April 9, 2024

Intel Unleashes Enterprise AI with Gaudi 3, AI Open Systems Strategy and New Customer Wins

At Vision 2024, Intel goes all-in on open and more secure enterprise AI with new customers, partners and collaborations across the AI continuum.

NEWS HIGHLIGHTS

- Intel unveiled a comprehensive AI strategy for enterprises, with open, scalable systems that work across all AI segments.
- Introduced the Intel® Gaudi® 3 AI accelerator, delivering 50% on average better inference[1] and 40% on average better power efficiency[2] than Nvidia H100 – at a fraction of the cost.
- Intel announced Gaudi 3 availability to original equipment manufacturers (OEMs) – including Dell Technologies, Hewlett Packard Enterprise, Lenovo and Supermicro – broadening the AI data center market offerings for enterprises.
- Announced new Intel Gaudi accelerator customers and partners, including Bharti Airtel, Bosch, CtrlS, IBM, IFF, Landing AI, Ola, NAVER, NielsenIQ, Roboflow and Seekr.
- Intel announced the intention to create an open platform for enterprise AI together with SAP, RedHat, VMware and other industry leaders to accelerate deployment of secure generative AI (GenAI) systems, enabled by retrieval-augmented generation (RAG).
- Through the Ultra Ethernet Consortium (UEC), Intel is leading open Ethernet networking for AI fabric. The company introduced an array of AI-optimized Ethernet solutions, including the AI NIC (network interface card) and AI connectivity chiplets.

PHOENIX--(BUSINESS WIRE)-- At the Intel Vision 2024 customer and partner conference, Intel introduced the Intel Gaudi 3 accelerator to bring performance, openness and choice to enterprise generative AI (GenAI), and unveiled a suite of new open scalable systems, next-gen products and strategic collaborations to accelerate GenAI adoption. With only 10% of enterprises successfully moving GenAI projects into production last year, Intel's latest offerings address the challenges businesses face in scaling AI initiatives.

This press release features multimedia. View the full release here: https://www.businesswire.com/news/home/20240409077438/en/

Intel tackles the generative AI gap by introducing the Intel Gaudi 3 AI accelerator at the Intel Vision event on April 9, 2024, in Phoenix, Arizona. Gaudi 3 gives customers choice with open community-based software and industry-standard Ethernet networking to scale their systems more flexibly. (Credit: Intel Corporation)

“Innovation is advancing at an unprecedented pace, all enabled by silicon – and every company

is quickly becoming an AI company," said Intel CEO Pat Gelsinger. "Intel is bringing AI everywhere across the enterprise, from the PC to the data center to the edge. Our latest Gaudi, Xeon and Core Ultra platforms are delivering a cohesive set of flexible solutions tailored to meet the changing needs of our customers and partners and capitalize on the immense opportunities ahead."

More: Intel Vision 2024 (Press Kit) | Intel Vision 2024 Keynote (Livestream/Replay) | Intel Tackles the GenAI Gap with Gaudi 3 (News)

Enterprises are looking to scale GenAI from pilot to production. To do so, they need readily available solutions, built on performant and cost- and energy-efficient processors like the Intel Gaudi 3 AI accelerator, that also address complexity, fragmentation, data security and compliance requirements.

Introducing Gaudi 3 for AI Training and Inference

The Intel Gaudi 3 AI accelerator will power AI systems with up to tens of thousands of accelerators connected through the common standard of Ethernet. Intel Gaudi 3 promises 4x more AI compute for BF16 and a 1.5x increase in memory bandwidth over its predecessor. The accelerator will deliver a significant leap in AI training and inference for global enterprises looking to deploy GenAI at scale.

In comparison to Nvidia H100, Intel Gaudi 3 is projected to deliver 50% faster time-to-train on average[3] across Llama2 models with 7B and 13B parameters, and GPT-3 175B parameter model. Additionally, Intel Gaudi 3 accelerator inference throughput is projected to outperform the H100 by 50% on average[1] and 40% for inference power-efficiency averaged[2] across Llama 7B and 70B parameters, and Falcon 180B parameter models.

Intel Gaudi 3 provides open, community-based software and industry-standard Ethernet networking. And it allows enterprises to scale flexibly from a single node to clusters, super-clusters and mega-clusters with thousands of nodes, supporting inference, fine-tuning and training at the largest scale.

