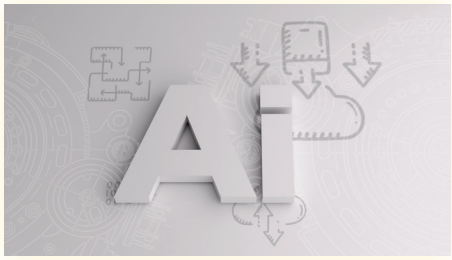
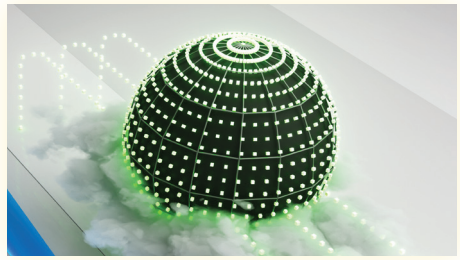
Deploy AI Everywhere



Taking AI from concept to production is a challenge

With complexity, cost, risk



Intel's open approach can help you leverage the best of AI

Dramatic advances in AI are making it ubiquitous in various forms. Increased sophistication of machine learning models is paying new dividends in established usages such as demand forecasting, recommendations and fraud detection. Deep learning goes further, emulating the layered structure of the human brain for novel interpretive capabilities such as reading medical images and chatbot-based customer support. More recently, generative AI has disrupted the industry with the ability to semi-autonomously generate original text and images, with profound implications for augmenting human work of all kinds.

The challenges of managing huge volumes of data and converting it into meaningful insights and actions continue to increase. End-to-end AI pipelines comprise many stages including data ingestion, data cleaning, feature engineering, model training and model inference. Each stage must be highly performant, and seamlessly integrate with the others to deliver an optimized solution for real world scenarios. Despite the massive opportunity, too many companies are failing to succeed with their AI ambitions.

Common challenges organizations face taking AI from concept to production include:

- **Complexity.** The number of methods, capabilities and infrastructure requirements to run AI is growing.
- **Costs**. Controlling expenses associated with AI contributes to overall business efficiency.
- **Operationalization.** Many steps are needed to get AI proofs of concept through to production.
- **Data security and privacy.** Activating sensitive data is challenging while remaining secure and compliant.

Intel's approach and commitment 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.

Intel AI technology extends from client to cloud and edge to data center. This paper explores processor and software technologies for the data center.

CONTENTS

Optimize performance on open standards tools you already use

An open software stack and ecosystem accelerate innovation and maximize flexibility and freedom of choice. A software-defined, silicon-enhanced ecosystem accelerates multiarchitecture, multivendor programming for AI, with Intel providing optimized performance and productivity:

- **Build faster with optimized, pre-trained models.** Intel AI Reference Kits provide foundations to streamline projects, from data ingest to benchmarking.
- **Get drop-in acceleration with familiar models.** Intel® AI Tools provide optimized industry-standard tools and frameworks, including PyTorch, TensorFlow, scikit-learn, XGBoost, Modin and other tools that accelerate end-to-end data science and analytics pipelines.
- **Streamline transformation with an open ecosystem.** Open solutions help drive flexibility and innovation across the computing, data and AI ecosystem to simplify adoption and success.

Solutions that run on Intel are the fast path to scale AI everywhere. Intel AI optimizations for Spark, PyTorch, TensorFlow, scikit-learn, NumPy, XGBoost, Modin, Intel AI analytics toolkit, Intel Distribution of OpenVINO toolkit, BigDL and Intel Neural Compressor offer a substantial performance advantage,[1] 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 Advanced Matrix Extensions (Intel AMX) to run out-of-box with minimal or no code changes.[2]

PERFORMANCE PROOFPOINT

UP TO **10X** HIGHER PYTORCH REAL-TIME INFERENCE PERFORMANCE[3]

UP TO **10X** HIGHER PYTORCH TRAINING PERFORMANCE[4]

on 4th Gen Intel Xeon Scalable processors with built-in Intel AMX (BF16) versus the prior generation (FP32)

Choose the hardware optimized for your needs

Heterogenous architectures consisting of CPUs, accelerators and GPUs provide Intel customers and partners with high-performance, open-standards solutions to quickly deploy AI at scale across AI workloads and usage models. This diversity enables deployments to balance requirements across factors such as scale, training time, latency, power and cost.

