Accelerating HPC and AI with Intel

Bruno Riva – Intel
We remain confident we will **regain process leadership**

<table>
<thead>
<tr>
<th>Intel 7</th>
<th>Intel 4</th>
<th>Intel 3</th>
<th>Intel 20A</th>
<th>Intel 18A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shipping Now</td>
<td>Manufacturing Ready in 2H'22</td>
<td>Manufacturing Ready in 2H'23</td>
<td>Manufacturing Ready in 1H'24</td>
<td>Manufacturing Ready in 2H'24</td>
</tr>
</tbody>
</table>

**2022 Milestones**

- Meteor Lake CPU tile production stepping tape out
- Lead server product test wafers running in fab
- IP Test Wafers running in Fab
- 1H'22: Foundry Customer Test Chips
- 2H'22: First IP shuttle

Tick Tock development model enables execution innovation and **5 nodes in 4 years**
Expanding the Intel® Xeon® Processor Roadmap

- **Sapphire Rapids**: Intel 7, 2022
- **Emerald Rapids**: Intel 7, 2023
- **Granite Rapids**: Intel 3, 2024
- **Future Gen**: Intel 20A and beyond

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

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

**Future Gen**
- **Sierra Forest**: Intel 3, 2024
- **Future Gen**
HPC - AI Super Compute Strategy

- **4th Gen Intel® Xeon® Scalable processors**
  - codenamed Sapphire Rapids

- **Intel® Xeon® Max Processor**
  - codenamed Sapphire Rapids HBM

- **Intel® Data Center GPU**
  - codenamed Ponte Vecchio

*Open Ecosystem*

*Compute Workloads*

- HPC Workloads
- AI ML/DL Training
- Rendering Visualization
- Scientific Research
Intel® Xeon® CPU Max Series

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

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

Leading performance and efficiency for our customers
## Super Compute Product Portfolio

**HPC-AI**

<table>
<thead>
<tr>
<th>CPU</th>
<th>GPU BOARDS</th>
<th>OAM SUBSYSTEMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel Branded</td>
<td>Intel Branded</td>
<td>Vector Compute Platforms</td>
</tr>
<tr>
<td>CPU with HBM</td>
<td>PCIe Add In Cards</td>
<td>x4 GPU Subsystem</td>
</tr>
<tr>
<td>CPUs optimized for HPC</td>
<td>OAM Modules</td>
<td>x8 GPU Subsystem</td>
</tr>
</tbody>
</table>

**Vector Compute Platforms**

- **CPU with HBM**
- **PCie Add In Cards**
- **x4 GPU Subsystem**
- **x8 GPU Subsystem**
Accelerate with Xeon

4th Gen Intel® Xeon®
Scalable Processors
Accelerating Data Center Growth

Delivering Leading Platforms for our Customers and Partners

Innovating for the Future of the Data Center

Continuing to Advance Products and Services
Focus on Customer Real World Workloads

- Artificial Intelligence
- Networking 5G
- Storage
- HPC
- Data Analytics
Intel’s Differentiated Approach

- Workload-First
- CPU Cores + Built-In Accelerators Wins
- Open Software Ecosystem + oneAPI & AI Tools
- Higher Performance
- Increased Efficiency
- Optimal TCO
# 4th Gen Intel® Xeon® Scalable Processors

<table>
<thead>
<tr>
<th>Feature</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 8 socket scalability</td>
<td></td>
</tr>
<tr>
<td>Up to 60 cores per processor</td>
<td></td>
</tr>
<tr>
<td>Most built-in accelerators of any CPU</td>
<td></td>
</tr>
<tr>
<td>Increased memory bandwidth with DDR5</td>
<td></td>
</tr>
<tr>
<td>Increased I/O bandwidth with PCIe 5</td>
<td>80 lanes</td>
</tr>
<tr>
<td>Increased inter-socket bandwidth with UPI 2.0</td>
<td></td>
</tr>
<tr>
<td>Compute Express Link (CXL) 1.1</td>
<td></td>
</tr>
<tr>
<td>Hardware enhanced security</td>
<td></td>
</tr>
</tbody>
</table>
Accelerate with Xeon

Network (N)
Storage (S)

Most Workload Optimized SKUs on the Market

Expanded Options for Workload Optimized SKUs

Cloud (-P,-V,-M)
Network (N)
Storage (S)

1-Socket (-U)
Long-Life Use (IOT) (-T)
IMDB Analytics (-H)

HPC (w/HBM)
Liquid Cooled (-Q)
CSP Custom

>56%

Intel® Xeon® Processor Volume supports customer specific or workload specific demand*

* Source: Intel Xeon CPU billings on 3rd Gen Intel Xeon Processors, 2022 YTD
### 4th Gen Intel® Xeon® Scalable Processors

