Redbooks
ibm.com/redbooks

Building a SAN-less Private Cloud with IBM PowerVM and IBM PowerVC

Javier Bazan Lazcano

Stephen Lutz

Cloud

Power Systems

IBM

Redpaper

International Technical Support Organization

Building a SAN-less Private Cloud with IBM PowerVM and IBM PowerVC

July 2018

REDP-5455-02

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

Third Edition (July 2018)

This edition applies to Version 1, Release 4, Modification 1 of IBM Cloud PowerVC Manager for Software-Defined Infrastructure V1.1 (product number 5765-VCD).

Contents

Notices

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

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

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

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

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

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

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

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

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

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at http://www.ibm.com/legal/copytrade.shtml

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

Redbooks (logo) ®
AIX®
GPFS™
IBM®
IBM Cloud™
IBM Spectrum™
IBM Spectrum Protect™
IBM Spectrum Scale™
Power Systems™
POWER9™
PowerVM®
Redbooks®
Redpaper™

The following terms are trademarks of other companies:

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group 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 a software-defined infrastructure (SDI) solution with IBM PowerVC. In IBM PowerVC SDI, you combine scale-out IBM Power Systems™ servers with software that creates the whole stack that is needed to provide virtual machines (VMs) for applications such as open source databases or Hadoop.

The SDI solution uses base IBM Power Systems technologies such as IBM PowerVM® NovaLink and the open source hypervisor kernel-based virtual machine (KVM). The solution combines with sophisticated storage technologies, such as IBM Spectrum™ Scale, and with the powerful networking capabilities that are provided by the Open vSwitch (OVS) technology. IBM PowerVC “hides” much of this software so that it is not apparent to your daily cloud operations. By using IBM PowerVC, you can manage scale-out SDI-based systems along with traditional PowerVM systems.

This publication describes how to install and configure the SDI solution using PowerVM and IBM PowerVC running on the Power Systems platform. This publication also presents the essentials to help existing Power Systems technical specialists use existing "under the covers" disk space to build a cost-effective cloud solution.

Authors

This paper was produced by a team of specialists from around the world working at the International Technical Support Organization, Austin Center.

Javier Bazan Lazcano is an IT Specialist at IBM Argentina. He has been with IBM since 2007, and has provided support and implemented solutions for North American accounts. He has more than 15 years of hands-on experience working with UNIX and virtualization platforms. His areas of expertise include IBM Power Systems, PowerVM, AIX®, HMC, Linux Red Hat, VMWare, and HPUX products.

Stephen Lutz is a Certified Senior Technical Sales Professional for Power Systems working at IBM Germany. He holds a degree in Commercial Information Technology from the University of Applied Science Karlsruhe, Germany. He is an IBM POWER9™ champion and has 19 years of experience in AIX, Linux, virtualization, and Power Systems and its predecessors. He provides pre-sales technical support to clients, IBM Business Partners, and IBM sales representatives in Germany. Stephen is an expert in IBM PowerVC. He is involved with many IBM PowerVC customer projects in DACH (Germany, Austria, and Switzerland), and has given presentation about IBM PowerVC at the IBM Technical Universities in Europe for several years.

The project that produced this publication was managed by:
Scott Vetter, PMP

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

Hsien Chang, Rebecca Dimock, Eric Larese, Andrew (Drew) Thorstensen, Sridhar Venkat
IBM US

Shyama Venugopal
IBM India

Thanks to the authors of the previous editions of this paper.

- Authors of the first edition, Building a SAN-less Private Cloud with IBM PowerVM and IBM PowerVC, published in February 2018, were:

  Drew Thorstensen and Shyama Venugopal

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 ITSO 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, International Technical Support Organization
  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`

1

Architecture of a software-defined infrastructure solution with IBM Power Systems servers and IBM PowerVC

This chapter provides an overview of software-defined infrastructure (SDI) and some architectural implementation examples with Power Systems and IBM PowerVC.

1.1 Software-defined infrastructure

The term *SDI* has many meanings to various individuals. But, it generally means using software-based systems to serve as an abstraction to the hardware. Beyond that, it also means controlling many different pieces of infrastructure (such as servers) as a single entity through a control point.

In general, you can break SDI into three different areas:

- *Software-defined compute* is the oldest area of SDI, and is often considered virtualization. Being able to carve processors into subunits and increase the density on a system is nothing new, but the cloud takes it to the next level by providing key capabilities, such as live migration, capacity on demand, and automated remote restart.

  A SDI-based solution with IBM PowerVC 1.4.1 provides integration of the following hypervisors:

  – PowerVM NovaLink

  – Kernel-based virtual machine (KVM) on Power through Ubuntu 16.04

  – KVM on Power through Red Hat Enterprise Server 7.5

  Environments can mix IBM PowerVM and KVM on Power Systems servers within a single IBM PowerVC environment. For more information about what hypervisor versions PowerVC SDI will integrate with in the future, see IBM Knowledge Center IBM Knowledge Center and change the version to the current PowerVC version.

  > **Note:** IBM PowerVC also supports traditional Power Systems environments with HMC and Virtual I/O Servers (VIOSes) that are not in an SDI context. An SDI-based infrastructure can be managed along with a traditional infrastructure from a single IBM PowerVC, as shown in Figure 1-1 on page 3.

- *Software-defined networking* (SDN) enables more flexible network configurations. By using it, cloud administrators can define virtual networks (private networks that do not require any virtual local area networks (VLANs)), control quality of service (QoS), and limit within the hypervisor which ports/protocols over which a workload can communicate.
- *Software-defined storage* (SDS) is a flexible storage back end that is managed and controlled by using software instead of a traditional storage array. This document uses IBM Spectrum Scale™ as its example storage back end.

Starting with IBM PowerVC 1.4.0, administrators can use SDI in their environments with their existing PowerVM environments.

1.2 The motivation behind software-defined infrastructure

You use SDI to integrate new types of infrastructure with your existing environments. With native KVM support, you can take advantage of the Power Systems LC models running standard KVM Linux distributions, such as Red Hat and Ubuntu. SDI is also supported on PowerVM servers.

You also can use the SDI infrastructure to run an Ethernet-only-based cloud infrastructure, which reduces the physical complexity of your infrastructure. These types of clouds are good fits for a cost/performance-focused infrastructure.

This type of infrastructure is a good fit for new cloud native workloads, such as Mongo DB, PostgreSQL, Cassandra, Redis, Hadoop, or Spark. These types of workloads are built for a scale-out cloud solution. Power Systems servers are suitable platforms for these kinds of workloads. Benchmarks show that Power Systems servers are far better than an x86-based solution, both from performance point and price/performance points of view. For example, Hadoop runs better on Power Systems serves by many factors in terms of price/performance.

From a management point of view, it is easy to manage an SDI environment because of the IBM PowerVC front end. IBM PowerVC 1.4 and later version can manage both the robust and classic Power Systems environment and the new SDI solution from one interface, as shown in Figure 1-1.