5th Gen Intel® Xeon® Scalable Processor

General-Purpose and Mixed AI Workloads

Intel® Gaudi® 2 AI Accelerator

Deep Learning Training & Inference of large scale models

Intel® Data Center GPU Max Series

AI Acceleration for High Performance Computing

General-purpose AI

Intel® Xeon® Scalable Processors

As AI becomes more prevalent in enterprise applications, mainstream general-purpose computing platforms must deliver high-performance training and inference, without discrete hardware accelerators. The reigning enterprise workhorse, with its massive installed base, is ideal for most AI usages, including analytics and mixed workloads. The latest Intel Xeon Scalable Processors have the most built-in accelerators of any CPU in the world for key workloads including AI:

- Intel® Advanced Matrix Extensions (Intel® AMX) accelerates AI capabilities, speeding up deep learning training and inference without additional hardware.
- 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.
- 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® Trust Domain Extensions (Intel® TDX) offers confidentiality at the virtual machine (VM) level. Intel TDX isolates the guest OS and all VM applications from the cloud host, hypervisor and other VMs on the platform. Intel TDX is designed so confidential VMs are easier to deploy and manage at scale than application enclaves.

Intel®Xeon® CPU Max Series with high-bandwidth memory (HBM) provides more memory for specialized HPC and AI workloads, with 64 GB in-package for ~1 TB/s memory bandwidth and ~1 GB HBM per core.

PERFORMANCE PROOFPOINT

3.9X HIGHER AVERAGE MACHINE LEARNING TRAINING AND INFERENCE PERFORMANCE[5]

on 4th Gen Intel Xeon Scalable processors vs. Nvidia A100 GPU

CUSTOMER CASE STUDY

Numenta Accelerates LLM Inference on CPUs

Breaking new ground in applying neuroscientific principles to AI, Numenta innovations make inference dramatically more efficient for models such as GPT-3 and BERT-Large, without impacting accuracy.

The company's ongoing testing and development shows cost-effective performance for production on Intel CPU platforms.

> Read more

Numenta

Faster Inference than AMD Milan

Short Text Sequences
Increased Throughput by 100x[6]
on 4th Gen Intel® Xeon® Scalable Processors

Long Text Sequences (Documents)
Run 20x Faster[7]

intel. GAUDI

Deep learning training and inference of large scale models

Habana® Gaudi®2 Deep Learning Processor

The growth of training compute requirements for deep learning is accelerating dramatically, with 100x – 1000x growth from 2016 to today.[8] Massive large language models (LLMs) such as GPT-3 epitomize these requirements, for which GPUs have been the only viable training and inference option.

The Habana Gaudi2 deep learning processor is Intel's next-generation hardware platform to improve on key performance metrics for large scale models, such as generative AI, excelling in measures such as time- and cost-to-train, fine-tuning efficiency and inference throughput and latency.

Habana SynapseAI® software suite, like OneAPI, integrates PyTorch and TensorFlow frameworks to enable developers to work at the abstraction level that they are accustomed to. With Gaudi accelerators, developers can build new or migrate existing code with minimal code changes. In addition to abstracting away complexity for the vast majority of implementations, SynapseAI enables developers who have specialized needs to write their own custom kernels.

Gaudi2 performance continues to advance with software maturity. With the implementation of FP8 software, these metrics get a substantial boost, making Gaudi2 increasingly competitive to Nvidia's leading product.

CUSTOMER CASE STUDY

Taboola Accelerates AI-Driven Recommendations

A TensorFlow-based recommendation engine is entrusted with generating Taboola's fast, targeted content recommendations for mobile customers. To drive as many as 40 billion recommendations each day, the company accelerates throughput using 4th Gen Intel® Xeon® Scalable processors.

High per-core performance, Intel® Advanced Matrix Extensions (Intel® AMX) and Intel® Advanced Vector Extensions 512 (Intel® AVX-512) combine to power fast recommendations and continued success.

