AI is transforming how we work. Businesses are looking to save money and time and create new opportunities using state-of-the-art AI models for language, text-to-image and code generation. But many AI applications never make it to deployment as businesses quickly run into challenges with hardware availability, cost, integration and scale. Furthermore, while these workloads require increasing compute power, a majority do not require the full power of a GPU-based system.

37% GLOBAL AI MARKET[1]
CAGR through 2030

40% INVESTING IN GENERATIVE AI[2]
of C-level execs

69% HAVE AI IN PRODUCTION[3]
of organizations

PERFORMANCE PROOFPOINT

UP TO **2.34x** HIGHER RECOMMENDATION INFERENCE PERFORMANCE

UP TO **2.26x** HIGHER PERFORMANCE PER WATT

5th Gen Intel Xeon Platinum 8592+ with Intel AMX INT8 vs. AMD EPYC 9654 (Genoa)[4]

intel
XEON

Your business objective: Grow revenue and innovate

AI is providing new business opportunities and accelerating time to value that will generate new revenue streams while enhancing customer experiences. The opportunities are without limit, but their success ultimately depends on delivering results with the performance demanded by future-focused AI models.

How Intel can help

5th Gen Intel® Xeon® processors are designed for AI, with unique capabilities that deliver unmatched performance. With AI acceleration in every core, 5th Gen Intel Xeon processors are ready to handle your demanding AI workloads.

Built-in Intel Advanced Matrix Extensions (Intel AMX) speed up deep learning inference and accelerate small model training. Intel oneAPI Deep Neural Network Library (oneDNN) software optimizations already integrated into TensorFlow and PyTorch simplify access by developers to the benefits of built-in AI acceleration.

Intel lets you:

- **Deliver fast, personalized product or content recommendations** that don't slow down the user experience, with a deep learning-based recommender system that accounts for real-time user behavior signals and context features such as time and location.
- **Enable more responsive interactions** with smart assistants, chatbots, predictive text, language translation and more with a performance leap in natural language processing (NLP) inference.
- **Run generative AI models** that mimic human-generated content, such as large language models (LLMs) and text-to-image generation. With Intel AMX, 5th Gen Intel Xeon processors make generative AI more accessible on CPUs, so you can take advantage of their ubiquity in the data center, and additional acceleration can be added with discrete accelerators.

PERFORMANCE PROOFPOINT

UP TO **3.87x** HIGHER REAL-TIME NLP INFERENCE PERFORMANCE (DISTILBERT, BS>1)

PERFORMANCE PROOFPOINT

UP TO **2.56x** HIGHER PERFORMANCE PER WATT

5th Gen Intel Xeon Platinum 8592+ with INT8 vs. AMD EPYC 9754 (Bergamo)[5]

Benefit *your* business

Taboola has directly monetized inference in the form of AI-powered content recommendations, made to readers in real time, that helped expand the company's revenue. Taboola engaged with Intel software engineers to optimize their code for Intel® Xeon® processors.

For Meituan, vision AI has become the key to driving business model innovation, delivering more accurate and personalized internet services to users and enhancing competitive advantages. The company improved vision AI inference throughput without sacrificing accuracy by drawing on platform capabilities including Intel AMX.

PERFORMANCE PROOFPOINT

UP TO **3.4x** FASTER BATCH DLSA BERT-LARGE FINETUNING FOR SST2 DATASET[6]

5th Gen Intel Xeon Platinum 8592+ with AMX BF16 compared to FP32

For dedicated AI or inference and training on larger models, add purpose-built Intel Gaudi® AI accelerators for an expanded platform that's ready for anything.

Ready to refresh

There is a lot of pressure for organizations to achieve efficiency, TCO, security and sustainability goals, which is amplified with the added capabilities and requirements for AI. Microsoft Server 2016 and 2019 end of life and end of support is also creating pressure to upgrade. Simply put, systems from four years ago do not meet today's demands.

A refresh strategy based on Intel® Xeon® processors enables you to look holistically at investments across software, hardware and infrastructure, with updated elements optimized to bring out the best of each other. It also enables you to lower costs while reducing your carbon footprint, creating a more sustainable compute infrastructure.

PERFORMANCE PROOFPOINT

UP TO **62%** REDUCED TCO THAT CAN BE REINVESTED FOR GROWTH[7]

WITH **3:1** SERVER CONSOLIDATION FOR A SMALLER DATACENTER FOOTPRINT AND LOWER COSTS[7]

Your business objective: Cost Reduction

Organizations are being challenged to reduce total cost of ownership across both capital and operating expenses — hardware purchases, software licenses, power consumption and cloud costs — even while keeping up with increasing workload demands. Significant cost reductions can result from improving operational efficiency through optimized performance and refreshing old technology.

