IBM Canada Newsroom

IBM z17: The First Mainframe Fully Engineered for the AI Age

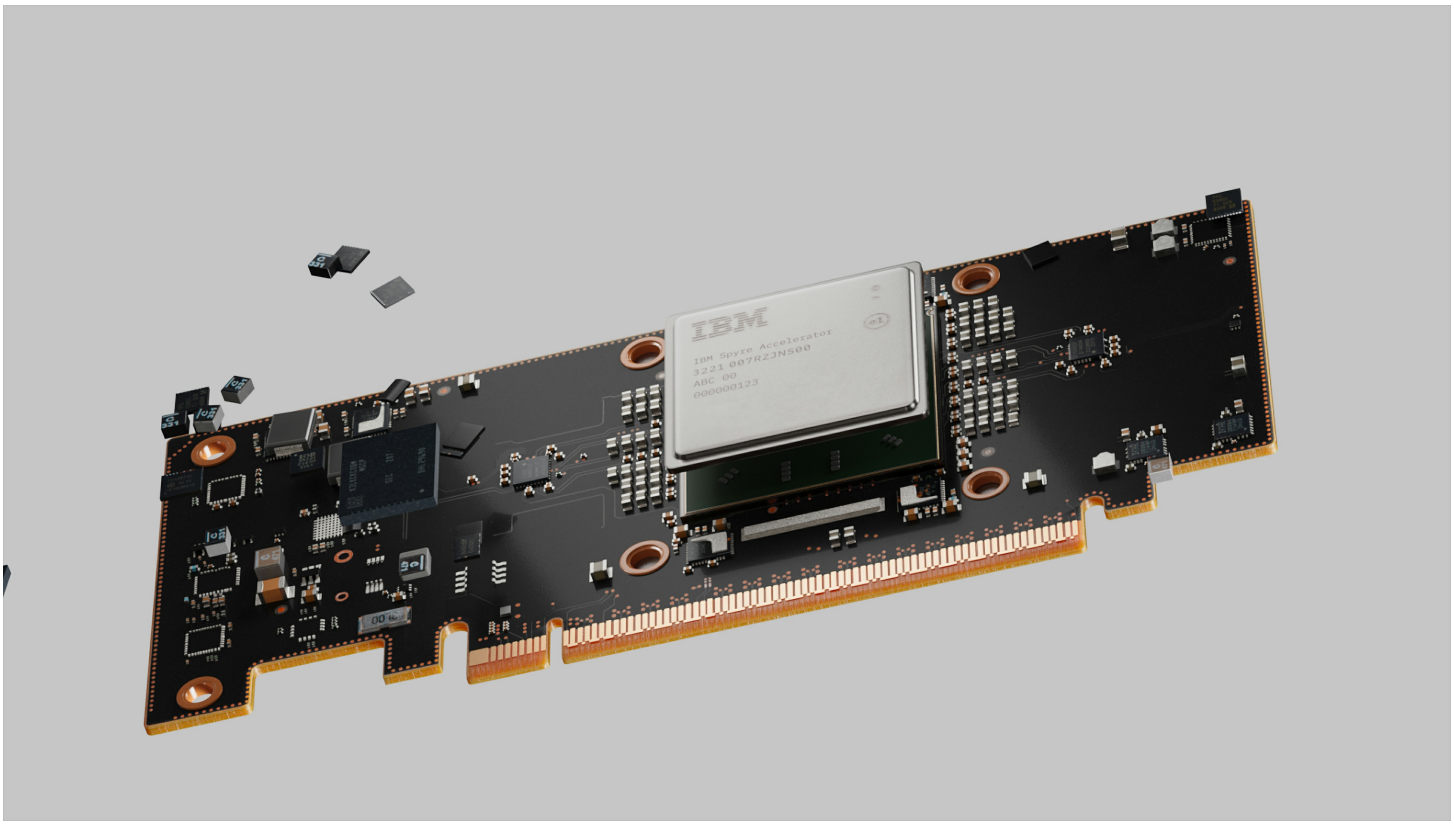
New Innovations Unlock Capabilities for Enterprise-Scale AI, Including Large Language Models and Generative AI

