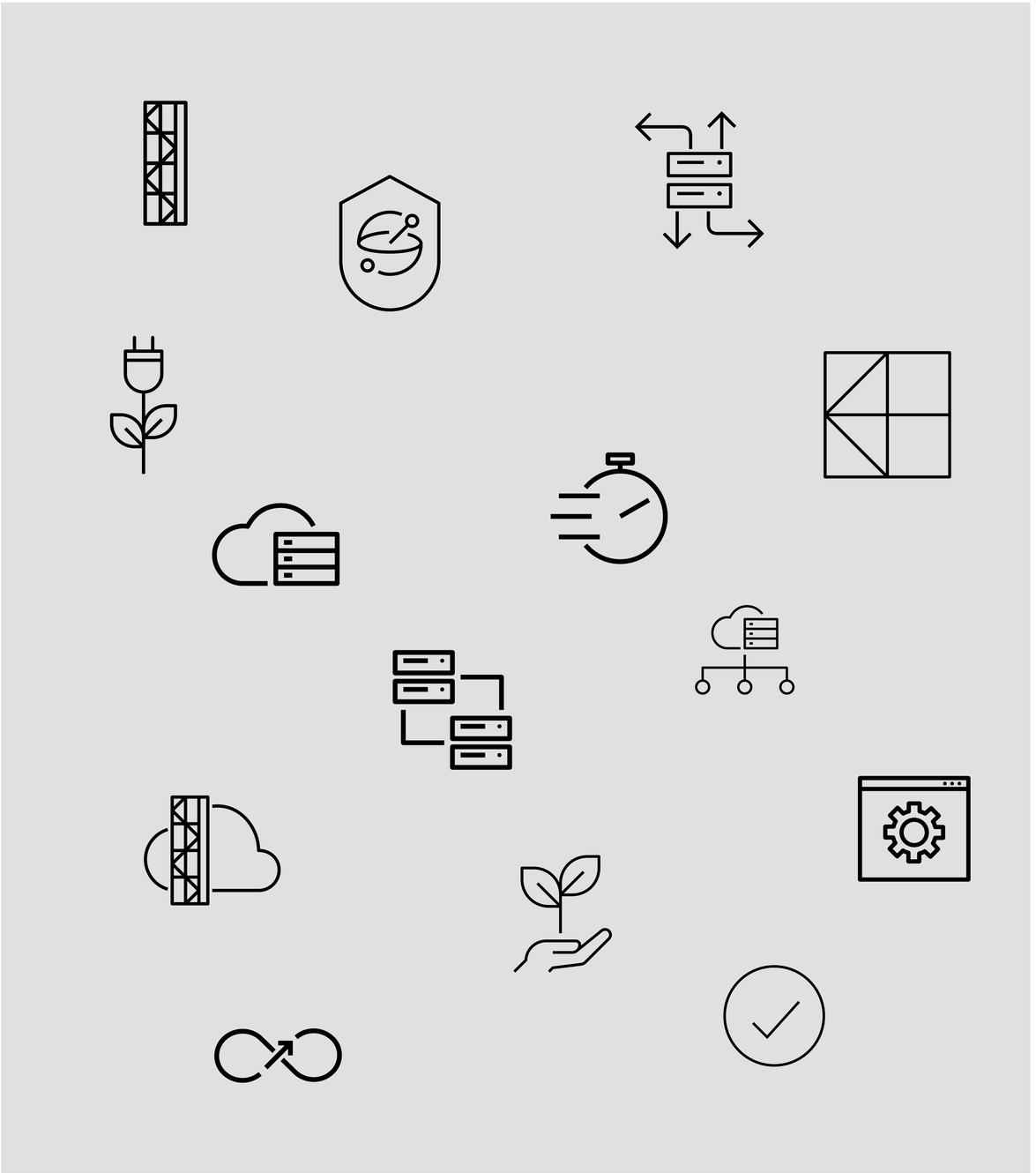
IFL use cases

on IBM Z and IBM LinuxONE



IFL use cases

Consolidation Virtualization Multi-tenant isolation	Agile connectivity via API and architecture IBM z/OS® colocation	Application hosting Transactional workloads	Infrastructure management and automation
Data serving Software-Defined Storage		Cloud-native deployments Container hosting	Security & Cryptography

IFL use cases solutions

Data serving

- IBM® Db2®
- IBM Cloud Pak® for Data
- Oracle
- MongoDB
- PostgreSQL - Enterprise DB, Fujitsu Enterprise Postgres

Software-Defined Storage

- IBM Spectrum Fusion™ (Red Hat OpenShift)
- IBM Storage Scale (Hybrid & AI)

Cloud-native deployments & container hosting

- Red Hat® OpenShift®
- IBM Cloud Paks – e.g., Cloud Pak for Integration, Cloud Pak for Data
- IBM z/OS Cloud Broker
- IBM Z Security and Compliance Center
- … containerized apps

Security & Cryptography

- IBM Crypto Express adapter (HSM) support
- IBM Hyper Protect Servers
- IBM Secure Execution for Linux®
- Linux Secure Boot
- IBM Z Security and Compliance Center

Consolidation Virtualization Multi-tenant isolation

- IBM z/VM®
- KVM
  - Canonical
  - Red Hat
  - SUSE
- LPAR

Agile connectivity via API and architecture

- IBM Cloud Pak for Integration, incl. API Connect, App Connect, Aspera®, Event Streams, Event Endpoint Management, DataPower®, MQ
- IBM App Connect
- IBM MQ

IBM z/OS colocation

- Access to open source ecosystem
- Application hosting

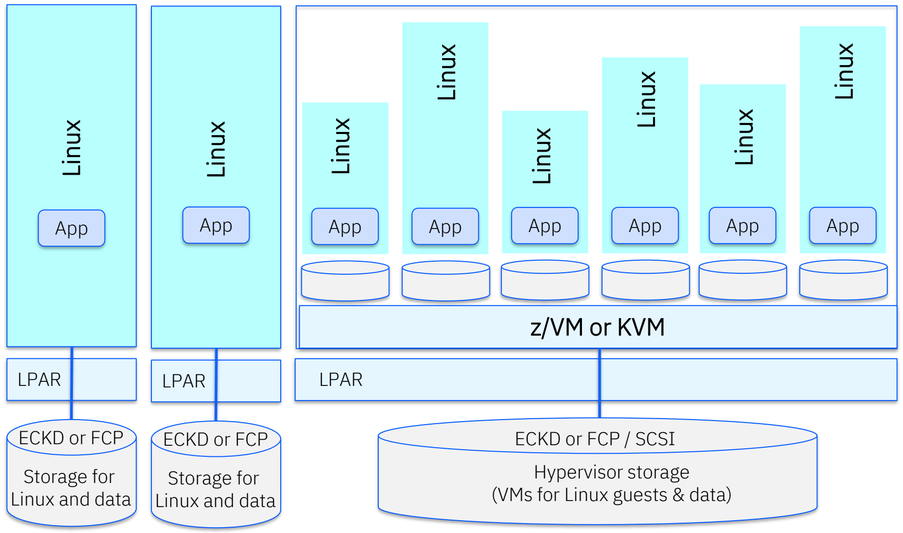
Application hosting Transactional workloads

- ISV apps, e.g., Temenos, Finacle, …, .NET
- IBM WebSphere® Liberty
- Open source apps
- AI
- IBM Db2 Analytics Accelerator

Infrastructure management and automation

- IBM Cloud Infrastructure Center (IaaS management)
- Red Hat Ansible® Automation Platform (use of Playbooks)

Consolidation /
efficient Linux deployment



Linux applications benefit from the strengths and capabilities of outstanding IBM Z / IBM® LinuxONE technology, such as extreme scalability, high resource utilization, encryption everywhere, unparalleled resiliency and availability, and energy efficiency.

IBM z16™/IBM® LinuxONE Emperor 4 have a maximum of 200 configurable IFLs running at 5.2 GHz.

An IFL can run many virtual servers, a consolidation onto an IFL can result in less IT costs, especially when the software licensing costs are priced per core.

The maintenance costs of a single IBM Z / IBM® LinuxONE server can replace the maintenance costs of many x86 servers and the related network components.

