Redbooks
ibm.com/redbooks

High Availability and Disaster Recovery Options for IBM Power Systems Cloud and On-Premises

Dino Quintero

Shawn Bodily

Manas Kunnathodika

Antony Steel

Kim Poh Wong

Cloud

Power Systems

IBM

Redpaper

IBM Redbooks

High Availability and Disaster Recovery Options for IBM Power Systems Cloud and On-Premises

December 2021

Note: Before using this information and the product it supports, read the information in “Notices” on page vii.

First Edition (December 2021)

This edition applies to Version:
PowerHA SystemMirror Standard Edition 7.2.5
PowerHA SystemMirror 7.2.3 SP3
IBM AIX 7.2.5.1
IBM AIX 7.2.4 SP2
IBM Spectrum Scale 5.1.1.0 (ppc64le)
IBM Power Systems Virtual Server
IBM Virtual Machine Recovery Manager 1.4

This document was created or updated on December 13, 2021.

Contents

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, 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.

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.

AIX®
Db2®
DB2®
DS8000®
Easy Tier®
FICON®
FlashCopy®
GDPS®
HyperSwap®
IBM®
IBM Cloud®
IBM FlashSystem®
IBM SmartCloud®
IBM Spectrum®
IBM z Systems®
MQSeries®
Parallel Sysplex®
POWER®
Power10™
POWER8®
POWER9™
PowerHA®
PowerVM®
Redbooks®
Redbooks (logo) ®
Resilient®
S/390®
Storwize®
SystemMirror®
Tivoli®
WebSphere®
XIV®
z Systems®
z/OS®

The following terms are trademarks of other companies:

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.

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

UNIX is a registered trademark of The Open Group in the United States and other countries.

VMware, and the VMware logo are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.

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

Preface

This IBM® Redpaper publication positions the new high availability (HA) and disaster recovery (DR) options in the cloud against those options on-premises.

Hybrid cloud applications on IBM Power Systems are known for its high performance and reliability. The flexibility and available services options of IBM cloud ensured high availability and disaster recovery of these hybrid applications on Power Systems affordable and easy to use. This publication is intended to help with the basic requirements to configure and implement the HA and DR for a number of on-premises and cloud configurations.

This book addresses topics for IT architects, IT specialists, sellers and anyone looking to implement and manage high availability and disaster recovery on-premises and in the Cloud. Moreover, this publication provides documentation to transfer the how-to-skills to the technical teams, and solution guidance to the sales team. This book compliments the documentation available at IBM Documentation and aligns with the educational materials provided by IBM Systems Technical Education.

Authors

This paper was produced by a team of specialists from around the world working at IBM Redbooks, Poughkeepsie Center.

Dino Quintero is an IBM Redbooks® Project Leader with IBM Systems. He has 25 years of experience with IBM Power Systems technologies and solutions. Dino shares his technical computing passion and expertise by leading teams developing technical content in the areas of enterprise continuous availability, enterprise systems management, high-performance computing (HPC), cloud computing, artificial intelligence including machine and deep learning, and cognitive solutions. He also is a Certified Open Group Distinguished IT Specialist. Dino is formerly from the province of Chiriqui in Panama. Dino holds a Master of Computing Information Systems degree and a Bachelor of Science degree in Computer Science from Marist College.

Shawn Bodily is a six-time IBM Champion for Power Systems and a Senior IT Consultant for Clear Technologies in Dallas, Texas. He has 28 years of IBM AIX® experience and the last 24 years specializing in HA and DR primarily focused around IBM PowerHA® SystemMirror®. He has co-authored AIX and PowerHA SystemMirror certification exams. He is an IBM Redbooks platinum author who has co-authored over a dozen IBM Redbooks and IBM Redpaper publications.

Manas Kunnathodika is a Presales Solution Architect Leader & Coach in Kyndryl Global Solutioning Hub. He has more than a decade of experience in IT Systems and Service Management Tools & Automation. His area of expertise is Architecture & Solutioning of AIOps and Hybrid/Multi-cloud Observability & Automation Platforms. He is an Open group Certified Expert Enterprise Architect and IBM Professional Certified Cloud Solution Architect. In his previous roles at IBM, he was Singapore Country Service line Manager for Tools & Automation and ASEAN Leader for Hybrid Service Technologies. Manas is a Professional mentor for Open-source technologies and Technical solution enabler for DevOps toolchain and SRE practices. Manas holds Bachelor of Technology degree in Electronics & Communication Engineering. He is a professional Member of Association of Enterprise Architects, IBM Cloud® Advisory and Eminence Board & Singapore Computer Society.

Antony Steel (Red) is a Senior technical staff member with an ASEAN IBM Business partner (Belisama) and is based in Singapore. He has over 30 years experience with UNIX, predominately AIX and Linux, even going to the extent of confounding the IBM team in Singapore by running Fedora on his Mac. After many years in IBM with support, ATS and lab services he set up a small company focussing on his passion - food. In between meals he installs, configures, troubleshoots and deploys AIX, IBM PowerVM®, PowerSC, PowerHA SystemMirror, PowerVS and GPFS (Spectrum Scale). He is also an IBM Champion and has co-authored a number of IBM Redbooks and assisted with the preparation of AIX and HA certification exams.

Kim Poh Wong is a Senior Technical Staff Member in Singapore. He has more than 30 years of experience in Information Technology field. He holds a Master of Business degrees in IT from Curtin University. His areas of expertise include Continuity Management and Critical Situation resolution. He has written extensively on Emergency Preparedness.

Thanks to the following people for their contributions to this project:

Wade Wallace
IBM Redbooks, Poughkeepsie Center

Jerry Cartwright, Neil Clark, Kyle Morrison, Joe Cox
Clear Technologies an IBM Business Partner

Dan Simms
Precisely

Mark Watts
Rocket Software

Tom Huntington
HelpSystems

Ash Giddings
Maxava

Brian Sherman
IBM Canada

Steven Finnes
Power Systems Product Management
PowerHA SystemMirror, VM Recovery Manager, CBU

A. Ravi Shankar
IBM Distinguished Engineer
Hybrid Cloud Resiliency
Cognitive Systems Software Development

Kevin R Gee
Capgemini Engineering

Maddison Lee
IBM Summit Intern

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our papers to be as helpful as possible. Send us your comments about this paper or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, IBM Redbooks
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

Introduction

This chapter addresses some of the common concepts and defines the associated terms used in IT infrastructure and generally referred to as Reliability, Availability, and Serviceability (RAS). Although this paper is focused on IBM Power Systems, the concepts and terms used here are fairly generic and applicable across most IT infrastructure.

The chapter contains the following topics:

- 1.1, “Overview” on page 2.
- 1.2, “High availability” on page 5.
- 1.3, “Continuous operations” on page 6.
- 1.4, “Continuous availability” on page 6.
- 1.5, “Business continuity” on page 6.
- 1.6, “Disaster recovery” on page 7.
- 1.7, “Review of planning” on page 11.

1.1 Overview

Today's enterprises can no longer afford planned or unplanned system outages. Even a few minutes of application downtime can result in considerable financial losses, eroded customer confidence, damage to brand image, and public relations problems.

To better control and manage their IT infrastructure, enterprises have concentrated their IT operations in large (and on demand) data centers. These data centers must be resilient (and flexible) enough to handle the ups and downs of the global market. They must also manage changes and threats with consistent availability, security and privacy, both around the clock and around the world. Most of the solutions are based on an integration of operating system clustering software, storage, and networking.

How a system, server or environment handles failures is characterized as its reliability, availability and serviceability (RAS). In today's world of e-business, the reliability, availability and serviceability of an operating system and the hardware on which it executes have assumed great importance.

Today's businesses require that IT systems be self-monitoring, self-healing, maintained without outage and support 7x24x365 operations. More and more IT systems are meeting this requirement through techniques such as redundancy and error correction, to achieve a high level of RAS.

The reliability, availability and serviceability characteristics are a significant market differentiator in the UNIX server space and one where IBM AIX and IBM i excel. This has resulted in IBM Power Systems servers attaining the RAS levels close to those considered to be available only on the mainframe systems. These levels are often referred to in measurements of *nines* of availability and the downtime associated with each level is shown in Table 1-1 on page 3.

1.1.1 Downtime

Downtime is simply any/all period(s) during which an application or service is unavailable to serve its clients. Downtime can be classified in two categories as follows:

- Planned:
  - Hardware upgrades.
  - Hardware or software repair or replacement.
  - Software (operating system and application) updates or upgrades.
  - Backups (offline backups).
  - Testing (periodic testing is required for cluster validation).
  - Development.
- Unplanned:
  - Administrator errors.
  - Application failures.
  - Hardware failures.
  - Operating system errors.
  - Environmental disasters.

Often times downtime is associated with unplanned outage time. However, in practice more downtime is the result of planned outages. Often planned outages are a necessary evil to help maintain systems to minimize unplanned outage risk. Downtime, as the name implies, is the exact inverse of uptime. Uptime is often a percentage reference of the amount of time a system's service(s) are available. Anything less than 100% obviously means that some amount of downtime has been experienced. Any downtime, planned or unplanned, generally counts against total uptime. A well thought out and implemented high availability solution can help minimize, and in some cases completely mask or prevent, planned maintenance requiring an outage. The amount of uptime by the nines, and its corresponding downtime in real time measurements are shown in Table 1-1.

Table 1-1 Six levels of nines and their availability times

Number of nines	Uptime%	Maximum annual downtime
Six (6)	99.9999	31.56 seconds
Five (5)	99.999	5 minutes 35 seconds
Four (4)	99.99	52 minutes 33 seconds
Three (3)	99.9	8 hours 46 minutes
Two (2)	99.0	87 hours 36 minutes
One (1)	90.0	36.5 days

Typically organizations will review their applications in terms of Recovery Time Objective (RTO - time till service resumes) and Recovery Point Objective (RPO - amount of data lost) to set the application's Service Level Agreement. Figure 1-1 shows the combination of events that make up RPO and RTO.

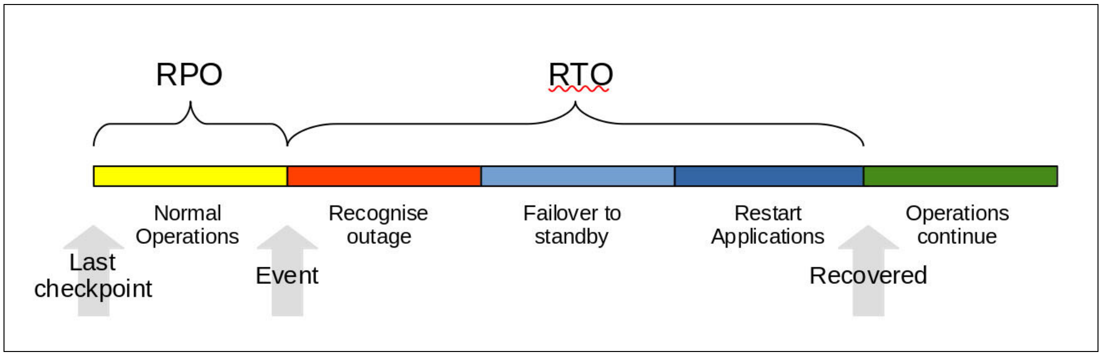


Figure 1-1 RTO and RPO

1.1.2 Single points of failure

One of the most common ways to increase availability, more so between one and two nines, is to minimize and preferably eliminate any *single points of failure* (SPOF) in the solution, providing a strong base on which to build higher levels of availability. Often the focus here is on the infrastructures and its multiple components. Many of these components are shown in Table 1-2.

Table 1-2 SPOF Components

Component	Method to eliminate or minimize SPOF
Server/Node	Use multiple servers/nodes.

Component	Method to eliminate or minimize SPOF
Power source	Use multiple power feeds and uninterrupted power supplies.
Virtualization	Duplicate the virtualization components.
Network adapter	Use multiple network adapters per server/node.
Network switch	Connect each server/node to multiple network switches.
Network	Have each server/node attached to multiple networks.
Storage	Use multiple storage subsystems and mirror across.
Disk/SAN adapters	Use multiple disk/SAN adapters per server/node.
SAN switch	Connect each server/node to multiple disk/SAN switches.
SAN	Connect each server/node to multiple SANs.
Disk	Use multiple disks with mirroring or RAID.
Application	Provide multiple instances of application on multiple servers if application allows. Otherwise utilize clustering and failover to another node/server.
Administrator/staff	Probably the most easily overlooked, so cross-train and utilize backups. Keep detailed operations documentation up to date and highly available (for example on company intranet and backed up).
Site	Provide an additional site.

In some cases there is the appearance of redundancy and somewhere along the way a SPOF is missed. For example, the two sites can be connected by multiple fiber links by way of one service provider. Well what if that service provider has an outage? To avoid this, it is recommended to use two separate service providers and ensure the communication links follow different routes between the sites and they also share no common entry point to each site. This type of diverse redundancy sometimes can seem like a never ending process.

1.1.3 Key recovery objectives

Generally there are a few key objectives to recoverability and they are as follows:

– Network Recovery Objective (NRO):

  How long it takes to switch over network access.

– Recovery Scope:

  This defines which resources are part of a backup. This will be defined according to the business goals and criticality of the business service.

– Recovery Time Objective (RTO):

  What is an acceptable amount of time to be without system access:

  - Minutes to a couple hours generally require automated recovery.
  - Hours to days can allow manual recovery steps.

– Recovery Point Objective (RPO):

  After an outage occurs, how much, if any, data is an acceptable to either recreate or do without.

  - If zero, then synchronous replication is required.
  - If greater than zero, then asynchronous replication might be suitable.

- Consistency:

  After successful recovery from backup, the data needs to be checked for consistency. There are two major consistency concepts are used across.

  i. Crash consistency

     The restored data bytes match the ones in the primary system at the time of the crash.

  ii. Application consistency

     Applications are able to access data from the time of the backup without failure.

- Service Level Agreements (SLA):

  Agreement between service provider and client defining the Disaster Recovery strategy and design for stated business continuity and service resiliency requirements.

To answer these questions accurately often a risk and requirement analysis needs to be performed in combination with a downtime cost analysis for each and every service. Organizations need to go beyond stating that their disaster recovery objectives are zero across the board as this is often simply unattainable, and does not recognize the different value of each application to the organization.

The following sections discuss the concepts of continuous availability features in more detail.

1.2 High availability

High availability is the attribute of a system which provides service during defined periods, at acceptable or agreed-upon levels, and masks both planned and unplanned outages from end users. It often consists of redundant hardware components, automated failure detection, recovery, bypass reconfiguration, testing, problem determination and change management procedures.

In addition, high availability is also the ability (and associated processes) to provide access to applications regardless of hardware, software, or system management issues. This is achieved through greatly reducing, or masking, planned downtime. As noted planned downtime often includes hardware upgrades, repairs, software updates, backups, testing, and development.

High availability solutions help eliminate single points of failure through appropriate design, planning, selection of hardware, configuration of software, and carefully controlled change management discipline. High availability does not mean *zero* interruption to the application; therefore, it is called fault *resilient* instead of fault *tolerant.*

A highly available environment typically includes more demanding recovery time objectives (seconds to minutes) and more demanding recovery point objectives than a disaster recovery scenario. High availability solutions provide fully automated failover to a alternate system so that users and applications can continue working with minimum disruption. HA solutions must have the ability to provide an immediate recovery point. At the same time, they must provide a recovery time capability that is significantly better than the recovery time that you experience in a non-HA solution.

1.3 Continuous operations

Continuous operations is an attribute of IT environments and systems which allows them to continuously operate and mask planned outages from end users. Continuous operations employs nondisruptive hardware, software, configuration and administrative changes.

Unplanned downtime is an unexpected outage and often is the result of administrator error, application software failure, operating system faults, hardware faults, or environmental disasters.

Generally, hardware component failure represents an extremely small proportion of overall system downtime. By far, the largest single contributor to system downtime is planned downtime. For example, shutting down a computer for the weekend is considered planned downtime. Stopping an application to take a full system backup (level 0) is also considered planned downtime.

1.4 Continuous availability

Continuous availability is an attribute of a system which allows it to deliver nondisruptive service to end users 7 days a week, 24 hours a day by preventing both planned and unplanned outages. The traditional view is that continuous availability or the elimination of downtime is the sum of continuous operations (the masking or elimination of planned downtime) and high availability (the masking or elimination of unplanned downtime).

Most of today’s solutions are based on an integration of the operating system with clustering software, storage, and networking. When a failure is detected, the integrated solution will trigger an event that will perform a predefined set of tasks required to reactivate the operating system, storage, network, and in many cases, the application on another set of servers and storage. This kind of functionality is defined as IT continuous availability. Scaled out solutions using multiple instances of the application can also provide continuous availability as the failure of a single instance will not impact the overall availability of the application.

The main goal in protecting an IT environment is to achieve continuous availability; that is, having no end-user observed downtime. Continuous availability is a collective term for those characteristics of a product which make it:

- Capable of performing its intended functions under stated conditions for a stated period of time (reliability).
- Ready to perform its function whenever requested (availability).
- Able to quickly determine the cause of an error and to provide a solution to eliminate the effects of the error (serviceability).

Continuous availability encompasses techniques for reducing the number of faults, minimizing the effects of faults when they occur, reducing the time for repair, and enabling the customer to resolve problems as quickly and seamlessly as possible.

1.5 Business continuity

The terms business continuity and disaster recovery are sometimes used interchangeably (as are business resumption and contingency planning). Here, *business continuity* is defined as the ability to adapt and respond to risks, and opportunities to maintain continuous business operations. However, business continuity solutions applied in one industry might not be applicable to a different industry, because they can have different sets of business continuity requirements and strategies.

Business continuity is implemented using a plan that follows a strategy that is defined according to the needs of the business. A total business continuity plan has a much broader focus and includes items such as a crisis management plan, business impact analysis, human resources management, business recovery plan procedure, test plan, documentation and so on.

1.6 Disaster recovery

For our purpose, *disaster recovery* is defined as the ability to recover a data center at a different site if a disaster destroys the primary site or otherwise renders it inoperable. The characteristics of a disaster recovery solution are that IT processing resumes at an alternate site on completely separate hardware.

Disaster recovery (DR) is a coordinated activity to enable the recovery of IT and business systems in the event of disaster. A DR plan covers both the hardware and software required to run critical business applications and the associated processes, and to (functionally) recover a complete site. The DR for IT operations employs additional equipment (in a physically different location) and the use of automatic or manual actions and methods to recover all of the critical business processes.

Every location, although different, will have some type of disaster to worry about. Fire, tornadoes, floods, earthquakes, and hurricanes can have far reaching geographical impacts. This drives remote disaster sites to be further and further apart. In some cases industry regulations can also determine the minimum distance between sites. Some important questions about designing for disasters are:

- What is the monetary impact to the business in case of a disaster?
- How soon can the business be back in production?
- At what point in time can it be recovered to?
- What communication bandwidth is required and can be afforded?
- What disaster recovery solution(s) are viable based on the inter-site distance requirements?
- What disaster recovery solution(s) are viable based on the application requirements?

Disaster recovery strategies cover a wide range from no recovery readiness to automatic recovery with high data integrity. Data recovery strategies must address the following issues:

- Data readiness levels:
  - Level 0:

    None. No provision for disaster recovery or off-site data storage.
  - Level 1:

    Periodic backup. Data required for recovery up to a given date is backed up and sent to another location.
  - Level 2:

    Ready to roll forward. In addition to periodic backups, data update logs are periodically sent to another location, either using physical media or electronically. Recovery point is up to the latest update log at the recovery site.

  - Level 3:

    Roll forward or forward recover. A shadow copy of the data is maintained on disks at the recovery site. Data update logs are received and periodically applied to the shadow copy using recovery utilities.

  - Level 4:

    Real time roll forward. Like roll forward, except updates are transmitted and applied at the same time as they are being logged in the original site. This real-time transmission and application of log data does not impact transaction response time at the original site.

  - Level 5:

    Real time remote update. Both the original and the recovery copies of data are updated before sending the transaction response or completing a task.

- Site interconnection options:

  - Level 0:

    None. There is no interconnection or transport of data between sites.

  - Level 1:

    Manual transport. There is no interconnection. For transport of data between sites, dispatch, tracking, and receipt of data is managed manually.

  - Level 2:

    Remote tape. Data is transported electronically to a remote tape. Dispatch and receipt are automatic. Tracking can be either automatic or manual.

  - Level 3:

    Remote disk. Data is transported electronically to a remote disk. Dispatch, receipt, and tracking are all automatic.

- Recovery site readiness:

  - Cold:

    A cold site typically is an environment with the proper infrastructure, but little or no data processing equipment. This equipment must be installed as the first step in the data recovery process. Both periodic backup and ready to roll forward data can be shipped from a storage location to this site when a disaster occurs.

  - Warm:

    A warm site has data processing equipment installed and operational. This equipment is used for other data processing tasks until a disaster occurs. Data processing resources can be used to store data, such as logs. Recovery begins after the regular work of the site is shut down and backed up. Both periodic backup and ready to roll forward data can be stored at this site to expedite disaster recovery.

  - Hot:

    A hot site has data processing equipment installed and operational, while the data can be restored either continually or regularly to reduce recovery time.

  - Active-active:

    A subset of the applications are active in both sites at the same time.

There are many common things to take into account in almost every disaster recovery solution. Some of these include:

- Systems provisioned for the purpose of disaster recovery are of a different type, size, and capacity than production.
- User and group permission problems.
- Application licenses tied to hardware.
- Some local high availability options such as multiple instances of an application no longer exist at the DR site if services are combined on same server.
- Production applications tied to a specific network address or network name during installation.
- Node name and host name conflicts (conflicts between existing systems in the disaster recovery site and the new systems being implemented under the disaster recovery plan).
- Multiple implementation standards for various functional system types such as stand-alone, high availability, and disaster recovery.
- Networking name or address conflicts.

  The best solution for avoiding networking conflicts during a disaster recovery implementation is to always ensure that each network address (TCP/IP) or name has a unique value across the enterprise. In organizations with multiple active data centers, network addresses (TCP/IP) from the production data center should not be failed over to the disaster recovery site. To do so requires reconfiguration of routers and switches, and it can endanger the existing production systems running in the data center accepting the disaster recovery workload. Therefore, the production applications should never be tied to or dependent upon a specific network TCP/IP address because, in a disaster, those network TCP/IP addresses change, causing the applications not to work. Applications and regular users should never use or specify a network service by its TCP/IP address, and they should only use a symbolic name. Furthermore, the symbolic name used by applications and regular users should only be an alias and point to a host name.
- User names.

  Each person in an organization should be assigned a unique identifier across the enterprise that is only assigned to that person and retired when they leave the organization. This ensures a seamless audit trail when evaluating problems, issues, and actions. The user name should consist of alphanumeric characters and be a valid structure for all systems within an organization so that each person only has one user name. Specifying a user name structure that works on all systems and provides enough variability can be a daunting task for organizations as typically they use a wide variety of operating systems, each with its own requirements for user name structures including password management.
- File system/mount point names.

  To ensure the ability to recover multiple instances of an application onto a single system in a disaster recovery scenario, each file system containing application files should have a unique mount point directory across the enterprise. The best way to achieve this is to use the resource group name or a substring of the logical volume name as the top-level directory, considering that typically a file system mount point is required for each logical volume.

Other considerations for planning disaster recovery vary a bit for each application environment. The connectivity options and the distance between sites will also dictate what type of data replication options are available to use. There is a careful balance required between the bandwidth required and latency encountered when traversing greater distance. Though technologies might support “unlimited” distance, this does not always mean it is possible or even feasible to implement it.

Now when combining these, you get the seven tiers of disaster recovery as shown in Figure 1-2. They are:

- Tier 0

  This one is simple, 0 means zero. There is no off-site, nor off-site data. Usually recovery must be local.

- Tier 1

  Usually backups only on tape, and should be offsite. However they are not kept at any site where hardware exists to be utilized in performing the recovery. This can even be a cold site but often is a storage data vault.

- Tier 2

  Often is offsite backups on tape and stored offsite at least at warm site, and more likely should be a hot site.

- Tier 3

  Generally means that data is transmitted electronically, at least critical data, to the hot recovery site. Allows for shorter recovery time of critical data and services.

- Tier 4

  Usually point-in-time copies, like a IBM FlashCopy®, to a hot site. This too can be in both directions.

- Tier 5

  Data is continuously copied to the remote hot site utilizing a two phase/sites commit. This can be storage, host, or application based replication.

- Tier 6

  From a data perspective it is zero or near zero data loss with instantaneous recover. This often is storage-based replication.

- Tier 7

  In addition to tier 6, automation of recovery procedures to restore the services is included. This is the highest level of protection available.

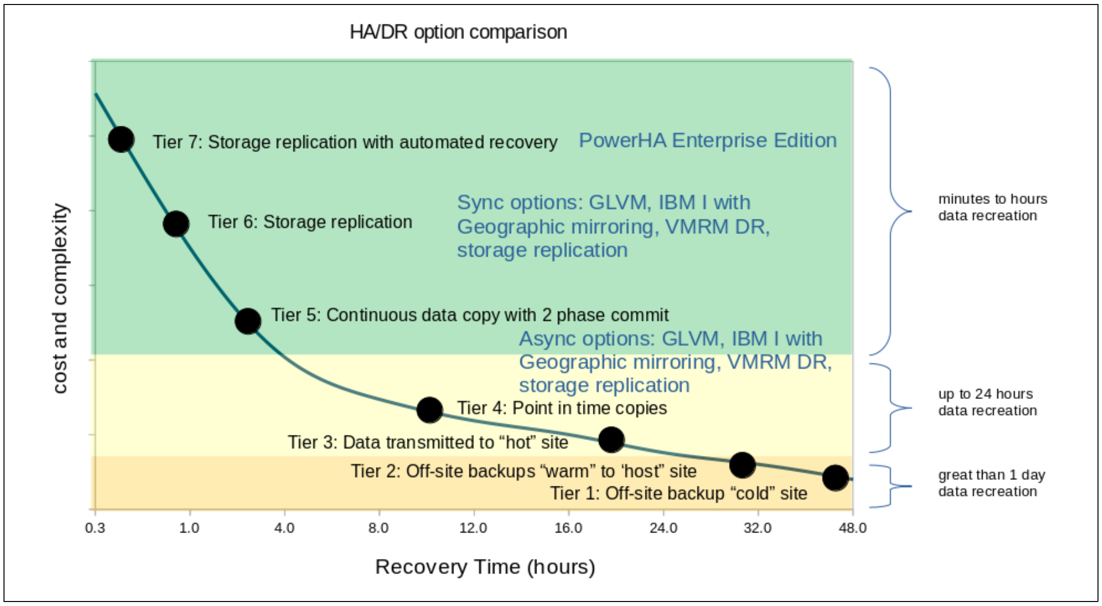


Figure 1-2 Tiers of HA/DR

1.6.1 Hybrid cloud disaster recovery

A hybrid cloud application is a mix of on-premises, private, or public cloud platforms with orchestration between these distributed platforms/workloads to perform as a single business service. The flexibility, agility, scalability, and interoperability of a hybrid cloud environment creates a perfect platform to run business-critical applications. The hybrid cloud applications built on IBM Power Systems are known for their high performance and reliability. Although public cloud service providers ensure high availability through data center redundancy, it is not always sufficient to protect from human or system errors or natural disasters hitting the services on a hybrid cloud application. The recent public cloud outages also point towards the need of a well-designed disaster recovery solution for your critical applications. Many protected environments, even IBM Power Systems, can fail due to a single or multiple failures. To prepare for those scenarios, it is important to proactively define your disaster recovery strategy and design.

In general, we consider two possible approaches for hybrid cloud disaster recovery: application driven and underlying technology driven. In application driven disaster recovery approach we assume the application is built to be DR ready with replication enabled across multiple instances. While technology driven disaster recovery approaches use data replication with DR site management or reprovisioning enabled by way of a recovery orchestrator.

1.7 Review of planning

The following are also critical components of a successful HA or DR environments:

- Planning.
- Monitoring.
- Maintaining.
- Documenting.
- Testing.

1.7.1 Planning

An important component of planning an overall HA or DR plan is to regularly review the organizations applications by the required RTO and RPO and ensure that the HA/DR solutions delivers to those requirements. Chapter 2, “High availability and disaster recovery concepts and solutions” on page 15 provides an overview of the options available and what they provide.

The other components are covered elsewhere, but include:

- Risk analysis and review of types of possible disaster.
- Planning network throughput and latency, while reducing risks of both data centers being impacted by the same disaster.
- Planning resources for normal operations and during recovery from disaster.

1.7.2 Monitoring

Monitoring the entire environment is important. It provides the capability to find and fix problems before they lead to an outage. For example, when a redundant component fails its imperative the component is fixed or replaced to continue to provide the original level of redundancy. Undetected or unresolved problems can accumulate over time removing redundancy and ultimately leading to an outage.

1.7.3 Maintaining

Although problems found by way of monitoring often leads to maintenance, it is not the only component of maintaining an environment. Normal maintenance often includes, but not limited to:

- Backups.
- Install operating system updates.
- Install application updates.
- User access and password management.
- Old data and files cleanup.
- Problem detect and fix.
- Security scans.

1.7.4 Documenting

Documenting can be a time consuming and often seen as a thankless task. However it is important even during normal operations but can prove to be even more crucial in an emergency. This can be done a variety of ways and recommended to be kept on the company's intranet when possible. Also just like the overall environment it too requires being constantly maintained. Often scripts and automation of tasks can help with keeping system documentation current.

Another critical component of documentation that is often ignored is the post-outage review and update. After every incident it is important to improve your organizations HA/DR by learning from the experience - improving your monitoring so the event will be captured and updating documentation, training and testing.

1.7.5 Testing

All the best laid plans and solutions are relatively worthless if they are never tested. Testing is as crucial as any other process. All change and management procedures need to be tested in a non-production environment before ever being implement in production. Then in staying with the theme of this paper, all HA and DR solutions need to be methodically tested on a regular basis. It is better to find a problem during planned testing than during an unplanned outage.

1.7.6 Comparison of options

While at a high level, the solutions presented in this publication apply equally to the following scenarios:

- Availability with a data center (on-premises).
- Availability / DR across two data centers (on-premises).
- Availability within the Cloud.
- Availability / DR between on-premises and the Cloud.
- Availability / DR between two Clouds (different providers or zones).

There are some differences and limitations offered by each solution as shown in Table 1-3 on page 13, However we will generally refer to HA and DR between data centers to cover all the options. Table 1-3 on page 13, Table 1-4 on page 13, and Table 1-5 on page 14 are intended to be a quick reference for users to see at a high level which solution will meet their particular requirements.

Table 1-3 Availability solution options for different data center configurations

Option	Within one data center	On-premises to on-premises	On-premises to cloud	Within cloud	Cloud to cloud
LPM	Yes	Yes			
SSR	Yes				
VMRM HA	Yes				
VMRM DR		Yes			
PowerHA Std	Yes				
PowerHA Std cross site		Yes		Yes	
PowerHA EE	N/A	Yes			
PowerHA EE with GLVM		Yes	Yes	N/A	Yes
PowerHA EE with I Geo Mirror		Yes	Yes	N/A	Yes
GLVM stand alone		Yes	Yes	N/A	Yes

Note: Although GLVM is available stand-alone, IBM i Geographic Mirror requires PowerHA.

Table 1-4 Replication options for different data center configurations

Replication	Within one data center	On-premises to on-premises	On-premises to cloud	Within cloud	Cloud to cloud
None (scale out)	Yes	Yes	Yes	Yes	Yes
Storage managed	Yes	Yes			
Application managed	Yes	Yes	Yes	Yes	Yes
GLVM Stand alone	N/A	Yes	Yes	N/A	Yes
Spectrum Scale stretched cluster	N/A	Yes	Yes	N/A	Yes
Spectrum Scale AFM/DR	N/A	Yes	Yes	N/A	Yes

When planning DR options, there will be some differences depending on the nature of the data centers. Typically for an on-premises solution the organization will have the ability to manage the whole infrastructure, while for a cloud solution, the organization will only be able to manage from the operating system up. This will restrict the replication choices available and can also limit the network bandwidth and choices.

Table 1-5 Management options for different data center configurations

Control	Within one data center	On-premises to on-premises	On-premises to cloud	Within cloud	Cloud to cloud
Manage the hypervisor and up	Yes	yes	only on-premises		
Manage virtualisation layer	Yes	yes	only on-premises		
Manage the storage	Yes	yes	only on-premises		
Provision the storage	Yes	Yes	Yes	Yes	Yes
Manage the OS and up	Yes	Yes	Yes	Yes	Yes
Shared network	Yes	Yes	Yes	Yes	Yes

2

High availability and disaster recovery concepts and solutions

This chapter discusses some of the basic requirements for building a highly available and a disaster recovery solution. The last section will look at how this foundation is used by some of the many highly available and disaster recovery solutions available. In a typical data center a range of solutions is required as applications vary, some have built in availability and each will have its own service level agreements and recovery times.

This chapter contains the following:

2.1 High availability and disaster recovery concepts

The following concepts will be used in this section:

Split-brain/Split cluster A cluster split-brain can occur when a subset of nodes in a cluster cannot communicate with the remaining nodes. While it is possible for this to occur within the data center, it is far more likely to happen to a cluster across data centers due to the greater exposure of the interconnecting networks to potential risk.

In a split-brain situation, the two partitions have no knowledge of each other's status, each of them believing that the nodes in the other partition are offline. As a consequence, each partition tries to bring online the other partition's applications and access the shared resources, an action highly likely to result in lost or corrupted data on the shared storage.

Tie breaker / 3rd site In HA/DR clusters it is recommended to use a tie breaker or a third site to protect from a split- brain. While it is still important to avoid this for clusters within a single data center, it is far less likely as multiple communication paths connect all nodes in the cluster and this is less common between sites.

The tie breaker feature uses a tie breaker resource to choose which partition survives and continues to operate when a cluster split-brain occurs. This feature prevents data corruption on the shared or replicated disks.

PowerHA SystemMirror uses tie breaker disks or a NFS share file to act as the tie breaker and split-merge policies to control the behavior of the cluster.

Split policy When a split-brain situation occurs, each partition attempts to acquire the tie breaker by placing a lock on the tie breaker disk or on the NFS file. The partition that holds the lock on the SCSI disk or reserves the NFS file wins, and the other loses.

All nodes in the winning partition continue to process cluster events, and all nodes in the losing partition attempt to recover according to the defined split and merge action plan. This plan most often implies either the restart of the cluster nodes, or merely the restart of cluster services on those nodes.

Merge policy There are situations in which, depending on the cluster split-brain policy, the cluster can have two partitions that run independent of each other. However, most often, it is a preferred practice to configure a merge policy that allows the partitions to operate together again after communications are restored between them.