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workload-first approach to innovation, design, and delivery</td>
<td></td>
</tr>
<tr>
<td>Most built-in accelerators of any CPU on the market</td>
<td></td>
</tr>
<tr>
<td>Leading performance and efficiency for our customers</td>
<td></td>
</tr>
<tr>
<td>Industry’s most comprehensive Confidential Computing portfolio</td>
<td></td>
</tr>
</tbody>
</table>
Maximize the Effectiveness of Every Core

New Integrated IP Acceleration Engines

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

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

New Instruction Set Architecture (ISA)
- Intel® Advanced Matrix (AMX)
- Intel® Advanced Vector Extensions for vRAN
# Intel® Accelerator Engines

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

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

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

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

## Intel® Network Engines
- Intel® QuickAssist Technology (Intel® QAT)
- Intel® Dynamic Load Balancer (Intel® DLB)
- Intel® Data Streaming Accelerator (Intel® DSA)
- Intel® Advanced Vector Extensions 512 (Intel® AVX-512)

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

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

---

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

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

- **Compilers, libraries & analysis tools** support built-in accelerators to unleash performance, and fast training and inference for AI workloads.
  - Intel® oneAPI Math Kernel Library for HPC and technical compute
  - Intel® oneAPI Deep Neural Network Library for deep learning training + inference
  - Intel® Query Processing & Intel® Data Mover Library* for query processing, compression and data movement
  - Intel® VTune™ Profiler helps locate time-consuming parts of code and identify significant issues affecting application performance

Learn more: [Software for 4th Gen Intel Xeon & Max Series Processors](#)

Intel® QAT Accelerates cryptography

Intel® AMX Built-in AI acceleration engine

Intel® DLB For efficient load balancing across CPU cores

Intel® DSA Optimizes streaming data movement & transformation operations

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

Intel® CPL is open source. Open source Intel® DML in beta, v1 coming soon.
# Intel® Quick Assist Technology

**Acceleration Engine**

<table>
<thead>
<tr>
<th>Function</th>
<th>Business Value</th>
<th>Software Support</th>
<th>Use Cases</th>
</tr>
</thead>
</table>
| - Accelerated cryptography and data de/compression | - Accelerated compression/decompression offloading leads to greater CPU efficiency  
- More encrypted connections and web secure connections between devices with less overhead | - Intel® QAT Engine for acceleration of cryptographic operations | - Distributed storage systems, file systems, RocksDB, Data lakes, Apache Spark, Hadoop, NGINX, IPSec |

## Performance gains

### Network Secure Gateway
- Up to 84% fewer cores to achieve same connections/s on NGINX with built-in QAT vs. out-of-the-box software

### Enterprise Storage and Data Analytics
- Up to 95% fewer cores and 2x higher level 1 compression throughput leveraging integrated QAT vs. prior generation

## Intel® Data Streaming Accelerator

**Acceleration Engine**

<table>
<thead>
<tr>
<th>Function</th>
<th>Optimizing streaming data movement and transformation operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Value</td>
<td>Accelerated data protection for NVMe/TCP improving efficiency for data storage applications via CPU offload</td>
</tr>
<tr>
<td>Software Support</td>
<td>Intel® Data Mover Library</td>
</tr>
<tr>
<td>Use Cases</td>
<td>Virtualization, fast replication across non-transparent bridge, ERP, In-Memory Databases</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Performance gains vs not using these accelerators</th>
<th>Performance gains vs prior generation products</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Integrity (Throughput)</strong></td>
<td>Up to <strong>1.7x</strong> higher IOPs for large packet sequential reads with built-in Intel® DSA vs. ISA-L software</td>
</tr>
<tr>
<td><strong>Data Integrity (Throughput &amp; Latency)</strong></td>
<td>Up to <strong>1.6x</strong> higher IOPs and <strong>37%</strong> latency reduction for large packet sequential reads with built-in Intel® DSA vs. prior generation</td>
</tr>
</tbody>
</table>

### Intel® Dynamic Load Balancer

**Acceleration Engine**

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

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

**Software Support**
- Intel® Data Mover Library

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

### Performance gains vs not using these accelerators

<table>
<thead>
<tr>
<th>Use Cases</th>
<th>Performance gains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microservices</td>
<td>Up to 96% lower latency at the same throughput with built-in Intel® DLB vs. software for Istio ingress gateway</td>
</tr>
</tbody>
</table>

### Performance gains vs prior generation products

<table>
<thead>
<tr>
<th>Use Cases</th>
<th>Performance gains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microservices</td>
<td>Up to 89% lower latency and 57% lower CPU utilization at same core count with built-in Intel® DLB vs. prior generation</td>
</tr>
</tbody>
</table>

