Business Brief

4th Gen Intel® Xeon® Scalable Processor
AI and Machine Learning

Advance Insights with Artificial Intelligence

With Intel you can get better insights for critical business outcomes.

Taking AI from concept to production at scale has been a challenge. Advanced AI models once required specialized hardware, advanced skills and custom tools to turn data into business results. 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. 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.

84% of executives **BELIEVE THEY NEED AI TO SUCCEED**[1]

70% of data center AI inference **RUNS ON INTEL® XEON® PROCESSORS**[2]

90% by 2025 **OF ENTERPRISE APPS WILL USE EMBEDDED AI**[3]

Accelerate your AI with Intel technologies

4th Gen Intel® Xeon® Scalable processors have the most built-in accelerators of any CPU on the market to deliver performance and power efficiency advantages across the fastest-growing workload types in AI, analytics, networking, storage and HPC. With all-new Intel Advanced Matrix Extensions (Intel AMX), 4th Gen Intel Xeon Scalable processors have exceptional AI training and inference performance. To enable new built-in accelerator features, Intel supports the ecosystem with OS-level software, libraries and APIs.

With built-in accelerators and software optimizations, previous generation Intel® Xeon® Scalable processors have been shown to deliver leading performance per watt on targeted real-world workloads.[4] This results in more efficient CPU utilization, lower electricity consumption, and higher ROI, while helping businesses achieve their sustainability goals.

PERFORMANCE PROOFPOINT

5.7X TO 10X HIGHER PYTORCH REAL-TIME INFERENCE PERFORMANCE[5]

3.5X TO 10X HIGHER PYTORCH TRAINING PERFORMANCE[6]

with built-in Intel AMX (BF16) versus the prior generation (FP32)

Key AI use cases

Deep learning: Recommender systems and Natural Language Processing (NLP)

To create personalized user experiences that account for real-time behavioral signals and contextual queues, businesses can deploy deep learning-based recommender systems and leverage natural language processing while balancing TCO. Recommender systems help companies better target their individual customers through personalized recommendations, and natural language processing enables machines to understand text in a more meaningful way so that companies can better comprehend and meet customer needs.

THE NEED:

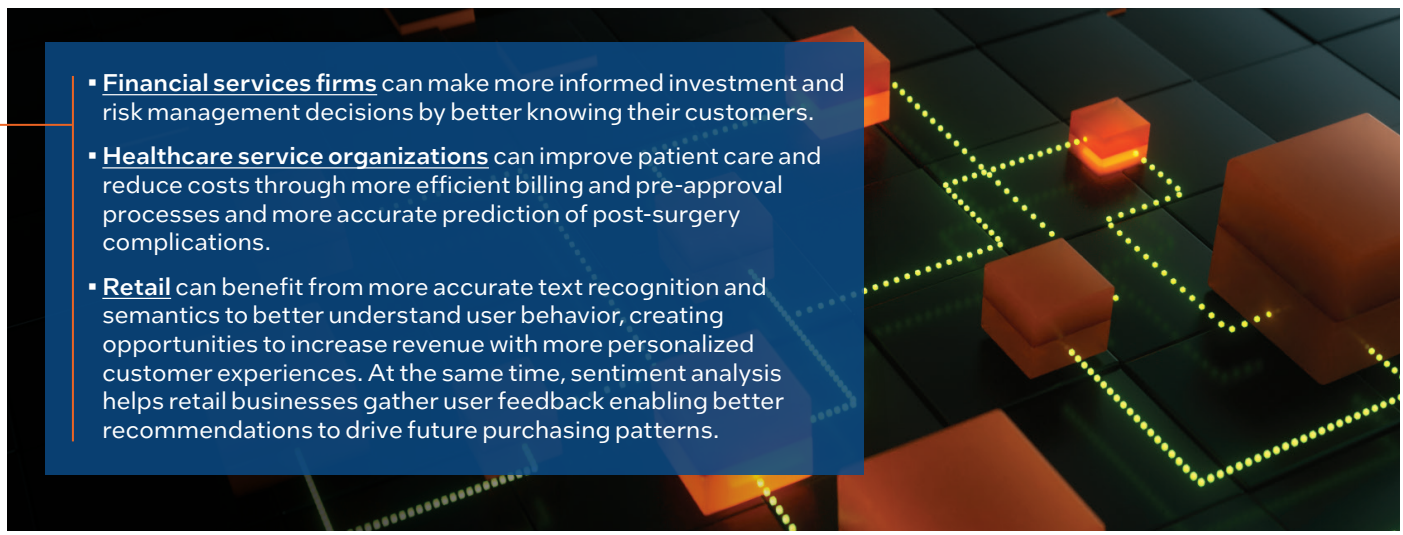
Providing customized user experiences drives customer demand and continued interest, with tremendous revenue potential across industries. To cost-effectively deliver a superior user experience with high-quality support, the sophistication of text interpretation by computers and recommender systems must continue to improve.

THE ANSWER:

When applied to unstructured data, sentiment analysis and content enrichment can enable businesses to pull insights from otherwise meaningless troves of data, for both real-time and retrospective analysis. Across industries, natural language processing can drive up user engagement, increase operational efficiencies and help take advantage of emerging revenue opportunities.

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.