In this second approach, when partitions that were part of the cluster are brought back online after the communication failure, they must be able to communicate with the partition that owns the tie breaker disk or NFS file. If a partition that is brought back online cannot communicate with the tie breaker disk or the NFS file, it does not join the cluster. The tie breaker device is released when all nodes in the configuration have rejoined the cluster.

The merge policy configuration must be of the same type as that for the split policy, for example using NFS based tie breaker.

Synchronous replication

Writes are committed at the remote storage before acknowledgement can be returned to the application. This delay significantly degrades the application performance and typically limits the distance between the application and the remote storage to around 80-120 km.

Asynchronous replication

Writes are cached locally in some form of non-volatile storage and acknowledgment is returned to the application. At some later point in time the write is committed to the remote storage and then the record removed from the local cache.
Asynchronous mode allows for much greater distances between sites and smooth peaks in I/O and allowing for a lower bandwidth network. It does however imply that, in a disaster, there will be data lost, with the cache size representing the maximum amount of data that can potentially be lost.

2.2 High availability and disaster recovery requirements

The underlying requirement is to remove all single points of failure in the environment. This includes redundancy options for servers, networks, storage, data centers and the surrounding infrastructure (people, printers, backups, and so on).

In this section we will examine:

- Basic system requirements.
- Network configuration.
- Storage configurations.
- Site requirements.
- AIX Cluster Aware AIX (CAA) for PowerHA SystemMirror.
- Other pre-requisites.

Generally applications can be broken down into two types:

- Scale out or concurrent.
- Clustered.

Scale out and concurrent solutions provide redundancy through the use of multiple instances, so the focus will be on ensuring that the surrounding infrastructure will allow client access to a number of the application instances, while assuring that sufficient instances of the application are available to meet workload requirements.

Clustered solutions rely more heavily on knowing the status of the infrastructure to keep the individual application available. The focus is on ensuring that the applications are only online as and where required, while ensuring that they have consistent access to the data. If the cluster splits, then nodes on either side should not start to operate independently.

2.2.1 Basic system requirements

There are many different ways to build a highly available environment. This chapter describes a subset of options.

Mirrored architecture

In a mirrored architecture, you have identical or nearly identical physical components in each part of the data center. You can have this type of setup in a single room (although this is not recommended), in different rooms in the same building, or in different buildings.

Figure 2-1 shows a high-level diagram of a typical cluster. In this example, there are two networks, two managed systems, two Virtual I/O Servers (VIOS) per managed system, and two storage subsystems. This example also uses the Logical Volume Manager (LVM) mirroring for maintaining a complete copy of data within each storage subsystem.

Figure 2-1 has a disk for the Cluster Aware AIX (CAA) repository disk on each storage subsystem. For details about how to set up the CAA repository disk, see 2.3.1, “Data replication latency and throughput challenges” on page 26.

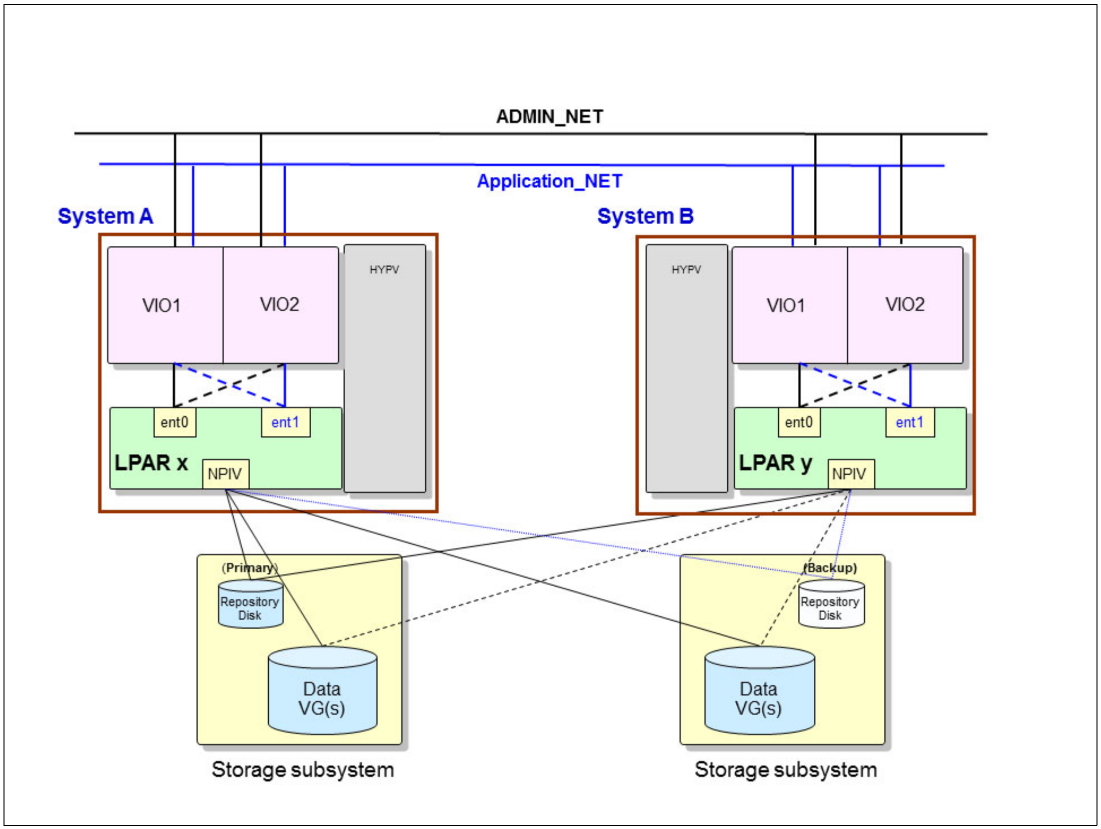


Figure 2-1 Cluster with multiple storage subsystems

2.2.2 Network configuration

This section focuses in the network considerations from an availability point of view and examines:

- Types of network, physical or virtual.
- Network adapters.
- Redundancy in networks.
- Inter-site considerations.

Physical or virtual

Use of technologies such as Live Partition Mobility (LPM), Simplified Remote Restart (SRR) and Virtual Machine Recovery Manager (VMRM) require the environment to be fully virtualized. Clustered solutions such as PowerHA SystemMirror, while there are some configuration differences, operate equally well in both the physical or virtual environment.

Network adapters

Network redundancy has been traditionally provided through the use of dual adapter networks. More recently single logical adapters are being used with their redundancy being provided by multiple physical backing devices. This can be done through bonding (Etherchannel), failover (Network Interface Backup), virtualisation (dual VIO Servers and Shared Ethernet Adapters) or a combination thereof.

Redundancy in networks

While in the past the use of redundant networks may have been common, it is now rarely the case as improvements in the design and operations of the network hardware, with the ability to have multiple paths has introduced greater redundancy.

Inter-site considerations

Care must be taken to ensure that the intersite connection does not become a single point of failure, so the following should be avoided:

- A single provider.
- A common entry point for client access to the application(s) at both sites.
- A common entry or exit point for the inter-site links.
- Common intermediate points.

Often different data centers will use different subnets. and while this can be handled by PowerHA SystemMirror and VM Recovery Manager DR, manual intervention may be required if other HA solutions are operated across sites.

This publication discusses the importance of planning the network bandwidth and latency to meet the application response time requirements. What is equally important is to plan a bandwidth that is sufficient for both normal operations as well the extra throughput required to recover and resynchronize a site after a disaster.

2.2.3 Storage configurations

This section describes different storage configurations.

Single storage architecture

In a single storage architecture, the common storage subsystem is shared by all the nodes. This solution can be used when there are lower availability requirements for the data, and is not uncommon when all nodes are in the same location.

If using storage based mirroring / replication, such as IBM SAN Volume Controller, the physical layout is similar to the mirrored architecture described in “Mirrored architecture” on page 18. However, from the Operating System perspective it is a single storage architecture as it is only aware of a single set of LUNs. However from the cluster management perspective it requires some extra administration to manage the underlying replication. For more information about the layout in a SAN Volume Controller stretched cluster, see “Stretched cluster” on page 20.

Figure 2-2 shows such a layout from a logical point of view.

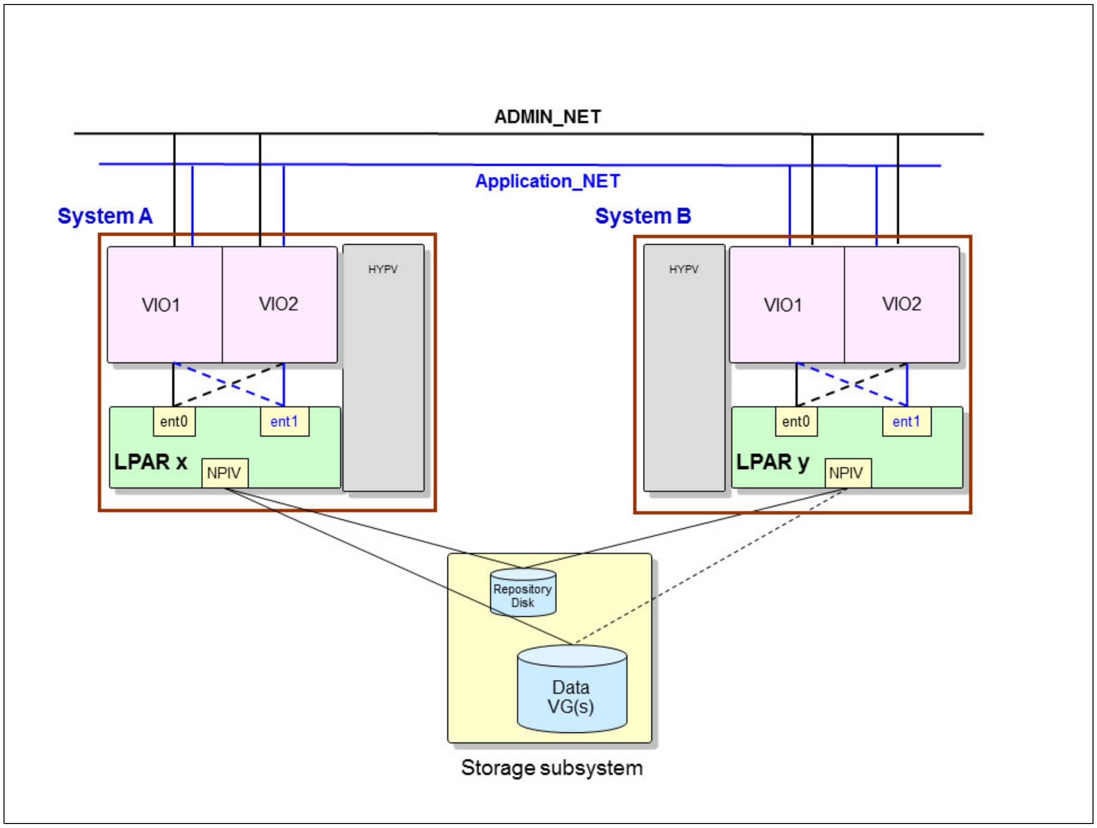


Figure 2-2 Cluster with single storage subsystem

Stretched cluster

A stretched cluster involves separating the cluster nodes into *sites*. A site can be in a different building within a campus or separated by typically less than 120 kilometers. In this configuration, the storage area network (SAN) spans the sites and storage can be presented across sites.

Having both SAN and TCP/IP connectivity between sites, removes the site network as a single point of failure. Steps must still be taken to ensure that both, different providers and routes, are used so that there is no a common point that can be broken, preferably for both SAN and IP networks.

Another main concern is having redundant storage and verifying that the data within the storage devices is synchronized across sites. The following section presents a method for synchronizing the shared data.

Storage subsystem using a stretched configuration

The SAN storage subsystems can be configured in a *stretched* configuration. In the stretched configuration, the storage controller presents the two storage devices as one unit even though they are separated by distance. The storage subsystem keeps the data between the sites consistent.

The storage subsystem in a stretched configuration allows the cluster software to provide continuous availability of the storage LUNs even through the failure of a single component. With this combination, the behavior of the cluster is similar in terms of function and failure scenarios in a local cluster (Figure 2-3).

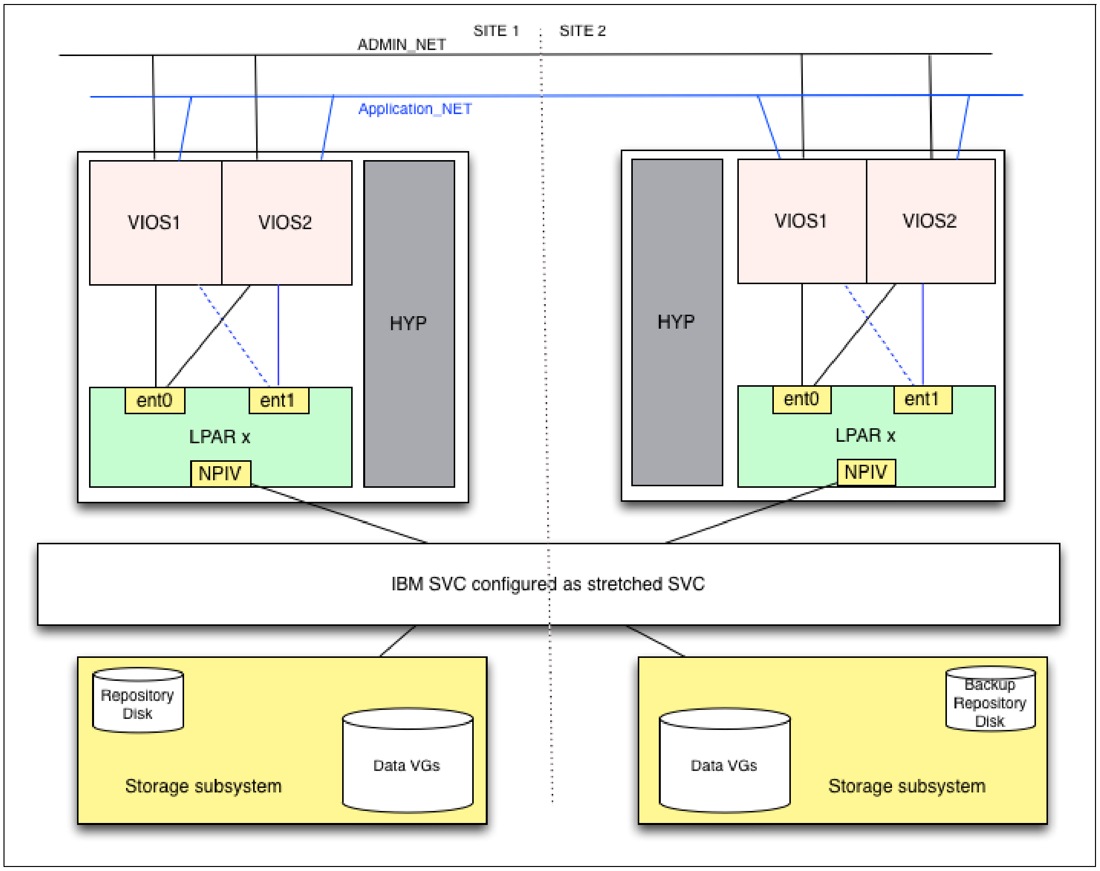


Figure 2-3 SAN Volume Controller stretched configuration

Linked cluster

A linked cluster is another type of cluster that involves multiple sites. In this case, there is no SAN link between sites due to a combination of cost and distance.

In this configuration, each site has its own copy of the repository disk and PowerHA SystemMirror keeps those disks synchronized.

As there is only one type of inter-site network, we have the IP network as a SPOF and must plan to reduce the possibility of it failing. It is especially important to ensure that there are multiple providers and routes to ensure that there is no loss of IP communications between the sites.

For more information about linked clusters see *IBM PowerHA SystemMirror 7.1.2 Enterprise Edition for AIX*, SG24-8106.

IBM supported storage that uses copy services

While here are several IBM supported storage devices with copy services capabilities, we will use the SAN Volume Controller (SVC) for the following example. The SVC can replicate data across long distances with the SVC copy services functions. The data can be replicated in either synchronous or asynchronous modes.

If there is a failure that requires moving the workload to the remaining site, the cluster software interacts directly with the storage to switch the direction of the replication. The LUNs are then presented to nodes at the surviving site and the clustering software actives the applications allowing access to users using the addresses appropriate for that site.

An example of this concept is shown in Figure 2-4.

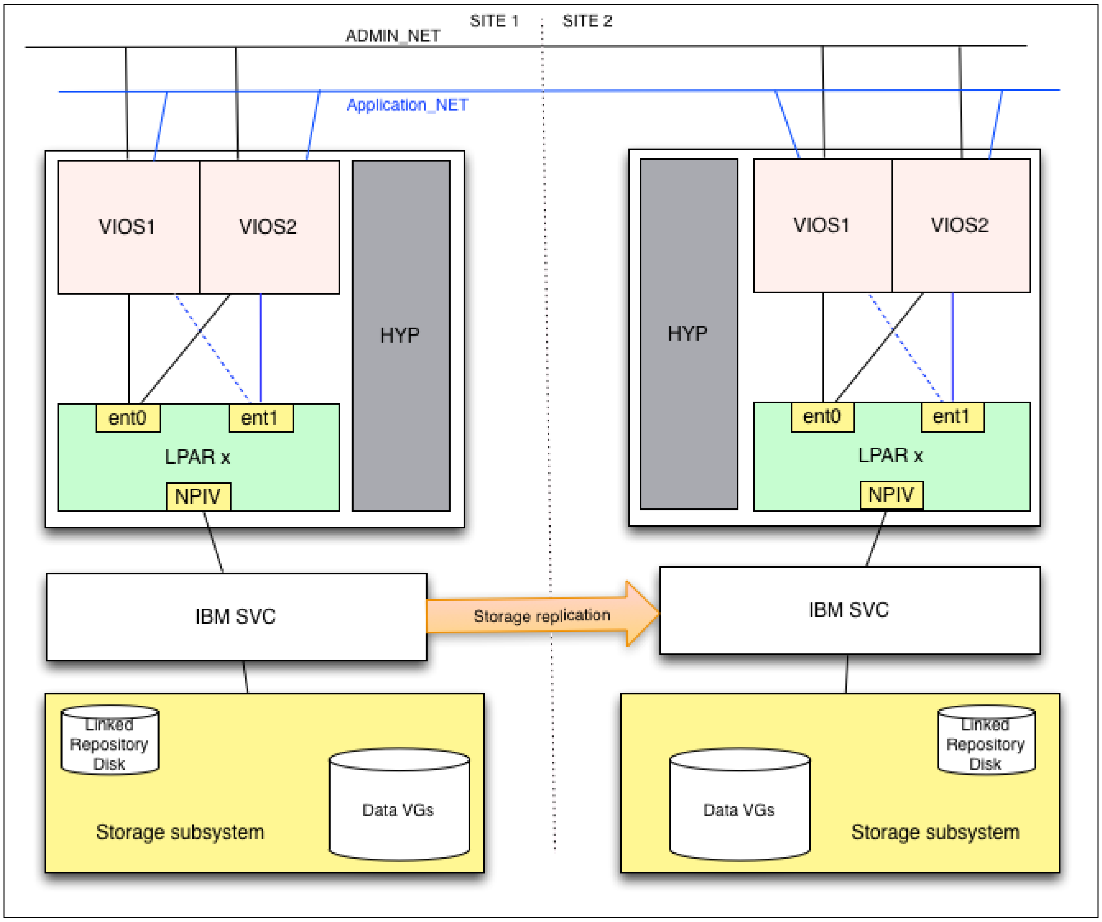


Figure 2-4 PowerHA SystemMirror and SAN Volume Controller storage replication

2.2.4 Site requirements

As discussed in the previous section there are some differences between using on-premises solutions compared to solutions in the cloud.

Other site planning should include:

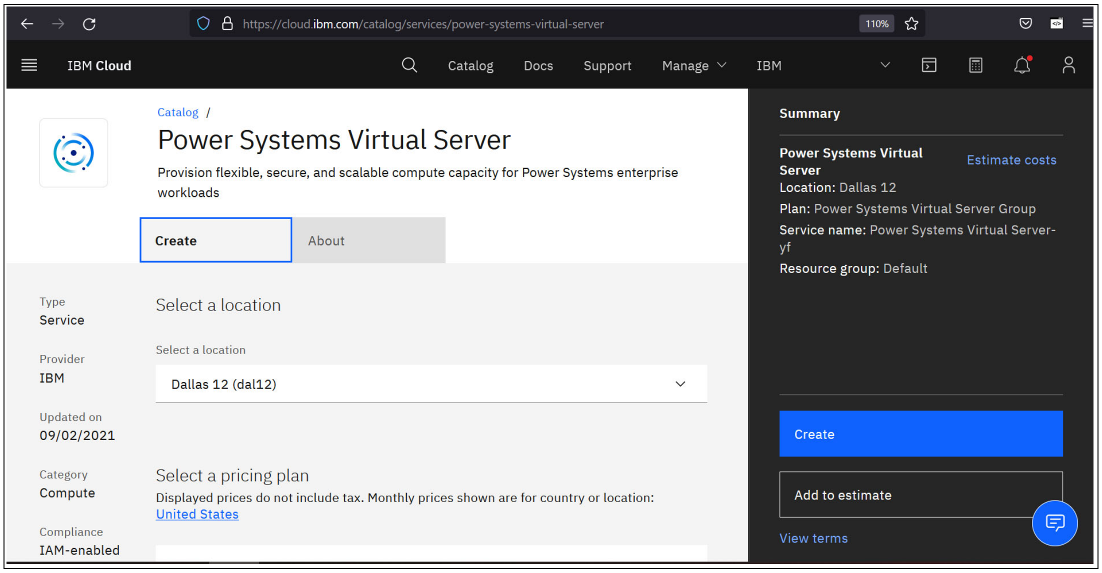
- Staff access and facilities at both sites.
- All the associated infrastructure that is required by the application.
- Availability of critical information, documents, backups and licenses.
- Distance between sites greater than what is affected by envisaged disasters.
- Access to backups.
- Access to qualified staff and documentation.
- Access to contracts, support services.
- Book outages for maintenance, at least 1/4ly for next 2 years[1]. If not tasks run test plan.

- Access to licenses and support contracts.
- Testing PowerHA SystemMirror assists with test tool (script your own test plan) VM Recovery Manager DR allows DR rehearsal.

IBM Power Systems Virtual Server (PowerVS) offering

Power Systems Virtual Server offering provides a secure and scalable server virtualization environment built on IBM Cloud platform for on-demand provisioning. The Power Systems Virtual Servers are located in the IBM data centers, distinct from the IBM Cloud servers with separate networks and direct-attached storage. The environment is in its own pod and the internal networks are fenced but offer connectivity options to meet customer requirements. This infrastructure design enables Power Systems Virtual Server to maintain key enterprise software certification and support as the Power Systems Virtual Server architecture is identical to certified on-premises infrastructure. The virtual servers, also known as logical partitions (LPAR), run on IBM Power Systems hardware with the PowerVM hypervisor.

Power Systems Virtual Server is available in IBM Cloud Catalog, under Compute → Virtual Machines. See Figure 2-5.



https://cloud.ibm.com/catalog/services/power-systems-virtual-server

IBM Cloud Catalog Docs Support Manage IBM

Catalog /

Power Systems Virtual Server

Provision flexible, secure, and scalable compute capacity for Power Systems enterprise workloads

Create About

Type
Service

Provider
IBM

Updated on
09/02/2021

Category
Compute

Compliance
IAM-enabled

Select a location

Select a location
Dallas 12 (dal12)

Select a pricing plan
Displayed prices do not include tax. Monthly prices shown are for country or location: United States

Summary

Power Systems Virtual Server Estimate costs
Location: Dallas 12
Plan: Power Systems Virtual Server Group
Service name: Power Systems Virtual Server-yf
Resource group: Default

Create

Add to estimate

View terms

Figure 2-5 IBM PowerVS

IBM PowerVS has the following features:

Powered by IBM Power Systems: Currently S-Class and E-Class systems running PowerVM.

Flexible compute: Configure your workloads with cores, types of cores, and memory resources, with dynamic resizing available.

AIX, IBM i, and Linux: Choose from a catalog of supported AIX, Linux and IBM i images or bring your own.

Reserved Instance Savings Plan: Leverage up to 45% discount with 3-year Reserved Instance Savings Plan. Up to 30% discount with 1-year Reserved Instance Savings Plan.

[1] Set an aspirational target and expect the business to negotiate it back

HA and DR options with PowerVS

The Power Systems Virtual Server instance restarts the virtual servers on a different host system if a hardware failure occurs. This process provides basic High Availability capabilities for the Power Systems Virtual Server service. For more advanced HA or DR options, deploy the following solutions in your environment.

- PowerHA SystemMirror Standard Edition (between pods).
- PowerHA SystemMirror Enterprise Edition (between data centers).
- IBM Cloud Disaster Recovery Solutions.

IBM Cloud disaster recovery solutions

IBM Cloud offers built-in capabilities and services for business continuity, resiliency and security. IBM Cloud DR solutions are categorized into three major areas.

- Management: Improve the management of infrastructure, apps, processes and entire cloud environments.
- Migration: Move existing applications and data to the cloud with a portfolio of disaster recovery focused migration tools and services.
- Storage: Scale capacity without interruption and deploy globally to achieve higher application performance.

IBM Backup as a Service

IBM Backup as a Service (BUaaS) from IBM offers fully managed, end-to-end data protection and data backup in a security-rich environment.

Benefits:

- Reliable data protection complies with government and industry regulations.
- Scalability based on your business needs.
- Remote management and operation.
- Monitoring solutions to ensure the health of data protection.

IBM Resiliency Services

IBM offers a full range of readily deployable services, solutions and technologies for data protection and recovery. This includes:

- Security & Resiliency Consulting Services.
- Disaster Recovery as a Service (DRaaS) for hybrid platform recovery.
- Data Protection with BUaaS.
- Cyber security and recovery.
- Data center services.

IBM Resiliency Disaster Recovery as a Service

IBM DRaaS offers continuous business resiliency of applications, infrastructure, data and cloud systems with health monitoring and comprehensive disaster recovery services.

Benefits:

- Less expensive OpEx based solution compared to self-managed on-premises model.
- Reliable disaster recovery orchestration with automation.
- Risk based approach to protect critical IT services.

- Data driven service environment for testing DR, patches and upgrades.

Migration

This section provides a few migration solutions options.

IBM Spectrum Protect Plus

IBM Spectrum® Protect Plus is a modern data resilience solution that provides recovery, replication, retention, and reuse for VMs, databases, applications, file systems, SaaS workloads, and containers in hybrid cloud environment.

Benefits:

- Easy to use and manage with SLA based policies, role-based access control (RBAC) and drill-down dashboards.
- Simple deployment as a virtual appliance or container application and easy to maintain with the agentless architecture.
- Seamless integration and data access by way of RESTful APIs.
- Supports data backup, recovery, replication for VMs, Windows file systems, databases, applications, SaaS workloads and containers; data retention and recovery on both on-premises and Cloud-based Object Storage.
- Available on IBM Cloud, Amazon Web Services and Microsoft Azure marketplaces.

Veeam on IBM cloud

Veeam on IBM Cloud can deliver reliable backup and predictable disaster recovery (DR) for virtual and physical workloads, wherever they reside, across your data center and the cloud.

Benefits:

- Supported with no-cost networking available between more than 60 global data centers for replication.
- Supports on-premises and on cloud backup and recovery.
- Available as software to use or as a Service model (BaaS).
- Long-term low-cost retention options with IBM Cloud Object Storage.

Zerto on IBM Cloud

Zerto provides disaster recovery and cloud mobility within a single, simple, scalable solution.

Benefits:

- Agentless, nondisruptive continuous data replication with journaling instead of snapshots, Zerto helps to deliver accelerated RTO in minutes and RPO in seconds.
- High speed global network backbone ensure resiliency with multi-site IBM Cloud DR environment without added cost.
- Easy to manage with application consistent recovery.
- Flexible SDDC and hardware configurations which can be automatically deployed.

Storage

This section provides a few storage solutions options.

Actifio GO on IBM Cloud

Actifio GO on IBM Cloud is the next generation multi-cloud Copy Data Management SaaS solution which enables customers to backup Enterprise workloads (VMware, Hyper-V, Physical Servers, SAP HANA, Oracle, SQL Server, and so on) directly to IBM Cloud while being able to instantly access the backup images within their data center.

IBM Cloud Backup

IBM Cloud Backup is a full-featured, automated, agent-based backup and recovery system managed through the IBM Cloud Backup WebCC browser utility.

Benefits:

- Implement and monitor backup policies from anywhere using web-based GUI.
- You can choose an IBM Data center or keep the backup outside the network.
- Recover from more than one facility using multi-vaulting capabilities.
- Scheduled backup with intelligent compression of data.
- End to end encryption with Deltapro Deduplication.
- Restoration options from previous backup or available multiple other recovery points.

IBM Cloud Object Storage

IBM Cloud Object Storage (COS) is a flexible, cost-effective and scalable cloud storage for unstructured data.

Benefits:

- Less expensive as you can save costs related to server, power, and data center space requirements.
- Streamlined storage environment for increased agility and reduced downtime.
- Supports exponential data growth and built-in high speed file transfer capabilities.
- Enhanced data security with role-based policies and access permissions.

2.3 Planning considerations

This section examines general planning considerations when planning highly available or disaster recovery configurations.

2.3.1 Data replication latency and throughput challenges

This section describes data replication latency and throughput challenges.

Network latency

Network latency is the time that it takes for messages to go across the network. Even when there is plenty of network bandwidth, it still takes a finite amount of time for the bits to travel over the inter-site link. The speed of the network is limited by the quality of the switches and the laws of physics and the network latency is proportional to the distance between the sites. Even if a network is capable of transmitting data at a rate of 120 kilometers per millisecond, it still adds up over a long distance. For example, if the sites are 60 km apart, all I/O must travel 60 km from the application to the remote storage. After the remote storage is updated, the result of the I/O request must travel 60 km back to the application. This 120 km round trip adds about 1 millisecond to each I/O request, and this time can be much greater depending on the number and quality of routers or gateways traversed. Suppose that the sites are 4000 km apart, so each I/O request requires an 8000 km round trip, adding approximately 67 milliseconds to each I/O. The resulting application response time is in most cases totally unacceptable. So synchronous mirroring is typically only practical, depending on the application, for metro distances, that is in the order of 100km or less. Greater distances typically necessitate asynchronous replication.

Network throughput

Another limitation on the operation of a DR site is the network bandwidth - think of this as the diameter of the pipe. The bigger the diameter, the more data that can be sent, but if the diameter is insufficient, then the data will backup until the flow is reduced - adding to the latency in the I/O, or filling the cache faster if using asynchronous replication.

Planning for the bandwidth to be sufficient to meet the peaks in your I/O may also mean that an expensive network is sitting idle for most of the time if peaks are rare, but as discussed, if the bandwidth is insufficient for peak I/O, then the application performance will suffer.

Planning both bandwidth and latency

Planning for latency is relatively simple and after the sites are selected can only be affected by the quality of the network hardware, as discussed. The application performance / user acceptance is the final arbiter in what is workable and, the I/O peak must not exceed the bandwidth of the network.

Planning the bandwidth is more difficult as not only should the bandwidth be sufficient for normal operations, but consideration must be given to recovery requirements. If there is a disaster, after recovered, the networks, depending on the topology, may have to support the extra activity as users catch up with lost processing and the system refreshing stale data at the recovering site.

2.3.2 Data divergence and recovery planning

Typically this problem is experienced if there is loss of access to the active site when asynchronous or time interval shipping of data is used. The organization has to decide whether to move production to the alternate data center using old data, or will waiting for the recovery of the failed active site fall within acceptable limits.

Should operations continue at the alternate site, the decision must be made when the failed site is recovered and if the *lost* data can be recovered:

- Move operations back to the recovered site and not recover the data cached there.
- Move operations back to the recovered site using the data there and discard the data created while running on the alternate site.
- Attempt to recover the cached data while using the recent data from the alternate site.

To make this decision, the organization will need to understand:

- The amount of data that can be lost and its potential value.
- Alternative (manual) methods to recover the data.
- Site recovery time.
- Is the failure localized or does it apply to the whole data center, and if localized, what will be the cost in moving all operations to the alternative site.

A good test plan, which is regularly executed, will assist in this planning and training staff in the procedures.

2.3.3 Quorum sites

As discussed, many automated DR solutions need to avoid creating a split-brain scenario due to the real possibility of losing or corrupting data. If nodes in the two data centers lose contact, then the clustering software will use the quorum site (also often called the third site or “laptop solution”) to determine which site should continue to operate.

The quorum site often will have a disk device (Fibre Channel or iSCSI) or file in an network shared file system and the ability to set a lock on this object will determine the surviving site.

2.4 Solutions

Over the last 10 years IT operations have evolved to the point where critical applications are rarely hosted on the same frame (server), nor in many cases, in the same data center. However, this development tends to be more piecemeal rather than being driven by a detailed review. A detailed review which examines the application requirements for high availability and disaster recovery, and then matches these requirements to the solutions available across all the whole infrastructure.

For many years IBM has been recognized as a leader in HA and DR solutions for workloads on IBM POWER® designed to meet the availability requirements of critical enterprise applications. In the recent years the portfolio has expanded to include protection for the “less critical” applications in the data center. These are the applications that can afford a slightly longer outage or have less stringent requirements around data loss. However if you are looking for a less complex and lower cost HA or DR solution, IBM now has a variety of LPAR restart options (for more details see “LPAR and VM restart options” on page 59).

It is worth noting that in ITIC’s 2020 Reliability poll[2] finds that 87% of respondents consider 99.99% (52.56 minutes) of unplanned per server/per annum downtime - to be the minimum acceptable level of reliability for mission critical servers and applications. This is coupled with a reported increase in mission critical business workloads by an average of 15% to 36% over the last three years. The same survey deals with estimated costs of outages, which while not under consideration here, needs to be taken into account when looking at pricing your HA or DR solution(s).

Now that IBM has a more comprehensive portfolio of HA and DR solutions, it is a good time to review what is available, what has changed and how these options will match your application availability requirements.

While the primary focus of HA and DR solutions is to work around failures in the infrastructure, these tools are equally useful in managing around maintenance and upgrade tasks. For example, PowerHA SystemMirror includes a tool on AIX to manage ifixes and Service Packs across the cluster. Over the last few years, PowerHA SystemMirror development has been focused on its ease of use and has successfully countered the old and often inaccurate perception that PowerHA SystemMirror is difficult to manage.

