

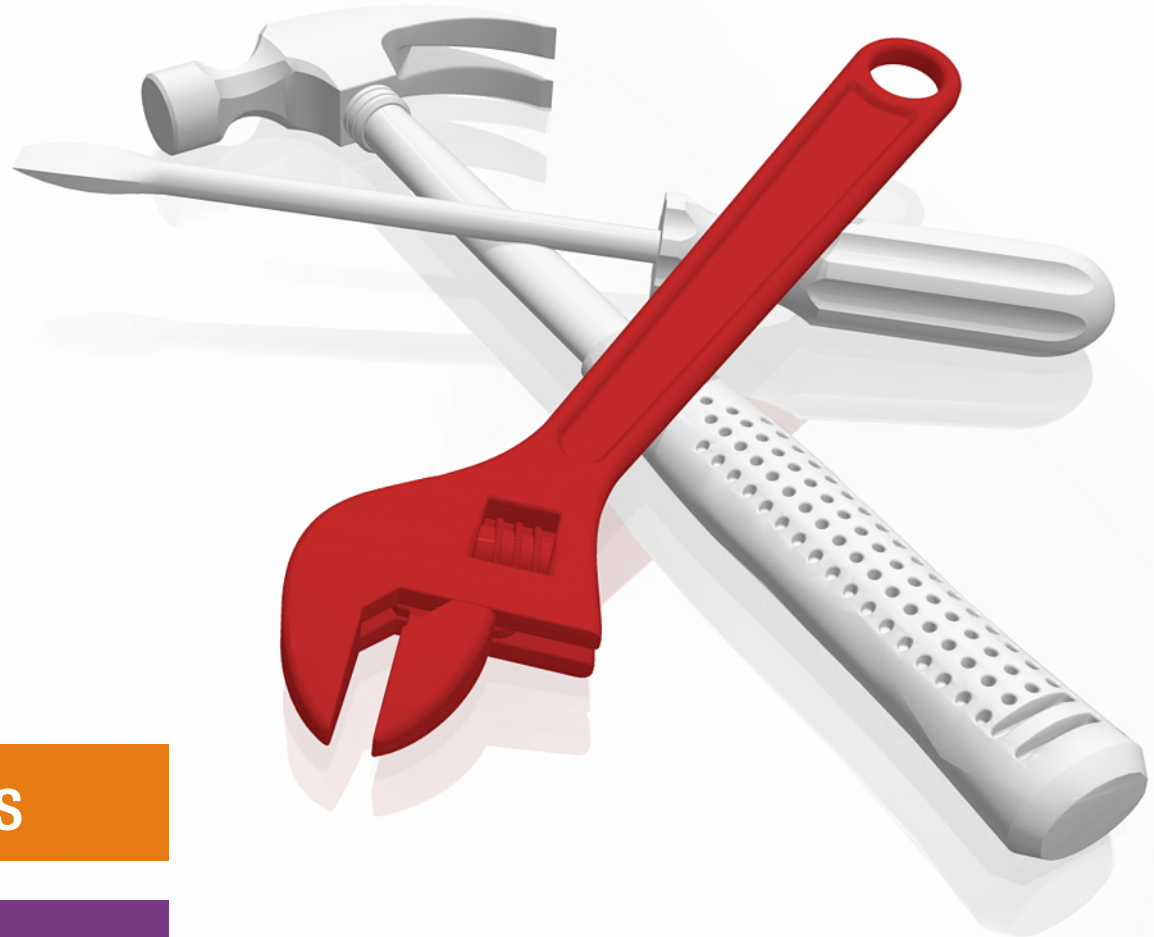
IBM Power Systems Infrastructure I/O for SAP Applications

Dino Quintero

Edmund Haefele

Gerd Kehrer

Katharina Probst



 Analytics

Power Systems



IBM Redbooks

**IBM Power Systems Infrastructure I/O for SAP
Applications**

April 2020

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

First Edition (April 2020)

This edition applies to:

SUSE Linux Enterprise Server 12 Service Pack 3

SUSE Linux Enterprise Server 12 Service Pack 4

SUSE Linux Enterprise Server 15

IBM Virtual I/O Server (VIOS) V3.1

IBM powerpc-utils-1.3.3-7.6.2

© Copyright International Business Machines Corporation 2020. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Notices	v
Trademarks	vi
Preface	vii
Authors	vii
Now you can become a published author, too!	viii
Comments welcome	viii
Stay connected to IBM Redbooks	ix
Chapter 1. Ethernet architectures for SAP workloads	1
1.1 Preferred Ethernet cards for Linux on IBM POWER9 processor-based servers for SAP workloads	2
1.2 Ethernet technology introduction	2
1.2.1 Dedicated and physical sharing	2
1.2.2 Network virtualization	4
1.2.3 Selecting the correct technology for SAP landscapes	8
1.3 Ethernet tuning for SAP networks	9
1.3.1 Optimizing the network configuration for throughput on 10 Gbps by using SEA and jumbo frames	10
1.3.2 Latency optimization	11
1.4 VIOS configuration for SAP network requirements	11
Chapter 2. Storage system, connectivity, and file system architecture	13
2.1 Filer infrastructures	14
2.2 PCIe Non-Volatile Memory Express (NVMe) enriched POWER servers	14
2.2.1 NVMe use cases and technology	15
2.3 SAN infrastructures	16
2.3.1 SAN use cases and architectures best practices	16
2.4 Fibre Channel infrastructure and VIOS options	18
2.4.1 vSCSI	18
2.4.2 N_Port ID Virtualization	20
2.4.3 Example setup of an NPIV virtual Fibre Channel configuration	21
2.5 iSCSI boot disk attachment with VIOS 3.1	30
2.5.1 Configuring iSCSI on the VIOS	33
2.6 Linux I/O	34
2.6.1 Multipathing	34
2.6.2 Sample multipath configuration	36
2.6.3 Linux file systems that are relevant to SAP applications	42
2.6.4 Logical Volume Manager	43
Related publications	47
IBM Redbooks	47
Online resources	47
Help from IBM	47

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](http://www.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®
IBM®
IBM Elastic Storage®

IBM Spectrum®
POWER®
POWER8®
POWER9™
PowerVM®

Redbooks®
Redbooks (logo) ®
System Storage™

The following terms are trademarks of other companies:

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.

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

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

Preface

This IBM® Redpaper publication describes practical experiences to run SAP workloads to take advantage of IBM Power Systems I/O capabilities. With IBM POWER® processor-based servers, you have the flexibility to fit seamlessly new applications and workloads into a single data center, and even consolidate them into a single server. This approach highlights all viable options and describes the pros and cons of each one to select the correct option for a specific data center.

The target audiences of this book are architects, IT specialists, and systems administrators deploying SAP workloads, who spend much time and effort managing, provisioning, and monitoring SAP software systems and landscapes on IBM Power Systems servers.

Authors

This paper was produced in close collaboration with the IBM SAP International Competence Center (ISICC) in Walldorf, SAP Headquarters in Germany and IBM Redbooks®.



Dino Quintero is an IT Management Consultant and IBM Level 3 Senior Certified IT Specialist with IBM Redbooks in Poughkeepsie, New York. He has 24 years of experience with 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, cloud computing, artificial intelligence (AI) (including machine and deep learning), and cognitive solutions. He is a Certified Open Group Distinguished IT Specialist. Dino holds a Master of Computing Information Systems degree and a Bachelor of Science degree in Computer Science from Marist College.

Edmund Haefele is a specialist in IT infrastructure for SAP environments with more than 20 years of experience in this field. Edmund is Level 2 Expert Certified as an IBM architect, and worked as an IT architect and specialist for many projects for both German and international clients. Edmund has a deep knowledge of IT systems, such as SAP HANA, servers, storage, network technologies, high availability (HA), backup and restore, disaster recovery (DR), service management, and automation. He is a Certified Open Group Master Certified IT Architect. Edmund holds a diploma in Physics and a PhD in natural sciences from the University of Heidelberg.

Gerd Kehrer leads the IBM IT-Admin team at SAP Headquarters, where he is responsible for SAP HANA on Power Systems development and the test infrastructure. He has 24 years of experience with SAP on IBM Systems in the areas of SAP backup and archive, SAP monitoring, and systems management on IBM AIX® and Linux.

Katharina Probst leads the IBM Development team at SAP Headquarters, where she is responsible for Datacenter Readiness and Ecosystem enablement of SAP HANA on Power Systems. She has 15 years of experience with SAP on Power Systems in the areas of storage, business continuity, DR, AIX, and Linux.

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

Wade Wallace
IBM Redbooks, Austin Center

Walter Orb and Tanja Scheller
IBM Germany

Paulo Sergio Lemes Queiroz
IBM Brazil

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 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>



Ethernet architectures for SAP workloads

This chapter summarizes several technologies and puts them into the context of different SAP landscapes. In SAP landscapes, Shared Ethernet Adapters (SEAs) on 10 Gbps infrastructures are the dominant and most viable deployment. In the future, new technology is coming that will change the landscape, especially for larger SAP S/4HANA/Business Suite environments.

This chapter describes the following topics:

- ▶ Preferred Ethernet cards for Linux on IBM POWER9 processor-based servers for SAP workloads
- ▶ Ethernet technology introduction
- ▶ Ethernet tuning for SAP networks
- ▶ VIOS configuration for SAP network requirements

1.1 Preferred Ethernet cards for Linux on IBM POWER9 processor-based servers for SAP workloads

At the time of writing, the preferred Ethernet cards that were tested in the context of SAP workloads in a Linux environment on IBM POWER9™ processor-based servers are shown in Table 1-1.

Table 1-1 Preferred Ethernet cards

Description	Low-profile Fibre Channel	Full-height Fibre Channel
PCIe3 LP 2-Port 10 GbE NIC & RoCE SR/Cu Adapter	EC2R	EC2S
PCIe3 LP 2-Port 25/10 GbE NIC & RoCE SR/Cu Adapter	EC2T	EC2U
PCIe3 LP 2-port 100/40 GbE NIC & RoCE QSFP28 Adapter x16	EC3L	EC3M
PCIe4 LP 2-port 100/40 GbE NIC & RoCE QSFP28 Adapter x16	EC67	EC66

With virtual network interface cards (vNICs) and 25 or 100 Gb cards, use the latest firmware levels when using vNICs to ensure the highest processor (core) savings and best performance.

Note: During the development of this publication, the team did not explicitly test the cards with transceivers.

1.2 Ethernet technology introduction

IBM PowerVM® provides different options for virtualizing the network connectivity of a Linux logical partition (LPAR). You must determine whether dedicated network adapters can be assigned to the LPAR to achieve the highest network bandwidth and lowest latency or whether you can use Virtual I/O Server (VIOS) and use its advanced flexibility and reliability options. IBM Power Systems servers can be configured in mixed modes with some LPARs that are configured with dedicated or shared adapters and other LPARs that use VIOS.

1.2.1 Dedicated and physical sharing

An Ethernet network card can be directly dedicated to one LPAR (all ports are bound to a single LPAR), or it can be shared by using Single Root I/O Virtualization (SR-IOV) technology. SR-IOV provides logical ports to share the physical ports across multiple LPARs.

The tradeoff for these deployment options are less latency and better throughput because of the absence of Live Partition Mobility (LPM). Without VIOS, all traffic goes outside the server and comes back to it, unlike internal virtual LAN (vLAN) configurations with VIOS.

The typical use cases that these two options provide are as follows:

- ▶ A single LPAR uses the full central electronics complex (CEC) without LPM (for example, database replication to a second node).
- ▶ Small deployments on S-class servers.
- ▶ Large SAP S/4HANA/Business Suite databases (appserver traffic).

Dedicated adapters

If you are using dedicated adapters, all physical adapters are assigned directly to the client LPAR. The adapter is exclusively bound to one particular partition, including all its ports. Dedicated adapters provide the best possible performance and latency, but do not allow any resource sharing.

Single Root I/O Virtualization

SR-IOV is an enhanced network virtualization technology on Power Systems servers. In SR-IOV shared mode, the physical network adapter is assigned to and managed by the IBM PowerVM Hypervisor. The physical SR-IOV adapter has multiple physical ports that are connected to external network switches. On POWER9, the different ports of the adapter can be equipped with different transceivers to allow operations with different network speeds.

In this case, no VIOS partition is required, so sharing is possible by enabling the SR-IOV adapter in *SR-IOV shared mode*. The ratio between LPARs and required adapters, occupied PCI slots, and used network ports is improved because of better resource utilization. Depending on the adapter type, a different number of *virtual functions* is possible. The number of virtual functions define the granularity for the partitioning of the adapter. For more information, see [How many logical ports/VFs are supported per adapter?](#)

Each LPAR receives an SR-IOV logical port (Figure 1-1) with ensured capacity and bandwidth that is assigned according to the defined number of virtual functions.

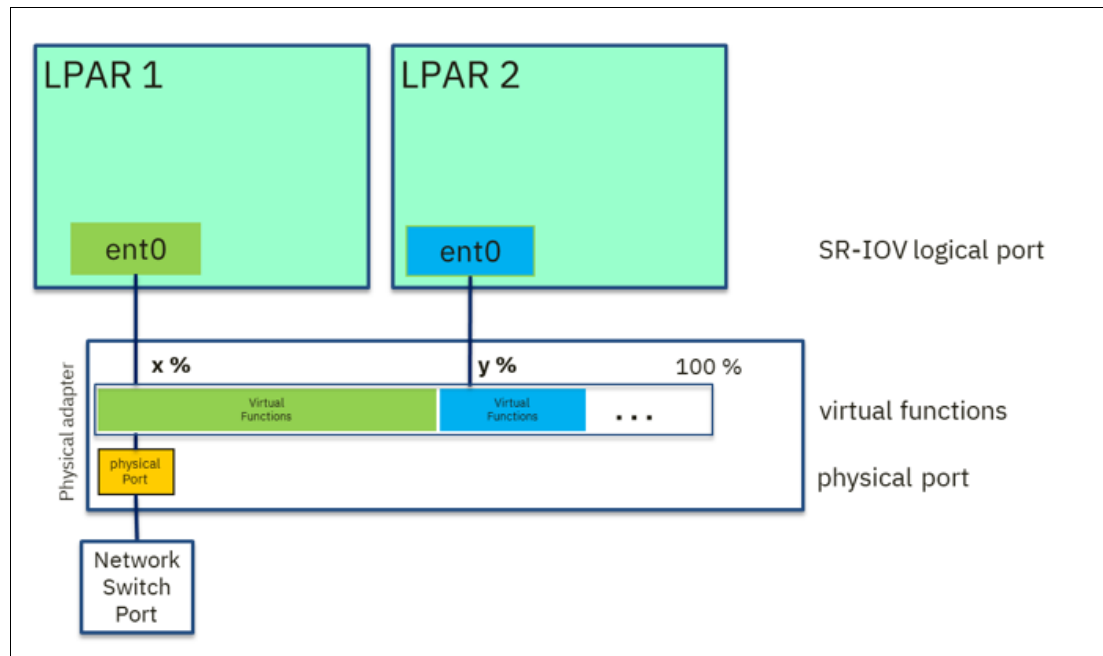


Figure 1-1 SR-IOV logical port view assignment per LPAR

The entitled capacity is the ensured amount of adapter bandwidth, which can be exceeded if the port has available bandwidth¹.

Remote direct memory access (RDMA) technology minimizes the required **memcpy** actions in the layers. So, SR-IOV provides more packets per second with lower latency and lower CPU consumption compared to SEA technology. Workloads that use many small packages can benefit from a latency perspective, such as transactional workloads where many appservers send their requests to a single database.

For each logical port that is assigned to an LPAR or configured for eventual use, a small amount of network bandwidth is reserved and is not available for any other LPAR. For example, a 10 Gbps port is assigned to 48 LPARs (2% each), and only one LPAR heavily communicates with the network (all other LPARs are idle in the network), which results in a maximum throughput value for the *busy* LPAR of about 5 Gbps. When all LPARs are actively communicating, the limit is not noticeable because the sum of all communication channels is limited by the total bandwidth of the adapter.

1.2.2 Network virtualization

There are three key technologies for network virtualization in PowerVM. All of them require the mandatory implementation of a dual-VIOS setup. They are valuable network sharing options and support LPM:

- ▶ *Virtual Ethernet* is used for internal LPAR to LPAR communication when all LPARs are within the same Power System server. Virtual Ethernet does not use a physical network adapter, and it provides high bandwidth and low latency.
- ▶ SEA has different means of implementation, and it has been the dominant network virtualization technique for more than a decade. Virtual Ethernet is extended to the external network infrastructure by using physical network adapters that are assigned to the VIOS.
- ▶ vNIC, a new technology, uses SR-IOV and addresses the disadvantages of SEA (high amount of CPU utilization, limits in high numbers of packets per second, and higher latency) when using high-speed network adapters. For 25 Gbps and faster ports, this technology is starting to appear in a few SAP deployments at clients.

