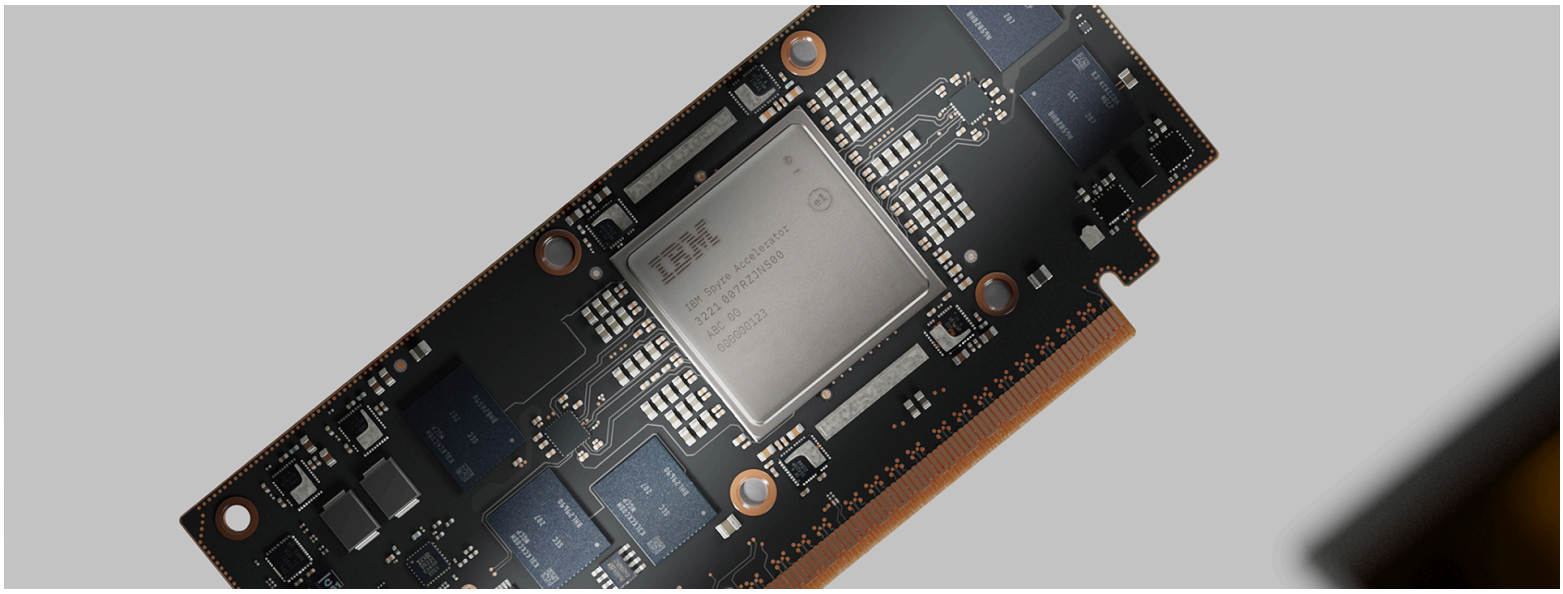
IBM Introduces the Spyre Accelerator for Commercial Availability

Coming this Fall to IBM Z, LinuxONE and Power, IBM Spyre Accelerator Enables Enterprises to Scale Generative and Agentic AI Workloads

Oct 7, 2025



ARMONK, N.Y., Oct. 7, 2025 /PRNewswire/ -- IBM (NYSE: IBM) today announced the upcoming general availability of the IBM Spyre Accelerator, an AI accelerator enabling low-latency inferencing to support generative and agentic AI use cases while prioritizing the security and resilience of core workloads. Earlier this year, IBM announced the Spyre Accelerator would be available in IBM z17, LinuxONE 5, and Power11 systems. Spyre will be generally available on October 28 for IBM z17 and LinuxONE 5 systems, and in early December for Power11 servers.

Today's IT landscape is changing from traditional logic workflows to agentic AI inferencing. AI agents require low-latency inference and real-time system responsiveness. IBM recognized the need for mainframes and servers to run AI models along with the most demanding enterprise workloads without compromising on throughput. To address this demand, clients need AI inferencing hardware that supports generative and agentic AI while maintaining the security and resilience of core data, transactions, and applications. The accelerator is also built to enable clients to keep mission-critical data on-prem to mitigate risk while addressing operational and energy efficiency.

