The benefits of co-locating workloads on IBM Z® and IBM® LinuxONE

Highlights

- Reductions in latency and improvements in throughput
- Improved security and reliability, and simplified compliance
- Savings in total cost of computing

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

Co-location refers to the deployment on a single physical server of the presentation, business logic, and data serving layers of a multitier workload.

Does co-location matter?

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.

The physical location or proximity of workloads can improve performance when these workloads have communication patterns that are network intensive and can help to secure and protect business critical workloads.

Network intensive means that the workloads either often communicate, exchange many messages to complete a single transaction, or exchange large amounts of data, for example, big data type solutions such as analytics related workloads.

The fact that the data does not have to be moved through the network to other servers but is processed between the logical partitions (LPARs) on the individual physical server also contributes to security.

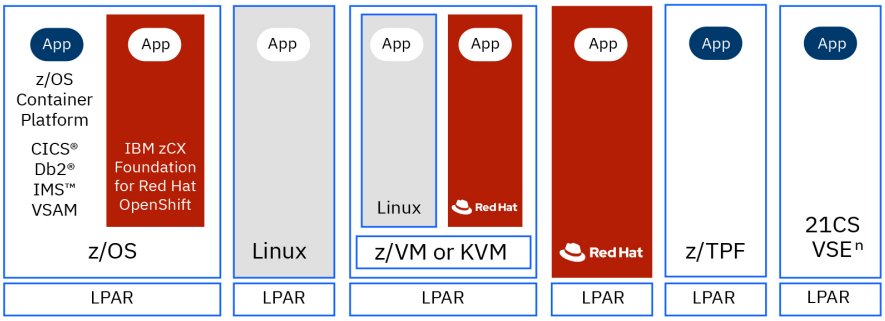
Protecting business critical data and workloads requires a comprehensive disaster recovery solution that can also meet compliance requirements. For example, the General Data Protect Regulation (GDPR) is mandating strong data protection and privacy controls. Having the IBM z/OS® and Linux® environments co-located makes it easier to process with a single solution.

Solutions benefitting from co-location

Maximizing the benefits of the IBM Z and IBM® LinuxONE platform strengths can deliver performance with co-location and cost savings through optimized infrastructure efficiency.

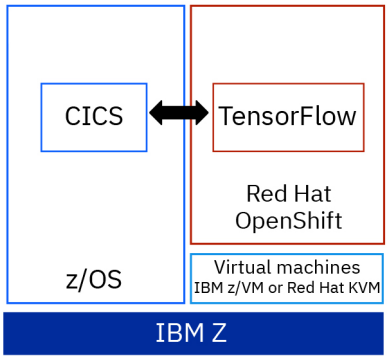
Workloads running on IBM z/OS, Linux, IBM z/TPF, IBM Hyper Protect Virtual Servers, Red Hat® OpenShift® Container Platform, and 21CS VSEn® can be co-located on IBM Z. On IBM® LinuxONE, the Linux, Hyper Protect Virtual Servers, and Red Hat OpenShift based workloads can be co-located.

Linux on IBM Z and IBM® LinuxONE provides you the choice to select from the growing ecosystem of open source and ISV software that has been ported and/or validated for building business applications.



AI processing

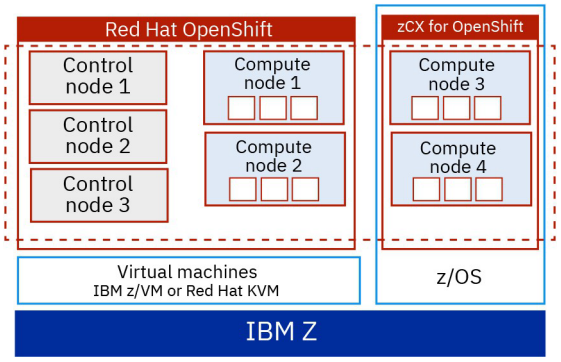
Having AI capabilities co-located with the applications that need the AI-processed data can provide immense performance improvements. With IBM z16™ / IBM® LinuxONE 4, based on the IBM Telum™ processor, you can get real-time insights when executing inference programs that exploit the Integrated Accelerator for AI. TensorFlow is one example of this kind of workload, an open source machine learning framework with tools that enable model development, training and inference.



For detailed information refer to AI on IBM Z.

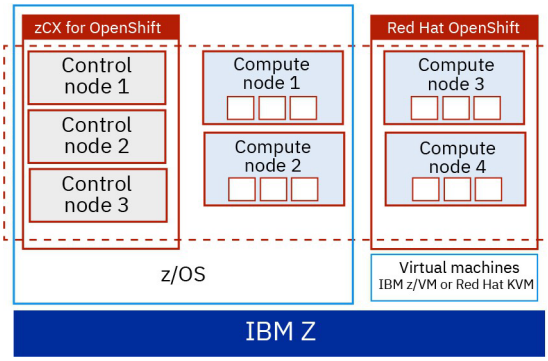
Red Hat OpenShift multi-architecture compute

Application deployment across multiple architectures in a single Red Hat OpenShift cluster is possible with the 'multi-architecture compute' feature. A Red Hat OpenShift cluster, comprising both control nodes and compute nodes, can be deployed e.g., on virtual machines (either running on z/VM® or Red Hat KVM on IFLs[1]), and compute nodes of the same cluster can run on IBM zCX Foundation for Red Hat OpenShift (zCX for OpenShift) in z/OS on zIIPs[2] – and vice versa. This co-location of compute nodes across architectures allows optimized composite solutions to be built for applications.