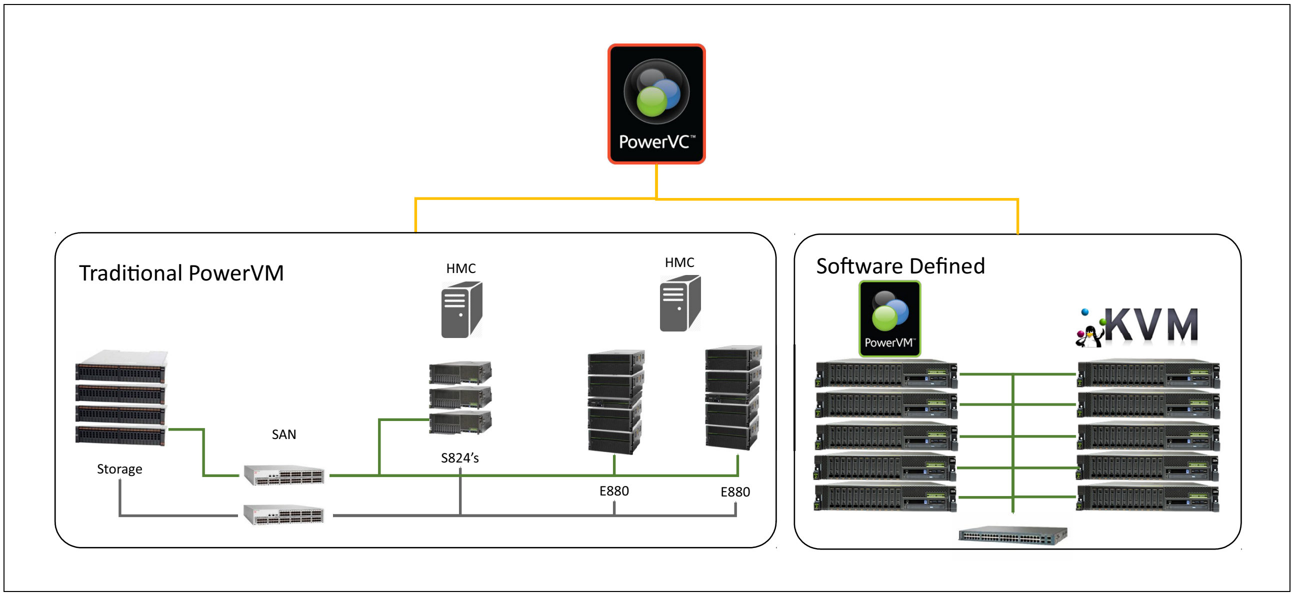


Figure 1-1 Power Systems servers combined management

A single IBM PowerVC 1.4+ installation can mix any of the following nodes:

- VIOS-based PowerVM systems:
  - Sample configurations might be either a Power S924, Power S824, or Power E880 system that uses VIOS with N_Port ID Virtualization (NPIV) or virtual Small Computer Systems Interface (vSCSI) connections to a Fibre Channel array.
  - Includes shared storage pool (SSP)-based systems.
  - Can be PowerVM NovaLink based or HMC-only managed systems.
- SDN/VIOS Hybrid PowerVM systems

  Systems that use VIOS for the storage back end and the SDN feature through PowerVM NovaLink.
- SDI-based PowerVM systems

  A sample configuration here might be a Power S922 running PowerVM that is installed in SDI mode. There are no traditional VIOSes in this environment, but PowerVM NovaLink fills the role of the VIOS.
- KVM on IBM Power Systems

  A sample configuration might be a Power 822LC system running Ubuntu 16.04 LTS with either local disks, iSCSI, or Fibre Channel-based storage area network (SAN) (to support the IBM Spectrum Scale cluster).

The next sections provide more details about use case scenarios and the components that are used in them.

1.3 Use cases

The following sections provide use cases that demonstrate what environments that you can build by using SDI.

1.3.1 Use case 1: Building a storage area network-less cloud

Assume that you have a few servers and you want to build a cloud quickly. You do not have a SAN, just Ethernet connections and servers. However, the servers have local disks.

You can now build a SAN-less cloud, as shown in Figure 1-2.

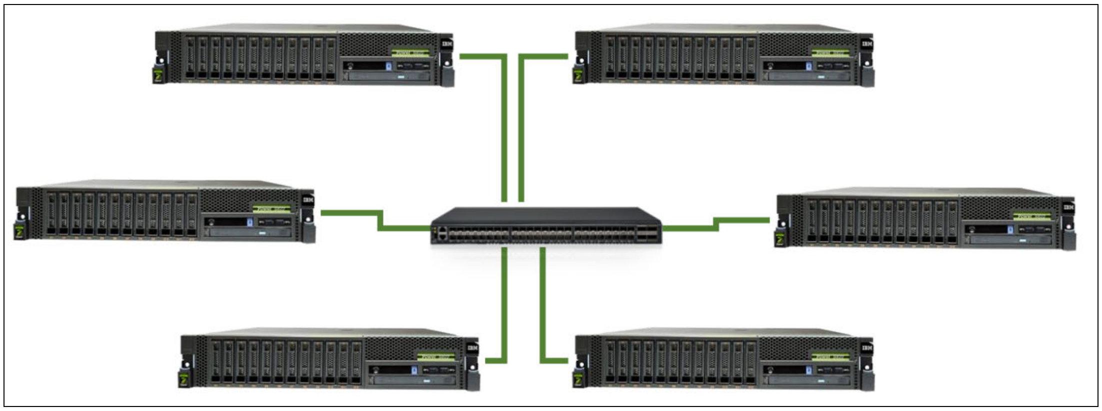


Figure 1-2 Example of SAN-less cloud configuration

In Figure 1-2, you have six Power S922 or Power L922 servers with PowerVM (or Power S822LC servers running KVM on Power) and link them together with 10+Gb Ethernet.

Note: The Power LC922 servers run Ubuntu 18, which is not supported for IBM PowerVC, so these servers cannot be used.

IBM PowerVC creates a virtual SAN across these nodes by using IBM Spectrum Scale. Redundancy is handled at the server level, so if any disk or server goes down there are extra copies and the cluster remains alive.

This is a flexible, Ethernet-only configuration. No external storage access is needed. The cloud storage is controlled by the server administrator, and the complexity of that storage is hidden behind IBM PowerVC.

This type of cluster can run on most Power Systems servers through KVM on Power and PowerVM support. You can even mix KVM servers to sit in the same cluster as your PowerVM servers.

1.3.2 Use case 2: Building an iSCSI-backed cloud

For many environments, keeping the storage external to the hosts makes sense. However, some environments need iSCSI-backed storage for their cloud. With SDI, this goal can be accomplished, as shown in Figure 1-3.

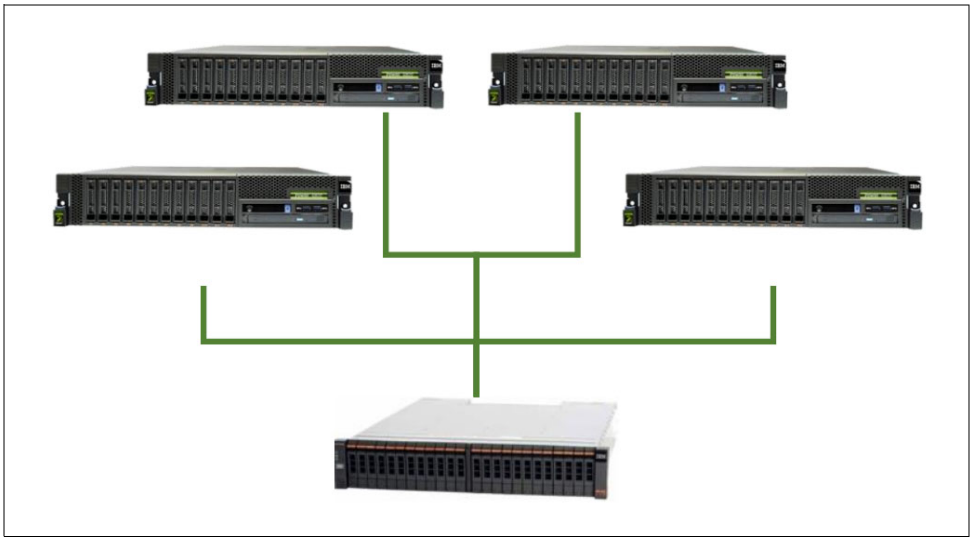


Figure 1-3 SDI iSCSI-backed storage for cloud

In Figure 1-3, you connect a few servers to a storage device over iSCSI. Therefore, these servers do not have any Fibre Channel cards in them. They have a dedicated (or shared) line to the storage device over Ethernet.