The IBM Spyre Accelerator reflects the strength of IBM's research-to-product pipeline, combining breakthrough innovation from the IBM Research AI Hardware Center with enterprise-grade development from IBM Infrastructure. Initially introduced as a prototype chip, Spyre was refined through rapid iteration, including cluster deployments at IBM's Yorktown Heights campus, and with collaborators like the University at Albany's Center for Emerging Artificial Intelligence Systems.

The IBM Research prototype has evolved into an enterprise-grade product for use in IBM Z, LinuxONE and Power systems. Today, the Spyre Accelerator is a commercial system-on-a-chip with 32 individual accelerator cores and 25.6 billion transistors. Produced using 5nm node technology, each Spyre is mounted on a 75-watt PCIe card, which makes it possible to cluster up to 48 cards in an IBM Z or LinuxONE system or 16 cards in an IBM Power system to scale AI capabilities.

"One of our key priorities has been advancing infrastructure to meet the demands of new and emerging AI workloads," said **Barry Baker, COO, IBM Infrastructure & GM, IBM Systems**. "With the Spyre Accelerator, we're extending the capabilities of our systems to support multi-model AI – including generative and agentic AI. This innovation positions clients to scale their AI-enabled mission-critical workloads with uncompromising security, resilience, and efficiency, while unlocking the value of their enterprise data."

"We launched the IBM Research AI Hardware Center in 2019 with a mission to meet the rising computational demands of AI, even before the surge in LLMs and AI models we've recently seen," said **Mukesh Khare, GM of IBM Semiconductors and VP of Hybrid Cloud, IBM**. "Now, amid increasing demand for advanced AI capabilities, we're proud to see the first chip from the Center enter commercialization, designed to deliver improved performance and productivity to IBM's mainframe and server clients."

For IBM clients, Spyre Accelerators offer fast, secured processing with on-prem AI acceleration. This marks a significant milestone, allowing businesses to leverage AI at scale while keeping data on IBM Z, LinuxONE and Power systems. In mainframe systems, coupled with the Telum II processor for IBM Z and LinuxONE, Spyre offers enhanced security, low latency, and high transaction rate processing power. Leveraging this advanced hardware and software stack, businesses can use Spyre to scale multiple AI models to power predictive use cases such as advanced fraud detection and retail automation.

On IBM Power-based servers, Spyre customers can leverage a catalog of AI services, enabling end-to-end AI for enterprise workflows. Clients can install the AI services from the catalog with just one click.[1] Spyre Accelerator for Power, combined with an on-chip accelerator (MMA), also accelerates data conversion for generative AI to deliver high throughput for deep process integrations. Additionally, with a prompt size of 128, it enables the ingestion of more than 8 million documents for knowledge base integration in an hour[2]. This performance, combined with the IBM software stack, security, scalability, and energy efficiency, supports clients on their journey to integrating generative AI frameworks into their enterprise workloads.

To learn more about the IBM Spyre Accelerator, visit http://www.ibm.com/solutions/ai-accelerator.

Additional resources:

- IBM LinuxONE blog
- IBM Power blog
- IBM Research blog
- IBM Z blog

About IBM
IBM is a leading provider of global hybrid cloud and AI, and consulting expertise. We help clients in more than 175 countries capitalize on insights from their data, streamline business processes, reduce costs and gain a competitive edge in their industries. Thousands of governments and corporate entities in critical infrastructure areas such as financial services, telecommunications and healthcare rely on IBM's hybrid cloud platform and Red Hat OpenShift to affect their digital transformations quickly, efficiently and securely. IBM's breakthrough innovations in AI, quantum computing, industry-specific cloud solutions and consulting deliver open and flexible options to our clients. All of this is backed by IBM's long-standing commitment to trust, transparency, responsibility, inclusivity and service. Visit www.ibm.com for more information.

Media Contacts
Willa Hahn, willa.hahn@ibm.com
Chase Skinner, Chase.Skinner@ibm.com

[1] AI service of the IBM-supported catalog is delivered as one or a set of containers that can be deployed with a single deployment command. The provided UI for the catalog executes such commands in the backend based on a single click within the UI page of the respective AI service.

[2] Based upon internal testing running 1M unit data set with prompt size 128, batch size 128 using 1-card container. Individual results may vary based on workload size, use of storage subsystems and other conditions.

SOURCE IBM

Release Categories

Artificial intelligence Corporate Hybrid cloud

more articles

IBM to Announce Third-Quarter 2025 Financial Results

→

S&P Global and IBM Deploy Agentic AI to Improve Enterprise Operations

→

IBM Spyre® Accelerator and Telum II® Processor: Capturing AI value at a trusted enterprise level

→

Assets