Linux on IBM z16

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

Consider the advantages

- Secure
- Scalable and flexible
- Efficient colocation
- Sustainable
- 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)
Auditing
Network Security
IBM Z® Security and Compliance Center

- IBM z16™ quantum-safe technology and key management services, were developed to help you protect data and keys against a potential future quantum attack like harvest now, decrypt later[1]
- The new Crypto Express8S is designed to meet the Federal Information Processing Standards (FIPS) 140-3 at Level 4 for cryptographic modules
- IBM z16 hardware accelerated encryption with Central Processor Assist for Cryptographic Functions (CPACF) is designed to provide fast encryption without any application changes
- IBM Secure Execution for Linux® is a hardware-based security technology designed to provide scalable isolation for individual workloads to help protect them from external attacks and insider threats
- Linux allows for 'protected key' encryption for data at-rest, and 'clear key' encryption for data in-flight/at-rest. With IBM z16 multi frame, scale up your I/O intensive Linux applications and protect your data at rest with up to 12 million read-only I/O operations per second and 10 million R/W operations per second to an encrypted filesystem with FCP attached storage[2]
- IBM Z clients surveyed anticipate that by using the IBM Z Security and Compliance Center on IBM z16, they can potentially reduce audit preparation timelines by 55%[3]
- IBM z/VM® 7.3 and 7.2 provide the prerequisite IBM z16 encryption support to enable exploitation by Linux guests
- IBM z16 is designed for Evaluation Assurance Level (EAL) 5+ hardware security certification

Environmental sustainability

IBM z16 is designed to make a powerful improvement in sustainability by decreasing electricity consumption, reducing the number of standing servers, and enabling high compute and resource utilization.

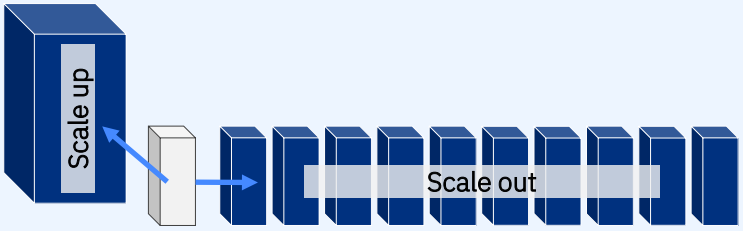
IBM z16 design is aligned with best practices for reducing electricity consumption, including that clients require a small number of physical systems. IBM z16 is high energy-efficiency system that enables high compute and resource utilization.

Consolidating Linux workloads on 5 IBM z16 multi frame systems instead of running them on compared x86 servers under similar conditions can reduce energy consumption by 75%, space by 50%, and the CO2e footprint by over 850 metric tons annually. This is equivalent to consuming about 362,000 fewer litres (95,600 gallons) of gasoline each year.[4]

An IBM z16 single frame or rack mount Max68 running Linux workloads can reduce the energy required by approximately 75% each year versus a compared configuration of 36 x86 servers with 1,440 cores running the same Linux workloads with similar conditions and location.[5]

Running workloads on a centralized infrastructure such as IBM Z can contribute to fewer green-house gas emissions and a more environmentally sustainable IT environment.

Multi-dimensional growth and scalability options



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

- Provision for peak utilization: dynamically add processors (cores), memory, I/O adapters, devices and network cards; unused resources automatically are reallocated after peak
- With IBM z16, execute up to 20 billion HTTPS transactions per day with OLTP microservice applications running on Red Hat® OpenShift® Container Platform[6]
- Outstanding scalability, horizontal and vertical, based on the immense total IBM z16 capacity
- On chip compression acceleration helps to reduce in the size of data to save storage space and also increase data transfer rates, with reduced CPU consumption
- Run the Yahoo Cloud Serving Benchmark (YCSB) on MongoDB without sharding on IBM z16 with 6 IFLs in total and achieve the same throughput as on MongoDB with 4 shards on compared x86 systems with 144 cores in total, which provides a 24:1 core consolidation ratio in favor of IBM z16[7]
- Run the HammerDB benchmark on EnterpriseDB Advanced Server 14 with on average 1.7x more throughput and 38% less response time on up to 24 cores on an IBM z16 LPAR versus on compared x86 system[8]
- Live virtual server migration capabilities provided with z/VM Single System Image (SSI) feature and KVM
- Goal-oriented approach for performance management of a hypervisor
- z/VM offers high levels of resource sharing, data-in-memory techniques and outstanding I/O
- Temporary activation of resources through ‘On/Off Capacity on Demand’ at a per-day per-core charge, permanent activation through ‘Capacity on Demand’

