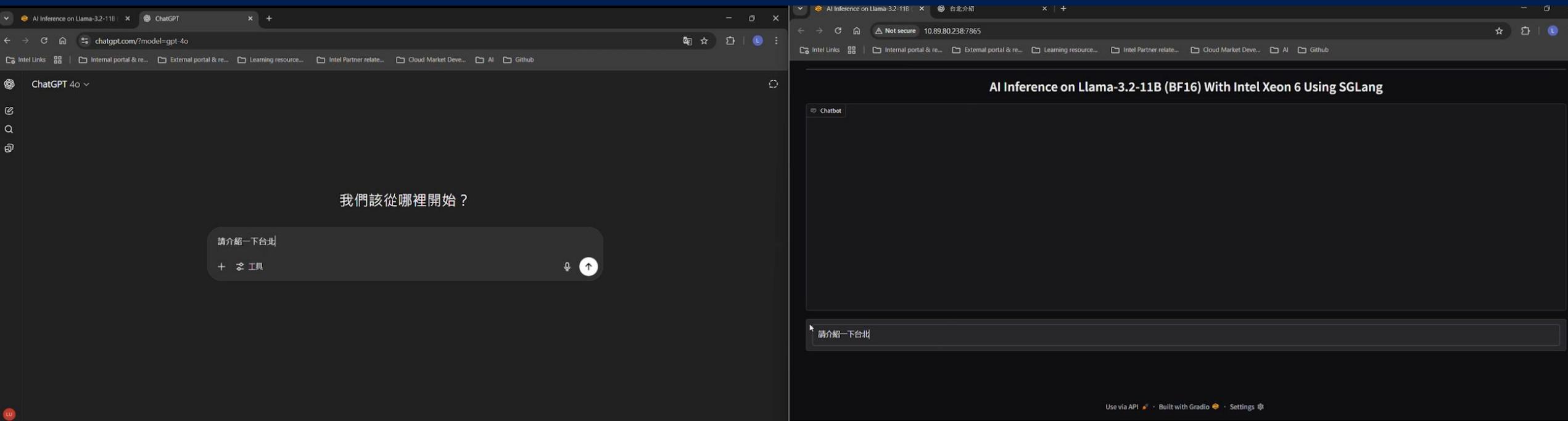
Chatbot

Solve the equation:  $2x + 5 = 15$ .

Submit 

Clear

# ChatGPT 4o v.s. Intel Xeon 6 CPU



# Call to action

Intel Xeon 6 with HPE Gen12 servers deliver the best TCO for your Enterprise applications



**4U, 2P**  
**HPE ProLiant Compute**  
**DL380a Gen12**  
Intel Xeon 6 processor



**1U, 1P**  
**HPE ProLiant Compute**  
**DL320 Gen12**  
Intel Xeon 6 processor



**2U, 1P**  
**HPE ProLiant Compute**  
**DL340 Gen12**  
Intel Xeon 6 processor



**HPE Synergy**  
**SY480 Gen12**  
Intel Xeon 6 processor

**VM/Data & Density**  
**optimized**



**1U, 2P**  
**HPE ProLiant Compute**  
**DL360 Gen12**  
Intel Xeon 6 processor



**SMB/Edge optimized**  
**2P Tower**  
**HPE ProLiant Compute**  
**ML350 Gen12**  
Intel Xeon 6 processor

**Big Data optimized**



**4U, 4P**  
**HPE ProLiant Compute**  
**DL580 Gen12**  
Intel Xeon 6 processor

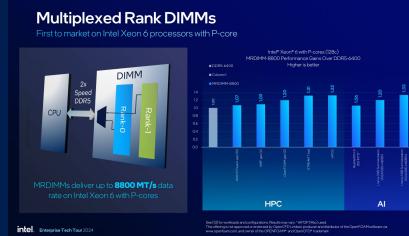


**2U, 2P**  
**HPE ProLiant Compute**  
**DL380 Gen12**  
Intel Xeon 6 processor

The Intel logo is displayed in white on a dark blue background. The word "intel" is written in a lowercase, sans-serif font. The "i" is unique, featuring a small blue square positioned above the top horizontal stroke of the letter. The "t" has a vertical stroke on its left side. The "e" and "l" are standard lowercase letters.

# Configurations

# Configuration: MRDIMM



[13] MRDIMMs Excel for Bandwidth Intensive Workloads

a) Ansys Fluent (aircraft\_wing\_14m, aircraft\_wing\_2m, combustor\_12m, combustor\_16m, combustor\_71m, exhaust\_system\_33m, fluidized\_bed\_2m, ice\_2m, landing\_gear\_15m, oil\_rig\_7m, pump\_2m, rotor\_3m, sedan\_4m)

6980P, MRDIMM: Test by Intel as of July 2024, 1 node, 2x Intel Xeon 6980P, HT On, Turbo On, SNC3, 1536 GB MRDIMM-8800, BIOS BHSDREL1.86B.0033.D40.2406180419, ucode=0x11000280, Ubuntu 24.04, Kernel 6.8.0, Ansys Fluent 2024R1  
6980P, DDR5: Test by Intel as of July 2024, 1 node, 2x Intel Xeon 6980P, HT On, Turbo On, SNC3, 1536 GB DDR5-6400, BIOS BHSDREL1.86B.0033.D40.2406180419, ucode=0x11000280, Ubuntu 24.04, Kernel 6.8.0, Ansys Fluent 2024R1

b) WRF (CONUS-12km, CONUS-2.5km)

6980P, MRDIMM : Test by Intel as of July 2024, 1 node, 2x Intel Xeon 6980P, HT On, Turbo On, SNC3, 1536 GB MRDIMM-8800, BIOS BHSDREL1.86B.0033.D40.2406180419, ucode=0x11000280, Ubuntu 24.04, Kernel 6.8.0, WRF v4.5.2, 6980P, DDR5: Test by Intel as of July 2024, 1 node, 2x Intel Xeon 6980P, HT On, Turbo On, SNC3, 1536 GB DDR5-6400, BIOS BHSDREL1.86B.0033.D40.2406180419, ucode=0x11000280, Ubuntu 24.04, Kernel 6.8.0, WRF v4.5.2,

c) OpenFOAM (motorbike-20m, motorbike-42m)

6980P, MRDIMM : Test by Intel as of July 2024, 1 node, 2x Intel Xeon 6980P, HT On, Turbo On, SNC3, 1536 GB MRDIMM-8800, BIOS BHSDREL1.86B.0033.D40.2406180419, ucode=0x11000280, Ubuntu 24.04, Kernel 6.8.0, OpenFOAM v2312  
6980P, DDR5: Test by Intel as of July 2024, 1 node, 2x Intel Xeon 6980P, HT On, Turbo On, SNC3, 1536 GB DDR5-6400, BIOS BHSDREL1.86B.0033.D40.2406180419, ucode=0x11000280, Ubuntu 24.04, Kernel 6.8.0, OpenFOAM v2312

d) Stream Triad

6980P with DDR5: Test by Intel as of July 2024, 1 node, 2x Intel Xeon 6980P, HT On, Turbo On, SNC3, 1536 GB DDR5-6400, BIOS BHSDREL1.86B.0033.D40.2406180419, ucode=0x11000280, Ubuntu 24.04, Kernel 6.8.0, App Version: v5.10  
6980P with MRDIMM : Test by Intel as of July 2024, 1 node, 2x Intel Xeon 6980P, HT On, Turbo On, SNC3, 1536 GB MRDIMM-8800, BIOS BHSDREL1.86B.0033.D40.2406180419, ucode=0x11000280, Ubuntu 24.04, Kernel 6.8.0, App Version: v5.10

e) HPCG:

6980P, MRDIMM : Test by Intel as of July 2024, 1 node, 2x Intel Xeon 6980P, HT On, Turbo On, SNC3, 1536 GB MRDIMM-8800, BIOS BHSDREL1.86B.0033.D40.2406180419, ucode=0x11000280, Ubuntu 24.04, Kernel 6.8.0, HPCG from Intel Optimized MKL v2024.1

