IBM Storage Fusion Data Foundation

Reference Architecture for using IBM Storage Fusion Data Foundation on IBM zSystems

June 2nd 2023

Table of Contents

Notices . . . 1
Trademarks . . . 3
Executive summary . . . 5
Chapter 1. About this document . . . 8
Chapter 2. Value proposition . . . 10
2.1 Introducing IBM Storage Fusion and IBM Storage Fusion Data Foundation . . . 10
2.2 Benefits and value proposition . . . 12
2.3 Use Cases . . . 13
Chapter 3. Planning and architecting overview . . . 17
3.1 Storage concepts . . . 17
3.2 IBM Storage Fusion Data Foundation architecture . . . 20
3.3 Deployment topologies . . . 26
3.4 Hardware requirements for installation . . . 28
3.5 Using storage in containerized applications . . . 29
3.6 Considerations for business continuity and resilience . . . 30
3.7 Comparing IBM Storage Fusion Data Foundation with other storage solutions . . . 31
Chapter 4. Design considerations . . . 35
4.1 Storage . . . 35
4.2 Networking . . . 38
4.3 Virtualization and hypervisors . . . 38
4.4 Deployment methods and tools . . . 39
4.5 Sizing and planning the hardware infrastructure . . . 41
4.6 Implementing high availability and disaster recovery . . . 43
Chapter 5. Operational considerations . . . 56
5.1 Logging and monitoring . . . 56
5.2 Performance considerations . . . 57
5.3 Security . . . 60
5.4 Backup and recovery . . . 61
Chapter 6. Summary . . . 70
Chapter 7. References . . . 71
7.1 OpenShift® Data Foundation . . . 71
7.2 Red Hat OpenShift Container Platform . . . 71
8.3 LinuxONE and IBM zSystems . . . 71
Chapter 8. Contributors . . . 72

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to: IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice. Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals,

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided “AS IS”, without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, 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 http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

- BLU Acceleration®
- Db2®
- DB2®
- DS8000®
- Easy Tier®
- FICON®
- FlashCopy®
- GDPS®
- HyperSwap®
- IBM®
- IBM BLU®
- IBM Cloud™
- IBM Spectrum®
- IBM Z®
- IBM zSystems®
- Insight®
- Interconnect®
- OMEGAMON®
- Parallel Sysplex®
- RACF®
- Redbooks®
- Redbooks (logo) ®
- Storwize®
- System Storage™
- System z®
- Tivoli®
- WebSphere®

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

- z Systems®
- z/Architecture®
- z/OS®
- z/VM®

Red Hat, Red Hat Enterprise Linux, the Red Hat logo, OpenShift and Ceph are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

The following terms are trademarks of other companies:

- Amazon
- Intel, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries. 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 worldwide basis.
- Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
- Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Other company, product, or service names may be trademarks or service marks of others.

https://www.ibm.com/legal/copytrade

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

Executive summary

Organizations across the globe continue to rapidly develop innovative software applications in hybrid and multi-cloud environments to achieve competitive advantages and ensure customer satisfaction. Adding statefulness and persistence to those applications is essential. Software defined storage (SDS) is emerging in a hybrid and multi-cloud strategy and provides persistence in a simple and flexible and cost-efficient way. Adding a powerful layer of persistence to containerized workload is the focus of IBM Storage Fusion Data Foundation, also known as OpenShift® Data Foundation.

IBM Storage Fusion Data Foundation provides a storage infrastructure for cloud users such as software developers, infrastructure administrators, data scientists, and solution architects.

Alongside with IBM Storage Scale, IBM Storage Fusion Data Foundation is the fundamental storage technology on which IBM Storage Fusion - IBM's offering for container-native storage - is based upon.

This software stack can run on the IBM zSystems and IBM® LinuxONE platform or within IBM® z/OS® Container Extensions (zCX).

The IBM zSystems platform can help to transform a customer's storage infrastructure. It adds innovative data privacy, enterprise security, and cyber resiliency capabilities – all delivered through a hybrid cloud. In particular IBM zSystems differentiates itself in four areas from other platforms running the Red Hat® OpenShift® software stack.

- **Protected business**
  - IBM zSystems prevent security threats, they can run up to 19 billion fully encrypted transactions per day. Hardware based encryption and compressing algorithm accelerate performance of workloads.
- **App and data integration**
  - IBM zSystems provide ultra-low latency for integration and modernization of co-located Red Hat OpenShift based workloads. This includes IBM Cloud Paks as well as existing apps and their data. The interaction between existing services and modern microservices is lightning fast.
- **Flexible and scalable**
  - IBM zSystems allows for non-disruptive growth in one physical server: Workload can scale out to 2.4 million containers. By increasing the resource capacity within the same server, you can build an entire data center in a single box.
- **Efficiency and cost savings**
  - IBM zSystems can reduce the life-cycle management costs and drive savings in operational management, security, and business continuity efforts. You can achieve an impressive reduction of energy and space consumption.

By combining the strength of IBM zSystems with the open-source ecosystem of Red Hat OpenShift, the power of the mainframe can be leveraged by modern containerized workloads in a consistent and easily consumable way, which is based on industry best practices.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

This reference architecture document showcases a prescriptive, pre-validated, private SDS solution based on IBM Storage Fusion and IBM Storage Fusion Data Foundation.

The following key characteristics highlight the value of IBM Storage Fusion Data Foundation and will be detailed further in this document:

- **Open Source**
  - As IBM Storage Fusion Data Foundation is based on a strong open-source community and implemented consistently across multiple architectures, exciting storage topologies can be realized. This includes bridging bare metal, hybrid cloud environments and different hardware architectures.
- **Dynamic provisioning** of Storage
  - Dynamic, shared, and highly scalable, production-grade persistent software defined storage for OpenShift applications, which can easily grow with the needs of the application workload.
- Developed, released, and deployed in **sync** with Red Hat OpenShift
  - Aligned release cycles and version schedule for OpenShift products on all platforms.
  - Aiming at an aligned feature delivery for IBM Storage Fusion Data Foundation.
- Storage **abstraction** and portability
  - Logical storage infrastructure abstract physical storage details, so data can be stored in many different places — but act as a single persistent repository.
  - Uniform set of Kubernetes storage services and hybrid cloud data services that abstract the customers' storage infrastructure.
- **Simplified** application development and administration approach
  - Tightly integrated into the developer and admin user experience of Red Hat OpenShift Container Platform.
  - Based on a common set of skills and well-known industry best practices.
  - Developers - even those with limited storage expertise - can provide storage directly from the OpenShift platform without switching to a separate user interface.
  - As part of IBM Storage Fusion, IBM Storage Fusion Data Foundation is part of IBM's preferred software stack for container-native storage.
- Support for **data-intensive workloads**
  - Support for all common storage types, including file, block, or object storage.
  - Build and run any kind of workload created by enterprise Kubernetes users.
- Take advantage of the **IBM zSystems & LinuxONE** platform
  - Achieve latency advantages.
  - Build secure cloud solutions.
  - Dynamic, granular resource allocation and scalability up to millions of container instances.
  - Total cost of ownership reduction through consolidation into a single hardware unit.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

- Future proof backward and forward compatible hardware.
- Allow for co-location.
- Can run on IBM LinuxONE or within the IBM z/OS Container Extensions (zCX).

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

Chapter 1. About this document

Container-based applications provide organizations with a new paradigm for application development and deployment. Kubernetes and containerized applications allow organizations to streamline their DevOps environment and rapidly deploy, develop, and maintain both stateless and stateful applications.

Red Hat® OpenShift® Container Platform (RHOCP) is Red Hat's enterprise Kubernetes distribution. It provides developers, IT organizations, and business leaders with a hybrid cloud application platform for developing and deploying new containerized applications on a secure platform. RHOCP also offers scalable resources that require minimal configuration and management. Developers can create and deploy applications by delivering a consistent environment throughout the lifecycle of the application including development, deployment, and maintenance. The self-service capabilities and minimal maintenance requirements reduce the burden on IT organizations associated with deploying and maintaining application platforms.

Container workloads are stateless by definition. This fundamental assumption makes their life cycle so easy and efficient to orchestrate. A container is easily portable and can be scaled up and down at ease. A container can be restarted and be moved between servers at any time. But this implies that any kind of permanent persistence needs to be provided separately.

Adding persistence to containers and making them stateful is exactly the focus of IBM Storage Fusion and IBM Storage Fusion Data Foundation: IBM Storage Fusion Data Foundation provides software-defined storage for containerized applications. IBM Storage Fusion Data Foundation manages storage resources used by application containers and ensures that persistence is strictly decoupled from the orchestration of applications. Engineered as the data and storage services platform for RHOCP, IBM Storage Fusion Data Foundation helps teams to develop and deploy applications quickly and efficiently across hybrid cloud and multi cloud container deployments.

Along with IBM Storage Scale, IBM Storage Fusion Data Foundation is the foundation of IBM Storage Fusion, which is IBM's preferred software stack for container-native storage. By licensing IBM Storage Fusion, a customer will have access to the IBM branded IBM Storage Fusion Data Foundation. It is also possible to purchase a subscription of a Red Hat branded OpenShift Data Foundation directly from Red Hat. Both options are based on the same open-source software.

NOTE	While this document will use IBM's branding as IBM Storage Fusion Data Foundation (which is a component of the offering IBM Storage Fusion), all content applies to Red Hat's Red Hat OpenShift Data Foundation as well (which is delivered from Red Hat's Operator Hub).

This reference architecture document provides an overview of the options available for implementing IBM Storage Fusion Data Foundation on IBM zSystems and IBM® LinuxONE as well as within IBM® z/OS® Container Extensions (IBM zCX).

NOTE	While this document refers only to IBM ZSystems, all information in it also applies to IBM® LinuxONE as well as IBM z/OS Container Extensions (zCX).

In addition, this document reviews topologies and configurations, as well as implementation

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

choices and outlines their pros and cons. This document also details the best practice for installing and configuring IBM Storage Fusion Data Foundation on IBM zSystems. This also includes important design considerations and key integrations between the products of the OpenShift Ecosystem.

The information provided by this document is intended to help organizations deploy IBM Storage Fusion Data Foundation on IBM zSystems quickly in a reliable and supported configuration.

This document should not be considered as product documentation. Official documentation exists for all the components included within this reference architecture and should be reviewed prior to deployment.

This document is intended for Systems Integrators, IT Architects, Systems Administrators, Application Developers. It is assumed that the reader has previous experience with Kubernetes, Red Hat OpenShift Container Platform, IBM zSystems and IBM® LinuxONE, and Red Hat® Enterprise Linux®.

NOTE	For more information regarding about Red Hat OpenShift Container Platform and OpenShift Data Foundation visit the documentation of IBM Storage Fusion, Red Hat OpenShift Container Platform Overview and Red Hat OpenShift Data Foundation Documentation

NOTE	This document is based on IBM Storage Fusion Data Foundation on IBM zSystems as released within IBM Storage Fusion 2.5, and released by Red Hat as Red Hat OpenShift Data Foundation version 4.12.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

Chapter 2. Value proposition

Before diving into the architecture and key technical considerations of a deployment of IBM Storage Fusion with IBM Storage Fusion Data Foundation, it is essential to understand the potential benefits and values for the business.

2.1 Introducing IBM Storage Fusion and IBM Storage Fusion Data Foundation

In most cases, adding persistence to a containerized application is a necessary task. Of course, you can achieve this yourself by building directly on Kubernetes concepts. But this typically leads to variations in capability, functionality, and performance from storage to storage. For enterprise-ready applications you generally need a more powerful and consistent infrastructure layer. This is where the combination of IBM Storage Fusion, Red Hat OpenShift Container Platform and IBM Storage Fusion Data Foundation will help you.

IBM Storage Fusion is a powerful infrastructure to add persistence to containerized workloads. It is IBM's strategic offering for container native storage. Storage Fusion adds a unified install and configuration experience on top of IBM Storage Fusion Data Foundation and IBM Storage Scale, featuring common user experience.

IBM Storage Fusion Data Foundation is not just a storage system itself, but rather a set of storage and data services running on top of OpenShift and abstracting the underlaying physical storage devices. IBM Storage Fusion Data Foundation exposes highly scalable production-grade persistent storage in a uniform way, regardless of where the actual physical storage resides. It is important here to emphasize the term uniform. IBM Storage Fusion Data Foundation is optimized to run stateful containerized applications and Cloud Paks with Kubernetes storage services in a hybrid cloud environment. Storage class services, data encryption, failover, backups, dynamic provisioning are just a few capabilities, which IBM Storage Fusion Data Foundation offers to the consuming applications.

IBM Storage Fusion Data Foundation can be directly deployed on the OpenShift platform. This gives you a big advantage in terms of usability and administration experience. You can benefit from common skills in the market.

When it becomes the agnostic layer for the storage, you can run workload anywhere and have the same experience, regardless of whether you run on bare metal or in the cloud. And of course, IBM zSystems are now supported platforms for IBM Storage Fusion Data Foundation as well.

NOTE	Subscriptions for IBM Storage Fusion Data Foundation can be purchased from IBM as part of IBM Storage Fusion or from Red Hat (branded as Red Hat OpenShift Data Foundation and subscribed from Red Hat). While this document will primarily use IBM's branding as IBM Storage Fusion Data Foundation, all content applies to Red Hat's Red Hat OpenShift Data Foundation as well (unless noted otherwise).

The following list sums up the key characteristics of IBM Storage Fusion Data Foundation:

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

- Based on open-source and implemented consistently across multiple architectures.
- Enables hybrid cloud portability.
- Dynamic provisioning of additional storage as needed by the application workload.
- Closely integrated with the OpenShift ecosystem.
- Abstracts details of a distributed physical storage infrastructure and exposes a single logical storage service.
- Based on a common set of skills and well-known industry best practices.
- Simplifies application development and infrastructure administration by being integrated into the OpenShift user experience.
- Supports any kind of data intensive workload regardless of whether the data is stored as files, block or object storage.
- Can be deployed on IBM zSystems and IBM® Linux ONE or within IBM z/OS Container Extensions (zCX).
- Part of IBM Storage Fusion, IBM's strategy platform for container native storage.

NOTE The principal role of IBM Storage Fusion Data Foundation is to provide uniformity to any consuming application or Cloud Pak, regardless of where they run. Clients can deploy and consume storage wherever it is needed: regardless of whether it is on premises or in a cloud environment.

NOTE IBM Storage Fusion Data Foundation separates hardware provisioning and administration from software-based consumption by developers. Furthermore, IBM Storage Fusion Data Foundation can deal with all kinds of storage classes.

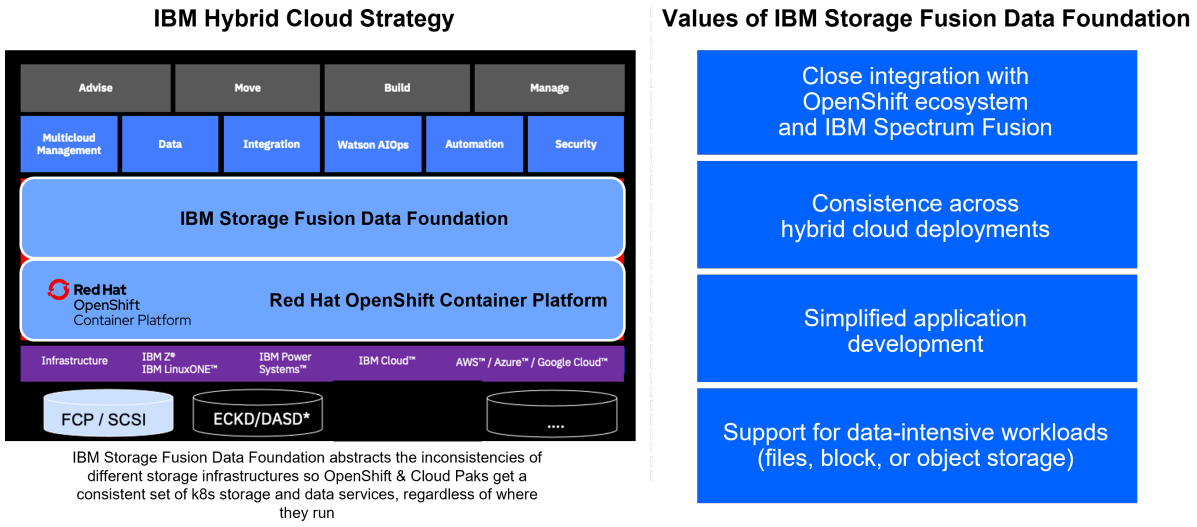

Figure 2.1-1: Introducing IBM Storage Fusion Data Foundation (I)

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

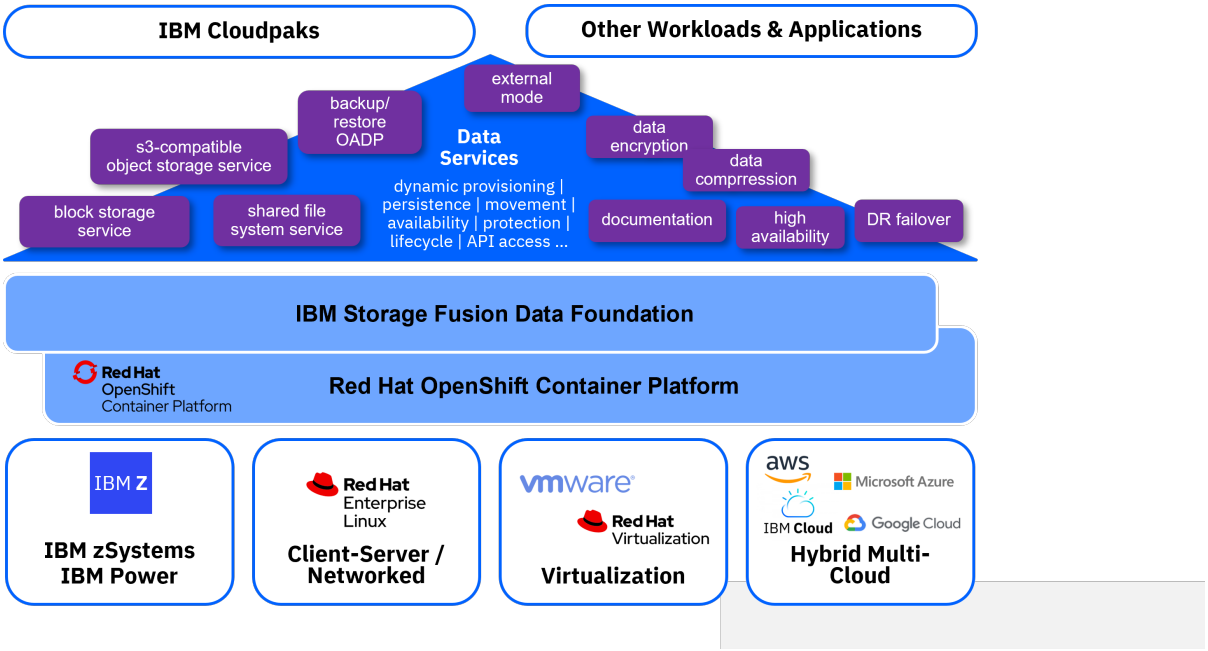

Figure 2.1-2: Introducing IBM Storage Fusion Data Foundation (II)

2.2 Benefits and value proposition

The IBM Storage Fusion Data Foundation value proposition boils down to the uniform consumption of storage by high-level services for stateful containerized applications.

There are 3 specific attributes.

- **Consistent user experience**
  - By using under the hood and well-known open-source components, you can take advantage of the abundant skills on the market.
  - As IBM Storage Fusion Data Foundation is tightly integrated into OpenShift you can unlock the Red Hat ecosystem.
  - You can easily install IBM Storage Fusion Data Foundation via operators from the OpenShift web user interface.
  - A consistent operation across hybrid cloud and the ability to manage and monitor from a single dashboard are key to efficiency.
  - IBM Storage Fusion Data Foundation is also integrated IBM Storage Fusion, which features a common install and configuration with IBM Storage Scale.
- **Dynamic scale**
  - IBM Storage Fusion Data Foundation unlocks your heterogenous infrastructure and lets it grow and expand as needed independently of application development.
  - IBM Storage Fusion Data Foundation is built on mature Ceph open-source technology that is proven to scale to tens of petabytes in production. This allows IBM Storage Fusion Data Foundation to scale linearly as workloads increase.
- **Simplified access**
  - You make your data accessible to applications across the hybrid cloud.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

    - All major storage protocols and storage classes for container workloads are supported, which makes IBM Storage Fusion Data Foundation suitable for any kind of application.
    - IBM Storage Fusion Data Foundation allows you to take advantage of a complete integrated solution, which covers the entire data lifecycle.

Specially, when running RHOCP on IBM zSystems, customers can take full advantage of the platform capabilities and characteristics:

- **Non-disruptive growth**
    - Vertical and horizontal scalability help to accommodate a substantial workload increase on demand.
- Highest **scalability**
    - Scaling up to millions of containers and thousands of Linux guests in one physical machine has been proven. See this article for more details: IBM unveils z15 .
- **Containerized workload**
    - Combines a state-of-the-art microservices architecture with traditional transactional z/OS services and databases.
- Takes advantage of **advanced security** and confidential cloud computing
    - Security includes FIPS 140-2 Level 4 certification with the IBM zSystems Cryptographic capabilities. See this article for more details on IBM Systems cryptographic HSMs.
- **Multi tenancy** with full LPAR isolation (EAL5+)
    - Administrators can share a single hardware securely as even virtual machines on the same hardware offer EAL4 certification.
- **Consolidation** and **business continuity**
    - In IBM zSystems the hardware consolidation means that capacity and operational analytics and analysis are simplified. This is because all results are consolidated on the single hardware machine for the entire RHOCP environment versus having a distributed environment with many physical servers. As a result, capacity planning is more predictable.
- Data **Gravity**:
    - IBM z/OS is designed to contain the back-end data required for the business.
- Platform agnostic **consumability**:
    - OpenShift on IBM zSystems is based on common Kubernetes skills for admins and developers and hides the hardware platform specific aspects.
- **Reliability** and **resilience**:
    - The IBM zSystems hardware is designed for error prevention, error detection, redundancy, and correction (ensuring 99.99999% platform uptime and reliability).

2.3 Use Cases

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

2.3.1 Generic use cases

In general, IBM Storage Fusion Data Foundation addresses critical use cases, when it establishes a storage infrastructure, and it adds persistence to application workloads.

- **Administer** storage hardware
  - Deploy storage services with maximum flexibility and mobility.
  - Expand storage as needed (for example, add capacity).
  - Provide out-of-the-box dashboards in the OpenShift Administrator Console to indicate storage health and status.
  - Alert users when there is a storage issue.
  - Self-healing and recovery of issues related to storage infrastructure.
  - Run a internal image registry on IBM Storage Fusion Data Foundation storage.
  - Add storage for OpenShift Logging.
  - Add storage for OpenShift Monitoring.
- **Develop** containerized applications which consume persistent storage
  - Implement simple and consistent data access whether on premise or in the cloud.
  - Leverage well-known standard Kubernetes APIs / operators and use of persistent volume claims and persistent volumes.
  - Develop, test and deploy end-to-end with one consistent platform.
  - Consume persistent volumes when deploying applications, which require storage.
  - Use shared and non-shared storage options.
  - Take advantage of common types of storage classes:
    - Block storage is suitable for databases and systems of record.
    - File storage is suitable for messaging, data aggregation, workloads machine learning, and deep learning.
    - Object storage is suitable for images, non-binary files, documents, snapshots, or backups. S3 APIs are the de facto standard for object storage.