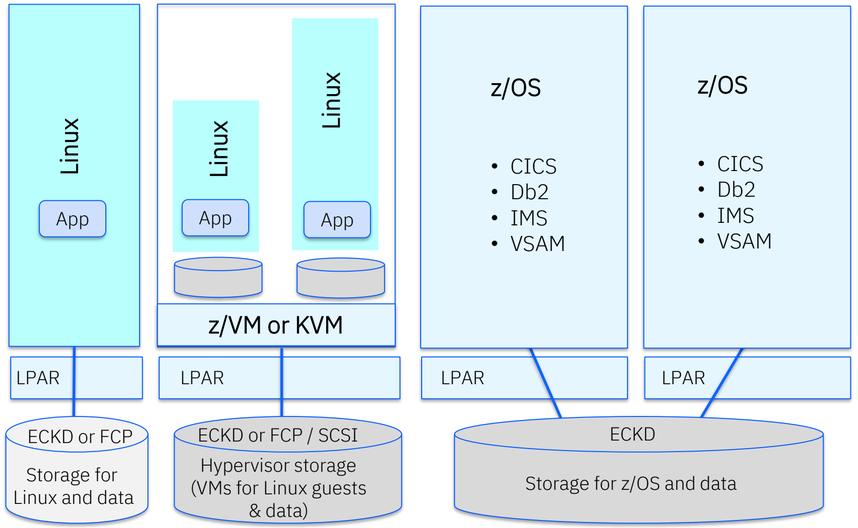
Consolidation can help on savings in physical space and electricity consumption.

Consolidating Linux workloads on 5 IBM z16 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.*

* 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.

Colocation

The term colocation is made up of the Latin prefix "co" (together) and the English "location" (place). Written: colocation or co-location



Linux applications runs side-by-side with IBM z/OS to fully leverage the application capabilities available for modernization on IBM Z / IBM® LinuxONE with low latency and high throughput, resource efficiency, centralized system administration, and energy efficiency.

Colocation can make a significant difference

The physical location or proximity of the workloads can make a difference when these multitiered workloads have communication patterns that are network intensive. This means, they either frequently communicate, exchange many messages to complete a single transaction, or exchange large amounts of data.

IBM HiperSockets, Shared Memory Communication (SMC), and zdsfs are IBM Z / IBM® LinuxONE technologies that provide benefits to customers who colocate workloads on the IBM Z / IBM® LinuxONE platforms.

The benefits of colocation on IBM Z / IBM® LinuxONE

- Reduction in latency and improvement in throughput
- Improved security and availability
- Cost savings

Security & Cryptography

- Protect and future-proof data with pervasive encryption and quantum safe technology
- Gain visibility and reduce risk with continuous compliance and intelligent security solutions
- Detect and report security events with solutions that integrate with the existing environment

IBM Crypto Express adapter (HSM) support

- Tamper-sensing and tamper-responding, high-performance cryptographic operations
- Supports specialized crypto APIs and functions, including those that are required with quantum-safe cryptography and in the banking industry
- Up to 60 Crypto Express adapters – up to thousands VMs can access a dedicated virtual HSM

CP Assist for Cryptographic Function (CPACF)

- Coprocessor that supports Pervasive Encryption by providing fast synchronous cryptographic services
- Encryption: DES, TDES, AES
- Hashing: SHA-1, SHA-2, SHA-3, SHAKE
- Random Number Generation: PRNG, DRNG, TRNG
- CPACF accelerator built into every core

IBM Hyper Protect Servers

- Provides hardware-level security for virtual servers
- Apps and data are always private and protected from internal and external threats
- Protects workloads throughout the app lifecycle – from build through to deploy and manage

IBM Secure Execution for Linux

- Protects workloads from internal and external threats
- Protects data at-rest, data in-flight, and also data in-use
- Supports boot images that are fully encrypted, and the user can measure the attestation results
- Provides trusted execution environment (TEE) and an attestation function

IBM Z Security and Compliance Center

- Simplifies compliance and helps reduce risk
- Centralized, interactive dashboard with out-of-the-box profiles built for regulatory requirements

Linux Secure Boot

- Ensures that only approved operating systems can be IPLed
- Verifies that malicious code hasn't been inserted
- Uses digital key pairs to check that SystemTap and other startup code hasn't been altered
- Reduces a Linux server security risk

Data serving

- Data volumes and formats are exploding
- On-prem is the lion's share in the DBMS market
- Open-source technology adoption is pervasive

IBM Cloud Pak for Data

- Leverage Integrated Accelerators for AI to provide machine learning acceleration
- Modernize your data infrastructure for simpler administration, faster time to value and lower operational costs
- Reduce cost and complexity by consolidating existing Db2 database applications without cloud specific rewrites
- Colocate AI/machine learning and data with business applications
- Access your most valuable enterprise data securely within modern hybrid cloud applications

IBM Db2

- Enable real-time insights
- Collaborate on data projects with governed data access and sharing
- Unify your data with a single engine that supports tight integration with your data lake over cloud object storage
- Simplify database administration

Oracle

- Scale and grow Oracle Database applications and data with confidence
- Benefit from mission-critical reliability
- Simplify IT operations with advanced virtualization
- Minimize network security vulnerabilities, and reduce latency
- Reduce cost of running Oracle Database

MongoDB

- NoSQL database
- Document model allows to avoid waisted efforts
- Rigorous resilience and governance controls
- IBM is the 2022 MongoDB Hybrid Cloud Partner of the Year
- Supporting transition to “Low Carbon” footprint

PostgreSQL

- Industry leading reliability, isolation, and integrity
- Superior performance and greater software value for a smaller, more sustainable data center footprint
- Lower database costs from 1/30th vs competitive DB
- Standardize with Open Source and PostgreSQL
- EnterpriseDB and FUJITSU Enterprise Postgres are supported

Agile connectivity
via API and architecture

IBM Cloud Pak for Integration

- Provides rich and comprehensive set of integration offerings in support of different types of data, applications, systems and services
  - API Management
  - Application integration
  - Enterprise messaging
  - Event streaming
  - Aspera: High-speed data transfer
- Integrate by building, managing and sharing APIs
- Connect to anything quickly, securely and at scale
- Unleash real-time insights to drive data-driven actions

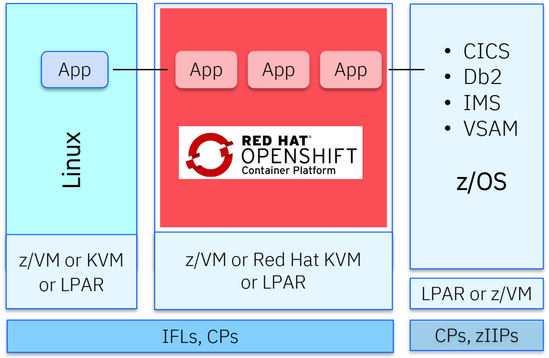
IBM App Connect

- Powers digital applications by unlocking business data and assets as APIs
- Connects business applications, integrates data, builds APIs
- Enable digital business
- Using new channels partners and monetizing data
- Speed application development
  Ongoing app evolution via reusable APIs
- Securely expose data and business applications on-premise and across clouds
- Expand brand reach
  By publishing APIs to tap into a broad developer community

IBM MQ

- Proven messaging for hybrid and multicloud environments
- Connectivity support for all kinds of applications
- Secure message delivery
  Once and only once delivery of messages with end-to-end message-level encryption
- High performance
  Message transport to deliver business-critical, high-value data with improved speed and reliability
- Deploy where needed
  On-prem, cloud, software, appliance

Cloud-native deployments & container hosting



Red Hat OpenShift

- Faster innovation and time-to-value with well-defined recommendations and best practices
- Simplification with a consistent set of APIs, defined for developers and administrators
- Adapt enterprise DevOps with consistent, cloud native development, and CI/CD tooling
- Safer deployments with automated and secured orchestration and life-cycle management
- Kubernetes device plug-in for IBM Crypto Express (CEX) cards
- Colocate containerized apps beside existing apps and data running in Linux or z/OS environments

IBM Cloud Paks

- Pre-certified containerized software and foundational services
- Build, modernize, and manage applications securely
- Simplify collection and analysis of data, apply intelligent automation, automate labor-intensive IT processes, increase operational agility

IBM Z Security and Compliance Center

- Increase productivity of compliance resources by reducing audit preparation time
- Interactive dashboard provides view of current compliance posture for PCI-DSS and NIST SP800-53 regulations
- Track compliance drift over time with dashboard style visualizations

