Search

It's Here: IBM Introduces z17 Mainframe with Integrated AI Capabilities

Apr 10, 2025

IBM , Mainframe , News , z/OS

Sonja Soderlund

Sonja Soderlund is an Oregon-based B2B freelance writer. Whether writing about mainframe computers, educational technology, or sustainable retail, she strives to bring clarity to complex issues. Connect with her or LinkedIn.

On April 8, 2025, after much public anticipation, IBM announced the launch of its next-generation mainframe, **the IBM z17**, designed with integrated artificial intelligence(AI) capabilities across hardware, software, and system operations. The new system builds on the company's longstanding presence in enterprise computing and reflects a broader shift toward embedding AI directly into core infrastructure.

At the center of the z17 is the **IBM Telum® II processor**, a custom chip featuring a second-generation on-chip AI accelerator. According to IBM, the processor enables the system to handle over 450 billion AI inference operations per day, delivering sub-millisecond response times. The company reports a **50% increase in AI inference** throughput compared to its previous generation, the z16.

This upgrade expands the platform's capabilities beyond traditional transactional AI, enabling support for more complex workloads, such as generative AI and agent-based AI models. IBM cites over **250 use cases across industries**, including fraud detection, customer service automation, medical imaging analysis, and retail loss prevention.

"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. Only 1% of the enterprise data is part of large language models (LLM). With z17, businesses can put their untapped enterprise data to work with AI in a secured, cost-effective way."

Built for AI Workloads

The z17 platform has been in development for five years and includes input from more than 100 clients. It introduces multi-model AI support (predictive AI with LLMs), new security features, and tools designed to improve usability and system management through AI.

One of the key additions is the upcoming **IBM Spyre™ Accelerator**, which will work alongside the **Telum® II processor** to provide additional compute power for AI tasks. Set for release in late 2025, the Spyre™ Accelerator is an optional add-on intended to support generative AI applications and IBM's growing suite of enterprise-focused AI assistants.

The system also incorporates AI tools, including IBM's **watsonx Code Assistant for Z** and **watsonx Assistant for Z**, designed to streamline coding and incident response. These tools will integrate with **Z Operations Unite**, enhancing the user experience for developers and IT operations by allowing for chat-based system issue detection and resolution

using real-time system data.

In a media briefing, **Tina Tarquinio**, Chief Product Officer, IBM Z and LinuxONE, underlined the energy efficiency of the new platform: "The z17 will offer **7.5x AI throughput compared to z16**, but we're going to do it with less energy than it would take to do a multi-model use case on another platform using a different type of accelerator.

"Given all the discussion around energy requirements and the difficulties in procuring it, I'm thrilled that we're able to offer our clients this capability right in their system that they already have running their mission critical applications without giving up anything they require in terms of scalability and availability."

Operating System and Integration Updates

IBM Spyre™ accelerator chip connecting to the PCIe card

Alongside the new hardware, IBM previewed **z/OS 3.2**, the upcoming version of its mainframe operating system. Scheduled for release in Q3 2025, z/OS 3.2 will include support for hardware-accelerated AI, modern data access methods, and hybrid cloud data integration, making it easier to incorporate enterprise data into AI workflows.

The company also introduced **Z Operations Unite**, a monitoring and operations tool that uses AI to analyze system data, identify anomalies, and reduce response times. According to IBM, the system will accelerate the time to detect anomalies, **reduce alert investigation by up to 9X**, isolate the impact of potential incidents, and help reduce resolution time.

Emphasis on Security and Resilience

Security continues to be a core focus for IBM. The z17 introduces new tools for **secrets management** through **IBM Vault**, a system based on technology from IBM subsidiary HashiCorp. It allows identity-based authentication and lifecycle management of sensitive data across hybrid environments. With the addition of Vault, clients can have a single solution for protecting critical workloads by securely managing the entire "secrets lifecycle" across their full IT estate.

IBM is also developing AI-based tools to **automatically identify and classify sensitive enterprise data**, using natural language processing and other AI techniques to help protect high-value information before it enters AI workflows.

"We heard from client after client that they wanted to be able to run Gen AI use cases on-prem, in a secure environment, where they had control over their data, as well as the model IP if they happened to fine-tune these models for enterprise-specific tasks," said **Elpida Tzortzatos**, IBM Fellow and CTO of z/OS and AI on IBM Z.

"So we wanted to make sure that our clients had the security as well as the AI governance in place to be able to manage the whole end-to-end AI life cycle and be able to capture model metadata throughout this life cycle that they can leverage for AI compliance."
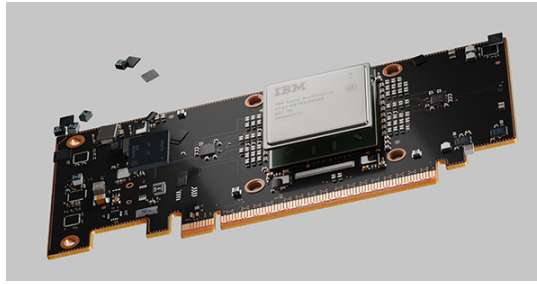
Storage and Availability

The z17 pairs with the **10th generation of IBM Storage DS8000**, which has been updated to support the data access demands of AI-heavy workloads. The combination is designed to deliver consistent performance, modular scalability, and secure data management for enterprise applications.

The IBM z17 will be generally available on June 8, 2025, with the Spyre™ Accelerator expected in the fourth quarter of 2025.

For more information, see the IBM **full news release**.

IBM z17 Blogs:

- **Transforming and simplifying the mainframe for greater productivity and efficiency on IBM z17**
- **Unleash the power of IBM z17 with comprehensive IT support services**
- **IBM z17 Ecosystem Blog**
- **How IBM Research helped prepare the IBM z17 for tomorrow's workloads**

0 Comments

Sign up to receive the latest mainframe information

First Name

Last Name

Email *(Required)*

Subscribe

2025 z/OS Performance Education

Scott Chapman's Essential z/OS Performance Tuning
   September 22 - 26, 2025

Peter Enrico's Parallel Sysplex & z/OS Performance Tuning
   October 21 - 22, 2025

Learn more now!

For more virtual education, visit www.pivotor.com!

Recently Published

Announcements at SHARE, Job-Ready Talent training, and more

by Sonja Soderlund

Open source on z/OS – Don't let breaking changes break you

by Ed Jaffe

Leveraging Data Insights To Strengthen Business Operations

by Ainsley Lawrence

A Decade in Purple Chucks: Celebrating Planet Mainframe
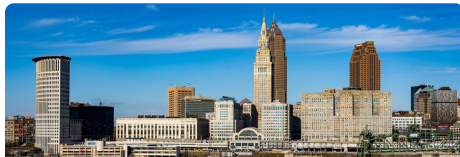
by Penney Berryman

Read More

Open source on z/OS – Don't let breaking changes break you
by Ed Jaffe

A Decade in Purple Chucks: Celebrating Planet Mainframe
by Penney Berryman

Join Planet Mainframe at SHARE Cleveland for Trivia, Prizes, and More
by Amanda Hendley

Cheryl Watson's Tuning Letter Delivers z/OS Tuning, Hardware Picks, and Performance Tips
by Mark Wilson

Latest

Become a Contributor

Tuning Letter

Events