Intel Gaudi 3 will be available to OEMs – including Dell Technologies, Hewlett Packard Enterprise, Lenovo and Supermicro – in the second quarter of 2024.

Read more at "Intel Tackles the GenAI Gap with Gaudi 3"

Generating Value for Customers with Intel AI Solutions

Intel outlined its strategy for open scalable AI systems, including hardware, software, frameworks and tools. Intel's approach enables a broad, open ecosystem of AI players to offer solutions that satisfy enterprise-specific GenAI needs. This includes equipment manufacturers, database providers, systems integrators, software and service providers, and others. It also allows enterprises to use the ecosystem partners and solutions that they already know and trust.

Intel shared broad momentum with enterprise customers and partners across industries to deploy Intel Gaudi accelerator solutions for new and innovative generative AI applications:

- **NAVER:** To develop a powerful large language model (LLM) for the deployment of advanced AI services globally, from cloud to on-device. NAVER has confirmed Intel Gaudi's foundational capability in executing compute operations for large-scale transformer models with outstanding performance per watt.
- **Bosch:** To explore further opportunities for smart manufacturing, including foundational models generating synthetic datasets of manufacturing anomalies to provide robust, evenly-distributed training sets (e.g., automated optical inspection).
- **IBM:** Using 5th Gen Intel® Xeon® processors for its watsonx.data™ data store and working closely with Intel to validate the watsonx™ platform for Intel Gaudi accelerators.
- **Ola/Krutrim:** To pre-train and fine-tune its first India foundational model with generative capabilities in 10 languages, producing industry-leading price/performance versus market solutions. Krutrim is now pre-training a larger foundational model on an Intel® Gaudi® 2 cluster.
- **NielsenIQ, an Advent International portfolio company:** To enhance its GenAI capabilities by training domain-specific LLMs on the world's largest consumer buying behavior database, enhancing its client service offerings while adhering to rigorous privacy standards.
- **Seekr:** Leader in trustworthy AI runs production workloads on Intel Gaudi 2, Intel® Data Center GPU Max Series and Intel® Xeon® processors in the Intel® Tiber™ Developer Cloud for LLM development and production deployment support.
- **IFF:** Global leader in food, beverage, scent and biosciences will leverage GenAI and digital twin technology to establish an integrated digital biology workflow for advanced enzyme design and fermentation process optimization.
- **CtrlS Group:** Collaborating to build an AI supercomputer for India-based customers and scaling CtrlS cloud services for India with additional Gaudi clusters.
- **Bharti Airtel:** Embracing the power of Intel's cutting-edge technology, Airtel plans to leverage its rich telecom data to enhance its AI capabilities and turbo charge the experiences of its customers. The deployments will be in line with Airtel's commitment to stay at the forefront of technological innovation and help drive new revenue streams in a rapidly evolving digital landscape.
- **Landing AI:** Fine-tuned domain-specific large vision model for use in segmenting cells and detecting cancer.
- **Roboflow:** Running production workloads of YOLOv5, YOLOv8, CLIP, SAM and ViT models for its end-to-end computer vision platform.
- **Infosys:** Global leader in next-generation digital services and consulting announced a strategic collaboration to bring Intel technologies including 4th and 5th Gen Intel Xeon processors, Intel Gaudi 2 AI accelerators and Intel® Core™ Ultra to Infosys Topaz – an AI-first set of services, solutions and platforms that accelerate business value using generative AI technologies.

Intel also announced collaborations with Google Cloud, Thales and Cohesity to leverage Intel's confidential computing capabilities in their cloud instances. This includes Intel® Trust Domain Extensions (Intel® TDX), Intel® Software Guard Extensions (Intel® SGX) and Intel's attestation service. Customers can run their AI models and algorithms in a trusted execution environment (TEE) and leverage Intel's trust services for independently verifying the trust worthiness of these TEEs.