> Read more

Higher Performance than Predecessors

Increased Throughput by **1.74x**[9]

on 4th Gen Intel® Xeon® Scalable processors

CUSTOMER CASE STUDY

Mobileye enables autonomous driving on Gaudi[10]

- Requires frequent training and retraining of computer vision models
- Enabling near real-time object detection, segmentation and tracking
- Up to 40% better price performance vs other instances
- Scales to more than 3500 nodes with near-linear scaling
- Delivers tangible TCO advantage

> Read more

PERFORMANCE PROOFPOINT

311 MINUTES GPT-TRAINING USING 384 ACCELERATORS[11]

95% NEAR-LINEAR SCALING FROM 256 TO 385 ACCELERATORS[11]

out-of-the-box MLPerf Training 3.0 Benchmark results on Habana Gaudi2 Processors

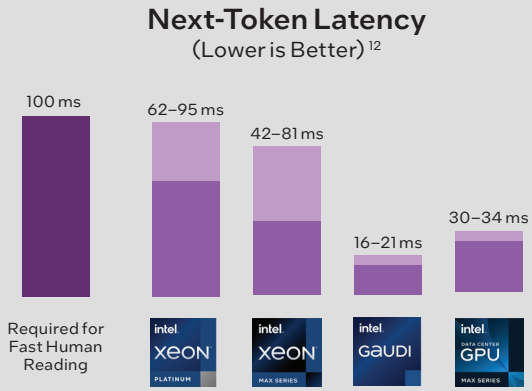
CUSTOMER CASE STUDY

Llama 2 Generates Text Faster than Humans Can Read It

Meta released Llama 2, a collection of LLMs up to 70 billion parameters that are hand-optimized to out-perform many open source models.

Llama 2 generates text in real time fast enough to keep up with fast human readers, across Intel® data center AI platforms.[12]

> Read more



PERFORMANCE PROOFPOINT

NEARLY 3X SPEEDUP FINE-TUNING BRIDGETOWER VISION-LANGUAGE MODEL[13]

on Habana Gaudi2 processors compared to Nvidia A100 GPUs

AI Acceleration for High Performance Computing

Intel® Data Center GPU Max Series

Engineering teams face substantial obstacles and inefficiencies when porting and refactoring code to deploy GPUs for AI in HPC environments. Proprietary programming models, such as CUDA, interfere with portability between GPUs and CPUs, as well as among GPU vendors.

oneAPI is an open, multiarchitecture programming model for CPUs and accelerators such as GPUs. Based on standards, oneAPI simplifies software development and delivers accelerated compute without proprietary lock-in, while enabling the integration of existing code.

The Intel Data Center GPU Max Series leverages oneAPI and offers high compute density, helping reduce deployment footprints.

- **Solve large problems with 408 MB L2 cache**. The industry's largest cache in a GPU keeps more hot data close to the processing cores.
- **Drive high throughput with accelerated hardware.** Built-in ray-tracing cores increase performance, and AI-boosting Intel® Xᵉ Matrix Extensions (XMX) provide deep systolic arrays to enable vector and matrix capabilities on a single device.
- **Increase density in the data center.** OpenCompute Accelerator Modules (OAMs) combine multiple Xᵉ HPC stacks in a single physical package for high scalability.

PERFORMANCE PROOFPOINT

UP TO 2X PERFORMANCE ON HPC AND AI WORKLOADS VERSUS THE COMPETITION[14]