Note: For LPARs running production or production-like workloads, a dual-VIOS configuration is mandatory to meet the availability requirements and limit both planned and unplanned downtime.

Virtual Ethernet adapter

The virtual Ethernet adapter (internal vLAN) allows for communication between LPARs within one physical Power Systems server. The IBM POWER Hypervisor is used as an internal network switch, which provides in traditional 10 Gb environments at least twice the throughput at lower latency without external network traffic. It can be also configured when using more than 10 Gbps port speeds, but the internal vLAN speeds did not increase with the tested stack that was used in 2019 for this document.

¹ Each used logical port reserves a tiny portion of the adapter that cannot be used by others. This portion is noticeable only when configuring many LPARs, but put a workload only on a single one and try to get to line speed.

VIOS Shared Ethernet Adapter

For the SEA, the physical network or SR-IOV (promiscuous mode) adapter and all its ports are assigned to a VIOS partition, and virtual adapters are provided to the client partitions by mapping inside the VIOS a single physical port or SR-IOV logical port in promiscuous mode to multiple virtual ports. The virtual ports are then assigned to an LPAR.

The throughput scalability for the multiple LPAR setup is excellent, but it comes with a CPU cost on the VIOS for doing the mapping between virtual and physical ports, and for using **memcpy** in the various layers. For environments up to 10 Gbps network speed, a SEA setup is a good tradeoff for optimizing utilization and providing redundancy at low cost. For environments with high-speed adapters (25 Gbps, 40 Gbps, and 100 Gbps), SEA implementation does not allow you to use fully that bandwidth from a single LPAR, but you can do it from multiple LPARs to reduce the number of physical adapters.

Note: You must still have redundant physical adapters.

SEA can be configured to share the load in a dual-VIOS environment or as a simple failover configuration. For more information, see “Shared Ethernet adapters for load sharing in [IBM Knowledge Center](#) and “Shared Ethernet Adapter failover” in [IBM Knowledge Center](#).

vNIC

vNIC is a new virtual adapter type that became available in December 2015 (it was restricted to the AIX operating system (OS) then). SUSE released the **ibmvnic** driver in 2019, which is described at the [SUSE Blog](#).

For SAP landscapes, vNIC is a future-oriented solution for a higher bandwidth adapter, lower latency, and reduced CPU usage. Because the SEA virtualization cost and latency is acceptable now, there is no technical need to move to this new technology yet.

The vNIC technology enables advanced virtualization features such as LPM with SR-IOV adapter sharing, and it uses SR-IOV quality of service (QoS).

To configure a vNIC client, the SR-IOV adapter must be configured in SR-IOV shared mode. Free capacity to feed the used logical ports must be available. When an LPAR is activated with a client vNIC virtual adapter (Figure 1-2), or when a client vNIC virtual adapter is added to a partition dynamically by a DLPAR operation, the Hardware Management Console (HMC) and the platform firmware automatically creates the vNIC server and the SR-IOV logical port backing device, and dynamically adds them to the VIOS.

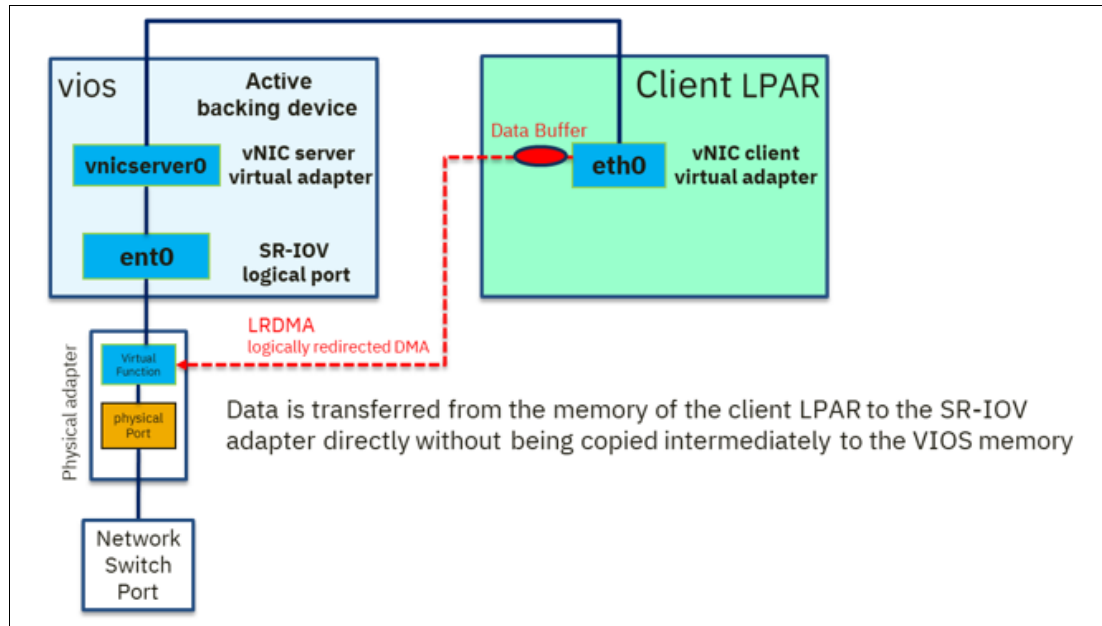


Figure 1-2 Client LPAR vNIC

The vNIC configuration requires the enhanced GUI in the HMC² or the HMC REST interface. When a vNIC adapter is added to an LPAR, all necessary adapters on the VIOS (SR-IOV logical port and vNIC server adapter) and on the LPAR (vNIC client adapter) are created in one step. No extra configuration is required on the VIOS.

vNIC has more concepts for failover (active-passive with multiple backing devices or link aggregation). Both concepts are compatible with LPM requirements and provide LPM capability.

vNIC failover

vNIC failover provides a high availability (HA) solution at the LPAR level. A vNIC client adapter can be backed by multiple logical ports (up to six) to avoid a single point of failure. Only one logical port is connected to the vNIC client concurrently (the active backing device has the highest priority). If the active backing device fails, then the hypervisor selects a new backing device according to the next highest priority.

Active-backup link aggregation technologies like Linux bonding active-backup mode can be used to provide network failover capability and sharing of the physical port (Figure 1-3 on page 7). To ensure detection of logical link failures, a network address to ping must be configured to monitor the link. For Linux active-backup mode, the `fail_over_mac` value must be set to active (`fail_over_mac=1`) or follow (`fail_over_mac=2`).

² Since 2019, the HMC comes with the enhanced GUI by default.

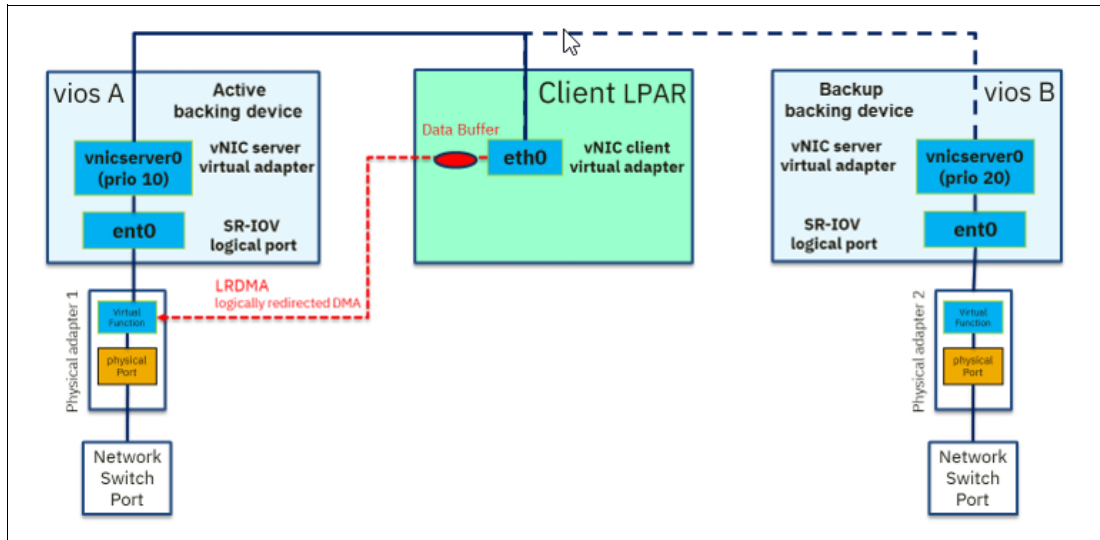


Figure 1-3 vNIC bonding

Multiple vNIC client virtual adapters can be aggregated to a single bonding device in the client LPAR to achieve higher bandwidth.

A set of requirements must be met to support network bonding (Figure 1-4):

- ▶ Each vNIC client must have a single backing device. When the vNIC client is defined with multiple backing devices, then link aggregation is not possible.
- ▶ Each SR-IOV physical port must not be shared with other vNIC servers. Per physical port, only one LPAR can be assigned. It is a best practice to configure the logical port with a capacity of 100% (to prevent sharing it with other LPARs).

Note: When using high-speed network adapters, check that the Linux service `irqbalance` is installed and active.

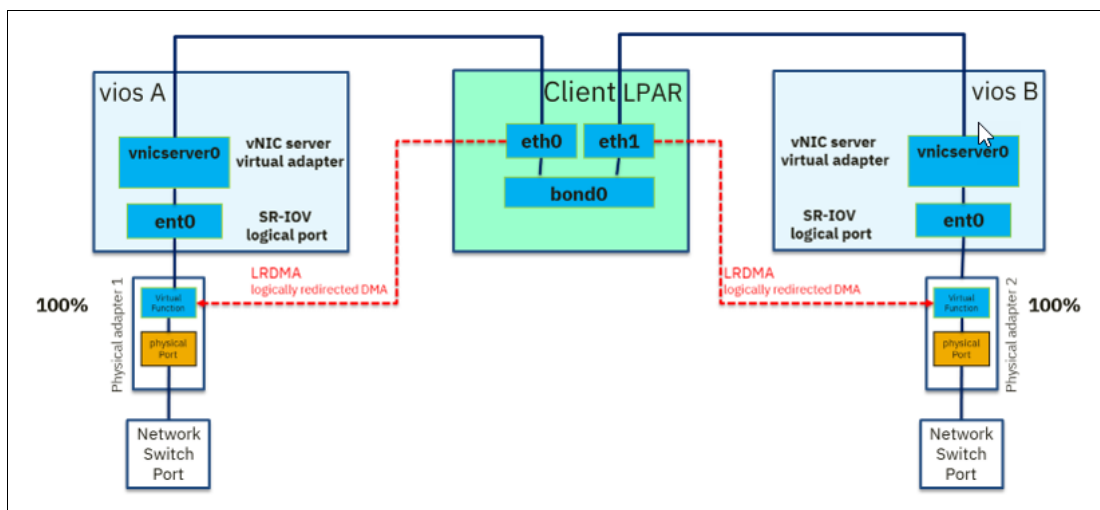


Figure 1-4 Sample architecture for using bonding and vNIC for file attachments

The target system must have an adapter in SR-IOV shared mode with an available logical port and available capacity (virtual functions) on a physical port. If labels are correctly set on the SR-IOV ports, then during LPM the correct physical ports are automatically assigned to the label name.

1.2.3 Selecting the correct technology for SAP landscapes

When you select the correct technology for your SAP landscapes, you typically have the following typical decisions to make for the client architecture workshops:

- ▶ Which IBM Ethernet virtualization and sharing capabilities are wanted (LPM capability, rolling maintenance, or sharing).
- ▶ Different network needs. For example, hot standby databases such as IBM DB2® High Availability Disaster Recovery (DR) (HADR) or SAP HANA System Replication (HSR) versus appserver communication. These needs can be defined by transmissions per second (TPS), latency, and packet sizes.

Client sample

SAP HANA is installed by using HSR for the Business Suite application. The application servers use up 160 cores in total with more than 300,000 TPS, and they have a low latency I/O requirement for small packages on the HANA LPAR. So, the application servers are on 10 Gbps SEA, but the HANA DB where all I/O is bundled is configured with SR-IOV.

Note: After this step is complete, cross-verify the planned number of adapters that fit into the selected server model, including the consideration of no network cards (for example, Fibre Channel (FC) cards).

Comparing the different network virtualization and sharing options

Table 1-2 summarizes the various options.

Table 1-2 Comparison of Ethernet technologies on Power Systems servers

Technology	LPM	QoS	Direct-access performance	Link aggregation	Requires VIOS	>400.00 TPS per 25 Gbps port	Physical adapter sharing
Dedicated network adapter	No	N/A	Yes	Yes	No	Yes	No. Each adapter is assigned directly to an LPAR.
SR-IOV	No	Yes	Yes	Yes*	No	Yes	Yes. An SR-IOV logical port is created, and virtual functions are assigned.

Technology	LPM	QoS	Direct-access performance	Link aggregation	Requires VIOS	>400.00 TPS per 25 GBps port	Physical adapter sharing
vNIC	Yes	Yes	No	Yes ^a	Yes	Not yet	For vNIC failover full sharing flexibility. For link aggregation, a full port must be dedicated.
SEA	Yes	No	No	Ye	Yes	No	Yes. A virtual Ethernet client.

a. IEEE802.3ad/802.1ax (LACP) is supported for SR-IOV and vNIC. The requirement is that there is a one-to-one relationship between the physical port and the logical port. Configure only one logical port per physical port by configuring the logical port with a capacity value of 100% to prevent configuration of more than one logical port per physical port.

1.3 Ethernet tuning for SAP networks

There are different networks in SAP landscapes, and some have different needs based on the application workload. This section highlights the key areas of Ethernet tuning, but does not necessarily cover all aspects:

- Appserver to DB server: 10 Gbps cards + SEA are often used on the application server.

Transactional workloads: The dominant deployment is based on 10 Gbps SEA with load sharing and internal vLAN. Transactional workloads can result in many small packages with lower latency. More interrupt queues can help improve the performance on the DB. In the past, the solution used dedicated adapters, but with the SR-IOV and vNIC options, more flexibility is available. Typically, it is sufficient to use this deployment on the DB side because all the application servers centralize their I/O requests (1-DB: n-application server).

Analytical workloads: These workloads tend to have fewer and larger packages that are sent to and from the DB server. In most cases, this communication does not require special handling, and unified connectivity is the objective. The dominant deployment is based on 10 Gbps SEA with load sharing. SEA still delivers best sharing characteristics when used in 10 Gbps environments for many applications with small bandwidth and no need for bandwidth control. When moving to higher speeds, SEA is not the preferred option, but it can be considered an intermediate step.

- Backup by using BACKINT.

SAP BACKINT is a backup mechanism where data is read directly from memory to the backup server. If this method is too slow, it can have an impact on the DB availability and responsiveness depending on the backup method that is configured. If the performance is insufficient, throughput must be increased (latency is not the problem because large packages are written). Throughput can be increased either by having multiple interrupt requests (IRQs) (more adapters inside the LPAR) or higher bandwidth (25 Gbps cards) by using jumbo frames and a large send offload (LSO) configuration. The offloading of large packages is mandatory to benefit from jumbo frames.

Note: The maximum speed cannot go beyond the storage speed to where the backup is written. For more information about using jumbo frames, see *Network Configurations for HANA Workloads on IBM Power Servers*, found at [SAP HANA on IBM Power Systems and IBM System Storage - Guides](#).

- ▶ Database to database.

Databases in scale-out deployments such as HANA have specific I/O patterns for internode communication. For SAP HANA, see *Network Configurations for HANA Workloads on IBM Power Servers*, found at [SAP HANA on IBM Power Systems and IBM System Storage - Guides](#). For other databases, contact your database vendor.

Database Replication for Hot Standby Solutions typically tends to create many small packages. Hence, the number of parallel interrupt queues determine the efficiency.

- ▶ Filer Attachment and internet Small Computer Systems Interface (iSCSI) boot have different patterns because this is storage I/O, which often requires high bandwidth and high-speed adapters. Existing client deployments are based on bonding 25 Gbps ports, but other deployment options are possible too. Also, InfiniBand can be used in some cases to benefit from RDMA but without LPM capability.

1.3.1 Optimizing the network configuration for throughput on 10 Gbps by using SEA and jumbo frames

These instructions focus on throughput optimization occurring in SAP landscapes, for example, for backup or HANA scale-out deployments when using SEA.