2.3.2 Specific use cases

This reference architecture is valuable for cross-industry enterprises that want to deploy a private cloud solution with programmable Infrastructure as a Service (IaaS), Containers as a Service (CaaS), and Platform as a Service (PaaS) capabilities. Specific target use cases, where IBM Storage Fusion Data Foundation can be applied, aim at:

- Data gravity
  - A key use case in which RHOCP can take advantage of the IBM zSystems platform is the co-location of containerized applications with traditional workloads (for example, data lakes, databases, transactional systems, or other traditional workloads running in Linux on IBM zSystems or z/OS). In such a scenario the applications can be located close to the data to

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

optimize latency, response time, deployment, security, service, and cost. In combination with Oracle databases as an example, the resulting synergy can dramatically improve the Total Cost of Ownership (TCO).

- Consolidation and TCO Reduction
  - When consolidating a RHOCP environment from x86 platforms to IBM zSystems, customers can achieve economic and operational advantages.
  - Most of the 3-dimensional scalability (vertical, horizontal and combined) results in high flexibility without the need for a new hardware footprint. This is a key advantage for dynamic workloads and unpredicted growth.
- Business continuity
  - For business continuity consider these two different aspects:
    - For high availability (HA), IBM zSystems provide an internal network with significantly more reliability based on a higher degree of redundancy in the hardware. Because virtualization happens within a single hardware environment, the networking traffic is more predictable with less latency compared to x86 environments.
    - Building a disaster recovery (DR) setup with IBM zSystems is much simpler in such an environment because you must deal with fewer hardware units.

2.3.3 Target industries

IBM Storage Fusion Data Foundation is intended for generic orchestration and life-cycle management of containerized workloads. While it is used widely among industries, most relevant background of IBM Storage Fusion Data Foundation for IBM zSystems customers is in industries such as:

- Banking and insurance (with a strong focus on high availability, transactions, and security).
- Government (with strong focus on high availability and security).
- Retail (with strong focus on scalability and coping with peak loads).
- Cloud and computing services (with strong focus on high availability and variable load / scalability).

Additional industries which can benefit include:

- Media
- Telecommunications
- Education
- Transportation
- Pharmaceutical & medical products, healthcare
- Hospitality & travel
- Manufacturing
- Utilities and energy
- Electronics

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

- Chemicals
- Consumer products
- Metals & natural resources
- Construction & engineering
- Food & beverage processing

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

Chapter 3. Planning and architecting overview

In this chapter you will learn architectural choices and understand how IBM Storage Fusion and IBM Storage Fusion Data Foundation can be deployed and used in typical environments on IBM zSystems and IBM® LinuxONE as well as within the z/OS Container Extensions.

3.1 Storage concepts

3.1.1 Adding persistence to containerized workloads

Containers are a very powerful concept to simplify application deployment and operation dramatically. Containers are small, lightweight bundles. They include exactly the subset of the operating system that they need. No more, no less.

- On the one hand, containers avoid the deployment of a full copy of an operating system on each container.
- On the other hand, containers package all required dependencies, such as runtime environments or libraries.

As a result, a customer deployment will face significantly fewer integration problems later because containers can run just about everywhere.

Furthermore, containers assume a strongly simplified lifecycle. This makes orchestrating a containerized workload extremely easy; a container can be restarted, scaled up and down, or moved back and forth between servers at any time.

However, this flexibility and ease comes at a price; the simplified lifecycle assumes that containers run stateless without any persistence; after exiting or restarting a container all data in it will be lost. It is the responsibility of the application workload to take care of storing data on its own.

When storing a containerized workload's data, a customer needs to consider how storage can be easily provided, mounted, and scaled. Kubernetes provides some basic capabilities to help here. Among those capabilities, the most notable are persistent volumes and claims, storage classes and a set of operators to deal with storage.

But for larger enterprise scale applications a customer will soon conclude that a more overarching and consistent approach is needed. To address this, various kinds of software-defined storage offerings build powerful infrastructure around the basic Kubernetes capabilities. Products such as IBM Storage Fusion Data Foundation or IBM Storage Scale feature a nice separation of storage hardware from the consuming software layers, which makes development easier. They also facilitate high availability and the scaling of storage hardware for serious production use.

3.1.2 Container-ready and container-native storage

For containerized applications and services, there are two kinds of storage for those that require stateful applications and tlong-term data persistence.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

- Container-ready
- Container-native storage

Container-ready storage includes Storage Area Network (SAN), Software-Defined Storage (SDS), and Network Attached Storage (NAS) devices. These devices support capabilities such as backups, snapshots, clones, and data replication. As container-ready storage they continue to operate in their environments unchanged and independently of the containerized workload itself. The storage environment and the OpenShift Container Platform cluster co-exist next to each other.

A benefit of choosing any of these options is that they often allow clients to leverage their existing infrastructure investments (either in networking or storage). Existing processes for management, scaling and monitoring can be used.

IBM Storage Scale, is one example of a product which offers “container-ready” storage options.

The alternative is container-native storage. Container-native storage types are already deployed within containers and are presented to containerized applications as ready-to-go storage. Processes for deployment, management and scaling are similar for storage and applications as both are containerized.

Productized versions of container-native storage include IBM Storage Fusion Data Foundation based on Red Hat OpenShift Container Platform.

What benefits do container-native storage provide? Container-native storage simplifies management by offering a single control plane within your entire Kubernetes containerized environment and the ability to manage the integrated storage just like the application workload. At the same time, this architectural paradigm allows storage and compute nodes to scale independently of one another.

In IBM Storage Fusion Data Foundation it is possible for container-ready and container-native storage to coexist within the same cluster, giving further flexibility to clients who might have a mixture of persistence options available from their previous investments. This co-existence is shown in the figure below.

- Container-ready storage connects the OpenShift compute nodes directly to the storage devices via a fabric of APIs and provisions volumes directly on the storage device.
- By contrast, container-native storage devices consume storage either as disk drives or external storage. Container-native storage creates a new abstracted storage pool, which is managed and scaled as containers in a Kubernetes or OpenShift environment. That storage pool is used to create persistent volumes, which is used by the application containers.

IBM Storage Fusion includes both IBM Storage Scale and IBM Storage Fusion Data Foundation and exposes a common usage interface for admins and developers.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

NOTE

IBM Storage Fusion Data Foundation is based on the same code base as Red Hat OpenShift Data Foundation While this document will use IBM's branding as IBM Storage Fusion Data Foundation (which is a component of the offering IBM Storage Fusion), all content applies to Red Hat's Red Hat OpenShift Data Foundation as well (which is delivered from Red Hat's Operator Hub). Both distributions are built from the same open-source code base.

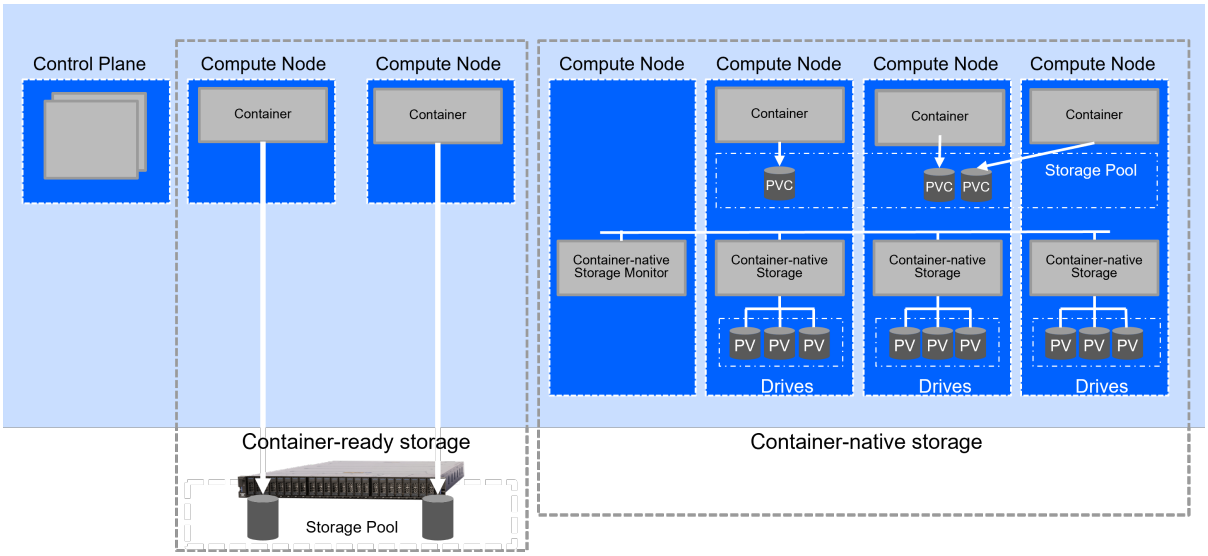

Figure 3.1-1: Combining container ready and container-native storage

3.1.3 Persistent volume claim and persistent volume

Adding persistence to a containerize application in a Kubernetes-based environment such as OpenShift, boils down to a simplistic approach and concept.

- One the one hand, developers make use of Persistent Volume Claims (PVCs) to request the storage resources they need for their applications. This PVC is an abstraction of the storage, which is available to applications, and declares how much and what kind storage is needed by the consumer. PVC is a common Kubernetes concept, and IBM Storage Fusion Data Foundation takes advantage of this concept.
- On the other hand, the storage administrator provisions storage for the cluster as persistent volumes (PV). PV resources on their own are not scoped to any single project. They can be shared across the entire OpenShift Container Platform cluster and claimed from any project.
- Under the hood, IBM Storage Fusion Data Foundation dynamic binds the claims to the actual available storage. It maps the physical storage to the needs of the developer. After a PV is bound to a PVC, that PV cannot be bound to any additional PVCs. Each PV represents a piece of existing storage in the cluster that was either statically provisioned by the cluster administrator or dynamically provisioned.

This simple fundamental concept makes IBM Storage Fusion Data Foundation easy to deploy, use and administer.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

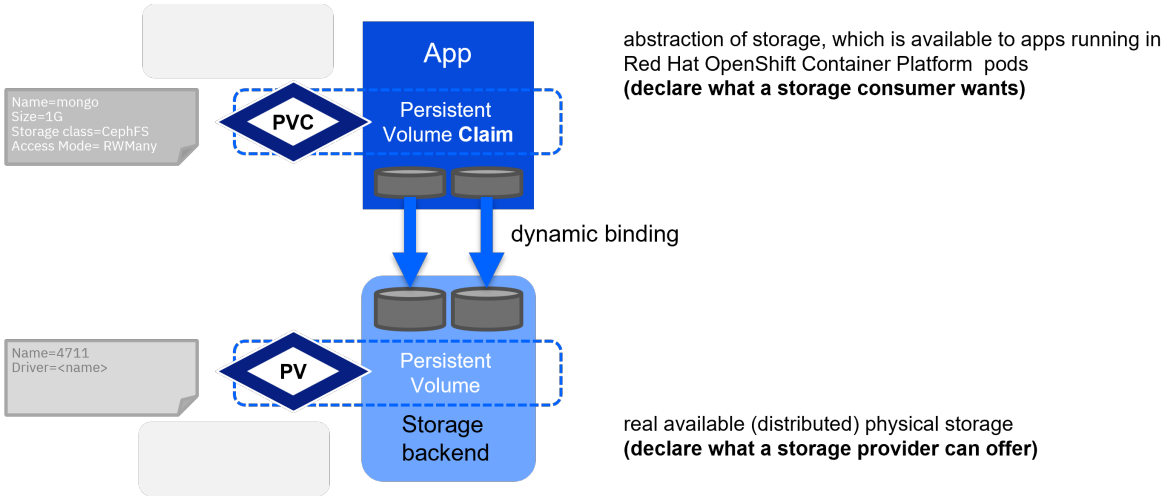

Figure 3.1-2: Persistent volume claim and persistent volume

3.2 IBM Storage Fusion Data Foundation architecture

3.2.1 Internal components

IBM Storage Fusion Data Foundation is based on a open-source technology stack including Rook, Ceph and NooBaa.

- **Ceph** is IBM Storage Fusion Data Foundation's core storage platform. Ceph is based on RADOS (Reliable Autonomic Distributed Object Store). RADOS is by itself is an open-source object storage service and integral part of Ceph distributed storage system.
- **Rook** is a storage orchestrator for Kubernetes and coordinates the services provided by Ceph. Rook has operators to support different storage backends.
- **NooBaa** (Multi Cloud Object Gateway) provides consistent S3 API endpoints across different backend infrastructures, such as AWS, Azure, GCP, IBM Cloud, Bare Metal, VMware, or OpenStack.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

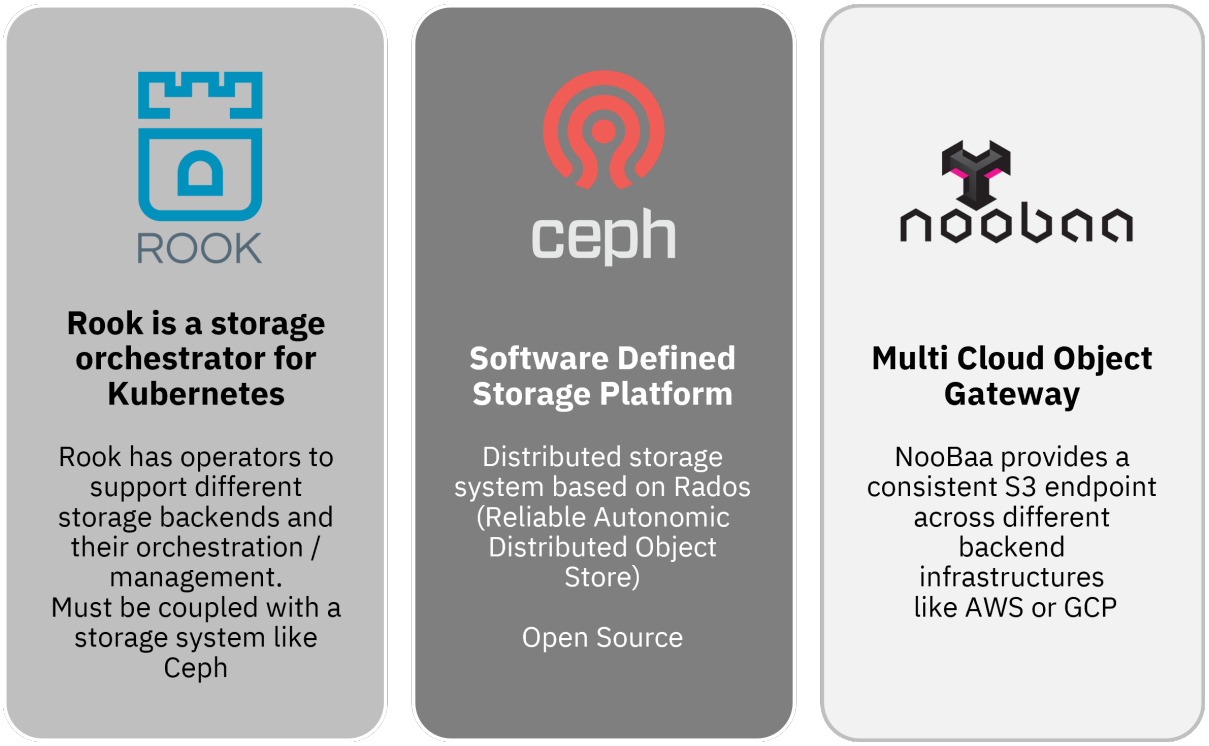

Figure 3.2-1: Components of IBM Storage Fusion Data Foundation

3.2.2 Ceph

After Ceph has been installed, it exposes the following storage classes

- RBD block storage (provisioned by openshift-storage-rbd.csi.ceph.com)
- CephFS file storage (provisioned by openshift-storage-cephfs.csi.ceph.com)
- RGW object storage (provisioned by openshift-storage-rgw.csi.ceph.com)

A key characteristic of Ceph is its massive scalability up to petabytes of data. Ceph includes mechanisms for self-management and self-healing, as well as for distribution of data across servers and disks. The storage entity, which Ceph orchestrates is referred to as Object Storage Daemons (OSD). An OSD holds the actual data exposed as persistent volumes, including multiple copies for redundancy.

NOTE	In addition to the RGW object storage provided by Ceph, the Noobaa Multi Cloud Gateway provides an alternate implementation of object storage. Both expose the S3 API, which is a REST-based and de facto industry standard to interface with object storage.

3.2.3 Rook

Rook as storage orchestrator is responsible for the packaging and management and scaling of the storage services provided by Ceph. Rook makes Ceph much easier to consume, as complex operations and configuration are taken care of by Rook's automation and abstraction. Rook also detects and repairs inconsistent objects ensuring data is protected and coherent.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

3.2.4 Multi Cloud Gateway (Noobaa)

While Ceph and Rook build the core software stack of IBM Storage Fusion Data Foundation, Noobaa provides the Multi Cloud Object Gateway (MCG) interface to interact with object storage. As mentioned above this is an alternate implementation of the object storage class, which IBM Storage Fusion Data Foundation offers in addition to Ceph Object Gateway (Ceph's native object storage interface). Using MCG, applications can create object bucket claims to store objects. MCG creates its corresponding object buckets and also generates a config map with the required credentials to access the bucket. Object buckets and object bucket claims are parallel and similar to PVs and PVCs.

At the same time, MCG acts as a proxy between the local OpenShift cluster and external object storage services offered by one or more cloud vendors, such as Amazon Web Services and IBM Cloud. The big advantage of this setup is that data can be kept in its place at the remote cloud vendor. When using object storage from applications running within the OpenShift cluster, there is no need to copy data back and forth. Instead, a highly efficient caching mechanism achieves significant performance improvements. You can use Noobaa to manage your s3 compatible object storage resources in cloud services such as IBM COS, AWS S3, or Azure Blob storage.

When the application claims an object storage as an object bucket claim, MCG creates its corresponding object bucket. Each bucket can have its own data placement and access control policy and can be changed over time. This allows adaptation to the changing needs of applications and environments. For example, a bucket can span multiple different cloud storage providers and manage the necessary authentication and caching.

For applications using multiple data sources residing in more than one cloud service or leveraging a combination of cloud and on premise, object storage can be achieved with Namespace buckets. MCG namespace buckets provide flexible data federation over multiple clouds and data object stores. Namespace buckets do not replicate or mirror data from the data resources and the storage admin decides the scope of the namespace.

In short: Noobaa Multi Cloud Object gateway is an exciting opportunity to open IBM Storage Fusion Data Foundation to a broader Hybrid Cloud topology.

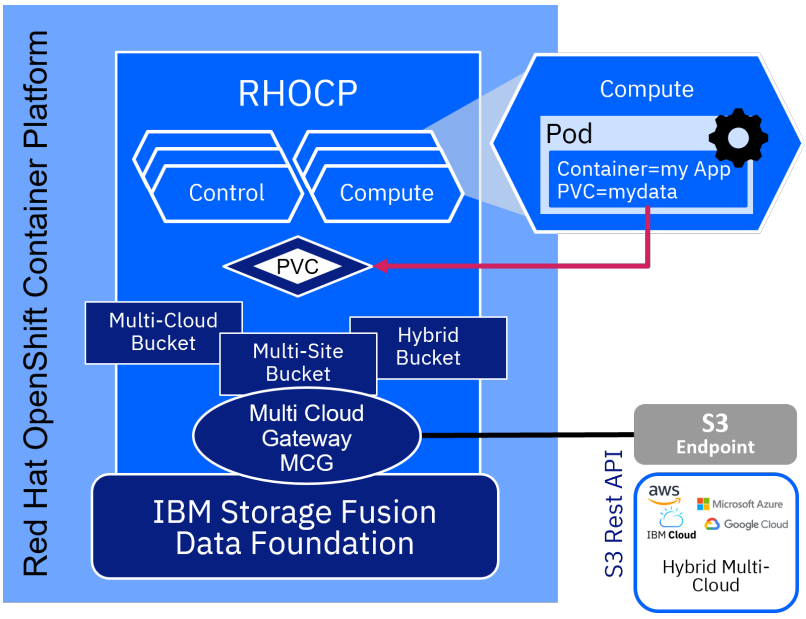

Figure 3.2-2: Noobaa

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

3.2.5 OpenShift Operators

IBM Storage Fusion Data Foundation is deployed inside OpenShift Container Platform using operators provided by IBM Storage Fusion as well as the Red Hat Operator Hub. In conjunction with the Local Storage Operator and the Rook-Ceph operator the installation creates a containerized Ceph storage cluster running on OpenShift Container Platform. The operators are available through the Red Hat OpenShift Container Platform Service Catalog.

Installing software add-ons to OpenShift via operators is extremely user-friendly and straight-forward. It is seamlessly integrated into the overall OpenShift administration experience. Making IBM Storage Fusion Data Foundation part of all of this is a key advantage and reason to choose IBM Storage Fusion Data Foundation as persistence layer for stateful containerized workloads.

NOTE

One reminder: Currently IBM Storage Fusion Data Foundation is the naming for the IBM branded code base obtained as paret of the larger offering IBM Storage Fusion. The code base is also available from Red Hat's Operator Hub as Red Hat branded Red Hat OpenShift Data Foundation. This also includes the naming of the operators in the OpenShift catalog.

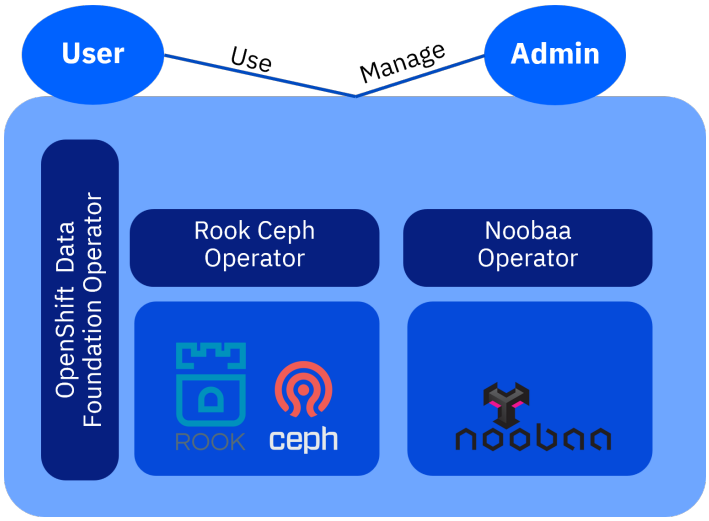

Figure 3.2-3: Operators of OpenShift Data Foundation

3.2.6 Storage classes provided by IBM Storage Fusion Data Foundation

IBM Storage Fusion Data Foundation supports a variety of storage classes, which give developers flexibility when they implement workloads. Differentiators between storage classes are the supported attachment modes and data handling characteristics. Storage classes provide a way for administrators to describe the characteristics of the storage they offer. Different classes might map to quality-of-service levels, backup policies, or arbitrary policies determined by the cluster administrators.