UP TO 256 INT8 OPS PER CLOCK WITH INTEL XMX

on Intel Max Series GPUs

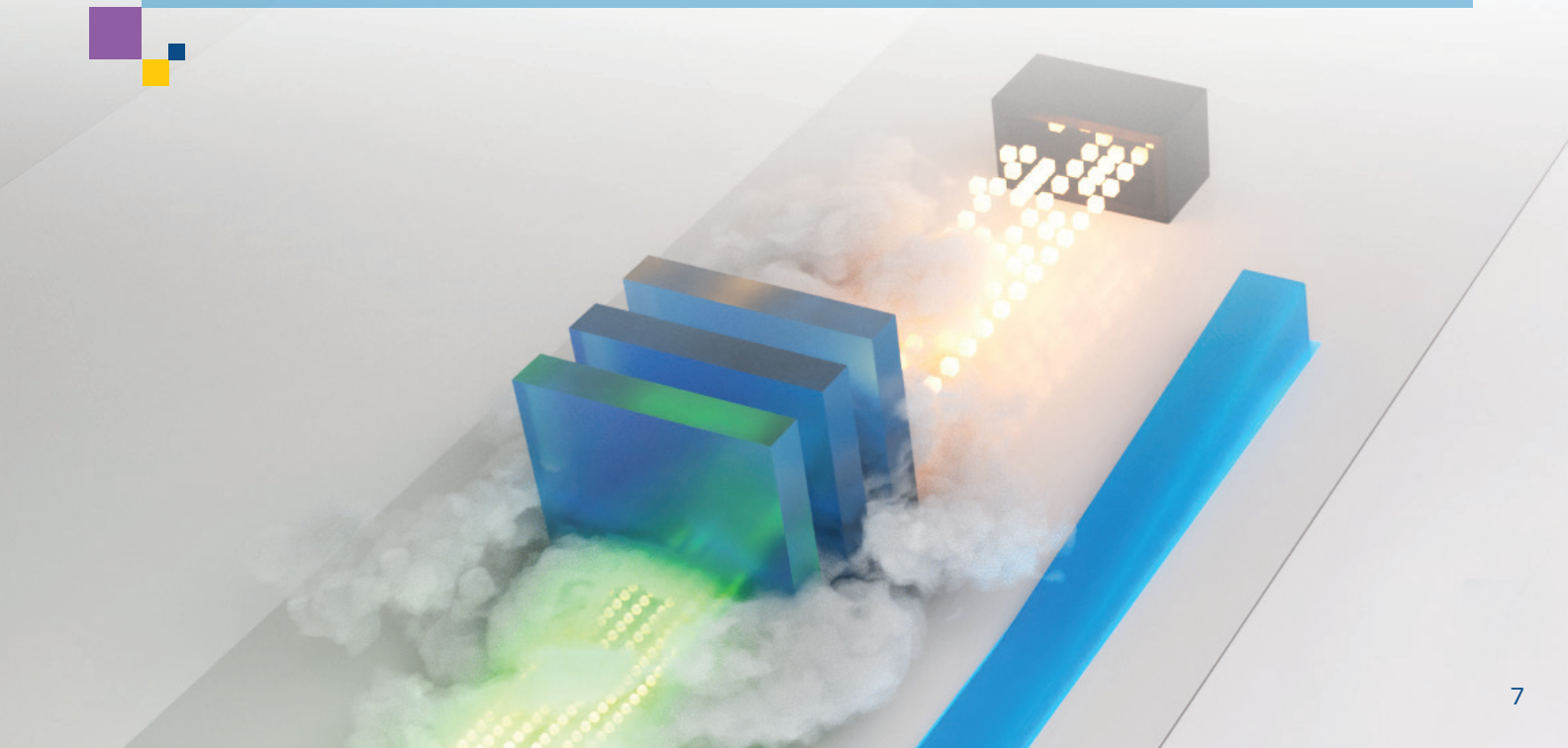
oneAPI — Open-standards software development without vendor lock-in

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 oneAPI tools include advanced compilers, libraries, profilers and code migration tools.

Intel's AI tools and optimized 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.

oneAPI™

- **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, powered by oneAPI. 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 frameworks** 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 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.)

