# Innovate **Faster** with **Integrated AI**

Taking AI from concept to production at scale has been a challenge. Making AI work across the entire end-to-end pipeline — whether on premises, in the cloud or using a hybrid approach — often meant additional expense coupled with difficulty recruiting the right talent.

## Why is AI so **Hard?**


**The Chasm Between Concept and Production**

**Specialized Hardware, Advanced Skills and Custom Tools**

**Complexity, Cost and Risk**

For business leaders struggling with how to scale AI across their businesses, reducing complexity is key. It's now more critical than ever for technology to deliver business value as organizations look to scale, drive down costs and deliver new services. Instead of customizing systems for new applications, which adds yet another layer of complexity, enterprises can achieve the performance they need to meet a wide variety of deployments — both today and in the future — with a scalable platform.

**EXECUTIVES WHO BELIEVE THEY NEED AI TO SUCCEED**
84%[1]

### Democratizing AI so everyone benefits

We are at a pivotal moment in AI technology. A new era of AI must provide an opportunity for the ecosystem at large to have access, visibility, transparency and trust. This is only possible with an approach that embraces industry standards and openness, is more cost-effective and breaks down proprietary, walled gardens.

**ENTERPRISE APPS THAT WILL USE EMBEDDED AI BY 2025**
90%[2]

# **AI** Starts with **Intel**.

Faster inference than 96-core AMD EPYC CPU by up to 1.8x[3]

Intel has built ready-to-deploy systems and optimized developer tools that run out of the box on the most widely used AI inference server platforms.

For those pushing the boundaries of what's possible, Intel's variety in domain-specific accelerators is unmatched. Habana deep learning training and inference processors, Intel's low-power VPUs, programmable FPGAs and discrete GPUs expand a CPU-based foundation with capabilities to address a variety of workloads.

- 4th Gen Intel® Xeon® Scalable Processors have the most built-in accelerators of any CPU in the world for key workloads including AI, along with end-to-end data science tools and an ecosystem of smart solutions. Intel solutions are optimized for cloud, enterprise, HPC, network, security and IoT workloads, with powerful cores and a wide range of frequency, feature and power levels.
- Intel® Advanced Matrix Extensions (Intel® AMX) accelerates AI capabilities on 4th Gen Intel Xeon Scalable processors, speeding up deep learning training and inference without additional hardware. This accelerator is ideal for natural language processing, recommendation systems and image recognition.
- Intel Software Guard Extensions (Intel SGX) is the most researched, updated and deployed confidential computing technology in data centers on the market today, with the smallest trust boundary of any confidential computing technology in the data center today.
- Intel Advanced Vector Extensions 512 (Intel AVX-512) can accelerate the preprocessing of unstructured data from multiple sources for training models, along with speeding up data movement for less time processing data sets. The Intel Extension for Scikit-learn, coupled with Intel AVX-512, also accelerates machine learning algorithms for both training and inference.
- Intel® Deep Learning Boost (Intel® DL Boost) is built in to run complex AI workloads on the same hardware as your existing workloads. Intel® Ultra Path Interconnect (Intel® UPI) channels increase platform scalability and improve inter-CPU bandwidth for I/O-intensive workloads.
- Habana® Gaudi®2 Accelerator delivers high-performance, high-efficiency deep learning training and inference, and it is particularly well-suited for the scale and complexity of generative AI and large language models.

## Intel® Security Solution for Fortanix Confidential AI Platform

Fortanix

- AI models and data can be shared without exposing intellectual property and sensitive data.
- Delivers a turnkey, enterprise-level, high-performance security solution without requiring application modifications.
- Addresses time-to-market concerns by providing a validated solution with an installation guide, containerized tools and sample workloads.

Up to 7.78x deep-learning inference performance[4]

The solution provides a pathway for deploying unmodified AI applications in secure enclaves in any cloud environment. It operates through three integrated pillars: Fortanix Confidential AI, Fortanix Confidential Computing Manager and new 4th Gen Intel® Xeon® Scalable processors with Intel® Software Guard Extensions (Intel® SGX).[5] As a turnkey service, it's great for immediate provisioning that's ready to use without requiring extensive AI skills. It also maintains model integrity by securing data and models in secure enclaves. This simplifies protection across all stages of the data security lifecycle so that organizations can easily get up and running with AI.