For data access, the characteristics are distinguished depending on how they are shared or non-shared across OpenShift Container Platform nodes.

- ReadWriteOnce (RWO): the volume can be mounted as read-write by a single node.
- ReadOnlyMany (ROX): the volume can be mounted read-only by many nodes.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

- ReadWriteMany (RWX): the volume can be mounted as read-write by many nodes.

A key differentiator is the type of data the storage class contains:

- **Block Storage**: Mainly appropriate for ReadWriteOnce access modes. However, ReadWriteMany can be appropriate if the application can maintain data consistency and integrity. Suitable for databases and record systems.
- **File Storage** (Shared and distributed file system): Appropriate for both ReadWriteOnce and ReadWriteMany access modes, as the underlying file system is designed for multiple threads and multi-tenancy. Suitable for messaging, data aggregation, workloads machine learning, and deep learning.
- **Object Storage**: Suitable for images, non-binary files, documents, snapshots, or backups. In IBM Storage Fusion Data Foundation either Ceph RGW or Noobaa Multi Cloud Gateway MCG can be used. Both expose S3 Rest API to the application.

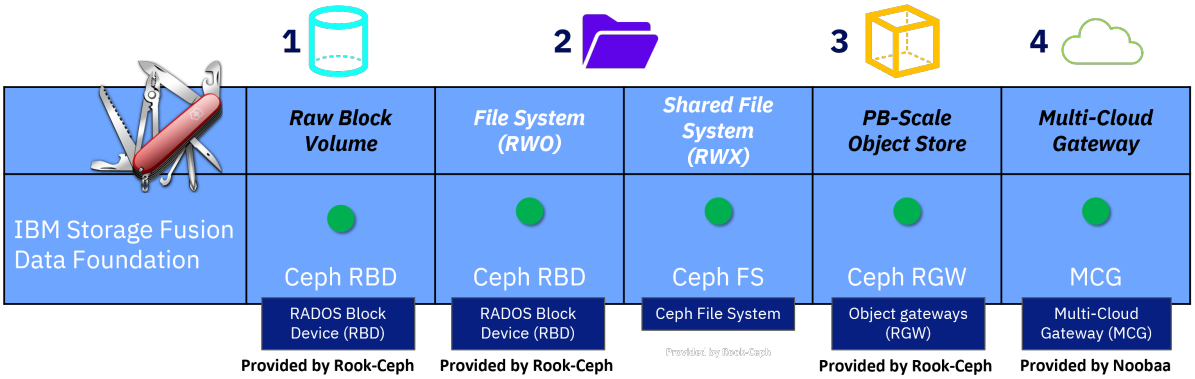

Figure 3.2-4: Storage classes of IBM Storage Fusion Data Foundation

3.2.7 How IBM Storage Fusion Data Foundation works internally

IBM Storage Fusion Data Foundation is deployed and running inside a OpenShift Container Platform cluster as storage nodes. Storage nodes can run either on a compute or infrastructure node. A complete IBM Storage Fusion Data Foundation installation needs 3 storage nodes.

A containerized application runs on OpenShift compute node. When there is a need to store state and data, the application creates a persistent volume claim (PVC) and references it in its configuration. Each PVC specifies the storage class to be used (for example, ocs-storagecluster-ceph-rbd), the amount of storage needed (for example, 5 Gi), and other characteristics.

A PVC is bound to a persistent volume (PV). A PV is the logical representation of a storage unit and a Kubernetes concept, but it is NOT yet real physical storage.

IBM Storage Fusion Data Foundation orchestrates the mapping of a persistent volume to real physical storage to satisfy the persistent volume claims requested by the applications:

- A PV is managed by a Ceph CSI driver on the node on which the consuming application is running. The Ceph CSI driver uses a RADOS Block Device or Ceph File System component from the Linux kernel to facilitate the interaction between the PV and a corresponding Ceph Object Storage Daemon (OSD) on the IBM Storage Fusion Data Foundation storage nodes. A OSD is an

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

entity of the storage managed by Ceph.

- Behind the scenes, the OSDs are mapping the logical PVs to physical PVs, representing real physical storage, which is exposed by the storage adapter (or network adapter) and the Linux kernel. The storage/network adapters can be connected to FCP or DASD/FICON storage. For this mapping, IBM Storage Fusion Data Foundation uses the OpenShift Local Storage Operator to interact with the physical storage.
- IBM Storage Fusion Data Foundation is implementing redundancy by ensuring that each block of a logical PV, which is used on the compute node, is stored within at least three OSDs on storage nodes. IBM Storage Fusion Data Foundation also provides capabilities for seamless operation, such as storage monitoring. All this is orchestrated by Rook. Rook creates several management components, which coordinate the life cycle of the OSDs, that is, Monitors (MON), Metadata servers (MDS) and the Storage Manager (MGR). Rook also provides an object gateway (RGW) for object storage.

The IBM Storage Fusion Data Foundation storage nodes also expose an alternate object storage API via Noobaa's MultiCloud Gateway (MCG). Noobaa MCG exposes an S3-compatible interface, which is a well-known industry best practice for application workloads interacting directly with object storage. The object storage can be spread out to multiple providers and Noobaa takes care of federating the data to a virtualized object storage. Noobaa presents capabilities to administer object storage and define buckets for access control and account management. Namespacing can be used to separate data between applications. A cache can be applied for optimized performance. In the case of NooBaa, Ceph and OSDs are not involved.

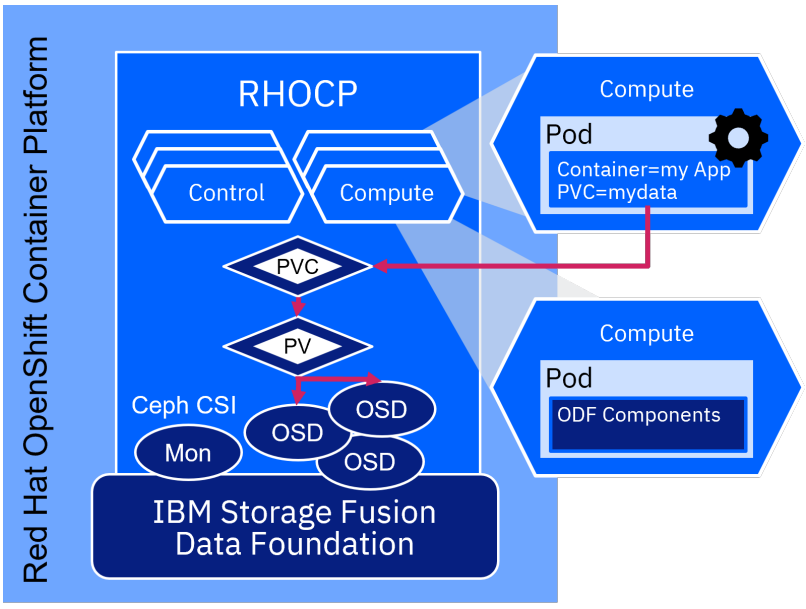

Figure 3.2-5: Internal architecture of IBM Storage Fusion Data Foundation

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

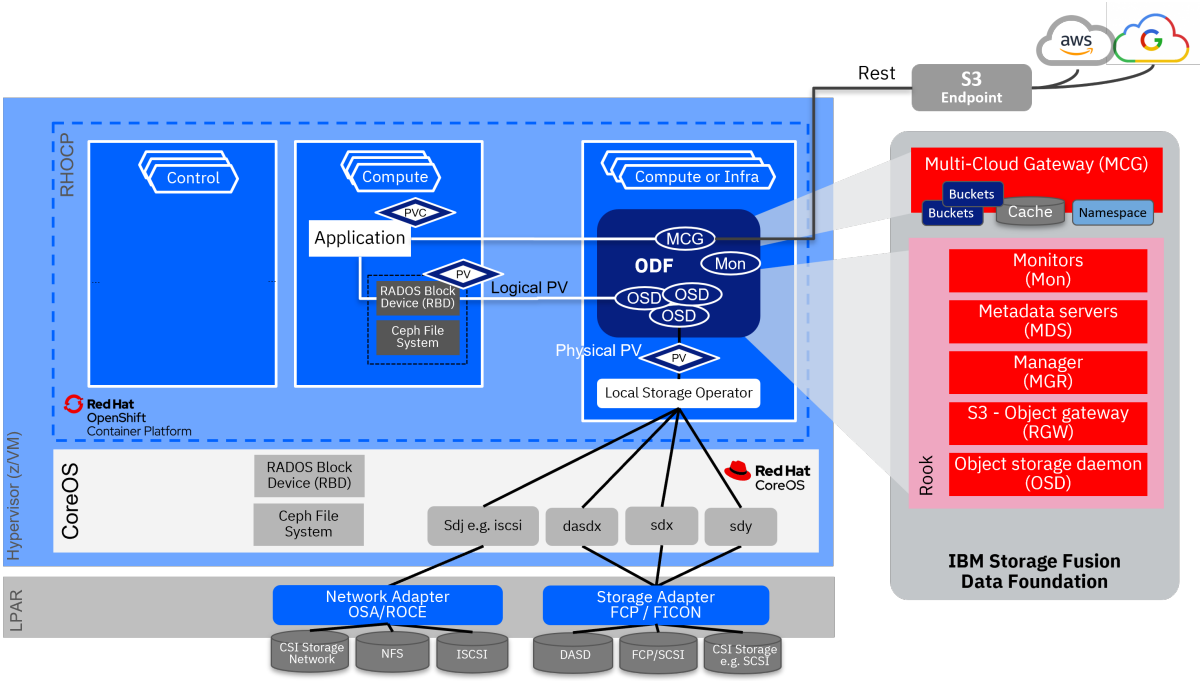

Figure 3.2-6: Internal architecture of IBM Storage Fusion Data Foundation (detailed view)

3.3 Deployment topologies

IBM Storage Fusion Data Foundation can be deployed in two different modes:

- **Internal mode**: IBM Storage Fusion Data Foundation is installed within a single OpenShift cluster. This provides highly scalable, enterprise grade storage, which is fully integrated into the OpenShift lifecycle, monitoring, and management.
  - In a simplified mode, application compute pods and IBM Storage Fusion Data Foundation pods can be scheduled on the same nodes. In this case, compute and storage infrastructure scale together within the same cluster. This setup is optimized for simplicity of management.
  - As an alternative, a more balanced mode allows the deployment of application pods and IBM Storage Fusion Data Foundation pods on different nodes. This implies that compute hosts and storage hosts scale independently within the same cluster. With this, a more balanced deployment can be achieved.
- **External mode**: IBM Storage Fusion Data Foundation pods run on one or more OpenShift clusters, while the actual storage is provided from an external Ceph cluster (IBM Storage Ceph). This implies that lifecycle, monitoring, and management of OpenShift Container Platform and IBM Storage Fusion Data Foundation are decoupled. Compute and storage infrastructure scales independently in entirely different clusters. This approach is optimized for scale and performance and intended for on-premises deployments.

NOTE	The external IBM Storage Ceph will need to be deployed on an x86 bare metal architecture platform.

NOTE	The advantage of internal mode is that you only need one single cluster to run both applications and storage infrastructure.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

NOTE The big advantage of external mode is that one single external IBM Storage Ceph cluster, can provide storage infrastructure to multiple consuming OpenShift clusters at the same time. It is also possible to cross boundaries of hardware architecture as the OpenShift cluster can run on IBM zSystems, while the Ceph cluster is running on x86 hardware.

NOTE External mode is support on IBM LinuxONE, but currently not supported on IBM zCX.

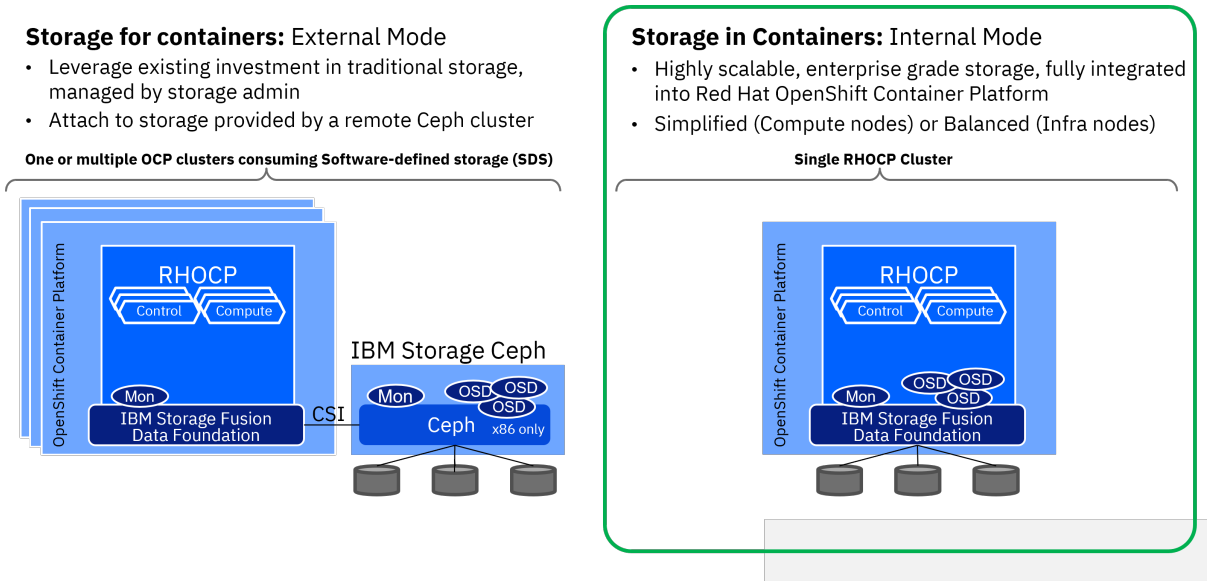

Figure 3.3-1: Internal and external Mode

With IBM Storage Fusion Data Foundation it is possible to build distributed storage architectures and mix on-premise with cloud to create a truly hybrid storage topology.

In external mode it is possible to connect IBM Storage Fusion Data Foundation to a remote Ceph storage cluster. Such an external Ceph cluster could be deployed on bare metal x86 system.

But even without external mode, there is another interesting option to connect IBM Storage Fusion Data Foundation deployment to external storage; Noobaa Multi Cloud Gateway allows the inclusion of 3rd party object storage. This object storage can be provided by a cloud and on-premise solution. It needs to adhere to the popular S3 API. For a developer it is completely transparent where the object storage comes from, where and how it got deployed. Noobaa facilitate the magic behind the developer's object storage class PVCs and that storage class can be use just the same as any local attached storage.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

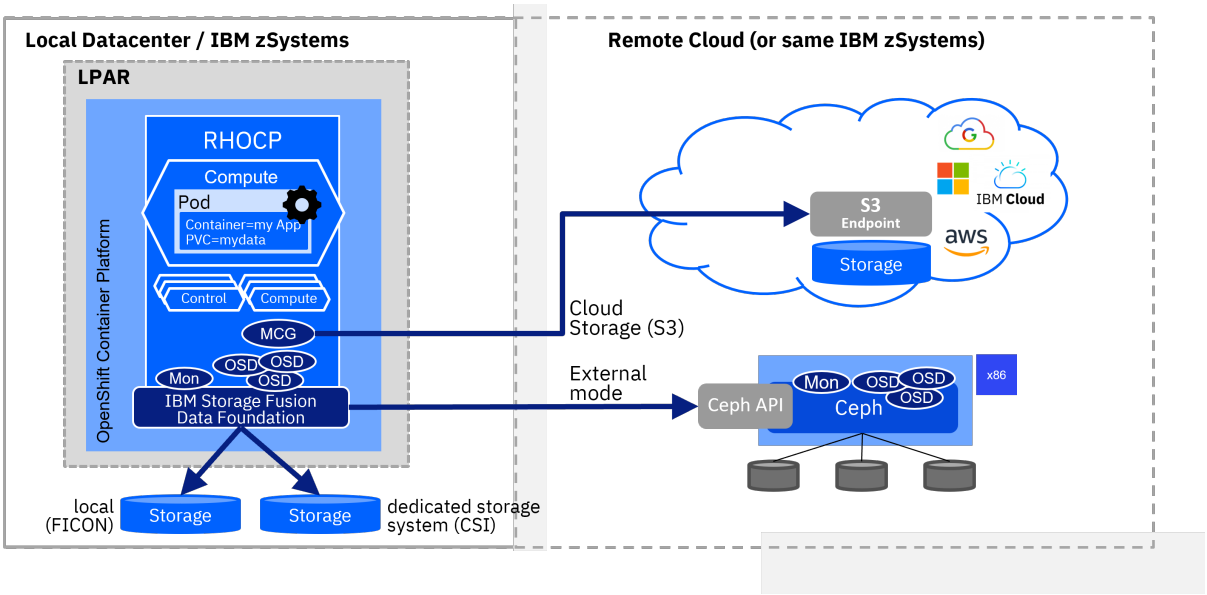

Figure 3.3-2: Extended topologies

3.4 Hardware requirements for installation

To install IBM Storage Fusion Data Foundation on IBM zSystems and IBM LinuxONE, the following hardware requirements must be met:

- As prerequisite, an OpenShift Container Platform cluster must be installed. This implies that requirements for an OpenShift installation must also be met.
  - For a minimal deployment on IBM zSystems hardware, approximately 6 IFLs (physical cores referred to as Integrated Facility for Linux) are required.
  - For each OpenShift control node, 16 GB of Memory and 120 GB of storage are required.
  - For OpenShift itself, a minimum of 2 compute nodes is required, each consuming 8 GB of memory and 120 GB of storage. But if IBM Storage Fusion Data Foundation Local Storage O is installed afterwards, a minimum of 3 compute nodes is necessary instead. Optionally, infrastructure nodes can be installed.
- IBM Storage Fusion Data Foundation needs to deploy 3 storage nodes in the OpenShift cluster. Storage nodes can be deployed on either 3 compute or 3 infrastructure nodes in the same OpenShift cluster.
  - IBM Storage Fusion Data Foundation adds at least the need for one more IFL to the entire deployment. Depending on the workload up to 4 IFLs can become necessary.
  - Each storage node requires another 64 GB Memory and at least 512 GB of storage.

NOTE For IBM zCX similar hardware needs apply. Please note that zCX leverages zIIPs instead of IFLs as processing units.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

NOTE	IBM Storage Fusion Data Foundation storage nodes can be either installed on compute nodes or on infrastructure nodes within the same cluster. When installing on infrastructure nodes, the application workload on the compute nodes can scale independently of the storage nodes. When installing on compute nodes, the application and the available storage scale together, and the management is bit easier.
NOTE	The described environment above addresses a minimal hardware setup, which is limited in terms of capacity and performance. For CPU intense and network intense workload the hardware needs increase.
NOTE	Depending on the subscription, IBM Storage Fusion Data Foundation can be deployed via Red Hat Operator Hub (using the branding Red Hat OpenShift Data Foundation) or by installing via IBM Storage Fusion from IBM (using the new branding IBM Storage Fusion Data Foundation)

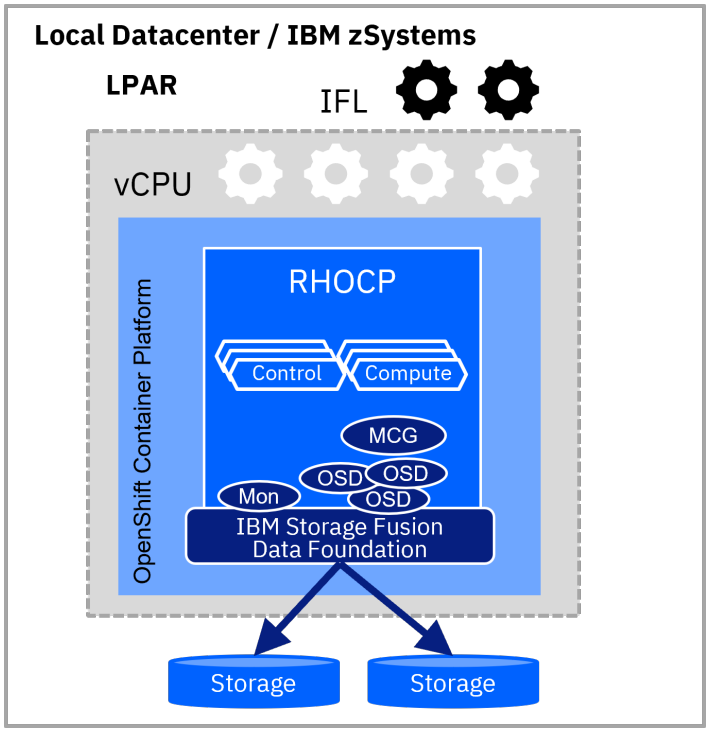

Figure 3.4-1: Hardware setup

3.5 Using storage in containerized applications

The following example shows what a typical setup of a sample application using persistence for its data looks like.

In this example the application uses a Postgres database. Both the application itself as well as the Postgres DB run in containers inside pods within the same OpenShift Container Platform cluster.

- In a first step, the PostgreSQL DB gets deployed in OpenShift. This can be done for example via ImageStream, which is pulled from quay.io or docker hub.
- Next, a persistent volume claim (PVC) for that PostgreSQL is defined. PVC details can be specified

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

in a YAML File and include access mode and storage class. The referenced storage class is provided by IBM Storage Fusion Data Foundation. PostgreSQL can run in its pod without worrying about how the actual physical storage is provided. IBM Storage Fusion Data Foundation takes care of mapping the specified PVC to real physical storage internally.

- Finally, the sample application is deployed in OpenShift as well. In this example the demo application interacts with the database via Web App Endpoints exposed by the PostgreSQL, which has been deployed and configured as described above.