The storage logical unit numbers (LUNs) are shared across all of the hosts. IBM PowerVC then sees these storage LUNs and creates a cluster (through IBM Spectrum Scale) on top of them. This process means that the LUNs are created and attached only once. The provisioning of workloads does not require any direct storage interaction, and is instead managed by IBM Spectrum Scale on top of those LUNs.

SDI can also be used to create a virtual SAN on top of Fibre Channel-backed environments. The topology is identical to iSCSI, but uses Fibre Channel instead. This topology can be useful for KVM on Power environments. However, in a PowerVM environment, it is a preferred practice to use the VIOSes.

1.3.3 More use cases

In addition to these use cases, SDI provides extra benefits:

- A private virtualized network through Virtual Extensible LAN (VXLAN) overlay networks

  This configuration can reduce the number of IP addresses or VLANs that must be provisioned on the physical Ethernet devices.

- Virtual routers with failover capabilities:
  - This configuration enables private IP addresses within the VXLAN overlay networks to access the wide area network (WAN).
  - The router can dynamically add external (or public) IP addresses to virtual network interface controllers (NICs) on the private VXLAN network.

  - Routers have failover capabilities, so that if the server hosting the router fails, it dynamically fails over.
- Support for PowerVM or KVM on Power:
  - Pick the hypervisor that makes sense for your environment.
  - If you have Power LC (such as OpenPOWER) systems, KVM on Power is the preferred pick.
  - You can use KVM on Power with your PowerVM systems.
  - Mixing types is allowed.
- Hypervisor-based network access control (also known as security groups):
  - Control what ports, IP addresses, or protocols can communicate with various virtual NICs on your workloads.
  - Define policies and dynamically attaches them to the virtual machines (VMs).
- At the time of writing, there is another use case that is in technology preview mode, which means that it is not yet supported by IBM PowerVC, but can be tested by administrators. QoS with Rate Limiting controls the throughput of your workloads by setting a maximum allowable bandwidth.

These capabilities require that the systems be set up in SDI mode. However, the SDN use cases can be used on PowerVM systems that have PowerVM NovaLink in SDN mode. All SDS or compute nodes require that a PowerVM system be in the SDI (software-defined environment (SDE)) mode.

KVM-based systems can be installed *as is*, and the SDI software is loaded as part of the IBM PowerVC host registration.

1.4 Rack topologies and components

In addition to understanding the use cases, you must understand the physical layout. This section shows three rack topologies that outline various software-defined cloud sizes, from a simple cloud to more complex ones.

All of these topologies assume that local disks are used for storage. However, you might want to connect to the existing SAN (either Fibre Channel or iSCSI) infrastructure. That configuration is supported. However, a connection to that storage network must be accounted for. The requirement from the SDI perspective is that the disks are available to the host, and it is easiest to visualize that configuration with direct attach local disks.

1.4.1 Starter cloud

This configuration is the smallest cloud possible, and is shown Figure 1-4. It offers no redundancy for data, but you may experiment with a cloud software stack.

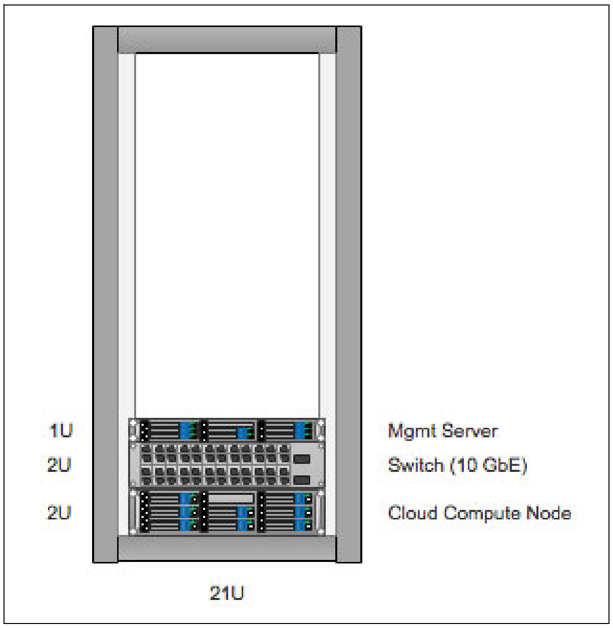


Figure 1-4 Starter cloud configuration

This cloud has a single compute node, which is a management server to house IBM PowerVC, and a switch. There are no separate management and data planes. Everything runs on the 10-GbE switch. Because there are no multiple nodes, there is no storage traffic in the network. There is also no disk redundancy. Therefore, this configuration should be used only for evaluation purposes.

Adding a second server adds storage redundancy and provides key capabilities, such as live migration and remote restart. However, even with two servers, use a 40-GbE switch so that the storage traffic has enough bandwidth.

The cloud compute node can be PowerVM (such as a Power S922, Power L922, Power S822, or Power S824L server) or KVM on Power System (such as a Power S822LC, Power S824L, or Power S821LC server). You can choose the hypervisor that meets your needs.

The management server is what houses the IBM PowerVC, and an example server type is a Power S821LC server.

1.4.2 Mini cloud

The mini cloud that is shown in Figure 1-5 offers more redundancy and is similar to a basic cloud. This cloud offers four compute nodes. Because there are more than three nodes in the cluster, there are three copies of the data. If one of the servers fails, that data is replicated on an active node.

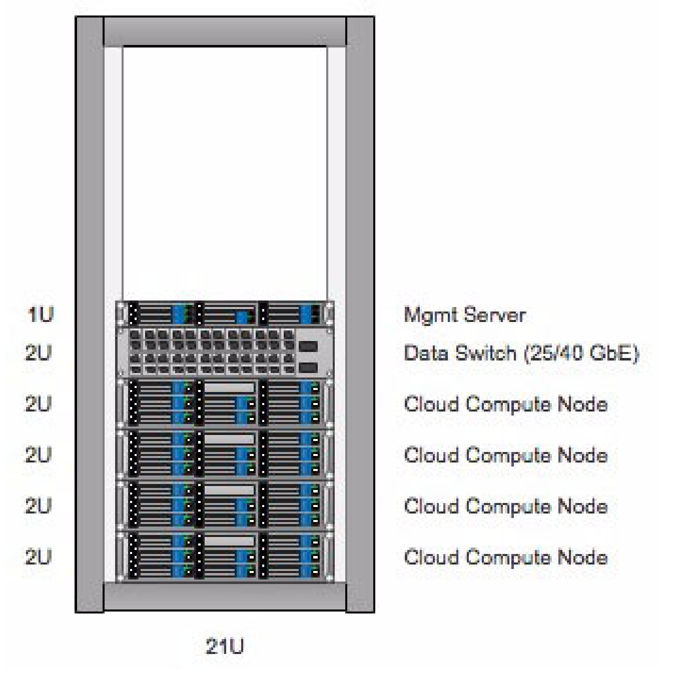


Figure 1-5 Mini cloud configuration

This topology introduces a 25/40-Gb Ethernet switch. The Ethernet switch must have a higher speed because it is still a converged switch that supports both the storage and the data traffic. If it is not fast enough, then the traffic from one area might impact the other (a high network load might slow down storage).