Typically organizations have a range of applications with related (but differing) Service Level Agreements (SLAs). To match this, IBM has a number of solutions, which can either work together or independently, to meet your different SLAs and the different operating systems that may be running in your Power Systems environment.

Addressing the cost of these solutions, which in most cases includes the duplication of some expensive infrastructure, is not easy. However, to be prepared, an organization needs to be

[2] https://www.ibm.com/downloads/cas/DVOXZV6R

able to calculate a realistic cost to their business of some of the more common failure scenarios. Fortunately the other side of the equation, the setup cost, is becoming easier to control by using some of the newer features of the products. For example, licenses now only need to be activated when needed and resources can be freed as required by automating the shutdown of less critical workloads.

This section examines the options available to replicate the application data and to manage the application, either through building availability around the management of the LPAR or the application. Availability can also be provided by scaling-out the application within the data center or across data centers.

The storage and application managed replication solutions include;

- Storage managed replication solutions.
- Application replication solutions.
- Geographic Logical Volume Manager (GLVM).
- IBM i Geographic Mirroring.
- Spectrum Scale stretched cluster.
- Spectrum Scale AFM/DR.

The options to manage the LPAR availability include:

- Live Partition Mobility or LPM (more of a useful tool for Administrators to move workloads for maintenance and some types of failure).
- Simplified Remote Restart (SRR).
- IBM Virtual Machine Recovery Manager HA (VMRM HA), management of SRR.
- IBM Virtual Machine Recovery Manager DR (VMRM DR), evolved from IBM Geographically Dispersed Resiliency for IBM Power Systems.

The clustering options to manage application availability include:

- IBM Tivoli® System Automation for Multi-Platform (AIX and Linux).
- PowerHA SystemMirror Standard Edition for AIX and i.
- PowerHA SystemMirror Enterprise Edition for AIX and i.

Scale out options:

- Red Hat OpenShift.

2.4.1 Introduction to data replication options

In general data replication is a process that provides multiple copies of data, often across sites for HA and DR purposes. There are many ways to replicate data. The most common are:

- Storage.
- Application.
- Server and operating system.

This section cover options in each of these areas and in no way is fully encompassing all options available today. Though we do focus primarily on options available for Power Systems. Many of these options can be used in combination with other high availability management options, like PowerHA SystemMirror and VM Recovery Manager.

Storage options

The following section primarily covers storage based replication options available from IBM storage. There may comparable options from other vendors.

IBM Spectrum Virtualize options

Following are details of each data replication option that is provided by IBM Spectrum Virtualize, formerly known as IBM Storwize® and originally known as code from IBM San Volume Controller (SVC).

The IBM Spectrum Virtualize system combines software and hardware into a comprehensive, modular appliance that provides symmetric virtualization.

Symmetric virtualization is achieved by creating a pool of managed disks (MDisks) from the attached storage systems and optional SAS expansion enclosures. Volumes can be created in a pool for use by attached host systems. System administrators can view and access a common pool of storage on the storage area network (SAN) or local area network (LAN). This functionality helps administrators to use storage resources more efficiently and provides a common base of advanced functions, not only for IBM storage, but for many heterogeneous storage environments.

Spectrum Virtualize offers many functions and features but for purposes of this document we are focusing on the Copy Services functionality specifically. For more details about all features and functions, consult the IBM documentation here.

IBM FlashCopy

FlashCopy makes an instant, point-in-time copy from a source volume to a target volume. While often times this is performed within the same storage unit, because of virtualizing many types of storage it is possible to create the copies across separate storage units.

Some of the reasons for using FlashCopy to make copies of data are:

- Backup processing.
- Data mining.
- Creating an environment for testing.
- Creating an environment for development.
- Creating data for reporting.
- Archiving.

In its basic mode, the FlashCopy function creates copies of content on a source volume to a target volume in a mapping. The function associates a source volume and a target volume in a mapping. If data exists on the target volume, that data is replaced by the copied data. After the copy operation has completed, the target volumes contain the contents of the source volume(s) as they existed at a single point in time, unless target writes have been processed. FlashCopy is sometimes described as an instance of a time-zero copy (T 0) or point-in-time copy technology. Although the copy operation takes some time to complete, the resulting data on the target volume is presented so that the copy appears to have occurred immediately, and all data is available immediately. However, if needed, data that is still in the process of being copied can be accessed from the source.

Figure 2-6 on page 31 shows an overview of the FlashCopy process.

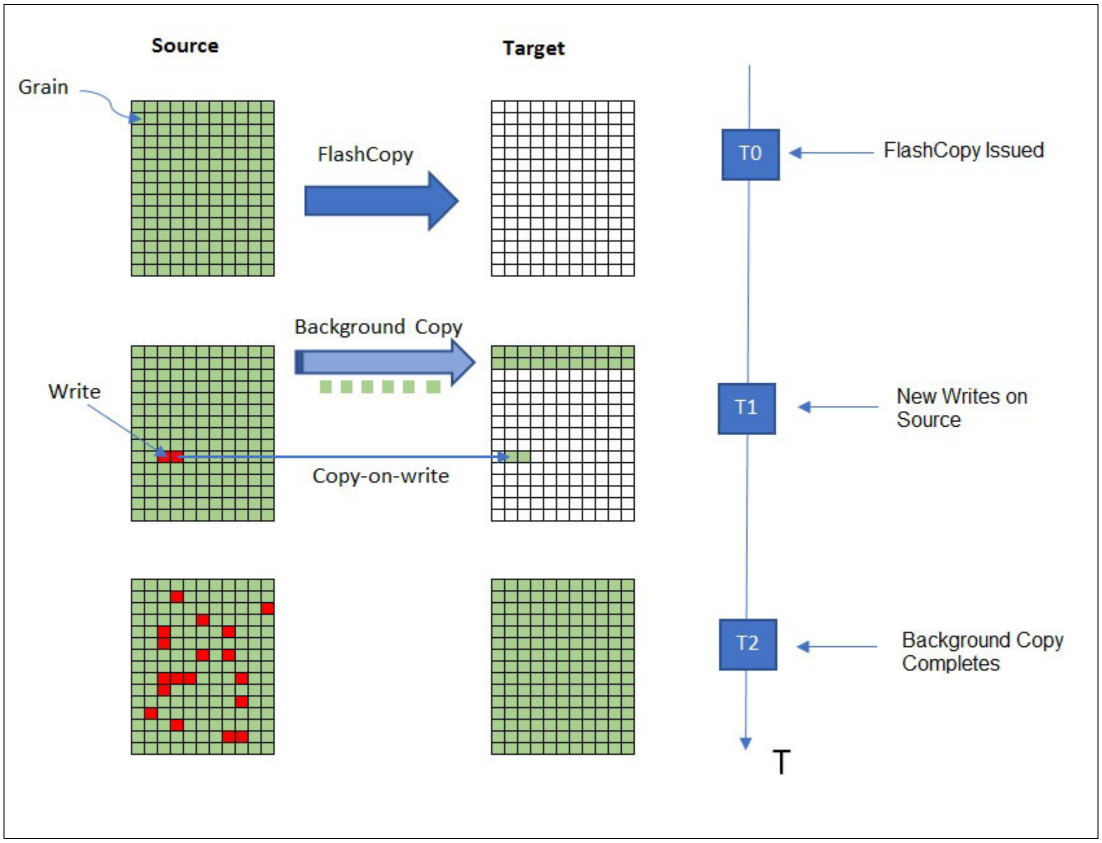


Figure 2-6 FlashCopy example

Although it is difficult to make a consistent copy of a data set that is constantly updated, point-in-time copy techniques help solve this problem. If a copy of a data set is created using a technology that does not provide point-in-time techniques and the data set changes during the copy operation, the resulting copy might contain data that is not consistent. For example, if a reference to an object is copied earlier than the object itself and the object is moved before it is copied, the copy contains the referenced object at its new location, but the copied reference still points to the previous location. You can also assign background copy and cleaning rates to a FlashCopy mapping to control the rate at which updates are propagated to the remote system. FlashCopy mapping copy rate values can be from 128 KBps to 2 GBps and can be changed when the FlashCopy mapping is in any state.

More advanced functions allow operations to occur on multiple source and target volumes. Management operations are coordinated to provide a common, single point-in-time for copying target volumes from their respective source volumes. This creates a consistent copy of data that spans multiple volumes. The function also supports multiple target volumes to be copied from each source volume. This can be used to create images from different points in time for each source volume.

FlashCopy can also utilize *consistency groups*. Consistency groups are a container for FlashCopy mappings to help manage related copies and ensure consistency. You can add many mappings to a consistency group.

The consistency group is specified when the FlashCopy mapping is created. You can also add existing FlashCopy mappings to a new consistency group or change the consistency group later. When you use a consistency group, you prepare and start that group instead of the individual FlashCopy mappings. This process ensures that a consistent copy is made of all the source volumes. FlashCopy mappings to control at an individual level are known as stand-alone mappings. Do not place stand-alone mappings into a consistency group because they become controlled as part of that consistency group.

When you copy data from one volume to another, the data might not include all that you need to use the copy. In many applications, data spans multiple volumes and requires that data integrity is preserved across volumes. For example, the logs for a particular database usually reside on a different volume than the volume that contains the data.

Consistency groups address the problem of applications having related data that spans multiple volumes. In this situation, copy operations must be initiated in a way that preserves data integrity across the multiple volumes. One requirement for preserving the integrity of data that is being written is to ensure that dependent writes are run in the intended sequence of the application.

For more details about FlashCopy see *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize Version 8.4*, SG24-8491.

Volume mirroring (VDisk Mirror)

Volume mirroring provides two physical copies, one on each of two LUNs. Each volume copy can belong to a different pool, and each copy has the same virtual capacity as the volume. In the management GUI, an asterisk (*) indicates the primary copy of the mirrored volume. The primary copy indicates the preferred volume for read requests.

When a server writes to a mirrored volume, the system writes the data to both copies. When a server reads a mirrored volume, the system picks one of the copies to read. If one of the mirrored volume copies is temporarily unavailable; for example, because the storage system that provides the pool is unavailable, the volume remains accessible to servers. The system remembers which areas of the volume are written and resynchronizes these areas when both copies are available.

You can create a volume with one or two copies, and you can convert a non-mirrored volume into a mirrored volume by adding a copy. When a copy is added in this way, the system synchronizes the new copy so that it is the same as the existing volume. Servers can access the volume during this synchronization process.

You can convert a mirrored volume into a non-mirrored volume by deleting one copy or by splitting one copy to create a new non-mirrored volume.

You can use mirrored volumes for the following reasons:

- Improving availability of volumes by protecting them from a single storage system failure.
- Providing concurrent maintenance of a storage system that does not natively support concurrent maintenance.
- Providing an alternative method of data migration with better availability characteristics. While a volume is migrated by using the data migration feature, it is vulnerable to failures on both the source and target pool. Volume mirroring provides an alternative because you can start with a non-mirrored volume in the source pool, and then add a copy to that volume in the destination pool. When the volume is synchronized, you can delete the original copy that is in the source pool. During the synchronization process, the volume remains available even if there is a problem with the destination pool.
- Converting fully allocated volumes to use data reduction technologies, such as thin-provisioning, compression, or deduplication.
- Converting compressed or thin-provisioned volumes in standard pools to data reduction pools to improve capacity savings.

When you use volume mirroring, consider how quorum candidate disks are allocated. Volume mirroring maintains some state data on the quorum disks. If a quorum disk is not accessible and volume mirroring is unable to update the state information, a mirrored volume might need to be taken offline to maintain data integrity. To ensure the high availability of the system, ensure that multiple quorum candidate disks are allocated and configured on different storage systems.

When a volume mirror is synchronized, a mirrored copy can become unsynchronized if it goes offline and write I/O requests need to be processed, or if a mirror fast failover occurs. The fast failover isolates the host systems from temporarily slow-performing mirrored copies, which affect the system with a short interruption to redundancy.

Figure 2-7 shows an example of Vdisk mirroring. For most highly available options the mirrored LUNs are each located in separate, even disparate, storage units. This providing redundancy in the event of storage unit access loss.

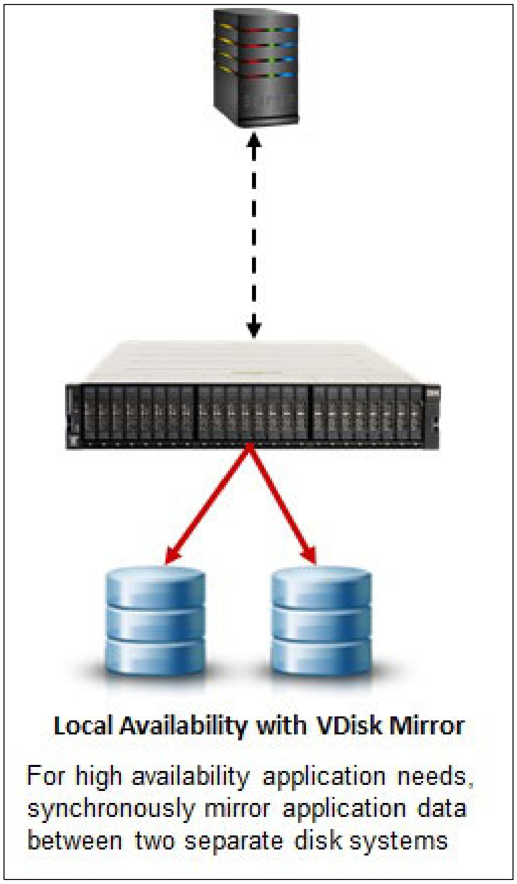

Local Availability with VDisk Mirror
For high availability application needs, synchronously mirror application data between two separate disk systems

Figure 2-7 Vdisk mirroring example

For more details about Vdisk mirroring see *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize Version 8.4*, SG24-8491*.*

Remote copy

Remote copy is a storage-based disaster recovery, business continuance, and workload migration solution that allows you to copy data to a remote location in real time. It is a blanket term that refers to the Advanced Copy Services covered in the remain.

Hyperswap

The IBM HyperSwap® HA feature in the IBM Spectrum Virtualize software enables business continuity during an array of failures. This includes hardware, power, connectivity, or even entire site disasters. It provides data access by utilizing multiple volume copies in separate locations or *sites*.

IBM Spectrum Virtualize 8.4 introduced support for three-site implementations. HyperSwap volumes consists of a copy at each site. Data that is written to the volume is automatically sent to all copies. If any site or storage unit is no longer available, another site can provide access to the volume.

To construct HyperSwap volumes, active-active relationships are made between the copies at each site. These relationships automatically run and switch direction according to which copy or copies are online and up to date. The relationships provide access to whichever copy is up to date through a single volume, which has unique ID. This is seen as a single volume to the operating system but is backed by many physical volumes and copies to provide continuous access.

Relationships can be grouped into consistency groups just like Metro Mirror and Global Mirror relationships. The consistency groups fail over consistently as a group based on the state of all copies in the group. An image that can be used for disaster recovery is maintained at each site.

An active-active relationship is used to manage the synchronous replication of volume data between sites. You must make the master volume accessible through either I/O group. The synchronizing process starts after change volumes are added to the active-active relationship.

Systems that are configured in a three-site topology have high DR capabilities, but a disaster might take the data offline until the system can be failed over to an alternate site. HyperSwap allows active-active configurations to maintain data availability, eliminating the need to failover if communication failures occur. This provides more resilience and provides up to 100% uptime for data.

To better assist with three-site replication solutions, IBM Spectrum Virtualize three-site Orchestrator coordinates replication of data for DR and HA scenarios between systems.

IBM Spectrum Virtualize three-site Orchestrator is a command-line based application that runs on a separate Linux host that configures and manages supported replication configurations on IBM Spectrum Virtualize products.

Figure 2-8 on page 35 shows the three-site replication solution with Hyperswap and Global Mirror.

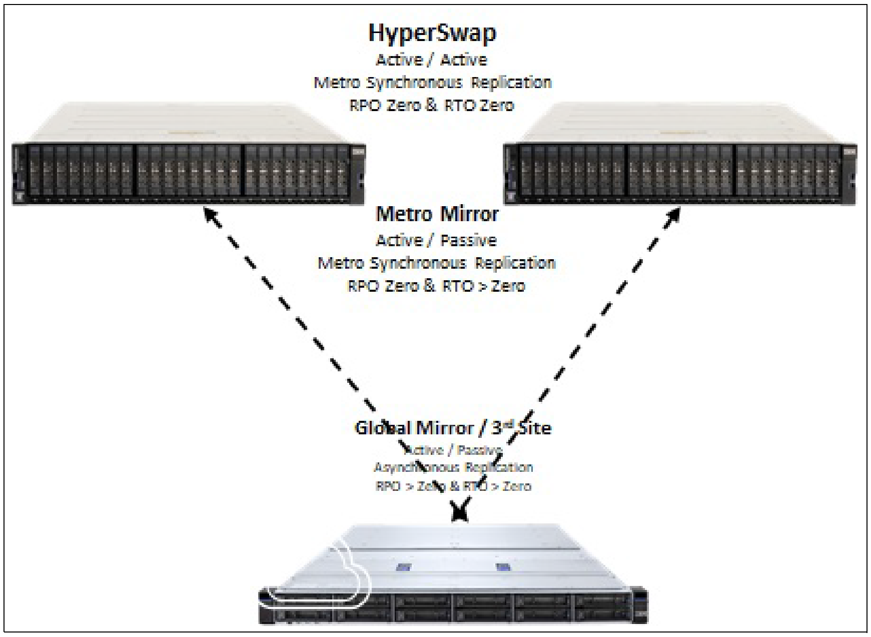


Figure 2-8 Spectrum Virtualize three-site solution

For more information, see the following publications:

- *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize Version 8.4, SG24-8491.*
- *IBM Storwize V7000, Spectrum Virtualize, HyperSwap, and VMware Implementation, SG24-8317.*
- *IBM Spectrum Virtualize HyperSwap SAN Implementation and Design Best Practices, REDP-5597-00.*
- *IIBM Spectrum Virtualize 3-Site Replication, SG24-8504.*

Metro Mirror

Metro Mirror is a type of remote copy that creates a *synchronous* copy of data from a primary volume to a secondary volume. Though a secondary volume can either be on the same system or on another system, it is more common to be on another system at a remote site.

With synchronous copies, host applications write to the primary volume but do not receive confirmation that the write operation has completed until the data is written to the secondary volume. This ensures that both the volumes have identical data when the copy operation completes. After the initial copy operation completes, the Metro Mirror function maintains a fully synchronized copy of the source data at the target site at all times.

The Metro Mirror function supports copy operations between volumes that are separated by distances up to 300 km. For disaster recovery purposes, Metro Mirror provides the simplest way to maintain an identical copy on both the primary and secondary volumes. However, like with all synchronous copies over remote distances, there can be a performance impact to host applications. This performance impact is related to the distance between primary and secondary volumes and depending on application requirements, its use might be limited based on the latency between sites.

Figure 2-9 on page 36 shows an example of a Metro Mirror configuration.

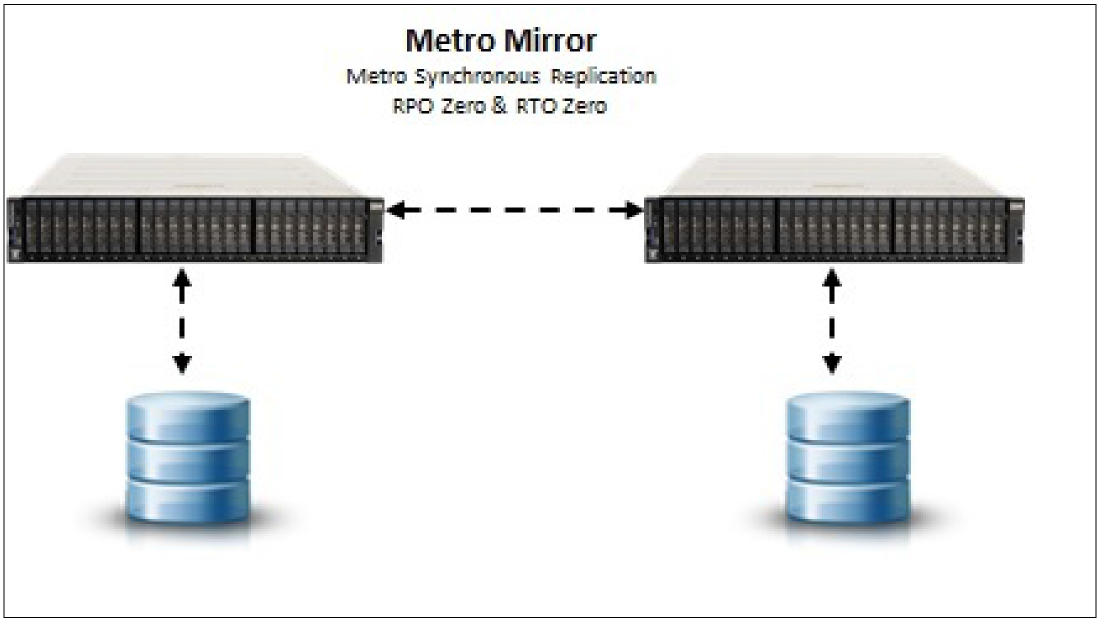


Figure 2-9 Metro Mirroring example

For more information, see the following publications:

- *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize Version 8.4, SG24-8491.*

Global Mirror

The Global Mirror function provides an asynchronous copy process. When a host writes to the primary volume, confirmation of I/O completion is received before the write operation completes for the copy on the secondary volume.

When a failover occurs the application must recover and apply any updates that were not committed to the secondary volume. If I/O operations on the primary volume are paused for a small length of time, the secondary volume can become an exact match of the primary volume. This function is comparable to a continuous backup process in which the last few updates are always missing. When you use Global Mirror for disaster recovery, you must consider how you want to handle these missing updates.

To use the Global Mirror function, all components in the network must be capable of sustaining the workload that is generated by application hosts and the Global Mirror background copy process. If all of the components in the network cannot sustain the workload, the Global Mirror relationships are automatically stopped to protect your application hosts from increased response times.

When Global Mirror operates without cycling, write operations are applied to the secondary volume as soon as possible after they are applied to the primary volume. The secondary volume is generally less than 1 second behind the primary volume, which minimizes the amount of data that must be recovered if a failover occurs. However, a high-bandwidth link must be provisioned between the sites.

For more information, see the following publications:

- *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize Version 8.4, SG24-8491.*

Global Mirror with change volumes

Global Mirror with change volumes (cycling mode set to Multiple) provides the same basic function of asynchronous copy operations between source and target volumes for disaster recovery.

If you are using Global Mirror with cycling mode set to Multiple, the copying process is similar to Metro Mirror and standard Global Mirror. Change volumes must be configured for both the primary and secondary volumes in each relationship. A copy is taken of the primary volume in the relationship using the change volume that is specified when the Global Mirror relationship with change volumes is created. The background copy process reads data from the stable and consistent change volume, copying the data to the secondary volume in the relationship. Copy-on-write technology is used to maintain the consistent image of the primary volume for the background copy process to read. The changes that took place while the background copy process was active are also tracked. The change volume for the secondary volume can also be used to maintain a consistent image of the secondary volume while the background copy process is active.

For more information, see the following publications:

- *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize Version 8.4, SG24-8491.*

Metro/Global Mirror

The Metro/Global Mirror function combines the capabilities of Metro Mirror and Global Mirror functions for greater protection against planned and unplanned outages.

Metro/Global Mirror is a three-site, high availability disaster recovery solution, which uses synchronous replication to mirror data between a local site and an intermediate site, and asynchronous replication to mirror data from an intermediate site to a remote site. The IBM DS8000® series supports the Metro/Global Mirror function on open systems and IBM z Systems® or IBM S/390® hosts. You can set up and manage your Metro/Global Mirror configurations using DS CLI and Time Sharing Option (TSO) commands.

In a Metro/Global Mirror configuration, a Metro Mirror volume pair is established between two nearby sites (local and intermediate) to protect from local site disasters. The Global Mirror volumes can be located thousands of miles away and can be updated if the original local site has suffered a disaster but has performed failover operations to the intermediate site. In the case of a local-site-only disaster, Metro/Global Mirror can provide zero-data-loss recovery at the remote site and at the intermediate site.

In some customer environments, it is necessary to mirror data from a local to a remote site within the distance that is supported for synchronous mirroring. This is especially true when synchronous I/O is required for high or near continuous availability and when a zero-data-loss configuration is required. However, in some cases, it is ideal to have more than a short distance synchronous mirroring solution. Sometimes the following mirroring solutions are required:

- A nearby two-site synchronous copy that can protect from local disasters.
- A longer distance asynchronous copy, at a third site, that can protect from larger scale regional disasters. The third site provides an extra layer of data protection.

The Metro/Global Mirror function provides this combination of synchronous and asynchronous mirroring. Metro/Global Mirror is an extension of Global Mirror, which is based on existing Global Copy (formerly known as PPRC XD) and FlashCopy functions. Global Mirror running at the intermediate site, using a master storage unit and optional subordinate storage units, internally manages data consistency, removing the need for external software to form consistency groups at the remote site.

Figure 2-10 shows the three sites that are used in a Metro/Global Mirror configuration. The configuration uses a minimum of three storage units, one each at the local, intermediate, and remote sites. A minimum of four groups of volumes (group A, group B, group C, and group D) are used in this configuration. An optional group E can be included for extra level of disaster protection.

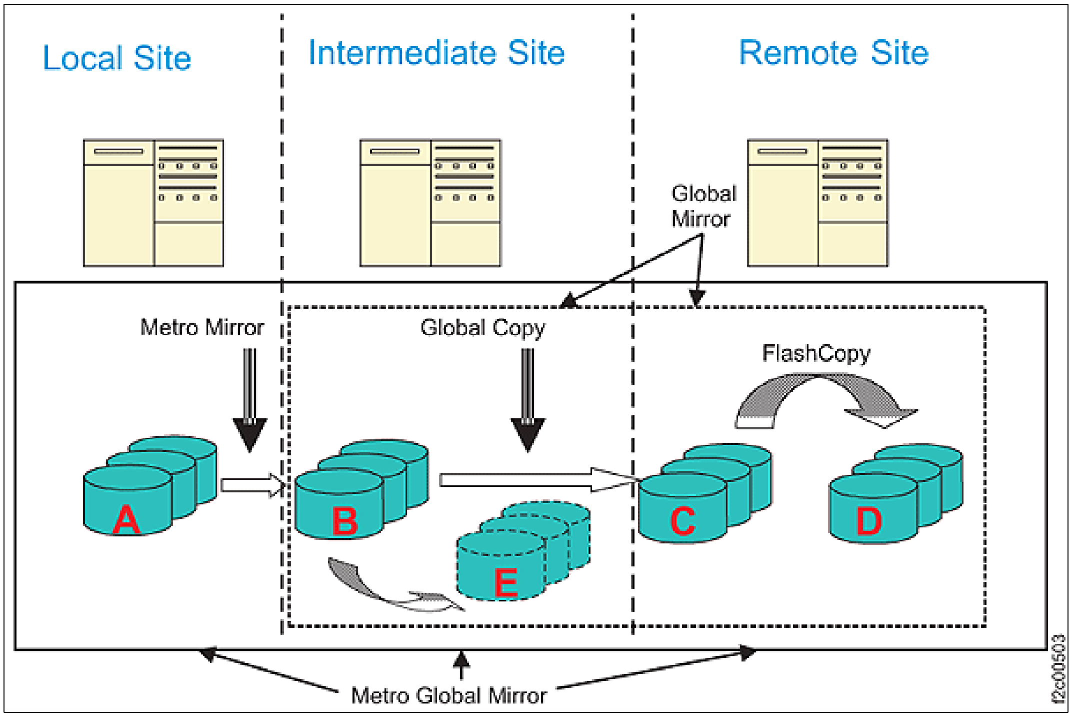


Figure 2-10 Metro/Global Mirror configuration with three sites

Data from the group A volumes at the local site is synchronously replicated to the group B volumes at the intermediate site using Metro Mirror. Data from the group B volumes at the intermediate site is asynchronously replicated to the group C volumes at the remote site using Global Copy. FlashCopy relationships are created with the group C volumes at the remote site as the FlashCopy source volumes and the group D volumes at the remote site as the FlashCopy target volumes, maintaining the consistent disaster recovery volumes using Global Mirror.

As an additional layer of disaster protection if Global Mirror processing were to fail at the remote site, you can use the storage at your intermediate site for a target copy. Setting up Global Mirror between the remote and intermediate sites requires an additional set of FlashCopy volumes at the intermediate site. Then, you can perform failover and restore operations at the remote site using these volumes at the intermediate site (acting as a remote site) to create Global Mirror consistency groups. These volumes, which are referred to as group E volumes, are used as FlashCopy targets for a Global Mirror consistency group.

For Global Mirror processing, one storage unit at the intermediate site is designated as the master storage unit. The master storage unit sends commands over Fibre Channel Protocol (FCP) links and coordinates the consistency group formation process. These links are required for the Global Mirror master storage unit to coordinate the consistency group formation process with the storage units and to communicate the FlashCopy commands to the remote site. All status is relayed back to the master storage unit.

With *Incremental Resync*, it is possible to change the copy target destination of a copy relation without requiring a full copy of the data. This function can be used, for example, when an intermediate site fails because of a disaster. In this case, a Global Mirror is established from the local to the remote site, which bypasses the intermediate site. When the intermediate site becomes available again, the Incremental Resync is used to bring it back into the Metro/Global Mirror setup.

For more information see:

- *IBM DS8000 Copy Services: Updated for IBM DS8000 Release 9.1, SG24-8367.*

For public cloud

Designed for SDS environments, IBM Spectrum Virtualize for Public Cloud represents the solution for public cloud implementations and includes technologies that complement and enhance public cloud offering capabilities.

For example, traditional practices that provide data replication simply by copying storage at one facility to largely identical storage at another facility are not an option regarding public cloud. Also, using conventional software to replicate data imposes unnecessary loads on application servers.

IBM Spectrum Virtualize for Public Cloud delivers a powerful solution for the deployment of IBM Spectrum Virtualize software in public clouds, starting with IBM Cloud. This new capability provides a monthly license to deploy and use IBM Spectrum Virtualize in IBM Cloud to enable hybrid cloud solutions, offering the ability to transfer data between on-premises data centers by using any IBM Spectrum Virtualize-based appliance and IBM Cloud.

With a deployment that is designed for the cloud, IBM Spectrum Virtualize for Public Cloud in any of the over 30 IBM Cloud data centers around the world, where after provisioning the infrastructure an installation script automatically installs the software and creates the cluster.

IBM Spectrum Virtualize for Public Cloud offers a powerful value proposition for enterprise and cloud users who are searching for more flexible and agile ways to deploy block storage on cloud. Using standard Intel servers, IBM Spectrum Virtualize for Public Cloud can be easily added to existing cloud infrastructures to deliver more features and functions, enhancing the storage offering available on public cloud catalog.

The benefits of deploying IBM Spectrum Virtualize on a public cloud platform are two-fold:

- Public cloud storage offering enhancement:

  IBM Spectrum Virtualize for Public Cloud enhances the public cloud catalog by increasing standard storage offering capabilities and features improving specific limitations:

  - Snapshots: A volume’s snapshots occur at high-tier storage with no options for lower-end storage tier. Using IBM Spectrum Virtualize, the administrator has more granular control so that they can provide a snapshot that is stored on lower-end storage for a production volume.
  - Volume size: Most cloud storage providers have a maximum volume size (typically a few TB) that can be provided which can be mounted by a few nodes. At the time of writing, IBM Spectrum Virtualize allows for up to 256 TB and up to 20,000 host connections.
  - Native storage-based replication: Replication features are natively supported but are typically limited to specific data center pairs, to a predefined minimum recovery point objective (RPO). They are accessible only when the primary volume is down. IBM Spectrum Virtualize provides greater flexibility in storage replication, which enables user-defined RPO and replication between any other system running IBM Spectrum Virtualize.
- New features for public cloud storage offering:

IBM Spectrum Virtualize for Public Cloud introduces IBM SAN Volume Controller and IBM Spectrum Virtualize capabilities to the public cloud catalog. These additional features mainly relate to hybrid cloud scenarios and the support to foster these solutions for improved hybrid architectures, which enhance data mobility and management flexibility:

- Replication or migration of data between on-premises storage and public cloud storage.

  In a heterogeneous environment (VMware, bare metal, Hyper-V, and others), replication consistency is achieved through storage-based replica-peering cloud storage with primary storage on-premises. Due to standardization of storage service model and inability to move its own storage to a cloud data center, the storage-based replica is achievable only by involving an SDS solution on-premises.

  In this sense, IBM Spectrum Virtualize for Public Cloud offers data replication among the Storwize family, IBM FlashSystem® 7200, IBM FlashSystem 9200, IBM SAN Volume Controller, and IBM VersaStack and Public Cloud, and it extends replication to all types of supported virtualized storage on-premises. Working together, IBM Spectrum Virtualize and IBM Spectrum Virtualize for Public Cloud support synchronous and asynchronous mirroring between the cloud and on-premises for more than 400 different storage systems from a wide variety of vendors. In addition, they support other services, such as IBM FlashCopy and IBM Easy Tier®.

- DR strategies between on-premises and public cloud data centers as alternative DR solutions. One of the reasons to replicate is to have a copy of the data from which to restart operations in case of an emergency. IBM Spectrum Virtualize for Public Cloud enables replication for virtual and physical environments, which adds new possibilities compared to software replicators in use today that handle virtual infrastructure only.
- Benefit from familiar, sophisticated storage functions in the cloud to implement reverse mirroring.

IBM Spectrum Virtualize enables the possibility to reverse data replication to offload from Cloud Provider back to on-premises or to another Cloud provider. IBM Spectrum Virtualize, both on-premises and in the public cloud, provides a data strategy that is independent of the choice of infrastructure. It delivers tightly integrated functions and consistent management across heterogeneous on-premises storage and cloud storage. The software layer that is provided by IBM Spectrum Virtualize on-premises or in the cloud can provide a significant business advantage by delivering more services faster and more efficiently, enabling real-time business insights and supporting more customer interaction.

Capabilities such as rapid, flexible provisioning, simplified configuration changes, nondisruptive movement of data among tiers of storage, and a single user interface help make the storage infrastructure (and the hybrid cloud) simpler, more cost-effective, and easier to manage can be deployed.

