June 3, 2024

Intel Accelerates AI Everywhere at Computex 2024; Redefines Compute Power, Performance and Affordability with new Xeon 6, Gaudi Accelerators and Lunar Lake Architecture to Grow AI PC Leadership

AI runs best on Intel across the compute continuum from the data center, cloud and network to the edge and PC.

NEWS HIGHLIGHTS

- Launches Intel® Xeon® 6 processors with Efficient-cores (E-cores), delivering performance and power efficiency for high-density, scale-out workloads in the data center. Enables 3:1 rack consolidation, rack-level performance gains of up to 4.2x and performance per watt gains of up to 2.6x[1].
- Announces pricing for Intel® Gaudi® 2 and Intel® Gaudi® 3 AI accelerator kits, delivering high performance with up to one-third lower cost compared to competitive platforms[2]. The combination of Xeon processors with Gaudi AI accelerators in a system offers a powerful solution for making AI faster, cheaper and more accessible.
- Unveils Lunar Lake client processor architecture to continue to grow the AI PC category. The next generation of AI PCs – with breakthrough x86 power efficiency and no-compromise application compatibility – will deliver up to 40% lower system-on-chip (SoC) power when compared with the previous generation[3].

TAIPEI, Taiwan--(BUSINESS WIRE)-- Today at Computex, Intel unveiled cutting-edge technologies and architectures poised to dramatically accelerate the AI ecosystem – from the data center, cloud and network to the edge and PC. With more processing power, leading-edge power efficiency and low total cost of ownership (TCO), customers can now capture the complete AI system opportunity.

This press release features multimedia. View the full release here: https://www.businesswire.com/news/home/20240603799554/en/

At Computex in Taipei, Taiwan, on June 4, 2024, Intel launched the Intel Xeon 6 processors with Efficient-cores (E-cores). For companies looking to refresh aging infrastructure to help reduce costs and free up space, Intel Xeon 6 with E-cores offers significant rack density advantages, enabling a 3-to-1 rack-level consolidation. (Credit: Intel Corporation)

“AI is driving one of the most consequential eras of innovation the industry has ever

seen," said Intel CEO Pat Gelsinger. "The magic of silicon is once again enabling exponential advancements in computing that will push the boundaries of human potential and power the global economy for years to come."

More: Intel at Computex 2024 (Press Kit)

Gelsinger continued, "Intel is one of the only companies in the world innovating across the full spectrum of the AI market opportunity – from semiconductor manufacturing to PC, network, edge and data center systems. Our latest Xeon, Gaudi and Core Ultra platforms, combined with the power of our hardware and software ecosystem, are delivering the flexible, secure, sustainable and cost-effective solutions our customers need to maximize the immense opportunities ahead."

Intel Enables AI Everywhere

During his Computex keynote, Gelsinger highlighted the benefits of open standards and Intel's powerful ecosystem helping to accelerate the AI opportunity. He was joined by luminaries and industry-leading companies voicing support, including Acer Chairman and CEO Jason Chen, ASUS Chairman Jonney Shih, Microsoft Chairman and CEO Satya Nadella, and Inventec's President Jack Tsai, among others.

Gelsinger and others made it clear that Intel is revolutionizing AI innovation and delivering next-generation technologies ahead of schedule. In just six months, the company went from launching 5th Gen Intel® Xeon® processors to introducing the inaugural member of the Xeon 6 family; from previewing Gaudi AI accelerators to offering enterprise customers a cost-effective, high-performance generative AI (GenAI) training and inference system; and from ushering in the AI PC era with Intel® Core™ Ultra processors in more than 8 million devices to unveiling the forthcoming client architecture slated for release later this year.

With these developments, Intel is accelerating execution while pushing the boundaries of innovation and production speed to democratize AI and catalyze industries.

Modernizing the Data Center for AI: Intel Xeon 6 Processors Improve Performance and Power Efficiency for High-Density, Scale-Out Workloads

As digital transformations accelerate, companies face mounting pressures to refresh their aging data center systems to capture cost savings, achieve sustainability goals, maximize physical floor and rack space, and create brand-new digital capabilities across the enterprise.

The entire Xeon 6 platform and family of processors is purpose-built for addressing these challenges with both E-core (Efficient-core) and P-core (Performance-core) SKUs to address the broad array of use cases and workloads, from AI and other high-performance compute needs to scalable cloud-native applications. Both E-cores and P-cores are built on a compatible architecture with a shared software stack and an open ecosystem of hardware and software vendors.