In this model, the compute nodes are directly on the WAN. This configuration means that only flat (untagged) or VLAN networking can be used. The Ethernet switch must also predefine the VLANs on the trunk ports of the switch. This configuration is common in smaller clouds because it does not require many extra pieces of infrastructure.

1.4.3 Rack scale

Figure 1-6 shows a cloud that is built to rack scale. At rack scale, you have enough servers that new challenges arise. For example, your network team might not want to use VLAN for all of the ports, so you must find a way to reduce the VLANs. You also have higher storage traffic in the network (due to the larger number of nodes) and might want to isolate storage and data in their own networks.

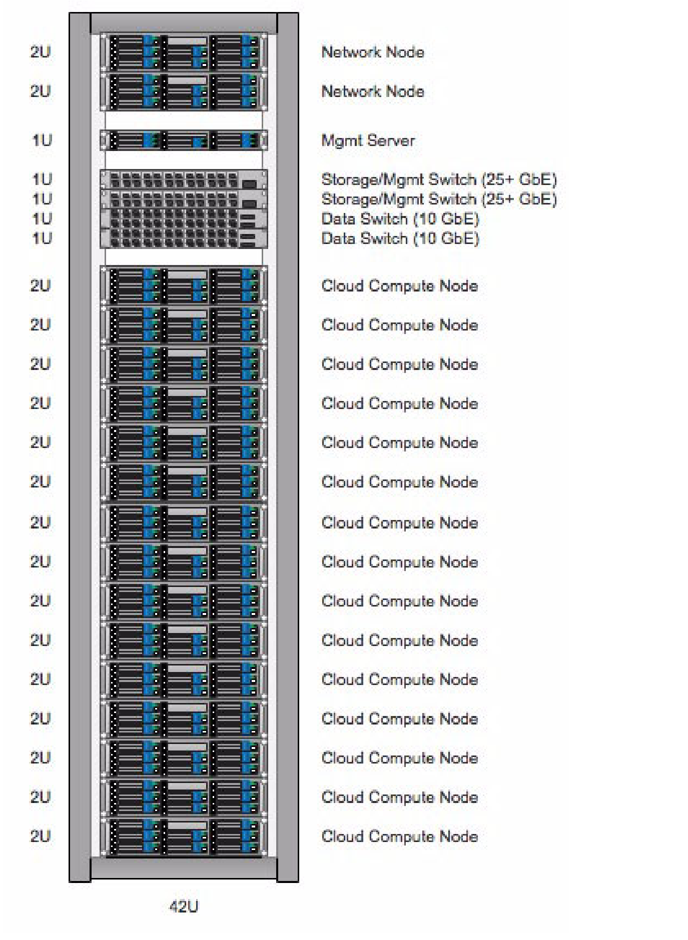


Figure 1-6 Rack-scale configuration

Because of these issues, there are two key additions to this topology:

- The separation of the storage and data networks into independent fabrics. This configuration means that your storage traffic cannot flood your data traffic and vice versa.
- The introduction of VXLAN overlay networks into the mix. VXLAN networks can reduce the number of changes that your network team must make to support your cloud by keeping their changes to the network nodes only.

For example, the network team works only with the ports that are plugged into the network nodes. The compute nodes are fully isolated from the WAN, and they can be independently scaled without having to engage the WAN team.

1.4.4 Storage/management switches

These fast switches support both the storage and management traffic. The lion's share of the traffic is storage-related if you use local disks (due to the replication of the storage devices) or iSCSI communication between the cloud compute nodes. However, there is minimal management-related traffic going over these lines.

When you use isolated storage/management switches (independent from the data network), the adapter in the cloud compute node should not be br-ex (an Open vSwitch (OVS); for more information, see 1.4.5, "Data switches" on page 10). It should instead be an independent port or link aggregation to support that traffic. IBM PowerVC should be given the IP address on that independent port to use for its traffic.

Two switches are included in this topology to support link aggregation.

1.4.5 Data switches

The data switches are redundant and serve as the high-speed links between your network nodes and your compute nodes. The following devices should be plugged into the data switches:

- Cloud compute nodes: At least one high-speed port should be linked to each data switch. A link aggregation should be built on top of the ports. The link aggregation should be an OVS that is named br-ex.
- Network nodes: At least one 10-GbE port should be linked to each data switch.

This process creates a high-speed network for your VXLAN devices. The WAN *does not* plug in to any of the switches in this topology. Instead, the WAN plugs directly into the network nodes.

Figure 1-7 is a conceptual view of the network topology.

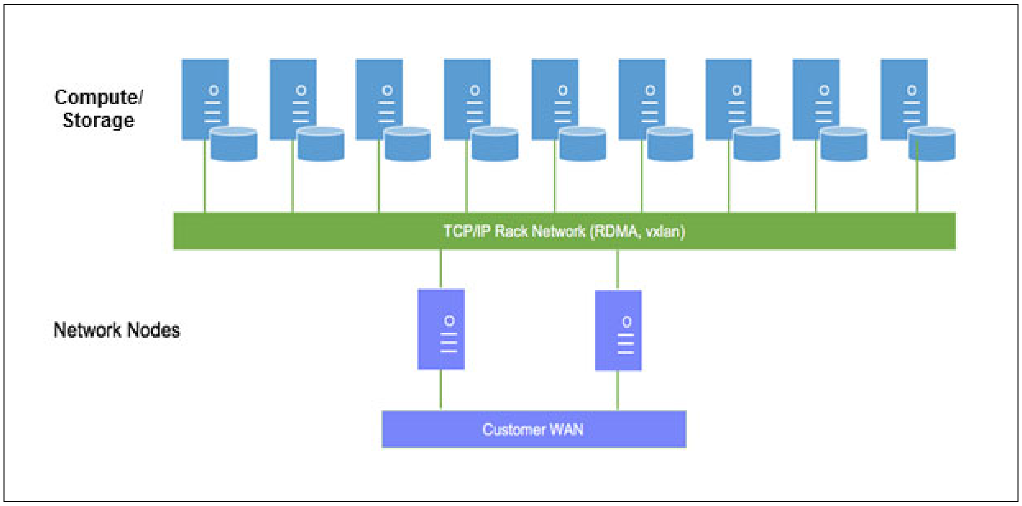


Figure 1-7 Network topology

The network nodes provide the VXLAN encapsulation and decapsulation for the VXLAN networks. This topology also isolates your converged network from any broadcast storms or other network instability that might occur on your broader network where you do not have as much control. This isolation is useful when running a converged network topology, especially when it is carrying your data and storage traffic.

1.4.6 Network nodes

The network nodes act as the bridge between your WAN and the private converged network. As such, they are physically plugged into two different domains.

Network nodes are required only if you want to use VXLANs. If you want to use a flat network, or use VLANs that are provided by your networking infrastructure, the network nodes are optional.