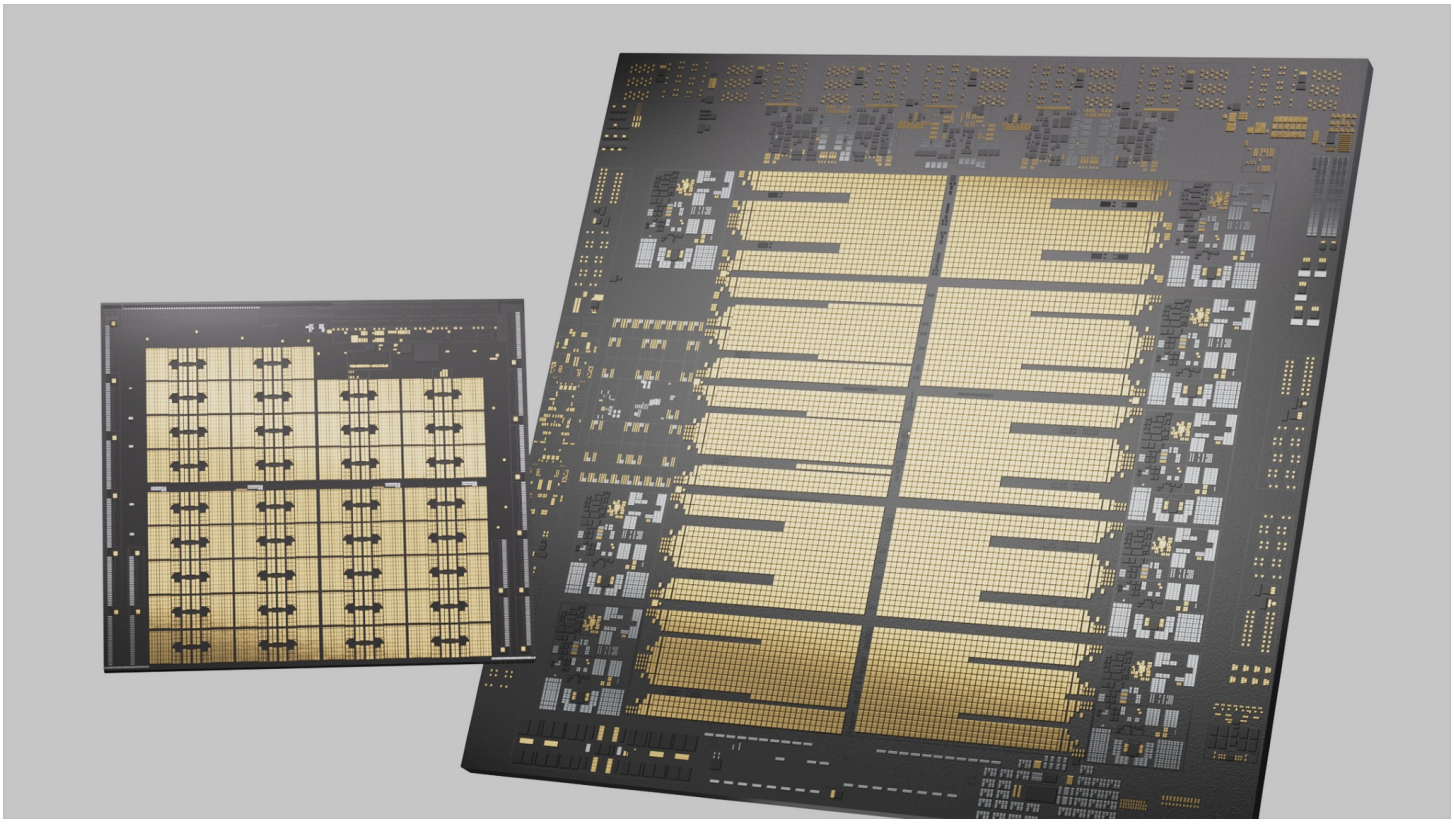
User Experience Reimagined via new Assistants and Agents

IBM's newest mainframe system, IBM z17, fully engineered for AI.

ARMONK, NY, April. 08, 2025 – IBM (NYSE: IBM) today announced the IBM z17, the next generation of the company's iconic mainframe, fully engineered with AI capabilities across hardware, software, and systems operations. Powered by the new IBM Telum® II processor, IBM z17 expands the system's capabilities beyond transactional AI capabilities to enable new workloads.