The first of the Xeon 6 processors to debut is the Intel Xeon 6 E-core (code-named Sierra Forest), which is available beginning today. Xeon 6 P-cores (code-named Granite Rapids) are expected to launch next quarter.

With high core density and exceptional performance per watt, Intel Xeon 6 E-core delivers efficient compute with significantly lower energy costs. The improved performance with increased power efficiency is perfect for the most demanding high-density, scale-out workloads, including cloud-native applications and content delivery networks, network microservices and consumer digital services.

Additionally, Xeon 6 E-core has tremendous density advantages, enabling rack-level consolidation of 3-to-1, providing customers with a rack-level performance gain of up to 4.2x and performance per watt gain of up to 2.6x when compared with 2nd Gen Intel® Xeon® processors on media transcode workloads[1]. Using less power and rack space, Xeon 6 processors free up compute capacity and infrastructure for innovative new AI projects.

Fact Sheet: Intel Xeon 6 Processors

Providing High Performance GenAI at Significantly Lower Total Cost with Intel Gaudi AI Accelerators

Today, harnessing the power of generative AI becomes faster and less expensive. As the dominant infrastructure choice, x86 operates at scale in nearly all data center environments, serving as the foundation for integrating the power of AI while ensuring cost-effective interoperability and the tremendous benefits of an open ecosystem of developers and customers.

Intel Xeon processors are the ideal CPU head node for AI workloads and operate in a system with Intel Gaudi AI accelerators, which are purposely designed for AI workloads. Together, these two offer a powerful solution that seamlessly integrates into existing infrastructure.

As the only MLPerf-benchmarked alternative to Nvidia H100 for training and inference of large language models (LLM), the Gaudi architecture gives customers the GenAI performance they seek with a price-performance advantage that provides choice and fast deployment time at lower total cost of operating.

A standard AI kit including eight Intel Gaudi 2 accelerators with a universal baseboard (UBB) offered to system providers at $65,000 is estimated to be one-third the cost of comparable competitive platforms. A kit including eight Intel Gaudi 3 accelerators with a UBB will list at $125,000, estimated to be two-thirds the cost of comparable competitive platforms[4].

Intel Gaudi 3 accelerators will deliver significant performance improvements for training and inference tasks on leading GenAI models, helping enterprises unlock the value in their proprietary data. Intel Gaudi 3 in an 8,192-accelerator cluster is projected to offer up to 40% faster time-to-train[5] versus the equivalent size Nvidia H100 GPU cluster and up to 15% faster training[6] throughput for a 64-accelerator cluster versus Nvidia H100 on the Llama2-70B model. In addition, Intel Gaudi 3 is projected to offer an average of up to 2x faster inferencing[7] versus Nvidia H100, running popular LLMs such as Llama-70B and Mistral-7B.

To make these AI systems broadly available, Intel is collaborating with at least 10 top global system providers, including six new providers who announced they're bringing Intel Gaudi 3 to market. Today's new collaborators include Asus, Foxconn, Gigabyte, Inventec, Quanta

and Wistron, expanding the production offerings from leading system providers Dell, Hewlett Packard Enterprise, Lenovo and Supermicro.

Accelerating On-Device AI for laptop PCs; New Architecture Delivers 3x AI Compute and Incredible Power-Efficiency

Beyond the data center, Intel is scaling its AI footprint at the edge and in the PC. With more than 90,000 edge deployments and 200 million CPUs delivered to the ecosystem, Intel has enabled enterprise choice for decades.

Today the AI PC category is transforming every aspect of the compute experience, and Intel is at the forefront of this category-creating moment. It's no longer just about faster processing speeds or sleeker designs, but rather creating edge devices that learn and evolve in real time – anticipating user needs, adapting to their preferences, and heralding an entirely new era of productivity, efficiency and creativity.

AI PCs are projected to make up 80% of the PC market by 2028, according to Boston Consulting Group. In response, Intel has moved quickly to create the best hardware and software platform for the AI PC, enabling more than 100 independent software vendors (ISVs), 300 features and support of 500 AI models across its Core Ultra platform.

Quickly building on these unmatched advantages, the company today revealed the architectural details of Lunar Lake – the flagship processor for the next generation of AI PCs. With a massive leap in graphics and AI processing power, and a focus on power-efficient compute performance for the thin-and-light segment, Lunar Lake will deliver up to 40% lower SoC power[3] and more than 3 times the AI compute[8]. It's expected to ship in the third quarter of 2024, in time for the holiday buying season.