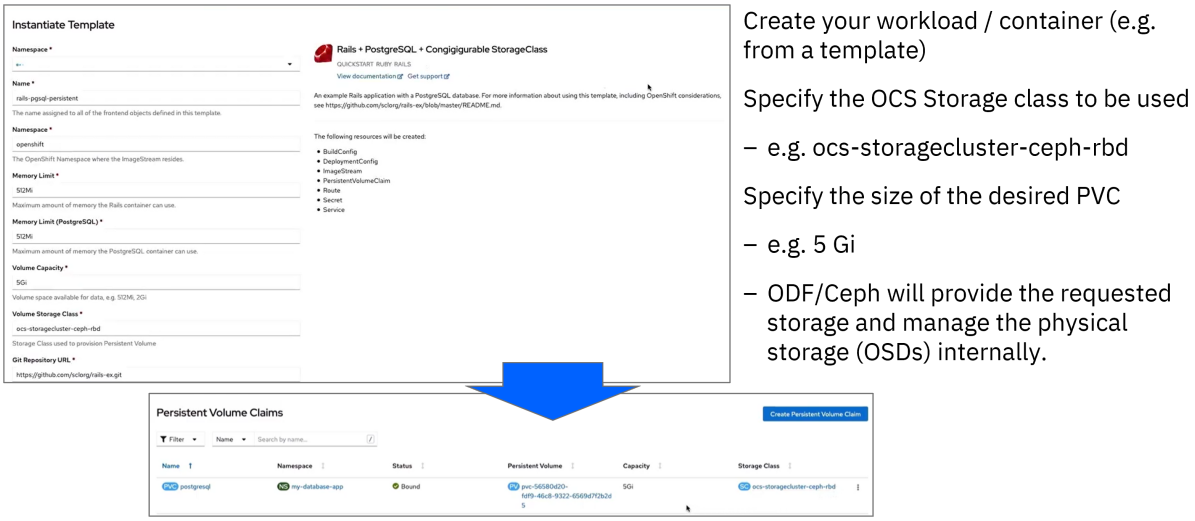

Figure 3.5-1: Using storage in applications (I)

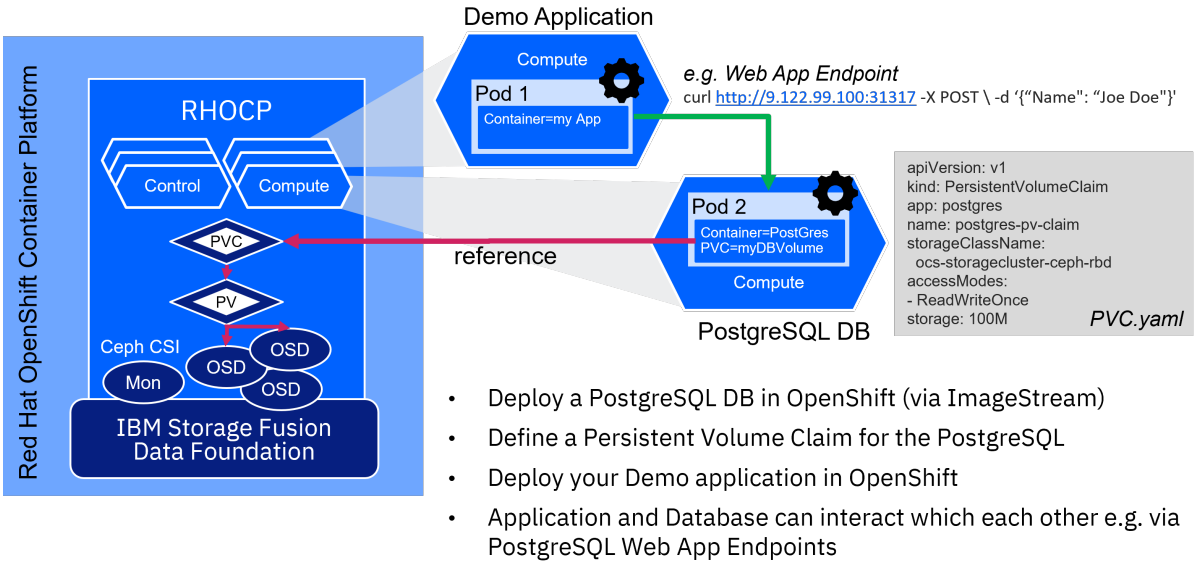

Figure 3.5-2: Using storage in applications (II)

3.6 Considerations for business continuity and resilience

In addition to the setup of a full production environment, it is essential to consider resiliency.

- To protect against logical failures, it is necessary to facilitate **backups**. Logical failures include bugs, malicious software, or even administration mistakes. A backup provides the option to

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

restore the application by a point-in-time copy of its data. For IBM Storage Fusion Data Foundation, the OpenShift API for Data Protection (OADP) is used for backup. OADP is a operator on its own. Its implementation is based on the open-source project Velero and uses CSI Snapshots or Restic under the hoods. Backup images are being saved as object storage in the cloud. It is important to emphasize that in the context of OADP, backups are snapshots of the application's persistent volumes data and all the Kubernetes meta data belonging to the application's namespace. This makes a backup always scoped by a namespace. The deployment of the entire OpenShift Container Platform cluster itself and its corresponding configuration, like the etcd database are not part of OADP backup functionality.

- Obviously, resiliency not only covers the data protection itself, but also the **high availability** of the entire system, including failover and **disaster recovery** in the case of major disruptions. IBM zSystems platform best practices can be applied here.
  - A deployment of a single OpernShift cluster can be stretched across datacenters within proximity. Network latency needs to be below 10 ms for round-trips. This setup addresses outages due to local system or hardware failures, as the individual nodes of the OpenShift cluster and the IBM Storage Fusion Data Foundation storage nodes are isolated from each other (in separate LPARs, hardware units or even data centers). The minimum of two control, compute, and storage nodes must always be maintained to ensure continuous operation.
  - IBM Storage Fusion Data Foundation can also be set up in a active / passive setup, where two datacenters are located further apart to cover broader regional disasters. Each datacenter has its own OpenShift cluster and its own IBM Storage Fusion Data Foundation deployment. One cluster is active, while the other remains as stand-by for use in case of a disaster. The 2 seperate clusters are orchestrated by Red Hat Advanced Cluster Manager (RHACM). The persistent volumes in the stand-by system are kept up to date by Ceph's asynchronous replication.

3.7 Comparing IBM Storage Fusion Data Foundation with other storage solutions

There are multiple options to add persistence to containerized workload. The needs of customers and applications differ. Besides IBM Storage Fusion Data Foundation, other options for providing persistence to an OpenShift cluster on IBM zSystems are listed in the table below.

With IBM Storage Fusion a comprehensive suite for storage infrastructure has been introduced.

Table 1. Storage options (I)

Solution	Suitable usage scenario	Storage classes	Access options
Local Storage Operator	Operator used to create PVs from Local Volumes with SCSI-over-FCP and DASD volumes		
iSCSI	manually provisioned		ReadWriteOnce, ReadOnlyMany

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

Solution	Suitable usage scenario	Storage classes	Access options
Fibre Channel	manually provisioned		ReadWriteOnce, ReadOnlyMany
IBM Block Storage CSI driver	ISV provided driver for storage attachments using iSCSI and FCP		
NFS	Generic and commonly used network attached persistent shared storage	file	ReadWriteMany
IBM Storage Scale	Extremely scalable IBM product for file and object storage access. Deployed external to the OpenShift cluster	file, object	ReadWriteOnce, ReadWriteMany
IBM Cloud Native Storage Access (CNSA)	a cross cluster mount capability to attach a RHOCP cluster to an IBM Storage Scale cluster	file, object	ReadWriteOnce, ReadWriteMany
IBM Storage Ceph	Scalable and resilient open-source storage for storage access (only available for x86)	object, file, and block	ReadWriteOnce, ReadWriteMany
IBM Storage Fusion Data Foundation	Versatile multi-purpose storage access based on Ceph. Tightly integrated into the OpenShift ecosystem	object, file, and block	ReadWriteOnce, ReadWriteMany
IBM Storage Fusion	IBM storage offering, which wraps IBM Storage Scale and IBM Storage Fusion Data Foundation in a unified install and configuration experience for container workload	object, file, and block	ReadWriteOnce, ReadWriteMany

Table 2. Storage options (II)

Use Case	Product offering	workload
Unified Software Defined Storage for OpenShift and Kubernetes workloads	IBM Storage Fusion based on IBM Storage Fusion Data Foundation	Container based workload (OpenShift)
Unified Software Defined Storage (File/Block/Object) for hybrid workloads	IBM Storage Ceph (For IBM zSystems: IBM Storage Fusion Data Foundation	Generic workload
Distributed File and Object storage for AI/Data intense workloads	IBM Storage Scale	Generic workload
Object storage for analytics and active archive workloads	IBM Storage Ceph (For IBM zSystems: IBM Storage Fusion Data Foundation	Generic workload

From that list above, NFS, IBM Storage Scale and IBM Storage Fusion Data Foundation are the most

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

relevant solutions, especially if storage needs to be share across applications (ReadWriteMany).

- NFS is well-known and popular. NFS can be integrated with the OpenShift platform. It has several limitations, such as supporting file storage only or being constrained with regards to scalability and high availability.
- IBM Storage Scale adds valuable capabilities in terms of parallel file systems and data sharing between architectures. A strength of IBM Storage Scale is its amazing scalability in terms of both performance and amount of data stored. This makes Storage Scale a preferred option for AI applications and data lakes across the globe. Spectrum scale is not limited to containerized workload and container-native storage, but also finds its use in classic legacy workload. Data tiering is another signature feature of IBM Spectrum Scale. IBM Spectrum scale also provides sophisticated patterns for high availability and disaster recovery using IBM technologies like IBM Geographically Dispersed Parallel Sysplex (GDPS). IBM Spectrum Scale focuses on general purpose file storage, but recently also added S3 Object Storage access to Cloud providers Spectrum Scale is deployed in a cluster of its own, external to the OpenShift cluster, which runs the containerized applications.
- IBM Storage Fusion Data Foundation supports all three common storage classes, file, block and object storage. One of the biggest advantages of IBM Storage Fusion Data Foundation is that it is closely coupled with the OpenShift ecosystem and its user experience. This makes is an excellent choice for general-purpose storage for containerized workload. Under the hood, it is based on well-known and well-accepted open-source components such as Noobaa, Rook, and Ceph. With Noobaa, a storage infrastructure can federate to build true hybrid cloud topologies, which span multiple cloud vendors as well as on-premises deployments. This allows for very flexible data tiering. Backup, disaster recovery and high availability are based on a Red Hat software stack, which includes Red Hat Advanced Cluster Manager (RHACM) and Open API for Data Protectiom (OADP).
- IBM Storage Fusion is IBM’s strategic offering for container-native storage. It integrates IBM Storage Fusion Data Foundation and IBM Storage Scale into a unified install and configuration experience. Additional data platform, protection, and cluster services include further value on top of the core storage platforms.

NOTE	IBM Storage Fusion Data Foundation can be licensed by IBM as part of IBM Storage Fusion. Nevertheless, subscriptions for Red Hat OpenShift Data Foundation Essentials and Red Hat OpenShift Data Foundation Advanced can also be purchased by Red Hat in parallel.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

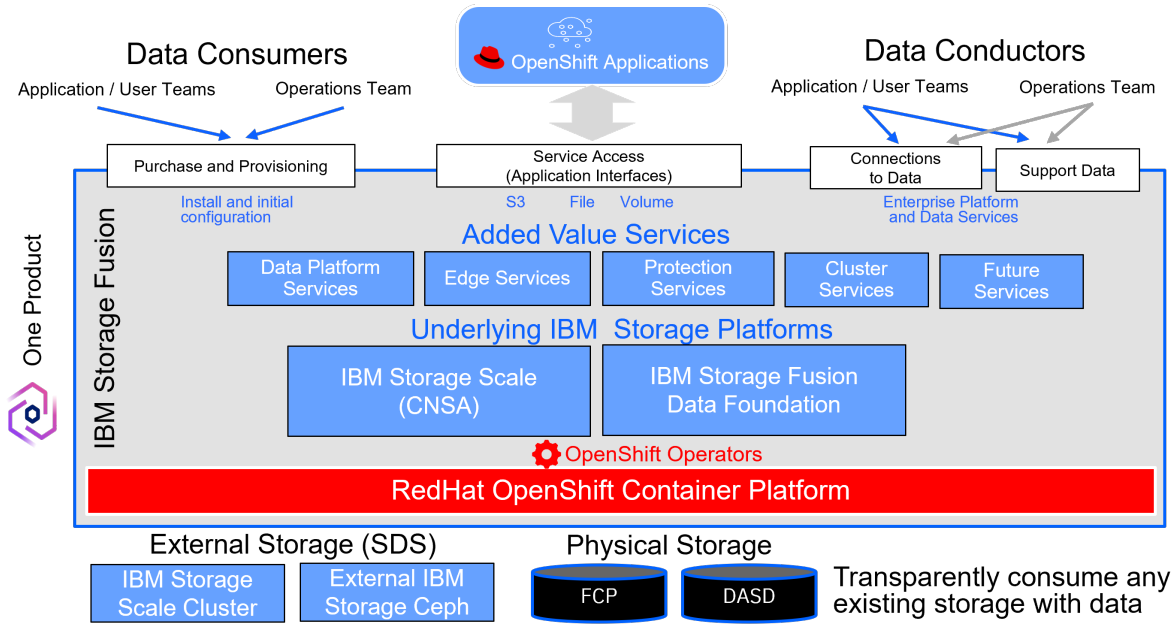

Figure 3.7-1: IBM's software stack for container-native storage based on IBM Storage Fusion

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

Chapter 4. Design considerations

This section describes in detail how this reference architecture addresses key design considerations when planning a cluster running IBM Storage Fusion Data Foundation. The recommendations in this reference architecture were developed in conjunction with Red Hat field consultants, engineers, and product teams.

4.1 Storage

4.1.1 Choosing storage devices for the Local Storage Operator

IBM Storage Fusion Data Foundation can consume physical storage, which is locally attached and accessed using the OpenShift Local Storage Operator (which therefore is a prerequisite for installing IBM Storage Fusion Data Foundation).

Considerations to keep in mind before deploying IBM Storage Fusion Data Foundation include:

- For IBM zSystems and IBM LinuxONE, both FCP/SCSI and DASD/FICON devices are supported.
- For IBM IBM z/OS Container Extension (zCX), only VSAM LDS (dasd) and NFS are supported. FCP and iSCSI are not supported since they are not available on z/OS.
- For each storage node at least one physical disk needs to be assigned. Multiple disks can be assigned when larger storage capacity is desired. But each storage node must have an identical disk configuration and a growing number of disks will increase the demand for the processing power assigned to IBM Storage Fusion Data Foundation.
- The local disks need to be made available to the Hypervisor. Disks and partitions are then automatically discovered during installation of IBM Storage Fusion Data Foundation.
  - Please note: for DASD, the disk must be formatted and partitioned (see section 4.1.2).
- The storage devices must be empty before they can be attached to IBM Storage Fusion Data Foundation.

The following images show the access to physical storage in a z/VM and KVM environment. These details are transparent to IBM Storage Fusion Data Foundation because the Local Storage Operator takes care of platform-specific aspects. For instance, disks are automatically discovered during storage node setup.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

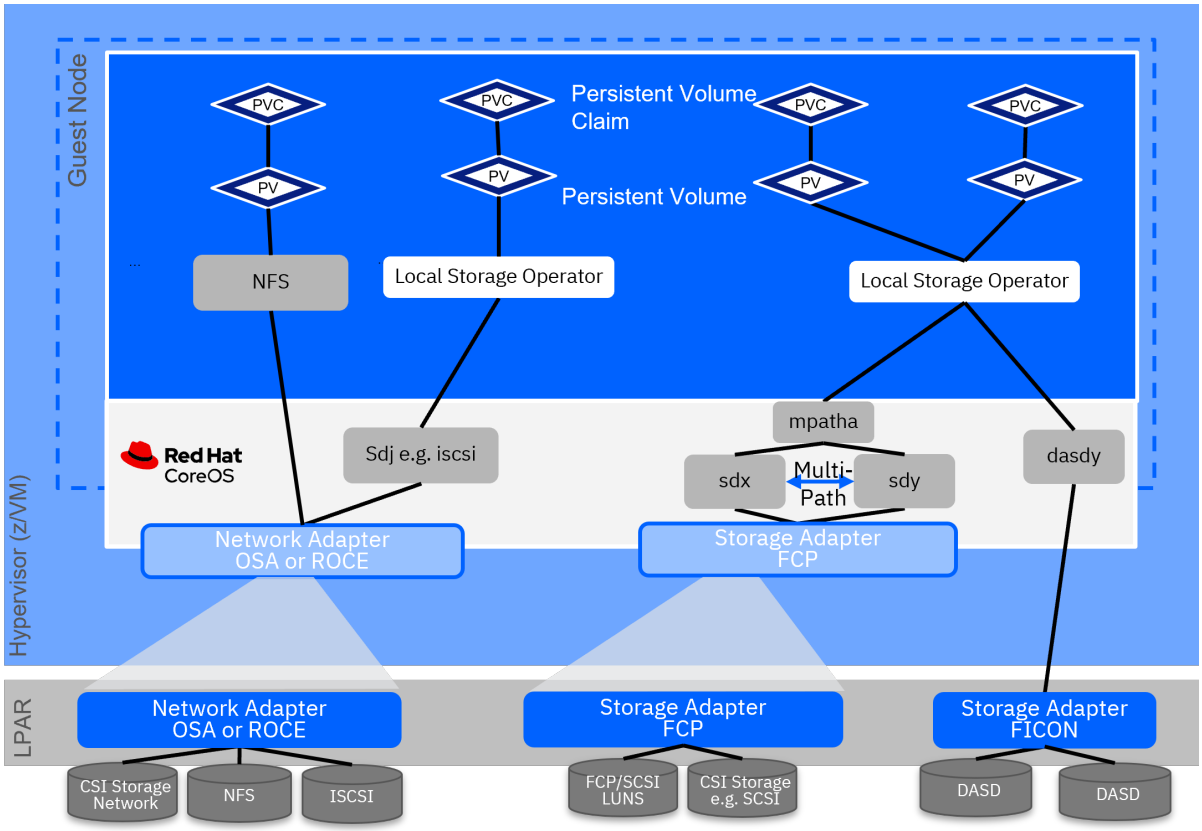

Figure 4.1-1: Local Storage Options using z/VM

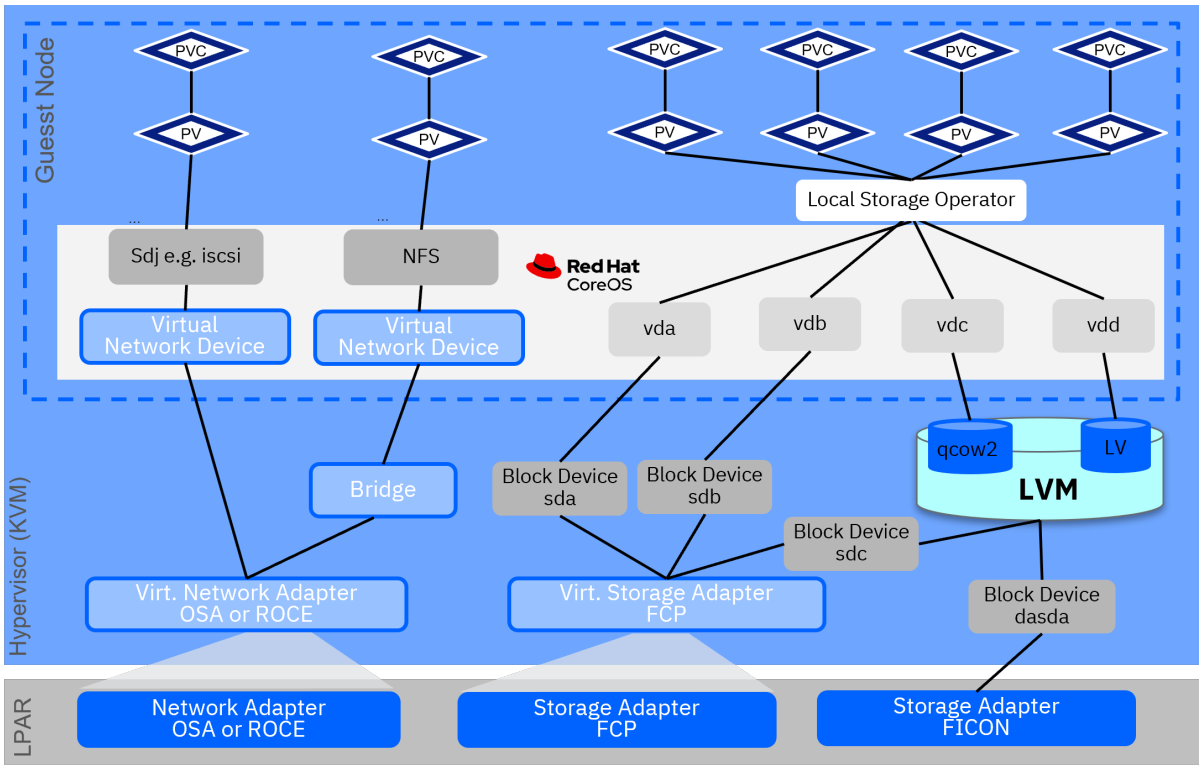

Figure 4.1-2: Local Storage Options using KVM

During the planning phase it is worth comparing the specific details of the FCP and DASD devices as described in this section of official documentation.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

4.1.2 Steps specific for DASD devices

Once the DASD devices are available to z/VM guests, complete the following steps from the compute or infrastructure node on which an OpenShift Data Foundation storage node is being installed.

1. **Enable Device**

From the compute node run below command to enable the device.

```
sudo chzdev -e <DASD ID>
```

Now the device will be listed and can be checked using the `lsdasd` and`lsblk` commands.

2. **Low-level format the Device**

DASD need to be formatted using below command specifying the disk name, e.g., `dasdb`.

```
sudo dasdfmt /dev/dasdb -b 4096 -p -y -F
```

Here `dasdb` is the assigned device letter once it is enabled.

3. **Partition the Device**

Create full partition using the command below.

```
sudo fdasd -a /dev/dasdb
```

Once these steps are done, the device will be available during IBM Storage Fusion Data Foundation deployment with device name `/dev/dasdb1`.

4.1.2 Deployment using IBM Block CSI driver instead of Local Storage Operator.

Instead of attaching disks to each compute node and then deploying the Local Storage Operator, the IBM Block CSI driver can be deployed from the operator hub. The CSI driver can provision the necessary storage devices from the configured storage system. This driver allows the admin to easily scale up the cluster size using the `add capacity` option from the administration user interface of IBM Storage Fusion Data Foundation.

NOTE	`add capacity` can be found in the administration user interface menu using `Operators -> Installed Operators -> Openshift Data Foundation -> Storage System -> ocs-storagecluster-storagesystem -> Actions -> Add Capacity`

To discover and mount dynamically provisioned volumes automatically, N-Port ID Virtualization (NPIV) should be enabled for those compute (or infra) nodes on which the IBM Storage Fusion Data Foundation storage nodes are installed on this cluster.

1. Follow these instructions to deploy "IBM Block CSI driver" and to configure it with the storage system after creating the storageclass.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

2. Deploy IBM Storage Fusion Data Foundation using the storageclass that you created previously.

4.2 Networking

For IBM Storage Fusion Data Foundation, there are no particular considerations for configuring the networking adapters on IBM zSystems.

Both OSA and RoCE are supported.

Nevertheless, one critical aspect to keep in mind is the network bandwidth: you need at least 10 Gbs networking speed to achieve reasonable performance and stability for your cluster and workload.

4.3 Virtualization and hypervisors

On IBM zSystems, IBM Storage Fusion Data Foundation supports both KVM and z/VM as hypervisors. Although the interaction with the hypervisor is mainly covered by the OpenShift Container Platform itself, there are a couple of considerations with regards to IBM Storage Fusion Data Foundation. Because IBM Storage Fusion Data Foundation is a storage-centric product, which requires high I/O throughput and responsiveness compared to OpenShift Container Platform which is a generic platform, it is essential to have native performance on disk devices and to keep the high level of disk utilization.

With z/VM, both FCP SCSI and DASD disks are in native/passthrough mode, which means that guests can utilize native performance of those devices. Enabling HyperPAV (aliases) for ECKD DASD disks would achieve higher throughput and overall better I/O performance. This might result in higher CPU utilization, which is acceptable for a storage-oriented product such as IBM Storage Fusion Data Foundation.