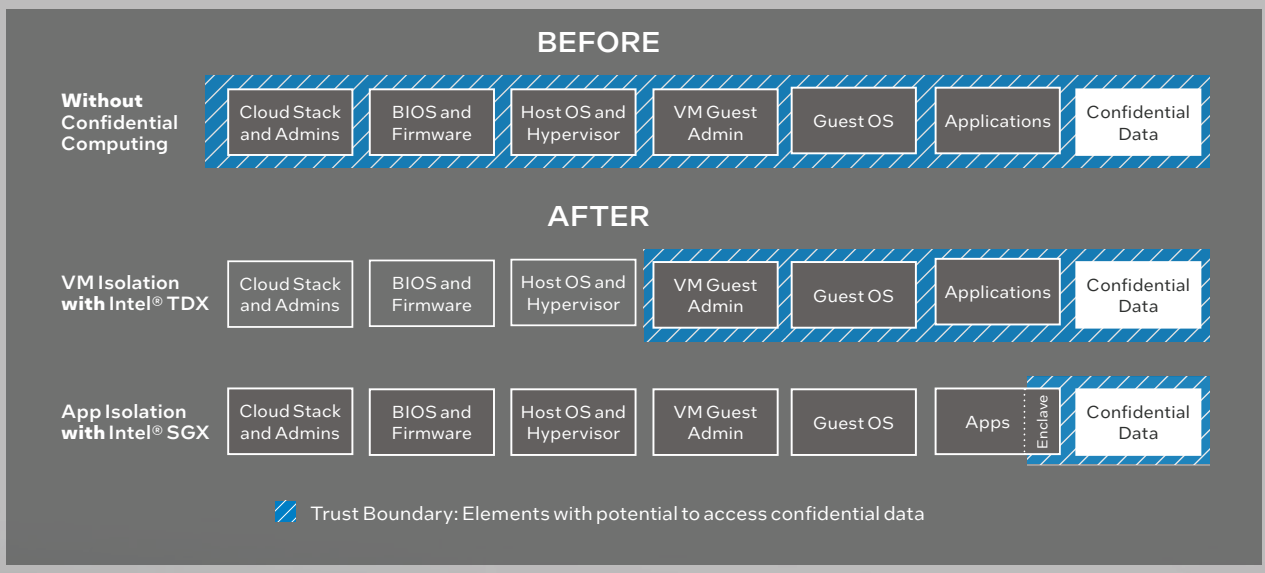


Built-in security for AI

To better realize the potential business benefits of AI, many organizations draw on confidential or sensitive data. Uploading private data to existing models is a common requirement for fine tuning. Likewise, having humans interact with models based on such data can be key to the application of reinforcement learning with human feedback (RLHF) to improve the accuracy of results (completions).

Confidential computing protects AI data, models and parameters while in use with hardware-based memory protections built into Intel Xeon Scalable processors. These technologies improve the isolation of data by enabling only trusted code to access it within hardware-protected memory enclaves. These privacy protections enable innovation around regulated workloads in hyper-scaled platforms and other distributed networks.

Intel Confidential Computing technologies help increase security around sensitive or regulated AI data and code by limiting which software is allowed to access it. Intel TDX prohibits access by any software or administrator outside the workload's virtual machine, essentially removing the cloud host's stack, OS and hypervisor from the trust boundary. Intel SGX shrinks the attack surface even further by limiting data access only to authorized application software or functions.



AI for good

AI has many benefits it can bring to humanity, from personalized medicine to improved stability of financial systems and the energy grid. AI's incredible power and untold potential is still relatively immature. 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.

To realize the entirety of that potential, the technology industry must make AI open, visible and accessible. The Intel AI portfolio provides competitive, high-performance, open standards solutions for our customers and partners to quickly deploy AI at scale across the full spectrum of workloads and usages.

More about AI with Intel.