Generally, the network nodes should have two network cards. One port from each network card should plug into the converged network and another into the customer WAN. This use case allows for redundancy in case of a card failure.

A basic topology is shown in Figure 1-8.

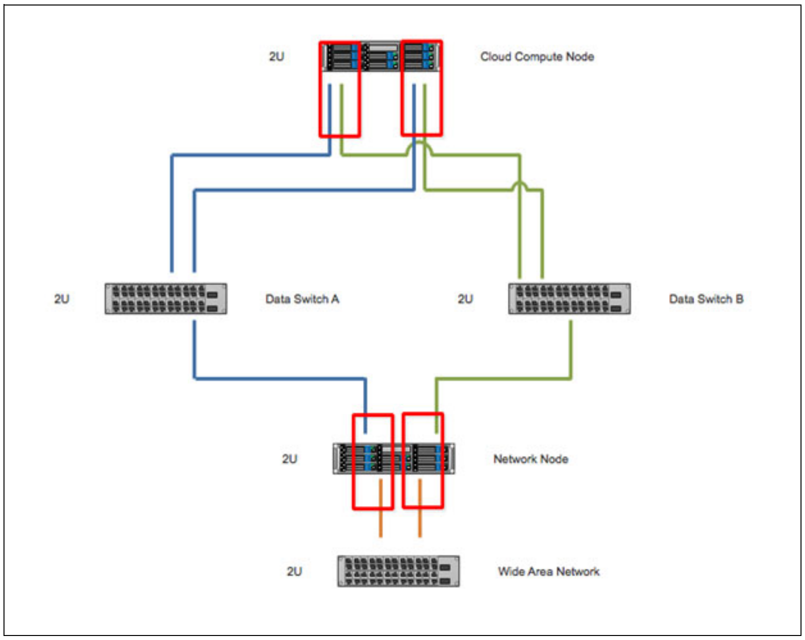


Figure 1-8 Network node topology

In this topology, the red boxes represent the Ethernet adapters in the system. Each line into the box represents a wire between a switch and an adapter port. The adapters are assumed to have at least two ports.

As you can see, each adapter on the cloud compute nodes has one port going to each switch. A single OVS bridge (named br-ex) is put on top of all these high-speed links. The link aggregation mode must be balance-slb or active-backup because the switches are not linked together. Some switch vendors have technologies that enable two switches to be linked together to act as one. If your switch supports that capability, then a standard Link Aggregation Control Protocol (LACP)-based link can be used.

A single internal network IP address should sit atop br-ex. The internal network IP can be a 10.x.x.x address to indicate that it is a private network.

The network nodes have two cards. Two OVS switches must be created:

- Br-ex: Sits atop the links that connect the ports to the WAN.
- Br-data: Sits atop the links connecting the ports from the data switch. Should be balance-slb or active-backup.

  The br-data switch needs a persistent IP address for the internal network. This address should be on the same subnet as the IP addresses on your cloud compute nodes' br-ex switch.

The management server also needs connections to the data switches. It also should have an internal network IP address (likely sitting atop a bridge).

Two 25/40-GbE switches are not required, but are ideal. Although this configuration adds a need for a second Ethernet card and extra bridges on each server, it helps protect the cluster if a switch dies. It also allows the administrator to update the firmware in the network switch without having to take down the cluster.

1.4.7 Understand the scaling requirements

A cloud is a group of devices that are connected through a common management plane. There is no one device that is the cloud; there are many. Each device can be impacted by the scaling in the environment.

In general, an IBM PowerVC server that manages up to 500 VMs requires significantly less capacity than one that manages up to 5,000 VMs. As the cloud grows, ensure that the memory, CPU, and disk capacities grow with it. For more information, see IBM Knowledge Center.

Also, be aware of the other components such as the compute servers and network. The compute servers indicate their capacity in two ways: Capacity allocated and utilization. Capacity allocated is the (generally minimum) amount of CPU that is available to a VM. For example, 2 vCPU with 0.2 Uncapped Processing Units means that the VM has two CPU cores that are available to it. Each core has at least 0.1 or 10% of the CPU cycles. Given that it is uncapped, it can use any remaining cycles that are available.

In that scenario, the capacity that is allocated is 0.2 CPUs. However, the utilization itself is tied to the workload within the VM. If the workload is idle, the CPU utilization is low. If the workload is busy and the cores have extra cycles available, the utilization might be over 100%. In this example, it might scale up to 1000% utilization because it can use a full core.

In general, keep some space available on your servers. The overall goal is to keep your utilization numbers high, not just allocated.

Make sure that you have extra capacity to support your maintenance operations. Maintenance is needed on servers, so having the capacity to support host maintenance mode is important. Host maintenance mode moves all the workloads off your server and prevents new workloads from going on to it. That way, you can perform maintenance on it with peace of mind.

Network capacity

Many SDI deployments use local disks or iSCSI-backed SANs. These run both your data and storage traffic over the IP network. You must ensure that there is enough capacity for each host. However, as you scale out your environment, you might find other challenges. To avoid bottlenecks, keep data and storage traffic separate so that they do not collide.

For example, assume that you have a 10-server cloud running with local disks that replicates the data on writes to two other servers. So, the communication between the servers must be quick. Perhaps a link aggregated through a 10-GbE bond works. But what happens when you add 10 more servers? At that point, you might have to hop between multiple servers, or the switches that you are plugged in to might not have enough switching capacity for all of that traffic.

Keep an eye on the overall network traffic to help ensure that your switches have capacity and that you do not introduce bottlenecks.

Several common bottlenecks can occur. Here is a set of bottlenecks that can affect local disk (such as SAS, SATA, or NVMe) storage environments:

- Having hosts running across multiple subnets. Running inter-server storage traffic through a router can cause significant performance impacts. If this is part of the design, ensure that the network team is aware. This configuration can also cause latency challenges. Therefore, it is a preferred practice to have all of the compute nodes running on the same subnet in a SAN-less (local disk) topology.
- Slow middle-man switches. Perhaps you have two racks of servers, each with its own set of 10-GbE top-of-rack switches. The connection between these two racks is important. A single 10-GbE or even 40- GbE link between the top-of-rack switches will not be enough. You should have eight or more times the throughput to the edge server (assuming 20 servers per rack). In this example, you need at least 80-GbE throughput between the racks.
- Latency hops. The number of hops you have between the racks can affect performance. If you have three or more racks, use a spine/leaf network design between the racks. This configuration reduces the number of hops between the racks.

If there is too much network congestion, and the servers cannot communicate with each other, that can affect the cluster's health. If the cluster is having stability issues, messages should appear in the IBM PowerVC console to indicate the issues.

1.5 Enabling technologies

Now that you understand the topologies that can be built with SDI, you can learn more about what that software is and what it is doing.

1.5.1 IBM Spectrum Scale

IBM PowerVC 1.4.1 relies on IBM Spectrum Scale (V5.1.0.1 and higher) for its SDS capabilities. IBM Spectrum Scale provides several key facilities for IBM PowerVC:

- A globally parallel file system across many servers.

  If a volume is shared by many VMs and they are accessing it at the same time, IBM Spectrum Scale ensures that access to the disk is kept consistent.
- Redundancy for local disks through its File Placement Optimizer (FPO) mode:
  - If you are using systems with local disks and one of those servers (or disks within the server) dies, you will not lose data.
  - IBM Spectrum Scale ensures that when there are multiple hosts, the blocks of data are kept on multiple servers to prevent failures, minimize outages, and maintain data integrity.