IBM z/OS Cloud Broker

- Integrates z/OS-based services and resources into private cloud
- Empowers organizations to operate z/OS with the cloud agility and flexibility
- Leverage the strengths of the z/OS platform while enhancing applications

More containerized apps

Hosted at the IBM Z and IBM® LinuxONE Container Registry

- Images are built from scratch by IBM
- IBM controls the contents of the channel
- Images are scanned by IBM Vulnerability Manager
- Image digest hashes published to enable secure pull

Application hosting

Access to open source ecosystem

- Open source software is secure and transparent
- Open-source software can be easily integrated
- Source code is downloadable and can be changed for specific needs
- Validated Open Source Software for IBM Z/IBM® LinuxONE
  - Open source packages have been ported and/or validated on corresponding distro versions
- Linux on IBM Z/ IBM® LinuxONE HOW-TOs
  - GitHub project hosts instructions and tutorials, helping to get software on Linux on IBM Z / IBM® LinuxONE (s390x architecture)

ISV applications

- Banking ecosystem
  - Temenos, EdgeVerve Finacle, Fiorano, DXC's Hogan, Luxoft's CAMS
- Data & AI ecosystem
  - Oracle, EnterpriseDB, Fujitsu Enterprise Postgres, MongoDB, Hazelcast, Clari5
- DevOps management ecosystem
  - Gitlab Runner, F5/NGINX, Flexera, Dynatrace
- Security tools ecosystem
  - Crowdstrike, Qualys, Illumio, Cisco Secure Workload, Sysdig, Opentext/encase

IBM WebSphere Liberty

- Cloud-native development
- Simple rapid inner-loop developer experience in any IDE
- Operational resource optimization
- Reduce costs with world-leading performance for microservices and monoliths
- Application modernization
- API & configuration compatibility for reduced effort and risk

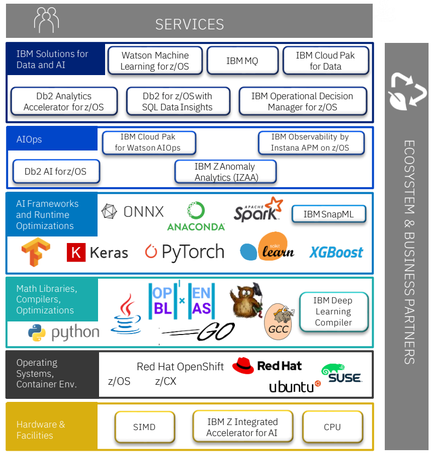
IBM Db2 Analytics Accelerator

- Integrated, hybrid workload, optimized database management system
- Runs each query workload efficiently in its optimal environment
- Ensures great performance and cost efficiency
- Exploits data-in-place to improve efficiency, drive smarter outcomes and gain competitive differentiation

AI

Optimized stack for AI

- Execute Inference Programs with Integrated Accelerator for AI
- Build and Train models on any platform
- Open source tools for framework interoperability
- New compilers optimize performance for Linux and z/OS



AI ecosystem - Seamlessly leverage AI accelerator

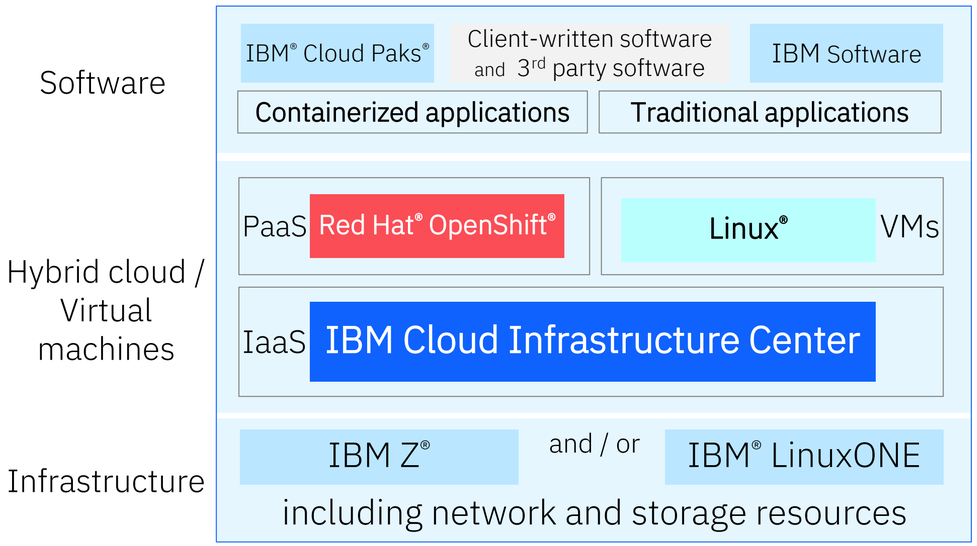
Build and train model in any popular framework on any platform of your choice

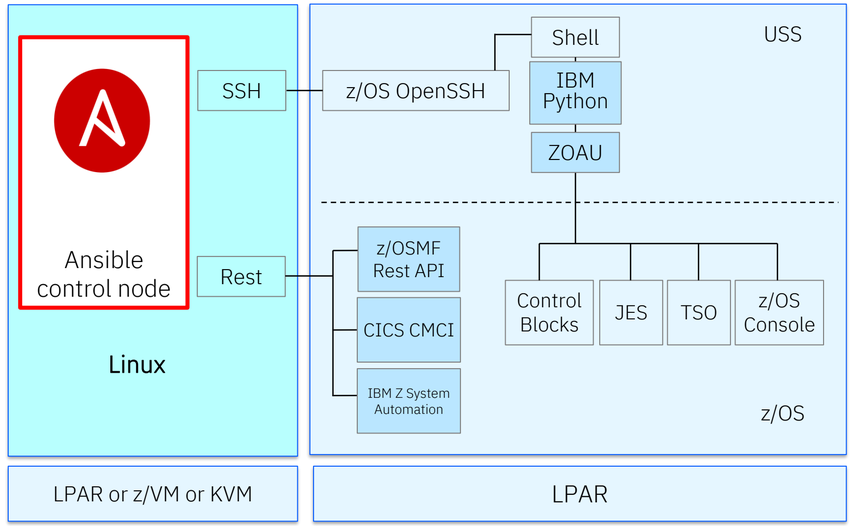
Use ONNX, an opensource tool for framework interoperability Models are converted to the ONNX interchange format

IBM Z Deep Learning Compiler, optimized for performance and new libraries, generates a program from the model for execution on Linux and z/OS

Deploy on IBM z16 and IBM® LinuxONE and infuse model into workload application

Infrastructure management and automation





IBM Cloud Infrastructure Center

Manages the lifecycle of virtual machines and integrates Infrastructure as a Service (IaaS), including the network and storage resources. This includes multi-tenancy, HA clustering, live migration, backup/restore, snapshot, etc. Industry-standard, compatible OpenStack APIs are used for simple and flexible automation and operations.

Red Hat Ansible Automation Platform

Ansible is an end-to-end automation platform to configure systems, deploy software, and orchestrate advanced workflows. This standardized approach empowers administrators to automate IBM Z / IBM® LinuxONE tasks without having platform specific skills.

Software-Defined Storage

A consistent software-defined storage foundation that extends across your hybrid cloud, can simplify operations by supporting all storage systems with the same set of APIs, procedures, and interfaces, giving applications and operations a consistent experience across on-premises and cloud infrastructure.

IBM Storage Fusion

An integrated Red Hat OpenShift solution to build, deploy, and run resilient cloud-native applications.