## Intel® Advanced Matrix Extensions

### Acceleration Engine

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

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

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

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

### Performance gains vs prior generation products

<table>
<thead>
<tr>
<th>Feature</th>
<th>Performance Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Speech Recognition Inference</strong></td>
<td>Up to 8.6x</td>
</tr>
<tr>
<td>Higher speech recognition inference performance with built-in Intel® Advanced Matrix Extensions (Intel® AMX) (BF16) vs. the prior generation (FP32)</td>
<td></td>
</tr>
<tr>
<td><strong>PyTorch Training and Inference</strong></td>
<td>Up to 10x</td>
</tr>
<tr>
<td>Higher PyTorch for both real-time inference and training performance with built-in Intel® Advanced Matrix Extensions (Intel® AMX) (BF16) vs. the prior generation (FP32)</td>
<td></td>
</tr>
</tbody>
</table>


**Intel® In-Memory Advanced Analytics Accelerator**

**Acceleration Engine**

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

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

**Software Support**
- Intel® Query Processing Library, Intel® Data Mover Library

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

**Performance gains vs not using these accelerators**

**Embedded Databases**
- Up to 2.1x higher RocksDB performance with built-in Intel® IAA vs. Zstd software

**Performance gains vs prior generation products**

**Embedded Databases**
- Up to 3x higher RocksDB performance with 66% latency reduction with built-in Intel® IAA vs. prior generation

See [D1] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.
A More Energy Efficient Server Architecture

Intel® Accelerator Engines Raise Performance Per Watt Ceilings

Baseline is 4th Gen Intel Xeon processor with No Acceleration

Relative Perf/W
Higher is Better

<table>
<thead>
<tr>
<th>Workload</th>
<th>IAA</th>
<th>AVX-512</th>
<th>DSA</th>
<th>AMX</th>
<th>QAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClickHouse (IAA vs LZ4)</td>
<td>1.12</td>
<td>2.01</td>
<td>1.61</td>
<td>8</td>
<td>1.22</td>
</tr>
<tr>
<td>ClickHouse (IAA vs ZSTD)</td>
<td>1.26</td>
<td>1.12</td>
<td>2.01</td>
<td>9.76</td>
<td>1.26</td>
</tr>
<tr>
<td>RocksDB (IAA vs ZSTD)</td>
<td>2.01</td>
<td>1.26</td>
<td>1.61</td>
<td>14.21</td>
<td>1.26</td>
</tr>
<tr>
<td>HPL Linpack (AVX-512 vs AVX2)</td>
<td>1.61</td>
<td>2.01</td>
<td>3.18</td>
<td>13.53</td>
<td>1.61</td>
</tr>
<tr>
<td>SPDK 128K QD64 (large media files) vs OOB</td>
<td>1.92</td>
<td>3.18</td>
<td>8</td>
<td>28.85</td>
<td>1.92</td>
</tr>
<tr>
<td>SPDK 16K QD256 (database requests) vs OOB</td>
<td>8</td>
<td>14.21</td>
<td>9.76</td>
<td>13.53</td>
<td>8</td>
</tr>
<tr>
<td>Real Time Image Recognition (AMX vs FP32) RN50</td>
<td>192</td>
<td>192</td>
<td>14.21</td>
<td>13.53</td>
<td>192</td>
</tr>
<tr>
<td>Batch Image Recognition (AMX vs FP32) RN50</td>
<td>8</td>
<td>192</td>
<td>14.21</td>
<td>13.53</td>
<td>8</td>
</tr>
<tr>
<td>Real Time Object Detection (AMX vs FP32) SSD-RN34</td>
<td>9.76</td>
<td>9.76</td>
<td>14.21</td>
<td>13.53</td>
<td>9.76</td>
</tr>
<tr>
<td>Batch Object Detection (AMX vs FP32) SSD-RN34</td>
<td>14.21</td>
<td>14.21</td>
<td>14.21</td>
<td>13.53</td>
<td>14.21</td>
</tr>
<tr>
<td>NGINX (65K cps Perf) QAT vs OOB</td>
<td>1.22</td>
<td>1.22</td>
<td>1.22</td>
<td>1.22</td>
<td>1.22</td>
</tr>
</tbody>
</table>

See backup for workloads and configurations. Results may vary.
## A More Cost-Efficient Server Architecture

**Benefits of Workload Optimized Products**

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

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

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

---

*Accelerate with Xeon*
### CPU + Accelerators: Differentiated Performance On Real Workloads