6980P, DDR5: Test by Intel as of July 2024, 1 node, 2x Intel Xeon 6980P, HT On, Turbo On, SNC3, 1536 GB DDR5-6400, BIOS BHSDREL1.86B.0033.D40.2406180419, ucode=0x11000280, Ubuntu 24.04, Kernel 6.8.0, HPCG from Intel Optimized MKL v2024.1

f) ResNet50

6972P, MRDIMM, 1-node, 2x Intel(R) Xeon(R) 6972P, 96 cores, 500W TDP, HT On, Turbo On, Total Memory 1536GB (24x64GB MRDIMM 8800 MT/s [8800 MT/s]), BIOS BHSDCRB1.IPC.0033.D57.2406240014, microcode 0x11000290, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 1.7T SAMSUNG MZQL1T79HCIR-00A07, Ubuntu 24.04 LTS, 6.8.0-31-generic, Test by Intel as of 07/10/24

6972P, DDR5: 1-node, 2x Intel(R) Xeon(R) 6972P, 96 cores, HT On, Turbo On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s), BIOS BHSDCRB1.IPC.0033.D57.2406240014, microcode 0x11000290, Dual-Media 10G RDMA Ethernet Controller, 1x 1.7T SAMSUNG MZQL12T9HC/IR-00A0Z, Ubuntu 24.04 LTS, 6.8.0-31-generic. Test by Intel as of 07/18/24.

Software: ResNet50 v1.5 Inference: int8 bs=1 (sla=15ms). Dataset: ImageNet. Framework: PyTorch 2.4.0, IPPEX 2.4.0, OneDNN: v3.4.2. Modelzoo: <https://github.com/intel/ai-reference-models>

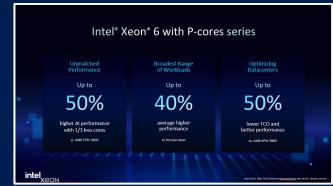
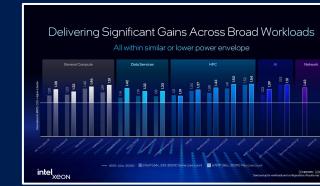
### g) Gen AI - LLM

9680P, MRDIMM: 1-node, 2x Intel Xeon 9680P, 128 cores, 500W TDP, HT On, Turbo On, Total Memory 1536GB (24x64GB MRDIMM 8800 MT/s [8800 MT/s]), BIOS BHSDCRB1.IPC.0033.D57.2406240014, microcode 0x11000290, 2x Ethernet Controller X710 for 10GBASE-T 1x 1.7T Micron 7400 MTEDKCC1T9TDZ, Ubuntu 24.04 LTS 6.8.0-31-generic, Test by Intel as of 07/11/24

X710-10G-10GBase-T, 1x 1.7 GHz Intel(R) Xeon(R) W-2135 Processor, 32GB DDR4-3200 MT/s, 3x 3.5" generic, Test by Intel as of 07/11/24.  
6980P, DDR: 1-node, 2x Intel Xeon 6980P, 128 cores, 500W TDP, HT On, Turbo On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6400 MT/s]), BIOS BHSDCRB1.IPC.0033.D57.2406240014, microcode 0x11000290, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 3.5 T. Micron Z450 M.2EDKCRB3T8TER Ubuntu 24.04 LTS 6.8.0-31-generic. Test by Intel as of 07/23/24.

Software: |lama3.8B|, int8, P90<=100ms, bs=1 x (1024/128), PyTorch: 2.3.0, IPEx: 2.3.0, OneDNN: v3.4.2, Modelzoo: <https://github.com/intel/ai-reference-models>

# Configurations



## [7A220] Up to 50% higher AI Performance with 1/3 less cores vs EPYC 9005

6787P: 1-node, 2x Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of November 2024.

9755: 1-node, 2x AMD EPYC 9755 128-Core Processor, 128 cores, 500W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 2x Ethernet Controller X550, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic., Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of January 2025.

Software: ResNet50 v1.5 inference, OpenVino 2024.4.0-16554-9c9778aba39-luocheng/mha\_fusion\_bhls, ww45 container, Python 3.8.20, BSX INT8, multi-instance, batched

## [7G20] Up to 40% higher average performance vs. 5th Gen Xeon:

[Geomean of Integer Throughput, Floating Point Throughput, Stream Triad, LINPACK, MongoDB(1S), MySQL(1S), Redis Memtier(1S), LAMMPS, WRF, BlackScholes, OpenFOAM, HPCG, BERT-Large, GPT-J 6B, Next Gen Firewall(1S), NGINX (1S) comparing 6787P vs. 8592+]

### a) General Compute : Integer throughput and Floating-point throughput

6787P: 1-node, 2x Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), microcode 0x1000311, 1x 1.7T 9660-16i, Ubuntu 24.04 LTS, 6.8.0-51-generic. Test by Intel as of January 2025.

6767P: 1-node, 2x Intel(R) Xeon(R) 6767P, 64 cores, 350W TDP, On [Off Linpack, Stream], Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), microcode 0x1000311, 1x 1.7T 9660-16i, Ubuntu 24.04 LTS, 6.8.0-51-generic. Test by Intel as of January 2025.

8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 3B08.TEL3P1, microcode 0x21000283, 2x Ethernet Controller X710 for 10GBASE-T, 1x 3.6T INTEL SSDPE2KX040T7, Ubuntu 24.04.1 LTS, 6.8.0-51-generic. Test by Intel as of Jan 2025.

Software: SPECcpu2017 (est): gcc14.2;

### b) Stream :

6787P: 1-node, 2x Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 512GB (16x32GB MRDIMM 8800 MT/s [8800 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x Ethernet Controller X550, 1.7T SAMSUNG MZ7L31T9, Ubuntu 24.04.1 LTS, 6.8.0-51-generic. Test by Intel as of December 2024.

6767P: 1-node, 2x Intel(R) Xeon(R) 6767P, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), BHSDCRB1.IPC.3544.P22.2411120403, microcode 0x1000341, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 1.7T Micron\_7450\_MTFDKCC1T9TFR, Ubuntu 24.04.1 LTS, 6.8.0-31-generic. Test by Intel as of November 2024..

8592+1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.3, microcode 0x21000240, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 1.7T SAMSUNG MZ1L21T9HCLS-00A07, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of November 2024.

STREAM: App Version: v5.10, Triad, icx:2025.0, running on physical cores

### c) Linpack :

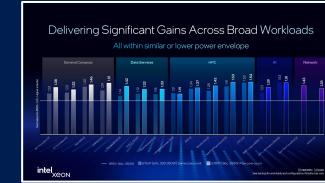
6787P: 1-node, 2x Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 6400 MT/s [6400 MT/s]), BIOS: BHSDREL1.IPC.3544.P15.2410300111, microcode: 0x81000341, 2x Ethernet Controller X550, 476.9G INTEL SSDPEKNW512G8, Ubuntu 24.04.1 LTS, 6.8.0-38-generic. Test by Intel as of December 2024.

6767P: 1-node, 2x Intel(R) Xeon(R) 6767P, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 6400 MT/s [6400 MT/s]), BHSDCRB1.IPC.3544.P22.2411120403, microcode 0x1000341, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 1.7T Micron\_7450\_MTFDKCC1T9TFR, Ubuntu 24.04.1 LTS, 6.8.0-31-generic. Test by Intel as of November 2024..

8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.3, microcode 0x21000240, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 1.7T SAMSUNG MZ1L21T9HCLS-00A07, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of November 2024.

HPL: App Version: Intel\_Optimized\_MKL\_v2024.1, running on physical cores.

# Configuration: Xeon 6787P/676xP vs Xeon 8592+



## [7D21] MongoDB:

6787P: 1-node, 2x (1 used) Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 6400 MT/s [6400 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x Ethernet Controller E830-CC for QSFP, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 4x 3.5T KIOXIA KCD8XPUG3T84, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of December 2024.

6760P: 1-node, 2x (1 used) Intel(R) Xeon(R) 6760P, 64 cores, 330W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 6400 MT/s [6400 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x Ethernet Controller E830-CC for QSFP, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 4x 3.5T KIOXIA KCD8XPUG3T84, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel December 2024.

