IBM z17

Enable hybrid cloud with advanced AI where it matters most

Highlights

Leverage AI-powered innovation to fuel business growth

Automate and transform for operational efficiency

Secure your critical data with the most resilient system

The AI+ era is enabling new opportunities, competitive advantage and growth; however, it can place significant demands on your IT infrastructure. With IBM® z17™, you can take advantage of AI to fuel growth today while keeping vital data and applications secure and still deliver the highest qualities of service. You can enable operational efficiencies, extracting the most value from your IT investments and talent, and accelerate digital transformation.

IBM z17 provides a vital foundation for your hybrid cloud and enables innovation such as quantum-safe cryptography and AI-powered security to reduce risk and multimodel AI for more precision and accuracy. It is engineered from the chip through the stack to optimize mission-critical transaction processing and data-intensive workloads. IBM z17 makes more possible, enabling AI where it matters most to drive efficiency, innovation and better business outcomes.

↑450 billion

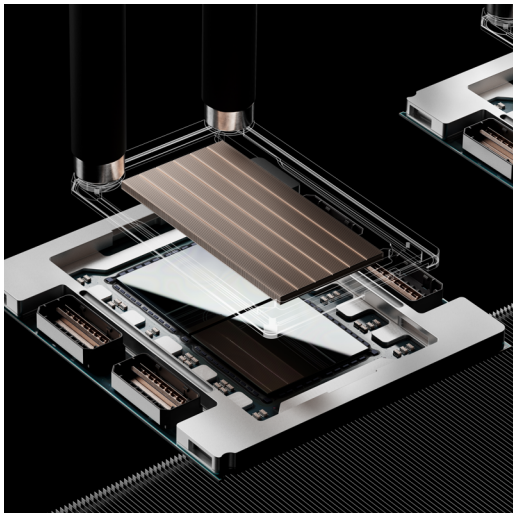
inferences per day with
a 1 ms response time[1]

↑99.999999%
availability

equivalent to 315 ms
of downtime per year[2]

Leverage AI-powered innovation to fuel business growth
IBM z17 can make more precision possible with industry-leading AI inferencing. The on-chip and PCIe-attached AI acceleration, provided by the IBM Telum® II processor and the IBM Spyre™ Accelerator card, provides enhanced high-speed inferencing with low latency and extends generative AI (gen AI) and large language models (LLMs) at enterprise speed and scale. Now you can unlock more value in transactional AI, deploying multiple models to improve accuracy and reduce false positives, such as enhanced fraud detection, anti-money laundering or anomaly detection. You can harness the power of increasingly larger, more complex models for in-transaction workloads to gain better outcomes while safeguarding data and the intellectual property of your models. Use cases requiring secure, compliant and protected AI operations—such as specialized financial applications, sensitive document summarization and information search and data extraction—can rely on IBM Z®, the most trusted platform, to safeguard their data and model intellectual property.

Automate and transform for operational efficiency
IBM z17, combined with IBM Spyre Accelerator, will make more productivity possible with AI assistants and AI agents. These systems are fundamentally changing the way users of all types are experiencing the mainframe, driving increased productivity and efficiency while easing concerns around skills. Mainframe practitioners will have access to AI assistants and agentic workflows running on premises—on IBM z17, with IBM Spyre Accelerator. These gen AI solutions will guide actions with increased understanding and confidence, saving initial learning time for new practitioners. Application developers can increase productivity across the application development lifecycle by leveraging gen AI and chat to more easily understand, refactor, optimize, create, test and deploy applications with increased speed and agility. Advanced AI insights from gen AI assistants and agents for IT operations can help developers quickly detect incidents and identify root cause, minimizing business impact by reducing time to resolution from hours to minutes.



Secure your critical data with the most resilient system
IBM z17 is designed to make more cyber resilience possible, with industry-leading eight nines availability[2]. IBM z17 helps you improve security management by simplifying implementation and leveraging AI to deliver better insights. AI-powered capabilities detect and mitigate potential internal and external threats, rapidly identifying data access anomalies that can reduce business risk, while helping meet emerging anti-malware regulatory mandates. AI is also used to automatically identify the most sensitive data so it can be protected, minimizing security risks. Additionally, IBM z17 enables progress on your post-quantum encryption journey by incorporating NIST-standardized algorithms and providing tools to automate cryptographic inventory creation.