For more information see:

- *Implementing IBM Spectrum Virtualize for Public Cloud Version 8.3.1.*
- *Implementing IBM Spectrum Virtualize for Public Cloud on AWS Version 8.3.1.*
- *Achieving Hybrid Cloud Cyber Resiliency with IBM Spectrum Virtualize for Public Cloud.*
- *Multicloud Solution for Business Continuity using IBM Spectrum Virtualize for Public Cloud on AWS.*

IBM Copy Services Manager

IBM Copy Services Manager (formerly IBM Tivoli Storage Productivity Center for Replication, a component of IBM Tivoli Storage Productivity Center and IBM SmartCloud®. Virtual Storage Center) manages copy services in IBM storage environments. Copy services are features that are used by storage systems to configure, manage, and monitor data replication functions. These copy services include IBM FlashCopy, Metro Mirror, Global Mirror, and Metro Global Mirror data replication.

IBM Copy Services Manager automates key replication management tasks to help you improve the efficiency of your storage replication. You can use a simple GUI or command line to configure, automate, manage, and monitor all important data replication tasks in your environment, including the following tasks:

- Manage and monitor multisite environments to meet disaster recovery (DR) requirements.
- Automate the administration and configuration of data replication features.
- Keep data on multiple related volumes consistent across storage systems in a planned or unplanned outage.
- Recover to a remote site to reduce downtime of critical applications.
- Provide high availability (HA) for applications by using IBM HyperSwap technology.
- Practice recovery processes while disaster recovery capabilities are maintained.
- Plan for replication when you are provisioning storage.
- Monitor and track replication operations.
- Automate the mapping of source volumes to target volumes.

Copy Services Manager runs on Windows, AIX, Linux, Linux on z Systems, and IBM z/OS® operating systems. When it is running on z/OS, Copy Services Manager uses the Fibre Channel connection (IBM FICON®) to connect to and manage count-key data (CKD) volumes.

For more information see the IBM Copy Services base publications at the following website:

https://www.ibm.com/docs/en/csm

There is also additional information and demos available on the IBM Copy Services YouTube channel.

Geographic Logical Volume Manager (GLVM)

GLVM is an IP-based replication facility that is both native and exclusive to the AIX operating system. The main function of GLVM is mirroring local/production site data across an IP-based network to a system at a remote/backup site. A total failure of the node at the local site does not cause the loss of data on the node at the remote site. GLVM itself does not provide any automated fallover to the remote site, however, it is a supported component that PowerHA SystemMirror for AIX can utilize to provide automated recovery. For more information about this combination see the GLVM configuration assistant topic in *IBM PowerHA SystemMirror V7.2.3 for IBM AIX and V7.22 for Linux*, SG24-8434*.*

GLVM is based upon AIX's native LVM facility, and as such, supports up to three total copies. However, typically only one copy resides at a remote backup site. To utilize GLVM an AIX instance with ample amount of storage for the replication must be online and available at each location and connected by at least one IP network. GLVM is intended to keep data volume groups synchronized and *not* the base operating system volume group, *rootvg*. GLVM also supports both raw logical volumes and file systems.

GLVM is storage type independent. As long as the storage is supported on AIX with native LVM then it also can be utilized with GLVM. This also means there is no requirement for similar storage types at each location. Internal disks locally can be mirrored across IP to either internal or external disks at the remote location. The same is true for mirroring external disk locally to either internal or external disks remotely. Though you can mix and match there can be performance implications with doing so. The key performance factor is the I/O rates will never be faster than the slowest common denominator. This often is the IP-network but also can be because of old internal disks at either site.

GLVM supports both synchronous and asynchronous forms of replication. Typically when recovery point objective (RPO) is zero, then synchronous is required. When RPO is greater than zero, then asynchronous mode is an option. Though technically there is no distance limitation in using either one the key factors for determining which method to use are bandwidth, latency, and cache logical volume size. While it may be desirable to have a RPO of zero with 5000 Km between sites, the latency will most likely result in unacceptable performance.

For more details about configuring GLVM see the following:

https://tinyurl.com/asyncglvm
https://tinyurl.com/redsglvm

IBM i Geographic Mirroring

Geographic mirroring is a function of the IBM i operating system. All the data that is placed in the production copy of the IASP is mirrored to a second IASP on a second, perhaps remote system. The replication is done within the operating system, so this solution can be used with any type of storage supported by IBM i. There is both a synchronous and an asynchronous version of geographic mirroring.

The benefits of this solution are essentially the same as the switched LUN solution with the added advantage of providing disaster recovery to a second copy at increased distance. The biggest benefit continues to be operational simplicity. The switching operations are essentially the same as that of the switched LUN solution, except that you switch to the mirror copy of the IASP, making this a straightforward HA solution to deploy and operate. As in the switched LUN solution, objects not in the IASP must be handled by some other mechanism such as administrative domain, and the IASP cannot be brought online to an earlier system. Geographic mirroring also provides real-time replication support for hosted integrated environments such as Microsoft Windows and Linux. This is not generally possible through journal-based logical replication.

Since geographic mirroring replication is within the IBM i operating system, a potential limitation of a geographic mirroring solution is performance impacts in certain workload environments. For synchronous geographic mirroring, when running I/O intensive batch jobs, some performance degradation on the primary system is possible. Also, be aware of the increased CPU overhead that is required to support geographic mirroring.

The backup copy of the independent disk pool cannot be accessed while the data synchronization is in process. For example, if you want to back up to tape from the geographically mirrored copy, you must quiesce operations on the source system and detach the mirrored copy. Then you must vary on the detached copy of the independent disk pool on the backup system, perform the backup procedure, and then reattach the independent disk pool to the original production host. Synchronization of the data that was changed while the independent disk pool was detached will then be performed. Your HA solution is running exposed, meaning there is no up-to-date second data set, while doing the backups and when synchronization is occurring. Geographic mirroring utilizes source and target side tracking to minimize this exposure.

The following list the characteristics of geographic mirroring:

- All data maintained in the independent disk pool will be replicated to a second copy of the data on a second system.
- Replication is a function of the IBM i OS so any type of storage can be used.

- The application can be switched to the backup system and operate on the independent disk pool copy.
- Two copies of the data eliminating single point of failure.
- When using synchronous geographic mirroring, both copies of the IASP are guaranteed to be identical. Synchronous geographic mirroring over a distance may impact application performance due to communication latency.
- Second copy of data can be geographically dispersed if using asynchronous geographic mirroring. In the case of an unplanned outage on the source system, a few seconds of data loss is possible.
- Data transmission over upto four TCP/IP communication lines for throughput and redundancy.
- It is also recommended that a separate line be used for the clustering heartbeat since sharing the heartbeat with data port can cause contention and possible time outs.
- Offline saves and queries to backup copy of the data while backup dataset is detached.
- Data resiliency not maintained while backup dataset is detached. Data resiliency is resumed after partial or full resynchronization has completed.
- Can be used in conjunction with the IBM i switch LUN technology.
- System performance overhead is associated with running geographic mirroring.
- It is strongly recommended that you configure separate main storage pools or user jobs that access independent disk pools to prevent those jobs from contending with other jobs on the system and using more main storage than desired. More specifically, independent disk pool jobs should not use the machine pool or base pool. If independent disk pool jobs use the same memory as jobs that are not accessing the independent disk pools, independent disk pool jobs can monopolize the memory pool, lock out other jobs, and in extreme situations deadlock the system. Exposure for this situation is greater when using geographic mirroring.
- Journaled objects in the independent disk pool will guarantee data update to target system.
- Simple monitoring of mirror process.
- Cost associated with a second set of disk.
- Replication is at a memory page level managed by IBM i.

Spectrum Scale

Spectrum Scale (previously called General Parallel File System - GPFS) is a cluster file system that provides concurrent access to a file system or file systems from multiple nodes. These nodes can all be SAN attached or a mix of SAN and network attached. This enables high performance access to this common set of data to support a scale-out solution or provide a high availability platform.

Spectrum Scale has many features beyond common data access including data replication, policy based storage management, and multi-site operations. You can create a GPFS cluster of AIX nodes, Linux nodes, Windows server nodes, or a mix of all three. GPFS can run on virtualized instances providing common data access in environments, taking advantages of e logical partitioning, or other hypervisors. Multiple GPFS clusters can share data within a location or across wide area network (WAN) connections.

Spectrum Scale is highly flexible, but we will look at only two configurations:

- Spectrum Scale stretched cluster - synchronous two data center configuration.

- Spectrum Scale Active File Management (AFM) DR.

Spectrum Scale stretched cluster

This configurations provides concurrent access to synchronously replicated data across two data centers with only IP connectivity. The cluster is configured using nodes from the two data centers, however to allow either site to keep operating should one site fail, a third site or “laptop solution” is required. Usually there will be the same number of quorum nodes in each data center and one quorum node at the third site to act as the tie breaker.

The GPFS storage, or the Network Shared Disks (NSDs), are configured at each of the main data centers and in the simplest case, assigned to a failure group, one for each data center. These NSDs can be metadataOnly, dataAndMetadata, dataOnly or a mixture of all three types. The NSD at the third site is configured as descOnly - containing no data, just a file system descriptor in a third failure group. Some or all of the nodes can be configured as NSD Servers, providing NSD access to the clients in the other data center. If the file system or file systems are configured with default data and metdata replicas of two, there will be a complete copy of all data/metadata in each data center. The file system(s) will remain available as long as a quorum of both nodes and file system descriptors are available. This access to the file system(s) will remain available through a single failure - either one of the data centers or the third site, but not both as shown in Figure 2-11.

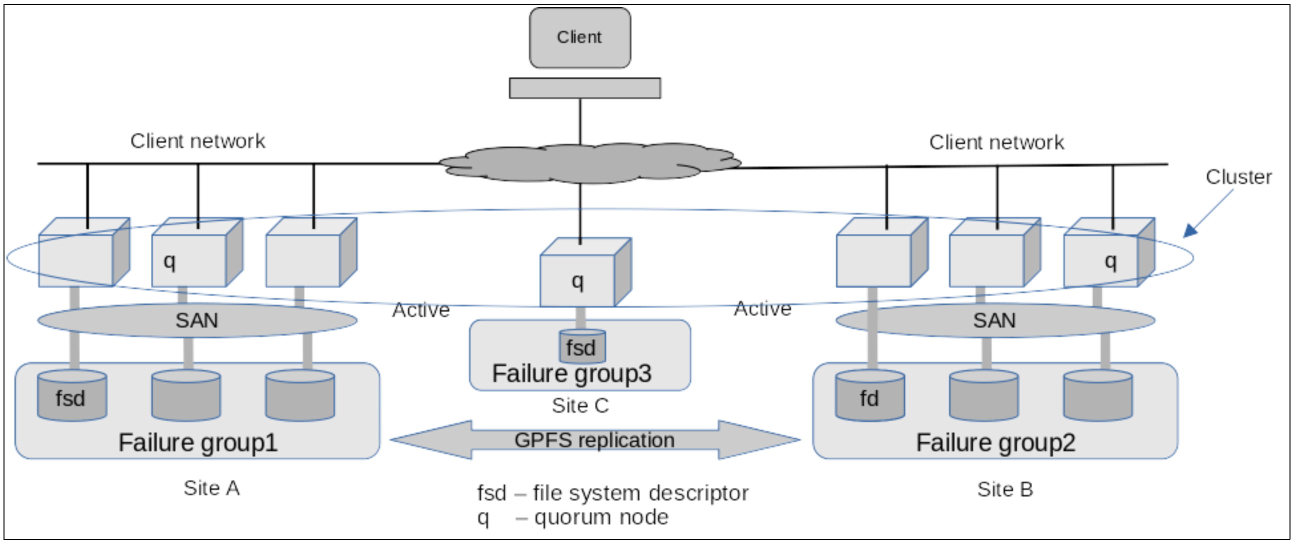


Figure 2-11 Spectrum Scale using one failure group per site

IBM Spectrum Scale Active File Management (AFM) DR

Active file management (AFM) is a feature available in IBM Spectrum Scale Standard Edition (or higher). It provides a scalable, high-performance, file system caching layer integrated with the GPFS cluster file system. AFM allows you to create associations from a local GPFS cluster to a remote cluster or storage, and to define the location and flow of file data to automate the management of the data. This allows you to implement a single namespace view across sites around the world.

Active File Management-based asynchronous disaster recovery (AFM DR) is a fileset-level replication disaster-recovery capability. This capability is a one-to-one active-passive model and is represented by two sites: primary and secondary.

The primary site is a read/write fileset where the applications are currently running and has read/write access to the data. The secondary site is read-only. All the data from the primary site is asynchronously synchronized with the secondary site. The primary and secondary sites can be independently created in storage and network configuration. After the sites are created, you can establish a relationship between the two filesets. The primary site is available for the applications even when communication or secondary fails. When the connection with the secondary site is restored, the primary site detects the restored connection and asynchronously updates the secondary site.

The following data is replicated from the primary site to the secondary site:

- File-user data.
- Metadata including the user-extended attributes except the inode number and a time.
- Hard links.
- Renames.

The following file system and fileset-related attributes from the primary site are not replicated to the secondary:

- User, group, and fileset quotas.
- Replication factors.
- Dependent filesets.

AFM DR can be enabled only on GPFS-independent filesets. An independent fileset that has dependent filesets *cannot* be converted into an AFM DR fileset.

A consistent view of the data in the primary fileset can be propagated to the secondary fileset by using fileset-based snapshots (psnaps). Recovery Point Objective (RPO) defines the frequency of these snapshots and can send alerts through events when it is unable to achieve the set RPO. RPO is disabled by default. The minimum time that you can set as RPO is 720 minutes. AFM-based asynchronous DR can reconfigure the old primary site or establish a new primary site and synchronize it with the current primary site.

Individual files in the AFM DR filesets can be compressed. Compressing files saves disk space. Snapshot data migration is also supported. For more information, see ILM for snapshots in the *IBM Spectrum Scale Administration Guide.*

When a disaster occurs on the primary site, the secondary site can be failed over to become the primary site. When required, the filesets of the secondary site can be restored to the state of the last consistent RPO snapshot. Applications can be moved or failed over to the acting primary site. This application movement helps to ensure stability with minimal downtime and minimal data loss. This makes it possible for applications to eventually be failed back to the primary site as soon as the (new) primary is on the same level as the acting primary.

Concepts

AFM DR does *not* offer any feature to check consistency of files across primary and secondary sites. However, you can use any third-party utility to check that consistency after files are replicated.

You can simultaneously configure a site for continuous replication of IBM Spectrum Scale data along with AFM DR site. With IBM Spectrum Scale continuous replication, you can achieve a near disaster recovery and a far disaster recovery with AFM DR site.

AFM DR uses the same underlying infrastructure as AFM. AFM DR is characterized by two modes: the fileset in the primary cluster uses the primary mode and the fileset in the secondary cluster uses the secondary mode.

AFM DR is supported over both NFS v3 and GPFS protocol. The primary fileset is owned by the primary gateway, which communicates with the NFS server on the secondary side. The primary-secondary relationship is strictly one-to-one.

AFM revalidation does not apply to primary filesets. All files are always cached because primary is the only writer and secondary is in the read-only mode.

You can convert the SW/IW relationship to a DR relationship. However, you cannot convert a DR relationship to an SW/IW relationship.

Features

The following AFM features are offered on AFM DR filesets:

- Force flushing contents before async delay.
- Parallel data transfers.
- Peer snapshot - psnap.
- Gateway node failure and recovery.
- Operation with disconnected secondary.
- Using IBM Spectrum Protect for Space Management (HSM).
- Disabling AFM DR.
- Using AFM DR with encryption.
- Stop and start replication on a fileset.

You can use **`mmbackup`** command to back up all files from primary, as all files are in a cached state on the primary fileset. Similar to AFM filesets, IBM Spectrum Protect (HSM) can be connected to primary or secondary, or both sides. When HSM is connected to the primary side, set AFMSKIPUNCACHEDFILES yes in dsm.sys file. AFM features such as revalidation, eviction, prefetch, partial file caching, expiration, resynchronization, failover, and showing home snapshots are not offered on AFM DR filesets.

AFM to cloud object storage

The AFM to cloud object storage is an IBM Spectrum Scale feature that enables placement of files or objects in an IBM Spectrum Scale cluster to a cloud object storage.

Cloud object services such as Amazon S3 and IBM Cloud Object Storage offer industry-leading scalability, data availability, security, and performance. The AFM to cloud object storage allows associating an IBM Spectrum Scale fileset with a cloud object storage. Customers use a cloud object storage to run workloads such as mobile applications, backup and restore, enterprise applications, and big data analytics, file server. These workloads can be cached on AFM to cloud object storage filesets for faster computation and synchronize back to the cloud object storage server.

The front-end for object applications is an AFM to cloud object storage fileset with the data exchange between the fileset and cloud object storage buckets through the AFM to cloud object storage in the background by providing high performance for the object applications. Object applications can also span across AFM to cloud object storage filesets and on a cloud object storage. Both the fileset and the cloud object storage can be used as a backup of important data.

The AFM to cloud object storage on an IBM Spectrum Scale fileset becomes an extension of cloud object storage buckets for high-performance or used objects. Depending upon the modes of AFM to cloud object storage fileset configurations, objects required for applications such as AI and big data analytics can be downloaded, worked upon, and can be uploaded to a cloud object storage. The objects that are created by applications can be synchronized to the objects on a cloud object storage asynchronously. An AFM to cloud object storage fileset can cache only metadata or both metadata and data.

The AFM to cloud object storage also allows data center administrators to release the IBM Spectrum Scale storage capacity by moving less useful data to the cloud storage. This feature reduces capital and operational expenditures. The AFM-based cache eviction feature can be used to improve the storage capacity manually and by using policies.

The AFM to cloud object storage uses the same underlying infrastructure as AFM. The AFM to cloud object storage is available on all IBM Spectrum Scale editions.

Spectrum Scale also supports using cloud object storage as a target for ILM and the feature is called transparent cloud tiering (TCT). This features allows for the creation of rules to move particular files (for example, those less frequently used), to cloud storage, leaving a small stub in the file system.

For more details about implementing AFM see *Spectrum Scale Concepts, Planning, and Installation Guide IBM SC28-3161.*

2.4.2 Comparison of the storage replication options

Table 2-1 compares Hyperswap, Metro Mirror, Global Mirror, GLVM synchronous and asynchronous, IBM i Geographic Mirror synchronous and asynchronous, Spectrum Scale concurrent, and Spectrum Scale AFM and DR.

Table 2-1 Storage replication comparison

Storage options	**Tier**	**Storage unit fail**		**Site failure**	
		RTO[a]	**RPO**[a]	**RTO**[a]	**RPO**[a]
Hyperswap	7	0	0	0	0
Metro Mirror	7	~0	0	~0	0
Global Mirror	6	>0	≤cache	>0l	≤cache
GLVM synchronous	7	~0	0	~0	0
GLVM asynchronous	6	>0	≤cache	>0	≤cache
IBM i Geographic Mirror (sync)	7	~0	0	~0	0
IBM i Geographic Mirror (async)	6	>0	≤cache	>0	>0
SpectrumScale stretched cluster	7	0	0	0	0
SpectrumScale AFM / DR[a]	6	>0	≤cache	>0	≤cache

[a] ~0 = almost 0
>0 = greater than 0, but still small
≤cache = up to amount of data in the cache
where range is from 0 less than ~0 less than >0 less than ≤cache

2.4.3 Concurrent databases

Concurrent access to a database, both within the data center and across data centers increases availability with zero downtime and data loss. Two popular example are IBM Db2® Mirror and Oracle RAC.

Db2 Mirror

IBM Db2 Mirror for i enables continuous availability for your mission-critical applications. It provides a Recovery Time Objective (RTO) of zero. Db2 Mirror for i synchronously mirrors database updates between two separate nodes by way of remote direct memory access (RDMA) over Converged Ethernet (RoCE) network. Applications can be deployed in an active-active or active-passive (with read access on the secondary) mode.

The Db2 Mirror architecture consists of two nodes that are paired together to create a synchronous environment. The nodes are independent, and both can access and update the data that is synchronously replicated in both directions. Db2 Mirror supports replication of data in SYSBAS and in independent auxiliary storage pools (IASPs). Applications can use either SQL or traditional record level access (RLA) to work with replicated data.

For example, Figure 2-12 shows separate instances of the same application running on each node using a synchronously replicated database file. The database file can exist either in SYSBAS or within an IASP. When Row A is changed on Node 1, it is synchronously written to the file on both Node 1 and Node 2. When Row B is changed on Node 2, it is synchronously written to the file on both Node 2 and Node 1.

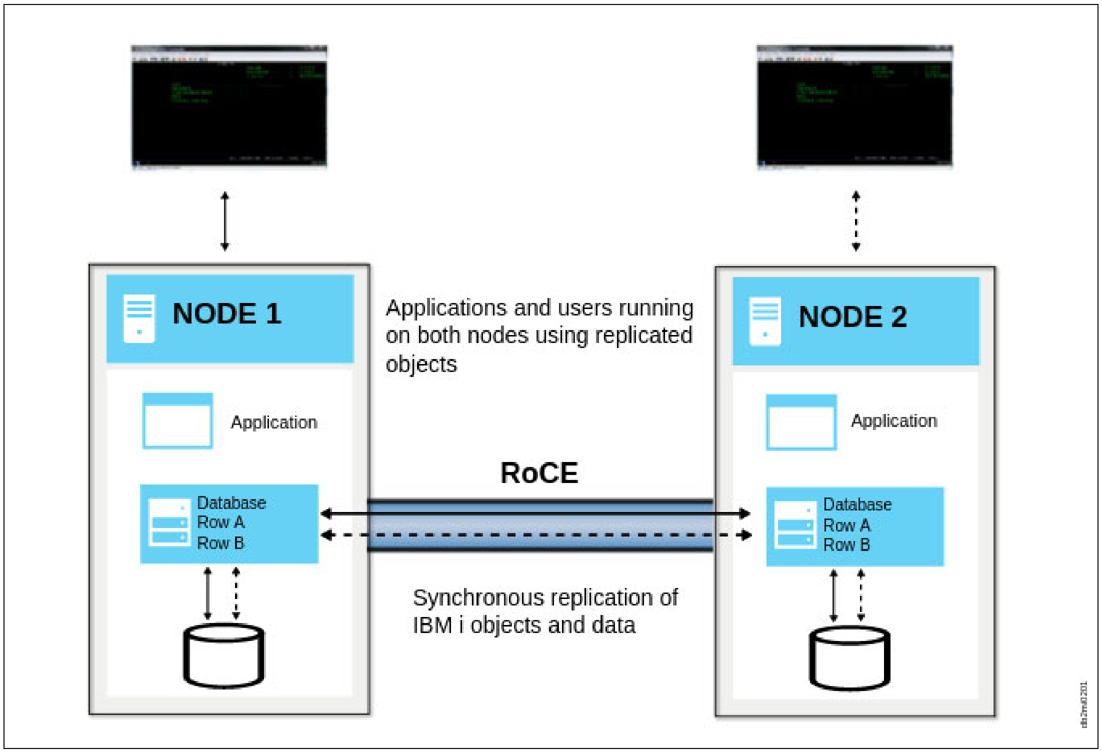

NODE 1
Application
Database
Row A
Row B
Applications and users running on both nodes using replicated objects
RoCE
Synchronous replication of IBM i objects and data
NODE 2
Application
Database
Row A
Row B
db2mi0201

Figure 2-12 Db2 Mirror

Oracle RAC

Oracle Real Application Clusters (RAC) allow customers to run a single Oracle Database across multiple servers to maximize availability and enable horizontal scalability, while accessing shared storage. User sessions connecting to Oracle RAC instances can failover and safely replay changes during outages, without any changes to user applications, hiding the impact of the outages from end users.

Oracle RAC enables customer databases to continue to run across component failures, reducing potential data loss and minimizing unplanned downtime created by single-point-of-failure designs.

Customers can eliminate planned, maintenance related downtime by using Oracle RAC to implement rolling upgrades and patching on a server-by-server basis.

Multiple interconnected computers or servers that provide a service but appear as only one server to end users and applications is commonly referred to as a *cluster*. Oracle RAC clusterizes an Oracle database often by providing an active-active configuration. Oracle RAC uses its own clustering, Oracle Clusterware, to simultaneously utilize multiple servers to provide database access.

Oracle Clusterware is a portable cluster management solution that is integrated with Oracle Database. Oracle Clusterware is also a required component for using Oracle RAC. In addition, Oracle Clusterware enables both noncluster Oracle databases and Oracle RAC databases to use the Oracle high-availability infrastructure. Oracle Clusterware enables you to create a clustered pool of storage to be used by any combination of noncluster and Oracle RAC databases.

Oracle Clusterware is the only clusterware that you need for most platforms on which Oracle RAC operates. You can also use clusterware from other vendors if the clusterware is certified for Oracle RAC.

IBM and Oracle have had a longtime agreement and team up together by way of the IBM Oracle International Competency Center (ICC). This collaboration has resulted in providing premier solutions and documentation of these solutions. In conjunction with the document listed as follows, there is also an email address of ibmoracle@us.ibm.com that be utilized for information.

For more details about implementing Oracle RAC on Power Systems see the following:

- *Installation of Oracle 12c on AIX and Spectrum Scale.*
- *Oracle Database 19c & Oracle Database 19c RAC on IBM AIX Tips and Considerations.*
- Oracle RAC on IBM AIX Best practices.
- Oracle 19c to 12c and 11.2.0.4 Database Performance Considerations with AIX on Power Systems including IBM POWER9™.

2.4.4 Application based / Log shipping

Many enterprise level applications today provide their own inherent high availability and disaster recovery capabilities. The following list is only a partial list of some of the current offerings.

Db2

IBM Db2 server contains functionality that supports many high availability strategies.

- Automatic client reroute roadmap.

  Automatic client reroute is an IBM Db2 server feature that redirects client applications from a failed server to an alternate server so the applications can continue their work with minimal interruption. Automatic client reroute can be accomplished only if an alternate server has been specified prior to the failure.

- Server lists.

  The server list is used by IBM Data Server drivers and clients for workload balancing (WLB) and automatic client reroute (ACR) operation. The server list contains a list of addresses and the relative priority of those addresses. When a client connects to a Db2 server over TCP/IP, the server list is returned to and cached by the client. The server periodically provides a refreshed server list to the client.

- Db2 fault monitor facilities for Linux and UNIX.

  Available on UNIX based systems only, Db2 fault monitor facilities keep IBM Db2 server databases up and running by monitoring Db2 database manager instances, and restarting any instance that exits prematurely.

- High availability disaster recovery (HADR).

  High availability disaster recovery (HADR) provides a high availability solution for both partial and complete site failures. HADR protects against data loss by replicating data changes from a source database, called the primary database, to the target databases, called the standby databases. HADR supports up to three remote standby servers.

- Db2 High Availability Feature.

  The Db2 High Availability Feature enables integration between IBM Db2 server and cluster managing software.

- High availability through log shipping.

  Log shipping is the process of copying whole log files to a standby machine either from an archive device, or through a user exit program that is running against the primary database. A scheduled job on the standby issues the ROLLFORWARD DATABASE command at a specified interval to keep the standby current in terms of log replay.

- Log mirroring.

  IBM Db2 server supports log mirroring at the database level. Mirroring log files helps protect a database from accidental deletion of an active log and data corruption caused by hardware failure.

- High availability through suspended I/O and online split mirror support.

  IBM Db2 server suspended I/O support enables you to split mirrored copies of your primary database without taking the database offline. You can use this to quickly create a standby database to take over if the primary database fails.

HADR Data Flow

Each rectangle in the Figure 2-13 on page 51 represents a thread (also known as EDU (Engine Dispatchable Unit)) in IBM DB2® engine. Threads relevant to HADR are:

- db2agent:

  Thread that serves SQL client connection. Multiple threads per database.

- db2loggw:

  Thread that writes log records to log files. One per database.

- db2hadrp:

  HADR primary side edu. One per database.

- db2hadrs:

  HADR standby side edu. One per database.

- db2lfr:

  LFR (Log File Reader) thread. One per database.

- db2shred:

  Shredder edu. Shreds log pages into log records. One per database.

- db2redom:

  Redo (replay) master thread. One per database.

- db2redow:

Redo (replay) worker thread. Multiple threads per database.

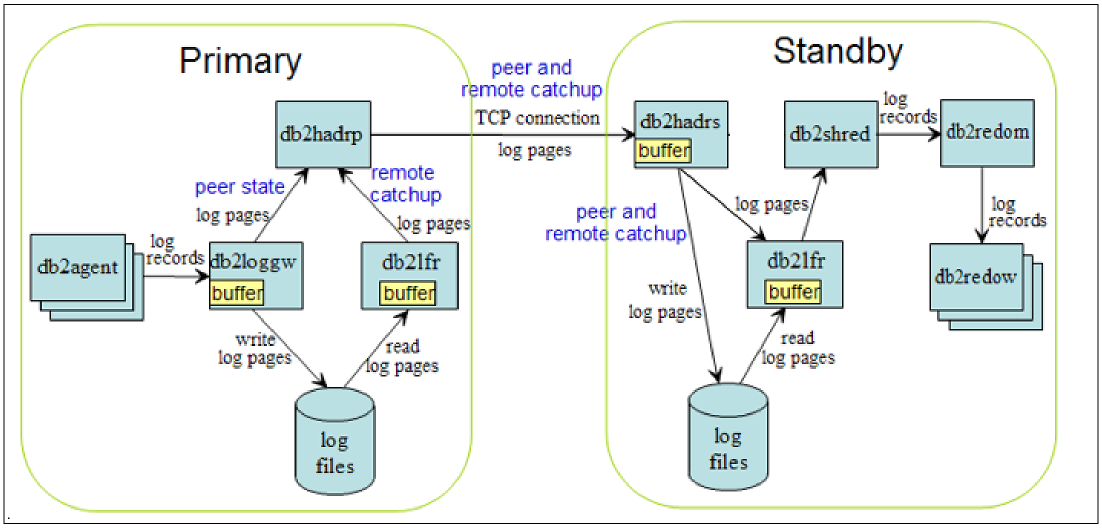


Figure 2-13 Db2 HA DR data flow

For more details see the Db2 wiki which can be found here.

WebSphere Application Server

IBM WebSphere® Application Server (WAS) is a flexible, secure Java server runtime environment for enterprise applications. Deploy and manage applications and services regardless of time, location or device type. Integrated management and administrative tools provide enhanced security and control, and support for multicloud environments lets you choose your deployment method. Continuous delivery capabilities and services help you to respond at the speed of your business needs.

The WebSphere Application Server high availability framework eliminates single points of failure and provides peer to peer failover for applications and processes running within WebSphere Application Server. The WebSphere Application Server high availability framework also allows integration of WebSphere Application Server into an environment that might be using other high availability frameworks, such as PowerHA SystemMirror to manage non-WebSphere Application Server resources.

A WebSphere Application Server cell (the main administrative domain) consists of one or more server processes, which host resources such as applications or messaging engines. The cell is partitioned into groups of servers known as core groups, which are defined by the user. Each core group has its own high availability manager and operates independently of other core groups. Core group boundaries do not overlap. Within each core group are dynamic logical groupings of servers known as high availability groups. The HAManager determines the membership of the HAGroups at runtime. Each core group can have a number of policies, which apply to particular HAGroups and determine the high availability behavior of resources running within the HAGroup.

Additional information about implementing high availability with WAS can be found here.

IBM MQ

IBM MQ, formerly known as MQSeries® and IBM WebSphere MQ, enables applications to communicate at different times and in many diverse computing environments. IBM MQ supports the exchange of information between applications, systems, services and files by sending and receiving message data by way of messaging queues. This simplifies the creation and maintenance of business applications. IBM MQ works with a broad range of computing platforms, and can be deployed across a range of different environments including on-premises, in cloud, and hybrid cloud deployments. IBM MQ supports a number of different APIs including Message Queue Interface (MQI), Java Message Service (JMS), REST, .NET, IBM MQ Light and MQTT.

IBM MQ provides:

- Versatile messaging integration from mainframe to mobile that provides a single, robust messaging backbone for dynamic heterogeneous environments.
- Message delivery with security-rich features that produce auditable results.
- Qualities of service that provide once and once only delivery of messages to ensure messages will withstand application and system outages.
- High-performance message transport to deliver data with improved speed and reliability.
- Highly available and scalable architectures to support an application’s needs.
- Administrative features that simplify messaging management and reduce time spent using complex tools.
- Open standards development tools that support extensibility and business growth.

IBM MQ provides a universal messaging backbone with robust connectivity for flexible and reliable messaging for applications and the integration of existing IT assets using a service-oriented architecture (SOA).

IBM MQ sends and receives data between your applications, and over networks.

Message delivery is assured and decoupled from the application. Assured, because IBM MQ exchanges messages transactionally, and decoupled, because applications do not have to check that messages they sent are delivered safely.

  – You can secure message delivery between queue managers with TLS.
  – With Advanced Message Security (AMS), you can encrypt and sign messages between being put by one application and retrieved by another.
  – Application programmers do not need to have communications programming knowledge.

An IBM MQ messaging system is made up of one or more queue managers. Queue managers are where messaging resources, such as queues, are configured and what applications connect to, either running on the same system as the queue manager or over the network.

A network of connected queue managers supports asynchronous routing of messages between systems, where producing and consuming applications are connected to different queue managers.

IBM MQ can be managed using a variety of tools, from the IBM MQ Explorer GUI, through scripted or interactive command line tools or programmatically.

The applications connecting to IBM MQ can be written in any one of many different programming languages and to many different APIs. From C and Cobol, to Java and .Net to NodeJS and Ruby.

If you want to operate your IBM MQ queue managers in a high availability (HA) configuration, you can set up your queue managers to work either with a high availability manager, such as PowerHA SystemMirror for AIX (formerly HACMP), or with IBM MQ multi-instance queue managers. On Linux systems, you can also deploy replicated data queue managers (RDQMs), which use a quorum-based group to provide high availability.

More information is available for implementing IBM MQ for high availability and disaster recovery.

Oracle Data Guard

Oracle Data Guard provides a solution for high availability, enhanced performance, and automated failover. You can use Oracle Data Guard to create and maintain multiple standby databases for a primary database. The standby databases can be started in the read-only mode to support reporting users and then returned to the standby mode. Changes to the primary database can be relayed automatically from the primary database to the standby databases with a guarantee of no data lost in the process. The standby database servers can be physically separate from the primary server.

In a Data Guard implementation, a database running in archivelog mode is designated as the primary database for an application. One or more standby databases, accessible through Oracle Net Services, provide for failover capabilities. Data Guard automatically transmits redo information to the standby databases over an IP network, where it is applied. As a result, the standby database is transactionally consistent.

Depending on how you configure the redo application process, the standby databases might be in sync with the primary database or might lag behind it. The redo log data is transferred to the standby databases through log transport services, as defined through your initialization parameter settings. Log Apply Services apply the redo information to the standby databases.

In case of a network outage, Data Guard can automatically synchronize the data by applying the redo data to the standby database that was archived at the primary database during the outage period. Data Guard ensures that the data is logically and physically consistent before it is applied to a standby database.