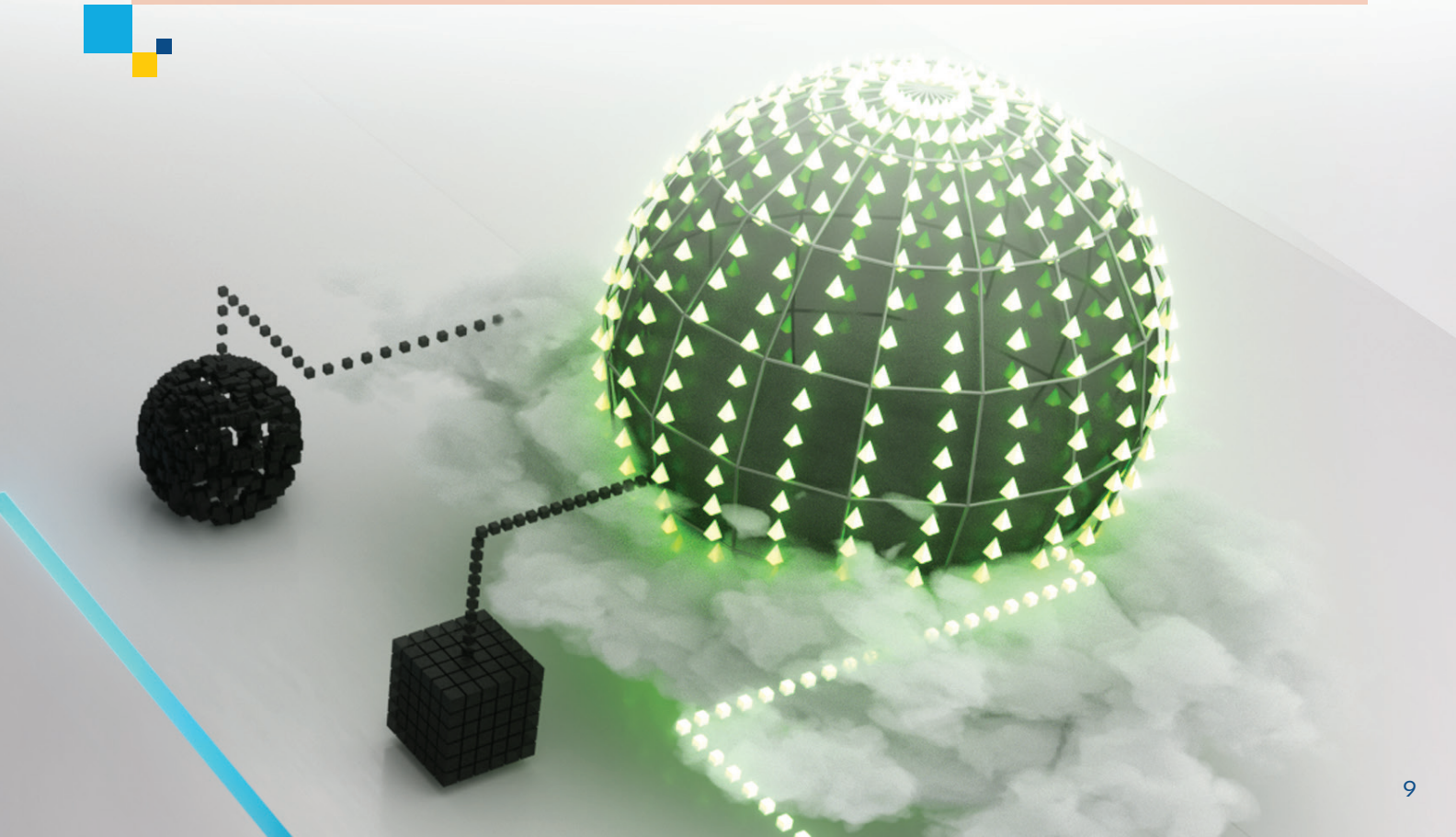
4th Gen Intel® Xeon® Scalable processors

Habana Gaudi 2 Deep Learning Processor

Intel® Data Center GPU Max Series

Responsible AI

Intel® AI News



[1] 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.

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

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

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

[5] See [A201] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

[6] For more, see: https://edc.intel.com/content/www/us/en/products/performance/benchmarks/4th-generation-intel-xeon-scalable-processors/. Numenta: BERT-Large: Sequence Length 64, Batch Size 1, throughput optimized 3rd Gen Intel® Xeon® Scalable: Tested by Numenta as of 11/28/2022. 1-node, 2x Intel® Xeon®8375C on AWS m6i.32xlarge, 512 GB DDR4-3200, Ubuntu 20.04 Kernel 5.15, OpenVINO 2022.3, Numenta-Optimized BERT-Large, Sequence Length 64, Batch Size 1 Intel® Xeon® 8480+: Tested by Numenta as of 11/28/2022. 1-node, pre-production platform with 2x Intel® Xeon® 8480+, 512 GB DDR5-4800, Ubuntu 22.04 Kernel 5.17, OpenVINO 2022.3, Numenta-Optimized BERT-Large, Sequence Length 64, Batch Size 1.

[7] For more, see: https://www.intel.com/content/www/us/en/products/details/processors/xeon/max-series.html. Numenta BERT-Large: AMD Milan: Tested by Numenta as of 11/28/2022. 1-node, 2x AMD EPYC 7R13 on AWS m6a.48xlarge, 768 GB DDR4-3200, Ubuntu 20.04 Kernel 5.15, OpenVINO 2022.3, BERT-Large, Sequence Length 512, Batch Size 1. Intel® Xeon® 8480+: Tested by Numenta as of 11/28/2022. 1-node, 2x Intel® Xeon® 8480+, 512 GB DDR5-4800, Ubuntu 22.04 Kernel 5.17, OpenVINO 2022.3, Numenta-Optimized BERT-Large, Sequence Length 512, Batch Size 1. Intel® Xeon® Max 9468: Tested by Numenta as of 11/30/2022. 1-node, 2x Intel® Xeon® Max 9468, 128 GB HBM2e 3200 MT/s, Ubuntu 22.04 Kernel 5.15, OpenVINO 2022.3, Numenta-Optimized BERT-Large, Sequence Length 512, Batch Size 1.