Co-location of Red Hat OpenShift cluster running in virtual machines and additional compute nodes running in IBM zCX Foundation for OpenShift.

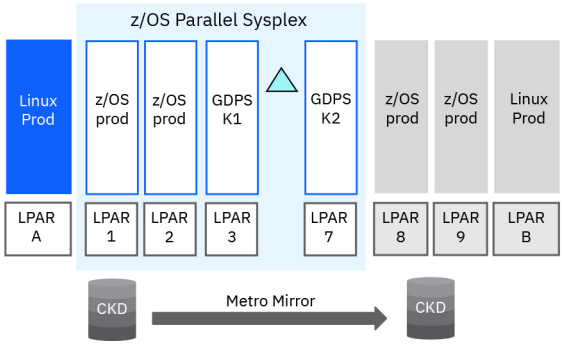
Co-location of Red Hat OpenShift cluster running in IBM zCX Foundation for OpenShift and additional compute nodes running in virtual machines.



For detailed information refer to the IBM Whitepaper:
Empower the strengths of different infrastructure architectures.

IBM GDPS®

GDPS Metro includes a function that is known as Multiplatform Resiliency for IBM Z, also known as xDR. This function provides near-continuous availability and disaster recovery capabilities in a single solution for z/OS and Linux running under z/VM or KVM or natively in a LPAR. GDPS can manage both z/OS and Linux environments simultaneously.



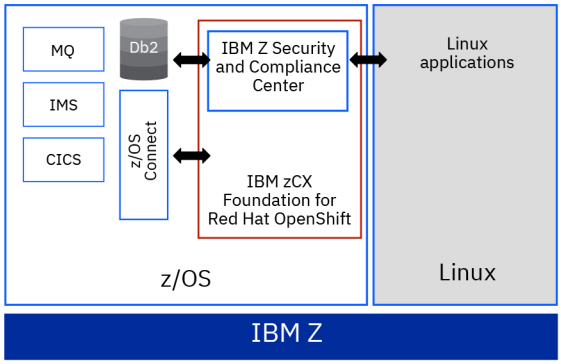
IBM GDPS® Multiplatform resiliency for z/OS and Linux in co-located LPARs

For more information refer to: IBM GDPS

IBM Z Security and Compliance Center

IBM Z Security and Compliance Center automates the collection of compliance-relevant data on z/OS and Linux on IBM Z, independent of whether it is installed on Red Hat OpenShift running in LPAR(s) or virtual machines on IFLs or on the IBM zCX Foundation for Red Hat OpenShift in z/OS on zIIPs.

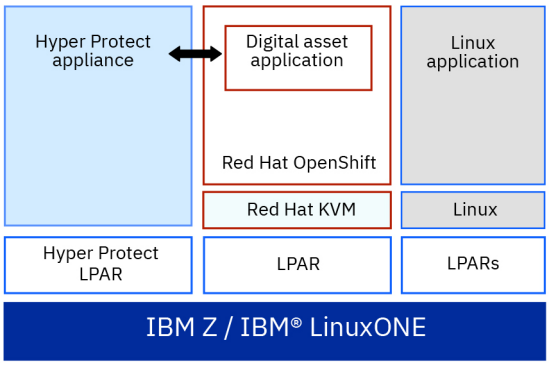
For more information refer to: IBM Z and LinuxONE Security and Compliance Center



IBM Hyper Protect Virtual Servers

IBM Hyper Protect Virtual Servers is a solution designed to protect mission-critical Linux workloads with sensitive data from both internal and external threats. It provides a confidential computing environment to protect sensitive data running in virtual servers and container runtimes by performing computation in a hardware-based, trusted execution environment (TEE). It provides secure services for on-prem and off-prem deployment of mission critical workloads. Clients using Hyper Protect Virtual Services on IBM Z and IBM® LinuxONE can benefit from co-location and integration with other workloads.

For more information refer to: IBM Hyper Protect Virtual Servers



Co-location of LPAR with Hyper Protect Virtiual Servers and LPAR with Red Hat OpenShift

Co-location supporting technologies

IBM HiperSockets™, Shared Memory Communication (SMC-D), and zdsfs are technologies that provide benefits to clients who co-locate workloads on IBM Z.

HiperSockets and HiperSockets Converged Interface provide a very efficient memory to memory transfer of standard packets between LPARs within the same central processor complex (CPC), without requiring additional or external physical networking hardware (e.g., channel adapters and LANs), thus providing low latency.

Shared Memory Communication - Direct Memory Access (SMC-D) can enable applications to replace the TCP/IP stack with an extremely thin layer, transmitting data using RDMA[3] technology with very low latencies and high throughput, while also reducing CPU consumption.

