Linux® on IBM z17™

Modernize for AI and hybrid cloud with the powerful and secure Linux platform for business

Consider the advantages

- Secure
- Scalable and flexible
- Efficient co-location
- AI acceleration
- Resilient and available
- Economical

Secure with a cyber resilient system

Security capabilities
Data encryption
Security integrated across the stack and lifecycle
Secured isolation
Clear Key, Secure Key, Protected Key & Public Key Infrastructure (PKI)
Network Security
Confidential Computing for data-in-use, data-at-rest, and data-in-flight
Compliance with automation, monitoring and reporting

- IBM Z® continues and extends its leadership in Quantum-safe cryptography, which is embedded in the system to improve the resiliency to cyber-attacks from bad actors with future access to quantum computing resources.
- IBM z17 hardware accelerated encryption with Central Processor Assist for Cryptographic Functions (CPACF) is designed to provide fast encryption without any application changes.
- Crypto Express 8S with quantum-safe APIs based on NIST PQC standards to modernize existing and build new applications.
- 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).

- IBM Hyper Protect Virtual Servers[1] designed to protect mission-critical Linux workloads with sensitive data from both internal and external threats. They take advantage of IBM Secure Execution for Linux.
- IBM Z Security and Compliance Center[2] helps 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 and NIST SP800-53.
- IBM z17 is designed for Evaluation Assurance Level (EAL) 5+ hardware security certification.

AI acceleration

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[3] for AI on IBM z17 is designed to process up to 24 trillion operations per second shared by all cores on the chip.

- With IBM z17, process up to 5 million inference operations per second with less than 1 ms response time using a Credit Card Fraud Detection Deep Learning model.[4]
- 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.[5]

The IBM Spyre™ AI Accelerator will extend and scale the IBM z17 AI capabilities by providing AI compute power to support LLMs and generative AI use cases for workloads that demand exceptional performance.

The IBM z17 supports Generative AI where the data resides and is best protected.

IBM z17 Model ME1 Telum® II processor – state-of-the-art technology

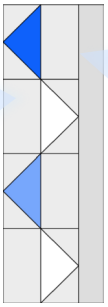
• 5nm technology • 5.5 GHz • up to 208 Cores • up to 64TB memory • 1 to 4 19-inch frames	• Improved on-chip AI Acceleration with 24 trillion operations per second per chip, 192 per drawer, and 768 per system • Up to 48 IBM Spyre AI Accelerator Adapters • On-chip Data Processing Unit implements complex I/O protocols and reduces latency • On-chip cryptography acceleration • On-chip compression acceleration	Redesigned cache subsystem: • 36 MB Level-2 - 3.6ns • 340 MB virtual Level-3 - 11.5ns • 2.8 GB virtual Level-4 - 48.5ns

- Reduced power for I/O management by 70%
- 24.1 Miles of wire per chip, 43B transistors
- ↑20% socket performance, ↓18% processor power consumption

Resiliency and availability

Not available on x86 servers

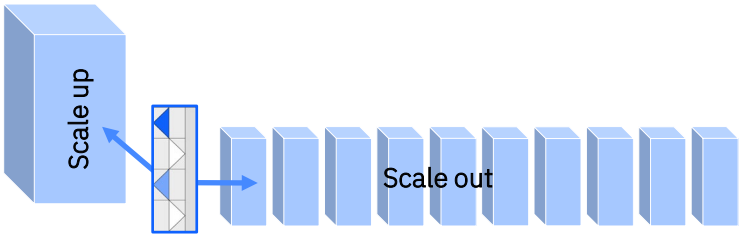
- Spare cores for transparent failover
- RAIM subsystem to eliminate memory failures
- Built-in bottom to top security



- Achieve scale and availability with elastic capacity
- Live Guest Relocation via IBM z/VM® SSI and KVM
- IBM GDPS® for near-continuous availability
- IBM Storage Scale for advance clustering

Error Prevention	• Hardware and firmware designed to protect against outages • Built-in redundancy eliminates single points of failure • Extensive testing and failure analysis at every level
Error Detection and Correction	• Error detection embedded in components • Built-in automated diagnostics; problem determination and isolation • Non-disruptive installation, upgrades and maintenance avoids outages
Error Recovery	• Automated failover to speed recovery and to minimize system impact • Business continuity and disaster recovery solutions – IBM GDPS, Live Guest Relocation via z/VM SSI and KVM, IBM Storage Scale, HiperDispatch, Call Home, etc.