A standby database is a transactionally consistent copy of an Oracle production database that is initially created from a backup copy of the primary database. After the standby database is created and configured, Oracle Data Guard automatically maintains the standby database by transmitting primary database redo data to the standby system, where the redo data is applied to the standby database.

The following types of standby databases are available from Oracle database version 11g onwards:

- Physical.
- Logical.
- Snapshot.

Physical standby database[3]

A physical standby database is an exact, block-for-block copy of a primary database. A physical standby is maintained as an exact copy through a process called *Redo Apply*, in which redo data received from a primary database is continuously applied to a physical standby database using the database recovery mechanisms.

A physical standby database can be opened for read-only access and used to offload queries from a primary database. If a license for the Oracle Active Data Guard option has been purchased, Redo Apply can be active while the physical standby database is open, thus

[3] https://docs.oracle.com/database/121/SBYDB/standby.htm#SBYDB00110>

allowing queries to return results that are identical to what is returned from the primary database. This capability is known as the real-time query feature.

A physical standby database provides the following benefits:

- Disaster recovery and high availability.

  A physical standby database is a robust and efficient disaster recovery and high availability solution. Easy-to-manage switchover and failover capabilities allow easy role reversals between primary and physical standby databases, minimizing the downtime of the primary database for planned and unplanned outages.

- Data protection.

  A physical standby database can prevent data loss, even in the face of unforeseen disasters. It also supports all data types, and all DDL and DML operations that the primary database can support. It also provides a safeguard against data corruptions and user errors. Storage level physical corruptions on the primary database are not propagated to a standby database. Similarly, logical corruptions or user errors that otherwise cause data loss can be easily resolved.

- Reduction in primary database workload.

  Oracle Recovery Manager (RMAN) can use a physical standby database to off-load backups from a primary database, saving valuable CPU and I/O cycles.

  A physical standby database can also be queried while Redo Apply is active, which allows queries to be offloaded from the primary to a physical standby, further reducing the primary workload.

- Performance.

  The Redo Apply technology used by a physical standby database is the most efficient mechanism for keeping a standby database updated with changes being made at a primary database because it applies changes using low-level recovery mechanisms which bypass all SQL level code layers.

Logical standby database

A logical standby database is initially created as an identical copy of the primary database, but it later can be altered to have a different structure. The logical standby database is updated by executing SQL statements. The flexibility of a logical standby database lets you upgrade Oracle Database software (patch sets and new Oracle Database releases) and perform other database maintenance in rolling fashion with almost no downtime. From Oracle Database 11g onward, the transient logical database rolling upgrade process can also be used with existing physical standby databases.

Oracle Data Guard automatically applies information from the archived redo log file or standby redo log file to the logical standby database by transforming the data in the log files into SQL statements and then executing the SQL statements on the logical standby database. Because the logical standby database is updated using SQL statements, it must remain open. Although the logical standby database is opened in read/write mode, its target tables for the regenerated SQL are available only for read-only operations. While those tables are being updated, they can be used simultaneously for other tasks such as reporting, summations, and queries.

A logical standby database is ideal for high availability (HA) while still offering data recovery (DR) benefits. Compared to a physical standby database, a logical standby database provides significant additional HA benefits:

- Minimizing downtime on software upgrades.

A logical standby database is ideal for upgrading an Oracle Data Guard configuration in a rolling fashion. Logical standby can be used to greatly reduce downtime associated with applying patch sets and new software releases. A logical standby can be upgraded to the new release and then switched over to become the active primary. This allows full availability while the old primary is converted to a logical standby and the patch set is applied. Logical standbys provide the underlying platform for the DBMS_ROLLING PL/SQL package, which is available as of Oracle Database 12c Release 1 (12.1). The DBMS_ROLLING package provides functionality that allows you to make your Oracle Data Guard configuration highly available in the context of rolling upgrades and other storage reorganization.

- Support for reporting and decision support requirements.

  A key benefit of logical standby is that significant auxiliary structures can be created to optimize the reporting workload; structures that can have a prohibitive impact on the primary's transactional response time. A logical standby can have its data physically reorganized into a different storage type with different partitioning, have many different indexes, have on-demand refresh materialized views created and maintained, and can be used to drive the creation of data cubes and other OLAP data views. However, a logical standby database does not allow for any transformation of your data (such as replicating only a subset of columns or allowing additional columns on user tables). For those types of reporting activities, Oracle GoldenGate is Oracle's preferred solution.

Snapshot standby database

A snapshot standby database is a type of updatable standby database that provides full data protection for a primary database. A snapshot standby database receives and archives, but does not apply, redo data from its primary database. Redo data received from the primary database is applied when a snapshot standby database is converted back into a physical standby database, after discarding all local updates to the snapshot standby database.

A snapshot standby database diverges from its primary database over time because redo data from the primary database is not applied as it is received. Local updates to the snapshot standby database cause additional divergence. The data in the primary database is fully protected however, because a snapshot standby can be converted back into a physical standby database at any time, and the redo data received from the primary is then applied.

A snapshot standby database is a fully updatable standby database that provides disaster recovery and data protection benefits that are similar to those of a physical standby database. Snapshot standby databases are best used in scenarios where the benefit of having a temporary, updatable snapshot of the primary database justifies the increased time to recover from primary database failures.

The benefits of using a snapshot standby database include the following:

- It provides an exact replica of a production database for development and testing purposes, while maintaining data protection at all times. You can use the Oracle Real Application Testing option to capture primary database workload and then replay it for test purposes on the snapshot standby.
- It can be easily refreshed to contain current production data by converting to a physical standby and resynchronizing.

The ability to create a snapshot standby, test, resynchronize with production, and then again create a snapshot standby and test, is a cycle that can be repeated as often as desired. The same process can be used to easily create and regularly update a snapshot standby for reporting purposes where read/write access to data is required.

More information about best practices for high availability and maximum availability for Oracle can be found here.

Oracle GoldenGate[4]

Oracle GoldenGate is an additional licensed software product that provides the capabilities to replicate, filter, and transform data between databases. Oracle GoldenGate enables replicating data between Oracle databases to other supported heterogeneous database, and between heterogeneous databases. Also, the Java Messaging Queues, Flat Files, and to Big Data targets in combination with Oracle GoldenGate for Big Data can be replicated. While it has many uses it often is utilized for data migrations, high availability and disaster recovery to help achieve business continuity.

More details about GoldenGate can be found here.

SAP HANA

SAP HANA is inherently designed for high availability. It can recover from most hardware faults, errors, and entire system or data center failure. Like many enterprise class application HANA provides three main levels disaster recovery support. They are as follows:

- Backups.

  SAP HANA database is in-memory for performance, it uses persistent storage to survive server outages without loss the of data. Two types of persistent storage are used:

  - Transaction redo logs.

    Changes are recorded, so that after an outage, the most recent consistent state of the database can be restored. This is achieved by replaying the changes recorded in the log, re-doing the completed changes and rolling back the incomplete ones.

  - savepoints for data changes.

    A savepoint is a consistent point in time across all SAP HANA processes when all data is written to storage. One goal is to reduce the time to recover from an outage, as the logs only need to be replayed from the latest savepoint.

    Normally savepoints overwrite previous savepoints, but they can be preserved for future use - equivalent to a snapshot, that can be used to rollback to a specific point in time.

  Shipping both the savepoints and transaction redo logs will allow recovery of the SAP HANA database after a disaster, and depending on the technology used, recovery time can range from hours to days.

- System replication.

  In general there is a single HANA instance at the primary site and another one at the secondary site. Each site has their own independent storage areas for the HANA data, log, and shared areas. In this DR scenario, the DR site has a fully duplicated environment for protecting your data from a total loss of the primary site. So, each HANA system has its own IP address, and each site has its own SAP application infrastructure pointing to that site's HANA DB IP address.

  The system replication technology within SAP HANA creates a unidirectional replication for the contents of the data and log areas. The primary site replicates data and logs to the secondary site, but not vice versa. The secondary system has a replication receiver status (secondary system), and can be set up for read-only DB access, thus not being idle.

[4] https://docs.oracle.com/goldengate/c1230/gg-winux/GGCON/introduction-oracle-goldengate.htm#GGCON-GUID-EF513E68-4237-4CB3-98B3-2E203A68CBD4

If there is a failure in the primary site, all you need to do is perform a takeover operation on the secondary node. This is a DB operation that is performed by the basis team and informs the secondary node to come online with its full range of capabilities and operate as a normal, and independent instance. The replication relationship with the primary site is broken. When the failed node comes back online, it is outdated in terms of DB content, but all you need to do is create the replication in the reverse order, from the secondary site to the primary site. After your sites are synchronized again, you can choose to perform another takeover operation to move the DB back to its original primary site.

Storage ReplicationA problem with backups is always the loss of data between the time of failure and the last backup. A common preferred method is to replicate all data. Many storage vendors offer storage-based replication solutions. There are some certified SAP vendor-specific solutions that provide synchronous replication. This means that the transaction is only marked completed when the locally persisted transaction log has been replicated remotely. Synchronous storage replication technically has no distance limitation per se, but often it 100 kilometers or less. This is primarily for performance reasons to keep round trip latency to a minimum and acceptable level.

High availability for SAP HANA

SAP HANA is designed for high availability and supports recovering from hardware and software errors. High availability is achieved by eliminating single points of failure and is designed to rapidly resume operations with minimum business loss after a system outage. SAP HANA also supports a disaster recovery configuration with multiple data centers.

Since SAP HANA is an in-memory database, it is designed to both manage the integrity of data in memory in the event of a failure and loading it back as quickly as possible after the failure.

SAP HANA uses the following components for high availability:

- A watchdog to automatically restart any stopped services.
- Ability to failover from a crashed host to a standby host.
- System replication.

  This process replicates the in-memory databases from the primary system to a secondary system. This configuration offers a number of solutions:

  - High availability with pre-load allowing faster recovery.
  - Disaster recovery with replication to another site.
  - Load sharing with reporting running against the secondary system.

  System replication supports database replication at the system level or tenant databases.

SAP HANA supports the following for disaster recovery:

- Off-site storage of backups.
- Storage replication to remote data center (synchronous or asynchronous).
- System replication.
- Virtual Persistent Memory (VPMEM).

  Virtual Persistent Memory (PMEM) is an enhancement to PowerVM that introduces the ability to configure persistent volumes using the existing DRAM technology. This persistent memory solution on Power Systems is being made available on existing IBM POWER9 (and soon to be releases IBM Power10™) processor-based systems. There are no special or additional HW components or memory modules required on IBM Power Systems with this solution. This functionality is built on top of the standard memory DIMMs that are available on IBM Power Systems.

Virtual PMEM solution reduces both shut down and start-up time of SAP HANA, thus significantly reducing maintenance related outage time as long as the VIO Servers remain active. More information about virtual PMEM can be found here.

Using secondary servers for non-production systems

With SAP HANA system replication, you can use the servers on the secondary system for non-production SAP HANA systems under the following conditions:

– Table pre-load is turned off in the secondary system.

– The secondary system uses its own disk infrastructure. In the case of single node systems this means, the local disk infrastructure needs to be doubled.

– The non-production systems are stopped with the takeover to the production secondary.

Summary of replication and log shipping options

Table 2-2 summarizes the features of each option.

Table 2-2 Replication and log shipping options

Database options	Tier	Storage unit fail		Site failure	
		RTO	RPO	RTO	RPO
Concurrent databases	7	0	0	0	0
Log shipping	6	log[a] freq	log freq	log freq	log freq

[a] log freq = frequency at which the logs are shipped

2.4.5 LPAR (or VM) availability management options

System administrators require the ability to move LPARs in the normal course of maintaining the environment to manage repairs or VIO Server and firmware updates and for load balancing and server resource constraints. However this will not help if the server unexpectedly halts. In that case, Simplified Remote Restart (SRR) and Virtual Machine Recovery Manager (VMRM) can help restart LPARs on other server(s).

> **Note:** IBM Laboratory Services have developed a GUI based tool to simplify the management of LPM and SRR called the *IBM PowerVM LPM/SRR Automation tool*. A demo of this tool is available here.

Live Partition Mobility (LPM)

Live Partition Mobility (LPM) is a component of the PowerVM Enterprise Edition hardware feature moves AIX, IBM i, and Linux LPARs from one system to another one. The mobility process transfers the system environment, which includes the processor state, memory, attached virtual devices, and connected users. All OS types (AIX, IBM i, and Linux) on Power Systems can utilize LPM. However a VIOS LPAR *cannot* as it has dedicated adapter resources. The single biggest key requirement for an LPAR to use LPM is that its adapter devices must all be virtualized. There are two primary mobility methods listed as follows.

Active partition mobility moves LPARs that are running, including the operating system and applications, from one system to another. The LPAR and the applications running on that migrated LPAR do not need to be shut down.

Inactive/Static partition mobility moves a powered-off AIX, IBM i, or Linux LPAR from one system to another.

Partition mobility provides systems management flexibility and can be utilized to improve system availability. For example:

- Planned outages for hardware or firmware maintenance can be avoided by migrating LPARs to another server and then performing the maintenance. Partition mobility can help because you can use it to work around scheduled maintenance activities.
- Outages for server hardware upgrades can be mitigated by migrating LPARs to another server and then performing the upgrade. This allows work to continue without disruption.
- In the event of a predictive server, LPARs can be migrated to another server before the failure occurs. Partition mobility can help avoid unplanned downtime.
- Consolidating workloads running on several small, under utilized servers onto a single large server.
- Workload balancing from server to server to optimize resource use and workload performance within your computing environment. With active partition mobility, you can manage workloads with minimal, if any, downtime.
- On some Power Systems applications be moved from one server to an upgraded server by using IBM PowerVM Editions LPM or the *AIX Live Application Mobility software* without affecting availability of the applications.

However, while partition mobility provides many benefits, it does *not* provide the following functions:

- Automatic workload balancing.
- Provide a bridge to new functions. LPARs must be restarted and possibly reinstalled to take advantage of new features.
- High availability.

During an LPM event a matching profile/clone partition is automatically created on the target server. The partition's memory is asynchronously copied from the source system to the target server. Any changed memory pages from the partition ("dirty" pages) are recopied at the end. After a threshold is reached that indicates that enough memory pages were successfully copied to the target server, the LPAR on that target server becomes active, and any remaining memory pages are copied synchronously. The original LPAR is then automatically is then deleted from the source server.

An inactive LPAR that has never been activated cannot be migrated because the HMC always migrates the last activated profile. In this case to utilize inactive partition mobility you can either select the partition state that is defined in the hypervisor or select the configuration data that is defined in the last activated profile on the source server.

Detailed LPM requirements can be found here.

Application mobility for WPARs

Workload partitions (WPARs) are virtualized operating system environments within a single instance of the AIX operating system. The mobility feature was managed by IBM System Director which is no longer in support.

LPAR and VM restart options

In this section the focus is upon the ability to relocate and activate an LPAR in the primary event of a hard outage where LPM inactive mobility cannot be used.

Remote restart

Remote restart is a high availability option for AIX, IBM i or Linux logical partitions when using PowerVM Enterprise Edition, PowerVM, or Linux edition. When an error causes a server outage, a partition that is configured with the remote restart capability can be restarted on a different physical server. Sometimes, it might take longer to start the server, in which case the remote restart feature can be used for faster re-provisioning of the partition. This operation completes faster as compared to restarting the server that failed and then restarting the partition. Remote restart is supported on POWER7 and newer processor-based systems.

The following are the characteristics of the remote restart feature:

- The remote restart feature is not a Suspend/Resume or migration operation of the partition that preserves the active running state of the partition. During the remote restart operation, the logical partition is shut down and then restarted on a different system.
- The remote restart feature preserves the resource configuration of the partition. If processors, memory or I/O are added or removed while the partition is running, the remote restart operation activates the partition with the most recent configuration.

The remote restart feature requires a reserved storage device that is assigned to each partition. You must manage a reserved storage device pool on both the source and the destination servers, and maintain a record of the device that is assigned to each partition. The simplified remote restart feature does not require a reserved storage device that is assigned to each partition.

The remote restart feature (including the simplified version) is not supported from the HMC for logical partitions that are co-managed by the HMC and PowerVM NovaLink. However, you can run simplified remote restart operations by using PowerVC with PowerVM NovaLink. However this feature has been mostly superseded by *simplified remote restart*.

Simplified Remote Restart

Similar to remote restart, Simplified Remote Restart (SRR), is feature available in PowerVM Enterprise Edition that can restart AIX, IBM i, and Linux LPARs on a different physical server when the original server is no longer active.If the source physical server has an error that causes it to halt, you can restart the LPARs on another (target) server. This may sound similar to inactive partition mobility but the key difference is that the source physical server itself is no longer available or accessible.

If the source server has a physical fault, SRR can be utilized to recover the key LPARs quickly. In some instances restarting the Server may be a lengthy process, in this case, the use of SRR can provide a shorter recovery time.

SRR with HMC Version 8.2.0 and later running on IBM POWER8® FW 8.2.0 and later removes the need to assign reserved storage to each LPAR and is recommended.

The characteristics of SRR are as follows:

- SRR is *not* a suspend and resume or migration operation of the partition that preserves the active running state of the partition. During the remote restart operation, the halted/crashed LPAR is started on a different system.
- SRR preserves the resource configuration of the partition. If processors, memory, or I/O are added or removed while the partition is running, the remote restart operation activates the partition with the most recent configuration.

When an LPAR is restarted by way of SRR, a new profile is automatically created on the target server that matches the profile on the source server. That new profile is then mapped to the storage LUNs that were being used by the original partition (that partition being inactive). The new profile on the target server is then activated and the partition is again active. When the source server becomes active, you must remove the old profile to ensure that the partition is not accidentally restarted on that server (if it restarts automatically). The automatic cleanup runs without the force option, which means that if a failure occurs during the cleanup (for example, RMC communications with the VIOS fails), the LPAR is left on the original source server and its status marked as *Source Side Cleanup Failed*.

The prerequisites for SRR are similar to LPM. In short if LPM does not work for an LPAR, then SRR will not work either.

Other than the minimum required FW, HMC versions, and VIOS versions, the high-level SRR prerequisites include:

- Remote restart must be enabled on the virtual machine. You can set this option while deploying or resizing the virtual machine.
- Remote restart must be enabled on the host.
- The hosts and virtual machines must be capable of simplified remote restart capability.
- The source system must be in a state of *Initializing, Power Off, Powering Off, No connection, Error,* or *Error - Dump in progress*.
- The source systems VIOSs that provide the I/O for the LPAR must be *inactive*.
- The target system must be in an *active* state.
- The target systems VIOSs that provide the I/O for the LPAR must be *active*.
- The LPAR that will be restarted must be in an *inactive* state.
- The LMB size is the *same* on the source and the target system.
- The target system must have enough available resources (processors and memory) to host the partition.
- The target system VIOSs must be able to provide the networks that are required for the LPAR.

The simplified version of the remote restart feature is recommend, when the firmware is at level 8.2.0, or later, and the HMC is at version 8.2.0, or later.

PowerVC automated remote restart

SRR is available by way of both an HMC and PowerVC. However, PowerVC also adds another level of high availability by adding an automated operation for SRR. This is because the HMC only has a hosts view.

Automated remote restart monitors hosts for failure by using the PRS (Platform Resource Scheduler) HA service. If a host fails, PowerVC automatically remote restarts the virtual machines from the failed host to another host within a host group.

Without automated remote restart enabled, when a host goes into *Error* or *Down* state, you must manually trigger the remote restart operation, but you can manually remote restart virtual machines from a host at any time, regardless of its automated remote restart setting. More details about automated remote restart with PowerVC can be found here.

> **Demo:** A demonstration of automated remote restart capability is available at `https://www.youtube.com/watch?v=6s72ZR5OLr8`.

IBM VM Recovery Manager HA

IBM VM Recovery Manager HA (VMRM HA) for Power Systems is a high availability solution that is easy to deploy and provides an automated solution to recover the virtual machines (VMs), also known as logical partitions (LPARs). It supports all three of the OS types supported on Power Systems of AIX, IBM i, and Linux.

The VM Recovery Manager HA solution implements recovery of the virtual machines based on the VM restart technology. The VM restart technology relies on an out-of-band monitoring and management component that restarts the VMs on another server when the host infrastructure fails. The VM restart technology is different from the conventional cluster-based technology that deploys redundant hardware and software components for a near real-time failover operation when a component fails.

The VM Recovery Manager HA solution is ideal to ensure high availability for many VMs. Additionally, the VM Recovery Manager HA solution is easier to manage because it does not have clustering complexities.

The VM Recovery Manager HA solution provides the following capabilities:

Host health monitoring The VM Recovery Manager HA solution monitors hosts for any failures. If a host fails, the virtual machines in the failed host are automatically restarted on other hosts. The VM Recovery Manager HA solution uses the host monitor module of the VIOS partition in a host to monitor the health of hosts.

VM and app monitoring The VM Recovery Manager HA solution monitors the virtual machines, its registered applications, and its hosts, for any failures. If a virtual machine or a critical application fails, the corresponding virtual machines are started automatically on other hosts. The VM Recovery Manager HA solution uses the VM monitor agent that must be installed in each virtual machine to monitor the health of virtual machines and registered applications.

Unplanned HA events During an unplanned outage, when the VM Recovery Manager HA solution detects a failure in the environment, the virtual machines are restarted automatically on other hosts. You can also change the auto-restart policy to advisory mode. In advisory mode, failed VMs are not relocated automatically, instead email or text messages are sent to the administrator. Administrator can use the interfaces to manually restart the VMs.

Planned HA events During a planned outage, when you plan to update firmware for a host, you can use the Live Partition Mobility operation of the VM Recovery Manager HA solution to vacate a host by moving all the VMs in the host to the remaining hosts in the group. After the upgrade operation is complete, you can use the VM Recovery Manager HA solution to restore the VM to its original host in a single operation.

Advanced HA policies The VM Recovery Manager HA solution provides advanced policies to define relationships between VMs such as collocation and anti-collocation of VMs, priority in which the VMs will be restarted, capacity of VMs during failover operations.

GUI and CLI mgmt You can use GUI or command-line interface to manage the resources in the VM Recovery Manager HA solution. For GUI, you can install the UI server and then use the web browser to manage the resources. Alternatively, the **ksysmgr** command and the **ksysvmmgr** command on KSYS LPAR provide end-to-end HA management for all resources.

IBM VM Recovery Manager DR

The IBM VM Recovery Manager DR (VMRM DR) for Power Systems, formerly known as IBM Geographically Dispersed Resiliency (GDR), consists of both HA and DR offering in the same package. This solution is a disaster recovery solution that is easy to deploy and provides automated operations to recover the production site. The VM Recovery Manager DR solution is based on the IBM Geographically Dispersed Parallel Sysplex® (IBM GDPS®) offering concept that optimizes the usage of resources. This solution does not require you to deploy the backup virtual machines (VMs) for disaster recovery. Thus, the VM Recovery Manager DR solution reduces the software license and administrative costs.

Clustered HA and DR solutions typically deploy redundant hardware and software components to provide near real-time failover when one or more components fail. The VM restart-based HA and DR solution relies on an out-of-band monitoring and management component that restarts the virtual machines on other hardware when the host infrastructure fails. The VM Recovery Manager DR solution is based on the VM restart technology.

The VM Recovery Manager DR solution automates the operations to recover your production site. This solution provides an easy deployment model that uses a controller system (KSYS) to monitor the entire virtual machine (VM) environment. This solution also provides flexible failover policies and storage replication management.

Table 2-3 identifies the differences between the conventional cluster-based disaster recovery model and the VM Recovery Manager DR solution.

Table 2-3 Clustered DR versus VM Recovery Manager DR

Parameters	**Cluster-based disaster recovery model**	**VM restart disaster recovery model that is used by the VMRecovery Manager DR solution**
Deployment method	Redundant hardware and software components are deployed in the beginning of implementation to provide near real-time failovers when some of the components fail.	With virtualization technology, many images of the operating system are deployed in a system. These virtual machines are deployed on physical hardware by the hypervisor that allocates and manages the CPU, memory, and I/O physical resources that are shared among the VMs.
Dependency	This solution relies on monitoring and heartbeat capabilities within the cluster to monitor the health of the cluster and take recovery action if a failure condition is detected.	This solution relies on an out-of-band monitoring software that works closely with the hypervisor to monitor the VM environment and to provide a disaster recovery mechanism for the VM environment.
Workload startup time	The workload startup time is faster because the virtual machines and the software stack are already available.	The VMs require additional time to restart in the backup environment.
Cluster administration required	Yes	No

Parameters	Cluster-based disaster recovery model	VM restart disaster recovery model that is used by the VMRecovery Manager DR solution
Error coverage	Comprehensive. This solution monitors the entire cluster for any errors.	Limited. This solution monitors the servers and the virtual machines for errors.
Deployment simplicity	This solution must be set up in each VM.	Aggregated deployment at the site level.
Protected workload type	Critical workloads can be protected by using this solution.	Critical workloads can be protected by using this solution.
Software license and administrative costs	This solution costs more because redundant software and hardware are required to deploy this solution.	This solution costs less because of optimized usage of resources.

Demo: A demonstration of VMRM DR, under its original name of GDR, capability is available at https://www.youtube.com/watch?v=kTeOTzpOghs&t=8s.

Summary of LPAR availability management options

Table 2-4 compares the features of the different LPAR management options.

Table 2-4 Comparing features of the LPAR management solutions in the IBM portfolio

	Live Partition Mobility	Simplified remote restart	VM Restart HA	VM Restart DR
Support	≥ p6	≥ p7	≥ p7+	≥ p7
Frame failure	N	Y	Y	Y
VM Monitor	N	N	Agent (AIX)	Agent (AIX)
Auto failover	N	N	Y	Y
Storage	Shared	Shared	Shared	Replicated
Clustering	N	N	N	N
Active-passive	Y	Y	Y	Y
DR	N	N	N	Y
Automated Failover	N	N	Y	N
Source Server Status	Active	Inactive	Active or Inactive	Active or Inactive
Source VIO Server Status	Active	Inactive	Active or Inactive	Active or Inactive
VM/Application Outage	No (if LPAR active)	Y	Only if Frame/LPAR outage	Yes

	Live Partition Mobility	Simplified remote restart	VM Restart HA	VM Restart DR
RTO	N/A	Operator + IPL + App start	IPL + App start	VMRM HA time if local, DR+
RPO	N/A	0	0	sync 0; async cache
Tier	N/A	5[a]	6[a]	6(async); 7(sync)
License usage	N/A	N + 0	N + 0	N + 0
Cost	N/A[a]	$	$$	$$

[a] Within one data center.

2.4.6 Clustering options

The following section covers some application agnostic clustering options available from IBM. Though some of them offer additional tight integration with specific applications they are generally considered a one size fits many solution.

Tivoli System Automation for Multiplatform

IBM Tivoli System Automation for Multiplatforms (TSA MP) is cluster managing software on Linux and AIX that facilitates automatic switching of users, applications, and data from one database system to another in a cluster. Tivoli SAMP automates control of IT resources such as processes, file systems, and IP addresses. It generally is a separate licensed product but does come bundled with some application like IBM Db2.

High availability and resource monitoring

IBM Tivoli System Automation provides a high availability environment for applications and business systems. High availability describes a system which is continuously available and which has a self-healing infrastructure to prevent downtime caused by system problems. Thus it relieves operators from manual monitoring, remembering application components and relationships, and can eliminate operator errors.

Policy-based automation

IBM Tivoli System Automation for Multiplatforms allows you to configure high availability systems through the use of policies that define the relationships among the various components. After you establish the relationships, IBM Tivoli System Automation for Multiplatforms will assume responsibility for managing the applications on the specified nodes as configured per policy.

Automatic recovery

IBM Tivoli System Automation for Multiplatforms quickly and consistently performs an automatic restart of failed resources or whole applications either in place or on another system of a Linux or AIX cluster.

Automatic movement of applications

IBM Tivoli System Automation for Multiplatforms manages the cluster-wide relationships among resources for which it is responsible. If applications need to be moved among nodes, IBM Tivoli System Automation for Multiplatforms automatically handles the start and stop relationships, node requirements, and any preliminary or follow-up actions.

Resource grouping

You can group resources together in IBM Tivoli System Automation for Multiplatforms. After grouped, all relationships among the members of the group can be established, such as location relationships, or start and stop relationships. After you complete configuration, operations can be performed against the entire group as a single entity.

End-to-end automation management

IBM Tivoli System Automation for Multiplatforms now provides all the features for a heterogeneous server environment (z/OS, Linux, and AIX) enabling true business application automation.

Tivoli SA MP provides a framework to automatically manage the availability of what are known as resources. Here are some examples of resources:

- Any piece of software for which start, monitor, and stop scripts can be written to control
- Any network interface card (NIC) to which Tivoli SA MP was granted access. That is, Tivoli SA MP manages the availability of any IP address that a user wants to use by floating that IP address among NICs that it has access to. This is known as a floating or virtual IP address.

TSA MP can use these resources for local data storage:

- Raw disk (for example, /dev/sda1).
- Logical volume that is managed by Logical Volume Manager (LVM).
- File system (for example, ext3, jfs).

For more information about TSA for multiplatforms see the base publications at:

`https://www.ibm.com/docs/en/tsafm/4.1.0`

IBM PowerHA SystemMirror

IBM PowerHA SystemMirror (PowerHA) for AIX, IBM i, and Linux is a separate licensed product that provides HA clusters on IBM Power Systems. A PowerHA cluster must contain a minimum of two LPARs (called nodes) that communicate with each other by using heartbeats and keepalive packets. The cluster contains many resources, such as IP addresses, shared storage, and application scripts, that are grouped to form a resource group.

If PowerHA detects an event within the cluster, it automatically acts to ensure that the resource group is placed on the most appropriate node in the cluster to ensure availability. A correctly configured PowerHA cluster after setup requires no manual intervention to protect against a single point of failure, such as failures of physical servers, nodes, applications, adapters, cables, ports, network switches, and storage area network (SAN) switches. The cluster can also be controlled manually if the resource groups must be balanced across the clusters or moved for planned outages.

PowerHA for AIX comes in two editions, *Standard* and *Enterprise*. Standard edition is generally more synonymous with local high availability, and in some configurations even near distance disaster recovery. It is dependent on both shared LAN and SAN connectivity between servers and storage. A basic local cluster is shown in Figure 2-14 on page 67.

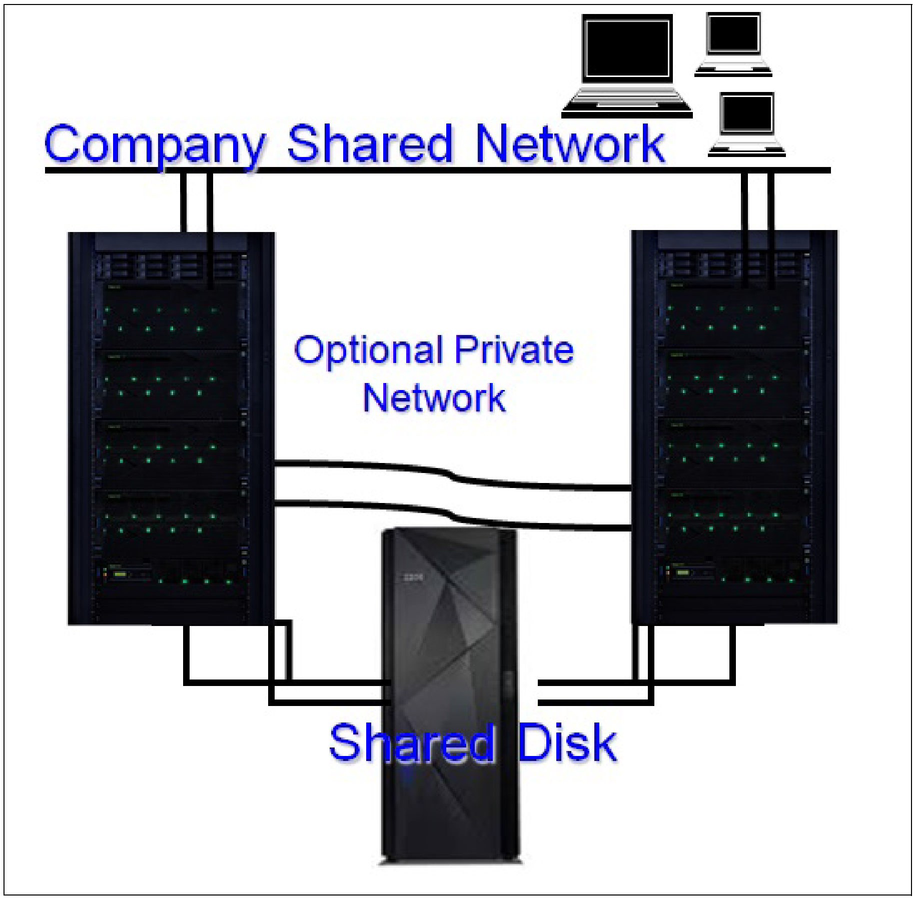


Figure 2-14 PowerHA Standard Edition Cluster

PowerHA Enterprise Edition includes everything standard edition does but also provides cross-site clustering where shared storage is not an option but SAN-based replication is available. In this environment, PowerHA uses the remote copy facilities, either IP or storage based, to ensure that the nodes at each site have access to the same data, but on different storage devices. It is possible to combine both local and remote nodes within a PowerHA cluster to provide local HA and cross-site DR.

PowerHA clusters can be configured in many ways:

- Active/Passive: One node in the cluster runs the resource group, and its partners are in standby mode waiting to take on the resources when required. The passive nodes in the cluster must be running in order for them to participate in the cluster.
- Active/Active: All nodes in the cluster are running a resource group but are also the standby node for another resource group in the cluster. Many resources groups can be configured within a cluster, so how they are spread out across the nodes and in which order they move is highly configurable.
- Concurrent: All nodes in the cluster run the same resource group. This historically was most common with Oracle RAC environments, however, some application servers can also be used this configuration.

AIX version

PowerHA, formerly known as High Availability Cluster Multi-Processing (HACMP), has been popular in its over 30 year history. Originally designed as a stand-alone product (known as HACMP classic) after the IBM high availability infrastructure known as RSCT) became available, HACMP adopted this technology and became HACMP Enhanced Scalability (HACMP/ES) because it provides performance and functional advantages over the classic version. Starting with HACMP V5.1, there are no more classic versions. Later HACMP terminology was replaced with PowerHA in Version 5.5 and then PowerHA SystemMirror V6.1.

PowerHA V7.1 was the first version to utilize Cluster Aware AIX (CAA) component of AIX. This major change improves the reliability of PowerHA because the cluster service functions now run in kernel space rather than user space. CAA was introduced in AIX 6.1 TL6 and AIX 7.1 TL0. At the time of writing, the current release of PowerHA is V7.2.5.

