Linux on IBM Z

Modernize for hybrid cloud, AI, security, and efficiency

Linux at its best

Using open-source Linux® solutions is a smart way to run your IT.

Linux on IBM Z®, available for 25 years now, provides an impressive Linux and hybrid cloud environment, especially for workloads that require high levels of security, flexibility, and resilience.

Linux on IBM Z benefits from the exceptional strengths and capabilities of IBM Z technology, including AI acceleration, quantum-safe cryptography, and confidential computing, as well as extreme scalability, high utilization, and unmatched resiliency and availability.

Co-locating workloads on IBM Z, where Linux workloads run side-by-side with workloads on IBM z/OS®, IBM z/TPF, 21CS VSEn, or container platforms like Red Hat® OpenShift®, not only delivers exceptional performance and operational efficiency but also maximizes the value of your existing investments in existing assets.

Linux on IBM Z offers a robust environment for modernizing your infrastructure, enabling hybrid cloud, AI, security, and efficiency, and providing a solid foundation for innovation and growth.

Linux on IBM z17™

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.

Highlights

- Data protection and privacy
- Industry-leading AI inferencing
- Efficiency, scalability and flexibility
- Quality of services with low cost

“It was critical that we invested in and moved to a technology that not only allowed us to optimize our core banking but that will help us move forward on our digital transformation and pivot to cloud.”

IBM z17 provides a strong Linux and hybrid cloud environment, as well as other IBM Z servers:

- Some basic facts: IBM z17 scales up to 208 cores running at 5.5 GHz and offers up to 64 TB of RAIM.[1]
- IBM z17 has an I/O Data Processing Unit (DPU) which simplifies system operations and can improve key performance.
- The IBM Telum® II processor reduces latency and delivers outstanding performance for in-transaction inferencing through its on-chip AI coprocessor.
- The IBM Integrated Accelerator for AI on IBM z17 at full utilization is designed to process up to 24 trillion operations per second shared across all cores on the chip.[2]
- IBM z17 enables progress on post-quantum encryption.
- The Crypto Express8S adapter is designed to meet FIPS 140-3 certification at Level 4.
- The IFL processors on the IBM z17 also provide an optional IBM z17 multi-threading technology capability; with the multi-threading function enabled, the performance capacity of a IFL is expected to typically be up to 25% higher than without the multi-threading function enabled.[3]
- On IBM z17 running Red Hat Enterprise Linux with KVM, deploy up to 3,000,000 NGINX containers.[4]
- IBM Hyper Protect Virtual Servers[5] are designed to protect Linux workloads with sensitive data from both internal and external threats.
- IBM Z Security and Compliance Center[6] helps to simplify the compliance workflow and auditing to help organizations meet security standards such as PCI DSS and NIST SP800-53.

Linux on IBM Z provides an on-premises infrastructure foundation for application modernization and hybrid cloud.

IBM Z and the accompanying Linux software provide a broad set of open and industry-standard tools, including container and Kubernetes technologies, as well as agile DevOps methodologies, to accelerate modernization.

As an example, simplified infrastructure-as-a-service (IaaS) management and the integration to cloud automation tools is provided with IBM Cloud Infrastructure Center[7]. It delivers an industry-standard user experience for the IaaS management, enables the automation of infrastructure services such as the deployment of a database-as-a-service, and enables the integration of IBM Z into an enterprise hybrid cloud model.

Data protection and privacy at scale and confidential computing

Quantum-safe cryptography is embedded in the latest IBM Z systems, alongside classical cryptography, to safeguard Linux-based applications and data against both current and future quantum-computing threats, ensuring long-term protection and security.

Pervasive encryption is enabled with Linux on IBM Z. It is transparent to existing applications and by leveraging the advanced on-processor cryptography and Crypto Express accelerators, encryption and decryption processes are optimized for improved usability and efficiency.

The IBM RACF® Security Server for IBM z/VM® provides a comprehensive security system that meets regulatory and auditing requirements, offering features such as access control, auditing, resource authorization, privileged command access, and logon controls.

IBM Secure Execution for Linux is a hardware-based security technology designed to provide a confidential computing environment to protect sensitive data running in virtual servers and container runtimes by performing computation in a trusted execution environment (TEE). It enables workloads to run in full isolation with protection from both internal and external threats across a hybrid cloud, ensuring the integrity of each application and its data. The IBM Hyper Protect Virtual Servers[5] are based on this technology, thus providing a confidential computing environment.

