

IBM Spectrum Virtualize HyperSwap SAN Implementation and Design Best Practices

David Green

Jordan Fincher



Storage



Introduction

IBM® Spectrum Virtualize HyperSwap® and Stretched Cluster configuration information is available at [IBM Knowledge Center](#).

Note: In this IBM Redpaper™ publication, for brevity, we use *HyperSwap* to refer to both HyperSwap and Stretched Cluster.

The documentation details the minimum requirements. However, it does not describe the design of the storage area network (SAN) in detail, nor does it describe the recommended way to implement those requirements on a SAN.

In this publication, we outline some of the best practices for SAN design and implementation that leads to optimum resiliency of the SAN Volume Controller cluster, and we explain why each recommendation is made.

This paper is SAN vendor-neutral wherever possible. Any mention of a specific SAN switch vendor, or terms used by a specific switch vendor, is made only where relevant to a specific context, and does not imply an endorsement of a specific switch vendor.

Note: Some of the figures in this document might not depict redundant fabrics or storage configurations. This was done for simplicity, and it should be assumed that any recommendations made for fabric design assume that there are two redundant fabrics.

IBM Spectrum Virtualize HyperSwap SAN design best practices

The following list includes recommendations to follow for a best practice SAN design. Each of the recommendations is explained in more detail, including common implementation issues that have led to customer problems with HyperSwap clusters.

- ▶ Redundant fabrics, with each fabric divided into public and private dedicated SANs
- ▶ Dedicated inter-site links for the public and private SANs
- ▶ Implement each redundant fabric on a separate provider, rather than having both fabrics on both providers
- ▶ Size the inter-site links for the private SAN appropriately, because all of the writes for HyperSwap volumes traverse these links
- ▶ All links between sites on the public and private SANs should be trunked, rather than using separate ISLs

Figure 1 shows an implementation of a best practice design recommendation.