With KVM, besides the passthrough mode, disks on the KVM host could be partitioned into libvirt/QEMU image files and be used as disks for KVM guests. In either cases, depending on the type of libvirt/QEMU driver being used - be it VirtIO Block (virtio-blk) or VirtIO SCSI (virtio-scsi) - the naming of the disk device in the guest might be different, where /dev/vdX for VirtIO Block and /dev/sdX for VirtIO SCSI. Using passthrough mode for disks in KVM is preferred, especially for the case of using the disks as block devices for OSD disks. Using I/O threads could improve the performance of I/O for disk devices, at a cost of higher CPU and memory utilization. Under certain events of the Ceph nodes performing rebalancing or recovering, the network load spikes might happen for extended period of time, depending on the size of the data sets. It is advisable to use multiple queues for VirtIO network interfaces, ideally matching the number of the guest's CPUs.

In general:

- With z/VM, enable HyperPAV.
- With KVM, depending on VirtIO drivers, the deployment plan needs to consider disk device naming convention.
- With KVM, enable I/O threads, as many as number of CPUs.
- With KVM, use multiple queues for VirtIO network interfaces, as much as number of CPUs.
- Use KVM passthrough mode for disks to achieve highest performance.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

- If not using passthrough disks, use KVM VirtIO Block driver for highest performance (being deprecated upstream).
- If not using passthrough disks, use KVM Virtio SCSI driver for highest compatibility and having a high number of (OSD) disks attached to the guest.

4.4 Deployment methods and tools

IBM Storage Fusion Data Foundation can be deployed on IBM zSystems and IBM® LinuxONE as well as within IBM z/OS Container Extensions (zCX).

Depending on the subscription, IBM Storage Fusion Data Foundation can be deployed in different ways:

- Installing via Red Hat Operator Hub (using the branding Red Hat OpenShift Data Foundation)
- Installing via IBM Storage Fusion from IBM (starting IBM Storage Fusion version 2.4 and using the new branding IBM Storage Fusion Data Foundation)

In any case a fully operational Red Hat OpenShift Container Platform cluster needs to be prepared and sufficient storage must be attached locally (see documentation).

4.4.1 Installation via Red Hat Operator Hub

IBM Storage Fusion Data Foundation is deployed via operators, which are available from the Red Hat Operator Hub.

Each release of IBM Storage Fusion Data Foundation can be based on a OpenShift Container Platform with either the same matching release number, or with one or two back level releases.

The deployment of IBM Storage Fusion Data Foundation within an OpenShift cluster creates:

- Noobaa,
- OpenShift Data Foundation Metrics,
- OpenShift Data Foundation Operator,
- Rook-Ceph Operator

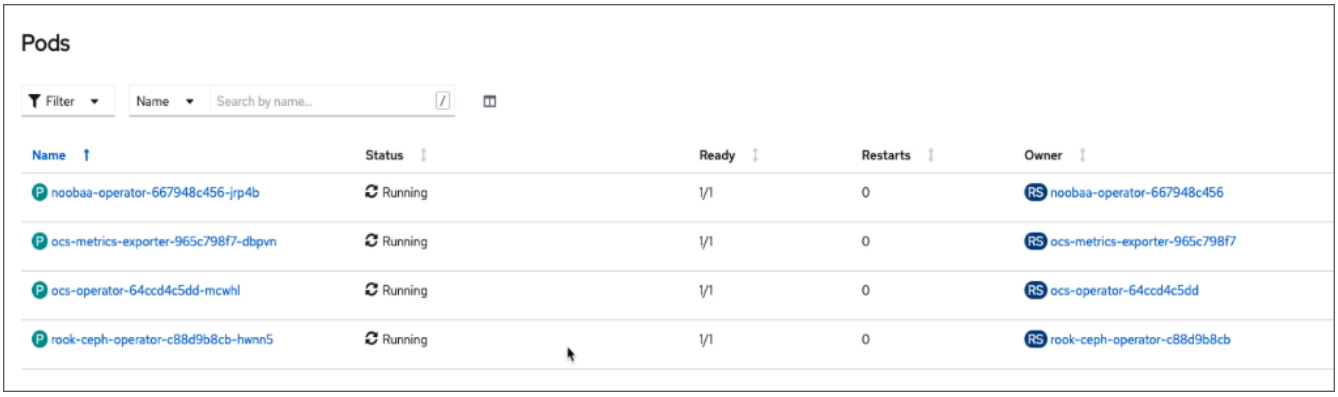

Figure 4.4-1: Deploying IBM Storage Fusion Data Foundation

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

After IBM Storage Fusion Data Foundation is deployed, the next step is to create a storage cluster.

- The internal attached storage devices are selected. As the Local Storage Operator is required to connect to the devices, that operator gets installed as part of this step as well.
- The OpenShift cluster nodes, which will be used as storage nodes, can be selected. A minimum of three nodes is required.
  - The available disks which can be used are discovered automatically.
    - There must be a minimum of one raw block device per node.
    - All disks must be empty and have a minimum size of 100 GiB.
    - Disks can be either DASD or FCP devices.
  - Storage classes for use by applications are created during this step. The default naming of the IBM Storage Fusion Data Foundation storage classes are:
    - ocs-storagecluster-ceph-rbd
    - ocs-storagecluster-cephfs
    - openshift-storage.noobaa.io
    - ocs-storagecluster-ceph-rgw

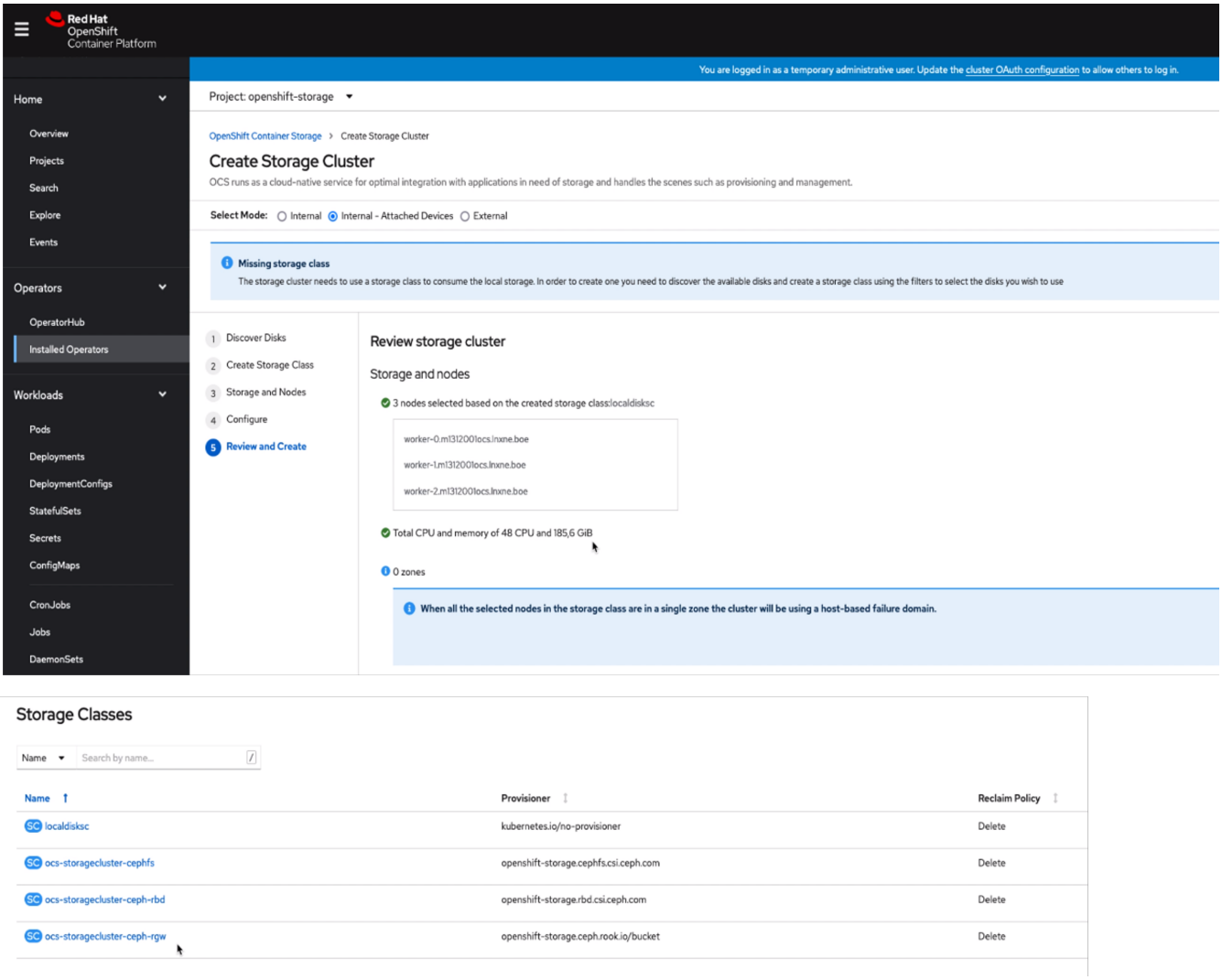

Figure 4.4-2: Deploying IBM Storage Fusion Data Foundation (III)

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

4.4.2 Installation via IBM Storage Fusion

It is also possible to install IBM Storage Fusion Data Foundation via the IBM Storage Fusion user interface. A new operator wraps the existing (Red Hat) OpenShift Data Foundation operators. The advantage of IBM Storage Fusion is to provide a common install and configuration experience for all underlaying storage infrastructure options regardless if IBM Storage Scale or IBM Storage Fusion Data Foundation is being used.

Once installation is complete, there is no functional difference between Red Hat OpenShift Data Foundation or IBM Storage Fusion Data Foundation. Nevertheless IBM Storage Fusion will provide a growing number of additional data services on top of IBM Storage Fusion Data Foundation (like a new install experience).

For more details on IBM Storage Fusion, please refer to the documentation: Documentation for IBM Storage Fusion

4.5 Sizing and planning the hardware infrastructure

4.5.1 Hardware requirements

The sizing for an infrastructure running IBM Storage Fusion Data Foundation needs to reflect the requirements of the underlying OpenShift Container Platform cluster, as well as the IBM Storage Fusion Data Foundation deployment.

From a licensing perspective, the subscriptions are calculated per used IBM LinuxONE IFL (or per zIIP instead of IFLs for IBM zCX).

The key hardware requirements are listed in the table below.

		Minimal Deployment		Typical Deployment	
OpenShift Container Platform					
IFL in total		6 IFLs		7 IFLs	
Control Plane	vCPU	**3 x**	4 vCPU	**3 x**	8 vCPU
	Memory	**3 x**	16 GB	**3 x**	16 GB
	Storage	**3 x**	120 GB	**3 x**	120 GB
Compute Node	vCPU	**Min 3 x**	2 vCPU	**Min 3 x**	4 vCPU
	Memory	**Min 3 x**	8 GB	**Min 3 x**	16 GB
	Storage	**Min 3 x**	120 GB	**Min 3 x**	120 GB
IBM Storage Fusion Data Foundation (additional resources added to the Worker or Infra Node on which deployed)					
IFL in total		1 IFL		3 IFL	
Storage nodes	vCPU	**3 x**	10 vCPU	**3 x**	14 vCPU
	Memory	**3 x**	24 GB	**3 x**	34 GB
	Storage	**3 x**	512 GB	**3 x**	512 GB

Figure 4.5-1: Hardware requirements

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

4.5.2 Using infrastructure or compute nodes

IBM Storage Fusion Data Foundation storage nodes can be either installed on compute nodes or on infrastructure nodes within the same OpenShift cluster.

When installing on infrastructure nodes, the application workload on the compute nodes can scale independently of the storage nodes used by IBM Storage Fusion Data Foundation. A further consideration is that infrastructure nodes will not need additional licenses for Red Hat OpenShift Container Platform.

When installing on compute nodes, the application and available storage scale together, and the management is bit easier.

4.5.3 Using external mode

If the external mode for IBM Storage Fusion Data Foundation is being used, a bulk of the CPU consumption for managing the stored data is being offloaded to the external Ceph cluster. For the stub of IBM Storage Fusion Data Foundation, which is running within the OpenShift cluster only the minimal amount of hardware prerequisites is needed (as specified in table 4.5-1).

4.5.4 Configuring local attached storage

When deploying IBM Storage Fusion Data Foundation, a minimum of 3 storage nodes is needed. But there is flexibility on the number of locally attached physical storage hardware devices, which are being used.

A minimum of 3 physical storage devices are required. At least one for each storage node. For each device, Ceph will need a logical Object Storage Devices (OSD), which abstracts the hardware underneath. While the capacity of the storage disks is not relevant for CPU consumption of the storage cluster, the number of used OSD drives is extremely relevant:

- 1 – 3 IFLs are roughly needed for IBM Storage Fusion Data Foundation to drive the minimum of 3 OSDs (in addition to the resources needed for RHOCP itself). This implies a minimum of 1 physical storage hardware device and 1 logical OSD per storage node.
- If more physical storage devices are required (e.g. to achieve 1000 TiB +), additional OSDs need to be configured (as each physical storage device must be mapped to only a single OSD). This will increase the demand for further IFLs.
- The need to increase the processing power with a growing number of OSDs is shown in figure 4.5-2.

For resilience it is recommended to distribute OSDs equally across storage nodes. At one point, adding more OSDs will make additional storage nodes necessary (to provide sufficient hardware resources).

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

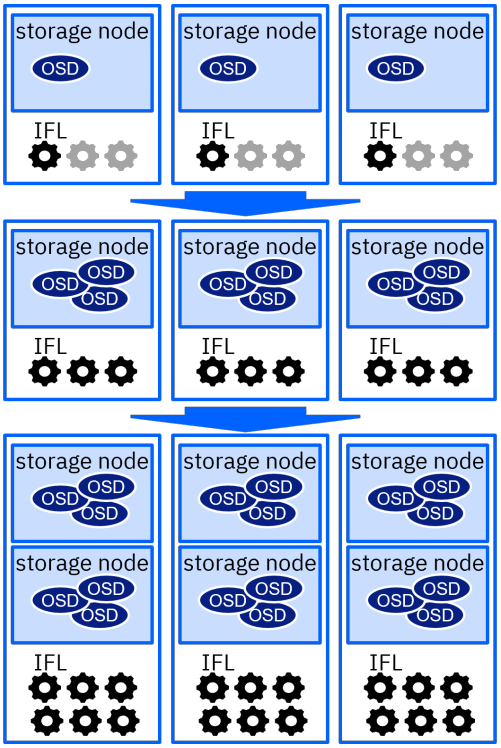

Figure 4.5-2: Scaling IBM Storage Fusion Data Foundation

4.6 Implementing high availability and disaster recovery

High availability (HA) and disaster recovery (DR) are requirements for production deployments using IBM Storage Fusion Data Foundation. A crucial consideration is removing single points of failure on the various layers of the environment. The goal is to avoid planned and unplanned outages and ensure continuous operation.

The considerations for HA and DR include the following:

- All layers of the environment must be included to secure continuous operations of the workload.
- Application development and DevOps need to ensure that new services can be deployed and operated without downtime aka as rolling updates.
- The operational procedures are well defined and consider HA and DR.

IBM virtualization and the IBM zSystems platform provide best practices and mechanisms for setting up highly available environments, including replication, failover and shifting of workload between systems. This makes it a perfect match for ensuring continuous operations (e.g., redundancy of hardware elements). Red Hat OpenShift Container Platform (RHOCP) adds with its cluster-based implementation, the ability for automatic failover of workloads within the same cluster.

On top of the considerations for high availability, there are several topologies, which can be applied to implement disaster recovery and involve a second data center site for redundancy. In case of a disaster the secondary site will take over and run the entire workload. Each site is referred to as an availability zone.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

The first set of topologies is based on Red Hat's OpenShift software stack. High availability and disaster recovery are implemented in software. This implementation adds an abstraction layer, which is application scoped and hardware agnostic.

- High availability with a single stretched cluster across 3 availability zones.
- Regional disaster recovery: 2 separate clusters with asynchronous replication of data and active/passive configuration
- Metro disaster recovery: 2 separate clusters in a active/active configuration and with synchronous replication of data in an external IBM Storage Ceph environment, which is stretched across the data centers.

NOTE	For IBM zSystems and IBM LinuxONE regional DR is a Tech Preview starting ODF 4.12.1. Metro DR is not yet tested and supported.

A hardware based option to achieve disaster recovery is based on IBM's Geographically Dispersed Parallel Sysplex (GDPS) technology, which is implemented based on mirroring of the storage hardware and agnostic to the running applications. GDPS allows synchronous and asynchronous mirroring for metro and global distance. GDPS also adds management capabilities for nodes and network switches.

NOTE	GDPS is supported based on VSAM LDS as local disks and running on IBM zCX.

The different topologies for HA and DR are described in the following sections.

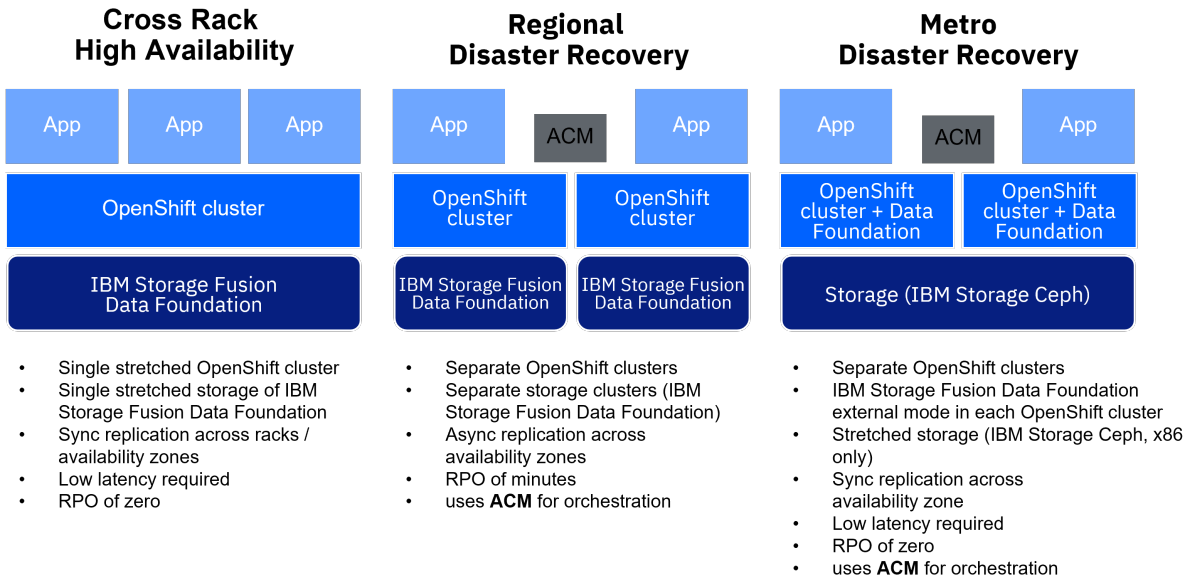

Figure 4.6-1: Topologies for setting up disaster recovery on (logical) software level

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

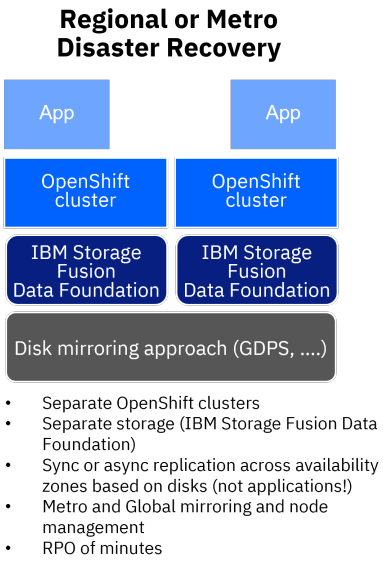

Figure 4.6-2: Topologies for setting up disaster recovery on hardware level

Software Resilience

Layer	DR Options	What is covered in case of disaster	Considerations
Persistent Volume	• IBM Storage Fusion Data Foundation/Ceph metro DR (active/active) • IBM Storage Fusion Data Foundation/Ceph regional DR (active/warm)	all OpenShift PVs are replicated between data centers (focus on application namespace only)	• agnostic to physical storage • Ceph RBD snapshot-based mirroring • VolSync for CephFS
OpenShift cluster Meta data (etcd)	• single-cluster stretch • ACM multi-cluster	restart, reinstall or move to new cluster	• via automation based on OCP configuration
Based on a hardware layer, which can fail. New hardware can be spun up by software orchestration quickly (**re-install software**)			

Hardware Resilience

Layer	DR Options	What is covered in case of disaster	Considerations
Hardware provides fall-backs, which can be used in case of outages (**re-start software**)			
IBM Z CEC & hypervisors	• virtualization	restart, reinstall or move to new CEC/hypervisor	• via automation based on HW / SW configuration
Storage Subsystem	• GDPS Metro • GDPS Global – XRC • GDPS Global – GM *(requires z/OS)*	content of disks is replicated between data centers (focus on physical disks not OpenShift applications)	• **Formal support pending** • agnostic to individual containerized applications / namespaces

Figure 4.6-3: Differences between disaster recovery based on logical software and hardware layers

It needs to be emphasized that the mentioned DR topologies using Red Hat's OpenShift software stack are based on the fundamental assumption that hardware is unreliable and software must implement resilience. As this assumption is not true for IBM zSystems, the combination of resilience in software combined with resilience in hardware results in an extremely reliable overall system.

It is also worth mentioning that while hardware-based resilience is focused on keeping a system alive by all means (using redundancy), the software-based resilience replaces failing instances immediately with new software deployments created fresh from a CI/CD pipeline and automation. This different point of view is often referred to as dealing with precious "pets" versus abundant "cattle".

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

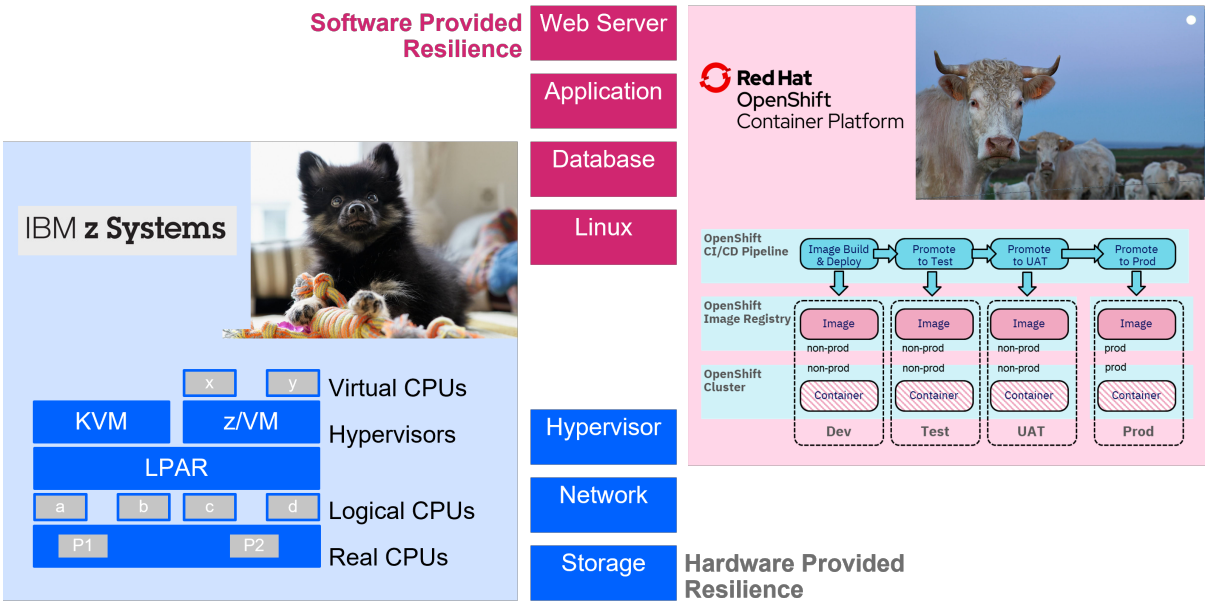

Figure 4.6-4: Differences hardware-based resilience ("pets") versus software-based resilience ("cattle")

4.6.1 Stretching an OpenShift cluster across availability zones

The basic set up to increase resilience of an OpenShift cluster is to stretch its nodes across multiple hardware units (or availability zones).