- User-driven deployment of software components on bare metal and virtual environments
- Flexible configuration of data services
- Multiple architectures
- Public cloud compatibility

Accelerate time to value with Red Hat OpenShift

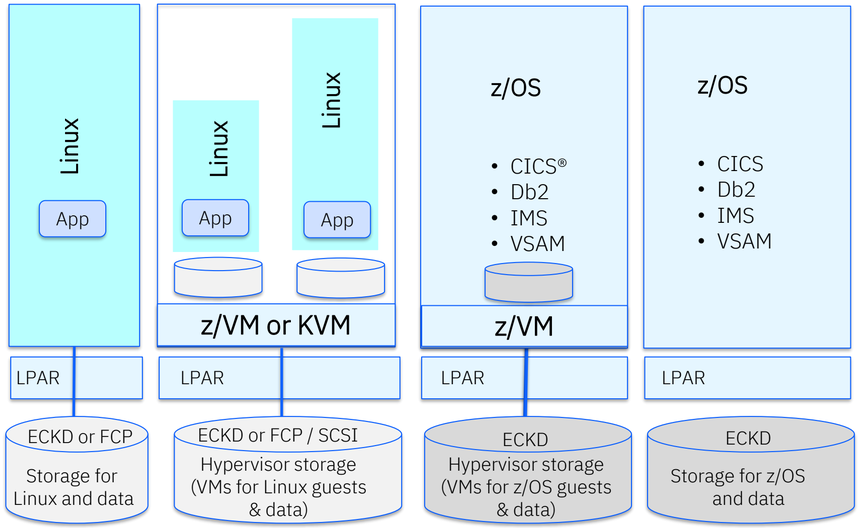
IBM Storage Scale

Get high performance, low latency storage with cloud-scale capacity for data-intensive workloads.

- Build a globally connected high-performance infrastructure
- Protect data and harden resiliency from attacks or unforeseen downtime
- Create an open hybrid cloud and connected data platform that modernizes infrastructure with global data unification

Software-defined file and object storage for AI and data intensive workloads

Virtualization



Unlike other hardware platforms, IBM Z / IBM® LinuxONE operate on virtualized hardware by default, which results in incredible performance and efficiency. IBM z/VM and Kernel-based Virtual Machine (KVM) are highly secure and scalable hypervisors to run critical applications and providing the cloud infrastructure.

IBM Z / IBM® LinuxONE virtualization technology allows to create virtual processors, communications, memory, I/O, and networking resources. It also simplifies the procedures to provide reliable, highly available, and seamless serviceability for the virtualized infrastructure.

Important, IBM Z / IBM® LinuxONE virtualization is not an afterthought; it is designed in from the bottom up.

- *IBM z/VM*: IBM virtualization that can be traced back to the beginning of virtualization in computing
- *KVM*: Open source virtualization that supports multiple hardware architectures
- *IBM Hyper Protect Virtual Server*: A fully encrypted partition with limited and encrypted network access and no access for system administrators
- *IBM Processor Resource/System Manager (PR/SM™)* or *IBM Dynamic Partition Manager (DPM)*: Firmware based virtualization to securely share and partition hardware resources

Workloads that fit well

Workloads with per core pricing

Due to different server architectures and processor speeds, Linux workloads on IBM Z/ IBM® LinuxONE can requires less processor cores than on distributed servers.

Workloads with variable resource requirements and activity fluctuations

IBM Z / IBM® LinuxONE provides compute elasticity and resource sharing, so that memory, CPU, and I/O can be allocated to workloads with diverse peak requirements.

Workloads with high security requirements

Workloads that access sensitive data can leverage IBM Z's/ IBM® LinuxONE's unique security benefits to lower the risk of a data or privacy breach.

Workloads with high availability requirements

IBM Z/ IBM® LinuxONE provides built-in redundancy and resiliency. Capacity Backup allows hardware engines to be used for disaster recovery without incurring additional software charges.

Workloads with low latency and high transaction requirements

Colocation can benefit of low latency and high throughput, since the workloads must not constantly access another system over the network.

Workloads with I/O demands

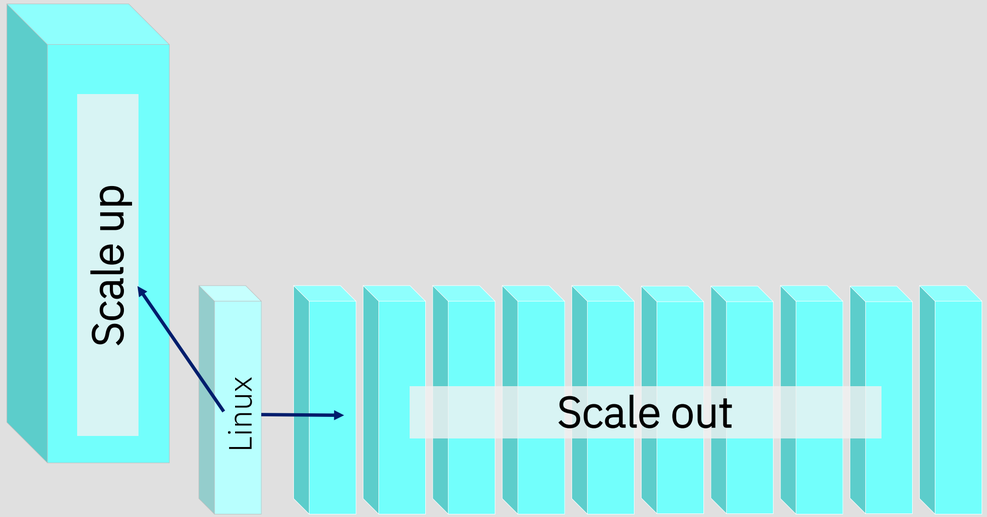
Leveraging IBM Z'/ IBM® LinuxONE's FICON® or Fibre Channel Protocol (FCP), designed to enhance data transfer and to increase sustained CPU utilization, can accelerate response times.

Scale up & out

of Linux workloads

Multi-dimensional growth and scalability options

- Scale horizontally and vertically without disruption
- Outstanding scalability, horizontal and vertical, based on the immense total IBM z16 / IBM® LinuxONE 4 capacity
- Live virtual server migration capabilities provided with z/VM Single System Image (SSI) feature and KVM
- IBM z/VM and KVM offer 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’