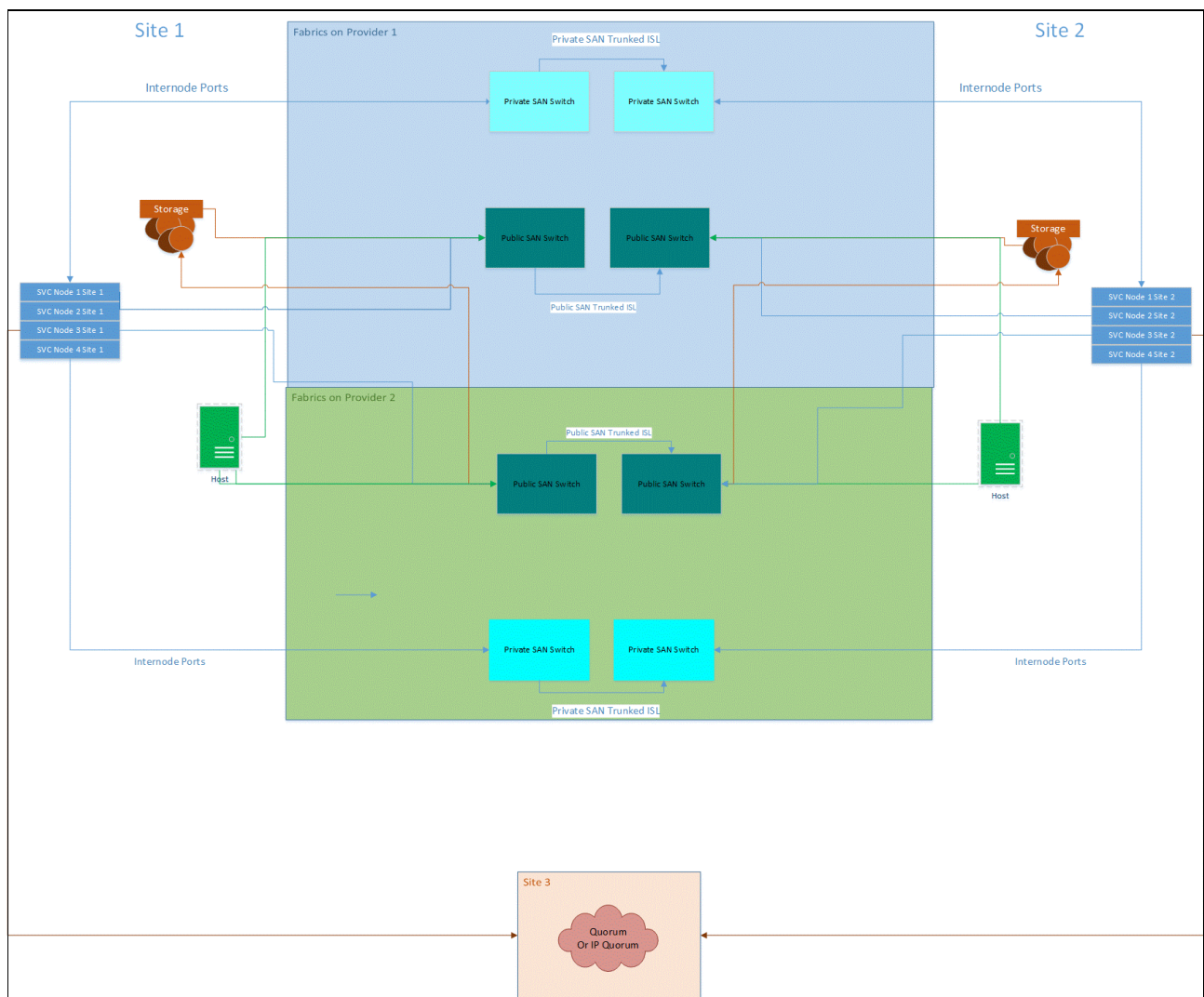


Figure 1 SAN best practice design recommendation

Figure 1 on page 2 shows an SAN Volume Controller cluster. However, the same design applies to IBM Storwize®, FS9100, and the FlashSystem storage systems if they are configured in an IBM HyperSwap topology. The IBM Spectrum® Virtualize HyperSwap cluster at each site is attached to two redundant fabrics. Each fabric is further divided into public and private SANs, and these SANs can be virtual or physical, with separate switches.

The third site pictured is the quorum site; however, this is optional. The quorum can also be one or more IP quorums, but there must either be a third site with a quorum, or an IP quorum that the nodes at both sites have access to.

Redundant fabrics

Redundant fabrics, as a best practice, are not unique to HyperSwap. However, what is unique to these IBM Spectrum Virtualize cluster configurations is a further subdivision of each of the redundant fabrics into public and private fabrics, as shown in Figure 1 on page 2.

The public SAN contains the IBM Spectrum Virtualize node ports used for host and controller connections, and the hosts and controllers.

The private SAN is a dedicated SAN, and must only contain the node ports that are used to communicate between the nodes within the same cluster. No other device ports should be attached to this private SAN, and the node ports that are used for the internode communication should not be used for copy services to a partner IBM Spectrum Virtualize cluster.

One of the common mistakes made is to use the private SAN for other unrelated purposes in *addition* to the internode communications. For example, an incorrect configuration would be using the private SAN to run tape backups, and this implementation has two issues.

Firstly, the private SAN is no longer a dedicated SAN; secondly, (and more problematic) some of the backup traffic traverses the same ISLs between the sites that the HyperSwap write data is using. This causes bandwidth issues on these ISLs and negatively affects the reliability and performance of the HyperSwap solution.

Another mistake is if the cluster is configured for copy services, such as Global Mirror or Global Mirror with Change Volumes (GMCV), and the IBM Spectrum Virtualize node ports dedicated to copy services are subsequently connected to the private SAN. These ports should not be connected to the private SAN. They can be connected either to the public SAN, or to their own dedicated virtual fabric, but they must not be connected to the private SAN.

Separate inter-site links for the public and private SANs

One of the requirements is that the public and private SANs must be completely separate from each other. However, one of the most common misconfiguration problems in HyperSwap is that this requirement is not adhered to. The public and private SANs must be completely separate and, as such, the inter-site/inter-switch links must also be completely separate. If separate physical switches are used, this requirement is nearly always automatically met.

Figure 2 shows this common mistake.