Many organizations are considering how to reduce power consumption in their technology infrastructure and cloud purchases. Their successes have positive impacts on operating costs, support corporate sustainability initiatives and help mitigate the climate impacts of operations.

How Intel can help

5th Gen Intel Xeon processors offer a balance of performance and cost on the hardware you already use for your other workloads. Lower your carbon footprint — and your TCO — by upgrading to 5th Gen Intel Xeon processors with improved performance per watt and features for managing power efficiency.

PERFORMANCE PROOFPOINT

UP TO **2.1x** PERFORMANCE SPEEDUP

UP TO **1.58x** HIGHER PERFORMANCE PER WATT

5th Gen Intel Xeon processor vs 3rd Gen Intel Xeon on Llama 2 13B first token latency (int8)[8]

Benefit *your* business

Mobileye, an autonomous driving technology company, improved price-performance training with complex computer vision systems using Amazon EC2 DL1 instances based on Intel Gaudi AI accelerators combined with Intel® Xeon® processors as host CPUs. The deployment yielded 40% better price performance than Mobileye experienced with alternative AI solutions and accelerated the company's deep-learning development cycle.[9]

Numenta demonstrates the capacity to dramatically reduce the overall cost of running LLMs in production on Intel architecture instead of AMD, unlocking entirely new natural language processing capabilities for customers.

PERFORMANCE PROOFPOINT

UP TO **5.7x** HIGHER REAL-TIME NLP INFERENCE PERFORMANCE (DISTILBERT)

UP TO **6X** HIGHER PERFORMANCE PER WATT

5th Gen Intel Xeon Platinum 8592+ with AMX INT8 compared to FP32[10]

Your business objective: Risk Mitigation

Businesses must adopt new models to enable AI in the context of distributed computing to realize its full value. Proprietary machine learning models must be protected as intellectual property alongside other critical business assets, while data privacy must also be addressed to enable collaboration across sensitive data held by multiple organizations. Securing AI is now a central strategic concern for companies in order to obtain and protect their competitive advantage.

How Intel can help

Until recently, data security has focused on protecting data at rest (in storage) and in flight (while moving between locations). Confidential computing, powered by Intel Software Guard Extensions (Intel SGX) and Intel Trust Domain Extensions (Intel TDX), goes a step further, helping to ensure that data is also protected while it is being processed. This is possible through the creation of a Trusted Execution Environment (TEE). Not only is all critical data stored inside the TEE, but so are the applications and algorithms that access and process that data.

5th Gen Intel Xeon processors improve dramatically on the confidential computing technologies of their predecessors with the general availability of VM-level in addition to application-level isolation. This range of choice allows you to match a solution to your specific business and regulatory needs, deployable across the infrastructure to protect data and IP at the data center, cloud and edge.

- **Application isolation with Intel SGX.** Intel SGX is the most researched and updated 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.
- **VM-level isolation with Intel TDX**. Intel TDX offers isolation and confidentiality at the virtual machine (VM) level. Within an Intel TDX confidential VM, the guest OS and VM applications are isolated from access by the cloud host, hypervisor and other VMs on the platform. Because Intel TDX does not require application code changes, it offers a relatively simple migration path for existing VMs to move to a TEE.
- **Independent attestation**. Intel provides independent attestation services in a public/private multi-cloud environment with Intel Trust Authority. Designed to remotely verify and assert trustworthiness of compute assets such as TEEs, devices and roots of trust, the service is operationally independent from the cloud/edge infrastructure provider hosting the confidential computing workloads.

Read the report "The Future of Risk is Upon Us And We Can Manage It if We Secure AI" for more on the importance of security in AI.

Benefit *your* business

Confidential computing lets Equideum Health engage in multi-party analysis collaborations using protected, sensitive data such as full-sequence human genomes and other biometrics. Drawing on Fortanix software technology and Intel® SGX, Equideum has opened the door to innovations in personalized precision medicine.

Innovate freely with open technology

Intel is a top contributor to the open source community, especially when it comes to AI. In fact, 90% of developers are using software developed or optimized by Intel.[11] Rather than building models from scratch, many developer teams prefer working with open source models. Intel has dozens of pre-trained, optimized AI models that are ready out of the box and easy to customize. And there's no need to wait for hardware, thanks to the largest processor ecosystem in the world and broad supply availability. For more demanding workloads, Intel® Xeon® processors have an open-standards framework to add accelerators and GPUs.