8592+: 1-node, 2x (1 used) INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.3, microcode 0x21000240, 2x Ethernet Controller X710 for 10GBASE-T, 2x Ethernet Controller E810-C for QSFP, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, 4x 3.5T KIOXIA KCD8XPUG3T84, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of December 2024.

MongoDB: MongoDB 6.0.4 ycsb-0.17.0

## [7D23] MySQL HammerDB:

6787P: 1-node, 2x (1 used) Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 6400 MT/s [6400 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x Ethernet Controller E830-CC for QSFP, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 4x 3.5T KIOXIA KCD8XPUG3T84, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of December 2024.

6760P: 1-node, 2x (1 used) Intel(R) Xeon(R) 6760P, 64 cores, 330W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 6400 MT/s [6400 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x Ethernet Controller E830-CC for QSFP, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 4x 3.5T KIOXIA KCD8XPUG3T84, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel December 2024.

8592+: 1-node, 2x (1 used) INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.3, microcode 0x21000240, 2x Ethernet Controller X710 for 10GBASE-T, 2x Ethernet Controller E810-C for QSFP, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, 4x 3.5T KIOXIA KCD8XPUG3T84, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of December 2024.

MySQL : HammerDB 4.7, TPROC-C, on MySQL 8.0.33, multi-instance

## [7D22] Redis Memtier:

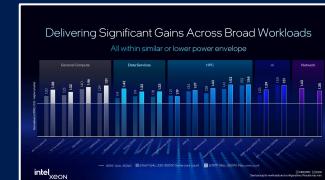
6787P: 1-node, 2x (1 used) Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x Ethernet Controller E830-CC for QSFP, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of December 2024.

6767P: 1-node, 2x (1 used) Intel(R) Xeon(R) 6767P, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x Ethernet Controller E830-CC for QSFP, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of December 2024.

8592+: 1-node, 2x (1 used) INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.3, microcode 0x21000240, 2x Ethernet Controller X710 for 10GBASE-T, 2x Ethernet Controller E810-C for QSFP, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of December 2024.

Redis Memtier: Redis: 7.0.5 Memtier: 1.4.0, multi-instance, 1 instance per core

# Configuration: Xeon 6787P/676xP vs Xeon 8592+



## [7H21] HPCG:

6787P: 1-node, 2x Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 512GB (16x32GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x Ethernet Controller X550, 1.7T SAMSUNG MZ7L31T9, Ubuntu 24.04.1 LTS, 6.8.0-51-generic. Test by Intel as of December 2024.

6767P: 1-node, 2x Intel(R) Xeon(R) 6767P, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), BHSDCRB1.IPC.3544.P22.2411120403, microcode 0x1000341, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 1.7T Micron\_7450\_MTFDKCC1T9TFR, Ubuntu 24.04.1 LTS, 6.8.0-31-generic. Test by Intel as of November 2024.

8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.3, microcode 0x21000240, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 1.7T SAMSUNG MZ1L21T9HCLS-00A07, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of November 2024.

HPCG: App Version: Intel\_Optimized\_MKL\_v2024.1, icx:2025.0, impi:2021.14, running on physical cores.

## [7H22] WRF:

6787P: 1-node, 2x Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 512GB (16x32GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x Ethernet Controller X550, 1.7T SAMSUNG MZ7L31T9, Ubuntu 24.04.1 LTS, 6.8.0-51-generic. Test by Intel as of December 2024.

6767P: 1-node, 2x Intel(R) Xeon(R) 6767P, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), BHSDCRB1.IPC.3544.P22.2411120403, microcode 0x1000341, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 1.7T Micron\_7450\_MTFDKCC1T9TFR, Ubuntu 24.04.1 LTS, 6.8.0-31-generic. Test by Intel as of November 2024.

8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.3, microcode 0x21000240, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 1.7T SAMSUNG MZ1L21T9HCLS-00A07, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of November 2024.

WRF: App Version: v4.5.2, conus2.5km, ifx:2025.0 impi:2021.14

## [7H23] BlackScholes:

6787P: 1-node, 2x Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 512GB (16x32GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x Ethernet Controller X550, 1.7T SAMSUNG MZ7L31T9, Ubuntu 24.04.1 LTS, 6.8.0-51-generic. Test by Intel as of December 2024.

6767P: 1-node, 2x Intel(R) Xeon(R) 6767P, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), BHSDCRB1.IPC.3544.P22.2411120403, microcode 0x1000341, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 1.7T Micron\_7450\_MTFDKCC1T9TFR, Ubuntu 24.04.1 LTS, 6.8.0-31-generic. Test by Intel as of November 2024.

8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.3, microcode 0x21000240, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 1.7T SAMSUNG MZ1L21T9HCLS-00A07, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of November 2024.

Black Scholes: App Version: v1.4, icx:2025.0, tbb:2022.0

## [7H24] LAMMPS:

6787P: 1-node, 2x Intel(R) Xeon(R) 6987P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 6400 MT/s [6400 MT/s]), BIOS: BHSDREL1.IPC.3544.P15.2410300111, microcode: 0x81000341, 2x Ethernet Controller X550, 476.9G INTEL SSDPEKNW512G8, Ubuntu 24.04.1 LTS, 6.8.0-38-generic. Test by Intel as of December 2024.

6767P: 1-node, 2x Intel(R) Xeon(R) 6767P, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 6400 MT/s [6400 MT/s]), BHSDCRB1.IPC.3544.P22.2411120403, microcode 0x1000341, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 1.7T Micron\_7450\_MTFDKCC1T9TFR, Ubuntu 24.04.1 LTS, 6.8.0-31-generic. Test by Intel as of November 2024.

8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.3, microcode 0x21000240, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 1.7T SAMSUNG MZ1L21T9HCLS-00A07, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of November 2024.

LAMMPS: App Version: v2024-03-07\_dev, cmkl:2025.0, icx:2025.0, impi:2021.14, tbb:2022.0, geomean of Atomic Fluid, Copper, DPD, Liquid\_crystal, Polyethylene, Protein, Stillinger-Weber, Tersoff, Water

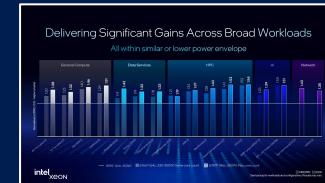
## [7H25] OpenFOAM:

This offering is not approved or endorsed by OpenCFD Limited, producer and distributor of the OpenFOAM software via [www.openfoam.com](http://www.openfoam.com), and owner of the OPENFOAM® and OpenCFD® trademark.

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

6787P: 1-node, 2x Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 512GB (16x32GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x Ethernet Controller X550, 1.7T SAMSUNG MZ7L31T9, Ubuntu 24.04.1 LTS, 6.8.0-51-generic. Test by Intel as of December 2024.

# Configuration: Xeon 6787P/676xP vs Xeon 8592+



## [7A21] Bert Large:

6787P: 1-node, 2x Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of November 2024.

6767P: 1-node, 2x Intel(R) Xeon(R) 6767P, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of December 2024.

8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.3, microcode 0x21000240, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 1.7T SAMSUNG MZ1L21T9HCLS-00A07, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of October 2024.

Software: BERT-Large inference, ww42 dlboost container, Python 3.10.14, Pytorch 2.5.0.dev20240903+cpu, IPEx 2.5.0+gitf5417a3, INT8, multi-instance, batched

## [7A28] GPT-J 6B

6787P: 1-node, 2x Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of November 2024.

6767P: 1-node, 2x Intel(R) Xeon(R) 6767P, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 1.7T Micron\_7450\_MTFDKCC1T9TFR, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of November 2024. 8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.3, microcode 0x21000240, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 1.7T SAMSUNG MZ1L21T9HCLS-00A07, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of October 2024. GPT-J 6B: Intel Model Zoo Optimized Benchmark, Docker 24.0.7; Pytorch/IPEx 2.6.0.dev20241016+cpu, Python 3.10.15. 1 instance per NUMA node; 2nd token P90 latency < 100ms, Chatbot: input token 128, output token 128. Summarization: input token 1024, output token 128.BSX, INT8

## [7N20] Next Gen Firewall (NGFW)

6787P: 1-node, 2x (1 socket used), Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo Off, Total Memory 512GB (16x32GB DDR5 6400 MT/s [6400 MT/s]), BIOS BHSDCRB1.IPC.3544.P22.2411122234, microcode 0x81000360, 1x I210 Gigabit Network Connection, 2x Ethernet Controller E810-C for QSFP, 1x 223.6G INTEL SSDSC2KB240G8, 1x 120M Disk, Ubuntu 22.04 LTS, 5.15.0-27-generic. NGFW 2403, gcc 11.3, Snort 3.1.36. Test by Intel as of 12/30/24.