Key new innovations in IBM z17 include

IBM Telum II Processor

AI Acceleration on chip coprocessor
The IBM Telum II processor integrates improved AI acceleration through an on-chip AI coprocessor to reduce latency and deliver outstanding performance for in-transaction inferencing. It now supports small language models (SLMs) where the number of variables is less than 8 billion. Organizations can embed AI directly into business processes and existing IBM Z applications to help improve business outcomes and deliver customer value in each interaction at unprecedented scale and speed within stringent SLA response times.

I/O acceleration unit
A completely new data processing unit (DPU) on the Telum II processor chip is engineered to accelerate complex input/output (I/O) protocols for networking and storage on the mainframe. The DPU simplifies system operations and can improve key component performance.

IBM Spyre Accelerator
When available, the IBM Spyre Accelerator card will provide additional AI compute capability to complement the Telum II processor. This component will extend and scale the IBM z17 AI capabilities by providing compute to support gen AI use cases needing unstructured data such as text. Up to 48 Spyre Accelerators will be supported to scale gen AI for enterprise workloads that demand exceptional performance and rigorous security and resilience.

AI-powered security
IBM z17 represents a step forward in AI-powered security. With Sensitive Data Tagging for IBM z/OS®, which leverages AI with natural language processing (NLP) to distinguish between sensitive and non-sensitive data, you now have a robust yet simple to implement solution for data identification. Another AI-powered security feature, IBM Threat Detection for z/OS, runs routine scans and leverages AI to identify potential threats, giving you the opportunity for early detection and mitigation to contain damages associated with cyberattacks. Both of these features can assist in meeting cybersecurity regulations.

Configurations Table

IBM z17 Overview	**IBM z17 ME1** This configuration is designed for general-purpose use, offering a balance of performance, scalability and security suitable for a wide range of applications.
Specifications	
Maximum number of engines	208
Maximum number of drawers	4
Maximum number of IO drawers	12
Number of frames	4
Colocate with storage/switch	No
Frequency	5.5 GHz
Telum chip	Yes
Maximum memory	64 TB
Sizes	43, 90, 136, 183 and 208
Resources	
Specification sheets	Multiframe specifications
Energy efficiency	Multiframe carbon footprint
Technical guides	Multiframe guide
Interactive tour	Multiframe tour

With AI at its core, the new IBM z17 supports multimodel AI, at scale, while continuing to provide the highest levels of performance, resilience and security for mission-critical workloads. IBM z17 and AI technology will fundamentally transform transaction processing and data to unlock additional productivity and efficiency for your business. IBM Z can also be an integral part of your hybrid cloud. With IBM Z integrated into your hybrid cloud, your workloads will benefit from a seamless experience across infrastructure platforms while retaining the security, resilience and AI-driven insights of IBM Z systems. IBM Z also offers a variety of software designed to optimize hybrid cloud, with the security, resilience, AI and application modernization you need.

To learn more about IBM z17, contact your IBM Business Partner:

Arbor Solutions, Inc.

6169011714 | Rafton@arbsol.com

arbsol.com

1. With IBM z17, you can process up to 450 billion inference operations per day with a 1-millisecond response time using a Credit Card Fraud Detection Deep Learning model.
   DISCLAIMER: This performance result is extrapolated from IBM internal tests running on IBM system hardware machine type 9175. The benchmark was executed with one thread performing local inference operations using an LSTM-based synthetic Credit Card Fraud Detection model to exploit the Integrated Accelerator for AI on a batch size of 160. IBM system 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.
2. ITIC 2023 Global Server Hardware, Server OS Reliability Report, August/September 2023.

© Copyright IBM Corporation 2025

Produced in the
United States of America
April 2025

© Copyright IBM Corporation 2025. IBM, the IBM logo, IBM Spyre, IBM Telum, IBM Z, IBM z/OS, and IBM z17 are trademarks or registered trademarks of International Business Machines Corporation, in the United States and/or other countries. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on ibm.com/legal/copyright-trademark

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Red Hat® is a registered trademark of Red Hat, Inc. or its subsidiaries in the United States and other countries.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

Examples presented as illustrative only. Actual results will vary based on client configurations and conditions and, therefore, generally expected results cannot be provided.