Intel hardware platforms are unified by a common, open-standards programming model based on oneAPI, built for productivity and performance across CPUs and GPUs. Intel Software Development Tools include advanced compilers, libraries, profilers and code migration tools. Intel's optimized AI frameworks make it easier for data scientists, HPC/AI researchers and developers to get out-of-the-box performance using Intel machine and deep learning optimizations, exploit cutting-edge features of hardware, optimize AI inference with streamlined deployment and implement powerful end-to-end solutions more productively.

- **Optimize AI inferencing and increase performance** by taking advantage of Intel accelerators — CPU, GPU and VPU — to deploy at scale using the popular, open source OpenVINO™ toolkit from Intel. Start with a trained model from popular deep learning frameworks such as TensorFlow, PyTorch and others and seamlessly integrate with OpenVINO compression techniques for streamlined deployment across various hardware platforms. All with minimal code changes.
- **Accelerate fine tuning and inference** in deep learning and other AI use cases by enabling Intel AMX and Intel Advanced Vector Extensions 512 (Intel AVX-512) on the CPU and Intel XMX on the GPU using Intel oneAPI Deep Neural Network Library (oneDNN) and Intel oneAPI Data Analytics Library (oneDAL), part of the Intel oneAPI Base Toolkit.
- **Drive orders of magnitude performance improvement** for training and inference optimizations into TensorFlow and PyTorch using Intel-optimized deep learning AI frameworks.
- **Speed model development and innovate AI faster** across various industries using Intel-built open source AI reference kits (34 are available).

PERFORMANCE PROOFPOINT

UP TO **9.2x** HIGHER REAL-TIME NLP INFERENCE PERFORMANCE (BERT-LARGE)

UP TO **10.2x** HIGHER PERFORMANCE PER WATT

5th Gen Intel Xeon Platinum 8592+ with AMX INT8 compared to FP32[12]

Reimagine what's possible

Forward-looking decision makers must capture the full potential of AI to grow revenue, innovate, reduce costs and reduce risk. Intel uniquely provides the comprehensive hardware, software, tools and design patterns to realize that vision. 5th Gen Intel Xeon processors deliver more performance per core and per watt than predecessors, to meet emerging demands while delivering on key business metrics.

It starts with Intel.

Learn More

www.intel.com/xeon

www.intel.com/ai

Choose the CPU that organizations trust most for their AI — The majority of data center AI inference deployments today run on Intel Xeon processors.[13]

intel XEON

[1] Grand View Research. "Artificial Intelligence Market Size, Share & Trends Analysis Report By Solution, By Technology (Deep Learning, Machine Learning), By End-use, By Region and Segment Forecasts, 2023 - 2030." https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market.

[2] McKinsey, August 1, 2023. "The state of AI in 2023: Generative AI's breakout year." https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year.

[3] S&P Global Market Intelligence, August 2023. "2023 Global Trends in AI Report." https://www.weka.io/resources/analyst-report/2023-global-trends-in-ai/.

[4] Up to 2.34x higher batched Recommendation System inference performance (DLRM) and 2.26x higher performance/watt on 5th Gen Intel Xeon Platinum 8592+ with AMX INT8 vs. AMD EPYC 9654 (Genoa). See [A208] at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary.

[5] See [A209] at intel.com/processorclaims: 5th Gen Intel Xeon processors. Results may vary.

[6] See [A5] at intel.com/processorclaims: 5th Gen Intel Xeon processors. Results may vary.

[7] Claims based on 4th Gen Intel Xeon processors. Intel, June 29, 2023. "Minimize Total Cost of Ownership with a Server Refresh." https://www.intel.com/content/www/us/en/content-details/783070/minimize-total-cost-of-ownership-with-a-server-refresh.html.

[8] See [A2] at intel.com/processorclaims: 5th Gen Intel Xeon processors. Results may vary.

[9] https://aws.amazon.com/solutions/case-studies/mobileye-ec2-dl1-case-study/.

[10] See [A24] at intel.com/processorclaims: 5th Gen Intel Xeon processors. Results may vary.

[11] Global Development Survey conducted by Evans Data Corp., 2021.

[12] See [A19] at intel.com/processorclaims: 5th Gen Intel Xeon processors. Results may vary.

[13] Based on Intel market modeling of the worldwide installed base of data center servers running AI Inference workloads as of December 2022.

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.

1123/MH/MESH/353951-001US