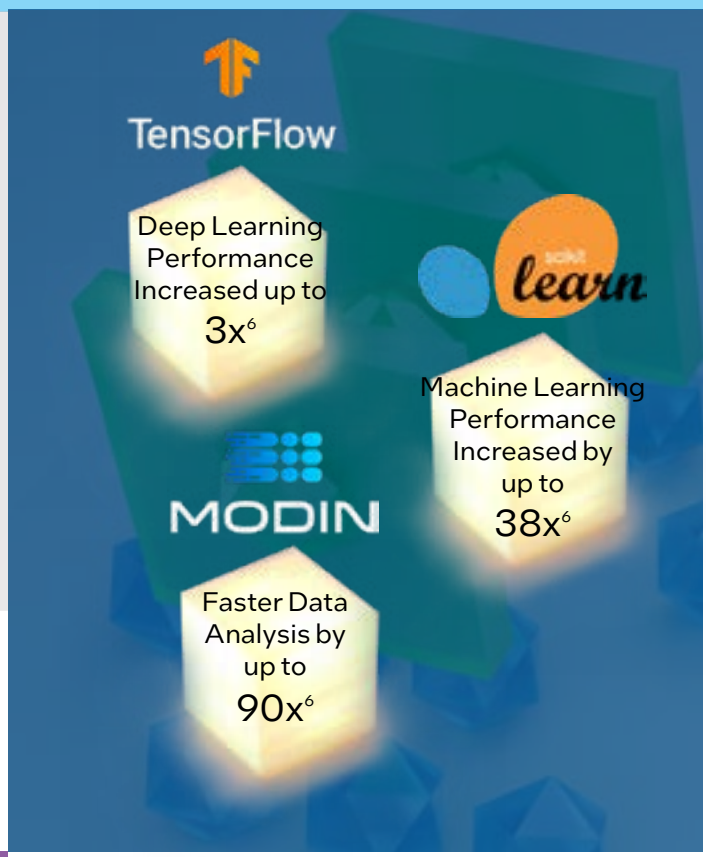
Find out more

# **Commitment** to Build AI into All Our Platforms, and Keep it **Open**.

Our approach and commitment to customers is to build AI into all our platforms, and to keep it open driving software optimizations upstream into AI/ML frameworks to promote programmability, portability and ecosystem adoption. AI as a workload needs to be accessible and part of every application. In this way, we will provide choice and compatibility across architectures, vendors and cloud platforms in support of an open accelerated computing ecosystem.

Our software platforms and heterogenous architectures consisting of CPUs, GPUs and deep learning accelerators make it easier for customers to quickly deploy AI at scale from the cloud to the network, out to edge and to the client. Our end-to-end AI pipeline lets developers write once and deploy anywhere and we deliver a unified programming model that makes it easy to go from a Xeon-based deployment to our GPUs, as well as to our dedicated accelerators. At the foundation of our platforms and solutions is trust. Customers can confidently secure diverse AI workloads in the data center and inference at the edge with confidential computing.

Solutions that run on Intel are the fast path to scale AI everywhere. Intel® AI optimizations for Spark, TensorFlow, PyTorch, scikit-learn, NumPy and XGBoost offer a substantial performance advantage,[6] while the Intel® Distribution of OpenVINO™ toolkit simplifies deep learning inference deployment for hundreds of pretrained models. And Intel® oneAPI toolkits allow maximum code reuse across stacks and architectures, with the ability for accelerators such as Intel AMX to run out-of-box with minimal or no code changes.[7] For more information, see www.intel.com/content/www/us/en/developer/topic-technology/artificial-intelligence/



## Key Intel® Open Source Optimizations for AI

**PyTorch:**
Develop and deploy models from research to production using Intel® optimizations in the default PyTorch framework and Intel® Extension for PyTorch.

**TensorFlow:**
Accelerate training and inference with upstreamed default optimizations and an extension for the latest performance improvements.

**Scikit-learn:**
Speed up your classical machine learning scikit-learn workflows by changing two lines of code.

**Modin:**
Scale your pandas workflows by changing a single line of code.

**Intel® AI Analytics Toolkit:**
Accelerate end-to-end data science and machine learning pipelines using Python-based tools and frameworks.

**Intel® Distribution of OpenVINO™ Toolkit:**
Deploy high-performance inference applications from devices to the cloud.

**BigDL:**
Seamlessly scale your AI models to big data clusters with thousands of nodes for distributed training or inference.