THE IMPLEMENTATION:



- **Financial services firms** can make more informed investment and risk management decisions by better knowing their customers.
- **Healthcare service organizations** can improve patient care and reduce costs through more efficient billing and pre-approval processes and more accurate prediction of post-surgery complications.
- **Retail** can benefit from more accurate text recognition and semantics to better understand user behavior, creating opportunities to increase revenue with more personalized customer experiences. At the same time, sentiment analysis helps retail businesses gather user feedback enabling better recommendations to drive future purchasing patterns.

Confidential computing: Multi-party machine learning

THE NEED:

Leverage the power of machine learning without compromising the confidentiality and privacy of sensitive customer data.

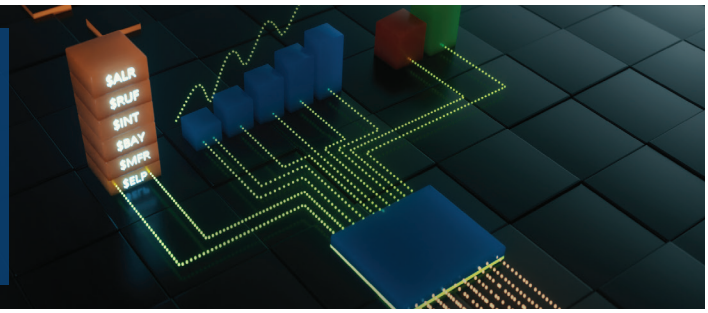
THE ANSWER:

Multi-party machine learning with confidential computing is especially useful in financial services, fraud detection and research.

- **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)** is a general-purpose performance-enhancing accelerator with a wide range of data and machine learning uses that continues to be supported in 4th Gen Intel Xeon Scalable processors. Intel AVX-512 can accelerate the preprocessing of unstructured data from multiple sources for training models, along with speeding up data movement so that it takes less time to prepare data sets for processing. The Intel Extension for Scikit-learn, coupled with Intel AVX-512, also accelerates machine learning algorithms for both training and inference.

THE IMPLEMENTATION:

- **Healthcare service organizations** can leverage the power of data to conduct more advanced research without exposing confidential patient information.
- **Financial services** can better predict potentially fraudulent activities while also fighting money laundering and the financing of terrorism.



Do more

With libraries integrated into TensorFlow and PyTorch, developers can also gain access to the benefits of built-in AI acceleration without extra work. Additionally, by changing just a couple lines of code, developers can seamlessly speed up Scikit-learn applications across single- and multi-node configurations. Intel has been working for years with the AI community to optimize the most popular AI frameworks, software and tools, so their mainstream distributions run better on Intel. Take advantage of additional tools like the Intel Distribution of OpenVINO™ toolkit to optimize inference models, Intel BigDL for distributed deep learning on Apache Spark, or the cnvrg.io MLOps platform to orchestrate your machine learning pipeline on any infrastructure in data center or cloud.

Ease of integration with existing infrastructure

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.

Decades of ecosystem enablement help extend Intel's trusted computing foundation across the hyper-scaled data center and the new edge frontier, with everything needed to build, scale and transform for operational agility. The openness of this approach fosters unbridled choice among hardware, software, cloud and service providers.

With the most flexibility to choose different cloud services, shapes and sizes to support specific workload demands, Intel architecture scales globally, across all the major cloud providers.

Through the Intel Partner Alliance, access exclusive resources for AI, cloud, high performance computing and other solution areas to help plan, build and deliver more value to your customers.

Connect with the industry's premier ecosystem to create the most innovative solutions, empowering your business to grow faster and make extraordinary opportunities possible that will drive global progress and enrich lives.

SUPPORTING STAT

Get the most choice with Intel's **OVER 50,000** unique instance types, sizes and regions. **6X GREATER THAN** the competition.[7]

Leadership's top business priorities in the digital transformation journey

Investments in digital transformation by organization leaders (tech and business alike) are expected to total $6.3T between 2022 and 2024.[8] This business brief is part of a series that illuminates the top business priorities leaders are focused on to achieve their business success in this transformative future, and how Intel hardware, software and services, including the 4th Generation Intel Xeon processor, help achieve these priorities:

- **AI (this brief):** Adopt data analytics and AI to drive critical outcomes
- **Security:** Achieve rigorous security and contribute to your zero trust security strategy
- **Cloud:** Activate strategies across hybrid, multi-cloud and the intelligent edge
- **Redefined worker experiences:** Embrace boundaryless interactive worker experiences
- **ESG:** Foster equitable outcomes and responsibility in environment | social | governance (ESG)

Learn More

www.intel.com/xeon/scalable
www.intel.com/ai

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

[2] Based on Intel market modeling of the worldwide installed base of data center servers running AI Inference workloads as of December 2021.

[3] 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.

[4] 3rd Gen Intel Xeon Scalable Processor vs. AMD EPYC. See configuration details [126-130] at www.intel.com/3gen-xeon-config.

[5] See [A17] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

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

[7] Source: Historical Liftr Insights Component tracker data + Intel internal preliminary analysis as of 09/02/22.

[8] IDC, October 2021. "IDC FutureScape: Worldwide Digital Transformation 2022 Predictions." https://www.idc.com/getdoc.jsp?containerId=US47115521.

Availability of accelerators varies depending on SKU. Visit the Intel Product Specifications page for additional product details.

Performance varies by use, configuration, and other factors. Learn more at https://www.intel.com/PerformanceIndex.

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

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

Your costs and results may vary.

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

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a nonexclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

1122/MH/MESH/350497-002US