- Abstracts the block device:
  - SDI is about hardware choice. IBM Spectrum Scale enables IBM PowerVC to use a wider range of hardware devices, but retain a consistent experience on top of them.
  - Clusters can be backed by local disks or iSCSI disks, in addition to traditional Fibre Channel.
- Flash Cache

  IBM Spectrum Scale allows IBM PowerVC to accelerate certain workloads with a solid-state drive (SSD). It uses the High Availability Write Cache (HAWC) and Local Read-Only Cache (LROC) to provide faster access to data.

IBM Spectrum Scale is a mature product (it was originally branded as IBM General Parallel File System (IBM GPFS™)). As such, IBM PowerVC builds on this maturity to provide great capabilities with tested stability.

The IBM PowerVC integration of IBM Spectrum Scale makes it even easier for you by providing the following features:

- Automatic cluster deployment:
  - If you have a KVM or PowerVM SDI system, IBM PowerVC detects this configuration and automatically deploys the IBM Spectrum Scale cluster for you.
  - IBM PowerVC detects the components of the system and builds the correct configuration for you.
- Simple cluster upgrade:
  - SDI means that there are software components in your infrastructure. As such, they receive updates.
  - IBM PowerVC can roll out the updates to the IBM Spectrum Scale cluster for you by building on top of its Host Maintenance mode.

There is significantly more information available about IBM Spectrum Scale. However, the core idea is that IBM PowerVC uses IBM Spectrum Scale to provide new infrastructure deployment strategies. IBM PowerVC also fully manages the interface to IBM Spectrum Scale, so the system administrator should not be interacting with it directly. Instead, the administrator should be working through IBM PowerVC.

1.5.2 PowerVM NovaLink

The PowerVM NovaLink feature was released in 2015. This is a thin partition for PowerVM that provides virtualization management capabilities. It controls the VIOSes, creates partitions, deploys VLANs on Shared Ethernet, and more.

This PowerVM capability can be used by IBM PowerVC (or other OpenStack solutions) to distribute its management software across the environment. This feature allows for a dramatic increase in scalability for IBM PowerVC (up to 200 hosts) compared to managing through the HMC, where IBM PowerVC can manage only up to 30 hosts.

This capability can be used independently of any decision to use SDI.

1.5.3 PowerVM Open I/O

IBM released PowerVM Open I/O as a way to build cloud-based I/O models. The PowerVM Open I/O model extends the PowerVM NovaLink partition to also provide I/O virtualization.

Figure 1-9 provides an overview of the topology of a PowerVM Open I/O based system.

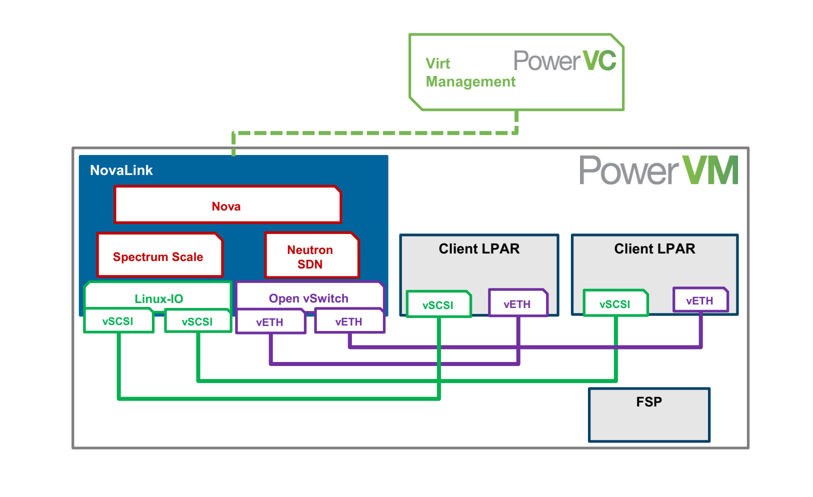


Figure 1-9 PowerVM Open I/O layout

In this model, all of the adapters in the system are attached to the PowerVM NovaLink partition. The PowerVM NovaLink partition then serves essentially as a VIOS, but is instead built on top of Linux.

When paired with SDI, PowerVM Open I/O uses IBM Spectrum Scale to virtualize its storage. It also uses OVS to virtualize its networking.

When running in SDI mode (or any configuration that uses PowerVM Open I/O), there is not a redundant LPAR like when running with a standard VIOS configuration. So, if the PowerVM NovaLink LPAR goes down, the client LPARs lose access to the network and storage.

Therefore, it is important to build these systems with redundancy in the I/O (such as link aggregations) and to use IBM PowerVC capabilities such as live migration and host maintenance mode to enable maintenance windows.

1.6 Planning for a software-defined infrastructure

To set up your environment for SDI, some planning is required. This section outlines the software and hardware requirements that are needed to build an IBM PowerVC V1.4.1 cloud that uses SDI.

1.6.1 Necessary software

IBM Spectrum Scale is the key component of an IBM PowerVC SDS solution. The IBM Spectrum Scale software can be bought independently. However, there is also an IBM offering that is called IBM Cloud™ PowerVC Manager for SDI that includes both the IBM PowerVC software and IBM Spectrum Scale V5.1 Data Management Edition. This offering can be used to gain initial entitlement to the required IBM software.

To run SDI, a few other pieces of software are needed. Table 1-1 describes the software requirements.

Table 1-1 Unified IBM PowerVC SDI software requirements

Software	Description
IBM PowerVC V1.4.1	This is the management controller for the cloud. It is used to deploy and manage workloads. It also serves as the controller for the compute, network, and storage planes.
IBM Spectrum Scale V5.1.0.1 (Data Management Edition)	Provides the storage virtualization in an SDI environment. This software must be put on the IBM PowerVC server. After it is there, IBM PowerVC distributes and installs it on all of the servers.
PowerVM NovaLink V1.0.0.10 (Part of PowerVM) with Ubuntu 16.04 - for PowerVM	Used to do all the virtualization tasks and run PowerVM Open I/O on the host servers. PowerVM NovaLink in a PowerVM environment is also available on top of Red Hat Linux. This combination is not yet supported in an IBM PowerVC SDI environment.
Ubuntu 16.04 LTS - for Open Power Abstraction Layer (OPAL)-based systems	For a compute node. Running the KVM hypervisor and some OpenStack components.
Red Hat Enterprise Linux 7.5 - for OPAL-based systems	For a compute node. Running the KVM hypervisor and some OpenStack components.

1.6.2 Environment planning

Before you start planning, you must determine what your requirements are for this environment:

- Will you use local disks or remote storage?
- What are your availability requirements (such as a dev/test environment or production)?
- What are your performance requirements?
- What hardware is available?

There are many more questions that must be asked, but each of these questions affects the architecture of the system.

1.6.3 Storage

One of the most challenging decisions in an environment is the storage. If you are using SDS within IBM PowerVC, you have two options:

- Local disks within the server
- Remote storage, either by using iSCSI or Fibre Channel LUNs

Local disks are a flexible storage solution. Use a minimum of three servers so that the system can retain three copies of the data in the event of a failure. The local disk servers also require at least one SSD so that the data access can be accelerated. An all flash strategy is also supported.