IBM z16 Model A01 Telum™ processor – state-of-the-art semiconductor technology

– 7nm technology – 5.2 GHz – up to 200 Cores – up to 40TB memory – 1 to 4 19-inch frames	– Centralized chip-based AI accelerator – On-chip cryptography acceleration – On-chip compression acceleration – Single Instruction Multiple Data (SIMD) – Out-of-order execution – Hardware transactional memory (HTM) – Separate cores for I/O processing	Redesigned cache subsystem: – semi-private 32 MB Level-2 caches per core – up to 256 MB virtual Level-3 cache per chip – up to 2 GB virtual Level-4 cache per drawer

- Using one Integrated Accelerator for AI on an OLTP workload on IBM z16 multi frame matches the throughput of a compared remote x86 server running inferencing on 18 cores[9]
- Hyperledger Digital Wallet workloads on IBM z16 can benefit from the security of IBM Secure Execution for Linux and achieve up to 1.5x more throughput and 31% lower latency versus on compared x86 platform using KVM[10]
- The IFL processors on the IBM z16 A02 also provide an optional multi-threading technology capability; with the multi-threading function enabled, the performance capacity of an IFL is expected to typically be up to 25% higher (range typically 10..40%) than without the multi-threading function enabled[11]

Resiliency and availability

Not available on x86 servers

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

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

Enterprise-grade security and resiliency features are architected throughout the IBM z16 and Linux on IBM z16 stack. The IBM Z Security and Compliance Center helps centralize monitoring of compliance related tacks and provides an interactive view of these tasks and potential severity of control deviations. It is designed to help with automated collection and validation of relevant facts from key areas of Linux.

Co-location efficiency

Co-location can make a big difference. IBM z16 enables the co-location, running workloads on Linux, IBM z/OS®, IBM z/TPF, 21CS VSE[n12] and Red Hat OpenShift, helping on resource efficiency, data access with low latency, eliminating network handling, and helping on centralized system administration.

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 due to dedicated I/O processors and memory buffer cache
- On digital currency transactions run inferencing for fraud 85% faster by co-locating your application with Snap ML on IBM z16 versus running inferencing remotely using Scikit-learn on a compared x86 server[13]

Cross-memory data and local network transfer advantages

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

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

IBM z16 provides the foundation for application modernization and hybrid cloud velocity by delivering leading hybrid cloud infrastructure to support the optimization of existing mission-critical applications and data.

IBM z16, Red Hat OpenShift Container Platform, and the accompanying IBM Cloud Paks® and cloud software, which is developed to support a cloud-native experience, delivers a broad set of open and industry-standard tools, including container and Kubernetes technologies, and an agile DevOps methodology to accelerate modernization.

When accessing your database while running an OLTP workload on OpenShift Container Platform, achieve 4.2x more throughput by co-locating the workload on IBM z16 multi frame versus running the workload on compared x86 platform connecting remotely to the IBM z16 multi frame.[14]

Simplified infrastructure-as-a-service (IaaS) management for all workloads, a self-service portal, and the integration to cloud automation tools is provided with IBM Cloud Infrastructure Center[15].

These capabilities deliver speed to market and agility for both development and operational teams as IBM z16 integrates as a critical component of hybrid cloud.

Economic advantages

IBM z16 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

Furthermore, the costs of using public cloud services are often underestimated and unpredictable, while the cost running cloud services on IBM z16 are almost always predictable.

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

Adding IFLs to an **IBM z16** means **low incremental costs**, and a more efficient and sustainable infrastructure.