Backup

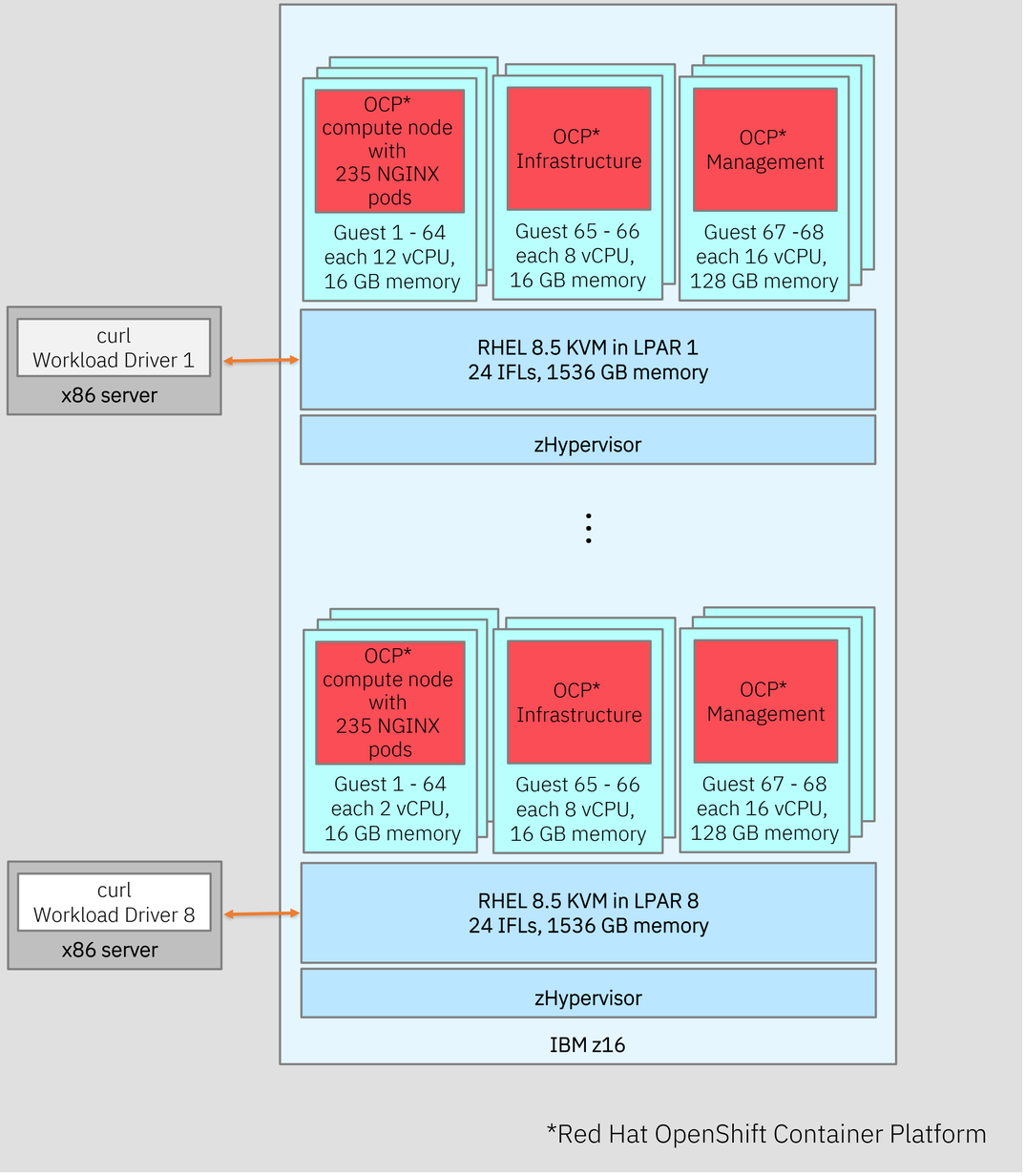
Proof points for IFL use cases

Outstanding total server capacity for large scale workload deployments

Consolidation makes the most sense when many workloads are consolidated.
IFLs and other resources capacity with z16 can ensure that each workload is optimal supported. -> **Scale-out** Red Hat OpenShift compute nodes and Pods

Example 1

On IBM z16, scale-out to **512** Red Hat OpenShift Container Platform Compute Nodes and deploy up to **100.000** NGINX pods



*Red Hat OpenShift Container Platform

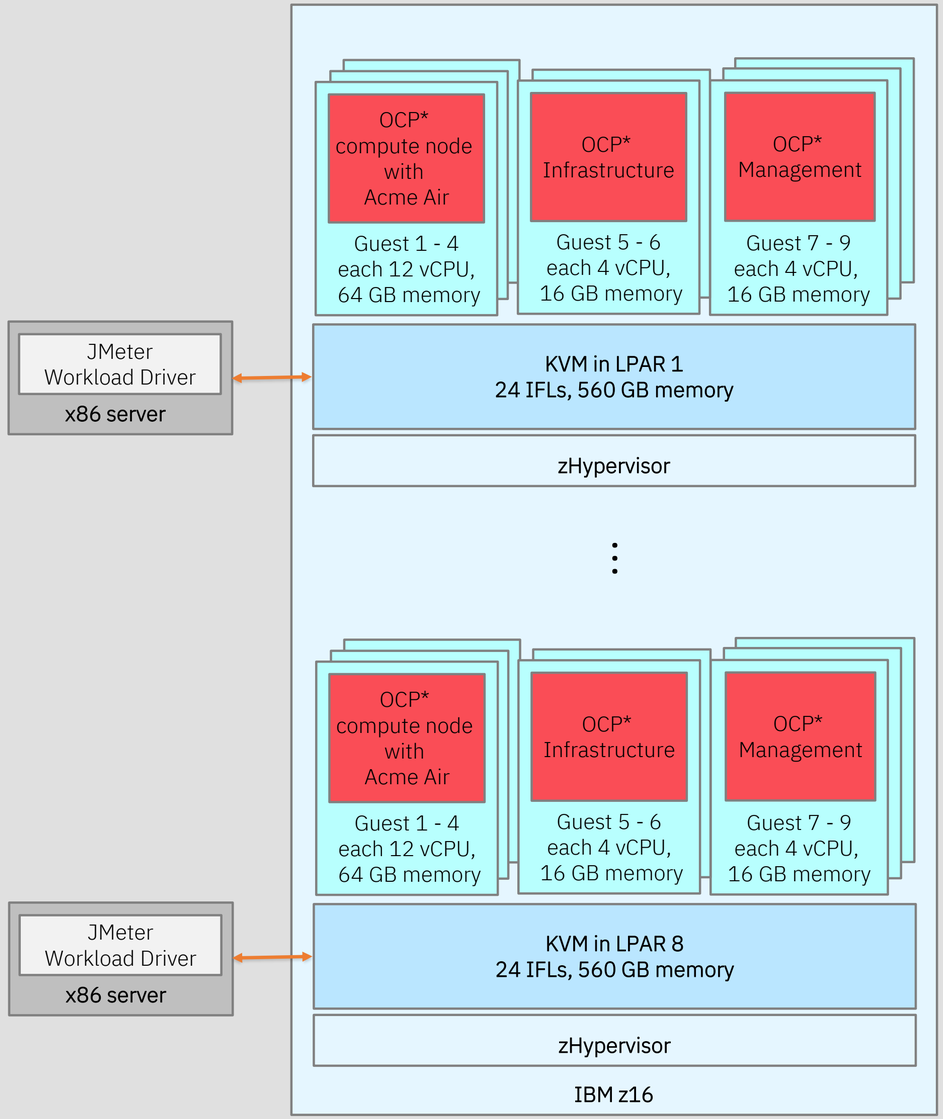
DISCLAIMER: Performance result is extrapolated from IBM internal tests running in an IBM z16 LPAR with 24 dedicated IFLs, 1536 GB memory and FS9200 storage NGINX pods on Red Hat OpenShift Container Platform (RHOCP) 4.10 running on a RHEL 8.5 KVM host. 64 RHOCP Compute nodes with 230 NGINX pods were running in parallel. The KVM guests with RHOCP Compute nodes were configured with 2 vCPUs and 16 GB memory each. The KVM guests with RHOCP Management nodes were configured with 16 vCPUs and 128 GB memory each. Results may vary.

Outstanding total server capacity for large scale workload deployments

Consolidation makes the most sense when many workloads are consolidated.
IFLs and other resources capacity with z16 can ensure that each workload is optimal supported. -> Number of **HTTPS transactions** with complex microservices

Example 2

With IBM z16, execute up to **20 billion HTTPS transactions per day** with OLTP microservice applications running on RedHat OpenShift Container Platform



*Red Hat OpenShift Container Platform

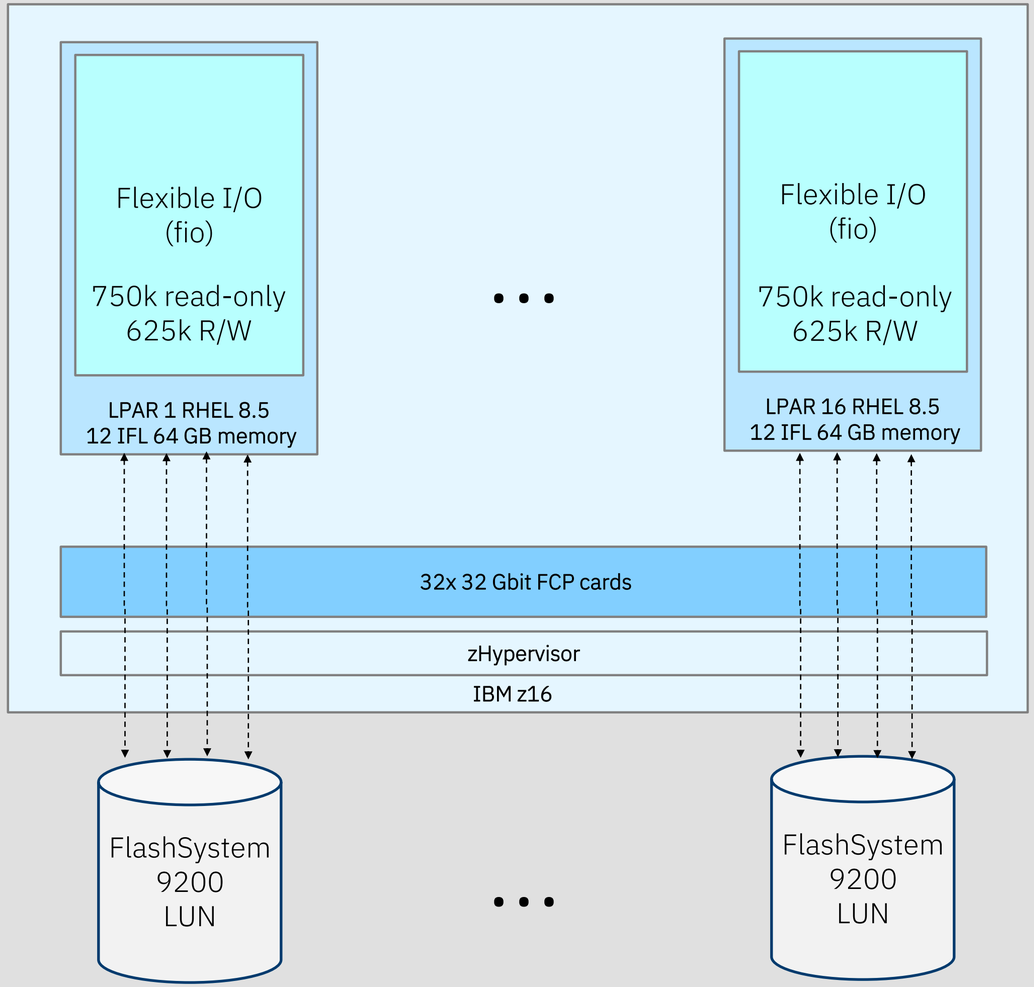
DISCLAIMER: Performance result is extrapolated from IBM internal tests running in an IBM z16 LPAR with 24 dedicated IFLs, 560 GB memory and DASD storage the Acme Air microservice benchmark (https://github.com/blueperf/acmeair-mainservice-java) on Red Hat OpenShift Container Platform (RHOCP) 4.9 using 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.