Thanks to the virtualization of IBM zSystems, high availability can be easily achieved even within a single physical machine by taking advantage of the concept of an LPAR, which provides highly secure hardware isolation. Given that LPAR virtualization is EAL5+ certified, each single LPAR can be considered as a logical hardware unit of its own. Deploying a single OpenShift cluster with IBM Storage Fusion Data Foundation across multiple LPARs allows you to achieve a high available distributed setup by spreading across largely independent logical units. By taking advantage of the redundancy built into IBM zSystems hardware, the deployment on multiple LPARs and hardware units (CECs) is the simplest setup for high availability.

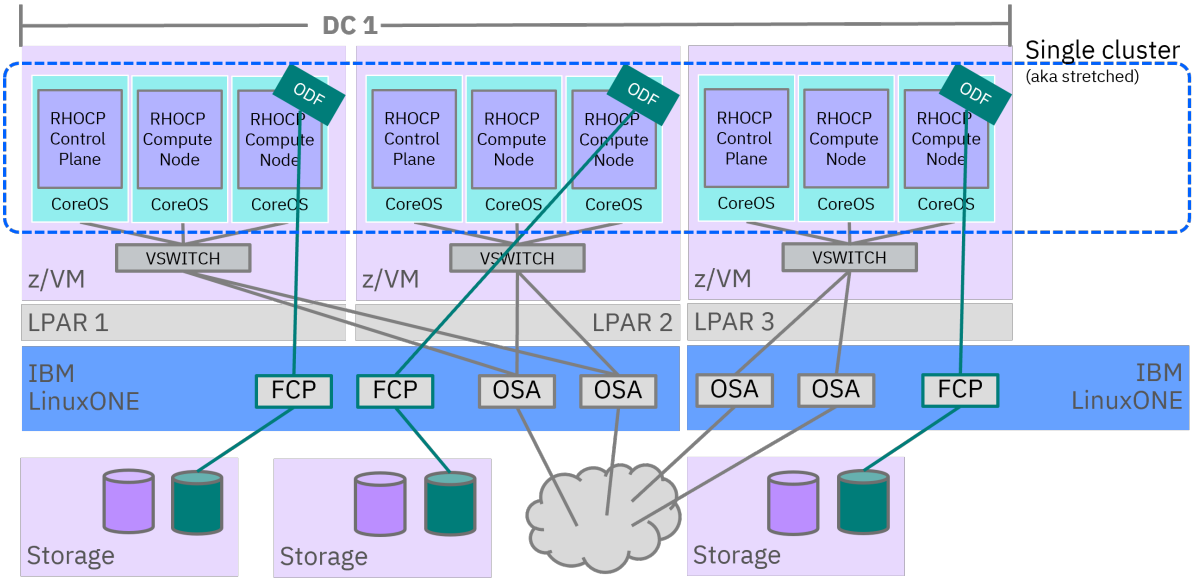

Figure 4.6-5: Single cluster setup for HA within a single data center

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

As soon as disaster recovery is needed in addition to high availability, a second site needs to be established and the nodes of the cluster will be spread across the sites.

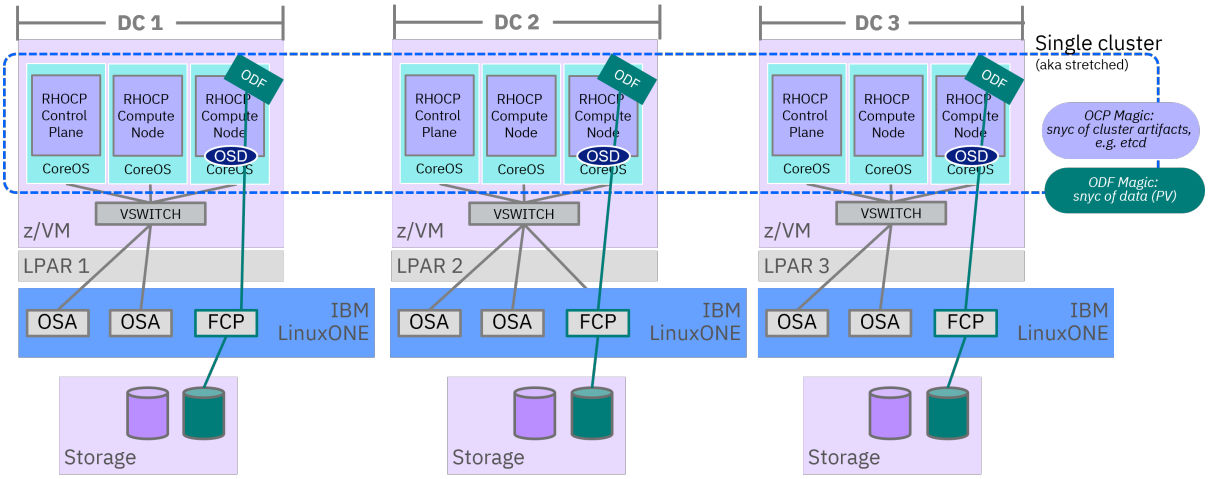

Figure 4.6-6: Single cluster setup for DR across 2 availability zones

There is flexibility on how the nodes of a single cluster can be spread across LPARs, hardware units, and data centers. For IBM Storage Fusion Data Foundation, there is no data replication necessary. Instead, the data in all storage nodes is mirrored to be kept instantly in sync. The overarching control plane of the OpenShift cluster schedules the pods and their applications and PVs across all nodes.

Nevertheless, there is one strict requirement imposed by OpenShift Container Platform: a network latency smaller than 10 ms round trip time must be ensured at all times. This is necessary to keep a quorum of *etcd* instances and avoid unnecessary *etcd* leader changes. The *etcd* instances manage a database of the control plane nodes in sync. The *etcd* database is the key-value store for OpenShift Container Platform, which persists the state of all resource objects.

If the latency is within the defined limit of 10 ms for a roundtrip, a stretched cluster can even span multiple data centers to cover aspects of disaster recovery. The flexibility of distributing the nodes can, for example, be used to spread a single cluster of OpenShift Container Platform and IBM Storage Fusion Data Foundation across typically 3 data centers, each representing an availability zone. In case of an outage of one availability zone, the other two availability zones will continue to operate without any downtime or recovery time at all. There is no recovery point nor data loss. This makes it possible to achieve a **Recovery Time Objective** (RTO) of zero, as well as a **Recovery Point Objective** (RPO) of zero.

NOTE	This solution for high availability can be applied if you can guarantee the low latency of the network between the data centers / availability zones. Each site / availability zone must be connected by low latency networks (latency < 5ms for a full round-trip).

NOTE	Despite the resilience features by this high available setup, it still does not cover larger disaster scenarios, where entire data centers fail. For disaster recovery a regional or metro DR topology is required (see section 4.6.2 and 4.6.3).

For normal operations, IBM Storage Fusion Data Foundation and OpenShift Container Platform

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

require a minimum of 3 storage nodes with attached storage but can still run healthy as long as 2 storage nodes and 2 control planes are operational.

In failure scenarios, storage nodes can become unavailable to the OpenShift cluster. This can be caused, for example, by:

- Unavailability of the **LPAR or physical hardware** in which a control node is installed.
- Failure of a **node** (e.g. storage, compute, or control node).
- **Network** failures interrupting the interaction between the nodes within the cluster.
- Failure of the **physical storage** or logical **OSDs** connected to the storage nodes.

The recommended best practices on a 3 LPAR cluster with 3 control nodes and 3 compute nodes includes the following constraints:

- To run a fully functional OpenShift cluster which includes IBM Storage Fusion Data Foundation, the following requirements need to be met:
  - Control plane: The number of 3 control plane nodes must always be active.
  - Compute nodes: Ensure that you have enough compute nodes to handle failures due to excessive load and lost infrastructure. In the case of deploying IBM Storage Fusion Data Foundation a minimum of 3 compute nodes is required.
  - Storage nodes: A minimum of 3 storage nodes must always be active.
- A single OpenShift cluster must ensure consistency even across sites:
  - For Red Hat OpenShift Container Platform, this includes the `etcd` database, which stores all cluster meta data. The OpenShift control plane takes care of keeping all copies of the `etcd` database on all control nodes within a cluster in sync.
  - to ensure consistency of *etcd* within a cluster the network latency must be less than 10 ms round trip time.
- Storage must be replicated across a multi-site deployment:
  - IBM Storage Fusion Data Foundation ensures that the persistent volumes on all storage nodes are kept in sync (provided by Ceph technology)
  - For all other storage, which is not managed by IBM Storage Fusion Data Foundation, the environment must ensure replication by itself.

In case of failure, a single OpenShift cluster can deal with failing nodes in a resilient way:

- If 1 OSD on one of the compute nodes is down, the cluster should continue its operation normally and the application access (read and write) should have no impact.
- If 2 out of 3 OSDs are down, you have no access to data (either read or write) to preserve or guarantee data integrity. Once the cluster rebalances data to other OSDs and you have 2 copies of the object again, you get back read-write access.
- If 1 compute node is down, the cluster should continue its operation normally and the application access (read and write) should have no impact.
- If 2 out of 3 compute nodes are down, the cluster will freeze scalability on all pods of the

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

remaining node. But the pods on that node will continue to run in healthy mode. Nevertheless, the applications will have no access to persistent volumes until the failing compute nodes are restarted/replaced and the IBM Storage Fusion Data Foundation storage nodes become available again. Once the failing compute nodes are restored (either by restart or replacement) the system will be fully functional again. Ceph will ensure that the storage volumes are in sync at any time.

- In case of network failure on 1 compute node, the cluster should continue its operation normally and the application access (read and write) should have no impact.
- In case of network failure on 2 compute nodes, the cluster will freeze scalability on all pods of the remaining node until the failing compute nodes are restarted/replaced. Again, the applications will have no access to persistent volumes until the IBM Storage Fusion Data Foundation storage nodes become available again.
- If 1 LPAR (control & compute) is down, the cluster should continue its operation normally and the application access (read and write) should have no impact.
- In case of node replacement due to 1 compute node failure, the cluster should continue its operation normally, the application access (read and write) should have no impact and the data recovery should happen automatically on the new node.
- In addition to the availability of the compute/storage node, OpenShift Container Platform requires 3 control plane nodes to be active. In case 2 control nodes are down, the cluster will be in error state and only allow read operations. Once failing control plane nodes are restored (either by restart or re-installation) the system will be fully functional again.

Table 3. Failure scenarios for IBM Storage Fusion Data Foundation

Resource	**3 instances**	**2 instances**	**1 instance**	**ODF Data recovery**
OSDs *(on one storage node)*	normal operation	normal operation	No access to application data	self-recovery
Storage / compute node	normal operation	normal operation	outage	self-recovery
network interface	normal operation	normal operation	outage	self-recovery
LPAR	normal operation	normal operation	outage	self-recovery

4.6.2 Regional disaster recovery

The regional disaster recovery topology creates an active/passive setup to cover regional data center failures. Such deployment becomes necessary once the latency requirements for a single cluster cannot be achieved, due to larger geographical distance between the data centers.

An active/passive setup is based on two independent and self-contained clusters of OpenShift Container Platform and IBM Storage Fusion Data Foundation. Each cluster running in a separate data center. One cluster is active, the other one is passive as stand-by in case of an outage.

The persistent volumes of IBM Storage Fusion Data Foundation are being replicated from the active

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

to the passive site. This happens asynchronously. OpenShift Data Foundation (Ceph) takes care of the replication implementation on a persistent volume level. The replication interval can be chosen as short as a minute.

This setup is targeted for data centers located at a larger geographical distance apart from each other (regional data centers). The network latency is too significant to allow a reliable synchronous replication, nor would it allow a stretched cluster.

To orchestrate the two independent clusters, an additional component is required: Red Hat Advanced Cluster Manager (ACM). RHACM is installed as the **hub cluster** and defines in its policies, to which of the **managed cluster** the traffic needs to be routed.

In case of an outage in the active site, an administrator can trigger a failover to the stand-by site. Because all namespaces of the active site have been synchronized to the stand-by site, it can be ensured that the applications can continue to run. As the failover procedure is automated, the **Recovery Time Objective** (RTO) is minimized and the transition, in case of an outage, will be smooth. It is possible to use the second site as "hot" stand-by with all applications up and running and ready to pick up the load in case of a disaster. As an alternate approach, the second site can decommission low-priority applications and only start them in case the stand-by site is being activated ("cold" stand-by).

On the stand-by site, the replicated data set will be used to continue the operations. As the replication between the persistent volumes on active and stand-by site can be configured with a high frequency, a solid value for the **Recovery Point Object** (RPO) can be achieved. Typically, the RPO is in the range of minutes, which implies that there could be a small amount of data loss to be considered.

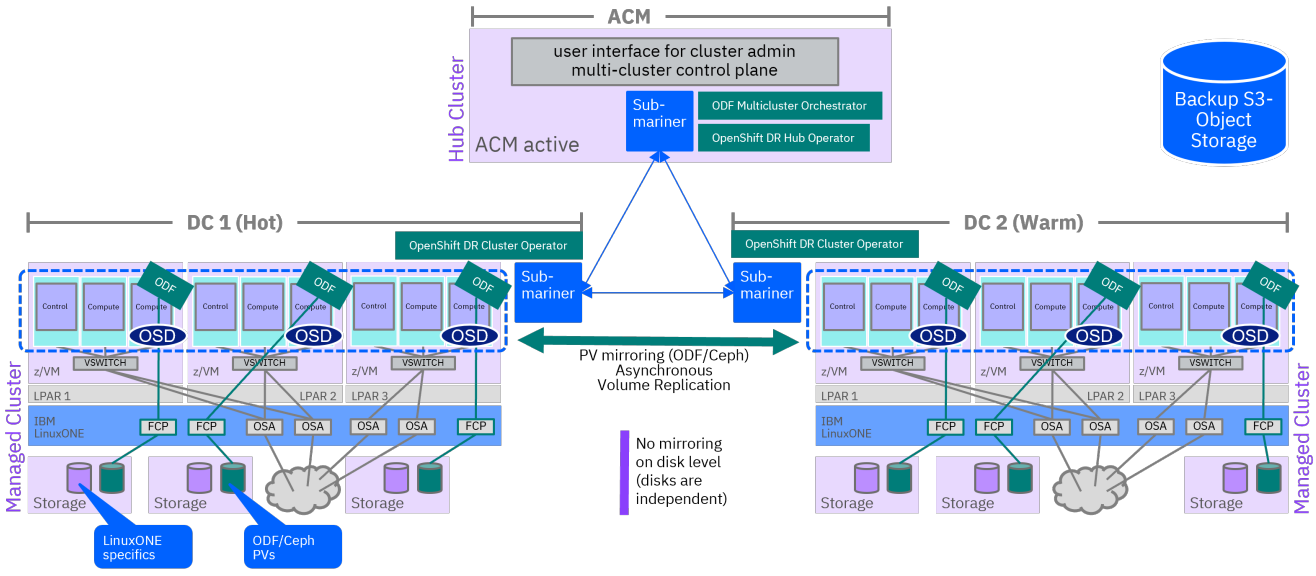

Figure 4.6-7: Regional DR setup

For more details and instructions how to setup this topology, please refer to the documentation for regional DR.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

NOTE	This setup takes advantage of OpenShift Application Data Protection API (OADP), which is currently available as a community operator for IBM zSystems. More details on OADP can be found in section 5.4.1.

NOTE	Please note that regional DR is currently only supported as Tech Preview for Block storage (CephRBD) starting ODF release 4.12.1. For IBM zCX this setup is not supported.

There are a several basic use cases for operating a regional DR environment. The following list describes the basic administrative procedures, when dealing with regional DR and the involved OpenShift clusters managed by Advanced Cluster Manager.

- **Normal operations:** Deployment of 2 managed clusters with OpenShift Container Platform and IBM Storage Fusion Data Foundation plus an additional hub cluster with Advanced Cluster Manager. Both managed clusters run in normal operations. One cluster is active and the other one is in passive mode.
- **Application resilience within a single managed cluster (high-availability):** As each managed cluster runs OpenShift, all resilience mechanism within the cluster are available. This includes the high availability capabilities like restart failing storage nodes, compute nodes, and control planes. As long as 2 storage nodes and 2 control planes are running, the cluster is fully operational and there is no need for fail over to another data center.
- **Switch active managed cluster:** Advanced cluster Manager allows administrators to select the active cluster among its managed clusters back and forth as needed. This is a manual task, which the administrator can perform. Applications can be re-located to the active cluster as well. The data, which is being used by the fail-over cluster (former passive cluster, which now will become active) had been replicated by Ceph before. (Applying DR placement control).
- **Create a backup of the artifacts:** At any time or based on a defined schedule, the artifacts of an application namespace within a managed cluster can be backed up and stored as an archive file in a safe place. Typically, all artifacts (meta data, images, PV data) are stored in object storage offered by a cloud provider and can be used later if a restore of the applications in the cluster becomes necessary. (see chapter 5.4 for more details on backup of applications in a OpenShift cluster)
- **Complete failure of a managed cluster (disaster recovery):** Once the active managed cluster is failing, it is switched over to the stand-by cluster, which has been passive so far. This task needs to be triggered by an administrator (identical to Scenario 3). The failing cluster can then be recovered from scratch and joined back to same DR environment, which is managed by Advanced Cluster Manager. Later, the selected active cluster can be switched back to the restored cluster. Assumption is, that the hub cluster of Advanced Cluster Manager will continue to run uninterrupted during this procedure.
- **Complete failure of both managed clusters (disaster recovery):** If both managed cluster are destroyed (either one after the other or both at the same time), those clusters need to be recovered from a backup (as created in Scenario 4) and added back to the hub cluster, managed by Advanced Cluster Manager. Assumption is, that the hub cluster of Advanced Cluster Manager will continue to run uninterrupted during this procedure.
- **Loss of network connectivity between the managed clusters:** As soon as one managed cluster

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

loses connectivity to the hub cluster with Advanced Cluster Manager, that cluster is considered as not operable and DR operations can be kicked-off (see previous scenarios)

- **Complete failure of the hub cluster running ACM:** In case of a failure of the hub cluster with ACM you will need to fall back to a redundant installation of a hub cluster, which can take over operations. Deployment of ACM for fail-over of ACM itself needs to be prepared upfront. The current configuration of the fail-over deployment of ACM can be restored from an OADP backup, which needs to be prepared upfront (e.g. from a S3 Cloud storage). More details on resilience and DR for ACM can be found here: backup and restore for ACM

4.6.3 Metro disaster recovery

Metro DR is a variation of the regional DR topology and also based on 2 managed OpenShift clusters and a ACM deployment for orchestration.

The key differences are:

- Low latency between the nodes is essential. Like a stretched cluster the roundtrip latency must be below 5ms roundtrip.
- IBM Storage Fusion Data Foundation must be deployed in external mode and uses externally deployed standalone Ceph storage cluster (IBM Storage Ceph), which is stretched (shared) across the 2 data centers. With the common storage the async replication can be avoided and a Recovery Point Objective (RPO) of zero can be achieved.
- The stretched Ceph cluster will need an arbiter node to resolve inconsistencies in case one of the Ceph OSDs fails.
- Both OpenShift clusters can be active.

NOTE	The external Ceph cluster is based on IBM Storage Ceph and can only be deployed in an x86 environment, as Ceph is currently not supported as standalone product on IBM zSystems.
NOTE	Red Hat Ceph Storage is currently being rebranded to IBM Storage Ceph.
NOTE	Metro DR is currently not supported neither on IBM zSystems and IBM LinuxONE nor on IBM zCX.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

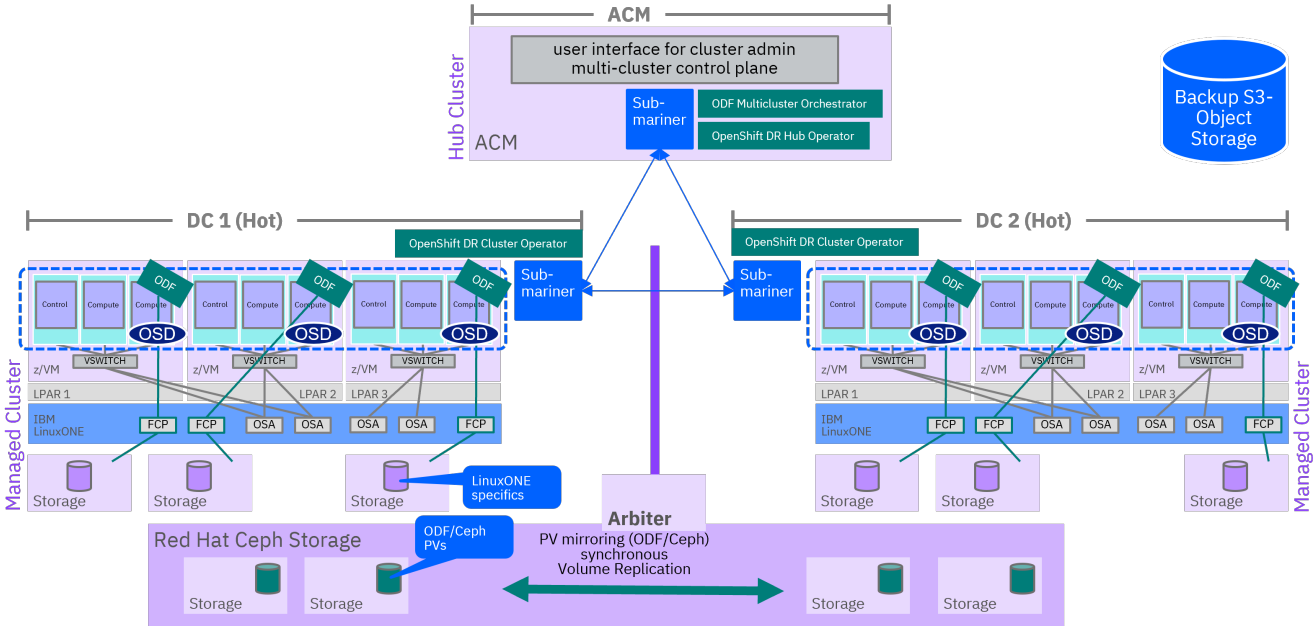

Figure 4.6-8: Metro DR setup

4.6.4 Using Geographically Dispersed Parallel Sysplex (GDPS)

While regional and metro DR are implemented as open-source software running on OpenShift, it is also possible to add disaster recovery to the hardware layers of the system.