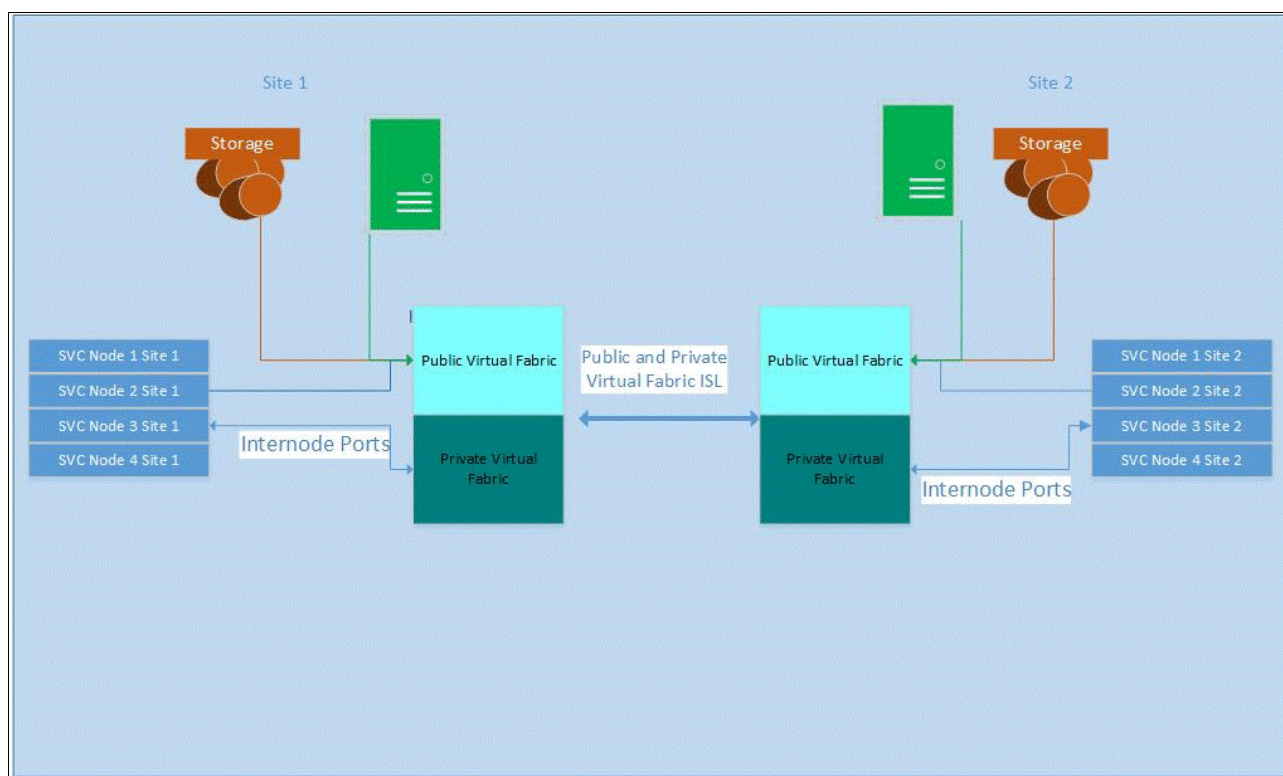


Figure 2 Common SAN mistake

In this design, the IBM Spectrum Virtualize node ports are connected to virtual fabrics, and the internode ports are on their own dedicated, private SAN. However, the mistake that has been made is to subsequently allow the public and private SANs to traverse the *same* physical links between the sites. Not only is it not a best practice but it is also not a supported configuration.

Figure 2 shows an implementation that is only seen on SANs that use virtual fabrics for the public and private SANs. Using virtual fabrics to separate the public and private traffic is a valid configuration. However, as can be seen in Figure 2, care must be taken to ensure that the public and private fabrics traverse dedicated ISLs.

On Cisco switches, customers implement public and private VSANs as shown in Figure 2, but they then configure the VSANs allowed on the port-channels to enable traffic from both VSANs on the same port-channels, and this is not correct. There must be at least *one* port-channel configured to *only* allow traffic from the private VSAN.

On Brocade switches, the equivalent mistake would be configuring XISLs to allow multiple virtual fabrics to traverse the same physical ISLs in the base switches. On Brocade switches, if virtual fabrics are used, the private virtual fabric *must* have its own, dedicated ISLs. Although using separate switches is more costly than implementing Cisco VSANs, or Brocade virtual fabrics, on the same physical switch, it eliminates the possibility of having the public and private fabrics traverse the same ISLs.

Different providers for the redundant fabrics

Figure 1 on page 2 shows two redundant fabrics spanning the two sites where the HyperSwap cluster is located, and each fabric is on a separate distance/WAN provider. Having each fabric on a separate provider increases the reliability of the solution because disruptions in service on one provider only affect a single fabric, and the second fabric remains untouched.

Putting both fabrics on the same provider introduces a single point of failure if that provider should suffer an outage. While provider outages are rare, they do happen, and it is something that should be considered. Using an analogy, someone who has a job where communication is critical might have two phones on two different providers to ensure that if one of the providers has an outage, the phone on the other provider will still work.