Here are items to check before planning for jumbo frames:

- ▶ Deployments require a backbone that can support large packages end-to-end to avoid performance impacts.
- ▶ Just setting the maximum transmission unit (MTU) to 9000 is not sufficient. For more information, see “Configuring your Network for SAP HANA”, found at [SAP HANA on IBM Power Systems and IBM System Storage - Guides](#).
- ▶ When small packages are sent, jumbo-frame-enabled landscapes and MTU=1500 landscapes do not show a difference in performance.
- ▶ SAP does not use jumbo frames by default.

For 10 Gbps adapters in an environment that can use jumbo frames (that use an MTU of 9000), see “Configuring your Network for SAP HANA”, found at [SAP HANA on IBM Power Systems and IBM System Storage - Guides](#). This information is also applicable to higher speeds but has not been verified.

Background

There are certain metrics that are described in this section that control the packaging of the network packets.

The *MTU* is the maximum size of a single data unit of digital communications that can be transmitted over a network. The MTU size is an inherent property of a physical network interface, and it is measured in bytes. The default MTU for an Ethernet frame is 1500. An MTU of 9000 is referred to as a *jumbo frame*. The maximum segment size (MSS) is the maximum data payload for a socket, and it is derived from the MTU. For a TCP session, each peer announces the MSS during the 3-way handshake.

The implementation of jumbo frames in with Platform Large Send Offload (PLSO) is the only way to reduce the impact of processing and CPU cycles when large packages are transmitted to achieve a throughput of more than 9 Gbps on a 10 Gb adapter, as verified by the SAP Hardware Configuration Check Tool (HWCCT) for SAP HANA multi-node.

One major prerequisite for implementing jumbo frames is that all network components across the whole chain from sender to receiver can handle the large MTU settings. Hosts or networks that have an MTU setting of 1500 can become unreachable after setting the MTU to 9000. If the infrastructure does not allow for MTU 9000, the MTU size must remain as the default value.

Setting only the MTU is not sufficient. Other techniques to improve network throughput and lower CPU utilization are large send offload (LSO) and large receive offload (LRO), which must be implemented.

For outbound communication, LSO aggregates multiple packets into a larger buffer to the network interface. The network interface then splits the aggregated packets into separate packets according to the MTU size. The server cannot send frames that are larger than the MTU that is supported by the network. When LSO is disabled, the OS is responsible for breaking up data into segments according to the MSS. With LSO enabled, the OS can bypass data segmentation and send larger data chunks directly to the adapter device.

LRO is the counterpart of LSO for inbound communication. Multiple incoming packets from a single stream are aggregated into a larger buffer before they are passed up the networking stack, thus reducing the number of packets that must be processed.

If the network adapter supports LSO and is a dedicated adapter for the LPAR, the LSO option is enabled by default. Especially for data streaming workloads (such as FTP), RCP, backup, and similar bulk data movement), LSO can improve performance on 10-Gigabit Ethernet and faster adapters.

If the default MTU size of 1500 is used, PLSO is still beneficial, but maximum throughput on a 10 Gb adapter can be expected to be 6 - 8 Gbps. Without PLSO, it goes down to 3 - 4.5 Gbps for a single LPAR.

For virtual Ethernet adapters and SEA devices, LSO is disabled by default because of interoperability problems with older OS releases. This issue must be addressed if LSO is configured.

1.3.2 Latency optimization

If your focus is to reduce latency and omit virtualization, then SR-IOV port sharing is the preferred option. For more information, see the latest publications at the [IBM Redbooks website](#) because this feature is constantly evolving.

1.4 VIOS configuration for SAP network requirements

The sizing requirements for every VIOS deployment for SAP landscapes are as follows:

- ▶ Use only *two* or *four* VIOSs, no more or less.
- ▶ Start with a minimum of two dedicated cores for E-class servers with production workloads.
- ▶ Implement VIOS Core utilization monitoring to prevent I/O bottlenecks by undersized VIOSs (stand-alone or by using **saphostagent** that is deployed on VIOS).

- ▶ For FC virtualization, use only N_Port ID Virtualization (NPIV) because some infrastructure functions rely on it. Also, NPIV saves on adding more cores inside the VIOS and provides better latency.
- ▶ Start with 16 GB of memory per VIOS:
 - NPIV uses a minimum of 128 MB of memory per virtual connection. This memory is used by the hypervisor. Add the calculated memory for NPIV.
 - Other requirements might apply.
- ▶ NUMA placement and the locality of the VIOS to the adapter matters.
- ▶ PLSO helps in all cases to reduce CPU, not only when using jumbo frames.
- ▶ For the convenience of the sysadmin and to not lose any virtualization capabilities, install the VIOSs on internal solid-state drives (SSDs) or with POWER9 processor-based supported cards on Non-Volatile Memory Express (NVMe) (consider redundancy).

The investment into VIOS pays off well for your landscape:

- ▶ As you share physical adapters with multiple VIO clients or LPARs, you need fewer adapters.
- ▶ You have increased flexibility because you can add an LPAR as needed at any time because no new hardware must be added.
- ▶ Provides a faster response to changing circumstances.
- ▶ Pooling physical adapters at the VIOS results in higher bandwidth than assigning exclusive adapters per client LPAR.
- ▶ You have better DR and administration of resources.
- ▶ Facilitates LPM, Simplified Remote Restart, and DR.
- ▶ Reduces planned downtime for server administration to zero.



Storage system, connectivity, and file system architecture

This chapter describes SAP applications and their requirements for a file system that is flexible and reliable, and for databases that deliver sufficient performance.

SAP applications on Linux are typically run on Scalable File System (XFS). Some possible alternatives are IBM Spectrum® Scale (formerly known as IBM GPFS) and Network File System (NFS). These filer options provide high availability (HA) to the SAP and SAP HANA shared file systems. Some clients also use these filers for the SAP HANA data and log files systems.

This chapter describes the following topics:

- ▶ Filer infrastructures
- ▶ PCIe Non-Volatile Memory Express (NVMe) enriched POWER servers
- ▶ SAN infrastructures
- ▶ Fibre Channel infrastructure and VIOS options
- ▶ iSCSI boot disk attachment with VIOS 3.1
- ▶ Linux I/O

2.1 Filer infrastructures

You can access a filer through Ethernet or InfiniBand.

InfiniBand provides high throughput and low latency by using remote direct memory access (RDMA). Because high-speed network cards now are used in data centers, Ethernet attachments also are used in SAP deployments.

Although InfiniBand cannot be virtualized, Ethernet can be virtualized (when it is not using RDMA over Converged Ethernet (RoCE)). For more information about Ethernet virtualization, see Chapter 1, “Ethernet architectures for SAP workloads” on page 1.

Filers provide three file systems that are relevant to SAP landscapes:

- ▶ NFS (for example, NetApp)
- ▶ IBM Spectrum Scale (for example, IBM Elastic Storage® Server (IBM ESS))
- ▶ Hadoop file system (for example, IBM ESS)

SAP NetWeaver can use filer-based shared file systems for `/sapmnt` and `/usr/sap/trans`. Especially in HA scenarios, filer-based shared file systems have higher availability with less operational cost compared to the traditional NFS cross-mount or software options.

SAP HANA can be deployed on IBM ESS or any other SAN HANA Tailored Data Center Integration (TDI) certified filer as a persistent back end. For more information about setting up and configuring this option, see the SAP HANA TDI documentation.

Regarding filers, the boot devices need special treatment. One option is to use internet Small Computer Systems Interface (iSCSI), which outlined in 2.5, “iSCSI boot disk attachment with VIOS 3.1” on page 30.

2.2 PCIe Non-Volatile Memory Express (NVMe) enriched POWER servers

Since IBM POWER8® processors were released, PCIe-attached NVMe cards are used in the field to accelerate predominately read I/O. This section focuses on highlights only, and does not cover all the options where NVMe drives play a role.

For more information about the available options for your Power Systems server model, see the e-config tool or contact your IBM representative or IBM Business Partner. Pay attention to the I/O characteristics of the NVMe cards because there are a broad variety of them based on endurance (determines whether this card is suitable for storing the data persistence of a database), speed, and capacity. With low-end cards, a SAN Volume Controller outperforms them and provide operational benefits that an internal solution cannot. Using internal disks eliminates in most cases Live Partition Mobility (LPM) capability because the data is bound to a single Power Systems server unless you use the disks as a read cache in a Shared Storage Pool (SSP) Virtual I/O Server (VIOS) deployment.

SAP-related documentation about how to configure NVMe along with other SAP HANA documentation can be found at [SAP HANA on IBM Power Systems and IBM System Storage - Guides](#).

2.2.1 NVMe use cases and technology

Although it is possible to use NVMe from an endurance point of view, the dominant use cases are to use them as fast cache.

Here are generic use cases of NVMe¹:

- ▶ VIOS boot image on NVMe: Customers can use an NVMe device to install and boot a VIOS image. Transferring VIOS boot images to an NVMe device can be done by using a Logical Volume Manager (LVM) mirror. Customers can add an NVMe mirror copy to rootvg and remove the old copy after a sync is done.
- ▶ Logical volume (LV)-backed virtual SCSI (VSCSI) device: Customers can install a NovaLink boot image on the device (An LV-backed device can be used to boot a NovaLink partition in greenfield deployment). A client logical partition (LPAR) can use the LV-backed device, which is on an NVMe volume group (VG), to host the read cache.
- ▶ Read cache device on VIOS: An NVMe device is perfect for the local read cache on VIOS. It can be used for SSP disk caching where data that is present in the SSP is cached on the NVMe disk. LPM with SSP where NVMe devices are used for caching is possible.
- ▶ No limitation on SSP operations: When you enable SSP caching that uses an NVMe disk, you can perform any kind of SSP operations, such as adding, removing, or replacing a disk to or from SSP; creating, modifying, or deleting a tier in SSP; and creating and deleting a mirror in SSP.
- ▶ No dependency on type of disk for client: You can create a VG that can spread across NVMe and some other type of devices. You can create an LV that can spread across NVMe and other devices. But for the client, the LV appears as a normal vSCSI disk even though the LV is spread between the NVMe disk and the other disk.
- ▶ Backup and restore of a VIOS configuration: You can create back ups of VIOS instances with NVMe disks, install a new VIOS build, and restore the configuration on the new VIOS build.
- ▶ Upgrade support from previous VIOS levels: You can upgrade VIOS instances from an older level to a new level and start using the NVMe device at the new level directly.

Db2

Depending on the workload and acceleration of the temporary file system of IBM Db2®, it can speed up processing by up to 30%.

SAP HANA

NVMe cards for SAP HANA can be used as an internal disk option and an accelerator for all read operations. Read acceleration has value when you restart the database, activate a standby node in an SAP HANA Auto Host Failover scenario, and with data tiering options.

For more information, see [SAP HANA on IBM Power Systems and IBM System Storage - Guides](#).

¹ Virtualization of NVMe adapters on IBM POWER9 processor-based systems, found at: <https://developer.ibm.com/articles/au-aix-virtualization-nvme/>

Notes:

- ▶ Because the NVMe adapters are attached locally to the Power Systems servers, LPM is no longer possible unless NVMe is used as cache acceleration in SSP configurations.
- ▶ NVMe performance tuning is different from SAN-based tuning. For example, PCIe NVMe does not perform well when using RAID 5 or RAID 6 configurations compared to RAID 1 or RAID 10. When using NVMe mostly as cache, no RAID protection is required, which boosts bandwidth and performance.
- ▶ Performance scales with the number of physical cards that is used and not with the number of modules on a single card. For up to four cards, the performance scales linearly. No testing was performed with more cards, but performance further increases if the file system is set up in a striped manner and a sufficient workload is run against it.

2.3 SAN infrastructures

This section describes best practices for the deployment of SAP on SAN.

2.3.1 SAN use cases and architectures best practices

The dominant type of attachment in SAP landscapes is using Fibre Channel (FC) to connect to storage devices, as described in 2.4, “Fibre Channel infrastructure and VIOS options” on page 18).

Today, most SAN designs use a variant of what is called a *core-to-edge SAN* design, as shown in Figure 2-1. In this design, the SAN elements (typically SAN switches) are designated as either core or edge switches. The edge switches connect servers, and the core switches connect the storage devices and the edge switches.

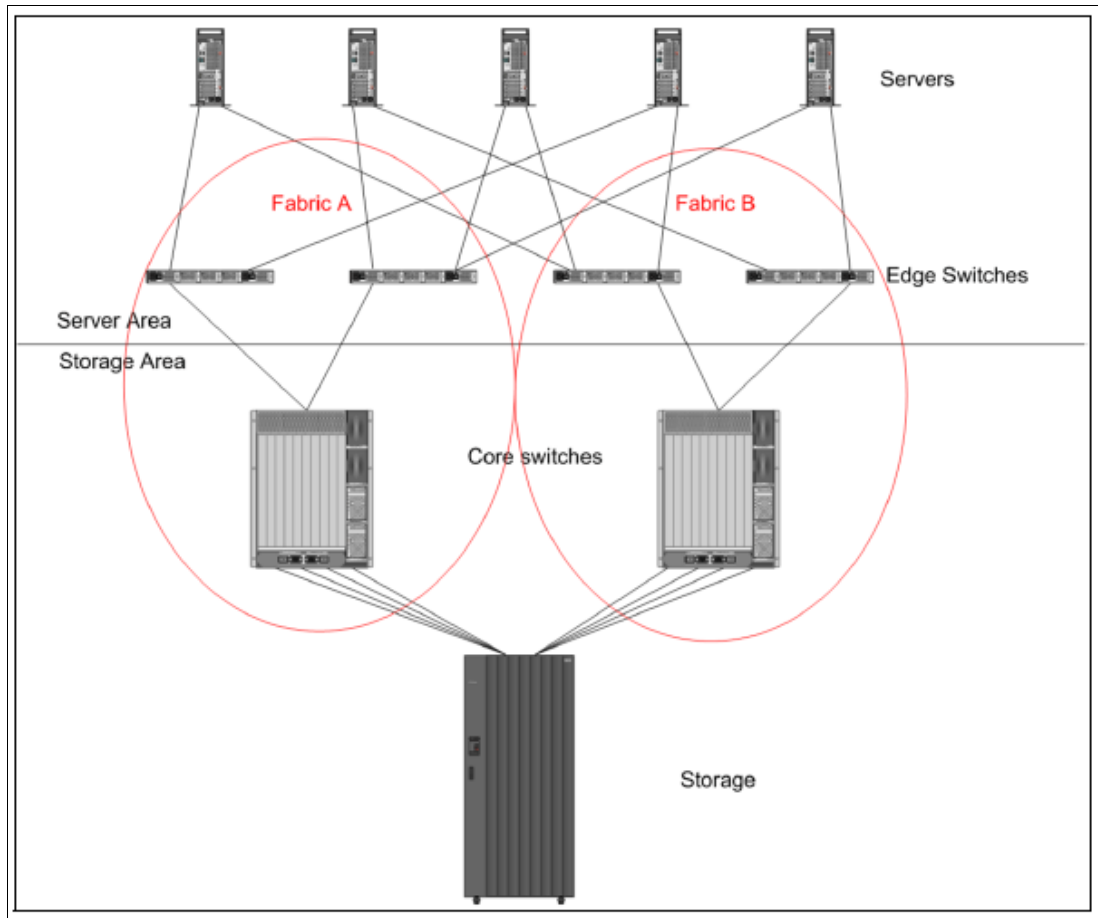


Figure 2-1 Highly redundant core-to-edge SAN architecture

For HA, the setup is divided into two fabrics that act as a failover configuration. All connections from the servers to the edge switches are handled by both fabrics. The storage systems are also connected to both fabrics.

This setup ensures that there are multiple independent paths from the servers to the storage systems. The maximum number of paths depends on the storage system, but must not be more than 16 for one server.

The number of allowed connections differs for different storage systems. Here are two examples, and both examples are sufficient to run SAP workloads. Configurations are listed so that you make the correct plans when ordering switch and server hardware:

- ▶ IBM XIV: Up to 12 FC connections to the core switches
- ▶ IBM SAN Volume Controller: Up to 16 FC connections to the core switches

SAN zoning is used to keep the servers isolated from each other and manage which server can reach which storage volumes on the storage systems.

There are multiple ways for a SAN zoning implementation, but the best practice method is to do the zoning by initiator port, that is, create zones for individual initiator ports (typically a single-server port) with one or more target (storage system) ports².

To configure the zone, use only the worldwide port name (WWPN), *not* the worldwide node name (WWNN). When a WWNN is used in place of a WWPN for a device, switches interpret the WWNN as designating *all* associated ports for the device. Using the WWNN in a zone can cause multipathing issues where there are too many paths between server and storage.