**Intel® Neural Compressor:**
Speed up AI inference without sacrificing accuracy through automated model compression.

## Meituan Accelerates Vision AI Inference Services and Optimizes Costs

Meituan needs to improve the throughput of its vision AI inference without compromising accuracy to support more intelligent operations. While discrete GPUs can meet performance requirements, their price is relatively high. For low-traffic longtail model inference services, CPUs are often more cost-effective. To accelerate AI inference, Meituan utilizes advanced hardware capabilities such as 4th Gen Intel® Xeon® Scalable processors and the built-in Intel® Advanced Matrix Extensions (Intel® AMX).

美团

**Up to 4.13X improvement**
in inference performance, converting models from FP32 to BF16[8]

**3X increase**
in overall efficiency of online resources and savings of 70% on service costs[8]

Hugging Face

ACCELERATED REAL-TIME INFERENCE by 5.7X[9]

AIBLE

CUT DATA EVALUATION TIME from weeks to 10 MINUTES PER DATASET[10]

Numenta

IMPROVED INFERENCE THROUGHPUT by 62X[11]

## Built-In accelerators in Intel® Xeon® Scalable processors


To enable new built-in accelerator features within a hyper-scaled environment, Intel supports the ecosystem with all of the most common cloud APIs, libraries and OS-level software. This results in more efficient CPU utilization, lower cloud electricity consumption and higher services ROI, while helping businesses achieve their sustainability goals.

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

AI is an incredibly powerful technology with untold potential, but it's still relatively immature. We must ensure AI technology advances responsibly. Industry, academia and global leaders must work together to shape our technological future, creating new possibilities that bring out the best in our human selves.

# Summon the future of AI.

### More Information

4th Gen Intel® Xeon® Scalable processors

Intel® AI and Deep Learning

Intel® AI News

[1]Accenture, November 19 2019. "AI: Built to Scale." https://www.accenture.com/us-en/insights/artificial-intelligence/ai-investments.

[2]Grand View Research, "Artificial Intelligence Market Size, Share & Trends Analysis Report By Solution, By Technology (Deep Learning, Machine Learning, Natural Language Processing, Machine Vision), By End Use, By Region, And Segment Forecasts, 2022 - 2030." https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market.

[3]DLRM recommendation system workload, BF16 data type. See https://edc.intel.com/content/www/us/en/products/performance/benchmarks/4th-generation-intel-xeon-scalableprocessors/.

[4]See configurations for more details.

CONFIGURATION DETAILS

TEST-1: Test by Intel as of 21 Nov 2022. 1-node, 2x Intel® Xeon® Platinum 8380 CPU @ 2.30GHz , 40 cores, HT Off, Turbo On, Total Memory 512 GB (16x32GB DDR4 3200 MT/s [run @3200 MT/s] ), BIOS version SE5C6200.86B.0022.D64.2105220049, ucode version 0xd000375, OS Version Ubuntu 22.04.1 LTS, kernel version 6.0.6-060006-generic, workload/benchmark Deep Learning inferencing in secure enclaves with Fortanix, framework version TensorFlow 2.11 , model name & version ResNet50v1.5/Bert-Large.

TEST-2: Test by Intel as of 21 Nov 2022. 1-node, 2x Intel® Xeon® Platinum 8480+ CPU @ 2.0GHz , 56 cores, HT Off, Turbo On, Total Memory 512 GB (16x32GB DDR5 4800 MT/s [run @4800 MT/s] ), BIOS version 3A05, ucode version 0x2b000070, OS Version Ubuntu 22.04.1 LTS, kernel version 6.0.6-060006-generic , workload/benchmark Deep Learning inferencing in secure enclaves with Fortanix, framework version TensorFlow 2.11 , model name & version ResNet50v1.5/Bert-Large.

[5]Intel® SGX is not vulnerable to most OS layer threats, and there are over 140,000 threats in the database today: https://cve.mitre.org.

[6]See intel.com/content/www/us/en/developer/articles/technical/code-changes-boost-pandas-scikit-learn-tensorflow.html for workloads and configurations. Results may vary.

[7]See technical documentation for software releases that support accelerators.

[8]For more complete information about performance and benchmark results, visit https://www.intel.com/content/www/us/en/customer-spotlight/stories/meituan-vision-ai-customer-story.html.

[9]See claim [A2] at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/4th-generation-intel-xeon-scalable-processors/.

[10]Visit https://enaible.aible.com/hubfs/Brochure_Aible_Intel_Overstock.pdf for more details.

[11]See [P6] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

Notices & 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.

Your costs and results may vary.

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

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

© 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.

0623/MH/MESH/353917-001US