To speed compliance and audit readiness, the IBM Z Security and Compliance Center6 helps to take the complexity out of your compliance workflow and the ambiguity out of audits through automated fact collection and mapping to help you comply with security standards PCI DSS, NIST SP800-53, and CIS – ready to strive for regulatory compliance of the Linux on IBM Z ecosystems.

It is important to mention that IBM Z is the world's only server with the high level EAL5+ hardware security certification. It guarantees that the IBM Z principal security features are reliably applied, allowing for isolation and protection of the deployed workloads, while the isolation capability inside the server offers significant operational simplicity.

Industry-leading AI inferencing

IBM z17 integrates improved AI acceleration via an on-chip AI coprocessor to reduce latency and deliver outstanding performance for in transaction inferencing. It now supports Small Language Models where the number of parameters is less than 8 billion. The IBM Telum II Integrated Accelerator for AI is designed to process up to 24 trillion operations per second shared by all cores on the chip.

Some examples that demonstrate the capacity

– Using a single Integrated Accelerator for AI on an OLTP workload on IBM z17 matches the throughput of running inferencing on a compared remote x86 server with 13 cores.[8]

– With IBM z17, process up to 450 billion inference operations per day using multiple AI models for credit card fraud detection.[9]

Efficiency with high levels of scalability, flexibility, and utilization

The high workload density of IBM Z, supporting up to thousands of virtual Linux servers, typically results in fewer components, reduced management effort, and lower software licensing costs compared to other platforms, making it a more efficient and cost-effective solution.

Impressive scalability—horizontal and vertical—is provided with the IBM Z capabilities in combination with the virtualization technologies[10]. Resources can be prioritized dynamically and efficiently between workloads, delivering them whenever and wherever they are needed.

– IBM z/VM[11] virtualization technology offers deep integration with IBM Z, allowing for high levels of resource sharing, data-in-memory techniques, outstanding I/O bandwidth, availability, and security.

– KVM[12] virtualization enables the use of Linux administration skills on IBM Z. KVM is delivered with the Linux distributions for IBM Z and is optimized to benefit from the IBM Z capabilities.

The high flexibility of IBM Z is not only shown in the hardware capabilities, the available solution portfolio for Linux on IBM Z, products and frameworks, from IBM, independent software vendors, and open source is immense.

IBM Dynamic Partition Manager provides a simplified configuration for Linux servers, allowing for a quick and easy adoption of Linux.

Simplified infrastructure management is provided with IBM Cloud Infrastructure Center3 for compute, network, and storage resources for virtual machines based on z/VM and Red Hat KVM running Linux and Red Hat OpenShift. Key use cases of Cloud Infrastructure Center are the deployment of a database-as-a-service, simplified experience with virtualization on IBM Z via the vendor-agnostic technology for simplified IaaS management, and the infrastructure-as-a-service management for service providers supporting to provide tenant-safe services.

Co-location on IBM Z can make a big difference in efficiency. Businesses and IT organizations must provide fast access to data, and when these multi-tiered workloads have communication patterns that are network intensive, meaning they either frequently communicate or exchange many messages to complete a single transaction, or they exchange large amounts of data, then the physical location and proximity of the tiers can make a difference.

The IBM Z technologies IBM HiperSockets™, Shared Memory Communication (SMC-D), and zdsfs enable to communicate efficiently. On IBM Z you can co-locate workloads not only to support reductions in latency and improvements in throughput, operational efficiency, and security, but also leverages investments in existing assets.

IBM Z technologies are engineered for high efficiency, with features like compression acceleration on the processor chip, which enables data compression to reduce storage space and increase data transfer rates, benefiting not only Linux applications and data but also overall system performance.

IBM Z servers also offer the flexibility to scale up by adding system resources, allowing Linux on IBM Z to grow 'on-demand' without disrupting existing business operations, providing a seamless and efficient way to adapt to changing needs.

IBM Z systems are designed with energy efficiency in mind, aligning with best practices that minimize electricity consumption, such as consolidating workloads onto a smaller number of physical systems, reducing the overall environmental footprint.

Deploying workloads on a centralized infrastructure like IBM Z can help reduce greenhouse gas emissions and promote a more environmentally sustainable IT environment.

Superior quality of service with low cost of computing

IBM Z helps to avoid or recover from failures to minimize business disruptions, realized through component reliability, redundancy and features that assist in providing fault avoidance and tolerance, as well as permitting concurrent maintenance and repair.

IBM Z systems are engineered to deliver exceptional performance for applications that require uncompromising transaction processing and data sharing. With massive scalability, IBM Z systems can dynamically add capacity on demand, grow processing power with minimal impact on energy consumption, floor space, and staffing requirements.