Outstanding total server capacity for large scale workload deployments

Consolidation makes the most sense when many workloads are consolidated.
IFLs and other resources capacity with z16 can ensure that each workload is optimal supported. -> Scale-up **encrypted I/O operations** with FCP

Example 3

On IBM z16, scale up your I/O intensive Linux application 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



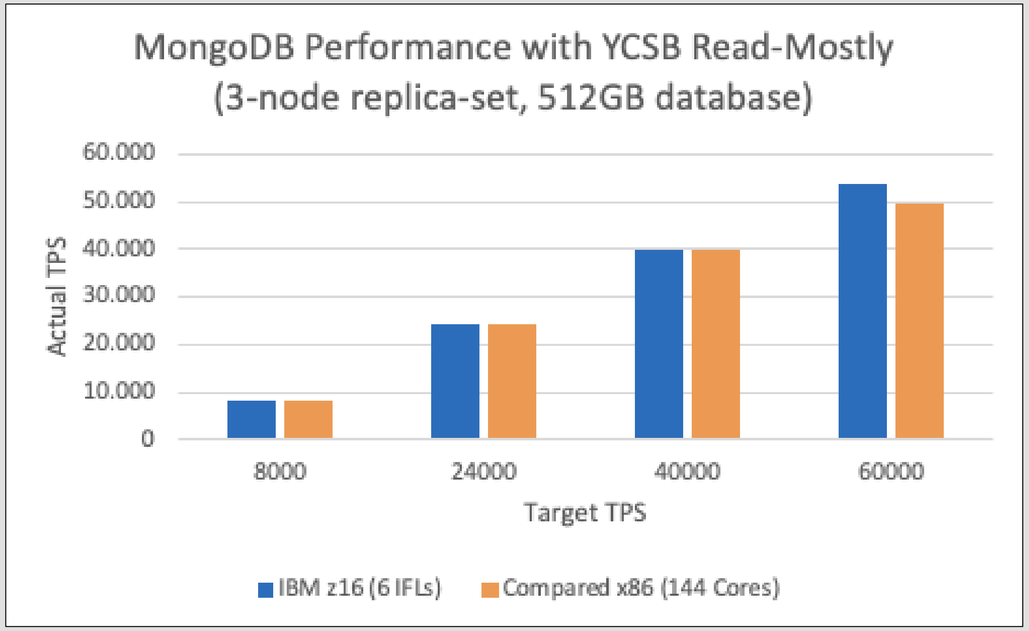
DISCLAIMER: Performance results is extrapolated based on IBM internal tests running the fio 3.19 benchmark tool. The fio benchmarking tool was run with 128 parallel threads using 8 volumes on FS9200 equally distributed over the two nodes and file size of 150GB on each volume. z16 configuration: LPAR with 12 dedicated IFLs, 64 GB memory, RHEL 8.5 (SMT mode) running fio 3.19. Two FICON Express 32S cards and Linux XFS file system encrypted with luks2 per LPAR. Results may vary.

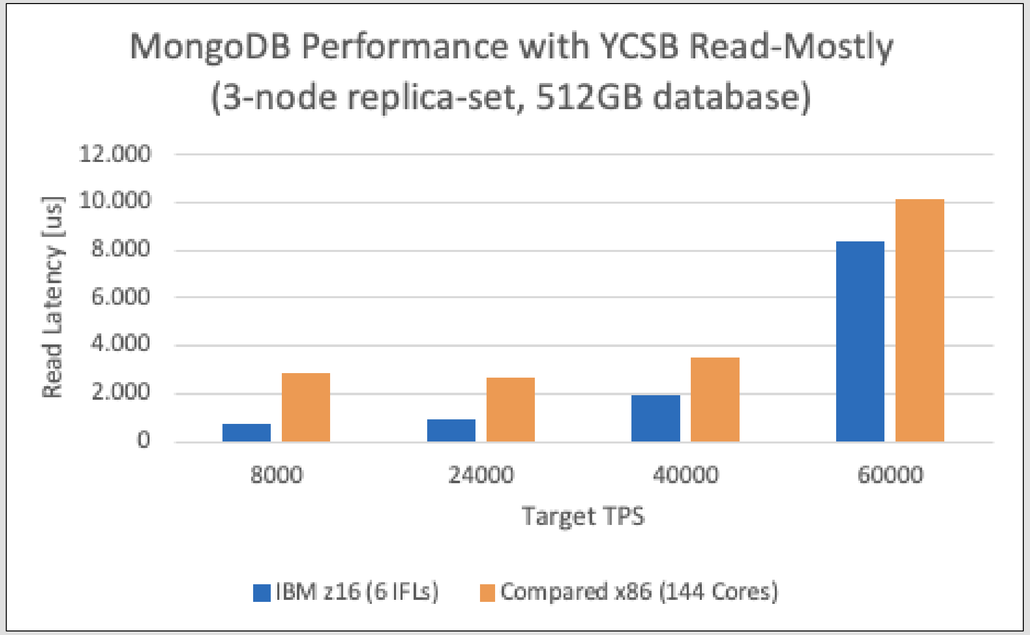
Outstanding total server capacity for large scale workload deployments

Consolidation makes the most sense when many workloads are consolidated.
IFLs and other resources capacity with z16 can ensure that each workload is optimal supported. -> MongoDB **scale-up vs scale-out** on x86

Example 4

On IBM z16 with 6 IFLs in total, run the Yahoo Cloud Serving Benchmark on MongoDB with no sharding and achieve the same throughput as compared x86 system running MongoDB with 4 shards and 144 cores in total, a **24:1 core consolidation ratio** in favor of IBM z16