For more information

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

Learn more:
Linux on IBM z16
IBM z16

1. *IBM z16 with the Crypto Express 8S card provides hardware enabled quantum-safe APIs. The quantum-safe public key technology used in IBM z16 has been summitted to the PQC standardization process conducted by NIST. https://csrc.nist.gov/Projects/post-quantum-cryptography/round-3-submissions system protected by quantum-safe technology across multiple layers of firmware.*
2. *Performance result is extrapolated from IBM internal tests running the fio 3.19 benchmark tool in an IBM z16 LPAR with 12 IFLs and 64 GB memory on RHEL 8.5 (SMT mode) using the XFS filesystem format with luks2 encryption and two FICON Express 32S features. The fio benchmarking tool was run with 128 parallel threads using 8 volumes on FS9200 equally distributed over the two nodes and file size of 150GB on each volume. Results may vary.*
3. *IBM does not ensure regulatory compliance. The intent is to provide a point in time statement of your current posture for a specific group of resources. The responsibility of ensuring systems are configured in accordance with regulatory controls is on the individual businesses who are using the IBM Z security and compliance Center and IBM does not take responsibility for any compliance oversights or penalties associated with data breaches. The survey consisted of 9 responses across 6 unique customers. Sourced from the IBM ZSCC Sponsor User Program and zDC.*
4. *Compared 5 IBM z16 Max 125 model consists of three CPC drawers containing 125 configurable cores (CPs, zIIPs, or IFLs) and two I/O drawers to support both network and external storage versus 192 x86 systems with a total of 10364 cores. IBM z16 power consumption was based on inputs to the IBM z16 IBM Power Estimation Tool for a memo configuration. x86 power consumption was based on March 2022 IDC QPI power values for 7 Cascade Lake and 5 Ice Lake server models, with 32 to 112 cores per server. All compared x86 servers were 2 or 4 socket servers. IBM Z and x86 are running 24x7x365 with production and non-production workloads. Savings assumes a Power Usage Effectiveness (PUE) ratio of 1.57 to calculate additional power for data center cooling. PUE is based on Uptime Institute 2021 Global Data Center Survey (https://uptimeinstitute.com/about-ui/press-releases/uptime-institute-11th-annual-global-data-center-survey). CO2e and other equivalencies that are based on the EPA GHG calculator (https://www.epa.gov/energy/greenhousegas-equivalencies-calculator) use U.S. National weighted averages. Results may vary based on client-specific usage and location.*
5. *Compared IBM Machine Type 3932 Max 68 model consisting of a CPC drawer and an I/O drawer to support network and external storage with 68 IFLs and 7 TB of memory in 1 frame versus compared 36 x86 servers (2 Skylake Xeon Gold Chips, 40 Cores) with a total of 1440 cores. IBM Machine Type 3932 Max 68 model power consumption was measured on systems and confirmed using the IBM Power estimator for the IBM Machine Type 3932 Max 68 model configuration. x86 power values were based on Feb. 2023 IDC QPI power values and reduced to 55% based on measurements of x86 servers by IBM and observed values in the field. The x86 server compared to uses approximately .6083 KWhr, 55% of IDC QPI system watts value. Savings assumes the Worldwide Data Center Power Utilization Effectiveness (PUE) factor of 1.55 to calculate the additional power needed for cooling. PUE is based on Uptime Institute 2022 Global Data Center Survey (https://uptimeinstitute.com/resources/research-and-reports/uptime-institute-global-data-center-survey-results-2022). Results may vary based on client-specific usage and location.*
6. *Performance result is extrapolated from IBM internal tests running in an IBM z16 LPAR with 24 dedicated IFLs, 560 GB memory and DASD storage the Acme Air microservice benchmark (https://github.com/blueperf/acmeair-mainservice-java) on Red Hat OpenShift Container Platform (RHOCP) 4.9 using RHEL 8.4 KVM. On 4 RHOCP Compute nodes 4 Acme Air instances were running in parallel, each driven remotely from JMeter 5.2.1 with 384 parallel users. The KVM guests with RHOCP Compute nodes were configured with 12 vCPUs and 64 GB memory each. The KVM guests with RHOCP Management nodes and RHOCP Infrastructure nodes were configured with 4 vCPUs and 16 GB memory each. Results may vary.*