IBM Z systems are architected for balanced performance with multiple layers of cache, massive I/O capabilities, and integrated accelerators to drive high utilization and processor efficiency.

Further strengthening resilience of the Linux and hybrid cloud workloads, are solutions such as:

– Live Guest Relocation, enabled with the z/VM SSI13 feature, allowing for the non-disruptive move of running virtual Linux servers from one member of a cluster to another.

– IBM GDPS® can provide multi-platform resiliency for Linux servers. It allows for disaster and failure recovery and ensures data consistency across multiple sites. When running GDPS with z/OS, you can benefit from a single point of control for the z/OS and Linux environments.

– IBM Storage Scale (former IBM Spectrum® Scale) is designed to provide high availability through advanced clustering technologies, dynamic file system management and data replication.

– Unlike with distributed systems or public clouds—resilience, availability, and failover capabilities can be expected for Linux on IBM Z.

IBM Z systems support operational efficiency by allowing multiple virtual Linux servers to run on a single system, leveraging its immense capacity and scalability, both horizontally and vertically. This typically results in reduced maintenance and administration efforts compared to other platforms and can lead to cost savings in various areas.

IBM Z systems also support operational efficiency through features such as dynamic resource addition, sharing, and reconfiguration, as well as the ability to run Linux alongside other operating systems, leveraging unique system arrangements to optimize resource utilization and streamline operations.

Red Hat Ansible® Automation Platform has emerged as a powerful solution on IBM Z, used to automate a wide range of IT tasks, from configuration management and application deployment to enforcing security and compliance.

Considering all the aspects mentioned above – reduced carbon footprint, privacy and protection, high levels of scalability, flexibility, and utilization, co-location benefits, superior quality of service – it seems obvious that they can also provide an economic advantage when running Linux on IBM Z compared to other platforms.

Workloads that fit well

Linux workloads with per-core pricing are ideal candidates for deployment on IBM Z from a financial perspective. Due to differences in server architectures and processor speeds, Linux workloads on IBM Z often require fewer processor cores compared to distributed servers, resulting in potential cost savings.

Workloads with high I/O demands, such as databases, messaging, and stream processing, can significantly benefit from IBM Z's advanced I/O capabilities, including FICON® and Fibre Channel Protocol (FCP), which are designed to accelerate data transfer and optimize CPU utilization, resulting in faster response times and improved overall performance.

Workloads with high availability requirements can take advantage of IBM Z's inherent redundancy and resiliency features, ensuring minimal downtime and maximum uptime. Additionally, Capacity Backup (CBU) for IBM Z enables hardware engines to be utilized for disaster recovery without incurring extra software costs, providing a cost-effective solution for maintaining business continuity.

Workloads that demand low latency and high transaction rates can significantly benefit from the co-location advantages of IBM Z, which offers ultra-low latency and high throughput, since the workloads must not constantly access another system over the network.

Workloads with high security requirements, particularly those that handle sensitive data, are often deployed on IBM Z to minimize the risk of a security breach. IBM Z offers unique security benefits that help reduce the risk of data or privacy breaches, providing a secure environment for sensitive workloads.

IBM Z is designed to provide a highly secure, resilient, and efficient Linux and cloud environment, with containers and Kubernetes to build and modernize cloud services, with potential competitive advantages in operational efficiency and business economics, extreme scalability, high resource sharing and utilization, encryption enablement, data privacy and server isolation, continuous operations, and cyber resiliency.

Why IBM?

As you transform your business and differentiate yourself in a trust economy, IBM remains your partner.

We have the total expertise in systems, software, delivery, and financing to help you create a secure and intelligent foundation for the future.

Our experts can help you configure, design, and implement Linux on IBM Z, optimized for your needs.

For more information

To learn more about Linux on IBM Z, please contact your IBM representative, your Red Hat representative, or IBM Business Partner.