8592+: 1-node, 2x (1 socket used), INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, HT On, Turbo Off, NUMA 2, Total Memory 512GB (16x32GB DDR5 5600 MT/s [5600 MT/s]), BIOS EGSDCRB1.SYS.0113.D55.240625, microcode 0x21000291, 1x Ethernet interface, 1x Ethernet Controller I225-LM, 1x 223.6G INTEL SSDSC2KB240G8, 1x 120M Disk, Ubuntu 22.04 LTS, 5.15.0-27-generic, NGFW 24.03, gcc 11.3, Snort 3.1.36, Test by Intel as of 12/31/24.

Performance measured on 1S . NGFW :Test run with Turbo Off, NGFW 24.03 / Snort 3.1.36/Hyperscan 5.4, HTTP 64K Packets / LightSpd Rules.

## [7N25] NGINX CPS

6787P: 1-node, 2x Intel(R) Xeon(R) 6787P (1 socket used), 86 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 8 [0], DSA 8 [0], IAA 8 [0], QAT 8 [0], Total Memory 512GB (16x32GB DDR5 6400 MT/s [6400 MT/s]), BIOS BHSDCRB1.IPC.3544.P07.2410011105, microcode 0x81000314, 10x Ethernet Controller E810-C for QSFP, 1x I210 Gigabit Network Connection, 1x 223.6G KINGSTON SA400S37240G, 1x 240M Disk, Ubuntu 22.04 LTS, 5.15.0-27-generic, Test by Intel as of 01/09/25..

8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+ (1 socket used), 64 cores, HT On, Turbo Off, NUMA 2, Integrated Accelerators Available [used]: DLB 5 [0], DSA 5 [0], IAA 2 [0], QAT 5 [0], Total Memory 512GB (16x32GB DDR5 5600 MT/s [5600 MT/s]), BIOS EGSDCRB1.SYS.0113.D55.2408280625, microcode 0x21000291, 1x Ethernet Controller I225-LM, 8x Ethernet Controller E810-C for QSFP, 1x 223.6G KINGSTON SUV400S37240G, 1x 240M Disk, Ubuntu 22.04 LTS, 5.15.0-27-generic, Test by Intel as of 12/29/24.

Software: NGINx v0.5.3, gcc 11.2.0, OpenSSL 3.3.2, TLSv1.3, Algorithm - ECDHE-X25519-ECDSA-P256

# Configurations: TCO Advantages, 128C: Xeon 6980P vs AMD 9755



[9T220] Intel® Xeon® 6 with P-cores delivers up to 24% lower total cost of ownership (TCO) than AMD EPYC 9755 based servers running a Stable Diffusion BS1 INT8 workload. Use 10 racks (100 servers) of Intel® Xeon® 6980P based servers running Stable Diffusion BS1 INT8 instead of 13 racks (130 servers) of AMD EPYC 9755 based servers and save 2618 MWh of energy, reduce carbon footprint by 1110 metric tons CO2, and save \$1651K in total cost of ownership over 4-years. Intel Xeon 6980P delivers 1.3x higher performance and 1.34x higher performance/watt per server vs Intel Xeon AMD EPYC 9755 on Stable Diffusion BS1 INT8.

6980P:1-node, 2x Intel(R) Xeon(R) 6980P, 128 cores, 500W TDP, HT On, Turbo On, Total Memory 1536GB (24x64GB MRDIMM 8800 MT/s [8800 MT/s]), BIOS 1.1, microcode 0x1000314, 2x Ethernet Controller X550, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of December 2024. 9755:1-node, 2x AMD EPYC 9755 128-Core Processor, 128 cores, 500W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 2x Ethernet Controller X550, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic., Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of January 2025. Software: Stable Diffusion inference, ww45 dlboost container, Python 3.10.14, 2.6.0.dev20241016+cpu, 2.6.0+git81c0d36, INT8, multi-instance, BS1

Assumptions: 1x 42U, 15KW Rack, supporting up to 20x 2U rack servers, 1x TOR switch, 1.6 PUE, kWh to kg CO2 factor 0.42394.

For 13 racks of AMD EPYC 9755 based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$3947k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$2715k, Energy use: 10545MWh, CO2 emissions: 4470 metric tons.

For 10 rack of Intel Xeon 6980P based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$2947k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$2064k: Energy use: 7926MWh, CO2 emissions: 3360 metric tons.

Costs based on Intel estimates, system pricing from major OEM, and information from [thinkmate.com](http://thinkmate.com) as of Feb 2025. Results may vary.

[9T221] Intel® Xeon® 6 with P-cores delivers up to 44% lower total cost of ownership (TCO) than AMD EPYC 9755 based servers running a Language Processing BERT-large workload. Use 10 racks (90 servers) of Intel® Xeon® 6980P based servers running Language Processing BERT-large instead of 15 racks (165 servers) of AMD EPYC 9755 based servers and save 4124 MWh of energy, reduce carbon footprint by 1748 metric tons CO2, and save \$3670K in total cost of ownership over 4-years. Intel Xeon 6980P delivers 1.85x higher performance and 1.51x higher performance/watt per server vs Intel Xeon AMD EPYC 9755 on Language Processing BERT-large.

6980P:1-node, 2x Intel(R) Xeon(R) 6980P, 128 cores, 500W TDP, HT On, Turbo On, Total Memory 1536GB (24x64GB MRDIMM 8800 MT/s [8800 MT/s]), BIOS 1.1, microcode 0x1000314, 2x Ethernet Controller X550, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of December 2024. 9755:1-node, 2x AMD EPYC 9755 128-Core Processor, 128 cores, 500W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 2x Ethernet Controller X550, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic., Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of January 2025. Software: BertLarge inference, ww42 dlboost container, Python 3.10.14, Pytorch 2.5.0.dev20240903+cpu, IPEx 2.5.0+gitf5417a3, BSX INT8, multi-instance, batched

Assumptions: 1x 42U, 15KW Rack, supporting up to 20x 2U rack servers, 1x TOR switch, 1.6 PUE, kWh to kg CO2 factor 0.42394.

For 15 racks of AMD EPYC 9755 based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$5010k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$3319k, Energy use: 12409MWh, CO2 emissions: 5261 metric tons.

For 10 rack of Intel Xeon 6980P based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$2652k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$2007k: Energy use: 8285MWh, CO2 emissions: 3512 metric tons.

Costs based on Intel estimates, system pricing from major OEM, and information from [thinkmate.com](http://thinkmate.com) as of Feb 2025. Results may vary.

# Configurations: TCO Advantages, 128C: Xeon 6980P vs AMD 9755



[9T222] Intel® Xeon® 6 with P-cores delivers up to 46% lower total cost of ownership (TCO) than AMD EPYC 9755 based servers running a Recommendation System DLRM workload. Use 10 racks (90 servers) of Intel® Xeon® 6980P based servers running Recommendation System DLRM instead of 17 racks (170 servers) of AMD EPYC 9755 based servers and save 5497 MWh of energy, reduce carbon footprint by 2330 metric tons CO2, and save \$4052K in total cost of ownership over 4-years. Intel Xeon 6980P delivers 1.87x higher performance and 1.65x higher performance/watt per server vs Intel Xeon AMD EPYC 9755 on Recommendation System DLRM.

6980P:1-node, 2x Intel(R) Xeon(R) 6980P, 128 cores, 500W TDP, HT On, Turbo On, Total Memory 1536GB (24x64GB MRDIMM 8800 MT/s [8800 MT/s]), BIOS 1.1, microcode 0x1000314, 2x Ethernet Controller X550, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of December 2024. 9755:1-node, 2x AMD EPYC 9755 128-Core Processor, 128 cores, 500W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 2x Ethernet Controller X550, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of January 2025. Software: DLRM v2 inference, ww42 dlboost container, Python 3.10.14, Pytorch 2.5.0.dev20240903+cpu, IPEX 2.5.0+gitf5417a3, BSX INT8, multi-instance, batched

Assumptions: 1x 42U, 15KW Rack, supporting up to 20x 2U rack servers, 1x TOR switch, 1.6 PUE, kWh to kg CO2 factor 0.42394.