2.4 Fibre Channel infrastructure and VIOS options

There are different connection types that are available for an FC connection that uses VIOS:

- ▶ vSCSI
- ▶ N_Port ID Virtualization (NPIV)

The recommended connection type is NPIV because it supports LPM without manual intervention. NPIV lowers latency and reduces the core consumption on the VIOS level. Some SAP solutions such as SAP HANA Auto Host Failover or third-party products require NPIV.

2.4.1 vSCSI

vSCSI is based on a client/server relationship. The VIOS owns the physical resources and acts as the server, or in SCSI terms the target device. The client LPARs access the vSCSI backing storage devices that are provided by the VIOS as clients.

Interaction between client and server

The virtual I/O adapters are configured by using a Hardware Management Console (HMC) or the Integrated Virtualization Manager on smaller systems. The interaction between a VIOS and a Linux client partition is enabled when both the vSCSI server adapter that is configured in the VIOS' partition profile and the vSCSI client adapter that is configured in the client partition's profile have mapped slot numbers, and both the VIOS and client operating system (OS) recognize their virtual adapter.

Dynamically added vSCSI adapters are recognized on the VIOS after running the `cfgdev` command. Linux OSs automatically recognize dynamically added vSCSI adapters.

After the interaction between the vSCSI server and the vSCSI client adapters is enabled, mapping storage resources from the VIOS to the client partition is needed. The client partition configures and uses the storage resources when it starts or when it is reconfigured at run time.

The process runs as follows:

- ▶ The HMC maps interaction between vSCSI adapters.
- ▶ The mapping of storage resources is performed in the VIOS.
- ▶ The client partition recognizes the newly mapped storage dynamically.

² IBM Support SAN Zoning Best Practices, found at:
<https://www.ibm.com/support/pages/san-zoning-best-practices>

Redundancy of vSCSI by using VIOS

Figure 2-2 shows one possible configuration in a PowerVM environment that shows the redundancy of vSCSI by using multipath I/O (MPIO) at client partitions. The diagram shows a dual-VIOS environment where the client partition has two vSCSI client adapters and each of them is mapped to two different vSCSI server adapters on different VIOSs. Each VIOS maps the same physical volume (PV) to the vSCSI server adapter on them.

The client partition sees the same PV (hdisk in Figure 2-2), which is mapped from two VIOSs by using vSCSI. To achieve this mapping, the same storage must be zoned to the VIOSs from the storage subsystem. This configuration also has redundancy at the VIOS physical FC adapter.

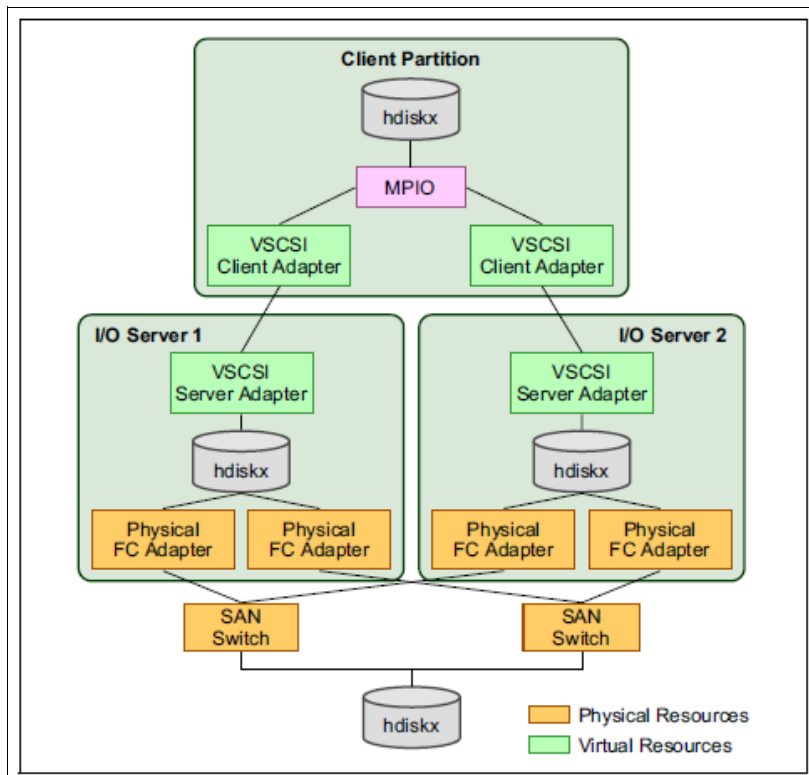


Figure 2-2 Redundancy of vSCSI using dual virtual I/O servers

2.4.2 N_Port ID Virtualization

NPIV is an industry-standard technology that allows an NPIV-capable FC adapter to be configured with multiple virtual WWPNs, as shown in Figure 2-3. This technology is also called *virtual FC*. Similar to the vSCSI function, virtual FC is another way of securely sharing a physical FC adapter among multiple VIOS client partitions.

From an architectural perspective, the key difference with virtual FC compared to vSCSI is that the VIOS does not act as a SCSI emulator to its client partitions, but acts as a direct FC pass-through for the FC Protocol I/O traffic through the IBM POWER Hypervisor. Instead of generic SCSI devices presented to the client partitions with vSCSI, with virtual FC the client partitions are presented with native access to the physical SCSI target devices of SAN disk or tape storage systems.

The benefit of virtual FC is that the physical target device characteristics like vendor or model information remain fully visible to the VIOS client partition so that device drivers like multipathing software, middleware such as copy services, or storage management applications that rely on the physical device characteristics do not need to be changed.

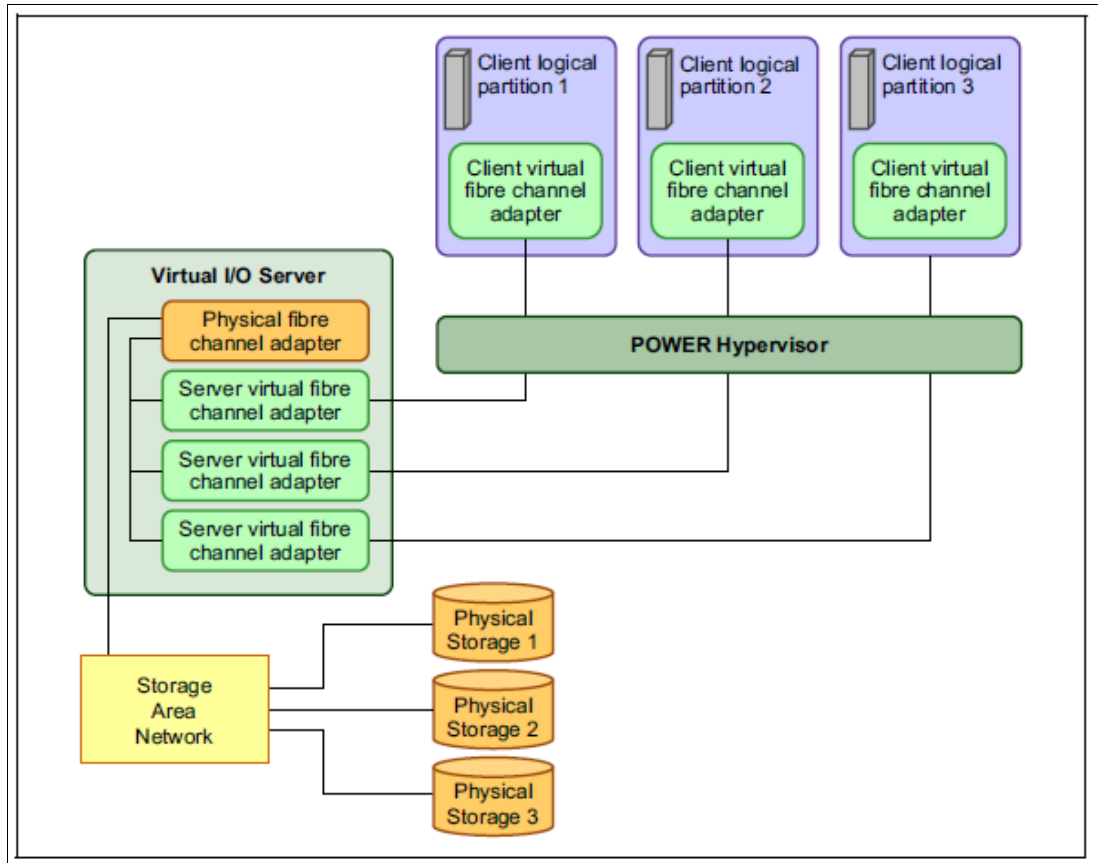


Figure 2-3 NPIV architecture

Redundancy of virtual Fibre Channel

A host bus adapter and VIOS redundancy configuration provides a more advanced level of redundancy for the virtual I/O client partition, as shown in Figure 2-4.

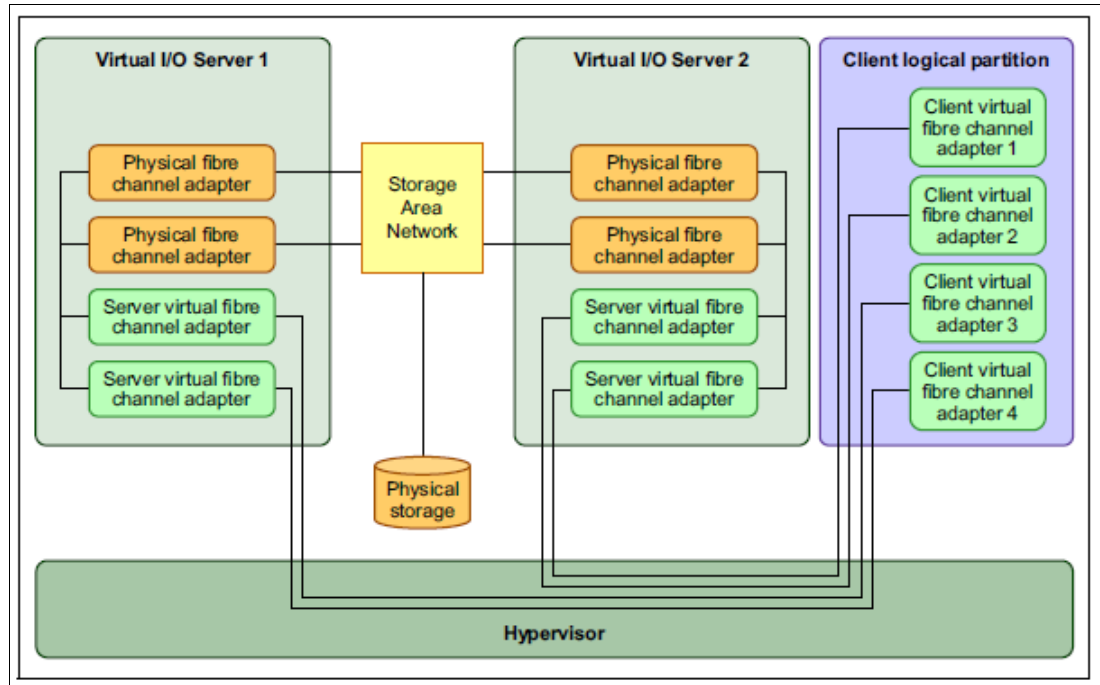


Figure 2-4 Redundancy of virtual Fibre Channel

2.4.3 Example setup of an NPIV virtual Fibre Channel configuration

This section describes how to configure SAN storage devices by using virtual FC for a Linux client of the VIOS. An IBM 2498-F48 SAN switch, an IBM Power Systems E980 server, and an IBM Spectrum Virtualize storage system were used in the lab environment to describe the setup of the virtual FC environment.

Complete the following steps:

1. Use a dedicated virtual FC server adapter (slot P1-C2-C1) in the VIOS partition ish400v1, as shown in Figure 2-5. Each client partition accesses physical storage through its virtual FC adapter, which must be configured in the profiles of the VIOS and the client.

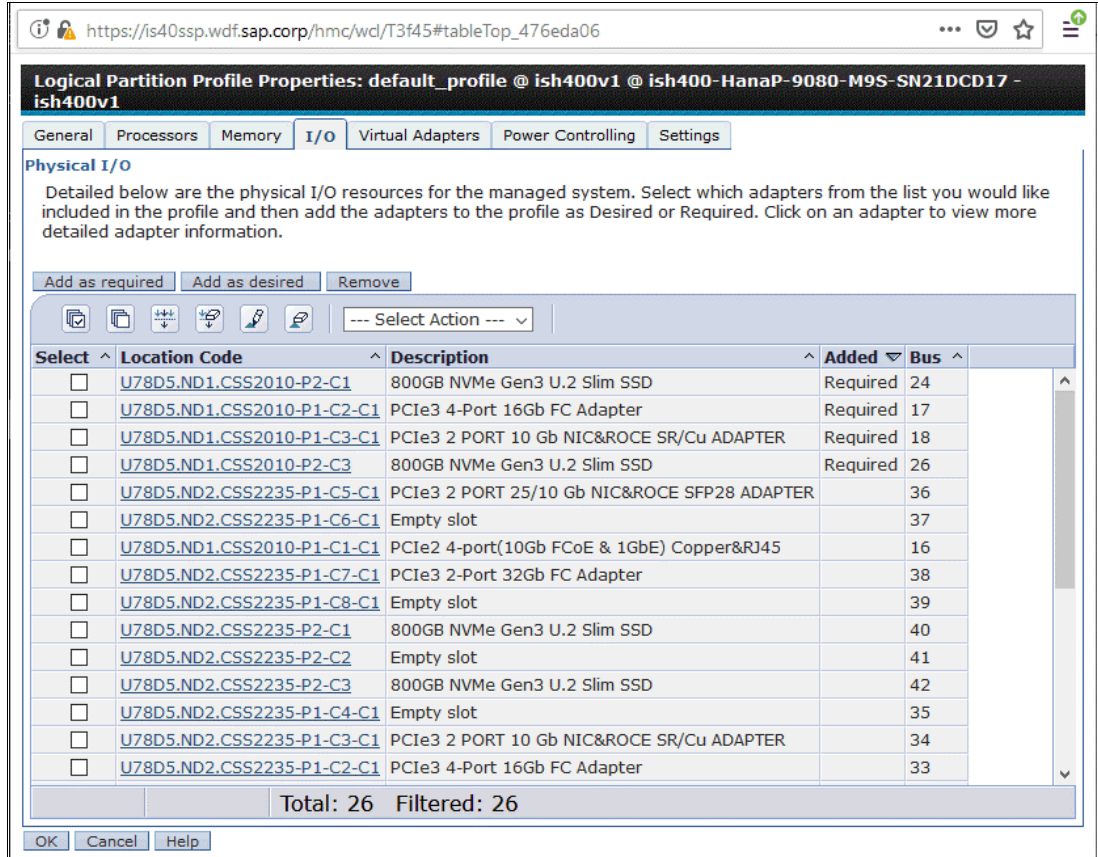


Figure 2-5 LPAR properties on HMC

2. Create the virtual FC server adapter in the VIOS partition:
 - a. On the HMC, select the managed server to be configured by clicking **All Systems** and then select <servername> (ish4001v1).
 - b. Select the VIOS partition on which the virtual FC server adapter will be configured. Then, select **Actions** → **Profiles** → **Manage Profiles**, as shown in Figure 2-6 on page 23.

