October 12, 2026

Intel Accelerates Open Infrastructure for the Agentic AI Era

SANTA CLARA, Calif.--(BUSINESS WIRE)-- At the Open Compute Project (OCP) Global Summit today, Intel corporation shared new technical details on the Intel GPU D-series, formerly code-named Crescent Island, introduced new Intel Xeon-based rack designs and highlighted open software and ecosystem innovations.

As open models reshape the economics of agentic AI, demand is growing for more flexible, cost- and power-efficient infrastructure. Together, these advancements reinforce Intel's approach to open, heterogeneous AI infrastructure, combining general-purpose compute, purpose-built acceleration, open software and standards-based systems to optimize performance and efficiency across AI workloads.

Agentic AI is transforming inference into a multi-step workflow spanning retrieval, reasoning, orchestration, tool use and execution, with x86 CPUs playing a central role. While open models give enterprises greater choice, power constraints, memory bottlenecks and cost pressures make scaling AI increasingly a systems challenge.

Meeting these demands requires turnkey, heterogeneous infrastructure that is easy to deploy and manage. Building on the established x86 software ecosystem, Intel brings CPUs, GPUs and specialized accelerators together through open, interoperable software to match the right compute resources to each task, improving utilization, performance and efficiency across the entire workflow.

"Open models are expanding the options enterprises have for deploying AI," said Nicolas Dubé, corporate vice president, Data Center Systems and Solutions Group, Intel. "To realize their full potential, organizations need infrastructure that can match different workloads with the right intelligence, right compute resources and right economics. That's what we mean by 'put every prompt in its place.'"

Intel Xeon: The Compute Foundation for Agentic AI

At OCP, Intel showcased new Xeon-based rack designs developed with ecosystem partners to meet the growing compute demands of agentic AI. Many leverage the Open Compute Project's Open Rack v3 (ORv3) standard to simplify deployment and integration within open data center infrastructure.

As agentic workflows grow more complex, they place greater demands on general-purpose compute to orchestrate workloads, process data, manage I/O and execute tasks that keep accelerators productive. To address these needs, select Xeon 6+ configurations scale beyond 46,000 cores per rack at 133.45 kW total rack power, delivering greater CPU compute density to support agentic AI at scale.

Building on two decades of rack-scale innovation and the broad x86 software ecosystem, Xeon provides the high-density, general-purpose compute foundation needed to scale agentic AI. Working alongside GPUs and specialized accelerators, Xeon orchestrates workloads, executes tasks and manages data movement to keep the entire system productive. As AI infrastructure evolves, performance is increasingly measured by the number of concurrent agents a rack can sustain at target latency – and Xeon helps maximize that delivered work by keeping data flowing, tools executing and accelerators fed.

Intel GPU D-Series: Purpose-Built for Efficient Inference at Scale

Intel GPU D-series, code-named Crescent Island, is designed to improve the economics of sustained, memory-intensive AI inference. Built on Intel's latest Xe3P architecture, the GPU combines up to 480 GB of LPDDR5X memory, 1.5 TB/s of memory bandwidth and 1.3 petaflops of FP8 compute within a 350-watt power envelope. This balance of memory capacity, compute performance and power efficiency reflects Intel's approach to optimizing infrastructure for real-world inference workloads.

Designed for large language models, long-context workloads and concurrent inference services, Intel GPU D-series supports models with up to 300 billion parameters on a single card and more than one trillion parameters in an eight-GPU system. Target applications range from multi-tenant AI services to robotics and other latency-sensitive workloads.

With a standard PCIe form factor, air-cooled design and common Xe software foundation, Intel GPU D-series is designed to simplify deployment within existing data center environments while supporting application portability across Intel GPU platforms. This allows customers to scale inference without specialized rack or cooling infrastructure.

“Inference is becoming the defining AI workload, and enterprises need infrastructure that delivers the right performance at the right cost,” said Anil Nanduri, vice president, AI Products & GTM, Intel Data Center Group. “With Intel GPU D-series, we're focused on making high-capacity inference practical to deploy and economical to scale.”