Lunar Lake's all-new architecture will enable:

- New Performance-cores (P-cores) and Efficient-cores (E-cores) deliver significant performance and energy efficiency improvements.
- A fourth-generation Intel neural processing unit (NPU) with up to 48 tera-operations per second (TOPS) of AI performance. This powerful NPU delivers up to 4x AI compute over the previous generation, enabling corresponding improvements in generative AI.
- An all-new GPU design, code-named Battlemage, combines two new innovations: X^e2 GPU cores for graphics and X^e Matrix Extension (XMX) arrays for AI. The X^e2 GPU cores improve gaming and graphics performance by 1.5x over the previous generation, while the new XMX arrays enable a second AI accelerator with up to 67 TOPS of performance for extraordinary throughput in AI content creation.
- Advanced low-power island, a novel compute cluster and Intel innovation that handles background and productivity tasks with extreme efficiency, enabling amazing laptop battery life.

As others prepare to enter the AI PC market, Intel is already shipping at scale, delivering more AI PC processors through 2024's first quarter than all competitors together. Lunar Lake is set to power more than 80 different AI PC designs from 20 original equipment manufacturers (OEMs). Intel expects to deploy more than 40 million Core Ultra processors in market this year.

Fact Sheet: Intel Unveils Lunar Lake Architecture

As Gordon Moore famously said, “Whatever has been done, can be outdone,” and Intel stands as the vanguard of this relentless pursuit of progress. With global scale spanning client, edge, data center and cloud, a robust ecosystem grounded in open standards, and powerful, cost-effective solutions, Intel is not just powering AI everywhere; it is shaping its future. Today’s announcements are not just a technological leap, but an invitation to customers and partners to seize unprecedented possibilities and pioneer the next era of their own innovations.

Forward-Looking Statements

This release contains forward-looking statements that involve a number of risks and uncertainties, including with respect to Intel’s product roadmap and anticipated product sales and competitiveness and projected growth and trends in markets relevant to Intel’s business. 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 the Middle East, 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 2023 Form 10-K, and our other filings with the SEC.

Given these risks and uncertainties, readers are cautioned not to place undue reliance on such forward-looking statements. Readers are urged to carefully review and consider the various disclosures made in this release and in other documents we file from time to time with the SEC that disclose risks and uncertainties that may affect our business.

Unless specifically indicated otherwise, the forward-looking statements in this release are based on management's expectations as of the date of this release, unless an earlier date is specified, including expectations based on third-party information and projections that management believes to be reputable. 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.

AI runs best on Intel across the compute continuum from the data center, cloud and network to the edge and PC as of May 2024, based on broad compatibility, extensive software options, unique architecture, and impressive performance of Intel offerings, which combine to deliver the best overall AI experience, including in comparison to competitive offerings. See intel.com/performanceindex for details. Results may vary.

[1] See [7T1] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.

[2] Pricing estimates based on publicly available information and Intel internal analysis.

[3] Disclaimer for footnote: Power measurements are based on Lunar Lake reference platform using YouTube 4K 30fps AV1. See backup for details. Results may vary.

[4] Pricing guidance for cards and systems is for modeling purposes only. Please consult your original equipment manufacturer (OEM) of choice for final pricing. Results may vary based upon volumes and lead times.

[5] Source for Nvidia H100 GPT 3 performance: https://mlcommons.org/benchmarks/training/, v3.1, closed division round. Accessed April 30, 2024.
Intel Gaudi 3 measurements and projections by Habana Labs, April 2024; Results may vary Intel Gaudi 3 performance projections are not verified by MLCommons Association. The

MLPerf name and logo are registered and unregistered trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See http://www.mlcommons.org/ for more information.

[6] Source for Nvidia H100 LLAMA2-70B performance https://developer.nvidia.com/deep-learning-performance-training-inference/training, April 29, 2024, a “Large Language Model” tab.
Intel Gaudi 3 measurements and projections by Habana Labs, April 2024; Results may vary

[7] Source for Nvidia performance: Overview — tensorrt_llm documentation (nvidia.github.io), May, 2024. Reported numbers are per GPU.
Intel Gaudi 3 projections by Habana Labs, April 2024; Results may vary

[8] Based on total number of platform Tops on Lunar Lake vs. prior generation.

© 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.

View source version on businesswire.com:
https://www.businesswire.com/news/home/20240603799554/en/

Cory Pforzheimer
1-805-895-2281
cory.pforzheimer@intel.com

Source: Intel Corporation