[8] Jaime Sevilla et al. "Compute Trends Across Three Eras of Machine Learning." https://arxiv.org/pdf/2202.05924.pdf.

[9] For more, see https://www.intel.com/content/www/us/en/customer-spotlight/stories/taboola-recommendation-engine-customer-story.html.

[10] For more, see: https://aws.amazon.com/solutions/case-studies/mobileye-ec2-dl1-case-study/

[11] MLCommons, June 27 2023. "MLCommons v3.0 Results."https://www.intel.com/content/www/us/en/newsroom/news/new-mlcommons-results-ai-gains-intel.html.

[12] Habana Gaudi2 Deep Learning Accelerator: All measurements were made using Habana SynapseAI version 1.10 and optimum-habana version 1.6 on a HLS2 Gaudi2 server with eight Habana Gaudi2 HL-225H Mezzanine cards and two Intel Xeon Platinum 8380 CPU @ 2.30GHz and 1TB of System Memory. Performance was measured in July 2023.

4th Gen Intel Xeon 8480: Intel 4th Gen Xeon Platinum 8480+ 2 socket system, 112-cores/224-threads, Turbo Boost On, Hyper-Threading On, Memory: 16x32GB DDR5 4800MT/s, Storage: 953.9GB); OS: CentOS Stream 8; Kernel: 5.15.0-spr.bkc.pc.16.4.24.x86_64; Batch Size: 1; Measured on 1 socket: 1; PyTorch nightly build0711; Intel® Extensions for PyTorch tag v2.1.0.dev+cpu.llm; Model: Llama 2 7B and Llama 2 13B; Dataset LAMBADA; Token Length: 32/128/1024/2016 (in), 32 (out); Beam Width 4; Precision: BF16 and INT8; Test by Intel on 7/12/2023.

Intel Xeon Max 9480: Intel Xeon Max 9480 2 socket system, 112-cores/224-threads, Turbo Boost On, Hyper-Threading On, Memory: 16x64GB DDR5 4800MT/s; 8x16GB HBM2 3200 MT/s, Storage: 1.8TB; OS: CentOS Stream 8; Kernel: 5.19.0-0812.intel_next.1.x86_64+server; Batch Size: 1; Measure on 1 socket; PyTorch nightly build0711; Intel® Extensions for PyTorch llm_feature_branch; Model: Llama 2 7B and Llama 2 13B, Dataset LAMBADA; Token Length: 32/128/1024/2016 (in), 32 (out); Beam Width 4; Precision: BF16 and INT8; Test by Intel on 7/12/2023.

Intel Data Center GPU Max Series: 1-node, 2x Intel Xeon Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS SE5C7411.86B.9525.D19.2303151347, microcode 0x2b0001b0, 1x Ethernet Controller X710 for 10GBASE-T, 1x 1.8T WDC WDS200T2B0B, 1x 931.5G INTEL SSDPELKX010T8, Ubuntu 22.04.2 LTS, 5.15.0-76-generic, 4x Intel Data Center GPU Max 1550 (measured solely using Single Tile of a single OAM GPU card), IFWI PVC 2_1.23166, agama driver: agama-ci-devel-627.7, Intel oneAPI Base Toolkit 2023.1, PyTorch 2.0.1 + Intel Extension for PyTorch v2.0.110+xpu (dev/LLM branch), AMC Firmware Version: 6.5.0.0, Model: Meta AI Llama 2 7B and Llama 2 13B, Dataset LAMBADA; Token Length: 32/128/1024/2016 (in), 32 (out); Greedy search; Precision FP16; Tested by Intel on 07/07/23.

[13] Hugging Face, June 29 2023. "Accelerating Vision-Language Models: BridgeTower on Habana Gaudi2." https://huggingface.co/blog/bridgetower.

[14] Visit intel.com/performanceindex (Events: Supercomputing 22) for workloads and configurations. 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.

0823/MH/MESH/353925-001US