<table>
<thead>
<tr>
<th>Category</th>
<th>4th Gen Intel® Xeon® Scalable processors</th>
<th>Intel® Xeon® CPU Max Series</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Purpose Compute</td>
<td>53%</td>
<td></td>
</tr>
<tr>
<td>Artificial Intelligence</td>
<td>Up to 10x</td>
<td></td>
</tr>
<tr>
<td>Network 5G vRAN</td>
<td>Up to 2x</td>
<td></td>
</tr>
<tr>
<td>Networking &amp; Storage</td>
<td>Up to 2x</td>
<td></td>
</tr>
<tr>
<td>Data Analytics</td>
<td>Up to 3x</td>
<td></td>
</tr>
<tr>
<td>HPC</td>
<td>Up to 3.7x</td>
<td></td>
</tr>
</tbody>
</table>

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

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

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

** Intel Xeon CPU Max Series vs. Intel Xeon 8380
Architected to Accelerate Real World Workloads

**Cloud**
- Up to 89% performance increase with Intel® QAT vs. prior gen.¹¹
- “We were pleased to observe a 20% increase in performance over the current generation C2 VMs from Google Cloud in testing with one of our key workloads.”¹²

**AI**
- Up to 2.48x performance improvement with Intel® AMX vs. prior gen.¹⁴
- Up to 4x performance gain with Intel® AMX vs. prior gen.¹⁵
- “Intel’s [4th Gen Xeon processor] provides unprecedented levels of performance for critical graph intelligence tasks.”

**HPC**
- Up to 4.3x performance improvement with Intel AMX® on Intel Xeon Max Series vs. prior gen.¹⁶
- Up to 8.57x performance improvement on Intel Xeon Max Series vs. Intel E5V4.¹⁷
- “The reason we use the 4th Gen Intel® Xeon® processor as the building block for immersion born systems is really because of its unrivaled power and efficiency.”

**Security**
- Intel® SGX performs up to 4.6x higher vs. prior gen.¹³

**5G**
- “It is not just a software, it is not interfaces, it is not only radio. It is how we can build all the pieces in our architecture.”

**Additional Comments**
- Intel’s 4th Gen Xeon® processor provides unprecedented levels of performance for critical graph intelligence tasks.
- Intel® SGX performs up to 4.6x higher vs. prior gen.
- The reason we use the 4th Gen Intel® Xeon® processor as the building block for immersion born systems is really because of its unrivaled power and efficiency.
Accelerate with Xeon

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

Using Intel® AMX, Intel and Tencent demonstrated BERT model throughput gains compared to the previous generation. Now, Tencent can use the optimized BERT model to deliver better service experiences and to help reduce TCO.
Intel’s Most Sustainable Data Center Processor Ever

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

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

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

Available immersion cooling warranty rider for Intel® Xeon® processors

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

Manufactured at sites with state-of-the-art water reclaim
facilities that in 2021 recycled 2.8 billion gallons of water
Thank you!
Learn more

- Intel® Xeon® Scalable Processors
- 4\textsuperscript{th} Gen Intel® Xeon® Scalable Processors
- 4\textsuperscript{th} Gen Intel® Xeon® Scalable Processor product brief
- Intel® Accelerator Engines
- Software for 4\textsuperscript{th} gen Intel Xeon Scalable and Intel® Xeon® Max Series
CPU + Accelerators: Groundbreaking Efficiency

- **Higher Performance per Watt**
  - 2.9x average improvement of perf/watt with built-in accelerators*

- **Lower Power Bills**
  - up to 70W power savings per CPU with Optimized Power Mode

- **Lower TCO More Sustainable**
  - 55% lower TCO and power consumption while reducing 524K kg of CO2 emissions*

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

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

*For selected workloads.
Acceleration Delivers TCO Value

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

Sustainable Compute: Optimized Power Mode

- up to 20% lower processor power savings
- <5% performance impact for select workloads
- up to 70W power savings per CPU with Optimized Power Mode

Resources & Configurations
Resources and Configurations

Architecting to Accelerate Customer Workloads

Leading Performance with the most built – in accelerators

- **Up to 3.7x on memory-bound workloads** - Intel® Xeon® 8380: Test by Intel as of 10/7/2022. 1-node, 2x Intel® Xeon® 8380 CPU, HT On, Turbo On, Total Memory 256 GB (16x16GB 3200MT/s DDR4), BIOS Version SE5C620.86B.0101.0006.2207150335, ucode revision=0xd00375, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crc1.x86_64, Stream v5.10; Intel® Xeon® CPU Max Series: Test by Intel as of 9/2/2022. 1-node, 2x Intel® Xeon® CPU Max Series, HT On, Turbo On, SNC4, Total Memory 128 GB (8x16GB HBM2 3200MT/s), BIOS Version SE5C7411.86B.8424.D03.201804100444, ucode revision=0x2c000020, CentOS Stream 8, Linux version 5.19.0-rc6.0712.intel_next.1.x86_64+server, Stream v5.10
Resources and Configurations