Some customers design the SANs such that both fabrics are on both providers, and the expectation is that if one provider fails, both fabrics are at least partially up. In practice, it is preferable to have one of the fabrics fail completely. End devices performing path recovery can nearly always handle complete path loss more effectively than partial or intermittent path loss.

Port masking considerations

As with all other types of SAN Volume Controller and FlashSystem HyperSwap implementations, it is imperative to isolate the local cluster inter-node traffic from all other forms of traffic. This is important to maintain cluster performance and stability, because lease renewal messages, cache-mirroring, and HyperSwap synchronization traffic all take place over the ports that are assigned for local node traffic on the system.

The general recommendations for port isolation and masking vary based on the number of ports per node that a specific configuration includes.

Figure 3 outlines the recommended port isolation and masking configurations based on the number of ports per node.

	4 port	8 port	12 port	16 port	SAN Fabric
Adapter 1 Port 1	Host+Storage	Host+Storage	Host+Storage	Host+Storage	A
Adapter 1 Port 2	Host+Storage	Host+Storage	Host+Storage	Host+Storage	B
Adapter 1 Port 3	Intracuster+Replication	Intracuster	Intracuster	Intracuster	A
Adapter 1 Port 4	Intracuster+Replication	Intracuster	Intracuster	Intracuster	B
Adapter 2 Port 1		Host+Storage	Host+Storage	Host+Storage	A
Adapter 2 Port 2		Host+Storage	Host+Storage	Host+Storage	B
Adapter 2 Port 3		Intracuster or Replication	Replication or Host+Storage	Replication or Host+Storage	A
Adapter 2 Port 4		Intracuster or Replication	Replication or Host+Storage	Replication or Host+Storage	B
Adapter 3 Port 1			Host+Storage	Host+Storage	A
Adapter 3 Port 2			Host+Storage	Host+Storage	B
Adapter 3 Port 3			Intracuster	Intracuster	A
Adapter 3 Port 4			Intracuster	Intracuster	B
Adapter 4 Port 1				Host+Storage	A
Adapter 4 Port 2				Host+Storage	B
Adapter 4 Port 3				Replication or Host+Storage	A
Adapter 4 Port 4				Replication or Host+Storage	B
localfcportmask	1100	11001100 OR 00001100	110000001100	0000110000001100	
remotefcportmask	1100	00000000 OR 11000000	000011000000	1100000011000000	
<p>Host refers to host objects defined in the system.</p> <p>Storage refers to controller objects defined in the system if external storage is being used.</p> <p>Replication refers to nodes which are part of a different cluster.</p> <p>Intracuster refers to nodes within the same cluster.</p> <p>The "+" indicates that both types are should to be used</p> <p>The word "or" indicates that one of the options must be selected. If using replication, preference should be given to replication.</p>					

Figure 3 Port isolation and masking recommendations

Note: In HyperSwap, replication ports are required only if replication to a remote cluster is taking place. Generally speaking, HyperSwap configurations will not be replicating in this way (unless participating in 3-site replication).

Only nodes that are participating in the HyperSwap cluster should be visible on the private SAN. Hosts, virtualized storage controllers, and replication partners should not be accessible on the private SAN.

Sizing the inter-site links for HyperSwap traffic

It is critical that the bandwidth requirements for the inter-site links be understood before beginning configuration of the fabric. This might be difficult under the following conditions:

- ▶ For a cluster that is being set up as a new cluster, unless hosts are being migrated from an existing cluster, because no historical performance data exists.
- ▶ For a new cluster with new hosts, no historical performance data exists.
- ▶ For a cluster that is being migrated to a HyperSwap configuration, or for a new cluster where hosts are being migrated from older storage, there might be historical performance data.

If a volume is not configured for HyperSwap, when a host writes to that volume the write data will not traverse the ISLs to the remote site, unless the storage where that volume resides is at the remote site.

Bandwidth requirement calculations do not need to include the write data rates for all of the volumes in an IBM Spectrum Virtualize cluster, unless all of the volumes are configured as HyperSwap volumes. Writes to volumes that are not configured as HyperSwap volumes do not normally traverse the links between the sites, assuming that the host is connected to the same site, because of the storage providing the non-HyperSwap volumes that it is accessing. HyperSwap volumes have all writes mirrored on the private fabrics, so these volumes must be included in the sizing.

When sizing links, a physical link is considered saturated at 75% of capacity. An 8 gigabit per second (Gbps) link is saturated at ~600 Megabytes per second (MBps) of throughput. A 16 Gbps link is saturated at 1200 MBps. An estimate can be made by measuring the peak write data rate for all of the HyperSwap volumes in a cluster, and using that as the bandwidth requirement for the inter-site links.