While most clusters are simple two-node active/passive cluster, PowerHA SystemMirror for AIX supports 16 nodes in a cluster allowing a varied of failover options. Some of these options are shown in Figure 2-15.

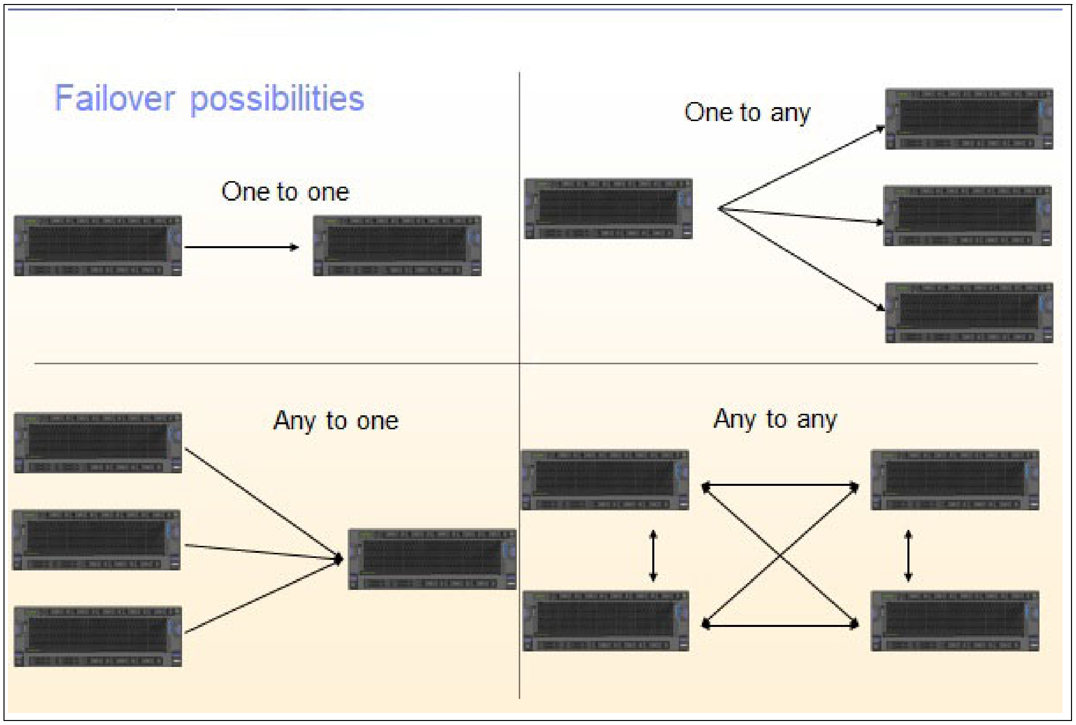


Figure 2-15 PowerHA failover options

PowerHA is packed full of options and features. Many of them are tightly integrated to both AIX and PowerVM specific features. Some of these Standard Edition features are shown as follows and also linked to online demos where available:

- Dynamic Node Priority (DNP).
  - Target and fallover node chosen by available resources such as:
    - Free CPU.
    - Paging space.
    - Disk I/O.
- Dynamic LPAR additional CPU or memory during startup or fallover.
  - Includes Resource Optimized failovers by using Enterprise Pools (Resource Optimized High Availability (ROHA)).
- LPM awareness.
- Live Kernel Update awareness.

- Resource group dependencies.
  - Great for multi-tier environments.
    - Parent/Child.
    - Same node or same site.
    - Different node or site.
- Resource group priorities.
  - Low.
  - Intermediate.
  - High.
- AIX LVM/JFS2 specialized option utilization.
  - File system Concurrent Mount Protection, known as Mount Guard.
  - Active/Passive mode of concurrent volume groups.
- Non-disruptive cluster updates and upgrades by way of cl_ezupdate.
- NovaLink managed LPARs support.
- Rootvg and critical volume group loss detection.
- User defined events.
- Customizable processing order.
- Automatic repository replacement.
- Cluster testing, both automated and customizable.
- Delayed Fallback Timer.

Enterprise Edition provides additional integrated supported primarily focused around disaster recovery. Some of these features are:

- IP-based replication, GLVM.
- IBM Spectrum Virtualize Storage Replication.
  - Metro Mirror.
  - Global Mirror.
  - HyperSwap.
- EMC.
  - SRDF - Synchronous or Asynchronous.
- Hitachi.
  - TrueCopy for synchronous.
  - Hitachi Universal Replicator for asynchronous.
- IBM XIV®.
  - Remote Mirror.
- User confirmation on split site failure.
- Site-specific service addresses.

Additional details about planning, installing and configuring PowerHA SystemMirror for AIX can be found in the following sources:

- *IBM PowerHA SystemMirror for AIX Cookbook*, SG24-7739.
- *Guide to IBM PowerHA SystemMirror for AIX Version 7.1.3*, SG24-8167.
- *IBM System Storage Solutions Handbook*, SG24-5250.

- *Base PowerHA 7.2.x publications.*

IBM i version

PowerHA SystemMirror for IBM i has been around since 2008 and shares many similarities with the AIX version. It is deeply integrated with IBM i and SLIC. However it offers three editions, *Express*, *Standard* and Enterprise.

Express Edition enables single-node, full-system HyperSwap with the DS8700 server. This provides continuously available storage through either planned or unplanned, storage outage events. Standard Edition is generally for local data center high availability and Enterprise Edition for multi-site, disaster recovery solutions.

PowerHA for i cluster configurations are also flexible. It is becoming more common for IBM i customers to deploy multi-site PowerHA clusters where the data is replicated either by IBM storage or by Geographic Mirroring. PowerHA integrates the IBM i operating system with storage replication technologies providing solutions that meet the high availability needs of clients, regardless of size.

Configurations range from a simple two-system two-site cluster using Geographic Mirroring with internal storage, to an IBM FlashSystem cluster or a three-site HyperSwap cluster with IBM DS8000 storage. Exploiting IBM storage adds the additional benefit of FlashCopy functionality, which is used to eliminate the backup window, conduct query operations and to create point in time copies for data protection purposes.

The production data, including the local journals, are contained within an Independent Auxiliary Storage Pool (IASP), planned switchovers between nodes in the cluster consists of a single command. Unplanned failovers can be configured to be automatic, requiring minimal operator intervention. The administration domain takes care of synchronizing security and configuration objects such as user profiles. This is all done with the integration between PowerHA and the IBM i operating system, and has no dependency on third-party replication tools. Since there is at least one active operating system on each node in the cluster, you are able to conduct software maintenance and OS upgrades on an alternate node without disrupting production.

Implementing IASPs is a simple task consisting of moving your application libraries and IFS data into the IASP, thus separating business data from the operating system. The application binaries do not change, and most users are completely unaware of the migration in their daily workflow as their jobs automatically have access to libraries both in the system ASP and the independent ASP simultaneously.

Demo: A demo of PowerHA for IBM i utilizing Geographic Mirroring can be found at https://www.youtube.com/watch?v=k_C8PbhSBCM.

Additional details about planning, installing and configuring PowerHA SystemMirror for i can be found in the following sources:

- *PowerHA SystemMirror for IBM i Cookbook*, SG24-7994.
- *IBM PowerHA SystemMirror for i: Preparation (Volume 1 of 4)*, SG24-8400.
- *IBM PowerHA SystemMirror for i: Using DS8000 (Volume 2 of 4)*, SG24-8403.
- *IBM PowerHA SystemMirror for i: Using IBM Storwize (Volume 3 of 4)*, SG24-8402.
- *IBM PowerHA SystemMirror for i: Using Geographic Mirroring (Volume 4 of 4)*, SG24-8401.
- Base 7.4 publications here.

Linux

The PowerHA SystemMirror for Linux offering was withdrawn from marketing as of September 29, 2020. The official IBM replacement is VMRM HA but its RTO is higher than general clustering.

However, additional information about PowerHA SystemMirror for Linux can be found at the following:

- *IBM PowerHA SystemMirror V7.2.3 for IBM AIX and V7.22 for Linux*, SG24-8434.
- IBM PowerHA SystemMirror for Linux base publications here.

Additional HA offerings for Linux are listed as follows.

Red Hat high availability add-on

The Red Hat High Availability Add-On is a clustered system that provides reliability, scalability, and availability to critical production services. It consists of several components and the major components are as follows:

- Cluster infrastructure.

  Provides fundamental functions for nodes to work together as a cluster: configuration file management, membership management, lock management, and fencing.
- High availability service management.

  Provides failover of services from one cluster node to another in case a node becomes inoperative.
- Cluster administration tool.

  Configuration and management tools for setting up, configuring, and managing the High Availability Add-On. The tools are for use with the cluster infrastructure components, the high availability and service management components, and storage.

You can supplement the High Availability Add-On with the following components:

- Red Hat GFS2 (Global File System 2).

  Part of the Resilient® Storage Add-On, this provides a cluster file system for use with the High Availability Add-On. GFS2 allows multiple nodes to share storage at a block level as if the storage were connected locally to each cluster node. GFS2 cluster file system requires a cluster infrastructure.
- LVM Locking Daemon (lvmlockd).

  Part of the Resilient Storage Add-On, this provides volume management of cluster storage. lvmlockd support also requires cluster infrastructure.
- HAProxy.

  Routing software that provides high availability load balancing and failover in layer 4 (TCP) and layer 7 (HTTP, HTTPS) services.

Pacemaker

Pacemaker is a cluster resource manager. It achieves maximum availability for your cluster services and resources by making use of the cluster infrastructure's messaging and membership capabilities to deter and recover from node and resource-level failure.

Pacemaker comprises separate component daemons that monitor cluster membership, scripts that manage the services, and resource management subsystems that monitor the disparate resources.

The following components form the Pacemaker architecture:

- Cluster Information Base (CIB).

  The Pacemaker information daemon, which uses XML internally to distribute and synchronize current configuration and status information from the Designated Coordinator (DC) to all other cluster nodes. DC is a node assigned by Pacemaker to store and distribute cluster state and actions by means of the CIB.

- Cluster Resource Management Daemon (CRMd).

  Pacemaker cluster resource actions are routed through this daemon. Resources managed by CRMd can be queried by client systems, moved, instantiated, and changed when needed.

  Each cluster node also includes a local resource manager daemon (LRMd) that acts as an interface between CRMd and resources. LRMd passes commands from CRMd to agents, such as starting and stopping and relaying status information.

- Shoot the Other Node in the Head (STONITH).

  STONITH is the Pacemaker fencing implementation. It acts as a cluster resource in Pacemaker that processes fence requests, forcefully shutting down nodes and removing them from the cluster to ensure data integrity. STONITH is configured in the CIB and can be monitored as a normal cluster resource.

- corosync.

  corosync is the component, and a daemon of the same name, that serves the core membership and member-communication needs for high availability clusters. It is required for the High Availability Add-On to function.

  In addition to those membership and messaging functions, corosync also:

  - Manages quorum rules and determination.
  - Provides messaging capabilities for applications that coordinate or operate across multiple members of the cluster and thus must communicate stateful or other information between instances.
  - Uses the kronosnet library as its network transport to provide multiple redundant links and automatic failover.

- pcs.

  The pcs command line interface controls and configures Pacemaker and the corosync heartbeat daemon. A command-line based program, pcs can perform the following cluster management tasks:

  - Create and configure a Pacemaker/Corosync cluster.
  - Modify configuration of the cluster while it is running.
  - Remotely configure both Pacemaker and Corosync and start, stop, and display status information of the cluster.

- pcsd Web UI.

  A graphical user interface to create and configure Pacemaker/Corosync clusters.

More detailed information about RHEL HA clustering on IBM Power Systems can be found here.

Red Hat OpenShift Container Platform cluster

Red Hat OpenShift is available on Linux and AIX LPARs both on-premises and in the Cloud with IBM Power Systems Virtual Server.

More detailed information about planning, installing and configuring Red Hat OpenShift Container Platform clusters on IBM Power Systems can be found here.

SUSE Linux Enterprise Server high availability extension

SUSE Linux Enterprise Server HA works in a similar fashion as PowerHA SystemMirror for AIX, as a simple comparison. There are virtual IP addresses, resource groups, heartbeating disks and networks. Cluster internals, virtual IP address placement and fail-over, and take-over operations, are all managed, operated, and controlled from within SUSE Linux Enterprise Server HA.

More detailed information about SUSE Linux Enterprise Server high availability extension can be found here.

Ubuntu

Ubuntu also offers numerous packages to created tailored HA solutions. More detailed information about Ubuntu HA core and community packages can be found here.

Summary of clustering options

Table 2-5 compares the features of the different clustering options discussed in this chapter.

Table 2-5 Comparing features of the HA/DR clustering options

	Tivoli Systems Automation	PowerHA	PowerHA EE	Linux clustering / pacemaker
Support	≥ p6	≥ p6	≥ p6	≥ p8
Frame failure	Y	Y	Y	Y
VM Monitor	Y	Y	Y	Y
Auto failover	Y	Y	Y	Y
Storage	Shared	Shared	Replicated	Replicated
Clustering	Y	Y	Y	Y
DR	Y	N (except Xsite LVM)	Y	Y
Automated Failover	Y	Y	Y	Y
VM/Application Outage	Yes	Yes	Yes	Yes
RTO	App start	App start	App start	App start
RPO	0	0	sync 0; async +	sync 0; async +
Tier	7[a]	7[a]	7	7
Node license usage	2N	N + 1	N + 1	N + 1
Cost	$$	$$	$$$	$$

[a] Within one data center.

2.4.7 Other IBM i offerings

This section provides additional IBM i offerings for high availability.

iCluster

Rocket iCluster is a software based HA/DR solution for IBM i to help maximized data availability and minimize downtime. It provides real-time, fault-tolerant, object-level replication that utilizes a "warm" mirror of a clustered IBM i system and can return production operations back into service within minutes.

Rocket iCluster can also be combined with IBM PowerHA SystemMirror. Rocket also has a community forum for iCluster that can be found at:

https://tinyurl.com/iclusterforum

More details about Rocket iCluster is also available at: can be found here.

Maxava HA

Maxava HA offers two editions, Enterprise+ and SMB. Maxava replicates data and objects in real-time (up to the last transaction) to multiple IBM i systems, regardless of location or configuration. Whether the backup server is in the same building, across town, interstate, in another country or in the cloud, Maxava HA can replicate data, objects, IFS, IBM MQ, QDLS and spooled files to a remote location of choice, maintaining data integrity at all times. Built on native IBM i Remote Journaling, Maxava HA keeps the overhead on the production server to an absolute minimum and comes complete with features which include:

- A highly-functional GUI usable for both the initial configuration and day-to-day monitoring.
- Unlimited concurrent apply processing built to handle enterprise-level transactional volumes.
- Multi-Threaded IFS which dynamically runs multiple IFS replication processes in parallel, increasing throughput so that replication is dramatically faster and more efficient in high-volume IFS environments.
- Simulated Role Swaps (SRS) designed for users to test their Disaster Recovery plan without downtime. SRS temporarily turns a backup system into a simulated primary system for role-swap readiness testing, while the primary system remains live and unaffected.
- Multi-Node Role Swap enabling role-swaps for customers with multiple target IBM i systems, which can include hardware replication options such as PowerHA.
- Remote Role Swap Capability allowing admins to perform a role swaps (in either direction) by way of command or from a mobile device.
- Flexible Autonomics designed for users to design their own self-healing requirements.
- User-definable Audits built to ensure data integrity at all times.
- Command Scripting Function enabling a predefined set of commands run to at failover to minimize role-swap times.

More details about Maxava HA can be found here.

Assure Mimix

Assure MIMIX provides full-featured, scalable High Availability and Disaster Recovery solutions by way of real-time logical replication. Assure MIMIX is IBM i journal based and includes extensive options for automating administration, comprehensive monitoring and alerting, data verification, customizable switch automation, and an easy to use graphical interface.

Assure MIMIX works across any combination of IBM i server, storage, and OS versions. It can provide HA and DR protection for just one IBM i server, or a multi-site mix of on-premises, remotely hosted, and Cloud Service-based systems. Assure MIMIX provides data protection and business continuity to help minimize, if not eliminate, planned and unplanned downtime.

Assure Mimix can also be combined with IBM PowerHA SystemMirror, Db2mirror and switchable Independent auxiliary storage pools (IASP) to provide options to manage risk and downtime.

More details about Assure Mimix can be found here.

Assure iTERA

Assure iTERA provides High Availability and Disaster Recovery solutions by way of real-time logical replication. It replicates IBM i data and objects in real time to local or remote backup servers. These servers then stand ready to assume the production role. It can also be used with a variety of IBM i OS levels and storage combinations, and is scalable from SMB to enterprise workloads.

More details about Assure iTERA can be found here.

Assure QuickEDD

Assure QuickEDD provides High Availability and Disaster Recovery solutions by way of real-time logical replication. It replicates IBM i data and objects in real time to local or remote backup servers. These servers then stand ready to assume the production role. It can also be used with a variety of IBM i OS levels and storage combinations, and is scalable from SMB to enterprise workloads.

More details about Assure QuickEDD can be found here.

Robot HA

Robot HA is a software-based high availability solution for IBM i 7.2 or higher that replicates important data, by way of IBM i remote journaling, to provide business continuity. Robot HA can provide fast, unplanned switchover to a target system, ideally at a remote location. Typical RTO is 15 to 30 minutes.

It provides many flexible options of how and what to replicate such as:

- Many-to-one.
- One-to-many.
- Object broadcast.
- Different library names.
- Only certain libraries.
- Only certain IFS directories.

It also provides:

- Simplified role swap for both audits and testing.
- Automatic resync.
- Automatic monitoring.

Robot HA can also be combined with IBM PowerHA SystemMirror.

More details about Robot HA can be found here.

2.4.8 Disaster recovery solution matrix

Table 2-6 on page 76 shows a summary of most options discussed throughout this chapter.

Table 2-6 DR Solution Matrix for Power Systems

Replication Method	Product	License on-prem	License cloud	License cost per core	Dedicated cloud capacity	RPO	RTO	Workload overhead	Automated	OPEX	Complexity	Cloud viable
Storage based	PHA EE (AIX & i)	N+1	N/A	$$	N/A	Sync 0 Async mins	App restart	0	Yes	1PH/Wk	low	No
	VMRM DR (AIX, i, Linux)	N+0	N/A	$	N/A	Sync 0 Async mins	System reboot	0	Yes	1PH/Wk	low	No
OS mirroring	PHA EE AIX GLVM	N+1	N+N[a]	$$	Yes	Sync 0 Async mins	App restart	20-40%	Yes	1PH/Wk	low	Yes
	PHA EE i Geo-mirror	N+1	N+N[a]	$$	Yes	Sync 0 Async mins	App restart	~10%	Yes	1PH/Wk	low	Yes
	PHA ihosting Geo-mirror	1+1 (hosting partition)	N+0 (guest partitions)	$$	Yes	Sync 0 Async mins	System reboot	~15%	Yes	1PH/Wk	low	Yes
Database replication (AIX)	Data Guard (Oracle) AIX	N+N	N+N	$$$	Yes	Sync 0 Async mins	Lag time		No	DBA manual failover	medium	Yes
	HADR (Db2)	N+N	N+N		Yes	Sync 0 Async mins	Lag time		No	DBA manual failover	medium	Yes
Middleware journal replication (IBM i)	iCluster Maxava Mimix Robot HA	N+M (for IBM i M=#licenses on target)	N+N (for IBM i)	Not published	Yes	Async mins	Variable lag time (queuing on target)	Variable 20-30% on target	No	Dedicated mgmt req manual failover	medium to high	Yes

[a] N+N for capacity = to the production side, you can choose to license target side at reduced capacity if desired.

Note: Cloud storage solutions for IBM i can be used for backup to the cloud. Bandwidth is a key factor.

3

Scenarios

This chapter delivers a series of case scenarios illustrating high availability and disaster recovery solutions.

This chapter contains the following solution scenarios:

- 3.1, “PowerHA for AIX Cross-Site LVM mirroring” on page 78.
- 3.2, “Standalone GLVM” on page 84.
- 3.2, “Standalone GLVM” on page 84.
- 3.3, “PowerHA for AIX Enterprise Edition with GLVM” on page 92.
- 3.4, “PowerHA for AIX Enterprise Edition with HyperSwap” on page 94.
- 3.5, “VM Recovery Manager High Availability” on page 98.
- 3.6, “VM Recovery Manager Disaster Recovery” on page 100.
- 3.7, “Tivoli System Automation for Multiplatform” on page 102.
- 3.8, “Spectrum Scale stretched cluster” on page 104.

3.1 PowerHA for AIX Cross-Site LVM mirroring

This section describes a disaster recovery solution, based on AIX LVM mirroring and a stretched PowerHA cluster. It is built from the same components generally used for local cluster solutions with SAN-attached storage. Cross-site LVM mirroring replicates data across the SAN between the disk subsystems at separate sites and PowerHA provides automated failover in the event of a failure. This solution can provide an RPO of zero, and RTO of mere minutes. The biggest determining factor in recovery time is application recovery and restart time.

Remote disks can be combined into a volume group by way of the AIX Logical Volume Manager (LVM) and this volume group can be imported to the nodes located at different sites. You can create logical volumes and set up a LVM mirror with a copy at each site. Though LVM mirroring supports upto three copies, PowerHA only supports two sites. However, it is possible to have two LVM copies locally, even two servers, locally at one site and one remote copy at another site however this generally is rare.

Though it is common to have the same storage type at each location, it is not a requirement. This is a perk for these configurations as they are storage type agnostic. As long as the storage is supported for SAN attachment to AIX and gives adequate performance it most likely is a valid candidate to be used in this configuration.

3.1.1 Compared to local cluster

The main difference between local clusters and clustered solutions with cross-site mirroring is as follows:

- For local clusters generally all nodes and storage subsystems are located in the same location.
- With cross-site mirrored, cluster nodes, and storage subsystems reside in two different site locations. Each site has at least one cluster node and one storage subsystem with all necessary IP network and SAN infrastructure.

This solution offers automation of AIX LVM mirroring within SAN disk subsystems between different sites. It also provides automatic LVM mirroring synchronization and disk device activation when, after a disk or site failure, a node or disk becomes available.

Each node in a cross-site LVM cluster accesses all storage subsystems. The data availability is ensured through the LVM mirroring between the volumes residing on separate storage subsystems at different sites.

In case of site failure, PowerHA performs a takeover of the resources to the secondary site according to the cluster policy configuration. It activates all defined volume groups from the surviving mirrored copy. In case one storage subsystem fails, data access is not interrupted and applications can access data from the active mirroring copy the on surviving disk subsystem.

PowerHA drives automatic LVM mirroring synchronization, and after the failed site joins the cluster, it automatically fixes removed and missing volumes (PV states *removed* or *missing*) and synchronizes data. Automatic synchronization is not possible for all cases, but C-SPOC can be used to synchronize the data from the surviving mirrors to stale mirrors after a disk or site failure.

3.1.2 General PowerHA requirements

The following AIX base operating system (BOS) components are prerequisites for PowerHA:

- bos.adt.lib
- bos.adt.libm
- bos.adt.syscalls
- bos.ahafs
- bos.cluster (CAA)
- bos.clvm.enh
- bos.data
- bos.net.tcp.client
- bos.net.tcp.server
- bos.rte.SRC
- bos.rte.libc
- bos.rte.libcfg
- bos.rte.libcur
- bos.rte.libpthreads
- bos.rte.lvm
- bos.rte.odm
- devices.common.IBM.storfwork.rte (optional, but required for sancomm)
- rsct.basic.rte
- rsct.compat.basic.hacmp
- rsct.compat.clients.hacmp
- rsct.core.rmc

Cluster Aware AIX

The Cluster Aware function is part of the AIX operating system. PowerHA SystemMirror 7.1 and later uses CAA services to configure, verify, and monitor the cluster topology. This is a major reliability improvement because core functions of the cluster services, such as topology related services, now run in the kernel space. This makes it much less susceptible to be affected by the workload generated in the user space.

Repository disk

CAA uses a shared disk, between 512 MB and 460 GB in size, to store its cluster configuration information. CAA requires a dedicated shared disk that is available to all nodes that are part of the cluster. This disk cannot be used for application storage or any other purpose.

It is recommended that the repository disk in a two-node cluster is at least 1 GB and be RAID protected.

> **Important:** The repository is *not* supported for mirroring by LVM.

Virtualization layer

This section describes the virtualization layer characteristics and considerations.

Important considerations for VIO Server

This section lists some new features of AIX and VIO Server that help to increase overall availability, and particularly apply to PowerHA environments.

Using poll_uplink

To use the **poll_uplink** option, the following versions and settings are required:

- VIOS 2.2.3.4 or later installed in all related VIOS.
- The LPAR must be at AIX 7.1 TL3, or AIX 6.1 TL9 or later.
- The option **poll_uplink** must be set on the LPAR on the virtual entX interfaces.

The option **poll_uplink** can be defined directly on the virtual interface if you are using shared Ethernet adapter (SEA) fallover or the Etherchannel device that points to the virtual interfaces. To enable **poll_uplink**, use the following command:

```
chdev -l entX -a poll_uplink=yes –P
```

Important: The LPAR must be restarted to activate **poll_uplink**.

Figure 3-1 shows an overview of how the option works. In a typical production environments with two physical interfaces on the VIOS in a dual-VIOS setup. In this environment, the virtual link is reported as down only when all physical connections on the VIOS for this SEA are down.

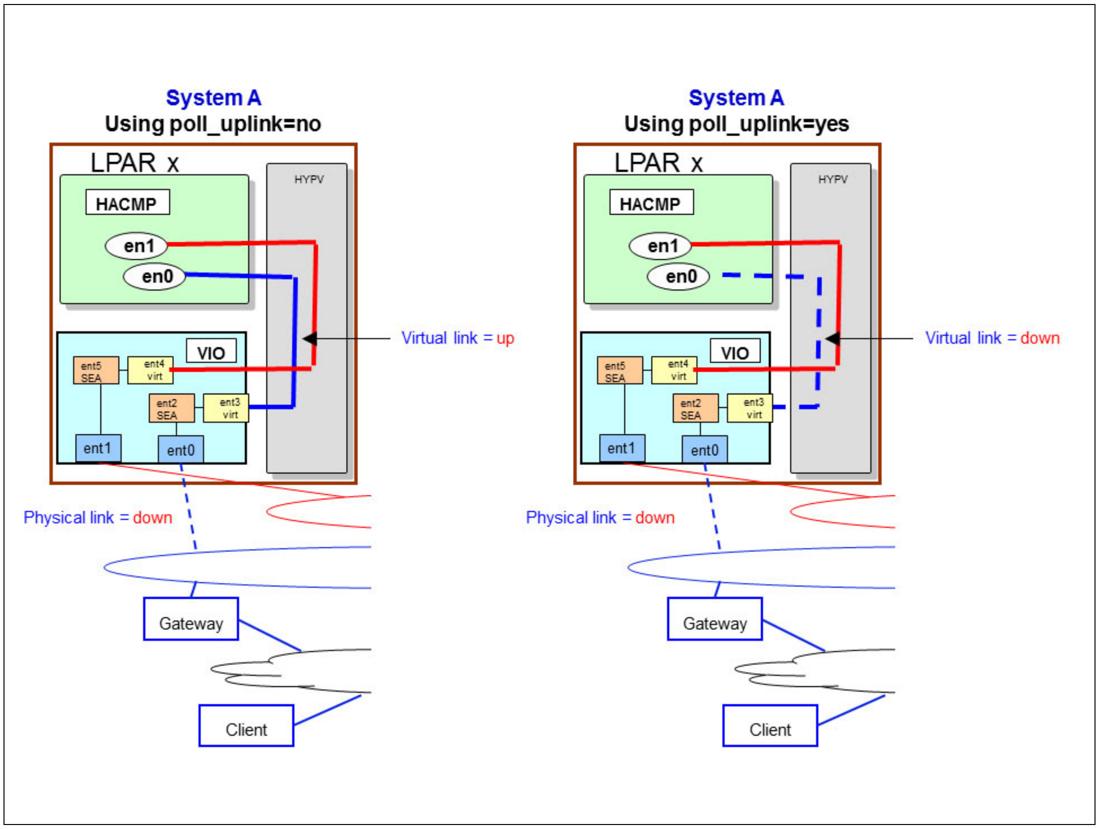


Figure 3-1 Using poll_uplink

Advantages for PowerHA when poll_uplink is used

In PowerHA V7, the network down detection is performed by CAA. CAA by default checks for IP traffic and for the link status of an interface. Therefore, using **poll_uplink** is advised for PowerHA LPARs, which helps the system to make a better decision when a given interface is up or down. The network down failure detection is much faster if **poll_uplink** is used and the link is marked as down.

PowerHA requirements for cross-site LVM

The following requirements are in addition to a typical PowerHA local cluster. They are necessary to assure data integrity and appropriate PowerHA reaction in case of site or disk subsystem failure:

- A server and storage unit at each of two sites.
- SAN and LAN connectivity across/between sites.
  - Redundant infrastructure both within and across sites also recommended.
- PowerHA Standard Edition (allows stretched clusters and support site creation).
- Configure a two site stretched cluster.
- The *force varyon* attribute for the resource group must be set to true.
- The logical volumes allocation policy must be set to *superstrict* (ensuring that LV copies are allocated on different volumes).
- The LV mirroring copies must be allocated on separate volumes that reside on different disk subsystem (on different sites).

Similar to a local cluster, a stretched cross-site LVM mirrored cluster consists of only a single repository disk. So a decision has to be made as to which site should the repository disk reside. An argument can be made to having it a either site. If it resides at the primary site, and the primary site goes down, a failover can and should still succeed. However it is also strongly recommended to define a backup repository disk to the cluster at the opposite site from the primary repository disk. In the event of primary site failure the repository disk will be taken over with the backup repository disk by way of the automatic repository replacement feature within PowerHA.

Though technically not a requirement, it is also strongly recommended to use the AIX LVM capability of mirror pools. Using mirror pools correctly helps to both create and maintain copies across separate storage subsystems ensuring a separate and complete copy of all data at each site.

Mirror pools

Mirror pools make it possible to divide the physical volumes of a volume group into separate pools. A mirror pool is made up of one or more physical volumes. Each physical volume can only belong to one mirror pool at a time. When creating a logical volume, each copy of the logical volume being created can be assigned to a mirror pool. Logical volume copies that are assigned to a mirror pool will only allocate partitions from the physical volumes in that mirror pool. This provides the ability to restrict the disks that a logical volume copy can use. Without mirror pools, the only way to restrict which physical volume is used for allocation when creating or extending a logical volume is to use a map file.

3.1.3 Configuration scenarios

This section provides a few sample scenarios.

Two sites, one server per site

The first scenario shown in Figure 3-2 on page 82 is a two-site, single server and storage unit at *each* site with fully redundant infrastructure both within and across the sites. It is actually a disaster recovery style configuration used more like a high availability one. In this scenario each server is providing some productive working service. They are *not* accessing the same data concurrently. However each server has access to both copies of its own data under normal circumstance. This type of configuration is referred to as *mutual takeover*.

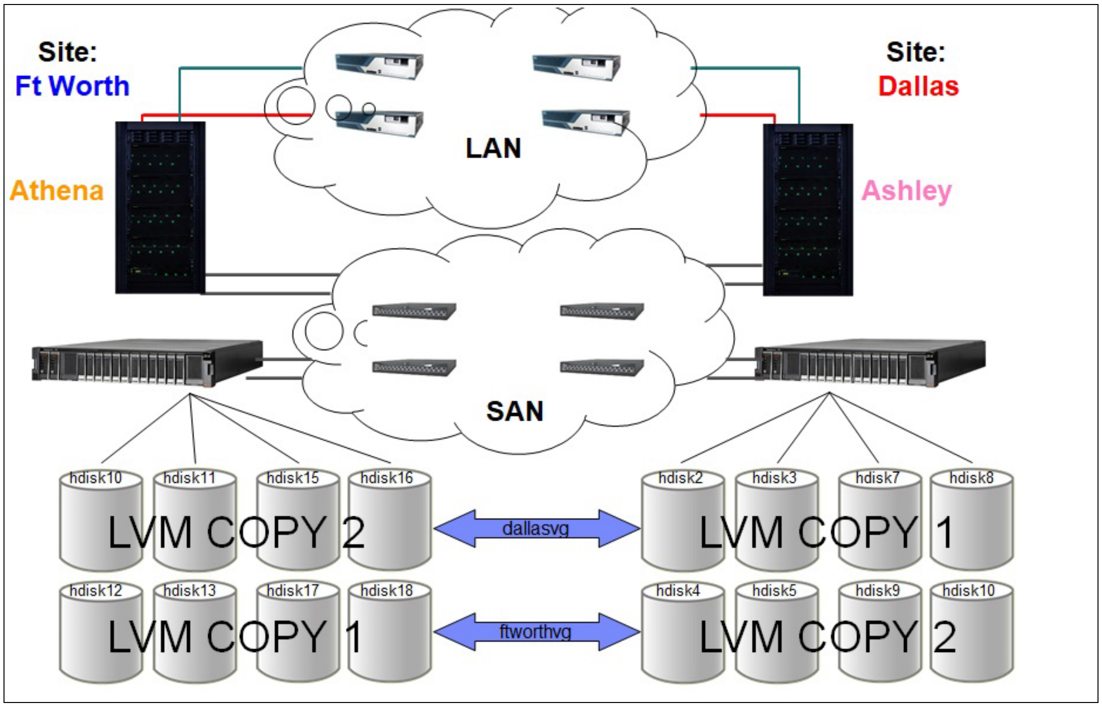


Figure 3-2 Cross-site LVM mirroring mutual takeover scenario

3.1.4 Failure scenario expectations

The following is the expected results based on each specific failure type:

- Application outage.

  PowerHA provides the capability to monitor application(s). In the event of an application failure it can either be restarted locally a specified number of times, or failed over to the next node at the remote if/as desired.

- Storage loss.

  In the event of lost storage access, as shown in Figure 3-3 on page 83, at either site normal operations should continue as access to one complete copy is still available. The disks will most likely go in the *missing* state and numerous reports of partitions going *stale* in the AIX error report.