Bringing the Architecture to Life (1 of 3)

- Get up to 53% faster results for life and material sciences for more effective research and Meet tight timelines with up to 45% faster results for options pricing
  - DeePM (Multi-Instance Training)
    - 8480+: Test by Intel as of 10/22/2022, 1-node, 2x Intel Xeon Platinum 8480+, Total Memory 512 GB, kernel 4.18.0-365.el8_3.x86_64, compiler gcc (GCC) 8.5.0, 20210514 (Red Hat 8.5.0-10), https://github.com/deepmodeling/deepmd-kit, Tensorflow 2.9, Harovod 0.24.0, oneCCL-2021.5.2, Python 3.9
  - 8380: Test by Intel as of 10/20/2022, 1-node, 2x Intel Xeon Platinum 8380 processor, Total Memory 256 GB, kernel 4.18.0-372.26.1.el8_6_x86_64, compiler gcc (GCC) 8.5.0, 20210514 (Red Hat 8.5.0-10), https://github.com/deepmodeling/deepmd-kit, Tensorflow 2.9, Harovod 0.24.0, oneCCL-2021.5.2, Python 3.9
  - LAMMPS
    - 8380: Test by Intel as of 9/30/2022, 1-node, 2x Intel Xeon Platinum 8380 CPU, HT On, Turbo On, NUMA configuration SNC2, Total Memory 256 GB (16x16GB 3200MT/s, Dual-Rank), BIOS Version SE5C620.86B.01.00006.2207150335, cuode revision=0xd000375, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6_x86_64, LAMMPS v2021-09-29 cmkl:2022.1.0, icc:2021.6.0, iomp:2021.6.0, threads/core: Turboon; BuildKnobs: -O3 -ip -xCORE-AVX512 -g -debug online-debug-info -qopt-zmm-usage=high;
  - Quantum Espresso (AUSURF112, Water_EXX)
    - 8480+: Test by Intel as of 9/22/2022, 1-node, 2x Intel Xeon Platinum 8480+, HT On, Turbo On, Total Memory 512 GB (16x32GB 4800MT/s, DDR5), BIOS Version SE5C7411.86B.8713.D03.2209091934, cuode revision=0xb20000070, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6_x86_64, LAMMPS v2021-09-29 cmkl:2022.1.0, icc:2021.6.0, iomp:2021.6.0, threads/core: Turbooff; BuildKnobs: -O3 -v -CORE-AVX512 -g -debug online-debug-info -qopt-zmm-usage=high;
    - 8380: Test by Intel as of 9/30/2022, 1-node, 2x Intel Xeon Platinum 8380 CPU, HT On, Turbo On, Total Memory 256 GB (16x16GB 3200MT/s, Dual-Rank), BIOS Version SE5C620.86B.01.00006.2207150335, cuode revision=0xd000375, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6_x86_64, Quantum Espresso 7.0, AUSURF112, Water_EXX
  - VASP (Geomet: CuC, Si, PdO4, PdO4, L221)
    - 8480+: Test by Intel as of 10/7/2022, 1-node, 2x 4th Gen Intel® Xeon® Platinum 8480+, HT On, Turbo On, SNC4, Total Memory 512 GB (16x32GB 4800MT/s, DDR5), BIOS Version SE5C7411.86B.8713.D03.2209091934, cuode revision=0xb20000070, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6_x86_64, LAMMPS v2021-09-29 cmkl:2022.1.0, icc:2021.6.0, iomp:2021.6.0, Usage=high; BuildKnobs: -O3 -v -CORE-AVX512 -g -debug online-debug-info -qopt-zmm-usage=high;
    - 8380: Test by Intel as of 10/7/2022, 1-node, 2x Intel® Xeon® 8380 CPU, HT On, Turbo On, NUMA configuration SNC2, Total Memory 256 GB (16x16GB 3200MT/s, Dual-Rank), BIOS Version SE5C620.86B.01.00006.2207150335, cuode revision=0xd000375, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6_x86_64, LAMMPS v2021-09-29 cmkl:2022.1.0, icc:2021.6.0, iomp:2021.6.0, Usage=high; BuildKnobs: -O3 -v -CORE-AVX512 -g -debug online-debug-info -qopt-zmm-usage=high;
  - GROMACS (geomet: benchMEM, benchPEP, benchPEP-h, benchRIB, hecosibio-3m, hecosibio-465k, hecosibio-6k, ion_channel_pme_large, lignocellulose_rf_large, mase_cubic, strw, water15m_pme_large, water15m_rf_large)
    - 8480+: Test by Intel as of 10/7/2022, 1-node, 2x 4th Gen Intel® Xeon® Scalable Processor, HT On, Turbo On, SNC4, Total Memory 512 GB (16x32GB 4800MT/s, DDR5), BIOS Version SE5C7411.86B.8713.D03.2209091934, cuode revision=0xb20000070, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6_x86_64, GROMACS v2021.4-SP
    - 8380: Test by Intel as of 10/7/2022, 1-node, 2x Intel® Xeon® 8380 CPU, HT On, Turbo On, NUMA configuration SNC2, Total Memory 256 GB (16x16GB 3200MT/s, Dual-Rank), BIOS Version SE5C620.86B.01.00006.2207150335, cuode revision=0xd000375, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6_x86_64, Converge GROMACS v2021.4-SP
Meet tight timelines with up to 45% faster results for options pricing