If you use a local disk strategy, be mindful of the durability of the disks that you are using. Do they support enough Drive Writes per Day (DWPDs)? What is their mean time between failures (MTBF)? Ensure that the drives that are used meet the resilience requirements for the workload.

The local disk strategy takes advantage of the FPO mode of IBM Spectrum Scale. In this mode, the environment retains three copies (assuming three servers or more) of the data. This feature ensures that if there is a disk or server failure, it does not impact the cluster or workloads. If you run with two nodes, only two copies of the data are retained. If you run with a single node, then data is potentially at risk during a drive failure. This configuration is not recommended and should be used only for evaluation purposes.

Remote storage is also supported. From each SAN device, either Fibre Channel or iSCSI, drives can be mounted to all of the servers within the cluster. These LUNs then are used to back the cluster. In this mode, IBM PowerVC and IBM Spectrum Scale assert that the backing SAN device provides the redundancy. Therefore, only one copy of the data is retained within the IBM Spectrum Scale file system.

If you use iSCSI LUNs, make sure that the LUNs are attached as part of host startup. For more information, see the Canonical Ubuntun documentation.

1.6.4 Hypervisor

The SDI capabilities are available on both PowerVM and KVM on Power. Your workload needs determine the hypervisor that you need.

KVM on Power requires Ubuntu 16.0, or with IBM PowerVC 1.4.1 and later it is also possible to use Red Hat Enterprise Linux 7.5. These operating systems are supported on the IBM Power Systems scale-out systems (such as the Power S922 or Power S822 servers), scale-out Linux (such as Power L922, Power S812L, or Power S822LC servers), and others. The key requirement is that the hypervisor is running on OPAL firmware.

To run PowerVM, your firmware must support the PowerVM mode (such as for the Power S922, Power S822, or Power S812L servers).

1.6.5 Networking

Although the networking plane can be dramatically simplified (down to a single wire if needed), it is preferable to separate the traffic onto different physical networks.

Table 1-2 lists three types of traffic running over the Ethernet.

Table 1-2 Three types of internet traffic

Traffic type	Description
Data	The data that is sent from the VMs.
Storage	The Ethernet network that drives any storage operations. This network might be similar to your Fibre Channel network, but runs over Ethernet.
Management	The network that runs the management operations, for example, the communication between IBM PowerVC and the servers.

For IBM PowerVC V1.4.1, there is a restriction that the storage and management data must be on the same network. This requirement simplifies the deployment model and generally matches the requirement where both storage and management traffic must be isolated.

A typical configuration is shown in Figure 1-10.

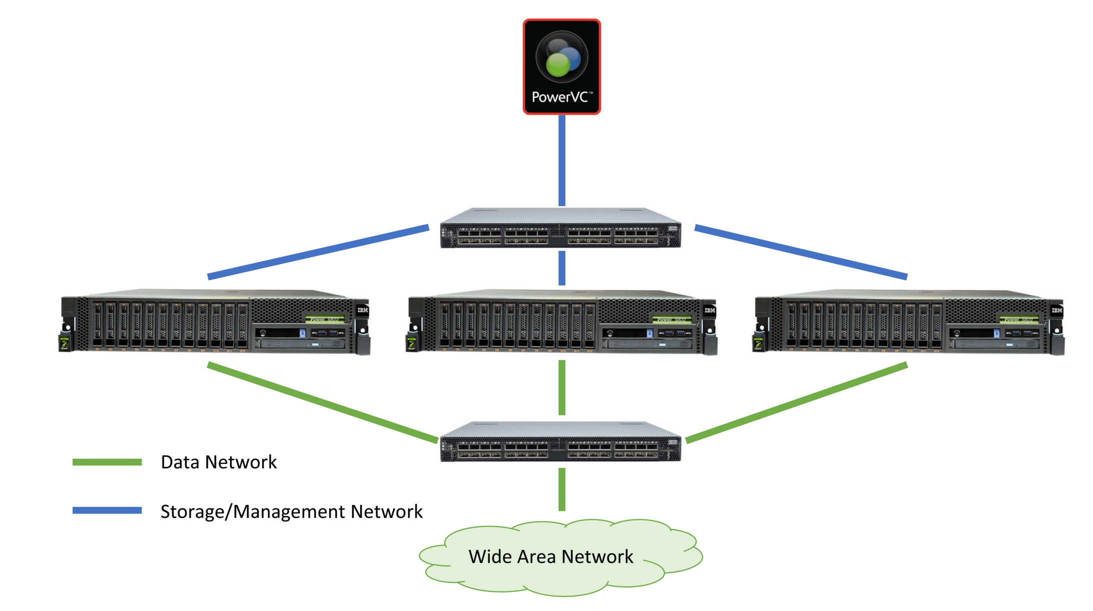


Figure 1-10 Typical IBM PowerVC configuration

In this environment, the servers run their data network independently of their storage/management network. However, with a network that is fast enough, this traffic can be converged into a single network.

A 10-Gb network is the minimum speed network that is supported. In general, run your storage network at 25 GbE or faster.

2

Planning and implementing an IBM PowerVC software-defined infrastructure environment

This chapter provides information about planning and setting up the prerequisites, nodes, and IBM PowerVC to create an IBM PowerVC software-defined infrastructure (SDI) environment.

2.1 Planning

This section describes some details about planning for an SDI cloud.

2.1.1 Cloud compute nodes

The cloud compute nodes are the physical servers where workloads run. Regarding these nodes, a few decisions must be made:

- What type of storage will be used?
  - Local disks can create a storage area network (SAN)-less, Ethernet only cloud. The cloud uses the disks in the system to create a cluster across all of the nodes.
  - External storage uses either iSCSI or Fibre Channel to create a storage pool that is backed by the SAN. The storage provides a flexible decision about what backing storage to use.
- Which hypervisor will be used?
  - A kernel-based virtual machine (KVM) on Power with Ubuntu 16.04 LTS or Red Hat Enterprise Linux 7.5 supports Linux based workloads and runs on various Power Systems servers, including the Power Systems LC line.
  - PowerVM supports Linux, IBM AIX, and has a technology preview that is available for IBM i when running with software-defined storage (SDS).

The cloud compute node has these hardware requirements:

- At least 64 GB of memory
- At least eight cores (ideally 16 or more)
- Storage that is provided either locally or externally
- At least 10 GbE:
  - If you use local disks or iSCSI, use at least 2 GbE.
  - It is preferable to have a separate Ethernet network for storage versus data, but it is not required.

2.1.2 Storage choice

You have two choices for storage: local disks or external storage. All of the disks within the cluster must be the same type. If the system detects that the server has any external storage, it assumes that external storage is the preferred type.

Local storage

If you are using local storage, you must make sure that each host has at least five disks. It is preferable to have more, but five is the minimum.

Table 2-1 shows the disk type requirements.

Table 2-1 Disk type requirements

Disk type	Minimum	Maximum	Purpose
Boot	1	2	The boot disk for the hypervisor. It is not used by the SDS pool.
IBM Spectrum Scale Metadata	1	1	Provides the metadata storage for the cluster. This disk must be a solid-state drive (SSD) for fast, random 4-KB operations. It does not need to be large (100 GB or more). The system automatically picks the metadata disk. The logic that it uses is the smallest SSD over 100 GB. The metadata disk is also used as a flash cache (IBM Spectrum Scale Local Read-Only Cache (LROC) and High Availability Write Cache (HAWC)).
IBM Spectrum Scale Data	3	24	Used for the physical storage of the data. These disks can be either hard disk drives (HDDs) or SSDs. Use the appropriate disk type for your workloads. It is highly preferable that all of the disks be the same size in the server.