For 17 racks of AMD EPYC 9755 based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$5162k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$3549k, Energy use: 13782MWh, CO2 emissions: 5843 metric tons.

For 10 rack of Intel Xeon 6980P based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$2652k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$2007k, Energy use: 8285MWh, CO2 emissions: 3512 metric tons.

Costs based on Intel estimates, system pricing from major OEM, and information from [thinkmate.com](http://thinkmate.com) as of Feb 2025. Results may vary.

[9T223] Intel® Xeon® 6787P delivers up to 28% lower total cost of ownership (TCO) than AMD EPYC 9755 based servers running a OpenFOAM geomean (2) workload. Use 10 racks (90 servers) of Intel® Xeon® 6980P based servers running OpenFOAM geomean (2) instead of 14 racks (126 servers) of AMD EPYC 9755 based servers and save 2623 MWh of energy, reduce carbon footprint by 1112 metric tons CO2, and save \$1887K in total cost of ownership over 4-years. Intel Xeon 6980P delivers 1.43x higher performance and 1.34x higher performance/watt per server vs Intel Xeon AMD EPYC 9755 on OpenFOAM geomean (2).

6980P with MRDIMM: 1-node, 2x Intel(R) Xeon(R) 6980P, 128 cores, 500W TDP, HT On, Turbo On, Total Memory 1536GB (24x64GB MRDIMM 8800 MT/s [8800 MT/s]), BIOS BHSDCRB1.IPC.3544.P15.2410232346, microcode 0x1000341, 2x Ethernet Controller X710 for 10GBASE-T, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of December 2024. 9755: 1-node, 2x AMD EPYC 9755 128-Core Processor, 128 cores, 500W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 2x Ethernet Controller X710 for 10GBASE-T, 1x 3.5T Micron\_7450\_MTFDKCB3T8TFR, Ubuntu 24.04.1 LTS, 6.8.0-48-generic. Test by Intel as of January 2025. OpenFOAM (Geomean of motorbike-20m, motorbike-42m) - Intel: App Version: v2312, icx:2025.0 impi:2021.14. AMD: App Version: v2312, <https://www.amd.com/en/developer/zen-software-studio/applications/spack/hpc-applications-openfoam.html>

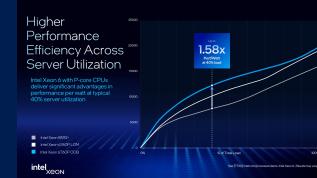
Assumptions: 1x 42U, 15KW Rack, supporting up to 20x 2U rack servers, 1x TOR switch, 1.6 PUE, kWh to kg CO2 factor 0.42394.

For 14 racks of AMD EPYC 9755 based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$3826k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$2726k, Energy use: 10949MWh, CO2 emissions: 4642 metric tons.

For 10 rack of Intel Xeon 6980P based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$2652k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$2013k, Energy use: 8326MWh, CO2 emissions: 3530 metric tons.

Costs based on Intel estimates, system pricing from major OEM, and information from [thinkmate.com](http://thinkmate.com) as of Feb 2025. Results may vary.

# Configurations: Performance Efficiency Intel Xeon 6760P vs Intel Xeon 8592+



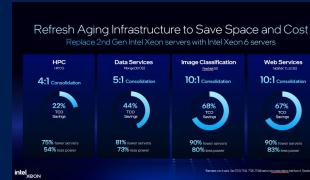
[7T20] Up to 1.58x higher Performance per Watt with Intel Xeon 6760P processor vs. Intel Xeon 8592+ at a typical 40% server utilization point

6760P: 1-node, 2x Intel(R) Xeon(R) 6760P, 64 cores, 330W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 6400 MT/s [6400 MT/s]), BIOS 1.0b, microcode 0x1000380, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, CentOS Stream 9, 5.14.0-529.el9.x86\_64. Test by Intel as of January 2025.

8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.3, microcode 0x21000240, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, CentOS Stream 9, 5.14.0-529.el9.x86\_64. Test by Intel as of January 2025.

Software Config: Power Efficiency workload

# Configurations: Server Consolidation, 86C Xeon 6787P vs 28C Xeon 8280



[7T21] Intel® Xeon® 6 with P-cores delivers up to 68% lower total cost of ownership (TCO) than 8280 based servers running a ResNet50 workload. Use 10 racks (100 servers) of Intel® Xeon® 6787P based servers running ResNet50 instead of 50 racks (1000 servers) of Intel® Xeon® 8280 based servers and save 30044 MWh of energy, reduce carbon footprint by 12737 metric tons CO2, and save \$10392K in total cost of ownership over 4-years. Intel Xeon 6787P delivers 9.98x higher performance and 4.79x higher performance/watt per server vs Intel Xeon 8280 on ResNet50.

6787P: 1-node, 2x Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of November 2024. 8280: 1-node, 2x Intel(R) Xeon(R) Platinum 8280M CPU @ 2.70GHz, 28 cores, 205W TDP, HT On, Turbo On, Total Memory 768GB (24x32GB DDR4 3200 MT/s [2666 MT/s]), BIOS Intel(R) Xeon(R) Platinum 8280M CPU @ 2.70GHz, microcode 0x4003605, 2x Ethernet Connection X722 for 10GBASE-T, 1x 1.4T INTEL SSDPE21K015TA, Ubuntu 24.04.1 LTS, 6.8.0-48-generic. Test by Intel as of December 2024. ResNet50 v1.5 inference, OpenVino 2024.4.0-16554-9c9778aba39-luocheng/mha\_fusion\_bhls, ww45 container, Python 3.8.20, INT8, multi-instance, batched

Assumptions: 1x 42U, 15KW Rack, supporting up to 20x 2U rack servers, 1x TOR switch, 1.6 PUE, kWh to kg CO2 factor 0.42394.

For 50 racks of 8280 based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$0k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$15272k, Energy use: 37967MWh, CO2 emissions: 16096 metric tons.

For 10 rack of Intel Xeon 6787P based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$2817k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$2063k, Energy use: 7922MWh, CO2 emissions: 3358 metric tons.

Costs based on Intel estimates, system pricing from major OEM, and information from [thinkmate.com](http://thinkmate.com) as of Feb 2025. Results may vary.

[7T22] : Intel® Xeon® 6 with P-cores delivers up to 22% lower total cost of ownership (TCO) than 8280 based servers running a HPCG workload. Use 10 racks (120 servers) of Intel® Xeon® 6787P based servers running HPCG instead of 24 racks (480 servers) of Intel® Xeon® 8280 based servers and save 9889 MWh of energy, reduce carbon footprint by 4193 metric tons CO2, and save \$1627K in total cost of ownership over 4-years. Intel Xeon 6787P delivers 3.95x higher performance and 2.15x higher performance/watt per server vs Intel Xeon 8280 on HPCG.

6787P: 1-node, 2x Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 512GB (16x32GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x Ethernet Controller X550, 1.7T SAMSUNG MZ7L31T9, Ubuntu 24.04.1 LTS, 6.8.0-51-generic. Test by Intel as of December 2024. HPCG: App Version: Intel\_Optimized\_MKL\_v2024.1, icx:2025.0, impi:2021.14, running on physical cores. 8280: HPCG :Intel Xeon 8280L: Test by Intel as of April 2024, 1 node, 2x Intel Xeon 8280L, HT On, Turbo On, 384 GB DDR3-2933, ucode=0x5003605, Ubuntu 23.10, Kernel 6.5.0, BIOS SE5C620.86B.02.01.0017.110620230543. HPCG from MKL\_v2022.1.0. running on physical cores.

Assumptions: 1x 42U, 15KW Rack, supporting up to 20x 2U rack servers, 1x TOR switch, 1.6 PUE, kWh to kg CO2 factor 0.42394.