Binomial Options, Black Scholes, Monte Carlo

8480+: Test by Intel as of 10/7/2022. 1-node, 2x Intel Xeon Platinum 8480+, HT On, Turbo On, SNC4, Total Memory 512 GB (16x32GB 4800MT/s, DDR5), BIOS Version SE5C741.86B.8713.D03.2209091345, ucode revision=0x2b000070, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, Binomial Options v1.1, Black Scholes v1.4, Monte Carlo v1.2

8380: Test by Intel as of 10/7/2022. 1-node, 2x Intel Xeon Platinum 8380 CPU, HT On, Turbo On, Total Memory 256 GB (16x16GB 3200MT/s DDR4), BIOS Version SE5C620.B6E.01.01.0006.2207150335, ucode revision=0xda000075, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, Binomial Options v1.1, Black Scholes v1.4, Monte Carlo v1.2
Resources and Configurations

Bringing the Architecture to Life (3 of 3)

- Run social network microservices up to 88% faster for better user experiences.
  - 8480+:4 (1 master, 3 worker)-node, each-node, pre-production platform with 2x Intel(R) Xeon(R) Platinum 8480+ on QuantaGrid D54Q-2U with GB (16 slots/64GB/DDR5 4800) total memory, ucode 0x2b000081, HT on, Turbo on, CentOS release 8.4.2105, 6.0, 6x 2.9T INTEL SSDPE2KE032T7T, 59383G AVAGO JBOD, 2x Ethernet Controller X710 for 10G/BASE-T, 2x Ethernet Controller E80-C for QSFP, DeathStarBench Social Network; wk2 - load generator, ICE driver (CWL) 6.0.6, Cilium CNI - 11.4, Kubernetes - 12.14, ContainerD - 1.12.12, deathstarbench/social-network-microservices:0.0.8, nginx:thrift:xenial, memcached:1.6.7, mongo:4.4.6, redis 7.0.5, dataset: DeathStarBench/socialNetwork/datasets/social-graph/socfb-Reed98, test by Intel on 11/2/2022.
  - 8380:1-node, 2x Intel(R) Xeon(R) Platinum 8360Y on Intel Whitley with GB (16 slots/32GB/DDR4 3200) total memory, ucode 0x0000375, HT on, Turbo on, CentOS release 8.4.2105, 6.0, 6x 2.9T INTEL SSDPE2KG96G, 2x Ethernet Controller X710 for 10G/Ethernet - T, 1x Ethernet Controller E80-C for QSFP, DeathStarBench Social Network, wk2 - load generator, ICE driver (CWL) 6.0.6, Cilium CNI - 11.4, Kubernetes - 12.14, ContainerD - 1.12.12, deathstarbench/social-network-microservices:0.0.8, nginx:thrift:xenial, memcached:1.6.7, mongo:4.4.6, redis 7.0.5, dataset: DeathStarBench/socialNetwork/datasets/social-graph/socfb-Reed98, test by Intel on 11/2/2022.

- Offer personalized product recommendations up to 6.3x faster for smoother e-commerce.
  - 8480+:1-node, pre-production platform with 2x Intel Platinum 8480+ on Archer City with 1024 GB (16 slots/64GB/DDR5-4800) total memory, ucode 0x2b0000a1, HT on, Turbo on, CentOS Stream 8, 5.15.0, 6x INTEL SSDPE2KG96G (PT)/Samsung SSD 860 EVO ITB (TF), DLRM: inf: bsz [socket/instance], Inference: bsz: fp32=128, amx int8=128, Training bsz:fp32/amx int8=324 [1 instance, 1 socket], Criteo Terabyte Dataset, Framework: https://github.com/intel-innersource/frameworks.ai.pytorch-private-cpu/tree/47607d0bb9830935d8a7077344b28a998b766a66, Modelzoo: https://github.com/IntelAI/models/tree/spr-launch-public; PT:1.13, IPEX: 1.13, OneDNN: v2.7, test by Intel on 10/24/2022.
Resources and Configurations