Each of the redundant fabrics must be able to carry all of the data, so the bandwidth requirement is actually 2x the peak write data rate. We have seen clients take this figure and implement the SAN such that each fabric can carry half the peak data rate. This is based on the assumption that each fabric will carry part of the data.

This works until there is an outage on one of the fabrics and the inter-site link drops. When this happens, all of the data being written at each site is then forced over the surviving fabric, but the links do not have enough bandwidth to handle that amount of traffic.

For the purposes of the bandwidth requirements calculations, a *link* is considered to be all of the physical links between a pair of switches. So, a pair of 8 Gbps links would therefore have the same bandwidth as a single 16 Gbps link. However, there are some design considerations for the inter-site link design.

Trunking the inter-site links

Inter-site links can be configured as either a group of physical links in a single logical link (also called a *trunk* or *port-channel*), or left as separate physical links. The best practice recommendation is to implement the inter-site links as multiple, redundant, physical links on each fabric and to combine them into a trunk, rather than leaving them as separate physical links.

For Fibre Channel switches, a trunk is considered an ISL between the two switches. This means that all the links in a trunk would have to go down before the switches would perform a fabric rebuild. A fabric rebuild is a reconfiguration of the fabric where switches update their routing tables and other fabric-related data.

However, some end devices are sensitive to a fabric rebuild, and excessive rebuilds of a flapping (alternating between up and down states) ISL have been seen to cause problems with end devices. If a single link in a trunk goes down, a rebuild is not performed. This prevents a trunked flapping ISL from initiating a rebuild. However, if multiple links are configured as separate physical links, then each time one of those links drops, a fabric rebuild is performed.

Load-balancing across the links in a trunk is more effective than load-balancing across a set of single links. If a link in a trunk drops, it might be possible for the switch to resend frames without having to abort an exchange, and have end devices do error recovery.

Zoning Considerations for HyperSwap

The following zoning requirements exist for all SAN Volume Controller, IBM Spectrum Virtualize, and FlashSystem clusters:

- ▶ Enable the NPIV feature on the cluster, and zone all hosts to the virtual WWPNs on each node port.
- ▶ Use the physical WWPNs on the node ports for inter-node communication. Do not zone any hosts or controllers to these ports.
- ▶ Use the physical WWPNs on the node ports that are used for inter-cluster (replication) communication. Do not zone any hosts or controllers to these ports.
- ▶ Use the physical WWPNs on the node ports for controller (storage) communication.
- ▶ If a hot-spare node is configured, all controllers must be zoned to the physical WWPNs of the hot-spare node.

For HyperSwap, the requirements are the same. The ports used for inter-node communication will be on dedicated private fabrics. If this is implemented correctly, it is not possible to zone hosts, controllers, or node ports for a partner cluster to these ports. The private fabrics should only have a single zone each that includes all of the cluster node ports attached to those fabrics. Because no other devices should be on these fabrics, there should not be any other zones.

Authors

This paper was produced remotely by a team working for the IBM Redbooks, San Jose Center.

David Green works with the IBM SAN Central team troubleshooting performance and other problems on storage networks. He has authored, or contributed to, a number of IBM Redbooks® publications. He is a regular speaker at IBM Technical University. You can find his blog at [Inside IBM Storage Networking](#) where he writes about all things related to Storage Networking and IBM Storage Insights.

Jordan Fincher works with the IBM SAN Volume Controller and IBM FlashSystem® support teams troubleshooting a wide variety of problems. He has contributed to several IBM Redbooks publications and periodically speaks at IBM Technical University events. You can find his blog [Supporting Spectrum Virtualize](#) where he writes about various support cases.

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

Chris Bulmer
Bill Passingham
Nolan Rogers
IBM Hursley, UK

This project was managed by:

Jon Tate
IBM ITSO

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

Stay connected to IBM Redbooks

- ▶ Find us on Facebook:
<http://www.facebook.com/IBMRedbooks>
- ▶ Follow us on Twitter:
<http://twitter.com/ibmredbooks>
- ▶ Look for 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>

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.

HyperSwap®


IBM Spectrum®

Storwize®

IBM®

Redbooks®

IBM FlashSystem®

Redbooks (logo) ®

The following terms are trademarks of other companies:

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



REDP-5597-00

ISBN 0738458716

Printed in U.S.A.

Get connected