For 24 racks of 8280 based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$0k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$7340k, Energy use: 18299MWh, CO2 emissions: 7758 metric tons.

For 10 rack of Intel Xeon 6787P based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$3380k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$2334k, Energy use: 8410MWh, CO2 emissions: 3565 metric tons.

Costs based on Intel estimates, system pricing from major OEM, and information from [thinkmate.com](http://thinkmate.com) as of Feb 2025. Results may vary.

# Configurations: Server Consolidation, 86C Xeon 6787P vs 28C Xeon 8280



[7T23] Intel® Xeon® 6 with P-cores delivers up to 26% lower total cost of ownership (TCO) than 8280 based servers running a Server Side Java Throughput workload. Use 10 racks (100 servers) of Intel® Xeon® 6787P based servers running Server Side Java Throughput instead of 21 racks (420 servers) of Intel® Xeon® 8280 based servers and save 7551 MWh of energy, reduce carbon footprint by 3202 metric tons CO2, and save \$1673K in total cost of ownership over 4-years. Intel Xeon 6787P delivers 4.09x higher performance and 1.9x higher performance/watt per server vs Intel Xeon 8280 on Server Side Java Throughput.

6787P: 1-node, 2x Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 6400 MT/s [6400 MT/s]), BIOS 3A08.OCT001, microcode 0x11000311, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, CentOS Stream 9, 5.14.0-467.el9.x86\_64. Test by Intel as of January 2025. 8280: 1-node, 2x Intel(R) Xeon(R) Platinum 8280L CPU @ 2.70GHz, 28 cores, 205W TDP, HT On, Turbo On, Total Memory 768GB (24x32GB DDR4 3200 MT/s [2666 MT/s]), BIOS Intel(R) Xeon(R) Platinum 8280L CPU @ 2.70GHz, microcode 0x4003604, 2x Ethernet Connection X722 for 10GBASE-T, 2x Ethernet Controller E810-C for QSFP, 1x 1.4T INTEL SSDPE21K015TA, 1x 1.5T INTEL SSDPE2KE016T7, CentOS Stream 9, 5.14.0-333.el9.x86\_64. Test by Intel as of December 2024. Server-side-java workload, JDK 23.0.2

Assumptions: 1x 42U, 15KW Rack, supporting up to 20x 2U rack servers, 1x TOR switch, 1.6 PUE, kWh to kg CO2 factor 0.42394.

For 21 racks of 8280 based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$0k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$6361k, Energy use: 15541MWh, CO2 emissions: 6588 metric tons.

For 10 rack of Intel Xeon 6787P based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$2617k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$2072k: Energy use: 7989MWh, CO2 emissions: 3387 metric tons.

Costs based on Intel estimates, system pricing from major OEM, and information from [thinkmate.com](http://thinkmate.com) as of Feb 2025. Results may vary.

[7T24] Intel® Xeon® 6 with P-cores delivers up to 29% lower total cost of ownership (TCO) than 8280 based servers running a MySQL OLTP (15 MySQL HammerDB) workload. Use 10 racks (180 servers) of Intel® Xeon® 6787P based servers running MySQL OLTP (15 HammerDB) instead of 39 racks (780 servers) of Intel® Xeon® 8280 based servers and save 14735 MWh of energy, reduce carbon footprint by 6247 metric tons CO2, and save \$3284K in total cost of ownership over 4-years. Intel Xeon 6787P delivers 4.29x higher performance and 2.73x higher performance/watt per server vs Intel Xeon 8280 on MySQL OLTP (15 HammerDB).

6787P: 1-node, 2x (1 used) Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 6400 MT/s [6400 MT/s]), BIOS 3A08.OCT001, microcode 0x11000311, 2x Ethernet Controller E830-CC for QSFP, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 4x 3.5T KIOXIA KCD8XPUG3T84, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of December 2024. 8280: 1-node, 2x (1 used) Intel(R) Xeon(R) Platinum 8280M CPU @ 2.70GHz, 28 cores, 205W TDP, HT On, Turbo On, Total Memory 768GB (24x32GB DDR4 3200 MT/s [2666 MT/s]), BIOS Intel(R) Xeon(R) Platinum 8280M CPU @ 2.70GHz, microcode 0x4003605, 2x Ethernet Controller E810-C for QSFP, 2x Ethernet Connection X722 for 10GBASE-T, 1x 1.4T INTEL SSDPE21K015TA, 4x 3.5T INTEL SSDPF2KX038TZ, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of December 2024. HammerDB 4.7, TPROC-C, on MySQL 8.0.33, multi-instance

Assumptions: 1x 42U, 15KW Rack, supporting up to 20x 2U rack servers, 1x TOR switch, 1.6 PUE, kWh to kg CO2 factor 0.42394.

For 39 racks of 8280 based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$0k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$11068k, Energy use: 23120MWh, CO2 emissions: 9802 metric tons.

For 10 rack of Intel Xeon 6787P based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$4833k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$2950k: Energy use: 8385MWh, CO2 emissions: 3555 metric tons.

Costs based on Intel estimates, system pricing from major OEM, and information from [thinkmate.com](http://thinkmate.com) as of Feb 2025. Results may vary.

# Configurations: Server Consolidation, 86C Xeon 6787P vs 28C Xeon 8280



[7T25] Intel® Xeon® 6 with P-cores delivers up to 44% lower total cost of ownership (TCO) than 8280 based servers running a MongoDB (1S) workload. Use 10 racks (200 servers) of Intel® Xeon® 6787P based servers running MongoDB (1S) instead of 54 racks (1080 servers) of Intel® Xeon® 8280 based servers and save 23036 MWh of energy, reduce carbon footprint by 9766 metric tons CO2, and save \$6720K in total cost of ownership over 4-years. Intel Xeon 6787P delivers 5.42x higher performance and 3.75x higher performance/watt per server vs Intel Xeon 8280 on MongoDB (1S).

6787P: 1-node, 2x (1 used) Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 6400 MT/s [6400 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 2x Ethernet Controller E830-CC for QSFP, 4x 3.5T KIOXIA KCD8XPUG3T84, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of December 2024. 8280: 1-node, 2x (1 used) Intel(R) Xeon(R) Platinum 8280M CPU @ 2.70GHz, 28 cores, 205W TDP, HT On, Turbo On, Total Memory 768GB (24x32GB DDR4 3200 MT/s [2666 MT/s]), BIOS Intel(R) Xeon(R) Platinum 8280M CPU @ 2.70GHz, microcode 0x4003605, 2x Ethernet Controller E810-C for QSFP, 2x Ethernet Connection X722 for 10GBASE-T, 1x 1.4T INTEL SSDPE21K015TA, 4x 3.5T INTEL SSDPF2KX038TZ, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of December 2024. YCSB, MongoDB 6.0.4, multi-instance

Assumptions: 1x 42U, 15KW Rack, supporting up to 20x 2U rack servers, 1x TOR switch, 1.6 PUE, kWh to kg CO2 factor 0.42394.

For 54 racks of 8280 based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$0k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$15250k, Energy use: 31437MWh, CO2 emissions: 13327 metric tons.

For 10 rack of Intel Xeon 6787P based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$5370k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$3159k: Energy use: 8400MWh, CO2 emissions: 3561 metric tons.

Costs based on Intel estimates, system pricing from major OEM, and information from [thinkmate.com](http://thinkmate.com) as of Feb 2025. Results may vary.

[7T26] Intel® Xeon® 6 with P-cores delivers up to 67% lower total cost of ownership (TCO) than 8280 based servers running a NGINX TLS (1S) workload. Use 10 racks (200 servers) of Intel® Xeon® 6787P based servers running NGINX TLS (1S) instead of 97 racks (1940 servers) of Intel® Xeon® 8280 based servers and save 35107 MWh of energy, reduce carbon footprint by 14883 metric tons CO2, and save \$17178K in total cost of ownership over 4-years. Intel Xeon 6787P delivers 9.71x higher performance and 5.81x higher performance/watt per server vs Intel Xeon 8280 on NGINX TLS (1S).