IBM Z is built to redefine AI at scale, positioning enterprises to score 100% of their transactions in real-time.[1] z17 enables businesses to drive innovation and do more, including the ability to process 50 percent more AI inference operations per day than z16.[2] The new IBM z17 is built to drive business value across industries with a wide range of more than 250 AI use cases, such as mitigating loan risk, managing chatbot services, supporting medical image analysis or impeding retail crime, among others.





IBM z17 is the culmination of five years of design and development which included the filing of more than 300 patent applications filed with the US Patent and Trademark Office. Designed with the direct input of more than 100 clients and in close collaboration with IBM Research and Software teams; the new system introduces multi-model AI capabilities, new security features to protect data, and tools that leverage AI for improving system usability and management:

- **Bringing AI to Data** – z17 AI inferencing capabilities are powered by a second-generation on-chip AI accelerator built into the IBM Telum® II processor, featuring increased frequency, compute capacity, a 40 percent growth in cache, enabling more than 450 billion inferencing operations in a day and one millisecond response time.[2]
- **Expanding Acceleration for AI** – The IBM Spyre™ Accelerator expected to be available 4Q 2025 via PCIe card, will provide additional AI compute capabilities to complement the Telum® II processor. Together, they will create optimized environments to support multi-model methods of AI. The Spyre™ Accelerator is specially engineered to bring generative AI capabilities to the mainframe including running assistants, leveraging enterprise data contained in the system.
- **Leveraging AI to Enhance User Experience** – z17 is designed to bolster the skills and efficiency of developers and IT operations with the use of AI assistants and AI agents, including IBM watsonx Code Assistant for Z and IBM watsonx Assistant for Z. In addition, for the first time, watsonx Assistant for Z will be integrated with Z Operations Unite to provide AI chat-based incident detection and resolution using live systems data.

“The industry is quickly learning that AI will only be as valuable as the infrastructure it runs on,” said Ross Mauri, general manager of IBM Z and LinuxONE, IBM. “With z17, we’re bringing AI to the core of the enterprise with the software, processing power, and storage to make AI operational quickly. Additionally, organizations can put their vast, untapped stores of enterprise data to work with AI in a secured, cost-effective way.”

Your browser does not support the video tag.

Fully Integrated Across Hardware and Software

IBM z17 is a system designed from the ground up to fully integrate into hybrid environments by tightly joining hardware innovations, software capabilities for AI, and rich support for open-standards and tooling. This enables differentiated performance and reliability while reimagining how developers and systems operators engage with and manage IBM Z, including:

- **Operating System for AI** – IBM also previewed z/OS 3.2, the next version of its flagship operating system for IBM Z, planned to be released in the third quarter of 2025. z/OS 3.2 is designed to support hardware-accelerated AI capabilities across the system and operational AI insights for system management capabilities. Additionally, z/OS 3.2 will provide support for modern data access methods, NoSQL databases, and hybrid cloud data processing. These new capabilities will help AI software tap into a broader set of enterprise data and derive predictive business insights.
- **Unified IT Operations** – Also announced today was IBM Z Operations Unite, which brings together key performance metrics and logs from multiple sources across IBM Z, in OpenTelemetry format, to streamline IBM Z operations with AI. The new solution is designed to accelerate the time to detect anomalies, isolate the impact of potential incidents, and reduce the resolution time. Used in conjunction with IBM Concert, operations teams can benefit from intelligent correlation of operational data across the entire enterprise. IBM Operations Unite will be generally available in May, 2025.
- **AI Accelerator for Business Efficiency** – With the expansion options for the IBM Spyre™ Accelerator expected to be available 4Q 2025 via PCIe card, IBM z17 aims to transform the user experience on the platform. Clients will be able to run IBM’s growing catalog of assistants and agents, based on IBM’s Granite models, natively on z17 without taking on the added risk associated with moving data or sensitive business logic off platform. Together, these solutions are engineered to form an optimized stack, to enable clients to drive more productivity with security and scale.