Ecosystem Rallies to Develop Open Platform for Enterprise AI

In collaboration with Anyscale, Articul8, DataStax, Domino, Hugging Face, KX Systems, MariaDB, MinIO, Qdrant, RedHat, Redis, SAP, VMware, Yellowbrick and Zilliz, Intel announced the intention to create an open platform for enterprise AI. The industrywide effort aims to develop open, multivendor GenAI systems that deliver best-in-class ease-of-deployment, performance and value, enabled by retrieval-augmented generation. RAG enables enterprises' vast, existing proprietary data sources running on standard cloud infrastructure to be augmented with open LLM capabilities, accelerating GenAI use in enterprises.

As initial steps in this effort, Intel will release reference implementations for GenAI pipelines on secure Intel Xeon and Gaudi-based solutions, publish a technical conceptual framework, and continue to add infrastructure capacity in the Intel Tiber Developer Cloud for ecosystem development and validation of RAG and future pipelines. Intel encourages further participation of the ecosystem to join forces in this open effort to facilitate enterprise adoption, broaden solution coverage and accelerate business results.

Intel's Expanded AI Roadmap and Open Ecosystem Approach

In addition to the Intel Gaudi 3 accelerator, Intel provided updates on its next-generation products and services across all segments of enterprise AI.

New Intel® Xeon® 6 Processors: Intel Xeon processors offer performance-efficient solutions to run current GenAI solutions, including RAG, that produce business-specific results using proprietary data. Intel introduced the new brand for its next-generation processors for data centers, cloud and edge: Intel Xeon 6. Intel Xeon 6 processors with new Efficient-cores (E-cores) will deliver exceptional efficiency and launch this quarter, while Intel Xeon 6 with Performance-cores (P-cores) will offer increased AI performance and launch soon after the E-core processors.

- Intel Xeon 6 processors with E-cores (code-named Sierra Forest):
  - 4x performance per watt improvement[4] and 2.7x better rack density[5] compared with 2nd Gen Intel® Xeon® processors.
  - Customers can replace older systems at a ratio of nearly 3-to-1, drastically lowering energy consumption and helping meet sustainability goals[6].
- Intel Xeon 6 processors with P-cores (code-named Granite Rapids):
  - Incorporate software support for the MXFP4 data format, which reduces next token latency by up to 6.5x versus 4th Gen Intel® Xeon® processors using FP16, with the ability to run 70 billion parameter Llama-2 models[7].

Client, Edge and Connectivity: Intel announced momentum for client and updates to its roadmap for edge and connectivity including:

- Intel® Core™ Ultra processors are powering new capabilities for productivity, security and content creation, providing a great motivation for businesses to refresh their PC fleets. Intel expects expect to ship 40 million AI PCs in 2024, with more than 230 designs, from ultra-thin PCs to handheld gaming devices.
- Next-generation Intel Core Ultra client processor family (code-named Lunar Lake), launching in 2024, will have more than 100 platform tera operations per second (TOPS) and more than 45 neural processing unit (NPU) TOPS for next-generation AI

PCs.
- Intel announced new edge silicon across the Intel Core Ultra, Intel® Core™ and Intel® Atom processor and Intel® Arc™ graphics processing unit (GPU) families of products, targeting key markets including retail, industrial manufacturing and healthcare. All new additions to Intel's edge AI portfolio will be available this quarter and will be supported by the Intel® Tiber™ Edge Platform this year.
- Through the Ultra Ethernet Consortium (UEC), Intel is leading open Ethernet networking for AI fabrics, introducing an array of AI-optimized Ethernet solutions. Designed to transform large scale-up and scale-out AI fabrics, these innovations enable training and inferencing for increasingly vast models, with sizes expanding by an order of magnitude in each generation. The lineup includes the Intel AI NIC, AI connectivity chiplets for integration into XPUs, Gaudi-based systems, and a range of soft and hard reference AI interconnect designs for Intel Foundry.

Intel Tiber Portfolio of Business Solutions

Intel unveiled the Intel® Tiber™ portfolio of business solutions to streamline the deployment of enterprise software and services, including for GenAI.

A unified experience makes it easier for enterprise customers and developers to find solutions that fit their needs, accelerate innovation and unlock value without compromising on security, compliance or performance. Customers can begin exploring the Intel Tiber portfolio starting today, with a full rollout planned for the third quarter of 2024. Learn more at Intel Tiber website.