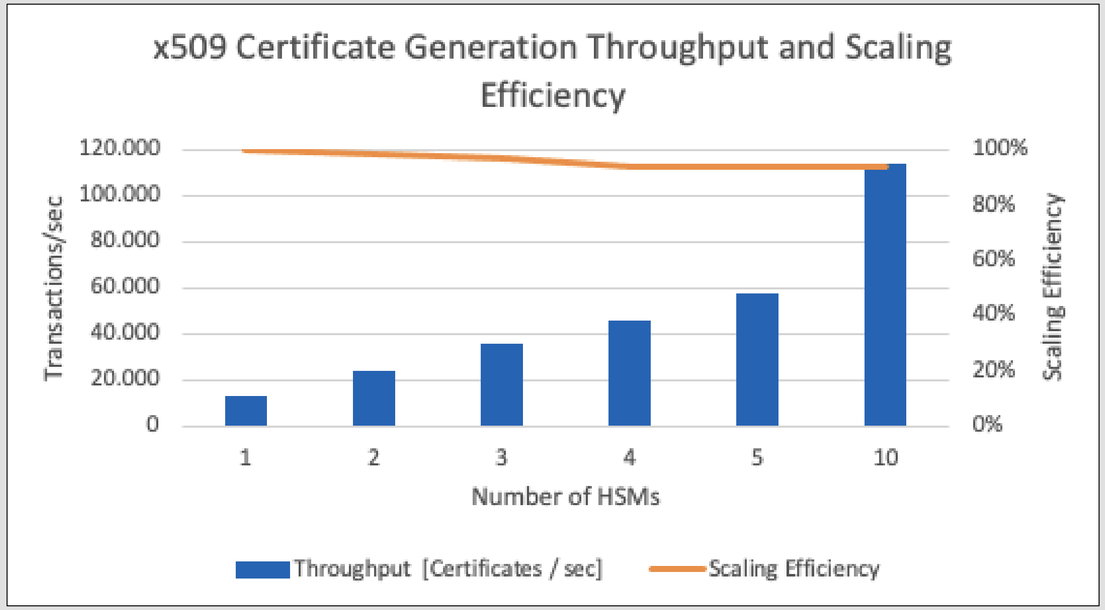
DISCLAIMER: Performance results based on IBM internal tests running YCSB 0.10.0 benchmark (read-mostly) on MongoDB Enterprise Release 5.0.6 with 3-node replication. On IBM z16 MongoDB was setup without sharding but with two replicas. IBM z16 configuration: LPAR with 4 dedicated cores and 2 LPARs with each 1 core, each with SMT and 128 GB memory, 1 TB FlashSystem 900. x86 config: 9 Intel® Xeon® Gold 5218 CPU @ 2.30GHz with Hyperthreading turned on with 144 cores in total on 5 servers, 5x 192 GB memory, 5x 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.

Outstanding total server capacity for large scale workload deployments

Consolidation makes the most sense when many workloads are consolidated.
IFLs and other resources capacity with z16 can ensure that each workload is optimal supported. -> Scale-up **certificate generation**

Example 5

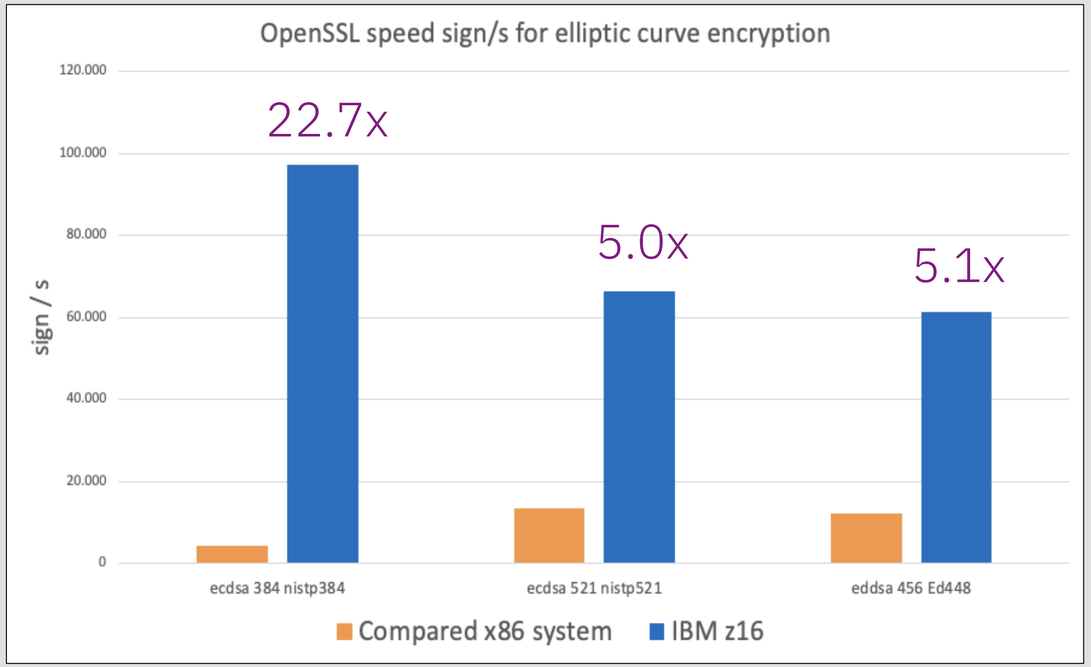
On a single IBM z16 drawer, generate up to **100,000 certificates** per second using protected keys exploiting Crypto Express 8S adapters when running application pods on Red Hat OpenShift Container Platform



DISCLAIMER: Performance results is extrapolated from an IBM internal study designed to replicate secure certificate generation with Java on Red Hat OpenShift Container Platform (RHOCP) 4.10 on IBM z16 using KVM. 2 microbenchmark pods (Signature RSA 2048 key is protected by AES master key in Crypto Express 8S adapters, certificate signatures are done with SHA-256 for x509 certificates) were run in parallel per compute node each driven locally with 20 parallel threads. IBM z16 configuration: The RHOCP Management and Compute nodes ran on RHEL 8.5 KVM using macVTap in a LPAR with 24 dedicated IFLs, 256 GB memory, FlashSystem 9200 storage, CEX8S adapters in “2 HSM version” mode, one HSM per compute node. Packages used for benchmark: IBM Semeru Open 11 JDK 11.0.14.1.1 0.30.1-1 using BouncyCastle packages for x509 certificate generation and SunPKCS11 JCE provider connected to Opencryptoki 3.16.0 CCA token for cryptographic operations. Results may vary.

Outstanding total server capacity for large scale workload deployments

Consolidation makes the most sense when many workloads are consolidated.
IFLs and other resources capacity with z16 can ensure that each workload is optimal supported. -> **OpenSSL elliptic curve signing** operation



Example 6

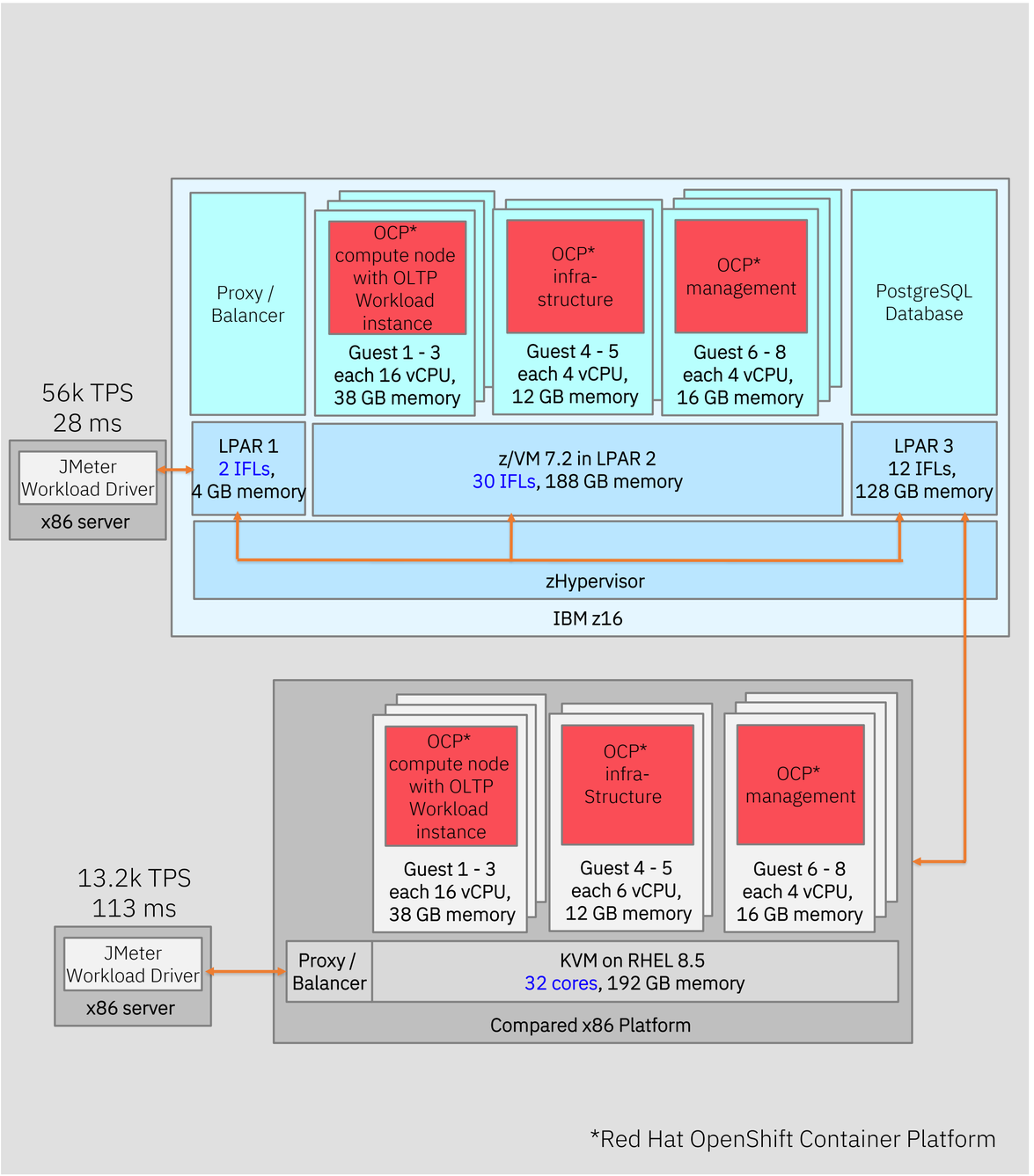
On IBM z16, run OpenSSL elliptic curve signing operations with ECDSA-NISTP384 with 22.7x more operations per second versus on compared x86 system