6787P: 1-node, 2x (1 used) Intel(R) Xeon(R) 6787P, 86 cores, 350W TDP, HT On, Turbo Off, Total Memory 1024GB (16x64GB DDR5 6400 MT/s [6400 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x Ethernet Controller E830-CC for QSFP, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 1.7T Micron\_7450\_MTFDKCC1T9TFR, Ubuntu 24.04.1 LTS, 6.8.0-48-generic. NGINX Webserver TLS1.3 ECDHE-X25519-RSA2K, NGINX Async v0.5.1, OpenSSL 3.1.3, IPP Crypto 2021.8, IPsec MB v 1.4, QAT Driver QAT20.L.1.2.30-00020, QAT\_Engine v 1.6.1; Test by Intel as of December 2024. 8280: 1-node, 2x Intel® Xeon® Platinum 8280M CPU @ 2.70GHz, 28 cores, HT On, Total Memory 768GB (12x64GB DDR4 3200 MT/s [2934 MT/s]), BIOS SE5C620.86B.02.01.0017.110620230543, microcode 0x4003605, 2x Ethernet Controller E810-C for QSFP, 2x Ethernet Connection X722 for 10GBASE-T, 1x 1.7T SAMSUNG MZWL1T9HBJR-00007, Ubuntu 22.04.4 LTS, 6.5.0-25-generic, Software NGINX,NGINX Async v0.5.1, OpenSSL 3.1.3,IPP Crypto 2021.8,IPsec MB v 1.4; Test by Intel as of Aug-2024.

Assumptions: 1x 42U, 15KW Rack, supporting up to 20x 2U rack servers, 1x TOR switch, 1.6 PUE, kWh to kg CO2 factor 0.42394.

For 97 racks of 8280 based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$0k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$25566k, Energy use: 42418MWh, CO2 emissions: 17983 metric tons.

For 10 rack of Intel Xeon 6787P based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$5370k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$3018k: Energy use: 7311MWh, CO2 emissions: 3099 metric tons.

Costs based on Intel estimates, system pricing from major OEM, and information from [thinkmate.com](http://thinkmate.com) as of Feb 2025. Results may vary.

# Configuration: Intel Xeon 6900P vs AMD EPYC 9005



## [9D220] MongoDB (1S):

6980P: 1-node, 2x (1 used) Intel(R) Xeon(R) 6980P, 128 cores, 500W TDP, HT On, Turbo On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6400 MT/s]), BIOS 1.1, microcode 0x1000314, 2x Ethernet Controller E830-CC for QSFP, 4x 3.5T KIOXIA KCD8XPUG3T84, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of January 2025.

9755: 1-node, 2x (1 used) AMD EPYC 9755 128-Core Processor, 128 cores, 500W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 1x Ethernet Controller E810-C for QSFP, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, 4x 3.5T KIOXIA KCD8XPUG3T84, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of January 2025.

Software: MongoDB 6.0.4 ycsb-0.17.0

## [9D221] Redis Memtier (1S):

6980P: 1-node, 2x (1 used) Intel(R) Xeon(R) 6980P, 128 cores, 500W TDP, HT On, Turbo On, Total Memory 1536GB (24x64GB MRDIMM 8800 MT/s [8800 MT/s]), BIOS 1.1, microcode 0x1000314, 2x Ethernet Controller E830-CC for QSFP, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of January 2025.

9755: 1-node, 2x (1 used) AMD EPYC 9755 128-Core Processor, 128 cores, 500W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 1x Ethernet Controller E810-C for QSFP, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, 4x 3.5T KIOXIA KCD8XPUG3T84, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of January 2025.

Software: 7.0.5 Memtier: 1.4.0, multi-instance, 1 instance per core

## [9D222] Redis Vector Similarity Search

6960P (sstpp 66c): 1-node, 2x Intel(R) Xeon(R) 6960P, 66 cores, 450W TDP, HT On, Turbo On, Total Memory 1536GB (24x64GB MRDIMM 8800 MT/s [8800 MT/s]), BIOS 1.0, microcode 0x11000311, 2x Ethernet Controller X710 for 10GBASE-T, 2x Ethernet Controller E830-CC for QSFP, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of November 2024. 9575F: 1-node, 2x AMD EPYC 9575F 64-Core Processor, 64 cores, 400 TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 2x I350 Gigabit Network Connection, 2x Ethernet Controller E810-C for QSFP, 1x 5.8T INTEL SSDPE2KE064T8, 1x 1.5T INTEL SSDPF21Q016TB, Ubuntu 22.04.5 LTS, 6.5.0-21-generic. Test by Intel as of January 2025. Software: Redis 8.0-m02, Redisearch 8, vectordb update.redisearch(1dcb421556448a285aaf84022302183749c459b7), Redis-scripts redisearch\_november(46f4d94eebab1efbf1836eafa30d2928a1ed4b3), 1 instance per core

## [9W220] NGINX TLS:

6972P: 1-node, 2x Intel(R) Xeon(R) 6972P, 96 cores, 500W TDP, HT On, Turbo Off, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6400 MT/s]), BIOS 1.1, microcode 0x1000314, 2x Ethernet Controller E830-CC for QSFP, 1x 1.7T SAMSUNG MZWL1T9HBJR-00007, Ubuntu 24.04 LTS, 6.8.0-49-generic. Test by Intel as of December 2024.

Software: NGINX Async v0.5.1, OpenSSL 3.1.3, IPP Crypto 2021.8, IPsec MB v 1.4

9655: 1-node, 2x AMD EPYC 9655 96-Core Processor, 96 cores, 400W TDP, SMT On, Boost Off, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 1x Ethernet Controller E810-C for QSFP, 1x 1.7T Micron\_7450\_MTFDKCC1T9TFR, Ubuntu 24.04 LTS, 6.8.0-47-generic. Test by Intel as of January 2025.

Software: NGINX Async v0.5.1, OpenSSL 3.1.3

# Configuration: Intel Xeon 6900P vs AMD EPYC 9005



## HPC:

6972P with MRDIMM: 1-node, 2x Intel(R) Xeon(R) 6972P, 96 cores, 500W TDP, HT On, Turbo On, Total Memory 1536GB (24x64GB MRDIMM 8800 MT/s [8800 MT/s]), BIOS BHSDCRB1.IPC.3544.P15.2410232346, microcode 0x1000341, 2x Ethernet Controller X710 for 10GBASE-T, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of December 2024.

9655: 1-node, 2x AMD EPYC 9655 96-Core Processor, 96 cores, 500W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 2x Ethernet Controller X710 for 10GBASE-T, 1x 3.5T Micron\_7450\_MTFDKCB3T8TFR, Ubuntu 24.04.1 LTS, 6.8.0-48-generic. Test by Intel as of January 2025.

[9H221] **HPCG: Up to 1.52x higher HPCG performance with Intel® Xeon® 6972P vs. AMD EPYC 9655**

Intel: App Version: Intel\_Optimized\_MKL\_v2024.1, icx:2025.0, impi:2021.14, running on physical cores. AMD: App Version: 2024\_10\_07, <https://www.amd.com/en/developer/zen-software-studio/applications/spack/hpcq-benchmark.html>

[9H222] **WRF (CONUS-2.5km): Up to 1.15x higher WRF performance Intel® Xeon® 6980P vs. AMD EPYC 9755**

Intel: App Version: v4.5.2, conus2.5km, ifx:2025.0 impi:2021.14. AMD: App Version: v4.5.2, conus2.5km, <https://www.amd.com/en/developer/zen-software-studio/applications/spack/hpc-applications-wrf.html>

[9H223] **LAMMPS (Geomean of Atomic Fluid, Copper, DPD, Liquid Crystal, Polyethylene, Protein, Stillinger-Weber, Tersoff, Water): Up to 1.23x higher LAMMPS performance with Intel Xeon 6972P vs. AMD EPYC 9655**

Intel: App Version: v2024-03-07\_dev, cmkl:2025.0, icx:2025.0, impi:2021.14, tbb:2022.0. AMD: App Version: v2024-03-07\_dev, cmkl:2025.0, icx:2025.0, impi:2021.14, tbb:2022.0

6980P with MRDIMM: 1-node, 2x Intel(R) Xeon(R) 6980P, 128 cores, 500W TDP, HT On, Turbo On, Total Memory 1536GB (24x64GB MRDIMM 8800 MT/s [8800 MT/s]), BIOS BHSDCRB1.IPC.3544.P15.2410232346, microcode 0x1000341, 2x Ethernet Controller X710 for 10GBASE-T, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of December 2024.