Intel's announcements at Vision 2024 underscore the company's commitment to making AI accessible, open and secure for enterprises worldwide. With these new solutions and collaborations, Intel is poised to lead the way in the AI revolution, unlocking unprecedented value for businesses everywhere.

For more information on Intel's AI solutions and Vision 2024 announcements, please visit the Intel Newsroom.

Forward-Looking Statements

This release contains forward-looking statements, including with respect to:

- our business plans and strategy and anticipated benefits therefrom;
- our AI strategy and AI accelerators;
- our open platforms approach and ecosystem support with respect to AI; and
- other characterizations of future events or circumstances.

Such statements involve many risks and uncertainties that could cause our actual results to differ materially from those expressed or implied, including those associated with:

- the high level of competition and rapid technological change in our industry;
- the significant long-term and inherently risky investments we are making in R&D and manufacturing facilities that may not realize a favorable return;
- the complexities and uncertainties in developing and implementing new semiconductor products and manufacturing process technologies;

- our ability to time and scale our capital investments appropriately and successfully secure favorable alternative financing arrangements and government grants;
- implementing new business strategies and investing in new businesses and technologies;
- changes in demand for our products;
- macroeconomic conditions and geopolitical tensions and conflicts, including geopolitical and trade tensions between the US and China, the impacts of Russia's war on Ukraine, tensions and conflict affecting Israel, and rising tensions between mainland China and Taiwan;
- the evolving market for products with AI capabilities;
- our complex global supply chain, including from disruptions, delays, trade tensions and conflicts, or shortages;
- product defects, errata and other product issues, particularly as we develop next-generation products and implement next-generation manufacturing process technologies;
- potential security vulnerabilities in our products;
- increasing and evolving cybersecurity threats and privacy risks;
- IP risks including related litigation and regulatory proceedings;
- the need to attract, retain, and motivate key talent;
- strategic transactions and investments;
- sales-related risks, including customer concentration and the use of distributors and other third parties;
- our significantly reduced return of capital in recent years;
- our debt obligations and our ability to access sources of capital;
- complex and evolving laws and regulations across many jurisdictions;
- fluctuations in currency exchange rates;
- changes in our effective tax rate;
- catastrophic events;
- environmental, health, safety, and product regulations;
- our initiatives and new legal requirements with respect to corporate responsibility matters; and
- other risks and uncertainties described in this release, our most recent Annual Report on Form 10-K and our other filings with the U.S. Securities and Exchange Commission (SEC).

All information in this release reflects management's expectations as of the date of this release, unless an earlier date is specified. We do not undertake, and expressly disclaim any duty, to update such statements, whether as a result of new information, new developments, or otherwise, except to the extent that disclosure may be required by law.

About Intel

Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore’s Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers’ greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel’s innovations, go to newsroom.intel.com and intel.com.

[1] NV H100 comparison based on https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8 , March 28, 2024. Reported numbers are per GPU. Vs Intel® Gaudi® 3 projections for LLAMA2-7B, LLAMA2-70B & Falcon 180B projections. Results may vary.
[2] NV H100 comparison based on https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8 , March 28, 2024. Reported numbers are per GPU. Vs Intel® Gaudi® 3 projections for LLAMA2-7B, LLAMA2-70B & Falcon 180B. Power efficiency for both Nvidia and Gaudi 3 based on internal estimates. Results may vary.
[3] NV H100 comparison based on: https://developer.nvidia.com/deep-learning-performance-training-inference/training, March 28, 2024. “Large Language Model” tab vs. Intel® Gaudi® 3 projections for LLAMA2-7B, LLAMA2-13B & GPT3-175B as of 3/28/2024. Results may vary.
[4] Based on architectural projections as of Feb. 14, 2023, vs. prior generation platforms. Your results may vary.
[5] Based on architectural projections as of Feb. 14, 2023, vs. prior generation platforms. Your results may vary.
[6] Based on architectural projections as of Feb. 14, 2023, vs. prior generation platforms. Your results may vary.
[7] See Vision 2024 section of intel.com/performanceindex for workloads and configurations. Results may vary.

View source version on businesswire.com:
https://www.businesswire.com/news/home/20240409077438/en/

Danielle Mann
1-973-997-1154
danielle.mann@intel.com

Source: Intel Corporation