DISCLAIMER: Performance results based on IBM internal tests running the OpenSSL-Speed benchmark for the ECDSA-NISTP384 cipher. OpenSSL version 1.1.1.g with a cipher block size of 4k has been used pinned to 4 IFLs with SMT/HT (8 threads). IBM z16 configuration: RHEL 8.4 in an LPAR with 4 dedicated IFLs, 64 GB memory. x86 system: RHEL 8.4 on 4 IceLake Intel® Xeon® Gold CPU @ 2.80GHz with Hyperthreading turned on, 1 TB memory, local SSDs. Results may vary.

Colocation can make a big difference

IFLs and other processors on IBM z16 enable colocation, helping on resource efficiency, data access with low latency, eliminating network handling, and helping on centralized system administration.

When accessing your database while running an **OLTP workload on Red Hat OpenShift Container Platform**, achieve **4.2x more throughput** by colocating the workload on IBM z16 versus running the workload on compared x86 platform connecting remotely to the IBM z16

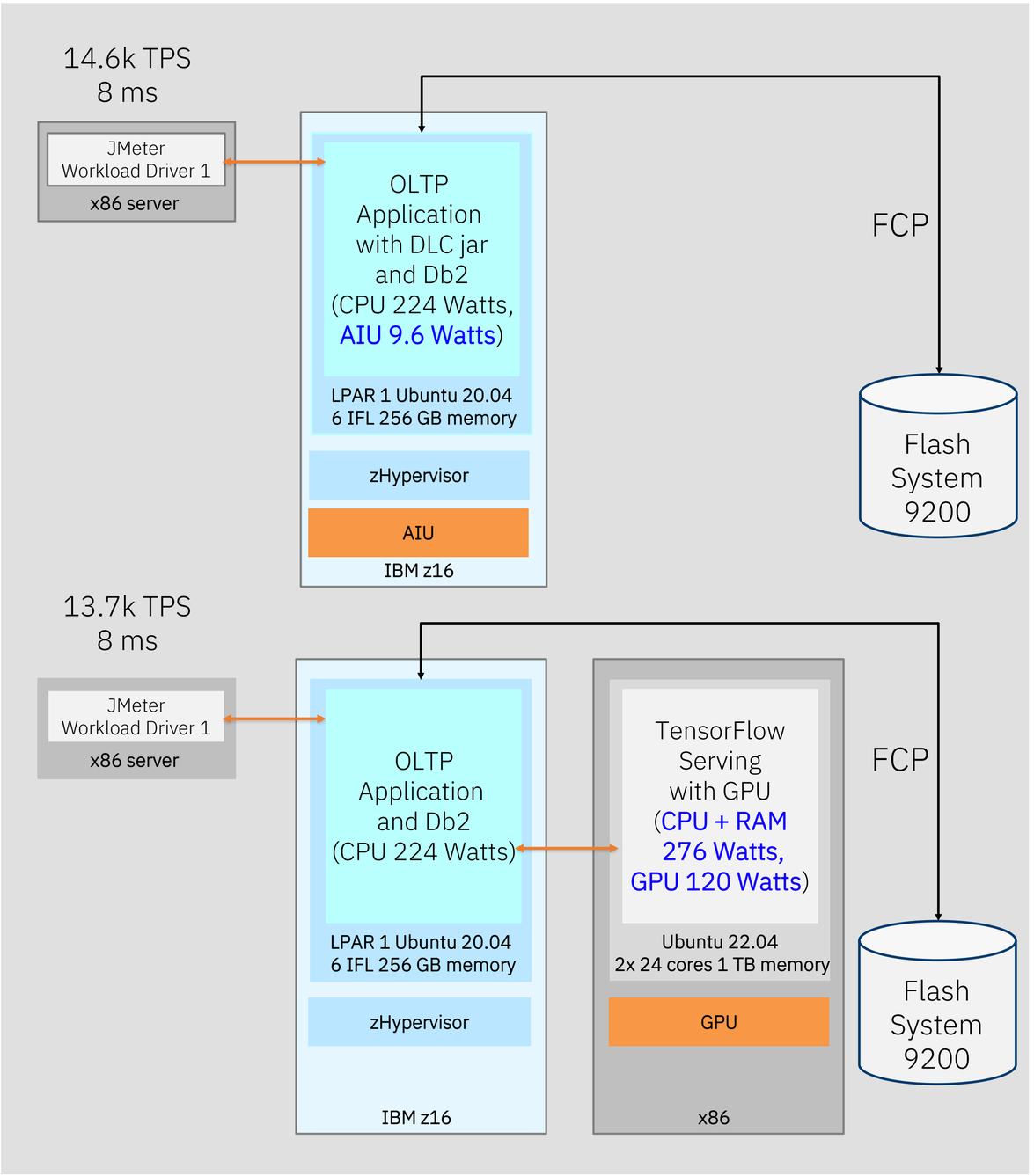


DISCLAIMER: This is an IBM internal study designed to replicate banking OLTP workload usage in the marketplace deployed on Red Hat OpenShift Container Platform (RHOCP) 4.9 on z16 using z/VM versus on compared x86 platform using KVM accessing the same PostgreSQL 12 database running in an IBM z16 LPAR. Results may vary. z16 configuration: The PostgreSQL database ran in a LPAR with 12 dedicated IFLs, 128 GB memory, 1TB FlashSystem 900 storage, RHEL 7.7 (SMT mode). The Compute nodes ran on z/VM 7.2 in a LPAR with 30 dedicated IFLs, 188 GB memory, DASD storage, and OSA connection to the PostgreSQL LPAR. LPAR with 2IFL, 4GB memory and RHEL 8.5 with RHOCP Proxy server. x86 configuration: The Compute nodes ran on KVM on RHEL 8.5 on 32 Cascade Lake Intel® Xeon® Gold CPU @ 2.30GHz with Hyperthreading turned on, 192 GB memory, RAID5 local SSD storage, and 10Gbit Ethernet connection to the PostgreSQL LPAR.

Centralized on-chip accelerator is shared with all IFLs and other processors

IFL workloads benefit from leveraging the on-chip accelerators coming with the server, not only regarding performance, but as well for sustainability.

On IBM z16, reduce the **energy consumption by 41x** using the **Integrated Accelerator for AI** to process inference operations of an OLTP workload versus running inference operations remotely on a compared x86 server using an NVIDIA GPU



DISCLAIMER: Results based on 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 using the Integrated Accelerator for AI to process inference operations vs running the OLTP workload (https://github.com/IBM/megacard-standalone) on IBM z16 with remote inferencing on a x86 server running TensorFlow serving. IBM z16 configuration: Ubuntu 20.04 in an LPAR with 6 dedicated cores, 256 GB memory, and IBM FlashSystem 9200 storage. x86 configuration: Ubuntu 22.04 on 2x 24 IceLake Intel® Xeon® Gold 6342 CPU @ 2.80GHz with Hyperthreading turned on, 1 TB memory, local SSDs, NVIDIA® A40 GPU, UEFI maximum performance profile enabled, CPU P-State Control and C-States disabled. Results may vary.

Thank you

Jerry (Gerald) Hosch
Sales enablement
hosch@de.ibm.com