A More Energy Efficient Server Architecture

Up to 112x and 125x higher performance/W using 4th Gen Xeon Scalable w/Intel Analytics Accelerator vs LZ4 and Zstd on ClickHouse

1 node, 2x pre-production 4th Gen Intel Xeon Scalable processor (60 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), Number of IAA device utilized=82 sockets active, on pre-production Intel platform and software, HT On, Turbo On, SNC off, Total Memory 1024GB (6x16GB DDR5 4800), microcode 0x2b0000ca, x3B4TB P5510 NVMe, 10GE e540-A12, Ubuntu 22.04.1 LTS, S5.I8-0.52-genic, QPL v3.4.6.4, gcc 11, ClickHouse 21.12, Star Schema Benchmark, tested by Intel November 2022.

Up to 2.0x higher performance/W using 4th Gen Intel Xeon Scalable w/Intel Analytics Accelerator vs RocksDB

1 node, 2x pre-production 4th Gen Intel Xeon Scalable Processor (60 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), on pre-production Intel platform and software, HT On, Turbo On, Total Memory 1024GB (6x16GB DDR5 4800), microcode 0x2b000000a, x3B4TB P5510 NVMe, 10GE e540-A12, Ubuntu 22.04.1 LTS, S5.I8-0.52-genic, QPL v3.4.6.4, gcc 11, ClickHouse 21.12, Star Schema Benchmark, tested by Intel November 2022.

Up to 1.6x higher performance/W using 4th Gen Intel® Xeon® Scalable Processor w/AVX-512 vs AVX2 on Linpack

1 node, 2x pre-production 4th Gen Intel® Xeon® Scalable Processor (60 cores) on pre-production Supermicro SYS-221H-TNR with 1024GB DDR5 memory (16x64 GB), microcode 0x2b000000c, HT On, Turbo On, SN4, CentOS Stream 8.2 2.3, tested by Intel November 2022.

Up to 3.18x and 192x higher performance/W using 4th Gen Intel® Xeon® Scalable Processor w/Data Streaming Accelerator vs out-of-box OS software on SPDK NVMe TCP

1 node, 2x pre-production 4th Gen Intel® Xeon® Scalable Processor (60 cores) with integrated Intel Data Streaming Accelerator (Intel ISA), ISA device utilized=4 sockets active, on pre-production Intel platform and software with 1024GB DDR5 memory (16x64 GB), microcode 0x2b000000d, x3B4TB P5510 NVMe, 10GE e540-A12, Ubuntu 22.04.1 LTS, S5.I8-0.52-genic, x3B4TB P5510 NVMe, 10GE e540-A12, One API BaseKit 22.02.062, One API HPC 22.02.0.191, Linpack ver 2.3, tested by Intel November 2022.
## Resources and Configurations

### A More Cost-Efficient Server Architecture

**RanNet3D Image Classification**

- **New Configuration**: 1 node, 2x pre-production 4th Gen Intel® Xeon® Scalable 8489H processor (60 cores) with Intel® Advanced Matrix Extensions (Intel® AMX), on pre-production SuperMicro SYS-220U-TNR with 1024GB DDR5 memory (16x64 GB), microcode 0xd000375, HT On, Turbo On, SNC Off, CentOS Stream 8, 3.10.16-301.63.7.66-64, 1x3.84TB P5510 NVMe, 10GbE x540AT2, Intel TF 3.10.408.1218, on production Intel platform and software, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR4 3200), microcode 0x2b0000c0, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR4 3200), microcode 0xd000375, 1x3.84TB P5510 NVMe, 10GbE x540AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generics, QPL v3.4.6.4, ZSTD v1.5.2, RocksDB v6.4.6 (db_bench), tested by Intel November 2022. Baseline: 1 node, 2x production 3rd Gen Intel Xeon Scalable 8380 Processor (40 cores) on SuperMicro SYS-220U-TNR, HT On, Turbo On, SNC Off, Total Memory 1024GB (16x64GB DDR4 3200), microcode 0xd000375, 1x3.84TB P5510 NVMe, 10GbE x540AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generics, ZSTD v6.4.6 (db_bench), tested by Intel November 2022.