Built for Resiliency: Security and Cyber Defense at the Core

IBM z17 furthers the platform's history of strong security and resiliency capabilities. New developments in AI have enabled the

deployment of added intelligence across this ever-growing area of importance for clients as new threats appear every day. This includes several new capabilities, including:

- **Secrets Management** – Capabilities from HashiCorp, an IBM Company, announced in March are now available on IBM Z to help standardize secrets management across hybrid cloud. IBM Vault uses identity-based security to authenticate and authorize access to secrets, certificates, keys, tokens and other sensitive data. With the addition of IBM Vault, clients can have a single solution to help protect critical workloads by managing the entire secrets lifecycle across their full IT estate.
- **AI-powered Data Security** – IBM intends to deliver new capabilities for discovering and classifying sensitive data on the platform. This would tap into Telum® II and utilize natural language processing so mission-critical data can be identified and protected. Additionally, our latest AI-driven security solution, IBM Threat Detection for z/OS, is designed to detect and identify potentially malicious anomalies that might be the result of a cyber-attack.

IBM Extends AI-Enabled Support to IBM z17

IBM's tailored, comprehensive support experience helps IBM Z clients meet demands beyond traditional maintenance. Delivered by IBM Technology Lifecycle Services, IBM Support for z17 helps clients optimize their environments for peak performance to address risk and disruptions for mission-critical operations. IBM's AI processes streamline incident remediation and help improve case resolution time, built on IBM watsonx, now support IBM Z systems.

IBM Delivers Secured and Agile Storage

IBM Storage DS8000 plays a key role as an integrated storage solution for IBM Z. The latest generation of IBM Storage DS8000 (10^{th} Generation) is designed to harness the full power of IBM z17, providing organizations access to critical workloads, consistent and optimized data performance, and a modular architecture to adopt the latest IBM research-backed technologies to fuel business growth while monetizing data. Together, IBM Z and IBM Storage offer a modern infrastructure delivering a secured and agile platform for mission-critical workloads.

Availability

IBM z17 will be generally available June 18, 2025 For more information, visitIBM.com/z17. The IBM Spyre™ Accelerator is expected to be available starting in Q4 2025.

Statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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 the competitive edge in their industries. Thousands of government 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.

Additional Sources

- Ecosystem & Skills blog
- Software blog
- z17 Support blog
- Research blog

Media Contact:

Chase Skinner
IBM Communications
chase.skinner@ibm.com

Aishwerya Paul
IBM Communications
aish.paul@ibm.com

1. Claim – Since the introduction of IBM z16 in 2022, IBM Z mainframes have been able to support AI inferencing directly in the mainframe, making it possible to score 100% of real-time transactions even in high-volume production environments.

Source - Celent report: "Mitigating Fraud in The AI Age" by Neil Katkov, 04/08/2025, commissioned by IBM

2. Claim - The percentage difference between IBM z17, that processes up to 450 billion inference operations per day with 1 ms response time using a Credit Card Fraud Detection Deep Learning model, and IBM z16, process up to 300 billion inference requests per day with 1ms response time using a Credit Card Fraud Detection model. For IBM z17 up to 450 billion inferences operations per day.

Disclaimer - For z17 performance result is extrapolated from IBM® internal tests running on IBM Systems Hardware of machine type 9175. The benchmark was executed with 1 thread performing local inference operations using a LSTM based synthetic Credit Card Fraud Detection model (https://github.com/IBM/ai-on-z-fraud-detection) to exploit the integrated Accelerator for AI. A batch size of 160 was used. IBM Systems Hardware configuration: 1 LPAR running Red Hat® Enterprise Linux® 9.4 with 6 IFLs (SMT), 128 GB memory. 1 LPAR with 2 CPs, 4 zIIPs and 256 GB memory running IBM z/OS® 3.1 with IBM z/OS Container Extensions (zCX) feature. Results may vary. For IBM z16, performance result is extrapolated from IBM internal tests running local inference operations in an IBM z16 LPAR with 48 IFLs and 128 GB memory on Ubuntu 20.04 (SMT mode) using a synthetic credit card fraud detection model (https://github.com/IBM/ai-on-z-fraud-detection) exploiting the IBM Integrated Accelerator for AI. The benchmark was running with 8 parallel threads each pinned to the first core of a different chip. The lscpu command was used to identify the core-chip topology. A batch size of 128 inference operations was used. Results were also reproduced using a z/OS V2R4 LPAR with 24 CPs and 256GB memory on IBM z16. The same credit card fraud detection model was used. The benchmark was executed with a single thread performing inference operations. A batch size of 128 inference operations was used. Results may vary.

https://canada.newsroom.ibm.com/2025-04-08-IBM-z17-The-First-Mainframe-Fully-Engineered-for-the-AI-Age