1. Redundant Array of Independent Memory; the RAIM design detects and recovers from failures of dynamic random access memory (DRAM), sockets, memory channels, or DIMMs
2. Result is the maximum theoretical number of trillion operations per second (TOPS) in 8bit precision that can be executed by a single IBM Integrated Accelerator for AI. Cores are running at 5.5GHz and have one IBM Integrated Accelerator for AI per chip. The IBM Integrated Accelerator for AI consists of 2 corelets, each with an array of 64 tensor cores capable of executing 4 integer-multiply-add operations (IMA) 8-way SIMD with no sparsity.
3. Based on internal measurements. Results may vary by customer based on individual workload, configuration and software levels. Visit LSPR website for more details at: http://www.ibm.com/support/pages/ibm-z-large-systems-performance-reference
4. Performance result is extrapolated from IBM® internal tests running on IBM Systems Hardware of machine type 9175. 1 LPAR with 12 dedicated IFLs (SMT) and 1.2 TB memory running Red Hat® Enterprise Linux® (RHEL) 9.5 with KVM. 24 RHEL 9.5 virtual machines with 1 vCPU and 64 GB memory. On each virtual machine, 7813 NGINX® 1.26.2 containers were deployed. Results may vary.
5. For more information refer to: https://www.ibm.com/products/hyper-protect-virtual-servers
6. For more information refer to: https://www.ibm.com/products/z-security-and-compliance-center
7. For more information refer to: https://www.ibm.com/products/cloud-infrastructure-center
8. Performance results are based on IBM® internal tests running on IBM Systems Hardware of machine type 9175. The OLTP application (https://github.com/IBM/megacard-standalone) and PostgreSQL was deployed on the IBM Systems Hardware. The Credit Card Fraud Detection (CCFD) ensemble AI setup consists of two models (LSTM: https://github.com/IBM/ai-on-z-fraud-detection, TabFormer: https://github.com/IBM/TabFormer). On IBM Systems Hardware, running the OLTP application with IBM Z Deep Learning Compiler (zDLC) compiled jar and IBM Z Accelerated for NVIDIA® Triton™ Inference Server locally and processing the AI inference operations on IFLs and the Integrated Accelerator for AI versus running the OLTP application locally and processing remote AI inference operations on a x86 server running NVIDIA Triton Inference Server with OpenVINO™ runtime backend on CPU (with AMX). Each scenario was driven from Apache JMeter™ 5.6.3 with 64 parallel users. IBM Systems Hardware configuration: 1 LPAR running Ubuntu 24.04 with 7 dedicated IFLs (SMT), 256 GB memory, and IBM FlashSystem® 9500 storage. The Network adapters were dedicated for NETH on Linux. x86 server configuration: 1 x86 server running Ubuntu 24.04 with 28 Emerald Rapids Intel® Xeon® Gold CPUs @ 2.20 GHz with Hyper-Threading turned on, 1 TB memory, local SSDs, UEFI with maximum performance profile enabled, CPU P-State Control and C-States disabled. Results may vary.
9. Performance result is extrapolated from IBM® internal tests running on IBM Systems Hardware of machine type 9175. The benchmark was executed with 64 threads performing local inference operations using a synthetic credit card fraud detection (CCFD) model based on an LSTM (https://github.com/IBM/ai-on-z-fraud-detection) and a TabFormer (https://github.com/IBM/TabFormer) model. The benchmark exploited the Integrated Accelerator for AI using IBM Z Deep Learning Compiler (zDLC) and IBM Z Accelerated for PyTorch. The setup consists of 64 threads pinned in groups of 8 to each chip (1 for zDLC, 7 for PyTorch). The TabFormer (tabular transformer) model evaluated 0.035% of the inference requests. A batch size of 160 was used for the LSTM based model. IBM Systems Hardware configuration: 1 LPAR running Ubuntu 24.04 with 45 IFLs (SMT), 128 GB memory. Results may vary.
10. Beside z/VM, KVM, and Dynamic Partition Manager, Red Hat OpenShift Virtualization is available to manage containers and virtual machines, providing a unified management experience with the Red Hat OpenShift tooling; for more information refer to: https://www.ibm.com/docs/en/rhocp-ibm-z
11. For more information refer to: https://www.ibm.com/products/zvm
12. For more information refer to: https://www.ibm.com/products/kvm
13. z/VM SSI = z/VM Single System Image, for more information: https://www.vm.ibm.com/ssi/

Learn more:

Linux on IBM Z
IBM Z

© Copyright IBM Corporation 2025
IBM Corporation
New Orchard Road
Armonk, NY 10504

IBM, the IBM logo, ibm.com, IBM Cloud, IBM Z, FICON, GDPS, HiperSockets, Spectrum, Telum, z15, z17, z/OS and z/VM are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml.
Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
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 world¬wide basis.

Red Hat®, JBoss®, OpenShift®, Fedora®, Hibernate®, Ansible®, CloudForms®, RHCA®, RHCE®, RHCSA®, Ceph®, and Gluster® are trademarks or registered trademarks 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 client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions. 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.
The client is responsible for ensuring compliance with laws and regulations applicable to it. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the client is in compliance with any law or regulation.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.