Geographically Dispersed Parallel Sysplex (GDPS) is a powerful and highly scalable solution by IBM, which mirrors physical storage between data centers. GDPS can run either on a z/OS system or within a appliance.

GDPS can control OpenShift nodes to be restarted, manage network failures and perform disk mirroring (ECKD/FICON/DASD only).

GDPS can also be used for multi-site replication and failover in case of the outage of one site.

Depending on the distance between the data centers, there are multiple options for setting up GDPS:

- GDPS Metro: based on the IBM Metro Mirror synchronous disk mirroring technology
- GDPS Global - XRC: based on the IBM Extended Remote Copy (XRC) asynchronous disk mirroring technology (also called IBM z/OS Global Mirror)
- GDPS Global - GM: based on the IBM System Storage Global Mirror technology, which is a disk subsystem-based asynchronous form of remote copy

GDPS features sophisticated options to ensure data consistency. In a metro distance setup, the application execution can be forced to wait for an acknowledgement before proceeding to the next task. GDPS Global includes capabilities for a background sync to ensure consistency.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

NOTE	GDPS is fully tested and supported for Red Hat OpenShift Container Platform applications running on IBM zSystems. Nevertheless, formal support for using GDPS with stateful workload using IBM Storage Fusion Data Foundation on IBM LinuxONE is not available yet. For IBM zCX, GDPS is formally supported as it uses VSAM LDS under the covers as the local disks.

NOTE	GDPS is limited to the use of ECKD/FICON/DASD disks only.

For more details on GDPS, please refer to this IBM Redbook: IBM GDPS: An Introduction to Concepts and Capabilities

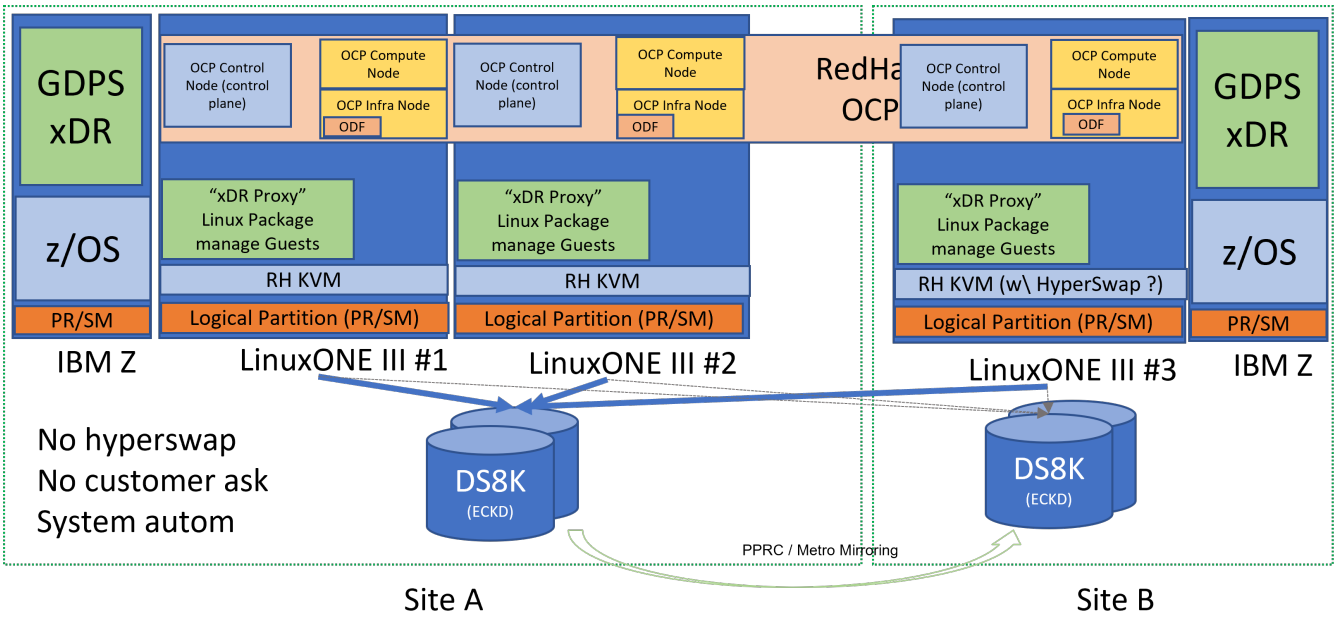

Figure 4.6-9: Using GDPS with KVM

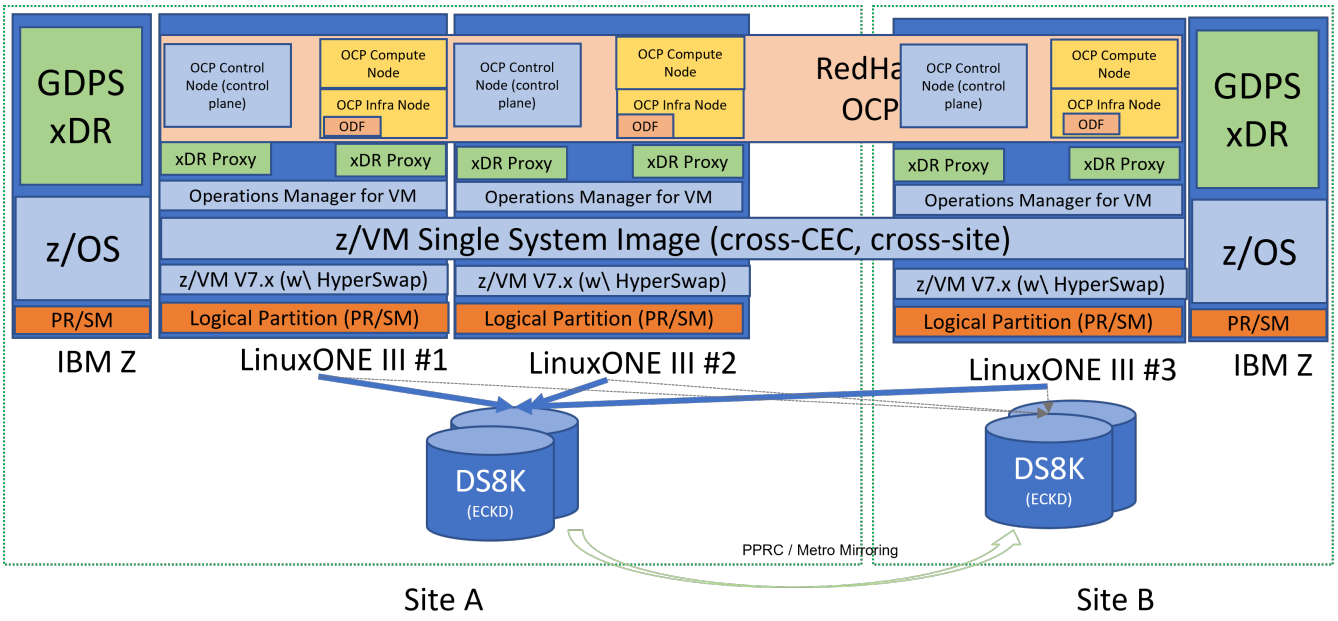

Figure 4.6-10: Using GDPS with z/VM

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

4.6.5 Further consideration for high availability and disaster recovery

Regardless, which topology is chosen for implementing a resilient disaster recover solution, the following considerations need to be taken into account:

- Coordinated backups of the application and data are essential to restore the system after a disaster or other kinds of data corruption:
  - IBM Storage Fusion Data Foundation supports snapshotting of persistent volumes using a Container Storage Interface (CSI).
  - IBM Storage Fusion Data Foundation embraces OpenShift APIs for Data Protection (OADP) in this context.
  - see section 5.4 for more details on backup and restore.
- Within a single OpenShift cluster, all interaction is managed synchronously and has strict demands with regards to networking performance and low latency (which implies that nodes within a cluster are in close geographical proximity):
  - A maximum latency of 5ms (roundtrip) must be ensured for all components within a single cluster. It must be emphasized: guaranteeing that maximum latency is essential for a reliable operation.
- Operational procedures need to be in place:
  - e.g., a plan to restart or re-install nodes within an OpenShift cluster or to switch to a failover datacenter in case of a disaster.
- The hardware of IBM zSystems includes the following mechanisms for fault tolerance:
  - RAID on storage server internal hard disks.
  - Redundant connections to storage servers.
  - Redundant server power supplies connected to different power sources.
  - Redundant cores in case of a core failure.
  - Bonded network interfaces connected to redundant network switches.
  - Extensive self-checking a self-recovery capability.

It is important to emphasize that there are further considerations of resilience applicable to IBM zSystems. IBM zSystems features all the aspects of reliability and resilience, eliminating most single points of failure at the hardware level.

For more information on reliability, availability, and serviceability of IBM zSystems refer to the IBM documentation covering reliability, availability, and serviceability.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

Chapter 5. Operational considerations

This chapter focuses on aspects of "Day 2" operations of IBM Storage Fusion Data Foundation like validation, security, persistence, backup, and performance tuning. This chapter enables you to become productive quickly. It also helps you to make your solution reliable and secure right from the beginning.

5.1 Logging and monitoring

IBM Storage Fusion Data Foundation uses built-in capabilities for monitoring and logging as provided by the Red Hat OpenShift Container Platform.

Once deployed in IBM Storage Fusion, the OpenShift Data Foundation operator manifests itself as a new `IBM Storage Fusion Data Foundation` navigation entry in the OpenShift Administrator's view under the `Storage` main entry. This provides a handy "USE" (Usage-Saturation-Events) dashboard for the operational status of IBM Storage Fusion Data Foundation.

NOTE	If deployed from Red Hat's Operator Hub, the Red Hat branding "Red Hat OpenShift Data Foundation" will appear.

Data for these visualizations is gathered through the built-in monitoring stack. Metrics are exported from the Noobaa and the S3 service monitors, the ODF operator for the default ODF storage cluster Storage System, and the Rook Ceph manager. Detailed metrics can be obtained through the Administrator's `Observe / Metrics` UI in the OpenShift Console.

In addition to the storage dashboard, the OpenShift Data Foundation Operator extends built-in monitoring with several alerts. These are named after the respective service. The interested administrator can find them by navigating to `Observe / Alert - Alerting Rules` and using `Ceph` or `Noobaa` as name filter.

Normal OpenShift Logging mechanisms are available. Log information from OpenShift Data Foundation-related Pods can be retrieved through the usual `oc logs <pod-name>` command. A consolidated logging experience, however, is achieved by using OpenShift Logging.

However, as OpenShift Data Foundation provides an integrated storage solution for the cluster, it is worth customizing the monitoring and, optionally, the logging environment.

5.1.1 Monitoring Data Persistence using OpenShift Data Foundation

When OpenShift's monitoring stack is deployed initially to a new OpenShift cluster, it does not start with persistent data stores for the collected metrics data. This means that whenever one of the provided Prometheus Pods is rescheduled, its own Time Series DataBase (TSDB) is re-created from scratch. In other words, historical monitoring data is lost. Notably for production clusters, it is essential to keep persistent monitoring data. This aspect is not only important for OpenShift Data Foundation monitoring, but also for monitoring the whole OpenShift cluster. Fortunately, with the presence of OpenShift Data Foundation, it is easy to persist monitoring data.

The OpenShift documentation on monitoring describes how to set up persistent storage for the

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

various components of the monitoring stack. While the documentation suggests the use of Physical Volumes (PV) obtained through the Local Storage Operator (LSO), the OpenShift Data Foundation solution provides some advantages. Notably consider the following: While LSO provided volumes are bound to distinguished cluster nodes, Pods for the monitoring components will be finally tied to the cluster node where their respective data resides. But with OpenShift Data Foundation volumes, this restriction is no longer true: The data can "move" to the node where the Pod resides, which dramatically improves the flexibility and resiliency of the monitoring stack.

Using `ocs-storagecluster-ceph-rbd` as `storageClassName` instead of the suggested `local-storage` value in the documentation example places the monitoring data to OpenShift Data Foundation provisioned volumes. As this storage class also provides volume expansion, it is easy to adjust the PVC size to the actual required size for the desired monitoring retention time. This is clearly a further advantage over the use of LSO provisioned PVCs.

5.1.2 Consolidated Logging using OpenShift Logging Operator with OpenShift Data Foundation

As mentioned above, plain Kubernetes / OpenShift logging on container level is by default available. However, a consolidated logging solution based on the OpenShift Logging Operator is a recommended way to manage production environments. Similar to the monitoring stack, logging data also should be persisted. Again, OpenShift Data Foundation is very helpful here.

The OpenShift documentation on logging shows how to connect the logstore entities with their Persistent Volume Claims. Using the `ocs-storagecluster-ceph-rbd` storage class for OpenShift Data Foundation provided storage gives the same flexibility and fault tolerance as discussed in the previous section on monitoring data.

5.2 Performance considerations

This section gives a brief overview about best practices to tune the performance of IBM Storage Fusion Data Foundation. It assumes that the minimal requirements of IBM Storage Fusion Data Foundation are met by your cluster. The performance characteristics of IBM Storage Fusion Data Foundation depend on the operating system, cpu, network, memory and disk configuration.

5.2.1 Hardware layer

This section covers performance considerations and tuning tips for the hardware layer.

IFL Configuration — Clusters operating IBM Storage Fusion Data Foundation should spend at least 1 IFL to operate IBM Storage Fusion Data Foundation. More IFLs increase the performance for cpu-intensive workloads, for example, with small block sizes. This can be less efficient for workloads with large block sizes and read operations, because they might be bottlenecked by network capacity.

Storage Guests — One should specify at least the recommended amount of CPU, memory (for example, frequency, size, etc) for nodes operating IBM Storage Fusion Data Foundation. The performance of the cluster can be impacted if a disk is too slow or broken. It is recommended to use SSD for journaling.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

Network Configuration — If the network configuration allows it, it is beneficial to enable jumbo MTU sizes, e.g. MTU 9200.

NOTE	Blog Post: Creating storage strategies for Red Hat Ceph Storage clusters and Hardware selection recommendations for Red Hat Ceph Storage

5.2.2 Operating system

This section introduces tuning tips for the operating system (OS) running IBM Storage Fusion Data Foundation. Most changes to the OS can be made via the Machine Config Operator of RHOCP.

Receive Package Steering — Receive Package Steering (RPS) directs packets to specific CPUs for processing. It aims to increase the CPU cache hit rate and thereby reduce network latency. Enabling RPS on the operating system on the storage nodes improves their network throughput and latency. This affects the Ceph performance noticeably, especially for workloads with large block sizes which tend to be network-bounded. You should consider that the increase in network performance comes at the expense of CPU demand, which might affect CPU-intensive workloads.

NOTE	Blog Post: https://developer.ibm.com/tutorials/red-hat-openshift-on-ibm-z-tune-your-network-performance-with-rfs/

Transparent Huge Pages — Transparent Huge Pages (THP) allows the operating system to use large memory pages, typically 2MiB in size. It aims to improve the performance by reducing the number of memory page table entries, memory fragmentation and memory overhead. However, it can cause memory allocation delays during runtime and is not beneficial for all workloads, especially databases. Therefore, it can be beneficial to disable THP for OCP Nodes to improve the performance and reduce CPU cycles.

NOTE	OCP Documentation: Disable THP on IBM Z OCP Clusters

5.2.3 OpenShift layer

This section introduces tuning tips for the nodes running IBM Storage Fusion Data Foundation.

Node Labeling — RHOCP enables the labelling of nodes as infrastructure to host infrastructure services on them. The infrastructure services include OpenShift monitoring, Prometheus, and Grafana, which are all CPU and memory intensive. Isolating these services from compute and storage nodes can prevent performance drops. It is recommended that you create at least 3 infrastructure nodes with 4 vCores each and 32GB memory.

5.2.4 Configuration of IBM Storage Fusion Data Foundation

This section lists general tuning tips for Ceph and IBM Storage Fusion Data Foundation.

5.2.4.1 Ceph Recovery

In the case of a failing OSD, the Ceph cluster has to rebalance the data after the OSD is online again. Ceph avoids cluster performance degradations by slowing down the rebalancing. You can speed up

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

the process by changing the following parameters:

Ceph Recovery parameters

```
ceph tell 'osd.*' injectargs --osd-max-backfills=12 --osd-recovery-max-active 4
ceph tell 'osd.*' config set osd_recovery_sleep_hdd 0
ceph tell 'osd.*' config set osd_recovery_sleep_ssd 0
```

NOTE Blog Post: How to speed up or slow down osd recovery: Throttling backfill and recovery and Handling a node failure

5.2.4.2 Ceph Replication Mechanism

Data availability in ceph comes with a price — the clusters available space, performance and efficiency is greatly affected by the object replication. Tuning it down to 2 improves the clusters performance and efficiency by over 30%. However, it greatly increases the risk of a loss of data on failures. A object replication of 2 can be beneficial for many workloads and can be considered in environments in which disk or node failures are rare.

Ceph Commands to configure number of replicas for CephRBD

```
POOL="ocs-storagecluster-cephblockpool"
NUM_REPLICAS=2
ceph osd pool set $POOL size $NUM_REPLICAS --yes-i-really-mean-it
```

5.2.5 Configuration of Hypervisors

This section gives an overview of the hypervisor-specific tunings.

5.2.5.1 z/VM Tuning

...

5.2.5.2 KVM Tuning

This section introduces KVM-specific tuning practices for IBM Storage Fusion Data Foundation.

KVM Guest I/O-Threads — Increasing the number of I/O threads for a KVM guest can improve the performance by enabling the guest operating system to handle more I/O operations in parallel. An IBM Storage Fusion Data Foundation storage node guest can operate one or more OSDs which increase the I/O contention. This contention can be reduced by increasing the number of I/O threads. This allows a more efficient use of the attached storage devices. However, it is important to consider that the increasing number of I/O threads could also lead to increased CPU and memory usage.

NOTE OCP Documentation: Use I/O Threads for your virtual block devices

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

5.2.6 Application-specific Tunings

...

5.2.6.1 Tuning an application based on MongoDB

DO WE HAVE SOMETHING NICE HERE?

5.2.7 Performance benchmarking

The performance of IBM Storage Fusion Data Foundation depends on disk, network, and compute resources. It is, therefore, important to gain insights into each resource. This section lists the performance benchmarks and general recommendations for IBM Storage Fusion Data Foundation.

- **FIO (Flexible I/O Tester)** is one of the most well-known tools to benchmark storage backends. Ceph itself has diagnostic tools to categorize OSDs as SSD or HDD and to warn for slow performance, but it can also be helpful to obtain the performance characteristics of a storage disk via a benchmark tool. You can characterize the performance capability of a raw storage disk using sequential read and write and random read and write workloads. Because Ceph strips data across many OSDs, the access pattern on the raw device resembles a random workload. FIO also contains a Ceph driver, which allows the benchmarking of CephRBD or CephFS. You should also consider multi FIO-server setups via the FIO client-server mechanism because this is a pattern which is typical for an RHOCP cluster where multiple pods consume the storage.
- **Yahoo! Cloud Serving Benchmark (YCSB)** is one of the most well-known NoSQL benchmark suites. It provides a framework to create benchmark workloads and collect performance data, e.g. transactions-per-second or latency. It supports many databases, e.g. MongoDB or Redis. This enables to evaluate the performance of databases backed by IBM Storage Fusion Data Foundation with cloud workloads.
- **Uperf** is a network benchmark tool, which can be used to estimate the network capacity. IBM Storage Fusion Data Foundation is network-intensive because data is striped across multiple OSDs on different storage nodes. Therefore, it is important to consider the network capability and configuration. With Uperf you can measure the performance in pod-to-pod scenarios on one compute node, between compute nodes, and across LPARs. You must ensure that the network between OSDs is sufficient.

CAN WE ADD SOME ROUGH COMPARISONS IN THIS PUBLIC DOCUMENT? Comparing with NFS Comparing with local storage on RHOCP Comparing with Intel Comparing with Ceph standalone Comparing with IBM Storage Scale

5.3 Security

5.3.1 Encryption

IBM Storage Fusion Data Foundation allows encryption of the stored data (data at rest) using the common Linux Unified Key Setup (LUKS2). There two levels of granularity possible:

- OSD level: an entire storage device is encrypted

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

- PV level: a specific persistent volume (used by a application) is encrypted individually

For an **OSD level encryption**, the keys used for the encryption process can be stored either internally or externally.

- Internal key management is done within the OpenShift cluster. For this purpose the key is stored as a name-value-pair inside the OpenShift etcd database.
- External key management can be provided by a 3rd party key store, such as Hashicorp vault.

An external key management solution is preferred because it ensures a clear separation between the keys and the protected data.

For a **PV level encryption**, external key management is mandatory. In this case it is not possible to store keys internally inside the etcd database. The PV level encryption allows scoping the encryption process to specific persistent volumes used by an application. This allows an instance to protect one application (or tenant) from another. IBM Storage Fusion Data Foundation has solved the separation between the persistent volumes, by creating a set of storage classes for each PV/tenant and assigning keys specifically to a storage class.

NOTE Currently, only RBD block storage is supported for PV level encryption. Support for the other storage classes will be available in a future release.

NOTE When storing keys internally, it is important to secure and backup the etcd database correctly!

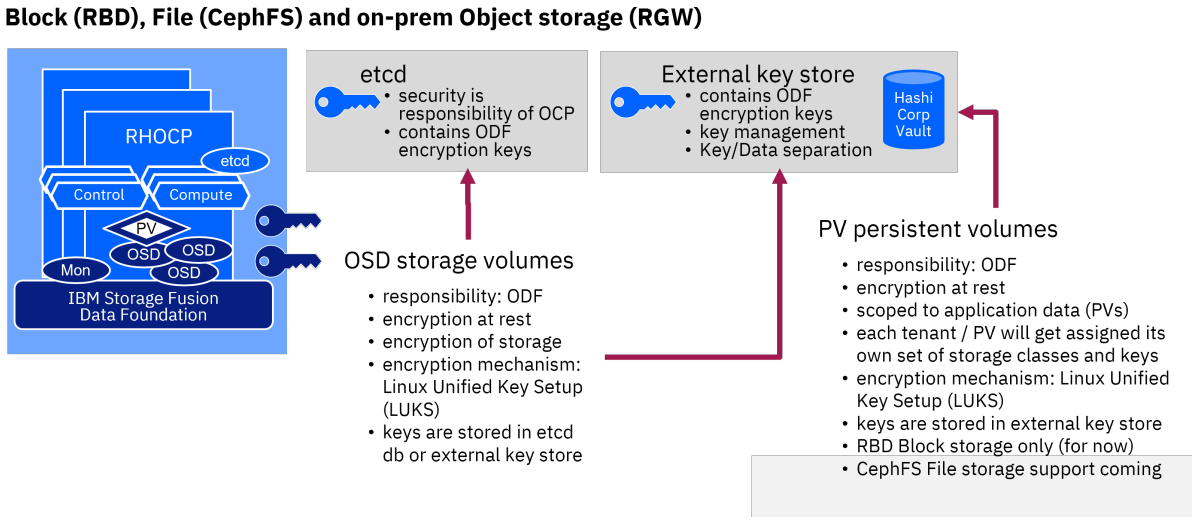

Figure 5.3-1: Encrypting storage

5.4 Backup and recovery

5.4.1 Considerations for Backup and Restore

An OpenShift cluster is focused on delivery containerized workload. Each application is built from self-contained containers which are orchestrated by Kubernetes. As part of that orchestration containers and scaled up and down or moved between cluster nodes (pods). From a Kubernetes

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

perspective, application instances can be recreated at any time within any cluster. This implies that a automated and reproducible CI/CD pipeline is established. Such pipeline is not only used to bring new code updates into a production environment, but also recreate a production environment as needed (e.g. in case of a data center disaster)

The following figure shows, which pieces of data are required for a Kubernetes application:

- Application image with all its dependencies (built from a source control system via automated CI/CD and provided via Image Registry)
- Static state of an application once deployed (stored in source control as scripts and part of the automated CI/CD. This initial state is applied once during deployment)
- PV / PVC (meta data of persistent storage used by the application)
- Storage (the actual data of the persistent application storage)
- Dynamic state (meta data created by Kubernetes during the lifetime of an application image, e.g. stored in the cluster etcd database)

It is important to mention that the dynamic state will be lost on each re-deployment of an application image, as containerized workload will always start as a new image with its static state as defined in the CI/CD pipeline.