9755: 1-node, 2x AMD EPYC 9755 128-Core Processor, 128 cores, 500W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 2x Ethernet Controller X710 for 10GBASE-T, 1x 3.5T Micron\_7450\_MTFDKCB3T8TFR, Ubuntu 24.04.1 LTS, 6.8.0-48-generic. Test by Intel as of January 2025.

[9H224] **OpenFOAM (Geomean of motorbike-20m, motorbike-42m): Up to 1.43x higher OpenFOAM performance with Intel® Xeon® 6980P vs. vs. AMD EPYC 9755**

Intel: App Version: v2312, icx:2025.0 impi:2021.14. AMD : App Version: v2312, <https://www.amd.com/en/developer/zen-software-studio/applications/spack/hpc-applications-openfoam.html>

# Configuration: Intel Xeon 6900P vs AMD EPYC 9005



## [9A221] ResNet50:

6980P: 1-node, 2x (1 used) Intel(R) Xeon(R) 6980P, 128 cores, 500W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of November 2024.

9755:1-node, 2x AMD EPYC 9755 128-Core Processor, 128 cores, 500W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 2x Ethernet Controller X550, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic., Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of January 2025.

Software:ResNet50 v1.5 inference, OpenVino 2024.4.0-16554-9c9778aba39-luocheng/mha\_fusion\_bhls, ww45 container, Python 3.8.20, BSX INT8, multi-instance, batched

## [9A222] BertLarge:

6980P: 1-node, 2x (1 used) Intel(R) Xeon(R) 6980P, 128 cores, 500W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of November 2024.

9755:1-node, 2x AMD EPYC 9755 128-Core Processor, 128 cores, 500W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 2x Ethernet Controller X550, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic., Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of January 2025.

Software: BertLarge inference, ww42 dlboost container, Python 3.10.14, Pytorch 2.5.0.dev20240903+cpu, IPEX 2.5.0+gitf5417a3, BSX INT8, multi-instance, batched

## [9A223] Vision Transformer:

6980P: 1-node, 2x (1 used) Intel(R) Xeon(R) 6980P, 128 cores, 500W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of November 2024.

9755:1-node, 2x AMD EPYC 9755 128-Core Processor, 128 cores, 500W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 2x Ethernet Controller X550, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic., Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of January 2025.

Software : Vision Transformer inference, ww45 dlboost container, Python 3.10.14, 2.6.0.dev20241016+cpu, 2.6.0+git81c0d36, INT8, multi-instance, batched

## [9A224] DLRM:

6980P: 1-node, 2x (1 used) Intel(R) Xeon(R) 6980P, 128 cores, 500W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of November 2024.

9755:1-node, 2x AMD EPYC 9755 128-Core Processor, 128 cores, 500W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 2x Ethernet Controller X550, 1x 3.5T SAMSUNG MZWLJ3T8HBL5-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic., Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of January 2025.

Software:DLRM v2 inference, ww42 dlboost container, Python 3.10.14, Pytorch 2.5.0.dev20240903+cpu, IPEX 2.5.0+gitf5417a3, BSX INT8, multi-instance, batched

# Configurations: TCO Advantages, 64C: Intel Xeon 676xP vs AMD EPYC 9535



[7T221] Intel® Xeon® 6 with P-cores delivers up to 52% lower total cost of ownership (TCO) than AMD EPYC 9535 based servers running a Vision Transformer BSN INT8 workload. Use 10 racks (140 servers) of Intel® Xeon® 6760P based servers running Vision Transformer BSN INT8 instead of 17 racks (289 servers) of AMD EPYC 9535 based servers and save 5788 MWh of energy, reduce carbon footprint by 2454 metric tons CO2, and save \$6367K in total cost of ownership over 4-years. Intel Xeon 6760P delivers 2.09x higher performance and 1.73x higher performance/watt per server vs Intel Xeon AMD EPYC 9535 on Vision Transformer BSN INT8.

6760P: 1-node, 2x Intel(R) Xeon(R) 6760P, 64 cores, 330W TDP, HT On, Turbo On, Total Memory 1024GB (16x64GB MRDIMM 8800 MT/s [8000 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 1x 1.7T Micron\_7450\_MTFDKCC1T9TFR, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of February 2025. 9535: 1-node, 2x AMD EPYC 9535 64-Core Processor, 64 cores, 300W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 2x Ethernet Controller X550, 1x 3.5T SAMSUNG MZWLJ3T8HBLS-00007, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Test by Intel as of January 2025. Software: Vision Transformer: Vision Transformer inference, ww45 dlboost container, Python 3.10.14, 2.6.0.dev20241016+cpu, 2.6.0+git81c0d36, INT8, multi-instance, batched

Assumptions: 1x 42U, 15KW Rack, supporting up to 20x 2U rack servers, 1x TOR switch, 1.6 PUE, kWh to kg CO2 factor 0.42394.

For 17 racks of AMD EPYC 9535 based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$7359k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$4807k, Energy use: 13999MWh, CO2 emissions: 5935 metric tons.

For 10 rack of Intel Xeon 6760P based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$3285k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$2514k, Energy use: 8210MWh, CO2 emissions: 3481 metric tons.

Costs based on Intel estimates, system pricing from major OEM, and information from [thinkmate.com](http://thinkmate.com) as of Feb 2025. Results may vary.

[7T223] Intel® Xeon® 6 with P-cores delivers up to 41% lower total cost of ownership (TCO) than AMD EPYC 9535 based servers running a NGINX TLS (1S) workload. Use 10 racks (200 servers) of Intel® Xeon® 6760P based servers running NGINX TLS (1S) instead of 16 racks (320 servers) of AMD EPYC 9535 based servers and save 5258 MWh of energy, reduce carbon footprint by 2229 metric tons CO2, and save \$5462K in total cost of ownership over 4-years. Intel Xeon 6760P delivers 1.55x higher performance and 1.71x higher performance/watt per server vs Intel Xeon AMD EPYC 9535 on NGINX TLS (1S).

6760P: 1-node, 2x Intel(R) Xeon(R) 6760P, 64 cores, 330W TDP, HT On, Turbo Off, Total Memory 1024GB (16x64GB DDR5 6400 MT/s [6400 MT/s]), BIOS 3A08.QCT001, microcode 0x11000311, 2x Ethernet Controller E830-CC for QSFP, 1x 1.7T Micron\_7450\_MTFDKCC1T9TFR, Ubuntu 24.04.1 LTS, 6.8.0-48-generic. Test by Intel as of November 2024. Software: NGINX Async v0.5.1, OpenSSL 3.1.3, IPP Crypto 2021.8, IPsec MB v 1.4 9535: 1-node, 2x AMD EPYC 9535 64-Core Processor, 64 cores, 300W TDP, SMT On, Boost Off, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), BIOS 1.1, microcode 0xb002116, 1x Ethernet Controller E810-C for QSFP, 1x 1.7T Dell Ent NVMe AGN RI U.2 1.92TB, Ubuntu 24.04 LTS, 6.8.0-51-generic. Test by Intel as of January 2025. Software: NGINX Async v0.5.1, OpenSSL 3.1.3

Assumptions: 1x 42U, 15KW Rack, supporting up to 20x 2U rack servers, 1x TOR switch, 1.6 PUE, kWh to kg CO2 factor 0.42394.

For 16 racks of AMD EPYC 9535 based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$8368k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$4882k, Energy use: 12110MWh, CO2 emissions: 5134 metric tons.

For 10 rack of Intel Xeon 6760P based servers over 4-years, estimated as of Feb 2025: CapEx costs: \$4830k, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): \$2958k, Energy use: 6851MWh, CO2 emissions: 2904 metric tons.

Costs based on Intel estimates, system pricing from major OEM, and information from [thinkmate.com](http://thinkmate.com) as of Feb 2025. Results may vary.

# Notices and Disclaimers

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

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure. Results have been estimated or simulated.

Your costs and results may vary.

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

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

All product plans and roadmaps are subject to change without notice.

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

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

# Thank You