The zdsfs command allows read access to z/OS data sets hosted on DASD[4]. Data can be transferred at speeds close to FICON® limits, without detouring through the networking stack as required by other solutions.

Benefits of co-location

Reductions in latency and improvements in throughput
Workloads that access data on IBM Z can do so at memory speeds with the lowest possible latency, bypassing network traffic and delays. Network latency can affect a workload's overall throughput capabilities.

The result of an IBM internal study shows the improved throughput: *When accessing your database while running an OLTP workload on Red Hat OpenShift Container Platform, achieve 4.2x more throughput by co-locating the workload on IBM z16™ versus running the workload on a compared x86 platform connecting remotely to the IBM z16.*[5]

Leveraging AI capabilities on IBM Z by pushing queries to where the data originates simplifies the infrastructure and reduces data movement. With the IBM z16 / IBM® LinuxONE 4 on-chip integrated AI accelerator, you can benefit from AI at speed, with scale and extremely low latency.

On IBM z16 with z/OS, co-locating applications with inferencing helps minimize delays caused by network latency, delivering up to 20x lower response time and up to 19x higher throughput versus sending the same inferencing requests to a compared x86 cloud server with 60ms average network latency.[6]

Improved security & resiliency and simplified compliance
External components are vulnerable to physical attacks. Because HiperSockets, SMC-D, and zdsfs do not have physical interfaces, these technologies are protected against attacks from the outside and therefore are more secure compared to solutions using external components. Because there are no network switches, routers, adapters, or wires that can fail or require maintenance, resiliency is improved.

Co-locating z/OS and Linux on IBM Z, the IBM GDPS provides disaster recovery and resiliency solutions for both environments with a single solution.

Continuous compliance readiness is important and with the IBM Z Security and Compliance Center, you can increase the productivity of compliance personnel by reducing audit preparation time for selected controls. The Z Security and Compliance Center works with z/OS and Linux products and elements running on IBM Z to collect and validate compliance data.

Cost savings
Co-location helps to reduce the number of physical servers, which can also reduce floor space and power consumption. With HiperSockets, SMC-D and zdsfs, there are no costs for cables or other external components to buy, replace or maintain. Also, the complexity associated with procuring and managing physical network equipment is reduced and simplified.

GDPS provides a single solution for the enterprise and because less infrastructure and software might need to be duplicated, it can be a cost-effective solution.

Non-compliance can lead to increasing costs. Regulatory violations involving data protection and privacy can have severe and unintended consequences. Preparing for compliance audits can be a major time drain, and maintaining and adding new processes for compliance can be a costly effort, which the IBM Z Security and Compliance Center can help reduce.

Bottom line
IBM Z and IBM® LinuxONE enables you to co-locate workloads and offers products that provide even more benefits when workloads are co-located, not only by reducing response times and thereby improving overall performance, improving security and operational efficiency, optimizing resource sharing and utilization, and making it easier to meet service levels, but also to leverage your investments in existing assets.

Co-location of new, cloud-native applications with existing, noncontainerized workloads empowers organizations to integrate and modernize without disrupting current services.

1. *IBM Integrated Facility for Linux*
2. *IBM z Integrated Information Processor*
3. *Remote Direct Memory Access*
4. *Direct Access Storage Device*
5. *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 IBM 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. IBM z16 configuration: The PostgreSQL database ran in an LPAR with 12 dedicated cores, 128 GB memory, 1TB FlashSystem® 900 storage, RHEL 7.7 (SMT mode). The Compute nodes ran on z/VM 7.2 in an LPAR with 30 dedicated cores, 188 GB memory, DASD storage, and OSA connection to the PostgreSQL LPAR. LPAR with 2 cores, 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 5218 CPU @ 2.30GHz with Hyperthreading turned on, 192 GB memory, RAID5 local SSD storage, and 10Gbit Ethernet connection to the PostgreSQL LPAR. Results may vary.*
6. *Performance results based on IBM internal tests using a CICS OLTP credit card workload with in-transaction fraud detection. A synthetic credit card fraud detection model was used: https://github.com/IBM/ai-on-z-fraud-detection. On z16, inferencing was done with WMLz on z/OS Container Extensions (zCX). Tensorflow Serving was used on the compared x86 server. A Linux on IBM Z LPAR, located on the same IBM z16, was used to bridge the network connection between the measured z/OS LPAR and the x86 server. Additional network latency was introduced with the Linux "tc-netem" command to simulate a remote cloud environment with 60ms average latency. Results may vary. IBM z16 configuration: Measurements were run using a z/OS (v2R4) LPAR with WMLz (OSCE) and zCX with APAR–oa61559 and APAR - OA62310 applied, 8 CPs, 16 zIIPs, and 8GB of memory. x86 configuration: Tensorflow Serving 2.4 ran on Ubuntu 20.04.3 LTS on 8 Skylake Intel® Xeon® Gold CPUs @ 2.30 GHz with Hyperthreading turned on, 1.5 TB memory, RAID5 local SSD Storage.*

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
04/2025