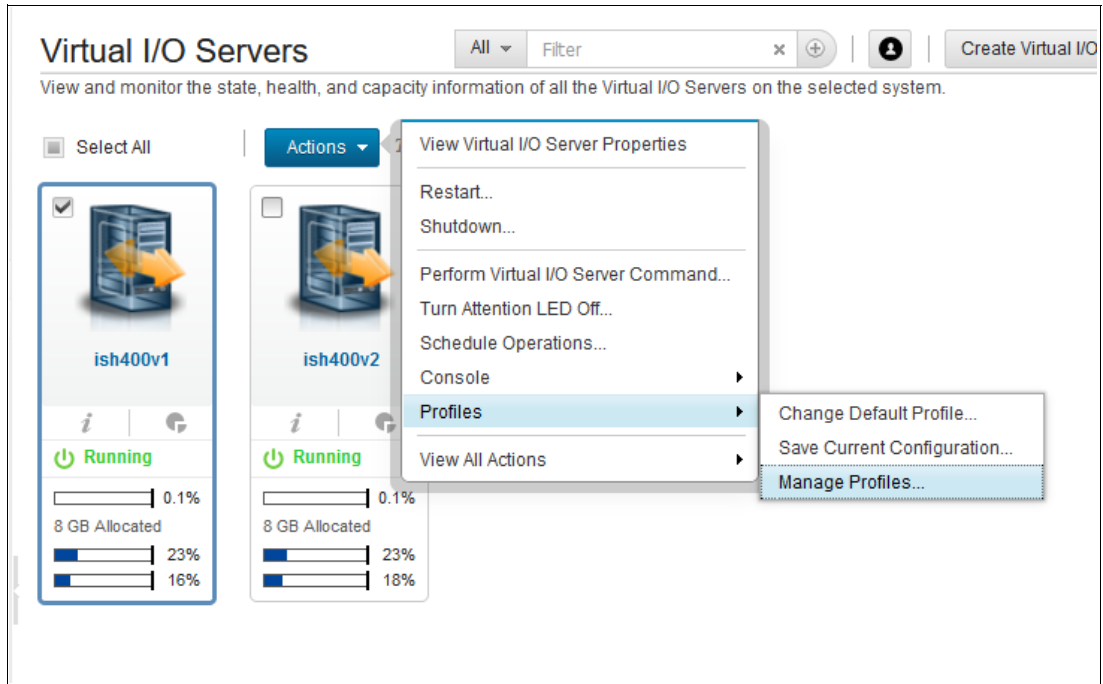


Figure 2-6 VIOS profile on the HMC

- c. To create a virtual FC server adapter, select the profile to use and open it by using the **Edit** action. Then, click the **Virtual Adapters** tab and select **Actions** → **Create Virtual Adapter** → **Fibre Channel Adapter**, as shown in Figure 2-7.

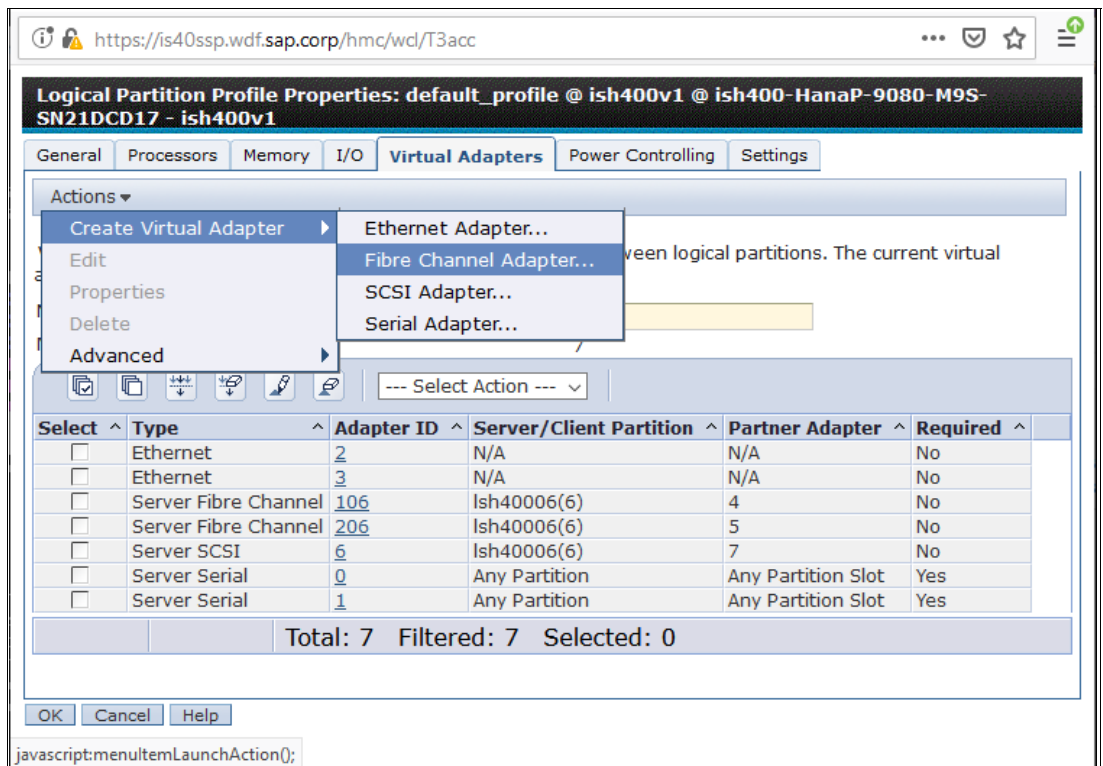


Figure 2-7 Creating a virtual Fibre Channel adapter on the HMC

- d. Enter the virtual FC adapter number for the virtual FC server adapter. Then, select the client partition to which the adapter can be assigned and enter the client adapter ID, as shown in Figure 2-8. Click **OK**.

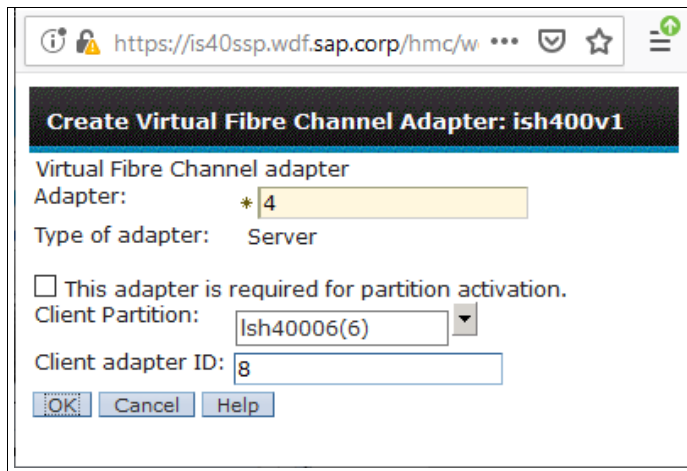


Figure 2-8 Adding the virtual Fibre Channel number on HMC

- e. Click **OK** in the Create Virtual FC Adapters dialog to save the changes.
- f. Update the partition profile of the VIOS partition by selecting **Profiles** → **Save Current Configuration**, as shown in Figure 2-9 to save the changes.

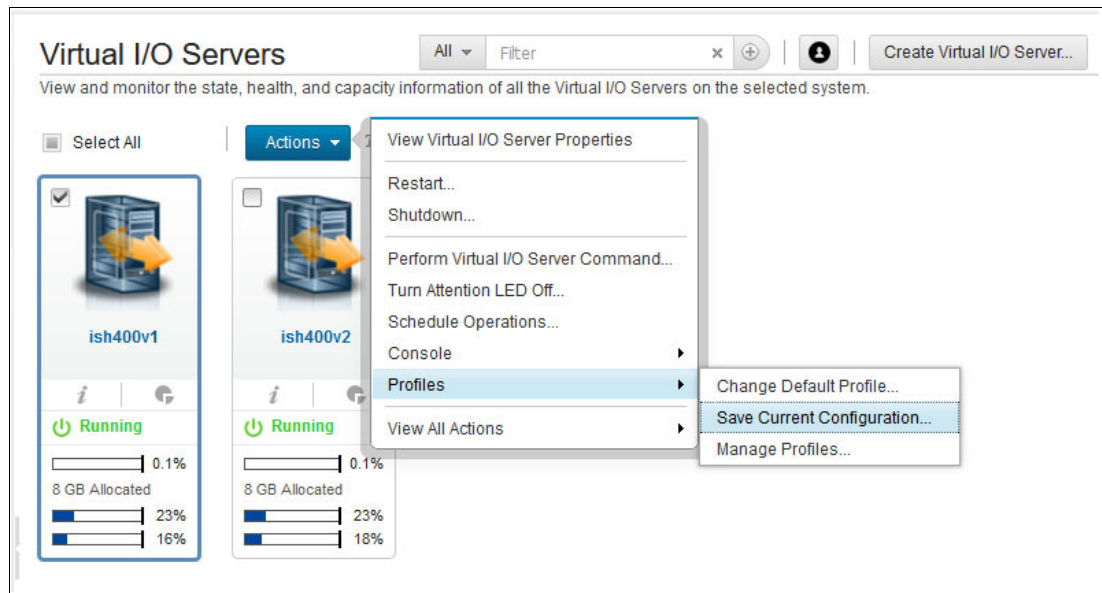


Figure 2-9 Save Current Configuration menu on the HMC

3. Create a virtual FC client adapter in the virtual I/O client partition:
 - a. Select the virtual I/O client partition on which the virtual FC client adapter will be configured. Change the partition profile by selecting **Profiles** → **Manage Profiles**, as shown in Figure 2-10.

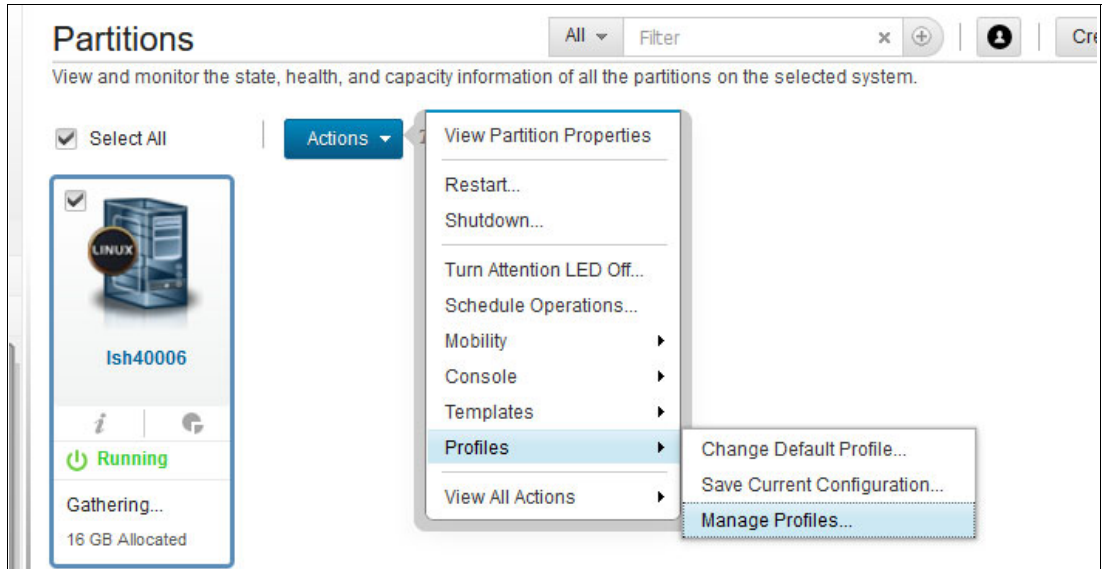


Figure 2-10 Selecting the virtual I/O client partition for the virtual Fibre Channel

- b. Click the profile name to edit it and select the **Virtual Adapters** tab in the Logical Partition Profile Properties dialog. Then, to create a virtual FC client adapter, select **Actions** → **Create Virtual Adapter** → **Fibre Channel Adapter**, as shown in Figure 2-11.

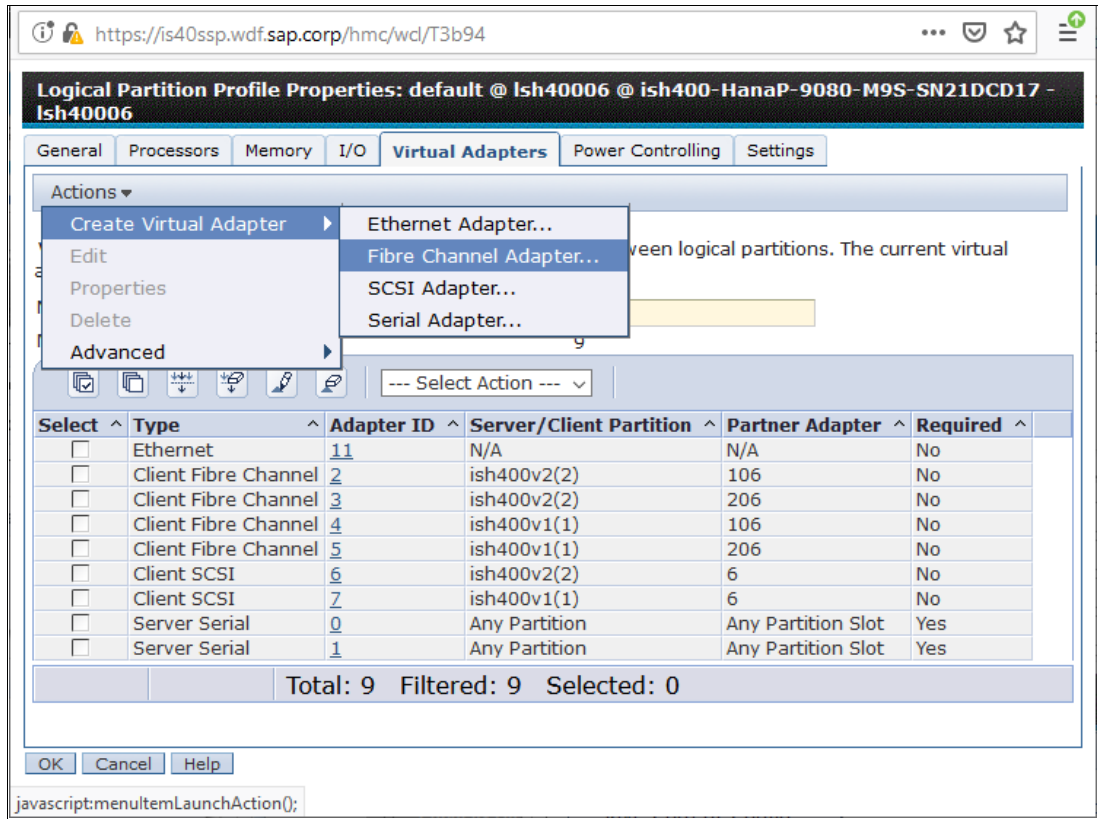


Figure 2-11 Assigning the virtual port on the HMC

- c. Enter the virtual slot number for the virtual FC client adapter. Then, select the VIOS partition to which the adapter can be assigned and enter the server adapter ID, as shown in Figure 2-12. Click **OK**.

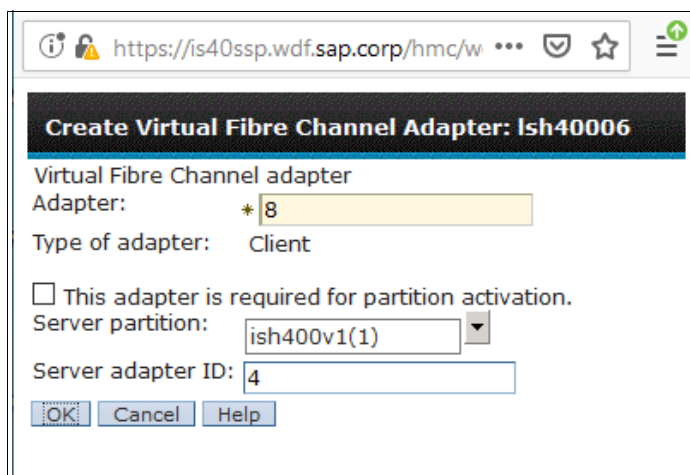


Figure 2-12 Adding a server adapter ID

- d. Click **OK** and then **Close** in the Managed Profiles dialog to save the changes.

4. Log in to the VIOS partition as user padmin.
5. Run the **cfgdev** command to configure the virtual FC server adapters.
6. Run the **lsdev -dev vfchost*** command to list all the available virtual FC server adapters in the VIOS partition before mapping to a physical adapter, as shown in Example 2-1.

Example 2-1 The lsdev -dev vfchost command on the Virtual I/O Server*

```
$ lsdev -dev vfchost*
name          status      description
vfchost0     Available  Virtual FC Server Adapter
vfchost1     Available  Virtual FC Server Adapter
```

7. The **lsdev -dev fcs*** command lists all available physical FC server adapters in the VIOS partition as shown in Example 2-2.

Example 2-2 The lsdev -dev fcs command on the Virtual I/O Server*

```
$ lsdev -dev fcs*
name          status      description
fcs0          Available  PCIe3 4-Port 16Gb FC Adapter (df1000e314101406)
fcs1          Available  PCIe3 4-Port 16Gb FC Adapter (df1000e314101406)
fcs2          Available  PCIe3 4-Port 16Gb FC Adapter (df1000e314101406)
fcs3          Available  PCIe3 4-Port 16Gb FC Adapter (df1000e314101406)
```

8. Run the **lsnports** command to check the virtual FC adapter readiness of the adapter and the SAN switch. Example 2-3 shows that the fabric attribute for the physical FC adapter in slot C1 is set to 1, which means that the adapter and the SAN switch are NPIV ready. If the value equals 0, then the adapter or SAN switch is *not* NPIV ready, and you must check the SAN switch configuration.