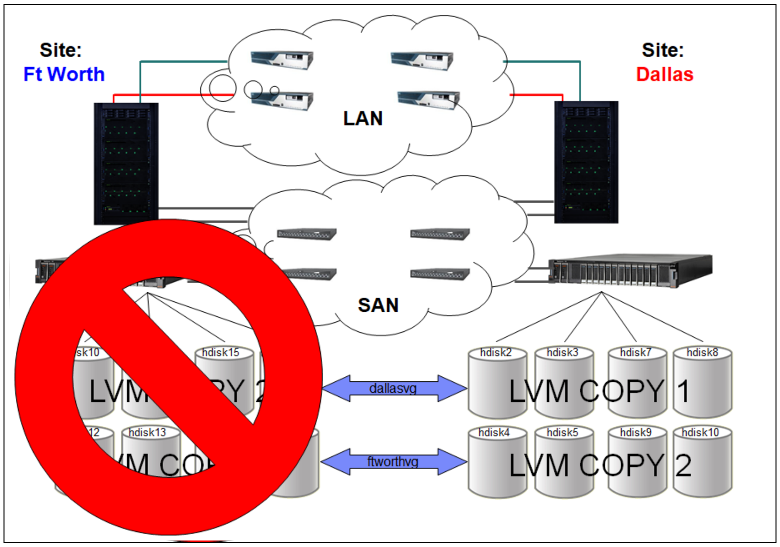


Figure 3-3 Storage loss in cross-site LVM mirroring

- Server/LPAR.

  If a server or LPAR fails within a site, let us say in Ft. Worth as shown in Figure 3-4, then a failover occurs over to its corresponding system in Dallas. The failover system in Dallas takes over, activates and have access to both copies. Both copies continue to be updated normally.

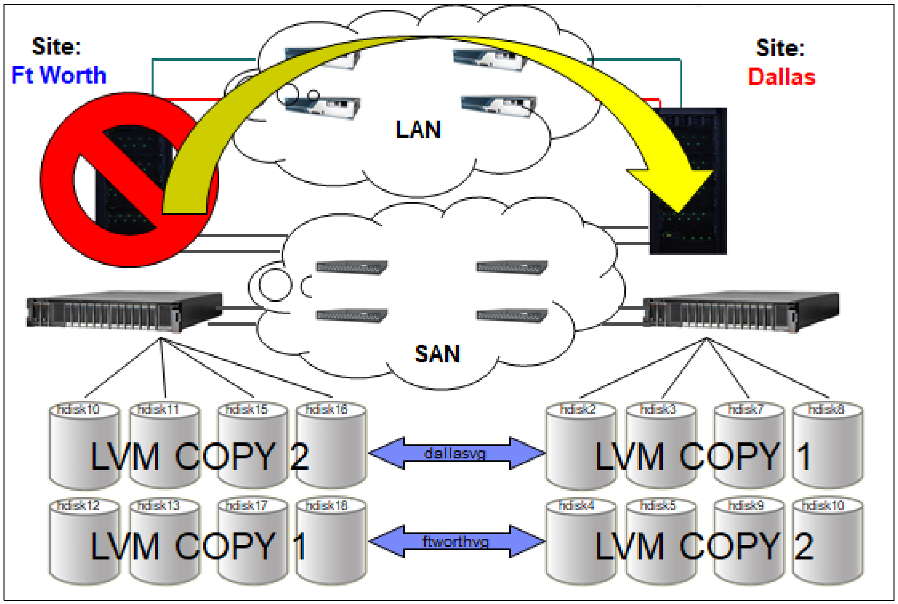


Figure 3-4 Server loss in cross-site LVM mirroring

- Site outage.

  In the event of a site outage, meaning both storage and server as shown in Figure 3-5 on page 84, then a failover occurs to the other site. The key difference is that now only one copy of the data is available, and this where using the *force varyon* attribute for the resource group is needed. This allows the failover server to start the volume group with only half the disks and one copy of data.

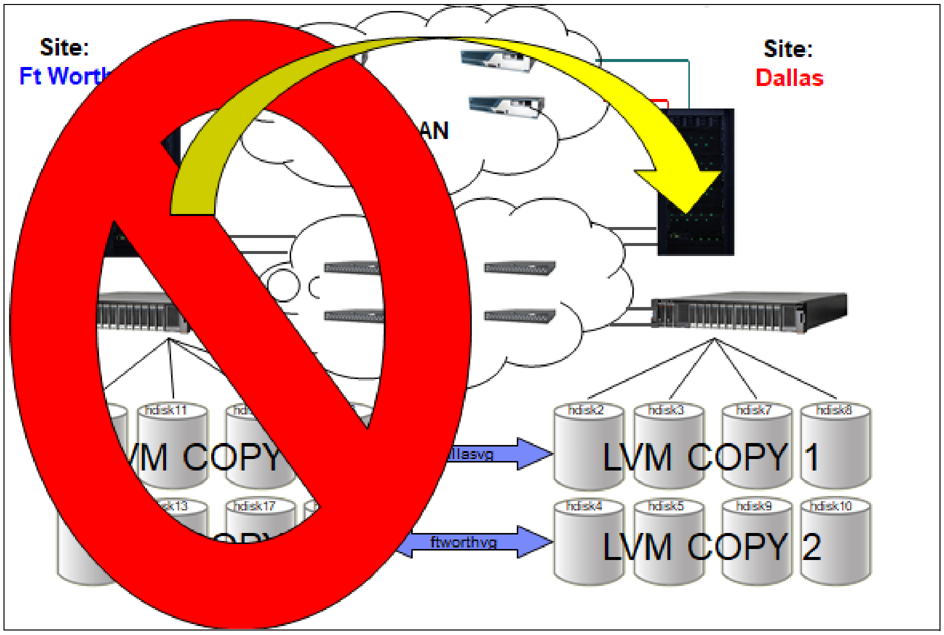


Figure 3-5 Site loss in cross-site LVM mirroring

Two sites, three servers, two data copies

This scenario is a continuation of the previous one. It adds another server locally to Ft. Worth to provide failover within the site first in the event of a server outage as shown in Figure 3-6. All previous failure scenarios and expectations remain the same.

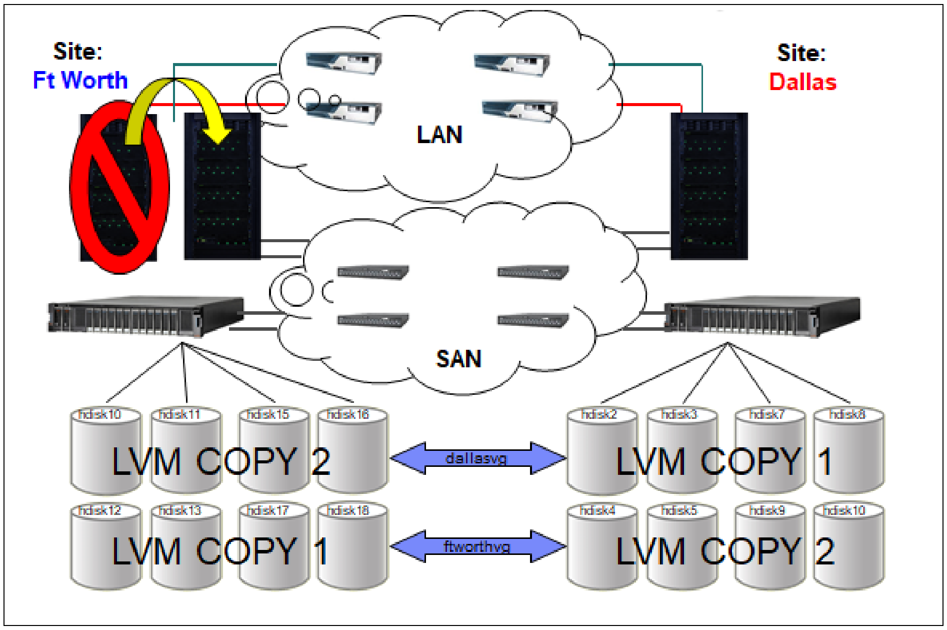


Figure 3-6 Three-node cross site LVM mirroring

3.2 Standalone GLVM

In this scenario we will explore the planning, implementation and monitoring an environment where stand alone GLVM is used to replicate data from one data center to another.

GLVM consists of the following components:

Remote Physical Volume (RPV)
: This is the pseudo local representation of the remote physical volume that allows the LVM to consider the physical volume at the remote site as another local, albeit slow, physical volume. The actual I/O operations are performed at the remote site.
The Remote Physical Volume consists of the RPV Client and the RPV Server - one for each remote physical volume.

The RPV client
: The RPV client is a pseudo device driver that runs on the active server/site, i.e. where the volume group has been activated. There is one RPV Client for each physical volume on the remote server/site and is named hdisk#. The LVM sees it as a disk and performs the I/Os against this device.
The RPV Client definition includes the remote server address and timeout values.

The RPV server
: The RPV server is an instance of the kernel extension of the RPV device driver that runs on the node on the remote server/site, that is, on the node which has the actual physical volume.The RPV Server receives and handles the I/O requests from the RPV client.
There is one RPV Server for each replicated physical volume and is named rpvserver#.

The GLVM Cache
: This is a special type of logical volume of type aio_cache that is designed for use in asynchronous mode GLVM.For asynchronous mode, rather than waiting for the write to be performed on the remote physical volume, the write is recorded on the local cache, and then acknowledgement is returned to the application.At some later point in time, the I/Os recorded in the cache are played in order against the remote disk(s) and then deleted from the cache after successful acknowledged.

Geographic Mirrored Volume Group
: This is an AIX Volume Group that contains both local physical volumes and RPV Clients.

See Figure 3-7 on page 86 for a diagram of the components. You can mirror your data across two sites by configuring volume groups that contain both local physical disks and RPVs. With an RPV device driver, the LVM does not distinguish between local and remote physical volumes - it maintains mirror copies of the data across attached disks. The LVM is, for the most part, unaware that some disks are located at a remote site. Refer to figure Figure 3-7 on page 86.

For PowerHA SystemMirror installations, the GMVGs can be added to resource groups and they will then be managed and monitored by PowerHA.

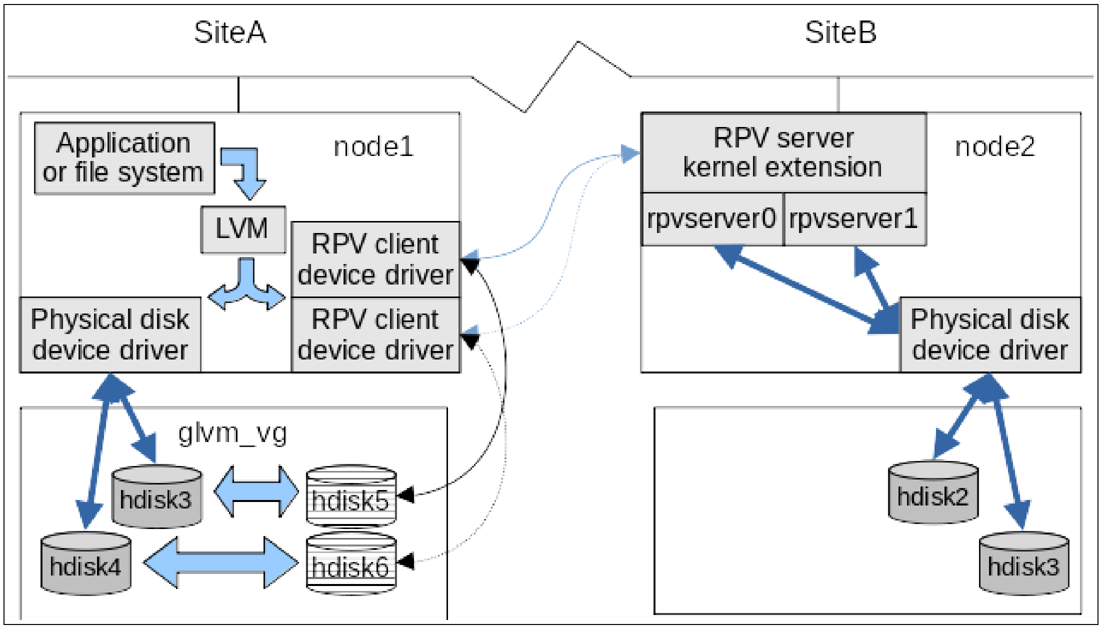


Figure 3-7 Example RPV Server and Client configuration replicating from SiteA to SiteB

3.2.1 Planning

Prior to implementing GLVM between two data centers, the following need to be addressed:

- AIX software requirements.
- Limitations.
- Sizing.
- Recommendations.

AIX software requirements

AIX with:

- glvm.rpv.client.
- glvm.rpv.server.
- glvm.rpv.util.

Limitations

GLVM imposes the following limitations:

- The inter-disk allocation policy for logical volumes in AIX must be set to *superstrict*. This policy ensures that there is a complete mirror copy on each set of either local or remote physical volumes. In GLVM, using the super strict policy for mirroring ensures that when you create a mirrored logical volume, you will always have a complete copy at each site.
- Up to three copies of the logical volumes can be created, with at least one mirror copy at each site. One of the sites may optionally contain a second copy. There will be extra considerations when moving back to the site with the two copies, as the write to each copy is sent separately over the network.
- Two sites, one local and one remote site. If using PowerHA, GLVM site names must correspond with the PowerHA site names.
- The `rootvg` volume group cannot be geographically mirrored.

- While asynchronous mode requires configuring Mirror Pools, it is recommended for synchronous mode.
- The asynchronous GLVM volume group cannot contain active paging space logical volume and it is not recommended for synchronous GLVM.
- You must use scalable volume groups. They can be non-concurrent or enhanced concurrent mode. The use of enhanced concurrent volume groups is required for use with PowerHA but doesn't really provide any advantage for stand-alone GLVM. If you do use enhanced concurrent in stand-alone, there will be extra steps to active the GMVG.
- The volume group should not be configured to auto active (varyon).
- Bad block relocation should be turned off. If a bad block is detected at one site and the block is relocated, then the block maps will differ between sites. This is only required for asynchronous replication as it will impact the playing of the cached I/O against the remote physical volume(s) if the block maps differ.
- IP Security (IPsec) can be configured to secure the RPV client-server network traffic between the sites.
- 1 MB of available space is required in /usr prior to installation.
- Port 6192 TCP/UDP is open between the two servers.

AIX LVM mirror pools

Mirror pools are just a way to divide the physical volumes in a volume group into distinct groups or "Pools" and then control the placement of the logical partition's mirrored copies. They were introduced in AIX 6.1.1.0 and only apply to scalable volume groups. Mirror pool names must be less than 15 characters and are unique within a volume group.

A mirror pool consists of one or more physical volumes and each physical volume can only belong to one mirror pool at a time. When defining a logical volume, each copy of the logical volume can be assigned to a specific mirror pool. This ensures that when a copy of a logical volume is assigned to a mirror pool, only partitions from physical volumes in that pool will be allocated. Prior to the introduction of mirror pools, the only way one can extend logical volumes and guarantee that partitions were allocated from the correct physical volume was to use a map file. Physical volumes can be assigned to a mirror pool with `chpv` or `extendvg`.

There cannot be more than three mirror pools in each volume group and each mirror pool must contain at least one complete copy of each logical volume that is defined in that pool.

> **Note:** After mirror pools have been defined, the volume group can no longer be imported into versions of AIX prior to AIX 6.1.1.0. If using enhanced concurrent mode volume groups, all nodes in the cluster must also be greater than AIX 6.1.1.0.

Mirror pool strictness can be used to enforce tighter restrictions on the allocation of partitions in mirror pools. Mirror pool strictness can have one of the following values:

off
This is the default setting and no restrictions apply to the use of the mirror pools.

on
Each logical volume created in the volume group must have all copies assigned to mirror pools.

super
This is specifically for GLVM and ensures that local and remote physical volumes cannot be assigned to the same mirror pool.

Mirror pool characteristics can be changed, however, any changes will not affect currently allocated partitions. This it is recommended to use the **`reorgvg`** command after any mirror pool changes so allocated partitions can be moved to conform to the mirror pool restrictions.

> **Note:** AIX LVM Mirror pools are only recommended for synchronous mode, but are required for asynchronous mode.

This mirror pools are used to ensure that:

- Each site has complete copy of each mirrored logical volume in the GMVG.
- The cache logical volume for asynchronous GMVGs are configured and managed correctly.

Sizing

There are a number of tools that are useful to examine the workload if GLVM is being planned for an existing application. While network latency and throughput is critical to the performance on synchronous GLVM, it is also important in planning an asynchronous configuration.

- **`gmdsizing`**

  Is a command to estimate network bandwidth requirements for GLVM networks. It was originally part of HAGeo / GeoRM and is part of the samples in PowerHA installations (find in /usr/es/sbin/cluster/samples/gmdsizing/gmdsizing). It monitors disk utilization over the specified period and produces a report to be used as an aid for determining bandwidth requirements.

- **`lvmstat`**

  Reports input/output statistics for logical partitions, logical volumes and volume groups. Also reports pbuf and blocked I/O statistics and allows pbuf allocation changes to volume groups.

  ```
  lvmstat { -l | -v } Name [ -e | -d ] [ -F ] [ -C ] [ -c Count ] [ -s ] [ Interval
  [ Iterations ] ]
  ```

- **`iostat`**

  Reports CPU statistics, asynchronous input/output (AIO) and input/output statistics for the entire system, adapters, TTY devices, disks CD-ROMs, tapes and file systems. Use flags **`-s -f`** to show logical and disk I/O.

- **`Other tools`**

  General monitoring commands such as **`nmon`** and **`topas`** can be used. For users that want a more graphical representation, statistics from rpvstat can be loaded into a time series database, such as influxDB and then presented with a tool such as Grafana.

Cache planning (asynchronous mode)

Asynchronous mode uses a local cache (logical volume in the mirror pool) to store locally the updates to the remote logical volumes. The size of this cache is critical in two ways:

- Too big
  The cache represents the maximum amount of data that can be lost in a disaster, so must be planned. Roughly 2 GB of data in the cache represents 1 GB of updates for the remote system.
- Too small
  After the local cache is full, GLVM will suspend all local writes until space is cleared in the cache (updates made to the remote copy). If there are sustained peaks in I/O activity

greater than that which the network can handle, the cache will fill faster than it can empty, eventually stopping local I/O until space can be cleared.

Planning for ongoing operations

As PowerHA SystemMirror will not be monitoring and managing the starting and stopping of the RPV Servers and Clients, this will be the task of the administrator. While many of these tasks can be scripted, carefully checking must be done of the status of the environment before starting or stopping any services as GLVM has no awareness of the environment on which it is operating.

It will be the task of the administrator to maintain the operations of the RPV Servers and Clients, monitor the health of the network(s) and the servers/LPARs, and sets preferred read for the local disks.

Planning quorum

In general, it is recommended to disable quorum for geographically mirrored volume groups in order to minimise the possibility of the volume group going offline should access to the remote copy be lost. Thus you will be able to keep operating in the event of an inter-site network failure or maintenance activity on the remote site.

> **Note:** If using PowerHA SystemMirror, it is a different discussion, since PowerHA detects quorum loss and manages the volume group.
>
> Disabling quorum will also require setting forced varyon for the volume group in PowerHA.

Planning for increase in CPU load

Implementing GLVM will increase the demand on system resources, so the following needs to be taking into account when planning:

- If compression is turned on, ensure that hardware compression is enabled (NX Crypto Accelerator) or this will add to CPU consumption on both the RPV Client and Server.
- Changes to **`io_grp_latency`** can increase CPU consumption.
- I/O wait will increase particularly for synchronous mode with the delay introduced by acknowledgement from the RPV Server. There is a longer I/O code path, a delay will be introduced for asynchronous mode.

Tuning options

The **`rpvutil`** command has the following options for tuning the operation of GLVM:

- **`rpv_net_monitor=1|0`**

  Setting rpv_net_monitor to 1, will turn on monitoring of the RPV network by rpvutil, so that the RPV client will detect any network failures and attempt to resume after the network recovers. The default is 0 (disabled).

- **`compression=1|0`**

  Before using compression, check that:

  - Both the RPV client and the RPV server are running AIX version 7.2.5, or later, with all the latest RPV device drivers.
  - Both the RPV server and the RPV client are IBM Power Systems servers with NX842 acceleration units.

  - The compression tunable parameter is enabled on both the RPV server and RPV client so that the I/O data packets are compressed when the workload is failed over between the RPV client and the RPV server.

  When the compression tunable parameter is set to 1, the **`rpvutil`** command compresses the I/O data packet before it is sent from the RPV client to the RPV server by using the cryptography and compression units (NX842) on IBM Power Servers. If the I/O data packet is compressed successfully, a flag is set in the data packet. When the RPV Server receives a packet with the compressed flag set, the packet will be decompressed. If the NX842 compression unit is not available, the RPV Server will attempt software decompression of the packet.

  By default this option is set to 0 (disabled).

- **`io_grp_latency=timeout_value`** (milliseconds)

  Used to set the maximum expected delay before receiving the I/O acknowledgement for a mirror pool that is configured in asynchronous mode. The default delay value is 10 ms and a lower value can be set to improve I/O performance, but may be at the cost of higher CPU consumption.

- **`nw_sessions=<number of sessions> (1 to 99)`**

  This is a new tunable (available in AIX 7.2.5.2) that controls the number of RPV sessions (sender and receiver threads) to be configured per network. This is used to increase the number of parallel RPV sessions per GLVM network, which will send more data in parallel, improve data transfer rate and more fully utilise the network bandwidth.

Setting hardware compression

Check that hardware compression is possible as shown in Example 3-1.

Example 3-1 Check hardware compression capabilities

```
pcha1:/:# prtconf
. . .
NX Crypto Acceleration: Capable and Enabled
. . .
```

General recommendations

The following are some general recommendation from the team experiences:

- Issues have been found with potential deadlocks if Mirror Write Consistency is set to **`active`** for asynchronous GMVG. Setting of **`passive`** is recommended for both asynchronous and synchronous modes.
- Configure RPV level I/O timeout value to avoid any issues related to network speed or I/O timeouts. This value can be modified, when RPV disk is in defined state. Default value is 180 seconds.
- AIX LVM allows the placement of disks in mirror pools, and then selecting read preference based on the mirror pool. A feature that was added for GLVM in PowerHA, is for physical volumes to be added to sites, and then the preferred read to be set to **`siteaffinity`**. This option is not available for stand-alone GLVM users, instead they will need to set the LVM preferred read to the local mirror pool before activating the volume group.
- Turn of quorum and have multiple networks in PowerHA or Etherchannel in the stand-alone. Ensure that all networks follow different paths and have no shared point of failure.
- **`rpvstat -n`** will give details with individual network and **`rpvstat -A`** will give details about asynchronous I/O.

- For better performance ensure that disk driver parameters are configured correctly for the storage deployed in your environment. Refer to AIX and storage documentation for setting those tunables (for example, `queue_depth`, `num_cmd_elems`, and so on).

Recommendations - asynchronous mode:

- Asynchronous GLVM is ONLY supported on scalable volume group(s). These may be in enhanced concurrent mode.
- You can lower the timeout parameter for the RPV client to improve application response times, but balance this against latency problems. This value can be changed when the RPV client is in a defined state.
- Reducing the `max_transfer` size for the remote device while there is data in the AIO cache can cause remote IO failures. (`lsattr -El hdiskX -a max_transfer`).
- In a stand-alone GLVM environment, you must ensure that all the backup disks in the secondary sites are in an active state before you bring the volume group online. During the online recovery of the volume group, if the RPV device driver detects that the RPV server is not online it updates the cache disk detailing a failed request and all subsequent I/Os will be treated as synchronous. To convert back to asynchronous mode after the problem is rectified, one must first convert the mirror pool to synchronous mode and then back to asynchronous mode using `chmp`.
- When an asynchronous GMVG it brought online, it will perform a cache recovery. If previously the node halted abruptly, say with a power outage, it is possible that the cache is not empty. In this case, cache recovery may take some time, depending upon amount of data in the cache and the network speed. No application writes are allowed to complete while cache recovery in progress to handle consistency at remote site. In this case, the application users may observe a pause.
- After a site failure, asynchronous mirror state on remote site will be inactive. After integrating back with primary site, mirror pool needs to be converted to synchronous and then back to asynchronous so as to continue in asynchronous mode.
- Monitor regularly whether the asynchronous mirroring state of the GLVM is active by using the `lsmp` command.
- `rpvstat -C` will give the details about IO cache Monitor and rpvstat will give details such as number of times the cache is full.
- For better performance, ensure that the disk driver parameters of the storage device that is deployed in your environment is configured correctly.

3.2.2 AIX modifications that support GLVM

The following changes have been made in AIX either for GLVM, or GLVM has taken advantage of:

Mirror pools
: Mirror pools were introduced in AIX 6.1 to ensure that one copy of a mirror is only placed on one group of physical volumes. One or more physical volumes can be defined as a member of a mirror pool so when logical volumes are created, each copy can be set to belong to a given mirror pool. GLVM uses mirror pools to guarantee that at a minimum there is a complete copy of one logical volume mirror at each site as each logical volume copy can only reside on one set of disks, that being the mirror pool at that site.

`varyonvg` command
: The `varyonvg` command has been modified to allow for instance where failures have led to the situation where the data at each site is not the same. You can now control the activation of a volume group specifying

where to use the local or remote copy data, which may be potentially stale.

varyoffvg command The **varyoffvg** command has been modified to ensure that all the outstanding I/Os in the aio_cache are drained before the command completes. This can have a considerable performance impact on the varyoffvg command if there are a large number of outstanding updates for the remote physical volumes.

3.3 PowerHA for AIX Enterprise Edition with GLVM

This scenario combines the automated failover of PowerHA and the AIX IP based replication of GLVM. It is mostly a combination of the previous two sections. The expected behavior during failures is nearly identical to that of 3.1, “PowerHA for AIX Cross-Site LVM mirroring” on page 78. It also is storage type agnostic since it does not utilize any storage specific data replication. This assumes synchronous based replication which can provide an RPO of zero, and RTO of mere minutes. The biggest determining factor in recovery time is application recovery and restart time.

3.3.1 Requirements

The following requirements are in addition to GLVM 3.2.1, “Planning” on page 86:

- PowerHA SystemMirror for AIX Enterprise Edition.

3.3.2 Configuration scenario

Two sites, three servers, and three data copies are shown in Figure 3-8 on page 93.

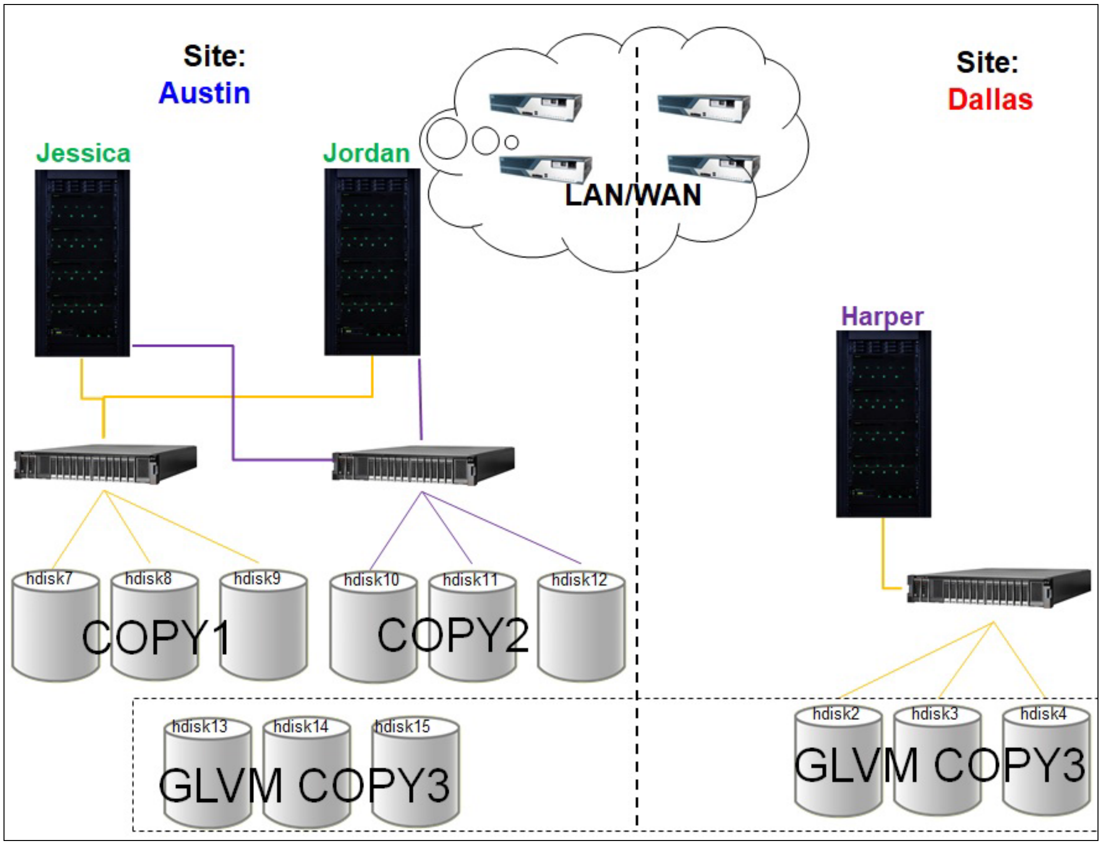


Figure 3-8 PowerHA Enterprise Edition with GLVM

3.3.3 Failure scenario expectations

The following is the expected results based on each specific failure type:

- Application outage.

  PowerHA provides the capability to monitor application(s). In the event of an application failure it can either be restarted locally a specified number of times, or failed over to the next node locally, and then remote if/as desired.

- Storage loss.

  In the event of lost storage access to any storage unit at either site normal operations should continue as access to at least one complete copy is still available. Actually in this exact scenario any *two* of the three storage units can be lost and operations still continue. The disks will most likely go in the *missing* state and numerous reports of partitions going *stale* in the AIX error report.

- Server/LPAR.

  If a server or LPAR fails, say *Jessica* within the primary site in *Austin* then a failover occurs over to next system, *Jordan* within the site. Jordan takeovers and activates all copies. All copies continue to be updated normally.

- Site outage.

In the event of a site outage, meaning all storage and servers are unavailable in Austin, then a failover occurs to the *Harper* server in *Dallas*. The key difference is that now only one copy of the data is available, and this where using the *force varyon* attribute for the resource group is needed. This allows the failover server to start the volume group with only one copy of data.

3.4 PowerHA for AIX Enterprise Edition with HyperSwap

The HyperSwap function in PowerHA SystemMirror for AIX Enterprise Edition 7.1.2, or later, provides for continuous availability against storage errors. HyperSwap is based on storage-based synchronous replication. HyperSwap technology enables the host to transparently switch an applications I/O operation to the auxiliary volumes, provided physical connectivity exists between the host and the auxiliary storage subsystem.

The HyperSwap function in PowerHA SystemMirror supports the following capabilities within your environment:

- Eliminates primary disk subsystems as the single point of failure.
- Provides maintenance for storage devices without any application downtime.
- Provides migration from an old storage device to a new storage system.

3.4.1 HyperSwap for PowerHA SystemMirror concepts

The HyperSwap function in PowerHA SystemMirror for AIX Enterprise Edition 7.1.2, or later, enhances application availability for storage errors by using IBM DS8000 metro mirroring. If you use the HyperSwap function in your environment, your applications stay online even if errors occur on the primary storage because application I/O to an auxiliary storage system.

The HyperSwap function uses a model of communication, which is called in-band, that sends the control commands to a storage system through the same communication channel as the I/O for the disk. The HyperSwap function supports the following types of configurations:

- Traditional Metro Mirror Peer-to-Peer Remote Copy (PPRC).

  The primary volume group is only visible in the primary site and the auxiliary volume group is only visible in the auxiliary site.

- HyperSwap.

  The primary and auxiliary volume group are visible from the same node in the cluster.

You typically configure the HyperSwap function to be used in the following environments:

- Single node environment.

  A single compute node is connected to two storage systems that are in two sites. This HyperSwap configuration is ideal to protect your environment against simple storage failures in your environment.

- Multiple site environment.

  A cluster has multiple nodes that are spread across two sites. This HyperSwap configuration provides high availability and disaster recovery for your environment.

Mirror groups in HyperSwap for PowerHA SystemMirror represent a container of disks and have the following characteristics:

- Mirror group contain information about the disk pairs across the site. This information is used to configure mirroring between the sites.
- Mirror groups can contact a set of logical volume manager (LVM) volume groups and a set of raw disks that are not managed by the AIX operating system.
- All the disks devices that are associated with the LVM volume groups and raw disks that are part of a mirror group are configured for consistency. For example, the IBM DS8800 views a mirror group as one entity regarding consistency management during replication.

- The following types of mirror groups are supported:
  - User mirror group.

    Represents the middleware-related disk devices. The HyperSwap function is prioritized internally by PowerHA SystemMirror and is considered low priority.
  - System mirror group.

    Represents critical set of disks for system operation, such as, rootvg disks and paging space disks. These types of mirror groups are used for mirroring a copy of data that is not used by any other node or site other than the node that host these disks.
  - Repository mirror group.

    Represents the cluster repository disks of that are used by Cluster Aware AIX (CAA).

3.4.2 Requirements

This section shows hardware, software and other requirements.

Hardware

Hardware requirements are as follows:

- POWER7 server or later (original support was POWER5 and later).
- DS8800, or later, with firmware R6.3sp4 (86.xx.xx.x) or higher.

Software

Software requirements are as follows:

- PowerHA 7.1.2 SP3 or higher.
- AIX.
  - Version 6, Release 1, Technology Level 8, Service Pack 2 or higher.
  - Version 7, Release 1, Technology Level 2, Service Pack 2 or higher.

Other considerations

There are a few additional considerations to know about:

- Metro Mirror (in-band) functions, including HyperSwap, are supported in Virtual I/O Server (VIOS) configurations by the N-Port ID Virtualization (NPIV) method of disk management.
- Metro Mirror (in-band) functions, including HyperSwap, are not supported by the virtual SCSI (VSCSI) method of disk management.
- To use Live Partition Mobility (LPM), you must bring the resource group that contains the mirror group into an unmanaged state by using the C-SPOC utility to stop cluster services with the Unmanage Resource Groups option. After you complete the LPM configuration process, you must bring the resource group back online by using SMIT. This process brings all mirror groups and resource groups back online.
- Disk replication relationships must adhere to a one-to-one relationship between the underlying LSS. An LSS that is already part of a mirror group, cannot be part of another mirror group.
- Repository disks require that you specify an alternate disk or a disk that is not configured to use the HyperSwap function when you set HyperSwap property to Disable.
- SCSI reservations are not supported for devices that use the HyperSwap function.

- You must verify and synchronize the cluster when you change the cluster configuration. If you change the mirror group configuration while cluster services are active (DARE), those changes might be interpreted as failures, which result in unwanted cluster events. You must disable the HyperSwap function before you change any settings in an active cluster environment.

For more detailed planning assistance see the planning section found here.