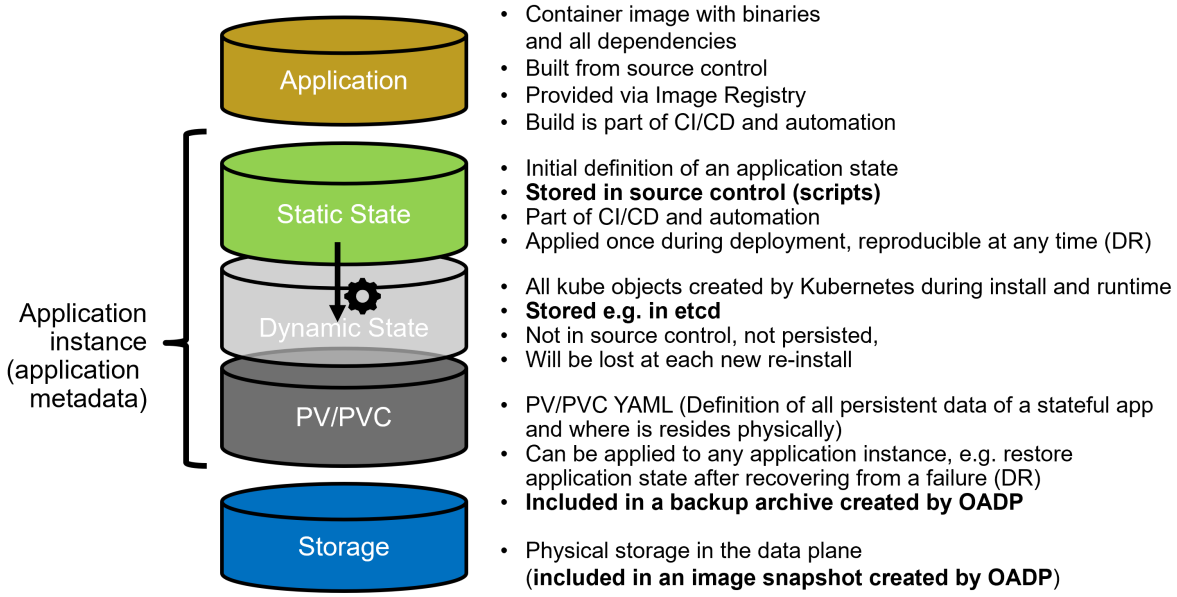

Figure 5.4-1: Data of a Kubernetes container applications

As a fundamental best practice in a Kubernetes world, there is no need to back up a cluster or application image, as it will always be possible to recreate that at any time using the CI/CD automation. This is why a solid CI/CD pipeline can also be considered as a key element of a disaster recovery approach for your workload.

Setting up an environment with infrastructure and application workload needs to be a fully automated and reproducible process allowing to restore an environment at any time in any place.

Nevertheless, the one essential thing to backup are the PVs and PVCs used by the application. Those include the precious stateful persistent data of the workload. The backup and restore of PVs (and it

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

meta-data) is covered by the Open API for Data Protection (OADP).

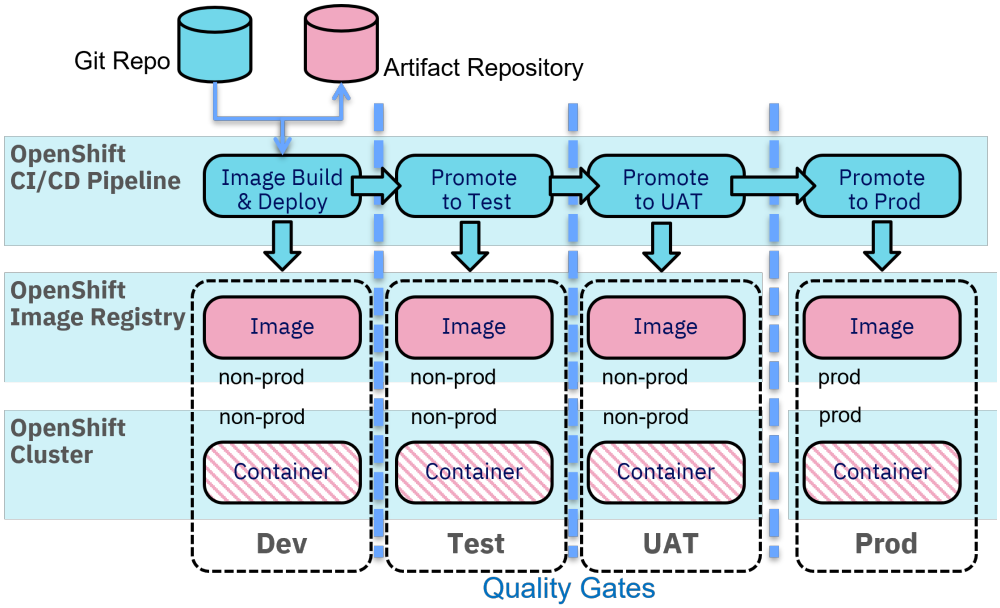

Figure 5.4-2: Fully automated and reproducible pipelines and CI/CD are the foundation of disaster recovery

5.4.2 Open API for Data Protection (OADP)

Creating and managing backups is a key aspect of data resiliency. For this purpose, IBM Storage Fusion Data Foundation can take advantage of the OpenShift Application Data Protection API (OADP) for backup and restore.

From an IBM Storage Fusion Data Foundation perspective, a backup covers the Kubernetes resources, meta data, and a snapshot of the **persistent volumes (PVs)** for a particular application namespace. The scoping by namespaces makes the backup specific to an application, rather than saving the entire OpenShift cluster in one piece.

OADP does not backup the cluster wide artifacts, nor does it backup install or topology configuration. For example, hypervisor settings or the specific compute/control node layout of the cluster (like the etcd database) need to be part of the automation/scripting which is used to install the cluster in the first place.

OADP-driven backups are independent of the release version of IBM Storage Fusion Data Foundation, the chosen architecture, and the storage providers.

OADP copies the created backups into a S3 object storage, which is typically hosted by a cloud provider like IBM Cloud Object Storage or AWS.

5.4.3 Velero, CSI Snapshots, and Restic

OADP uses a set of open-source components when creating a backup of an application namespace:

5.4.3.1 Velero and CSI Snapshots

- The open-source component Velero implements the backup of Kubernetes resources.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

- A snapshot of persistent volumes is based on Container Storage Interface (CSI). A prerequisite is a OpenShift cluster that supports either a native snapshot API or CSI snapshots (which is true for use with IBM Storage Fusion Data Foundation)
- The meta-data of a persistent volume is copied to an object store in the cloud as part of the CSI volume snapshot.
- The actual content of a CSI volume snapshot is an image stored as an archive file locally in the cluster from which it is produced. In a second step, OADP's data mover (based on VolSync) copies the CSI volume snapshot content to a object store in the cloud as well.
- The target location in which the backup is safely stored is typically a S3-compatible object store offered by a cloud provider, e.g., an IBM Cloud Object Store, Amazon Web Services, Microsoft Azure, or Google Cloud Platform. It is also possible to use the Noobaa Multicloud Object Gateway, which is included in IBM Storage Fusion Data Foundation. Additional custom OADP plugins can implement further backup target locations, such as NFS.

5.4.3.1 Restic

If CSI snapshots are not an option, there is a alternate approach based on the Restic open source project. Restic is a file system copying technique that is used by OADP in case the cluster does not support CSI APIs. Restic backups copy both volume data and Kubernetes metadata to S3-compatible object store.

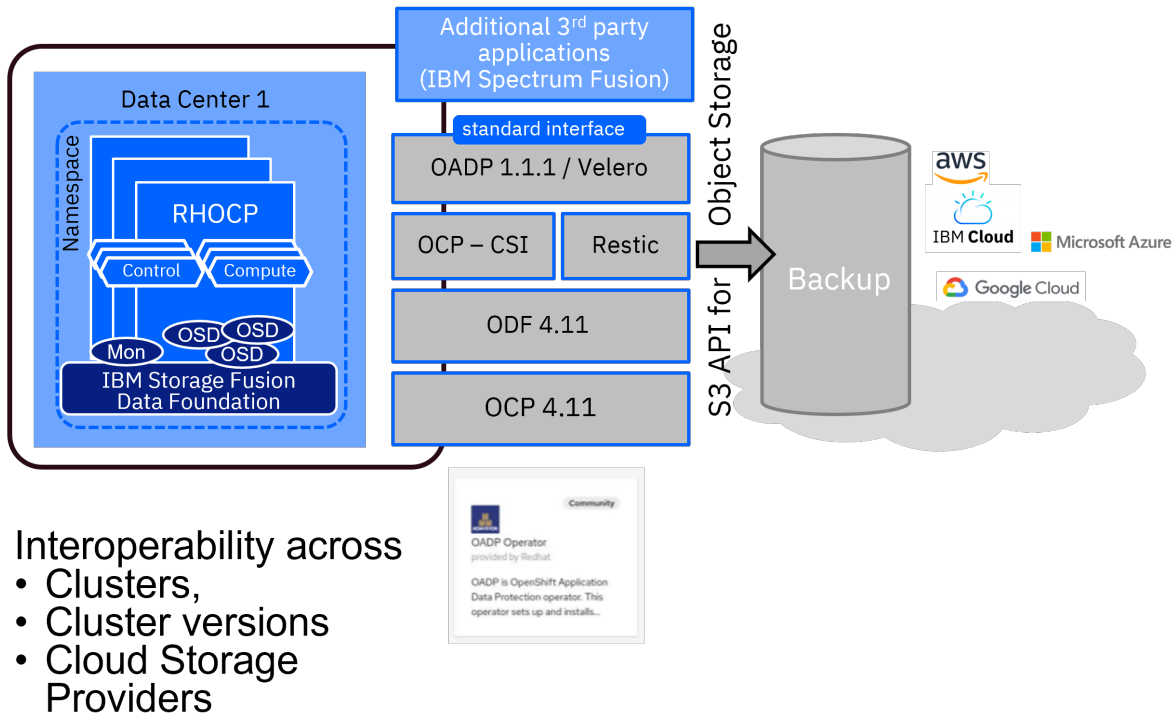

Figure 5.4-3: Backup and restore

5.4.4 Processing a backup

OADP backups up all resources in a cluster, which belong to a defined namespace. OADP supports both scheduled and on-demand backups. The scheduling capabilities allow specifying intervals of the backup snapshot.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

The configuration steps for setting up a backup include defining the target location and the access credentials for the backup archive, typically a S3 bucket in a cloud-based object store.

Once the location is defined, a `DataProtectionApplication` is created. If CSI snapshots are used for backing up the PersistentVolume (PV) then the `VolumeSnapshotClass` custom resource needs to be defined as well. `Volumesnapshotclass` specifies the storage provider specific parameters to use when creating a snapshot of the persisten volume.

Sample definition of a `DataProtectionApplication` using CSI snapshots

```
apiVersion: oadp.openshift.io/v1alpha1
kind: DataProtectionApplication
metadata:
  namespace: openshift-adp
  name: example-velero
spec:
  backupLocations:
    - velero:
        config:
          profile: 'default'
          region: us-east-1
        credential:
          key: cloud
          name: cloud-credentials
        default: true
        objectStorage:
          bucket: kvmtestbucket
          prefix: 'velero'
        provider: aws
configuration:
    restic:
      enable: false
    velero:
      defaultPlugins:
        - openshift
        - aws
        - csi
    featureFlags:
        - EnableCSI
```

In order to back up an application namespace, a `BackupCustomResource` has to be defined.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

Sample definition of a `BackupCustomResource` using CSI snapshots

```
apiVersion: velero.io/v1
kind: Backup
metadata:
  name: mysql-backup
  labels:
    velero.io/storage-location: default
  namespace: openshift-adp
spec:
  hooks: {}
  includedNamespaces:
  - mysql-persistent
  storageLocation: example-velero-1
  defaultVolumesToRestic: true
  snapshotVolumes: false
  ttl: 0h10m0s
```

Based on the parameters specified in the `BackupCustomResource`, the actual backup process will be triggered by OADP. In case of a scheduled backup, a `Schedule` CR has to be defined in addition.

5.4.4.1 Data Mover for CSI Snapshots

Another key feature in this release is the Volume Snapshot Mover or the Data Mover feature. The OADP Operator currently supports backup and restore of applications backed by CSI volumes by taking advantage of the Velero CSI plug-in. The problem with CSI snapshots on some storage solutions, such as IBM Storage Fusion Data Foundation is that these snapshots are local to the RHOCP cluster and cannot be recovered if the cluster get deleted accidentally or if there is a disaster. To overcome this issue, Data Mover is made available for users to save the snapshots in a remote storage.

VolSync is used as the default data mover for OADP. To use this feature the VolSync operator must be installed from the operator hub.

5.4.4.2 Multiple Label Selector

There is also support of the Multiple Label Selector. This enables the support for backing up resources based on multiple labels (or Logic) in a single backup config.

Multiple label selectors can be provided in the specification of the backup custom resource.

Sample definition of a `LabelSelectors` using CSI snapshots:

```
spec:
  orLabelSelectors:
  - matchLabels:
      app: mysql
  - matchLabels:
      testing: foo
```

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

More information on how to install and use the OADP Operator is included in the documentation of Red Hat OpenShift Container Platform:

Backing up applications

Backing up and restoring your OpenShift Container Platform cluster

Volume Snapshots

OADP Release Notes

NOTE	CSI backups might fail due to a Ceph limitation of VolumeSnapshotContent snapshots per PVC, and manual intervention might be needed to remove volumesnapshotcontents as the default policy of volumesnapshotclass is to retain and not delete.
NOTE	You can create many snapshots of the same persistent volume claim (PVC) but cannot schedule periodic creation of snapshots: For CephFS, you can create up to 100 snapshots per PVC. For RADOS Block Device (RBD), you can create up to 512 snapshots for each PVC. (OADP-804) and (OADP-975)

To further enhance the usage of OADP, it is also possible to use hooks to run commands in a container on a pod. This can give the application the chance to assume a well-defined state prior to the backup initiation, for example, `fsfreeze` to freeze a file system. Those hooks can be configured to run before or after a backup or restore. Restore hooks can run during initialization of the application container.

5.4.5 Restoring an application

OADP restores resources and persistent volumes from a backup archive. Restore can cover either all objects in the backup archive or it can filter the restored objects by namespace, persistent volume, or label.

For restoring an application, a `RestoreCustomResource` needs to be defined.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

Definition of a `RestoreCustomResource`

```
apiVersion: velero.io/v1
kind: Restore
metadata:
  name: mysql-restore
  namespace: openshift-adp
spec:
  backupName: mysql-backup
  excludedResources:
  - nodes
  - events
  - events.events.k8s.io
  - backups.velero.io
  - restores.velero.io
  - resticrepositories.velero.io
  restorePVs: true
```

OADP supports an incremental restore, which makes it possible to only restore data, which has been modified between two backups. To support incremental restore, a new parameter `existingResourcePolicy` has been added.

Definition of a `RestoreCustomResource` *for an incremental restore*

```
 apiVersion: velero.io/v1
kind: Restore
metadata:
  name: mysql-restore
  namespace: openshift-adp
spec:
  backupName: mysql-backup
  excludedResources:
  - nodes
  - events
  - events.events.k8s.io
  - backups.velero.io
  - restores.velero.io
  - resticrepositories.velero.io
  restorePVs: true
  ExistingResourcePolicy: update
```

Please see also this documentation for ExistingResourcePolicy: ExistingResourcePolicy

For a successful restore of an application to a new OpenShift cluster, it is important to maintain the consistency of User ID (UID) of the service account that runs the application. That UID will be used to run the application inside the Container.

It is a best practice that the application uses `securitycontextconstraints` targeting the service account in deployment to allow privileged containers. This means that the serviceaccount used in the pod is always same and have one UID on both backup and restore cluster. Create a SCC or use

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

an existing one that allows custom UID outside of normal allowed namespace UID ranges.

When the application also needs to be run under a specific custom UID, then the Pod definition needs to use a ServiceAccount with the RunAsAny privilege and the Pod needs to use the runAsUser to set the UID to the hardcoded UID. NOTE: Using a hardcoded UID is NOT recommended. Among issues with this approach is that it is prone to UID collisions with system UIDs or with UIDs of the same or different application running in a different Namespace expecting to use the same UID.

OpenShift commands to set UID

```
oc create sa sa-with-anyuid
oc adm policy add-scc-to-user anyuid -z sa-with-anyuid
```

Sample Deployment yaml

```
apiVersion: apps/v1
kind: Deployment
spec:
  template:
    spec:
      serviceAccountName: sa-with-anyuid
      securityContext:
        runAsUser: 1001
        fsGroup: 1001
```

For more information on Openshift UIDS, please refer Guide to OpenShift UIDS

The following section of a Red Hat FaQ document also go into more details on challenges and workarounds during the restore of an application on new cluster:

Can OADP restore routes with base domain from the restore cluster?

Can I turn off internal registry image backup?

5.4.6 Further considerations for backup and restore

The mechanism integrated into IBM Storage Fusion Data Foundation covers the backup of application-specific data and state. It is important to emphasize, that the underlying OpenShift Container Platform cluster implies additional backup considerations, which are outside of the scope of this document. Such considerations include the backup of the cluster's *etcd* database, which is the key-value store of OpenShift.

You can find more information about OpenShift backups here:

* Backing up the etcd database

Beyond the etcd database, there are additional backup strategies for the virtualization layer (e.g. z/VM or KVM).

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

Chapter 6. Summary

With IBM Storage Fusion and IBM Storage Fusion Data Foundation, organizations can add persistence storage to their Kubernetes based container infrastructure and build stateful applications. This document helps to ensure that your solution meets best-practices and recommendations when deployed on zSystems and LinuxONE or within z/OS Container Extensions (zCX). The Quality Engineering teams (QE) at Red Hat and IBM have tested and validated the consideration described in this document. This ensures organizations seeking to operationalize this reference architecture quickly can be assured that all options represented are both fully tested as well as fully supported.

In chapter 3 a solid understanding of the overarching concepts and architectural principles of IBM Storage Fusion Data Foundation has been sketched out. In addition, the chapter details how IBM Storage Fusion Data Foundation is positioned within IBM's strategic storage offering IBM Storage Fusion. This is the essential context technical leadership needs to have to guide real-life implementation projects towards a successful solution.

In chapter 4 the key design considerations have been detailed. This enables the technical team to make the correct decisions on networking, storage, installation, and sizing options based on a solid understanding of the concepts of IBM Storage Fusion Data Foundation. This also helps to avoid mistakes which need cumbersome mitigation in later stages of a project.

Chapter 5 focuses on aspects of "Day 2" operations like security, backup, and performance tuning. This chapter enables you to become productive quickly. It also helps to make your solution reliable and secure right from the beginning.

As IBM Storage Fusion Data Foundation evolves in the future and provides new capabilities, this document will be updated on a regular basis to include the latest best practices and considerations.

The information provided in this document is intended to encourage you to proceed with your IBM Storage Fusion Data Foundation based projects and deliver successful solutions to your customers.

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

Chapter 7. References

7.1 OpenShift® Data Foundation

IBM Storage Fusion documentation

* https://www.ibm.com/docs/en/storage-fusion/2.5

Red Hat OpenShift Data Foundation Documentation

- Red Hat OpenShift Data Foundation Documentation

Public Presentations

- Other Collateral

Blogs

- Introducing OpenShift Data Foundation, by Alex Handy
- OpenShift Data Foundation 4.3: Bringing dynamic storage to the cloud, by Karena Angell
- Introducing Red Hat Data Services With OpenShift Data Foundation 4.5, by Pete Brey
- Deploying OpenShift Data Foundation using Local Devices, by Annette Clewett, Kyle Bade
- OpenShift Services with OpenShift Data Foundation, by Marcel Hergaarden, Karan Singh

7.2 Red Hat OpenShift Container Platform

Red Hat OpenShift Container Platform Documentation

- Red Hat OpenShift Container Platform Documentation

Getting Started with Red Hat OpenShift

- Get started with Red Hat OpenShift
- Interactive Learning Portal
- Red Hat OpenShift Blog

8.3 LinuxONE and IBM zSystems

- IBM Knowledge Center
- Linux on IBM zSystems and IBM® LinuxONE, blog by Stefan Raspl

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.

Chapter 8. Contributors

The following individuals contributed to the production of this reference architecture. Feel free to send us any feedback about the document.

Maya Anilson	**Content author**	maya.anilson@in.ibm.com
Sravika Balusu	**Content author**	sravika.balusu@ibm.com
Nourhane Bziouech	**Content author**	nourhane.bziouech@de.ibm.com
Andrei Constantinescu	**Content author**	andrei.constantinescu@ibm.com
Manuel Gotin	**Content author**	manuel.gotin@ibm.com
Tuan Hoang	**Content author**	tuan.hoang1@ibm.com
Shiva Sai K	**Content author**	shiva.sai.k1@ibm.com
Abdul Kandathil	**Content author**	abdul.kandathil@ibm.com
Rishika Kedia	**Content author**	rishika.kedia@in.ibm.com
Melanie Manley	**Content author**	mjma@de.ibm.com
Michael Schaefer	**Content author**	schaefer.michael@de.ibm.com
Thomas Stober	**Project and content lead**	tstober@de.ibm.com

The following individuals contributed with content, reviews or feedback for this reference architecture.

Hendrik Brueckner	**Content reviewer**	brueckner@de.ibm.com
Daniel Dominguez Cuadrado	**Content reviewer**	dani@redhat.com
Murthy Garimella	**Content reviewer**	mgarimel@redhat.com
Wilhelm Mild	**Content reviewer**	wilhelm.mild@de.ibm.com
Filipe Miranda	**Content reviewer**	fmiranda@ibm.com
Daniel Alexander Parkes	**Content reviewer**	dparkes@ibm.com
Pedro Ibáñez Requena	**Content reviewer**	pedro@redhat.com
Eran Tamir	**Content reviewer**	etamir@ibm.com
Holger Wolf	**Content reviewer**	holger.wolf@de.ibm.com

© Copyright International Business Machines Corporation 2022, 2023. All rights reserved.