Example 2-3 The lsnports command on the Virtual I/O Server

```
$ lsnports
name  physloc          fabric tports  aports  swwpns  awwpns
fcs0  U78D5.ND1.CSS2010-P1-C2-C1-T1  1    64     63     3072   3069
fcs1  U78D5.ND1.CSS2010-P1-C2-C1-T2  1    64     63     3072   3069
```

9. Before mapping the virtual FC adapter to a physical adapter, obtain the **vfchost** name of the virtual adapter that you created and the **fcs** name for the FC adapter from the output of Example 2-2.
10. To map the virtual FC server adapter **vfchost0** to the physical FC adapter **fcs0**, use the **vfcmmap** command, as shown in Example 2-4.

Example 2-4 The vfcmmap command with vfchost0 and fcs0

```
$ vfcmmap -vadapter vfchost0 -fcp fcs0
vfchost0 changed
```

11. To list the mappings, use the **lsmmap -all -npiv** command, as shown in Example 2-5.

Example 2-5 The lsmmap -npiv -vadapter vfchost0 command

```
$ lsmmap -all -npiv
Name          Physloc          C1ntID C1ntName          C1ntOS
-----
vfchost0     U9080.M9S.21DCD17-V1-C206      6 lsh40006      Linux

Status:LOGGED_IN
```

FC name:fcs0 FC loc code:U78D5.ND1.CSS2010-P1-C2-C1-T1
 Ports logged in:3
 Flags:a<LOGGED_IN,STRIP_MERGE>
 VFC client name:host5 VFC client DRC:U9080.M9S.21DCD17-V6-C5

Name	Physloc	CIntID	CIntName	CIntOS
vfchost1	U9080.M9S.21DCD17-V1-C106	6	1sh40006	Linux

Status:LOGGED_IN
 FC name:fcs1 FC loc code:U78D5.ND1.CSS2010-P1-C2-C1-T2
 Ports logged in:3
 Flags:a<LOGGED_IN,STRIP_MERGE>
 VFC client name:host4 VFC client DRC:U9080.M9S.21DCD17-V6-C4

12. After you create the virtual FC server adapters in the VIOS partition and in the virtual I/O client partition, set the correct zoning in the SAN switch:

- a. Obtain the information about the WWPN of the virtual FC client adapter that was created in the virtual I/O client partition.
- b. Select the appropriate virtual I/O client partition, then from the task menu click **Properties**. Expand the **Virtual Adapters** tab, select the Client FC client adapter, and then select **Actions** → **Properties** to list the properties of the virtual FC client adapter, as shown in Figure 2-13.

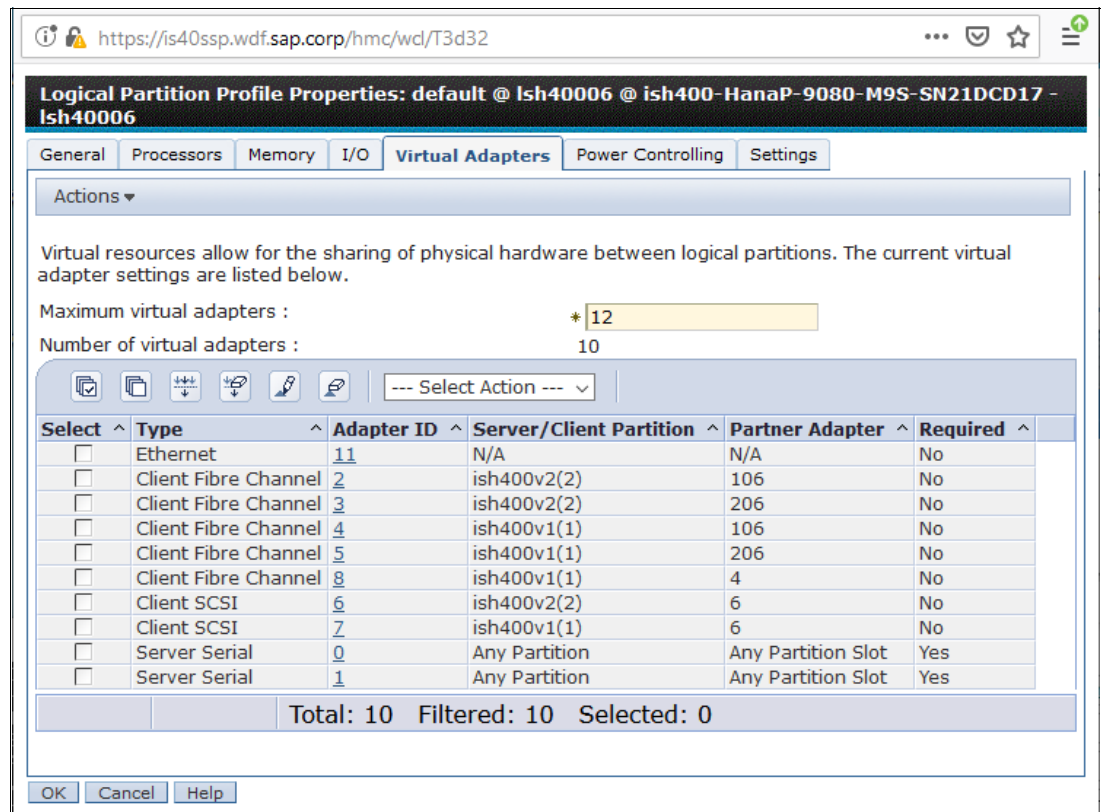


Figure 2-13 Zone the LPAR and virtual adapter

- c. Figure 2-14 on page 29 shows the properties of the virtual FC client adapter. Here you can get the virtual WWPN that is required for the zoning.

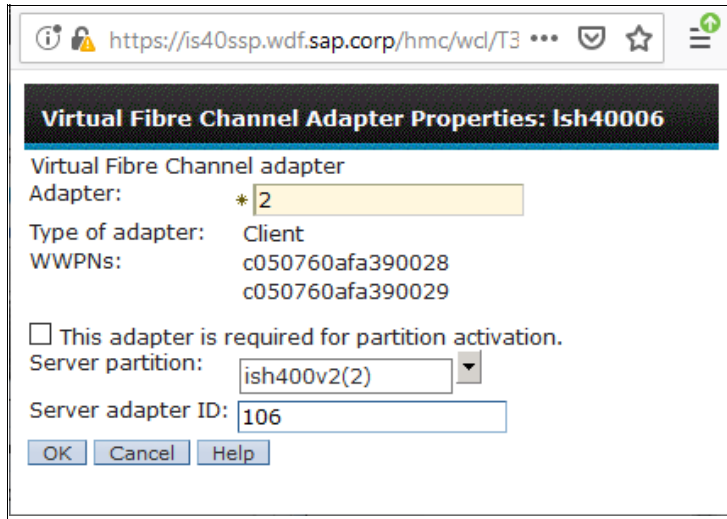


Figure 2-14 Getting the virtual WWPN that is required for zoning

- d. Log on to your SAN switches and create a zone for the virtual WWPN and the corresponding physical storage ports, or customize an existing one.

Only the first listed WWPN is used for the running LPAR and is considered for the SAN zoning and storage configuration. The second listed WWPN is inactive and used for LPM when the partition is moved to another system. After the move to another system, the second WWPN is the active one and the first WWPN is inactive.

- e. After completing the SAN switch zoning, create the storage configuration on your SAN storage system by mapping the LUNs to a host connection that is created with the virtual WWPN of the virtual FC client adapter.
- f. After completing the SAN storage configuration, the volumes that are configured in the virtual FC client adapter are now ready for use by the VIOS client partition.

From the Linux client perspective, virtual FC must look like a native physical FC device. There is no special requirement or configuration that is needed to set up a virtual FC on Linux.

After the `ibmvfc` driver is loaded and a virtual FC Adapter is mapped to a physical FC adapter on the VIOS, the FC port automatically shows up on the Linux partition. You can check whether the `ibmvfc` driver is loaded on the system by running the `lsmod` command, as shown in Example 2-6.

Example 2-6 Checking for the `ibmvfc` driver

```
[root@lsh40006: ~]# lsmod | grep ibmvfc
ibmvfc                79236  288
scsi_transport_fc     68048  1 ibmvfc
scsi_mod              293836  12
scsi_dh_emc,st,scsi_transport_srp,sd_mod,scsi_dh_alua,scsi_dh_rdac,ibmvfc,sr_mod,d
m_multipath,sg,ibmvscsi,scsi_transport_fc
```

You can also check the devices by looking at the kernel log in the `/var/log/messages` file or by running the `dmesg` command, as shown in Example 2-7.

Example 2-7 Checking `/var/log/messages` for the loaded driver

```
[root@lsh40006: ~]# dmesg | grep vfc
[ 1.932844] ibmvfc: externally supported module, setting X kernel taint flag.
```

```
[ 1.932858] ibmvfc: IBM Virtual Fibre Channel Driver version: 1.0.11 (April 12, 2013)
[ 2.036692] ibmvfc 30000002: Partner initialization complete
[ 2.040987] ibmvfc 30000002: Host partition: ish400v2, device: vfchost1
U78D5.ND2.CSS2235-P1-C2-C1-T2 U9080.M9S.21DCD17-V2-C106 max sectors 8192
[ 2.125175] ibmvfc 30000003: Partner initialization complete
[ 2.129065] ibmvfc 30000003: Host partition: ish400v2, device: vfchost0
U78D5.ND2.CSS2235-P1-C2-C1-T1 U9080.M9S.21DCD17-V2-C206 max sectors 8192
[ 2.195108] ibmvfc 30000004: Partner initialization complete
[ 2.198965] ibmvfc 30000004: Host partition: ish400v1, device: vfchost1
U78D5.ND1.CSS2010-P1-C2-C1-T2 U9080.M9S.21DCD17-V1-C106 max sectors 8192
[ 2.255039] ibmvfc 30000005: Partner initialization complete
[ 2.258851] ibmvfc 30000005: Host partition: ish400v1, device: vfchost0
U78D5.ND1.CSS2010-P1-C2-C1-T1 U9080.M9S.21DCD17-V1-C206 max sectors 8192
```

To list the virtual FC device, run the `lsscsi` command, as shown in Example 2-8.

Example 2-8 Listing the Fibre Channel devices

```
[root@lsh40006: ~]# lsscsi -H -v | grep fc
[2]    ibmvfc
[3]    ibmvfc
[4]    ibmvfc
[5]    ibmvfc
```

You can perform virtual FC tracing on Linux through the file system attributes in the `/sys/class` directories. The files containing the devices' attributes are useful for checking detailed information about the virtual device and also can be used for troubleshooting. These attributes files can be accessed in the following directories:

- ▶ `/sys/class/fc_host/`
- ▶ `/sys/class/fc_remote_port/`
- ▶ `/sys/class/scsi_host/`

2.5 iSCSI boot disk attachment with VIOS 3.1

For many years, FC-attached storage has been the data transmission technology of choice. It provides high reliability, high throughput, and low-latency storage access at moderate costs.

iSCSI provides block-level access to storage devices by carrying SCSI commands over a Internet Protocol network. iSCSI facilitates data transfers over the internet by using TCP, which is a reliable transport mechanism that uses either IPv6 or IPv4 protocols. TCP is used to manage storage over long distances.

Compared to FC-attached storage, iSCSI storage systems can be a cheaper option. Regarding infrastructure costs, iSCSI is less expensive because all the existing Ethernet infrastructure (network switches, host adapters, and network interface cards (NICs)) can be used with host-server-based iSCSI initiator software without needing extra FC adapters and SAN directors or switches. However, this situation is not a fair comparison because of performance and other considerations.

Some organizations are turning to iSCSI storage systems because they consider it a less expensive data transmission option because no extra FC components must be procured and operated. In the case of moderate performance and throughput requirements, for example, for system disk access or file services, iSCSI-attached storage can be a viable option.

With VIOS 3.1, iSCSI support was added to the VIOS, which you can use to export the iSCSI disks to client LPARs as virtual disks (vSCSI disks). This support is available in VIOS 3.1 and requires FW 860.20 or later. VIOS 3.1 also enables MPIO support for the iSCSI initiator so that you can configure and create multiple paths to an iSCSI disk (Figure 2-15).

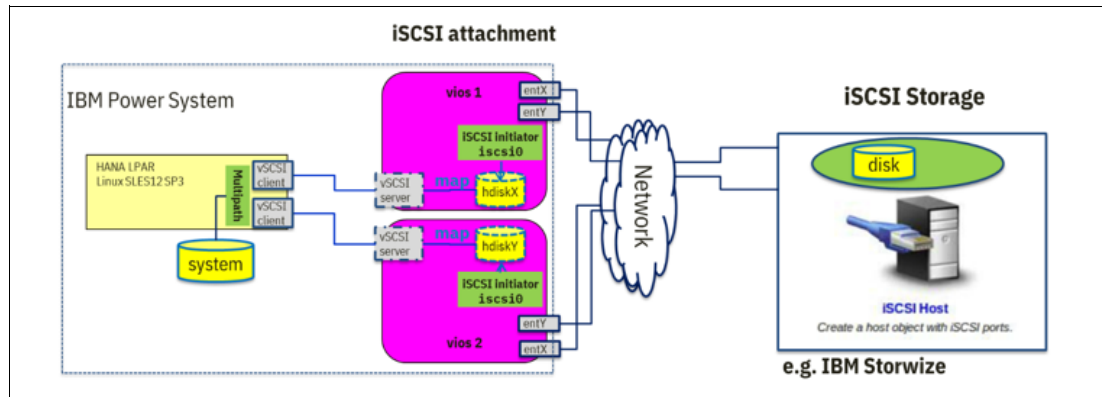


Figure 2-15 iSCSI boot architecture

Currently, the iSCSI disk support for VIOS has the following limitations:

- ▶ There is no support for booting VIOS by using an iSCSI disk. Instead, internal disks can be used for VIOS because a VIOS is always hardbound to a server and needs no LPM capability.
- ▶ Flat file-based discovery policy is not supported. The recommendation is to use discovery policy ODM.
- ▶ iSCSI disk-based LV backed devices are not supported. SSPs that use iSCSI disks as either repositories or shared pool disks are not supported. The iSCSI disks or iSCSI-based LVs or VGs cannot be used as paging devices for Active Memory Sharing (AMS) or remote restart.
- ▶ If the backing device is an iSCSI disk, the `client_reserve` and `mirrored` attributes are not supported for virtual target devices.

Several steps must be done to configure iSCSI and the iSCSI storage system on the VIOS, as shown in Figure 2-16.

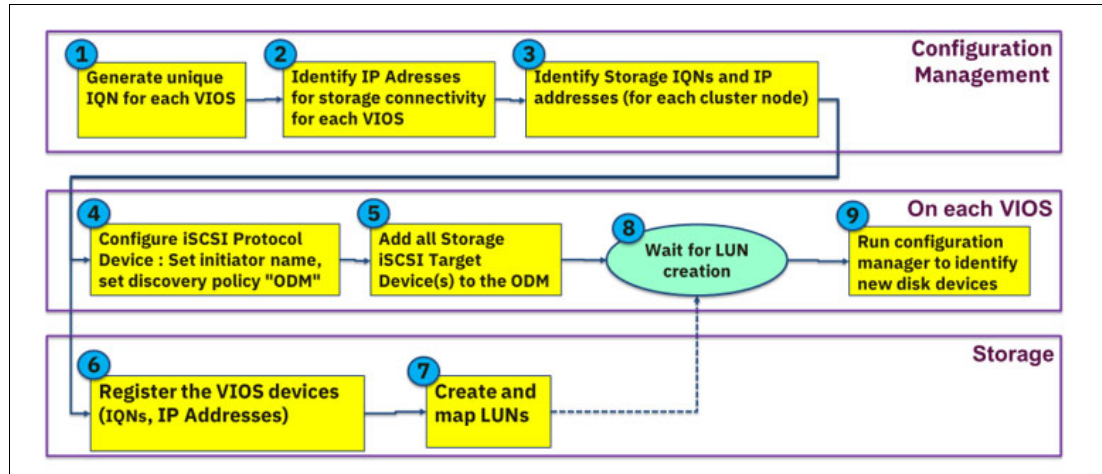


Figure 2-16 Overview of the iSCSI configuration flow

At first, all the configuration parameters must be defined: iSCSI Qualified Names (IQNs) and the IP addresses:

1. The default IQN for the iSCSI software initiator on the VIOS does not match all the IQN standards. A unique IQN must be defined for the iSCSI software initiator.
2. On each VIOS, two IP addresses (in different IP subnets) that are acquired on different network adapters must be created by forming a *private storage network* on high-bandwidth network adapters.
3. The IP addresses of the storage system need to be created as well.