Scalability options and flexible growth



Scale **horizontally** and **vertically without disruption**

- Outstanding scalability, horizontal and vertical, based on the immense total IBM z17 capacity. On IBM z17 running Red Hat® Enterprise Linux with KVM, deploy up to 3,000,000 NGINX containers.[6]
- Provision for peak utilization: dynamically add processors (cores), memory, I/O adapters, devices and network cards; unused resources automatically are reallocated after peak.
- On chip compression acceleration helps to reduce in the size of data to save storage space, also increase data transfer rates, with reduced CPU consumption.
- Simplified infrastructure-as-a-service (IaaS) management for both non-containerized and containerized workloads, a self-service portal, 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, provides lifecycle management for virtual machines that are based on IBM z/VM and Red Hat KVM, enables the automation of infrastructure services, and enables the integration of IBM Z into an enterprise hybrid cloud model.
- Goal-oriented approach for performance management of a hypervisor, and Live virtual server migration capabilities provided with z/VM Single System Image (SSI) feature and KVM.
- The redesigned I/O subsystem using the Data Processing Unit (DPU) on the Telum II chip to improve performance and I/O density with support for a 4 port FICON® adapter.

Co-location efficiency

Co-location can make a big difference. IBM z17 enables the co-location, running workloads on Linux, z/OS®, IBM z/TPF, 21CS VSEn, and Red Hat OpenShift®, helping on resource efficiency, data access with low latency, eliminating network handling, and helping on centralized system administration.

Co-location can make a significant performance difference since businesses and IT organizations need to provide fast access to data and applications, can contribute to security, and can be helpful when using disaster recovery and compliance solutions.

High performance and efficiency

- Optimized for data serving, quick response times and less application waits through optimized cache structure and large cache sizes
- High I/O bandwidth

Cross-memory data and local network transfer advantages

- High throughput and low latency by less hops
- Less network equipment (routes, switches) – network inside the server

Adding IFLs to an **IBM z17** means **low incremental costs** and a more efficient infrastructure.

Centralized management of co-located workloads

- Optimized resource utilization based on high levels of resource sharing
- Same arrangements for security, process monitoring, backup and disaster recovery, etc.

Modernize for hybrid cloud

Modernizing applications can unlock resources for enhancements and new applications that drive business growth. By adopting cloud-native and microservice architectures, your organization can achieve greater efficiency, flexibility, scalability, and security enabling to innovate faster, adapt more easily, and stay ahead.

Infrastructure-as-a-service (IaaS), infrastructure as code (IaC), container technologies, and DevSecOps approaches are key components and can help to speed innovation and improve agility, giving a competitive advantage.

Simplified IaaS management for all workloads is provided with IBM Cloud Infrastructure Center[7], as well as IaC, provided via the integration capabilities with management tools, such as Red Hat Ansible® or Terraform, based on OpenStack-compatible APIs.

Container technology and DevSecOps are available with Red Hat OpenShift, a security-focused hybrid cloud platform that empowers to develop, deploy, and manage applications across the IT environment. It is a trusted Kubernetes enterprise platform with support for cloud-native and containerized workloads. Red Hat OpenShift[8] provides a comprehensive set of optimized tools to secure, protect, and manage applications, streamlining the development and administration process. This toolset is ideal for building, deploying, and maintaining applications, and boosting efficiency. It also integrates with Red Hat OpenShift Virtualization to manage containers and virtual machines, providing a unified management experience.

These capabilities deliver speed to market and agility for both development and operational teams as IBM z17 integrates as a critical component of hybrid cloud.

Economic advantages

IBM z17 provides a scalable, highly available platform that delivers differentiated value to help enable business growth, reduce cost, and protect existing investments.

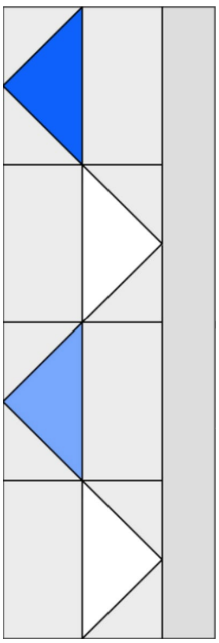
Cost advantages can be achieved in

- Operational management
- Security and business continuity
- Software acquisition and licenses
- Flexibility of configuration
- Floor space and energy
- Maintenance effort

Linux on IBM z17 Linux offers a robust solution that delivers scalability, agility, resiliency, high performance, and a secure environment, all while providing a lower Total Cost of Ownership.

IBM z17 can monitor the power consumption for the partitions in which the Linux images run.

Linux on IBM z17 can provide confidence in meeting the future, in a world of uncertainty.



Performance	
• Cores, memory	High
• Multi-tier cache	High
• I/O bandwith	High
Reliability	
• Spare cores	High
• Reliable memory	High
• Concurrent upgrades	High
Scalability	High
• Secure partitions	
• Capacity on demand	High
• Multiple workloads	High
Security	High

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. *For more information refer to: https://www.ibm.com/products/hyper-protect-virtual-servers*
2. *For more information refer to: https://www.ibm.com/products/z-security-and-compliance-center*
3. *For more information refer to: https://www.ibm.com/new/announcements/telum-ii*
4. *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 (CCFD) model (https://github.com/IBM/ai-on-z-fraud-detection) to exploit the IBM 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.*
5. *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.
6. *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.*
7. *For more information refer to: ibm.com/products/cloud-infrastructure-center*
8. *For more information refer to: https://www.ibm.com/docs/en/rhocp-ibm-z*

Learn more:
Linux on IBM Z
IBM z17

ZSP03194-USEN-21