3.4.3 Configuration scenario

Figure 3-9 on page 97 shows a cluster configuration using PowerHA SystemMirror Enterprise Edition for AIX that has the following characteristics:

- Two sites called *Site A* and *Site B*.
- Two nodes for each site for a total of four nodes.
- A concurrent application, like a DB2 application that is active on Node 1 and Node 2.
- Application disks are replicated by using IBM DS8800 metro mirroring.
- All four nodes can access both instances of the application disks that are being replicated.

Planned HyperSwap for PowerHA SystemMirror

A planned HyperSwap occurs when you initiate a HyperSwap from the primary storage subsystem to the auxiliary storage subsystem.

During a planned HyperSwap, I/O activity for an application stops after coordination occurs across the host in the cluster. The application I/O is switched to the auxiliary storage subsystem and the application I/O activity continues to function as normal.

A planned HyperSwap is ideal when you perform maintenance on the primary storage subsystem, or when you migrate from an old storage subsystem to a new storage subsystem. Figure 3-9 on page 97 shows the changes in your environment when a failure occurs and your sites are configured for a planned HyperSwap. The primary storage system on Site A is changed to the auxiliary storage system because the application is running on Node 1 and Node 2 can access the storage system on Site B. Therefore, the application that is running on Site A now stores data on the primary storage system at Site B.

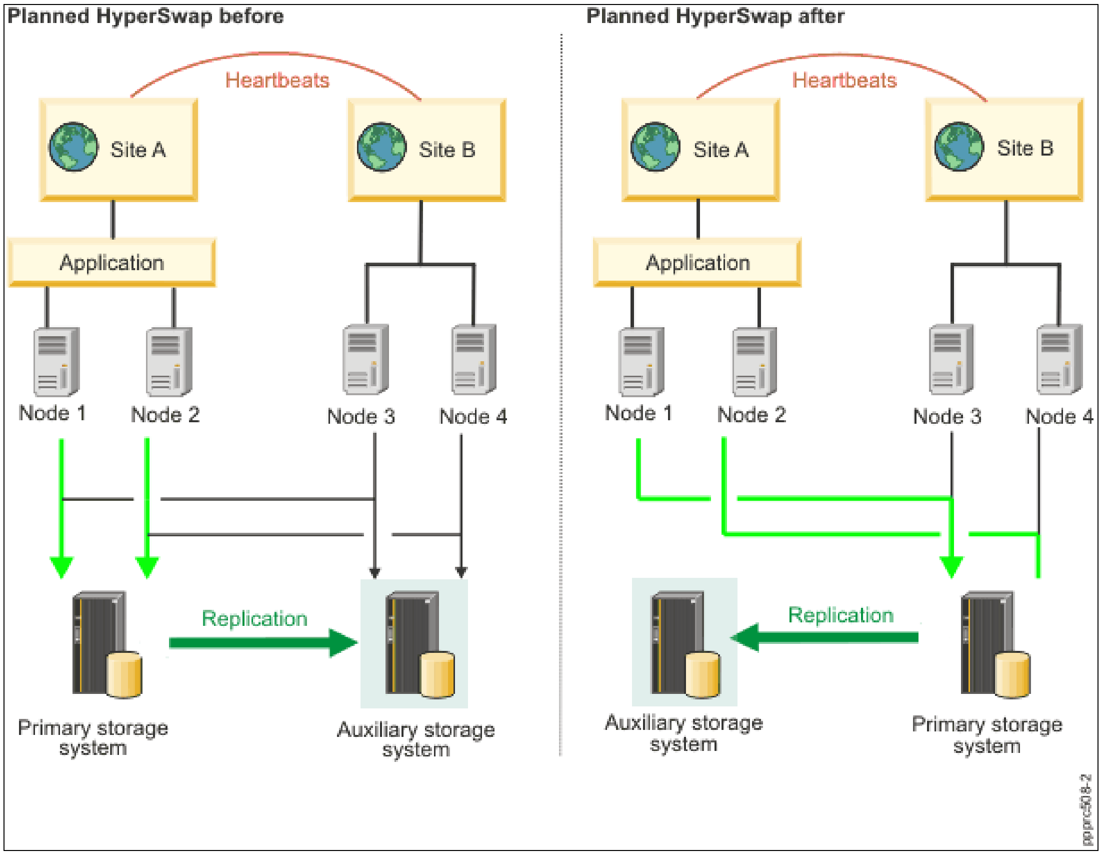


Figure 3-9 Planned Hyperswap

3.4.4 Failure scenario expectations

The following is the expected results based on each specific failure type:

- Application outage.

  PowerHA provides the capability to monitor application(s). In the event of an application failure it can either be restarted locally a specified number times, or failed over to the next node at the remote site.

- Storage loss/Unplanned HyperSwap.

  An unplanned HyperSwap occurs when a primary storage system fails, and the operating system detects and reacts by performing a failover. During the failover, the application I/O on the primary storage system is transparently redirected to a auxiliary storage system and the application I/O continues to run.

  During the HyperSwap process, when the applications are being redirected to an auxiliary storage system, the application I/O is temporarily suspended.

  If an unplanned HyperSwap does not complete successfully, the application I/O fails and a resource group fallover event starts based on the site policy. You cannot define a fallover event in a site policy for concurrent resource groups.

  There are multiple scenarios when an unplanned HyperSwap can occur and as with all PowerHA designs, carefully planning must be done to avoid communication failures leading to a split brain scenario and potential data corruption.

- Server outage.

A local failover with the site occurs just like with any PowerHA cluster with shared disk. The replication is unaffected as it is unaware that a host failover occurred.

- Site outage.

  A site failover occurs and services continue operating at Site B, and the auxiliary storage becomes the primary. The big difference now is that if Site A is unavailable it is unable to replicate back to the Site A. So careful planning is required to restore operations back to the original site when desired.

3.5 VM Recovery Manager High Availability

The VMRM HA solution has a unique benefit over most of the rest of the solutions in these scenarios and that is it can handle all LPAR OS types, and, handle them simultaneously. Since it is a high availability solution it does requires multiple hosts have access to the same set of data volumes. There is no replication in this case. It is mostly automation, not necessarily automatic, of remotely restarting the LPARs/VMs on another server. This means RTO can still be minutes, but since the LPAR/VM does have to boot up on another server it is longer than most clustered solutions like PowerHA.

For full details about planning, installing and configuring VMRM HA see *Implementing IBM VM Recovery Manager for IBM Power Systems*, SG24-8426.

3.5.1 Requirements

The following is a list of key requirements in addition to the license(s) needed to utilize VMRM HA:

- Additional LPAR with at least 1-core CPU and 8 GB memory running AIX 7.2 with Technology Level 2 Service Pack 1 (7200-02-01) or later for the controller system (KSYS).
  - 30 MB of disk space in the /opt directory and 200 MB of disk space in the /var directory.
- HMC(s) must be Version 9 Release 9.1.0 or later.
- Pair of VIOS per host with version 3.1.0.1 or later.
- LPM like requisites of:
  - All LPARs/VMs must have virtualized I/O resources.
  - Same VLAN(s) must be configured across hosts.
  - Storage area network (SAN) connectivity and zoning must be configured so desired target servers can access the host disks as required by way of VIOS.
- When using host groups the KSYS subsystem requires two disks for health cluster management. A disk of at least 10 GB, called the *repository disk*, is required for health monitoring of the hosts, and another disk of at least 10 GB, called the *HA disk*, is required for health data tracking for each host group. All these disks must be accessible to all the VIOSs on each of the hosts on the host group.
- LPARs/VMs must be at least one of the following operating systems:
  - AIX 6.1 or later.
  - Red Hat Enterprise Linux (Little Endian) Version 7.4 or later (kernel version 3.10.0-693).
  - SUSE Linux Enterprise Server (Little Endian) Version 12.3 or later (kernel version 4.4.126-94.22).

- Ubuntu Linux distribution Version 16.04.
- IBM i Version 7.2 or later.

3.5.2 VMRM HA configuration scenario

As is the case with many solutions there is a plethora of configuration options. In this scenario we show a combination of Power Systems and LPARs/VMs with different OS types to re-emphasize the flexibility the solution provides. In this scenario the KSYS controller node is on a separate physical Power Systems as shown in Figure 3-10. Though this is not a requirement it is not unusual, nor does it have to be dedicated to the KSYS. It can be used for other functions, like a NIM server.

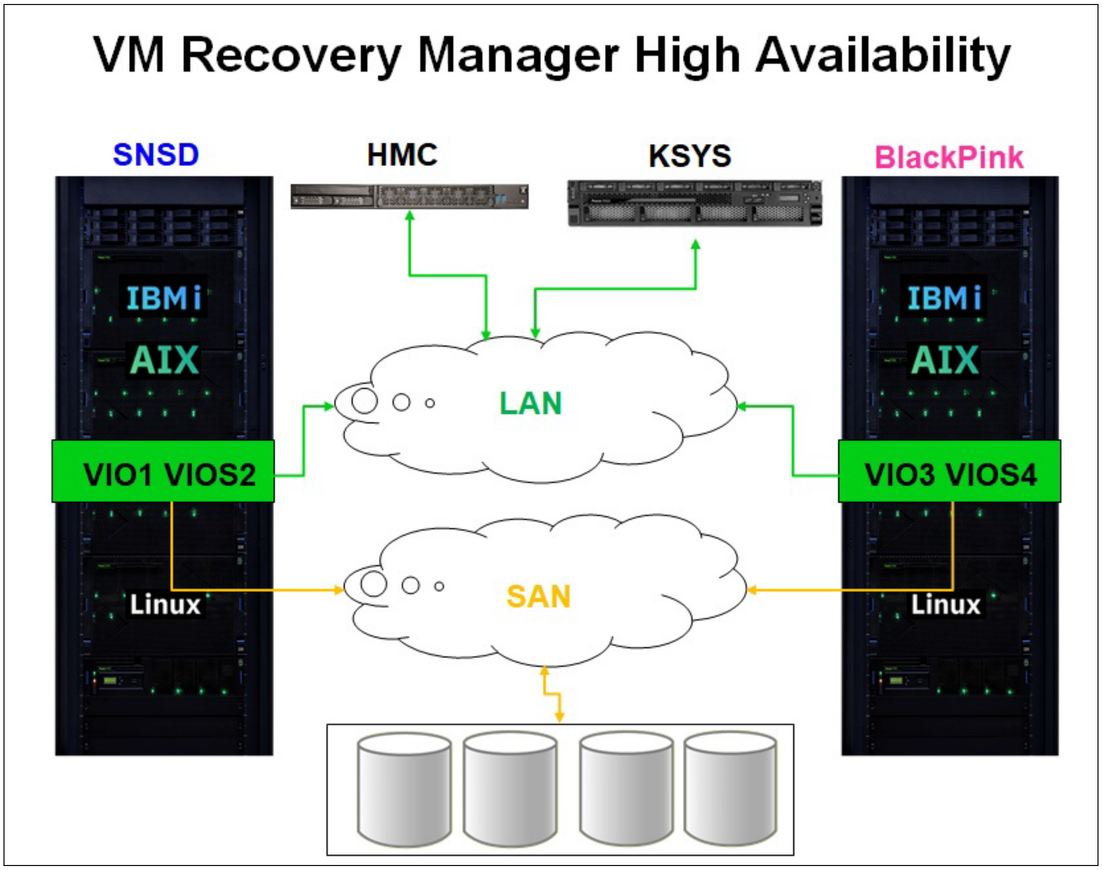


Figure 3-10 VM Recovery Manager High Availability - Single site, two servers, single data copy

3.5.3 Failure scenario expectations

The following is the expected results based on each specific failure type:

- Application outage.
- VMRM HA provides the capability to monitor application(s). In the event of an application failure it can either be restarted locally a specified number times, or failed over to the next node locally.
- Storage loss.

  In the event of lost storage access/storage failure, VMRM HA does *not* provide any additional facilities to help in recovery. Recovery is dependent on fixing the problem if it were just a connectivity problem, or in the event of full storage subsystem failure, recreating storage luns and restoring data if/as needed.

- Server/LPAR.

  If a server or LPAR fails it can be restarted on the other server by way of the KSYS controller node. So it is important that the KSYS controller node be available to perform this action.This can be either automatic or manually initiated.

- Site outage.

  In the event of an entire site outage this solution provides *no* additional recoverability as it is a single site w/o replication.

3.6 VM Recovery Manager Disaster Recovery

The VMRM DR solution shares similar benefits of VMRM HA over many other solutions in these scenarios and that is it can handle all LPAR OS types, and, handle them simultaneously. Since it is a disaster replication solution it is co-dependent on data replication across sites. It is mostly automation, not necessarily automatic, of remotely restarting the LPARs/VMs on another server. The RPO varies based on the replication type but the RTO can still be minutes. But since the LPAR/VM does have to boot up on another server it is longer than most clustered solutions like PowerHA.

For full details about planning, installing and configuring VMRM DR see the following:

- *IBM Geographically Dispersed Resiliency for IBM Power Systems*, SG24-8382.
- I*BM VM Recovery Manager DR version 1.5 base publications.*

3.6.1 Requirements

The following are in addition to the VMRM HA 3.5.1, “Requirements” on page 98.

- An HMC at each site.
- One of the supported storage replications:
  - IBM Spectrum Virtualize:
    - Metro Mirror.
    - Global Mirror.
  - IBM DS8000 Global Mirror:
    - DSCLI 7.7.51.48, and later.
  - IBM XIV and IBM Flashsystem A9000.
  - Dell EMC SRDF (VMAX).
    - SRDF/S (Synchronous).
    - SRDF/A (Asynchronous).
    - SYMCLI installed on KSYS node.
  - Dell EMC Unity Storage System:
    - Asynchronous replication with version 5.0.6.0.6.252, or later.
  - Hitachi Virtual Storage Platform (VSP) G1000 and Hitachi VSP G400:
    - CCI version 01-39-03/04 and model RAID-Manager/AIX.
    - Synchronous data replication.
    - Asynchronous data replication.

3.6.2 VMRM DR configuration scenario

Two sites, three servers, and two data copies are shown in Figure 3-11.

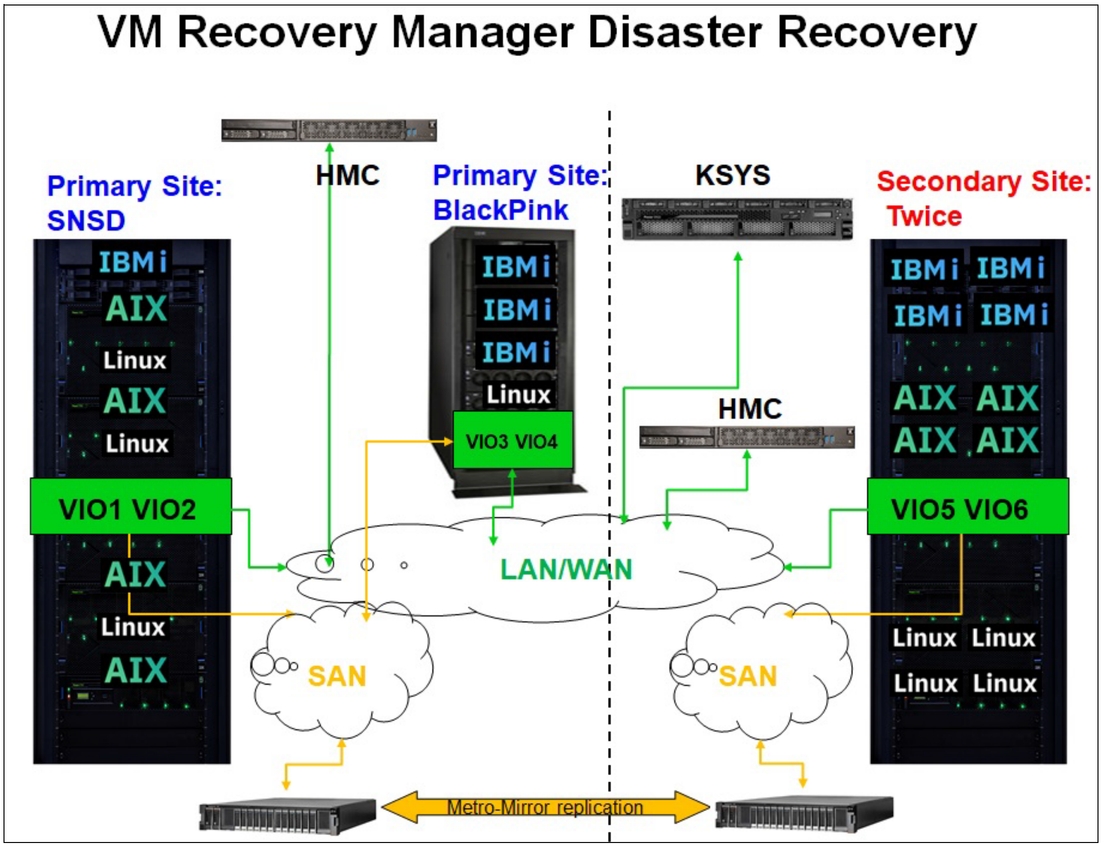


Figure 3-11 VM Recovery Manager Disaster Recovery

3.6.3 Failure scenario expectations

The following is the expected results based on each specific failure type:

- Application outage.

  VMRM DR provides the capability to monitor application(s). In the event of an application failure it can either be restarted locally a certain number of specified times, or failed over to the next node locally, and then remote if/as needed.

- Storage loss.

  In the event of lost storage access/storage failure, the LPAR/VM can be restarted at the secondary/DR site either automatic or manually initiated.

- Server/LPAR.

  It is possible to mix both the HA and DR options of VM Recovery Manager. Both are dependent on an active KSYS controller node. So it is important that the KSYS controller node be available to perform this action. The restart action can be done locally or remote and can be either automatic or manually initiated.

- Site outage.

  If the primary site fails, the LPARs/VMS can be restarted on the third server, *Twice* at the remote secondary site by way of the KSYS controller node. This action can be either automatic or manually initiated.

3.7 Tivoli System Automation for Multiplatform

TSA MP automates IT resources by starting and stopping resources automatically and in the correct sequence. Resources are located on a system, which is referred to as node in the context of a cluster. Resources controlled by System Automation can be applications, services, mounted disks, network addresses or even data replication. Basically anything on a node which can be monitored, started and stopped with help of commands or through an API. For each resource TSA MP provides an availability state and offers a way to start and stop them.

TSA MP allows customization of start and shut down dependencies between resources and resource groups. After the resources are described in an automation policy, the operator can start or shut down an application in an reliable way.

The following section covers a three node TSA MP cluster where a database or application server can failover to dedicated standby node.

3.7.1 Requirements

The following are in addition to the TSA MP code itself. TSA MP also provides a command, **prereqSAM**, to check if all prerequisites are installed.

AIX prerequisites

The following AIX prerequisites must be met:

- A 32-bit version of Java 7, Java 7.1 or Java 8 is required with the following minimum Service Refresh levels:
  - Java 7.0 SR8: AIX package Java7.jre/Java7.sdk 7.0.0.145.
  - Java 7.1 SR2: AIX package Java71.jre/Java71.sdk 7.1.0.25.
  - Java 8.0 SR0: AIX package Java8.jre/Java8.sdk 8.0.0.507.
  - System Automation for Multiplatforms Fixpack Version 4.1.0.7 supports Java 8 SR6 FP30: AIX package Java8.jre/Java8.sdk 8.0.6.30.
- System Automation for Multiplatforms Fixpack Version 4.1.0.7 on AIX, the RSCT 3.2.6.1 will be installed. The following AIX TL levels are only supported with this fixpack:
  - AIX 7.1 TL 5.
  - AIX 7.2 TL 3.
  - AIX 7.2 TL 4.
  - AIX 7.2 TL 5.
- RSCT packages.

Linux prerequisites

The following prerequisites must be met before System Automation for Multiplatforms can be installed on a Linux system:

- RSCT packages.
- The following package is required on each Red Hat Enterprise Linux v7.1 system:
  - perl-Sys-Syslog.
- The following package is required on each Red Hat Enterprise Linux v8 system:
  - perl-Net-Ping.

- The following package is required on each SUSE Linux Enterprise Server (12/15) system:
  - mksh.

For additional details about planning, installing and configuring TSA MP see the base publications here.

3.7.2 TSA MP configuration scenario

The following section covers a three node TSA MP cluster where a database or application server can failover to dedicated standby node within the same site.

One site, three servers are shown in Figure 3-12.

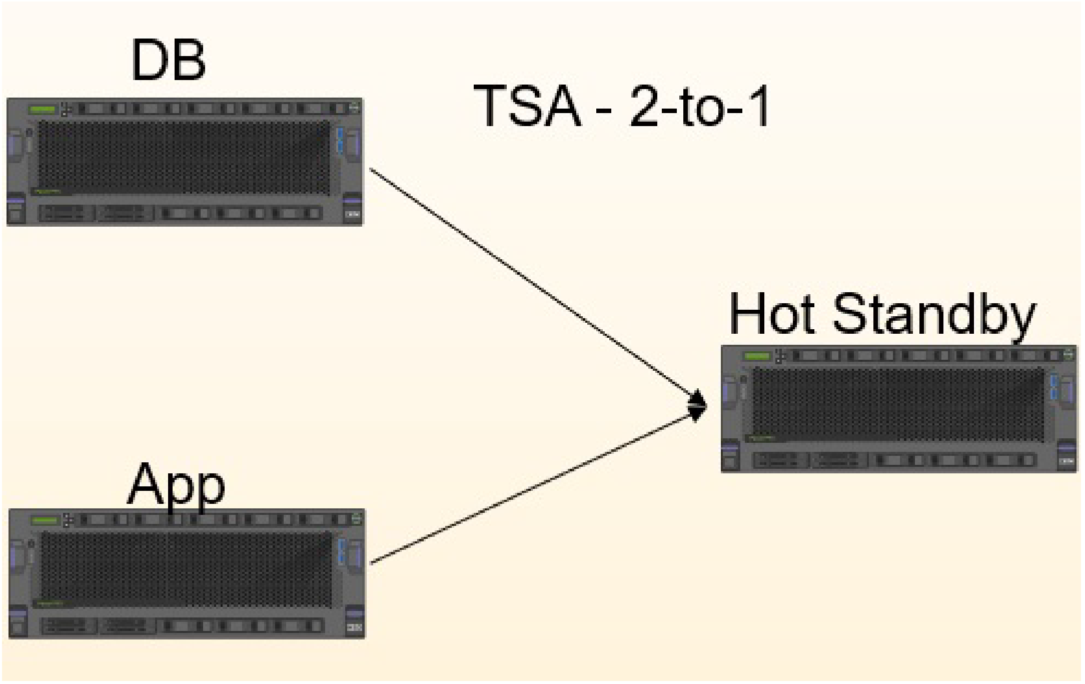


Figure 3-12 TSA MP configuration scenario

3.7.3 Failure scenario expectations

The following is the expected results based on each specific failure type:

- Application outage.

  TSA MP provides the capability to monitor application(s). In the event of an application failure it can either be restarted locally or failed over to the next node in the cluster.

- Storage loss.

  In the event of lost storage access/storage failure, TSA MP does *not* provide any additional facilities to help in recovery. Recovery is dependent on fixing the problem if it were just a connectivity problem, or in the event of full storage subsystem failure, recreating storage luns and restoring data if/as needed.

- Server/LPAR.

  If a server or LPAR fails it can be restarted on the other server by way of the KSYS controller node. So it is important that the KSYS controller node be available to perform this action.This can be either automatic or manually initiated.

- Site outage.

  In the event of an entire site outage this solution provides *no* additional recoverability as it is a single site without replication.

3.8 Spectrum Scale stretched cluster

This section looks at the configuration of a stretched Spectrum Scale cluster for DR. It is the ideal solution if your application requires a highly available, high performance file system active-active across two data centers. While the data centers can be connected by SAN or IP, we will focus on the IP solution as it is a solution for both on-premises and cloud.

Spectrum Scale supports a wide range of applications, for example:

- A highly available and scalable network file system.
- A tiered scalable storage solution.
- A multi-site concurrent database solution.
- A multi-media streaming solution.
- Persistent storage for Red Hat OpenShift or container solution.

Spectrum Scale has many advanced features such as GPFS RAID, storage tiering and information lifecycle management. Refer to the following document:

https://www.ibm.com/docs/en/spectrum-scale/5.1.1?topic=overview-spectrum-scale

This configuration makes use of 2 sites with Spectrum Scale quorum nodes and a separate failure group defined at each site. A third site with a quorum node and a local disk with just a file system descriptor is added for availability. Spectrum Scale is installed on 7 nodes, 6 of which will be providing cluster access to the data on their site’s SAN storage.

3.8.1 Configuration of the nodes and the NSDs

All sites in this example are connected by way of IP, which also provides client access to the clustered file system. At each of these site there are 3 nodes, one of which is a quorum node. Each of these nodes has SAN attached local storage providing the Spectrum Scale NSDs (backing storage). The NSDs at SiteA are in failure group 1 and the NSDs at SiteB are in failure group 2. The third site has a quorum node with only one local NSD of type **descOnly** (contains no data) in failure group 3. The shared file system is built using all these NSDs.

Spectrum Scale uses failure groups to determine where to place copies of both file data and metadata. Thus if a file system has the number of default data and metadata replicas set to 2, there will be one copy of all data/metadata in each failure group. In this example this means a complete mirror in each of SiteA and SiteB.

Spectrum Scale can use quorum nodes to determine if the cluster is active. For the cluster to be active, a majority of quorum nodes need to be available (**mmfsd** daemon active and reachable). Spectrum Scale also uses the concept of file system descriptors to determine if a file system should be mounted across the cluster. In this example we have 3 failure groups, so Spectrum Scale will set 3 descriptors, one in each failure group. As long as there are at least two sites available, there will be 2 quorum nodes (out of 3) and 2 file system descriptors (out of 3) so the cluster will be active and the file system can be mounted. The unreachable site will not be able to form an active cluster or mount the file system.

All nodes are connected by way of a single network, which is also used for client access. Each Spectrum Scale node is defined as a NSD Server, which allows the nodes at SiteA to access the NSDs at SiteB and visa-versa as shown in Figure 3-13.

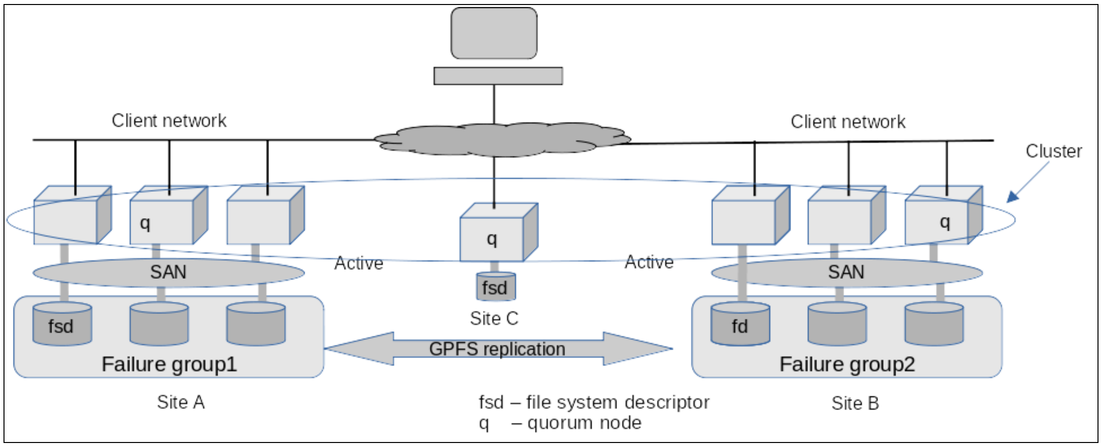


Figure 3-13 Spectrum Scale stretched cluster

3.8.2 Configuration of the file system

This scenario has one file system configured using all the NSDs from the 3 sites and the following settings:

- Maximum data replicas set to 2.
- Default data replicas set to 2.
- Maximum metadata replicas set to 2.
- Default metadata replicas set to 2.

This means that every file created in the file system will have a copy of its data and metadata at each site. The NSD at the third site does not contain data, and will only be used for file system quorum calculations.

3.8.3 Failure scenarios

As previously discussed, the configuration of quorum nodes and file system descriptors will ensure that should a single site fail or be unreachable, the remaining two sites will continue while the single site stops sharing the file system should any local clients be able to connect. Clients connecting to the surviving site will continue as normal. After the failed site is re-connected to the surviving nodes, the data on its local NSDs can be re-synchronised with the surviving site. During this process all clients will be using the latest copy of the data from the surviving site.

For example if SiteA fails, clients will still be able to access the file system at SiteB as shown in Figure 3-14 on page 106.

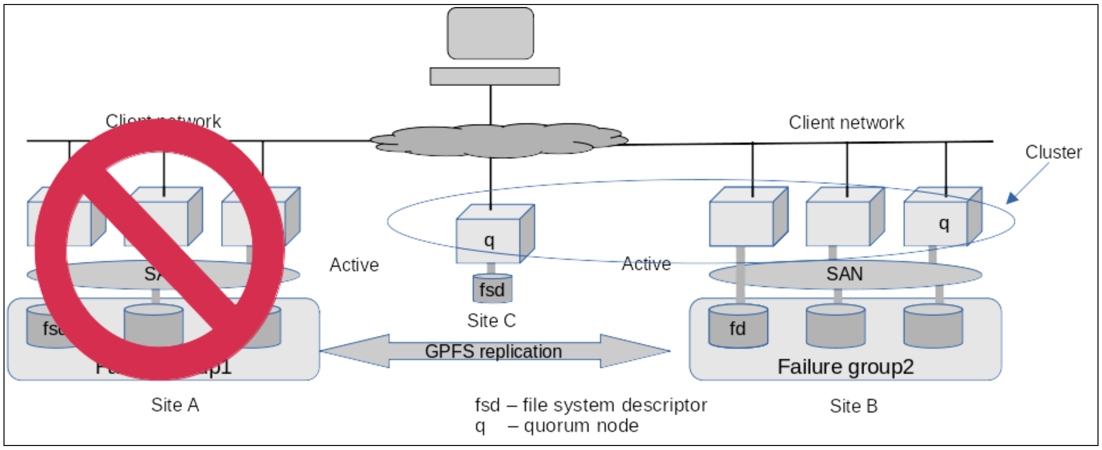


Figure 3-14 Spectrum Scale stretched cluster with SiteA failed

Similarly if the third site fails, the clients will still be able to access the file system at either SiteA or SiteB, see Figure 3-15.

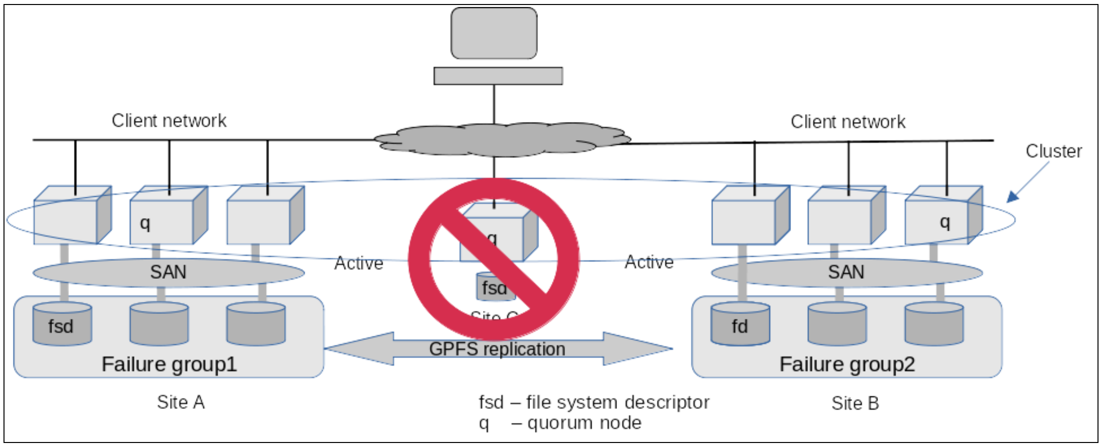


Figure 3-15 Spectrum Scale stretched cluster with quorum site failed

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this paper.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize Version 8.4*, SG24-8491.
- *IBM Storwize V7000, Spectrum Virtualize, HyperSwap, and VMware Implementation, SG24-8317.*
- *IBM Spectrum Virtualize HyperSwap SAN Implementation and Design Best Practices, REDP-5597.*
- *IBM SAN Volume Controller Stretched Cluster with PowerVM and PowerHA, SG24-8142.*
- *IBM Spectrum Virtualize 3-Site Replication, SG24-8504.*
- *IBM System Storage DS8000 Series: IBM FlashCopy SE, REDP-4368.*
- *Implementing IBM Spectrum Virtualize for Public Cloud Version 8.3.1, REDP-5602.*
- *IBM DS8000 Copy Services: Updated for IBM DS8000 Release 9.1, SG24-8367.*
- *Achieving Hybrid Cloud Cyber Resiliency with IBM Spectrum Virtualize for Public Cloud, REDP-5585.*
- *Multicloud Solution for Business Continuity using IBM Spectrum Virtualize for Public Cloud on AWS Version 1 Release 1, REDP-5545.*
- *High Availability and Disaster Recovery Planning: Next-Generation Solutions for Multiserver IBM Power Systems Environments, REDP-4669.*
- *Implementing IBM VM Recovery Manager for IBM Power Systems, SG24-8426.*
- *Introduction to Workload Partition Management in IBM AIX Version 6.1, SG24-7431.*
- *Exploiting IBM AIX Workload Partitions, SG24-7955.*
- *IBM Power System E980: Technical Overview and Introduction, REDP-5510.*
- *IBM Copy Services Manager Implementation Guide, SG24-8375.*
- *IBM PowerHA SystemMirror for AIX Cookbook, SG24-7739.*
- *End-to-end Automation with IBM Tivoli System Automation for Multiplatforms, SG24-7117.*
- *Implementing High Availability and Disaster Recovery Solutions with SAP HANA on IBM Power Systems, REDP-5443.*
- *PowerHA SystemMirror for IBM i Cookbook, SG24-7994.*
- *IBM AIX Continuous Availability Features, REDP-4367.*

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- IBM Copy Services base publications

  https://www.ibm.com/docs/en/csm
- Demonstration of automated remote restart capability

  https://www.youtube.com/watch?v=6s72ZR50Lr8
- Demonstration of VMRM DR

  https://www.youtube.com/watch?v=kTe0Tzp0ghs&t=8s
- TSA for multiplatforms

  https://www.ibm.com/docs/en/tsafm/4.1.0
- IBM Spectrum Virtualize for Public Cloud

  https://www.ibm.com/products/spectrum-virtualize-for-public-cloud

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

REDP-5656-00

ISBN DocISBN

Printed in U.S.A.