The configuration actions on the storage system and the VIOS start.

On the VIOS, complete the following steps:

1. Create an iSCSI protocol device on each VIOS.
The iSCSI initiator name and the discovery policy are defined in the iSCSI protocol device. The discovery methodology must be set to odm. The information about the iSCSI targets is then stored in the Object Data Manager (ODM) objects.
2. Add all iSCSI targets to the ODM.

On the storage system, complete the following steps:

1. Define all the IQNs of the VIOSs.
2. Create all the LUNs and map them to the VIOSs.

After the LUNs are created, the VIOS team should complete the following steps:

1. Wait for LUN creation.
2. Run the configuration manager to discover the new disk devices.
3. Map the new disk devices to the client LPARs by using the vSCSI mappings.

Actions on the storage system depend on the vendor, and those actions are not covered here. The required steps on the VIOS are described in more detail in the next section.

2.5.1 Configuring iSCSI on the VIOS

This section describes the iSCSI configuration on the VIOS.

Defining a unique iSCSI qualified name for each VIOS

For every iSCSI node, a node name that uses the IQN format must be set. The IQN-type designator is a logical name and has the following format:

```
iqn.<yyyy-mm>.<naming-authority>:<unique name>
```

<yyyy-mm> Year and month when the naming authority is established.

naming-authority The naming authority is built on the reverse of the internet domain name of the naming authority.

unique name Unique identifier for the iSCSI VIOS. The naming authority must make sure that any names that are assigned following the colon are unique.

Identifying IP addresses for storage connectivity for each VIOS

To get optimal performance from the iSCSI disks, establish the following items:

- ▶ A separate private network to access the iSCSI storage.
- ▶ High-speed network adapters and switches (at least 10 Gb Ethernet technology).
- ▶ A redundant network topology on the storage system and on the VIOS. Two IP addresses on different network adapters for each VIOS, connecting to two IP addresses on different storage cluster nodes.

Configuring the iSCSI protocol device on each VIOS

Log in to the VIOS as an admin user, and set the initiator name and the discovery policy to odm for the iSCSI protocol device by running the following command:

```
chdev -l 'iscsi0' -a initiator_name='<IQN of the VIOS>' -a disc_policy='odm'
```

Adding all the iSCSI target devices to the ODM on each VIOS

Log in to the VIOS as an admin user, and define the target devices in the ODM by running the following commands:

```
# mkiscsi -l iscsi0 -g static -t <Storage IQN #1> -n 3260 -i <Storage IP Addr #1>
# mkiscsi -l iscsi0 -g static -t <Storage IQN #2> -n 3260 -i <Storage IP Addr #2>
```

Running the configuration manager to configure the disk devices on each VIOS

Log in to the VIOS as an admin user, and run the configuration manager (`cfgvdev`).

As a result, all the iSCSI LUNs are discovered in the VIOS as hdisks, and they can be mapped through vSCSI to the client LPARs, as shown in Figure 2-17.

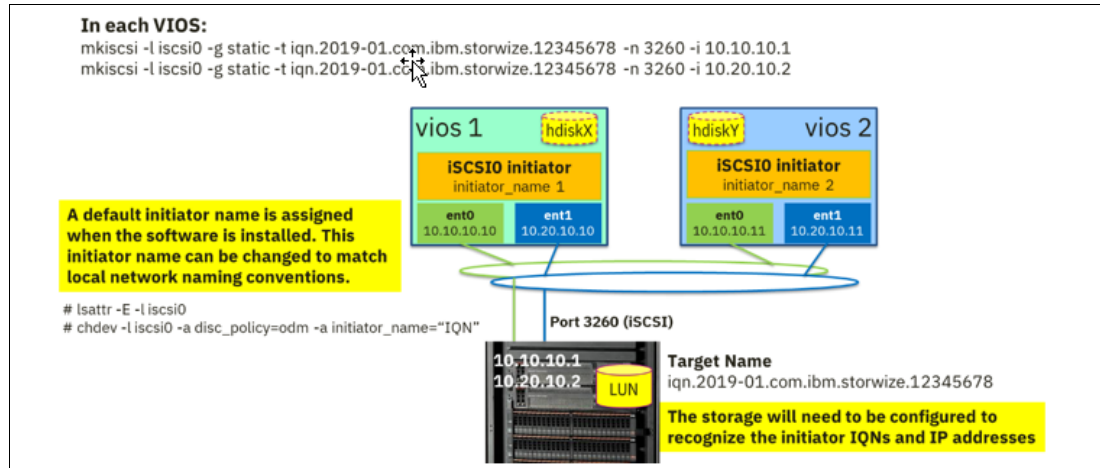


Figure 2-17 Diagram of the iSCSI configuration on the VIOS

2.6 Linux I/O

The I/O stack ranges from the physical device on the storage system up to the file system in the OS. This section focuses only on the OS portion. It does *not* describe the locations of the physical adapters in regard to LPARs, VIOSs, or other components. The focus is on multipathing. The objective is the elimination of single point of failures, and to add robustness to I/O scheduling, interrupt request (IRQ) balancing, the LVM, and the file systems that are relevant to the SAP applications.

2.6.1 Multipathing

The use cases for MPIO with VIOS on PowerVM are as follows:

- ▶ Reduced planned downtime by using rolling maintenance
- ▶ Reduced unplanned downtime by eliminating all single points of failure with less hardware
- ▶ Improved performance

For AIX, MPIO stacks are the dominant deployment option that works independently from the application type. In Linux, it is not widely used yet.

Regarding multipathing on Linux, consider the following important points:

- ▶ VIOS provides different options to virtualize I/O. Becoming familiar with the pros and cons of each option is essential to achieve the result that you want. The recommended default deployment is to use NPIV in dual-VIOS load-sharing setups. There are no limitations or rules from SAP for non- SAP HANA file systems, but using older technologies impose higher latency that is unwanted for database workloads.
- ▶ Differentiate between boot and SAP file systems. For SAP file systems, a good entry point is the default multipath settings that are provided by the OS vendor. For all IBM Storage Systems, the boot device needs special care.
- ▶ If performance is OK, do not start tuning because what helps in the one scenario can cause severe issue in others.

As a best practice, start with a dual-VIOS concept with load sharing that uses NPIV. NPIV does a direct pass through of all I/O requests to the LPAR instead creating a virtual device inside the VIOS, and then adds that virtual device to the LPAR, which results in a copy. This configuration is easier because you can add LUNs from the storage directly to the LPAR by using the host attachment function. Also, a leaner architecture enables lower latency and less processing impact in the virtualization because VIOS has less work to do and uses fewer cores.

The Linux MPIO driver enables multiple paths to a single device in FC environments for SAN storage. To configure it, complete the following steps.

Note: SUSE and Red Hat publish comprehensive documentation for each major release. For example, you can find the documentation for SUSE Linux Enterprise Server 15 SP1 at [SUSE Documentation](#).

1. Enable the daemon for MPIO (if this action was not done during installation).

This task must be performed for all partitions. Start the multipath daemon `multipathd` at boot time to enable automatically multipath services. To enable the daemon, run the following command:

```
systemctl enable multipathd
```

If the multipath services are enabled (or disabled), rebuild the `initrd` afterward by running the following command:

```
dracut --force --add multipath
```

2. For IBM System Storage™ servers, the default multipath settings work for boot and SAP HANA file systems for standard setups. Decide whether a tailored multipath configuration is needed. Candidates for a tailored `multipath.conf` file are:
 - Cluster managers demanding specific timeout settings.
 - Storage vendors not providing pretested defaults.
 - Administration tasks that require different timeout settings.

Check whether the `multipath.conf` file exists by running the following command:

```
/etc/multipath.conf
```

If the file does not exist, it can be created by running the following command:

```
multipath -T > /etc/multipath.conf
```

Check whether `multipath.conf` was created with the defaults for the storage back end.

Here is an excerpt from a sample `multipath.conf` file for SAN Volume Controller:

```
device {  
    vendor          "IBM"  
    product         "2145"
```

Note: When you change multipath versions, you must maintain these manual settings.

As the multipath configuration is not consistent between Linux OS versions, it is a best practice to conduct a series of tests to ensure that the timeout settings match your specific environment.

Note: The configuration of the `multipath.conf` file changes between versions. Check your setting after each service pack update.

Here are the known service pack or kernel updates where the defaults changed:

- SUSE Linux Enterprise Server 12 SP2 to SUSE Linux Enterprise Server 12 SP3.
- Linux kernel 2.6.31 to any newer kernel.

3. Treat special cases.

When using cluster software, you might need to change the defaults to what is described in the Linux distribution guides. For SUSE, these guides are published for each release and sometimes for selected service packs. The one for SUSE Linux Enterprise Server 15 SP1 can be found at [SUSE Documentation](#).

Linux supports different device name types: Worldwide identifier (WWID), user-friendly names, and aliases. For SAP landscapes either use the WWID or create for each WWID an alias. It is not recommended to rely on UUIDs for SAP landscapes.

The server is ready to accept the LUNs for SAP file systems.

4. Verify the bootlist.

Verify the bootlist for multipath boot devices and the LVM filters. Most outages occur due to not checking the boot devices to validate whether they are configured correctly for multipath. Losing the disks for the OS results in losing the SAP application too.

5. Make operational decisions.

Understand the timing of planned and unplanned events, and adjust the timeout and retry settings as needed. Planning maintenance is essential because of the timeout values. The speed at which the maintenance tasks are performed define the difference between an outage and successful maintenance. Covering unplanned outages of certain types require understanding the duration and adjusting the timing and retry settings in the `multipath.conf` file.

2.6.2 Sample multipath configuration

Here are configuration details for our sample multipath configuration:

- ▶ SUSE Linux Enterprise Server 12 SP3, kernel 4.4.162-94.72.
- ▶ Update `powerpc-utils` to at least SUSE Linux Enterprise Server 12-SP3 (src) `powerpc-utils-1.3.3-7.6.2`.
- ▶ Update to the latest `multipath-tools` rpm.
- ▶ VIOS 3.1 using NPIV.

These details assume that you are using IBM Spectrum Virtualize.

Modifying the filter in `/etc/lvm/lvm.conf`

When you use an LV that uses active and passive multipath arrays (not active/enabled multipath arrays as with IBM Spectrum Virtualize), they must be excluded from LVM scans. To do this task, configure filters in `/etc/lvm/lvm.conf` by completing the following steps:

1. Look at the device entries in which WWID patterns occur. In most cases, you find a pattern matching `/dev/mapper/360` or another 3-digit number. If such a pattern occurs, adjust the filter as described in step 2. Otherwise, use the filter that is described in [Troubleshooting boot issues \(multipath with lvm\)](#).
2. Between SUSE Linux Enterprise Server 12 SP2 and SP3, changes were made to `/etc/lvm/lvm.conf` regarding how `multipath_component_detection = 1` is handled. To address this change, use the following filters:

```
filter = [ "a|/dev/mapper/360.*|", "r/.*/"]
```


As most disks with WWIDs start with 360, this filter helps for almost all systems that use FC-attached storage on SUSE Linux Enterprise Server 12 SP3 and later.

Adjusting multipath.conf for LUNs that are used for SAP application file systems

Note: Adjust the file only if you must. Otherwise, use the default settings for the SAP-related LUNs, and ensure that the boot devices are multipath-capable.

In a sample IBM Spectrum Virtualize storage subsystem, by default the configuration must include the following settings:

```
path_grouping_policy    group_by_prio
prio                    alua
rr_weight                uniform          #for HDD
                        priorities        #for Flash/SSD
path_selector            "service-time 0"    #performance optimization
```

There are multiple parameters in `/etc/multipathd.conf` that affect error detection and failover times. These parameters must be set correctly to ensure that the OS and application on the LPAR is not impacted by a failure of a single path or during rolling maintenance activities on the VIOS or storage subsystem.

Number of I/Os that are routed to a single path before switching to the next one

For systems running kernels older than 2.6.31, use the following string in your `multipath.conf` file:

```
rr_min_io                1000
```

Otherwise, use the following string instead:

```
rr_min_io_rq            16
```

`rr_min_io` specifies the number of I/Os routed to one path before switching to the next path in the same path group (for systems running kernels older than 2.6.31). Newer systems use `rr_min_io_rq`. Larger values for `rr_min_io_rq > 32` can improve throughput while deeper queues then have a bigger impact on failure recovery.

In addition to these changes, you must adjust the queue depth of the devices by running a command, for example:

```
echo 64 > cat /sys/bus/scsi/devices/<device>/queue_depth
```

Also, increase `/sys/block/<device>/queue/nr_requests` if the default (128) results in blocked I/O submissions. This action indirectly helps to optimize the blocking inside SAP HANA.

For SAP applications: Go with the default of `rr_min_io_rq = 16` unless you value performance higher than recovery aspects.

For high availability setups

The parameter `no_path_retry` specifies the number of retries until queuing for that path is disabled. The `fail` (or 0) value prevents queuing and results in immediate failure. The default for `no_path_retry` in SUSE Linux Enterprise Server 12 is `fail`, but in SUSE Linux Enterprise Server 11, it is undefined.

Here is the string that is used to set **no_path_retry** to fail:

```
no_path_retry      "fail"
```

For SAP applications: For HA clusters, including SAP HANA Auto-Host-Failover, the typical setting is `no_path_retry = "fail"` to not hinder the take over.

Check with your HA vendor for up-to-date information and best practices if required for disk monitoring.

Timeout tuning for special purposes

Changing the timeout tuning helps in one situation, but can make another situation worse. When changing the timeout tuning, you must have the skills and understanding about the dependencies, for example, the **no_path_retry** parameter.

If `no_path_retry` is set to an integer greater than zero and is not set to queue, there are three different factors in a simplified model that define the timeout in the Linux kernel before an inaccessible volume leads to an I/O error. These factors are summarized in the following equation:

$$\text{Time until timeout} = (\text{number_of_active_paths} * \text{polling_interval} * \text{no_path_retry}) + \text{number of (recently seen) active paths (number_of_active_paths)}$$

The number of recently seen active paths depends on the number of paths that are set by the SAN and zoning configuration, and the MPIO parameter **dev_loss_tmo**:

```
dev_loss_tmo      typically between 120-300 for SAP applications
```

If a failure is detected on a multipath link, the SCSI layer waits for a timeout of **dev_loss_tmo seconds** before the multipath link is marked as failed. When the path is marked as failed, any I/O on that failed path is also marked as failed.

When a link problem is detected, the SCSI layer waits for a timeout of **fast_io_fail_tmo seconds** before the I/O to devices on that path is marked as failed. If I/O is in a blocked queue, the I/O does not fail until the **dev_loss_tmo** time elapses and the queue is unblocked. The value must be smaller than the value of **dev_loss_tmo**.

If there is a failure of one or more paths and there are more than **dev_loss_tmo seconds** before another path failure event, the number of recently seen active paths is reduced at first.

The LUN is still accessible if there are active paths that are available, and the number of recently seen active paths is reduced. Thus, if there is another path failure event afterward, the time until an I/O error is shown is reduced.

The parameter **polling_interval** is the interval between two path checks in seconds. For properly functioning paths, the interval between checks gradually increases to **max_polling_interval**.

For SAP applications: Using the defaults must be the starting point. Changes need special considerations and must be made based only on expert knowledge.

User-friendly names

When using aliases instead of WWPNs, set `user_friendly_names` to `yes` and add the list of WWPN aliases to `multipath.conf`. Typically, this parameter is added to the default profile and not the storage-specific portion.

Here is an example of using the `user_friendly_names` parameter:

```
user_friendly_names "yes"
```

Latency optimization

The `service-time-0` value selects the path with the potential to have the lowest time-reducing latency, which is a best practice for SAP applications when performance must be optimized.

Here is an example of using the `service-time-0` value:

```
path_selector "service-time 0"
```

Activating the changed multipath.conf file

After you change `multipath.conf`, run the following command:

```
systemctl reload multipathd
```

Then, verify that no errors occurred by running the following command:

```
dmesg -T | tail -16
```

To ensure that you have the latest configuration, update `initrd` by running the following command:

```
dracut --force --add multipath
```

Troubleshooting

If you have issues, see the following resources:

- ▶ [Troubleshooting boot issues \(multipath with lvm\)](#)
- ▶ [Multipath drive failed with queue_if_no_path after all pathes failed](#)

Enabling MPIO for an existing boot device

The preferred method for enabling MPIO is to enable multipath for all devices during installation. If this step was omitted, complete the following steps:

1. Mount the devices by using the `/dev/disk/by-id` path that you used during the installation.
2. Open or create `/etc/dracut.conf.d/10-mp.conf` and add the following line (mind the leading white space):

```
force_drivers+=" dm-multipath"
```

If you have problems with this process in SUSE Linux Enterprise Server 12, see [Systemd-udev-settle timing out](#).

