September 24, 2024

Intel Unveils Next-Generation AI Solutions with the Launch of Xeon 6 and Gaudi 3

Intel enables a new era of high-performance enterprise AI systems and solutions.

NEWS HIGHLIGHTS

- Intel launches Xeon 6 with Performance-cores (P-cores), doubling the performance for AI and HPC workloads.
- New Gaudi 3 AI accelerators offer up to 20 percent more throughput and 2x price/performance vs H100 for inference of LLaMa 2 70B[1].

SANTA CLARA, Calif.--(BUSINESS WIRE)-- As AI continues to revolutionize industries, enterprises are increasingly in need of infrastructure that is both cost-effective and available for rapid development and deployment. To meet this demand head-on, Intel today launched Xeon 6 with Performance-cores (P-cores) and Gaudi 3 AI accelerators, bolstering the company's commitment to deliver powerful AI systems with optimal performance per watt and lower total cost of ownership (TCO).

"Demand for AI is leading to a massive transformation in the data center, and the industry is asking for choice in hardware, software and developer tools," said Justin Hotard, Intel executive vice president and general manager of the Data Center and Artificial Intelligence Group. "With our launch of Xeon 6 with P-cores and Gaudi 3 AI accelerators, Intel is enabling an open ecosystem that allows our customers to implement all of their workloads with greater performance, efficiency and security."

Introducing Intel Xeon 6 with P-cores and Gaudi 3 AI accelerators

Intel's latest advancements in AI infrastructure include two major updates to its data center portfolio:

- **Intel® Xeon® 6 with P-cores:** Designed to handle compute-intensive workloads with exceptional efficiency, Xeon 6 delivers twice the performance of its predecessor[2]. It features increased core count, double the memory bandwidth and AI acceleration capabilities embedded in every core. This processor is engineered to meet the performance demands of AI from edge to data center and cloud environments.

- **Intel® Gaudi® 3 AI Accelerator:** Specifically optimized for large-scale generative AI, Gaudi 3 boasts 64 Tensor processor cores (TPCs) and eight matrix multiplication engines (MMEs) to accelerate deep neural network computations. It includes 128 gigabytes (GB) of HBM2e memory for training and inference, and 24 200 Gigabit (Gb) Ethernet ports for scalable networking. Gaudi 3 also offers seamless compatibility with the PyTorch framework and advanced Hugging Face transformer and diffuser models.

Intel recently announced a collaboration with IBM to deploy Intel Gaudi 3 AI accelerators as a service on IBM Cloud. Through this collaboration, Intel and IBM aim to lower the total cost of ownership to leverage and scale AI, while enhancing performance.

Enhancing AI Systems with TCO Benefits

Deploying AI at scale involves considerations such as flexible deployment options, competitive price-performance ratios and accessible AI technologies. Intel's robust x86 infrastructure and extensive open ecosystem position it to support enterprises in building high-value AI systems with an optimal TCO and performance per watt. Notably, 73% of GPU-accelerated servers use Intel Xeon as the host CPU[3].

Intel partners with leading OEMs including Dell Technologies and Supermicro to develop co-engineered systems tailored to specific customer needs for effective AI deployments. Dell Technologies is currently co-engineering RAG-based solutions leveraging Gaudi 3 and Xeon 6.

Bridging the Gap from Prototypes to Production with Co-Engineering Efforts

Transitioning generative AI (Gen AI) solutions from prototypes to production-ready systems presents challenges in real-time monitoring, error handling, logging, security and scalability. Intel addresses these challenges through co-engineering efforts with OEMs and partners to deliver production-ready retrieval-augmented generation (RAG) solutions.

These solutions, built on the Open Platform Enterprise AI (OPEA) platform, integrate OPEA-based microservices into a scalable RAG system, optimized for Xeon and Gaudi AI systems, designed to allow customers to easily integrate applications from Kubernetes, Red Hat OpenShift AI and Red Hat Enterprise Linux AI.

Expanding Access to Enterprise AI Applications

Intel's Tiber portfolio offers business solutions to tackle challenges such as access, cost, complexity, security, efficiency and scalability across AI, cloud and edge environments. The Intel® Tiber™ Developer Cloud now provides preview systems of Intel Xeon 6 for tech evaluation and testing. Additionally, select customers will gain early access to Intel Gaudi 3 for validating AI model deployments, with Gaudi 3 clusters to begin rolling out next quarter for large-scale production deployments.

New service offerings include SeekrFlow, an end-to-end AI platform from Seekr for developing trusted AI applications. The latest updates feature Intel Gaudi software's newest release and Jupyter notebooks loaded with PyTorch 2.4 and Intel oneAPI and AI tools 2024.2, which include new AI acceleration capabilities and support for Xeon 6 processors.

1 See intel.com/processorclaims: Intel Gaudi 3. Results may vary.

2 See intel.com/processorclaims: Intel Xeon 6. Results may vary.

3 Source: IDC Server Tracker report, based on Q1'24 system volume.

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.

View source version on businesswire.com:
https://www.businesswire.com/news/home/20240924255313/en/

Bats Jafferji
+1 603-809-5145
bats.jafferji@intel.com

Source: Intel Corporation