To help bring these capabilities into production, Intel is collaborating with a broad ecosystem of OEMs, ODMs, neoclouds, independent software vendors, appliance providers and AI-native companies. Together, these efforts focus on validated platforms, software integration and deployment-ready solutions that reduce complexity and help customers move more quickly from model evaluation to production.

With customer sampling underway, Intel continues to expand its hardware, software and open-source ecosystem, giving enterprises greater flexibility to deploy and scale inference across existing and new data center infrastructure.

Open Software for Heterogeneous AI

Making heterogeneous AI infrastructure work at scale requires open software that connects compute resources, simplifies deployment and gives developers the flexibility to run workloads on the right architecture. Intel's open software strategy emphasizes native upstream integration with industry-standard AI frameworks, inference runtimes and deployment environments across CPUs and GPUs. By building on the tools developers already use, Intel is helping simplify development, optimize performance and accelerate

deployment through several key capabilities:

Simplified deployment across Intel GPUs: Built on a common Xe software foundation, Intel GPU D-series supports software portability across Intel GPU platforms, helping developers reuse existing applications and reduce the effort required to deploy workloads on new hardware.

Open frameworks and performance software: Intel supports the AI frameworks, serving engines and development tools already used across the industry, complemented by performance libraries and programmable software layers to optimize inference throughput. These same frameworks support inference and agentic applications running on Intel Xeon processors, enabling developers to work across CPU and GPU architectures.

Day 0 support for emerging open models: Through upstream framework integration and model enablement, Intel is working to support new open models as they become available across Intel GPU D-series and Xeon processors. Recent efforts include Day 0 support for Meta's Muse Glimmer, helping developers more quickly evaluate and deploy emerging agentic AI workloads.

Built for real-world AI applications and agents: Intel's software stack supports diverse workloads, from high-throughput token generation to coding agents and complex enterprise workflows. The Intel AI for Enterprise Agent Toolkit is optimized and deployable on Xeon processors today, providing orchestration capabilities for production agentic applications. Its provider-based architecture is designed to extend to Intel GPUs, giving enterprises a flexible foundation for multimodal workflows and a path to scale across heterogeneous compute resources.

Open Infrastructure for Heterogeneous AI

Intel continues to advance open infrastructure through its work with the Open Compute Project, where it has made 116 contributions since 2011.

This year, Intel is showcasing its contribution to the Open Adapter Card workgroup, developed with industry partners to define a standardized sub-rack accelerator card form factor. The effort is designed to improve mechanical, electrical and management interoperability across system designs, giving system builders and cloud providers a more consistent path for integrating accelerators into open infrastructure.

For customers, the value of this approach is the freedom to evolve their AI systems without starting over. By building on existing investments and incorporating new compute technologies as workloads change, enterprises can scale inference more efficiently, manage costs and maintain choice over how their infrastructure evolves.

Underlining this, Kevork Kechichian, executive vice president and general manager, Data Center Group, Intel said: “Agentic AI is changing what customers need from their data centers. Success will depend not on any single component, but on open, heterogeneous infrastructure that brings together Xeon, purpose-built acceleration and interoperable software. Intel is helping customers build on their existing investments while creating a flexible foundation to scale emerging AI workloads efficiently.”

Attendees can visit Intel at booth B32 and hear from Intel leaders, including Nic Dubé's keynote, ‘Put Every Prompt in Its Place’, on October 13 at 10:30 a.m. PT. Intel will also participate in more than 30 technical sessions and speaking engagements. For more information, see Intel's OCP 2026 media alert.

Notes and Disclaimers Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates.

Intel branded PCIe card has 160GB LPDDR5x, design enables partners to build ODM branded cards with flexible options up to 480GB memory

Intel technologies may require enabled hardware, software or service activation.

About Intel

Intel (Nasdaq: INTC) designs and manufactures advanced semiconductors that connect and power the modern world. Every day, our engineers create new technologies that enhance and shape the future of computing to enable new possibilities for every customer we serve. Learn more at intel.com.

© 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/20261012581523/en/

Andrew Evangelista
Press Relations
andrew.evangelista@intel.com

Source: Intel Corporation