3. Show your boot list by running the following command:

```
#bootlist -m normal -o  
sd
```

```
# bootlist -m normal -r  
/vdevice/vfc-client@30000002/disk@500507680c326db2
```

→ only a single path is known from the bootlist as a single point of failure for the root device.

In this case the bootlist must be extended as described. The amount is limited depending on the OS used.

4. Find your root device in the VG column, with system providing the WWID by running the following command:

```
pvs | grep system
PV                               VG          Fmt Attr PSize PFree
/dev/mapper/360050768018087c5200000000000d68-part2 system      lvm2  a-- 49.99g 4.00m
/dev/mapper/360050768018087c5200000000000d6f          hn_lg_vg   lvm2  a-- 32.00g  0
```

5. Find the paths for the root device by running **multipath -ll**, as shown in Example 2-9.

Example 2-9 Paths for the root device

```
# multipath -ll
360050768018087c5200000000000d6f dm-2 IBM,2145
size=32G features='0' hwhandler='0' wp=rw
|+- policy='service-time 0' prio=50 status=active
| | - 1:0:0:3 sdc 8:32 active ready running
| | - 1:0:12:3 sddg 70:224 active ready running
| | - 1:0:4:3 sdaf 65:240 active ready running
| | - 1:0:8:3 sdbt 68:112 active ready running
| | - 2:0:0:3 sds 65:32 active ready running
| | - 2:0:12:3 sdej 128:176 active ready running
| | - 2:0:4:3 sdbg 67:160 active ready running
| | - 2:0:8:3 sdcv 70:48 active ready running
|+- policy='service-time 0' prio=10 status=enabled
| | - 1:0:10:3 sdcm 69:160 active ready running
| | - 1:0:14:3 sdea 128:32 active ready running
| | - 1:0:2:3 sdm 8:192 active ready running
| | - 1:0:6:3 sdaz 67:48 active ready running
| | - 2:0:10:3 sddo 71:96 active ready running
| | - 2:0:14:3 sdew 129:128 active ready running
| | - 2:0:2:3 sdam 66:96 active ready running
| | - 2:0:6:3 sdc b 68:240 active ready running
360050768018087c5200000000000d68 dm-9 IBM,2145
size=50G features='0' hwhandler='0' wp=rw
|+- policy='service-time 0' prio=50 status=active
| | - 1:0:1:0 sdj 8:144 active ready running
| | - 1:0:13:0 sddv 71:208 active ready running
| | - 1:0:5:0 sdat 66:208 active ready running
| | - 1:0:9:0 sdcg 69:64 active ready running
| | - 2:0:1:0 sdag 66:0 active ready running
| | - 2:0:13:0 sdet 129:80 active ready running
| | - 2:0:5:0 sdbv 68:144 active ready running
| | - 2:0:9:0 sddi 71:0 active ready running
|+- policy='service-time 0' prio=10 status=enabled
| | - 1:0:11:0 sddb 70:144 active ready running
| | - 1:0:15:0 sdeo 129:0 active ready running
| | - 1:0:3:0 sdz 65:144 active ready running
| | - 1:0:7:0 sdbn 68:16 active ready running
| | - 2:0:11:0 sded 128:80 active ready running
| | - 2:0:15:0 sdfd 129:240 active ready running
```

```
| - 2:0:3:0 sdba 67:64 active ready running
^- 2:0:7:0 sdcg 69:208 active ready running
```

6. The **multipath -ll** command shows the existing path (sdat - green). Now, select more paths. Example 2-10 is based on a SAN Volume Controller, where by default half of the paths are active and the other half are enabled. Check whether you have both paths enabled so that you can always boot. Example 2-10 shows the addition of the blue paths (sdet and sddb).

Example 2-10 The available path and adding new paths

```
360050768018087c5200000000000d68 dm-9 IBM,2145
size=50G features='0' hwhandler='0' wp=rw
| +- policy='service-time 0' prio=50 status=active
| | - 1:0:1:0 sdj 8:144 active ready running
| | - 1:0:13:0 sddv 71:208 active ready running
| | - 1:0:5:0 sdat 66:208 active ready running
| | - 1:0:9:0 sdcg 69:64 active ready running
| | - 2:0:1:0 sdag 66:0 active ready running
| | - 2:0:13:0 sdet 129:80 active ready running
| | - 2:0:5:0 sdbv 68:144 active ready running
| | ^- 2:0:9:0 sddi 71:0 active ready running
^-+ policy='service-time 0' prio=10 status=enabled
| - 1:0:11:0 sddb 70:144 active ready running
| - 1:0:15:0 sdeo 129:0 active ready running
| - 1:0:3:0 sdz 65:144 active ready running
| - 1:0:7:0 sdbn 68:16 active ready running
| - 2:0:11:0 sded 128:80 active ready running
| - 2:0:15:0 sdfd 129:240 active ready running
| - 2:0:3:0 sdba 67:64 active ready running
| ^- 2:0:7:0 sdcg 69:208 active ready running
```

7. Extend the bootlist by running the following command:

```
# bootlist -m normal -o sdat sdet sddb
sdat
sddv
sddb

# bootlist -m normal -r
/vdevice/vfc-client@30000002/disk@500507680c326db2
/vdevice/vfc-client@30000004/disk@500507680c526db2
/vdevice/vfc-client@30000002/disk@500507680c516db4
```

8. To verify that no errors occurred, run the following command:

```
dmesg -T | tail -8
```

Multipath configuration testing

Before you create the test plan, you need to understand the time that MPIO needs to put a path into the faulty state and then back into active mode. The duration differs depending on where the action was taken. For example, rolling VIOS maintenance is quickly detected, but for storage headnodes, the detection takes much longer because propagating the faulty path into MPIO requires more steps.

It is important to test the multipath configuration because there is a difference between the boot process and other file systems. Although the boot process cannot go through all paths during boot, by default all other file systems have higher robustness regarding redundancy.

Here are some sample candidates for testing:

► Performance

Performance tuning is a combination of physical redundancy (number of paths), multipath settings, and file system configuration. You must start with file system configuration optimization and physical redundancy *before* you start the multipath configuration. Complete the following steps:

- a. Verify the different service time settings for a performance difference.
- b. Verify different options on ratios among the number of active paths, numbers of LUNs per file system, and file system settings, such as blocksize and stripes.

► Rolling maintenance

When you pull and replug cables for different components, such as VIOS, switches, and storage, be careful to not do this task too quickly, or you encounter a race condition where all the paths are faulty.

► Path and storage failure

Define the outcome that you want based on your SLAs, and then test for these outcomes.

► HA of shared devices

When a cluster manager is installed for applications that use shared disks, you must test a disk failure with cluster handling, and differentiate among HA deployments on a shared-nothing architecture versus a deployment that is based on shared disks. Contact your cluster vendor for their recommendations.

2.6.3 Linux file systems that are relevant to SAP applications

Several different file system types are implemented within the Linux distributions, as shown in Figure 2-18.

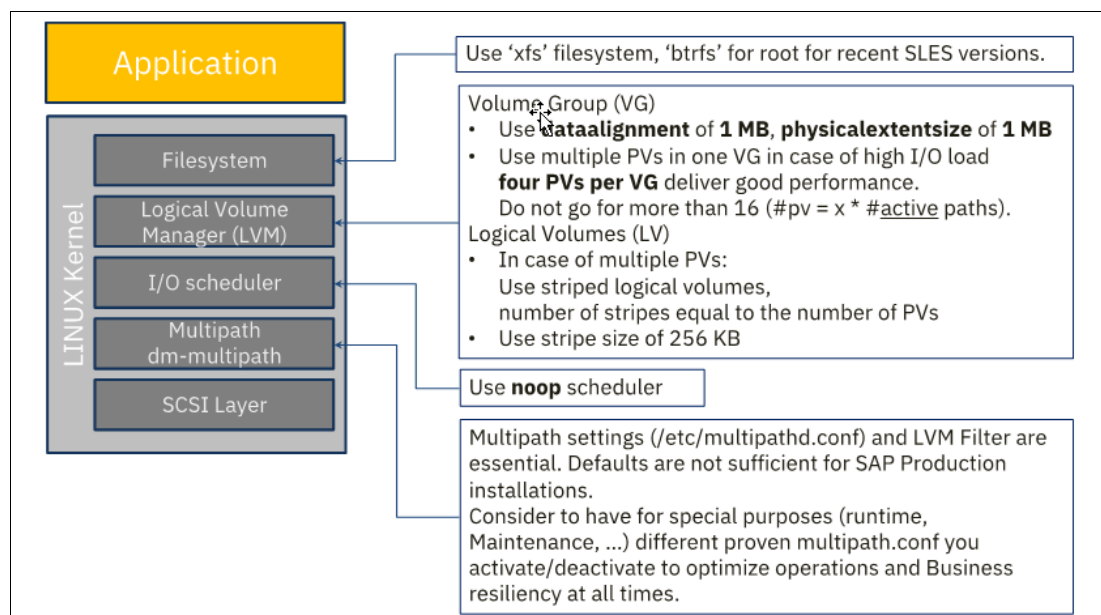


Figure 2-18 I/O stack for FC-attached multipath environments for SAP applications

For SAP workloads, use the default root file systems of the Linux distributor, and for SAP-related file systems, use XFS unless directed otherwise by an SAP Note. For the shared file systems, the most common deployment is an HA NFS server.

Here are the different file system types that are implemented within Linux distributions:

► Btrfs

Btrfs is a logging style, copy-on-write file system. A changed block is written to a new location, and then the links are updated to point to the new block. Changes are not committed until the last write. SUSE Linux Enterprise Server by default is installed by using Btrfs and with snapshots for the root partition. With snapshots, you can easily reset the system to a defined state, for example, in case of rolling-back applied updates, or to back up files. Before rolling back the system by using a snapshot, ensure that user and application data do not get lost or overwritten during a rollback. More Btrfs subvolumes are created on the root file system, and the subvolumes can be excluded from the snapshot.

Note: For production environments, plan for sufficient space to keep copies of the file system to benefit from the features of Btrfs. In the event more space is needed, use the Btrfs file system over LVM to add more volumes through LVM, or you can use another method advised by SUSE support team. For example, by using btrfs, changes in configurations can be found with snapshots and comparing these to the previous configuration.

► XFS

XFS is optimized for handling large files and provides high performance. In SUSE Linux Enterprise Server, XFS is the default file system for data partitions. XFS is supported for internal disks and SAN-attached storage. Multipathing with a matching file system configuration must be enabled to protect against path losses that impact the application and optimal performance.

► NFS

NFS is used in SAP landscapes to share binary files that are used to transport changes (/usr/sap/trans) or ensure that the same binary files are visible to all related code at the same time (/sapmnt and HA configurations, such as SAP HANA auto host failover). NFS requires a dedicated and redundant storage network with at least 10 Gb network connectivity.

► IBM Spectrum Scale (formerly known as GPFS)

IBM Spectrum Scale is a high-performance clustered file system. It can be deployed in shared-disk or shared-nothing distributed parallel modes. IBM Spectrum Scale requires an external storage server (or GPFS cluster) that is attached through InfiniBand or at least 10 Gb network connectivity. IBM Spectrum Scale File Placement Optimizer (FPO) or a self-build client/server IBM Spectrum Scale cluster is not supported for SAP HANA.

For more information, see [SAP Note 2055470](#).

2.6.4 Logical Volume Manager

With LVM, you can have a layer of abstraction between the Linux OS and the disk devices. One of the most interesting features of LVM is that you can use it to resize (extend or reduce) the various structured elements. Structures of the LVM consist of the following items:

- One or more entire LUNs or partitions are configured as PVs.
- A VG is created by using one or more PVs.
- One or multiple LVs can then be created in a VG.

- ▶ The OS then creates a file system by using the LV structure.

Because the VGs and LVs are not physically tied to disk devices, it is possible to dynamically resize and create disks and partitions. To add disk capacity to an LPAR, either create a VG or add disk space to an existing VG, and then either expand an existing LV or create one.

Note: Resizing the root VG by adding more PVs is not possible because the LPAR fails during the next restart (bootloader).

To add a new VG or LV, complete the following steps:

1. Map a new LUN to the LPAR.

The required steps depend on the attachment method of the disk. Eventually, the disk storage is presented by using vSCSI attachment from the VIOS to the client LPAR. In this case, the back-end device must be present to the VIOS. The device is attached to the LPAR by using VIOS device mapping commands.

Another possibility is that the client LPAR uses physical or virtual FC adapters. In this case, the disk must be masked to the WWPN of the FC adapter in the storage system. SAN zoning must allow access between the LPAR FC adapters and the storage system.

2. Make the new LUN visible to the Linux OS.

Run the **rescan-scsi-bus.sh** script to automatically update the logical unit configuration of the LPAR. For more information about how to use this script, run the following command:

```
rescan-scsi-bus.sh --help
```

If **rescan-scsi-bus.sh** does not work, run the following command instead:

```
echo "- - -" > /sys/class/scsi_host/host0/scan #iterate the "0" over the number of ports
```

3. Create a PV on the new LUN.

The new device for the disk now visible to the OS. To initialize the PV for use by the LVM, run the **pvcreate** command.

4. Assign the new PV to an existing VG or create a VG.

To add one or more PVs to an existing VG, run the **vgextend** command. This command increases the space that is available for LVs in the VG. To create a VG, run the **vgcreate** command.

5. Create an LV in the VG, or extend an existing LV by running the following command:

```
# lvcreate --size 5G -n testlv /dev/testvg  
Logical volume testlv created.
```

If the VG consists of multiple LUNs, it is beneficial to stripe the LV across all the LUNs. The command-line option **-i, --stripes Stripes** of the **lvcreate** command distributes the LV across multiple LUNs. The LV must be striped across all PVs of the VG. A best practice is to set the number of stripes equal to the number of PVs.

The command-line option **-I, --stripesize StripeSize** of the **lvcreate** command specifies the stripe size. The stripe size must be a power of 2, but must not exceed the physical extent size.

For file systems running a workload similar to a database log file (/hana/log), a stripe size of 64 KB delivers the best results. For file systems with a larger blocksize (/hana/data), 64 KB is as good as 128 KB. So, for HANA databases, a stripe size of 64 KB is recommended for all file systems.

6. Create a file system on the new LV by running the following command:


```
mkfs.xfs -L testlv /dev/testvg/testlv
```

7. Add the appropriate entries to /etc/fstab to mount the file system:

```
/dev/hdbvg/usr_sap      /usr/sap      xfs  defaults      1 2
/dev/hdbvg/hana_data_hn1 /hana/data/HN1 xfs  defaults      1 2
/dev/hdbvg/hana_log_hn1 /hana/log/HN1  xfs  defaults      1 2
/dev/hdbvg/hana_shared_hn1 /hana/shared/HN1 xfs  defaults      1 2
```

8. Mount the file system by running the following command:

```
mount -a # Will read /etc/fstab and mount file systems which are not mounted
```

References

In the context of SAP HANA and IBM, SUSE published an SAP Note about how to configure a striped XFS file system, which you can reuse independently of SAP HANA for all XFS file systems that are used by SAP applications in multipath SAN environments on Red Hat and SUSE Linux Enterprise Server OSs. You can find this SAP Note at [SAP Note 1944799](#).

Related publications

The publications that are listed in this section are considered suitable for a more detailed description of the topics that are covered in this paper.

IBM Redbooks

The following IBM Redbooks publications provide more information about the topics in this document. Some publications that are referenced in this list might be available in softcopy only.

- ▶ *IBM Power System E980: Technical Overview and Introduction, REDP-5510*
- ▶ *Live Partition Mobility Setup Checklist, TIPS1184*

You can search for, view, download, or order these documents and other Redbooks, Redpapers, web docs, drafts, and additional materials, at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- ▶ Configuring your Network for SAP HANA
<https://ibm.co/39eWfVj>
- ▶ IBM Network Configuration Guide for SAP HANA Supplemental Guide
<http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102502>
- ▶ SAP Support Software download portal
<https://support.sap.com/swdc>
- ▶ SEA failover
<https://ibm.co/2WGkAuk>
- ▶ Shared Ethernet adapters (SEAs) for load-sharing
<https://ibm.co/2vP3uiX>
- ▶ SUSE ibmvnic driver support
<https://bit.ly/2JkjjRz>

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services



REDP-5581-00

ISBN 0738458724

Printed in U.S.A.

Get connected