7. *Performance results based on IBM internal tests running YCSB 0.10.0 benchmark (read-mostly) on MongoDB Enterprise Release 5.0.6 with 3-node replication. On IBM z16 MongoDB was setup without sharding but with two replicas. IBM z16 configuration: LPAR with 4 dedicated cores and 2 LPARs with each 1 core, each with SMT and 128 GB memory, 1 TB IBM FlashSystem® 900. x86 config: 9 Intel® Xeon® Gold 5218 CPU @ 2.30GHz with Hyperthreading turned on, 192 GB memory, 1 TB local RAID5 SSD storage, RHEL 8.4 running MongoDB, driven remotely by YCSB using 2 x86 server with total 128 threads. Results may vary.*
8. *Performance results based on IBM internal tests running the HammerDB 4.5 TPC-C (https://github.com/TPCCouncil/HammerDB/releases/download/v4.5/HammerDB-4.5-RHEL8.tar.gz) benchmark on EnterpriseDB Advanced Server 14.4.0 on IBM z16 vs running it on a compared x86 server. The workload was driven by HammerDB from a remote x86 server using RoCE Express2 with 25 GbE connection. IBM z16 configuration: RHEL 8.7 in an LPAR with 6 - 24 dedicated IFLs, 128 GB memory, and IBM FS9200 series storage. x86 configuration: RHEL 8.7 on 2x 24 Ice Lake Intel® Xeon® Gold CPU @ 2.80GHz with Hyperthreading turned on, 1 TB memory, and IBM FS9200 series storage. Data in-flight (TLS v1.3: TLS_AES_256_GCM_SHA384 with 256-bit keys) and data at-rest encryption (AES-XTS with 256-bit keys) was used for both platforms. Results may vary.*
9. *Performance results is extrapolated from IBM internal tests running an OTLP workload with credit card transaction using the Credit Card Fraud Detection (https://github.com/IBM/ai-on-z-fraud-detection) model on IBM z16 vs running the OLTP workload https://github.com/IBM/megacard-standalone) on IBM z16 and running inferencing on a remote x86 server running Tensorflow serving. IBM z16 configuration: Ubuntu 20.04 in an LPAR with 6 dedicated IFLs, 256 GB memory, and IBM FlashSystem 900 storage. x86 configuration: Ubuntu 20.04 on 18 IceLake Intel® Xeon® Gold CPU @ 2.80GHz with Hyperthreading turned on, 1 TB memory, local SSDs. Results may vary.*
10. *Performance results based on IBM internal tests running Hyperledger Digital Wallets (https://github.com/xixuejia/digital-wallet) on IBM z16 versus on compared x86 platform using KVM. On both platforms, each Hyperledger Digital Wallet instance was running in a separate KVM guest with 2 vCPUs and 8 GB memory. On IBM z16, IBM Secure Execution for Linux was used to secure the KVM guest. The Hyperledger Fabric v2.4.1 test network was running on a single separate IBM z15® LPAR with 16 dedicated IFLs. The network consisted of 2 organizations, 1 peer per organization and 1 orderer, accessed using TLS encryption. Results may vary. IBM z16 configuration: LPAR with 2 dedicated IFLs, 128 GB memory, running Ubuntu 20.04, and running 2 KVM guests with IBM Secure Execution for Linux and filesystem encryption. X86 configuration: 2 IceLake Intel® Xeon® Gold CPU @ 2.80GHz w/ Hyperthreading turned on, 1.5 TB memory, running Ubuntu 20.04, with 2 KVM guests without filesystem encryption and without Intel® Software Guard Extension.*
11. *Based on internal measurements. Results may vary by customer based on individual workload, configuration and software levels. Visit LSPR website for more details at: https://www-40.ibm.com/servers/resourcelink/lib03060.nsf/pages/lsprindex*
12. *IBM z/VSE has achieved end of life and has been replaced by 21CS VSEn. Support for 21CS VSEn is provided by 21CS. For more information, see 21CS website.*
13. *Performance results based on IBM internal tests doing inferencing using a Scikit-learn Random Forest model with Snap ML v1.9.0 (tech preview) backend on IBM z16 and with Scikit-learn v1.0.2 backend on compared x86 server. The model was trained on the following public dataset https://www.kaggle.com/datasets/ellipticco/elliptic-data-set. BentoML v0.13.1 (https://github.com/bentoml/BentoML) was used on both platforms as model serving framework. IBM z16 configuration: Ubuntu 20.04 in an LPAR with 2 dedicated IFLs, 256 GB memory. x86 configuration: Ubuntu 20.04 on 9 IceLake Intel® Xeon® Gold CPU @ 2.80GHz with Hyperthreading turned on, 1 TB memory.*
14. *Performance result is based on IBM internal tests running on an IBM z16 A01 LPAR with 24 dedicated IFLs, 560 GB memory and DASD storage an OLTP workload on Red Hat OpenShift Container Platform (RHOCP) 4.10 using RHEL 8.4 KVM. On 4 RHOCP Compute nodes 4 OLTP workload instances were running in parallel, each driven remotely from JMeter 5.2.1 with 128 parallel users. The KVM guests with RHOCP Compute nodes for the OLTP workload were configured with 12 vCPUs and 64 GB memory each. The KVM guests with RHOCP Management nodes and RHOCP Infrastructure nodes were configured with 4 vCPUs and 16 GB memory each. Compliance Operator deployed a configured from (https://www.ibm.com/docs/en/zscc/1.1.1). Results may vary.*
15. *For more information refer to: ibm.com/products/cloud-infrastructure-center*

ZSP03194-USEN-20