**SuperMicro**

- **New Configuration**: 1 node, 2x pre-production 4th Gen Intel Xeon Scalable 8489H processor (60 cores) with integrated Intel® In-Memory Analytics Accelerator (Intel® IAA), on pre-production Intel platform and software, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR4 3200), microcode 0xd000375, 1x3.84TB P5510 NVMe, 10GbE x540AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generics, QPL v3.4.6.4, ZSTD v6.4.6 (db_bench), tested by Intel November 2022. Baseline: 1 node, 2x production 3rd Gen Intel Xeon Scalable 8380 Processor (40 cores) on SuperMicro SYS-220U-TNR, HT On, Turbo On, SNC Off, Total Memory 1024GB (16x64GB DDR4 3200), microcode 0xd000375, 1x3.84TB P5510 NVMe, 10GbE x540AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generics, ZSTD v6.4.6 (db_bench), tested by Intel November 2022.

**Intel® Advanced Matrix Extensions (Intel® AMX)**

- Costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $280.6K
- Energy use in kWh (4 year, per server): 52700, PUE 1.6
- For a 50 server fleet of 3rd Gen Intel® Xeon® Scalable 8380 processor (40 cores) on SuperMicro SYS-220U-TNR, estimated as of November 2022:
  - CapEx: $846.4K
  - OpEx: $507.2K
  - Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

**OpenFOAM**

- Costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $577.7K
- Energy use in kWh (4 year, per server): 32181, PUE 1.6
- For a 16 server fleet of Intel® Xeon® CPU Max Series 56 core, estimated as of December 2022:
  - CapEx: $550.5K
  - OpEx: $372.8K
  - Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

**OpenDAL**

- Costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $780.3K
- Energy use in kWh (4 year, per server): 58581, PUE 1.6
- For a 50 server fleet of 3rd Gen Intel® Xeon® Scalable 8380 Processor (40 cores) on SuperMicro SYS-220U-TNR, estimated as of November 2022:
  - CapEx: $280.6K
  - OpEx: $274.9K
  - Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

**OpenDAL**

- Costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $275.3K
- Energy use in kWh (4 year, per server): 44627, PUE 1.6
- For a 50 server fleet of 3rd Gen Intel® Xeon® Scalable 8380 Processor (40 cores) on SuperMicro SYS-220U-TNR, estimated as of November 2022:
  - CapEx: $1.64M
  - OpEx: $1.64M
  - Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

**Intel® Xeon® CPU Max Series 56 core**

- Costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $677.7K
- Energy use in kWh (4 year, per server): 6777K
- For a 17 server fleet of 4th Gen Intel® Xeon® Scalable 8380 Processor (60 cores), estimated as of November 2022:
  - CapEx: $970.4K
  - OpEx: $970.4K
  - Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

**Intel® Xeon® CPU Max Series 56 core**

- Costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $739.9K
- Energy use in kWh (4 year, per server): 7399K
- For a 11x3.5TB INTEL SSDPF2KX038TZ
  - Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

**Intel® Xeon® CPU Max Series 56 core**

- Costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $260.6K
- Energy use in kWh (4 year, per server): 2606K
- For a 17 server fleet of Intel® Xeon® CPU Max Series 56 core, estimated as of December 2022:
  - CapEx: $507.2K
  - OpEx: $507.2K
  - Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

**Intel® Xeon® CPU Max Series 56 core**

- Costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $780.3K
- Energy use in kWh (4 year, per server): 7803K
- For a 16 server fleet of Intel® Xeon® CPU Max Series 56 core, estimated as of December 2022:
  - CapEx: $550.5K
  - OpEx: $550.5K
  - Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394
Resources and Configurations

A More Cost-Efficient Server Architecture

SPDK
New Configuration: 1-node, 2x pre-production 4th Gen Intel Xeon Scalable processor (60 core) with integrated Intel Data Streaming Accelerator (Intel DSA), DSA device utilized=1 (1 active socket), on pre-production Intel platform and software with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000a1, 10GbE x40 AT2, Ubuntu 22.04.1 LTS, S.15.0.52-generic, 1x 1.92TB Intel® SSDSC2KG01, 4x 1.92TB Samsung PM1733, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 2x100GbE, FIO v3.30, SPDK 22.05, tested by Intel November 2022. Baseline: 1-node, 2x pre-production 3rd Gen Intel Xeon Scalable Processor (40 cores) on Supermicro SYS-220U-TNR, DDR4 memory total 1024GB (16x64 GB), HT On, Turbo On, SNC Off, microcode 0x000376, 10GbE x40-AT2, Ubuntu 22.04.1 LTS, S.15.0.52-generic, 1x 1.92TB Intel® SSDSC2KG01, 4x 1.92TB Samsung PM1733, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 2x100GbE, FIO v3.30, SPDK 22.05, tested by Intel November 2022.

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

For a 35 server fleet of 4th Gen Xeon 8490H (SPDK w/DSA), estimated as of November 2022:
CapEx costs: $743.8K
OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $220.1K
Energy use in kWh (4 year, per server): 43387, PUE 1.6
Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

Accelerate with Xeon