All of the servers in the cluster should have a similar storage configuration. This configuration reduces the risk of hot zones in the storage cluster. For example, it can create data loss opportunities and performance bottlenecks if you have one server with 100 TB of disk and another with 5 TB of disk.

Homogeny across the disks and servers is highly preferred.

External storage

Storage can be provided through an external connection, either Fibre Channel or iSCSI. Here are a few considerations when evaluating external storage:

- The storage must be attached to all of the hosts in the cluster. With this setup, IBM Spectrum Scale can optimize its reads/writes across the cluster. It also allows for quick failover of nodes if there is an error on a server.
- If using iSCSI, you must make sure that the disks are discovered as part of the server start sequence because they must be available at the end of the sequence. If not, the cluster fails to start on that node.
- If you are using iSCSI, also note the Ethernet port that is being used to host the disk. Is that port being shared with the data connections? If so, that might cause some bandwidth challenges between the two. Make sure that you separate the ports or provide enough bandwidth through the common port.

2.1.3 Cloud installation

Up to this point, we have focused on an overview of what a software-defined environment (SDE) is and its infrastructure layout. Now, the required software components must be installed and configured. This process is generally broken down into the following steps:

1. Prepare your cloud compute nodes.

   This step is covered in 2.2, “Installing PowerVM NovaLink for a system that uses PowerVM” on page 22.

2. (Optional) Prepare your network nodes.

   This step is needed only if you plan on taking advantage of network overlays within your cloud. If you plan to use virtual local area network (VLAN) or flat (untagged) networking, this step is not required. You do not need to use Virtual Extensible LAN (VXLAN) networks with SDI because the compute nodes can directly tag VLAN traffic.

3. Install IBM PowerVC.

   You install the cloud controller software. For more information, see *IBM PowerVC Version 1.3.2 Introduction and Configuration*, SG24-8199.

4. Connect the compute nodes to IBM PowerVC.

   Make the controller aware of the hardware. This process also configures the SDS pool. This step is covered in 3.3, “Adding compute nodes to IBM PowerVC” on page 66.

5. Configure IBM PowerVC.

   Make IBM PowerVC aware of the hardware, which includes building your networks, building images, and so on.

6. Deploy workloads.

2.2 Installing PowerVM NovaLink for a system that uses PowerVM

Before installing PowerVM NovaLink in our example, we prepared a network installation host to do the installation over the network. Because the hosts were connected to an HMC, some preparation was needed. Before the installation of the systems with PowerVM NovaLink, all LPARs, including Virtual I/O Server (VIOS) or predefined LPARs (for a new system) were deleted. The starting point was an empty system that is connected to an HMC.

In our lab environment, we used PowerVM NovaLink Version 1.0.0.10. This version is also the prerequisite for IBM PowerVC V1.4.1.

2.2.1 Preparing the resources for a PowerVM NovaLink network installation

The installation of PowerVM NovaLink may be done by using a USB stick, a DVD, or a network installation server. In the example environment, an installation server was created by using Ubuntu as the base operating system. It is also possible to use other operating systems if all the necessary services are working.

The preparations for an PowerVM NovaLink installation server are documented in IBM Knowledge Center.

For a default Ubuntu 16.04.04 system, complete the following steps:

1. Run the command that is shown in Example 2-1.

Example 2-1 Installing extra packets

```
$ sudo apt-get install bootp tftpd-hpa apache2
Reading package lists... Done
Building dependency tree
Reading state information... Done
...
Do you want to continue? [Y/n] Y
...
Setting up update-inetd (4.43) ...
Setting up bootp (2.4.3-18) ...
Setting up tftpd-hpa (5.2+20150808-1ubuntu1.16.04.1) ...
Processing triggers for libc-bin (2.23-0ubuntu10) ...
Processing triggers for systemd (229-4ubuntu21.2) ...
Processing triggers for ureadahead (0.100.0-19) ...
Processing triggers for ufw (0.35-0ubuntu2) ...
```

2. Add the boot information for the client into `/etc/bootptab`, as shown in Example 2-2.

Example 2-2 Editing /etc/bootptab and start the services

```
$ vi /etc/bootptab
...

<myNovaLink1>:\
   bf=core.elf:\
   ip=<IP of the NovaLink VM to be installed>:\
   sm=<Subnetmask>:\
   gw=<Gateway>:\
   dn=<Domain of the network>:\
   ns=<Nameserver>:\
   sa=<Install server IP - the IP of that VM>:

$ sudo systemctl daemon-reload

$ sudo service isc-dhcp-server stop

$ sudo service bootp start
```

Note: Do not forget the “\” at the end of each line, or write everything in one line separated by spaces.

Note: In our lab, `isc-dhcp-server` was not loaded, so it does not need to be stopped.

3. Copy the `core.elf` file into the `tftpboot` directory, which is `/var/lib/tftpboot`. Then, create a `grub.cfg` file. Copy the sample from the PowerVM NovaLink ISO, as shown in Example 2-3.

Example 2-3 Copying core.elf and grub.cfg

```
$ sudo mount PowerVM_NovaLink_V1.0.0.10.iso  /mnt
mount: /dev/loop0 is write-protected, mounting read-only

$ sudo cp /mnt/pvm/core.elf /var/lib/tftpboot
$ sudo cp /mnt/pvm/sample_grub.cfg /var/lib/tftpboot/grub.cfg
```

4. Edit the `grub.cfg` file, as shown in Example 2-4.

Example 2-4 Modifying /var/lib/tftpboot/grub.cfg

```
$ sudo vi /var/lib/tftpboot
#Uncomment/change/add:
set gateway=<Gateway>
set netmask=<Netmask>
set nameserver=<Nameserver>
set hostname=<hostname>
#Add:
pvm-installmode=SDE
```

Note: After the installation, the host name is overwritten with the name of the Domain Name Server (DNS) if the IP is resolvable.

Note: Depending on the documentation link, only software-defined networking (SDN) is mentioned as a technology preview. The sample `grub.cfg` in Version 1.0.0.10 did not have that line at all. If you want SDS + SDN, use the value `SDE` for SDE.

If the installation uses the wrong network boot, you can specify the network port that should be used, for example:

```
netcfg/choose_interface=enP30p128s0f0 \
```

For a first run, leave the default:

```
netcfg/choose_interface=auto \
```

After modifying the grub.cfg, the IBM Knowledge Center shows how to create the resources for an installation of VIOSes. As a true SDE does not use VIOS, this section was skipped.

5. Get the files from the PowerVM NovaLink ISO by using the HTTP server, as shown in Example 2-5. In our environment, we set the resources to be persistent across restarts and easy to switch between versions.

Example 2-5 Getting the PowerVM NovaLink ISO files over HTTP

```
$ sudo mkdir -p /var/www/html/novalink10010
#The PowerVM NovaLink ISO is still mounted on /mnt.
$ cd /mnt
$ sudo tar -cvf /tmp/novalink.tar ./*
$ cd /var/www/html/novalink10010
$ sudo tar -xvf /tmp/novalink.tar
#The real destination directory is PowerVM NovaLink without the version.
$ sudo ln -s /var/www/html/novalink10010 var/www/html/novalink
$ sudo mkdir /var/www/html/novalink-repo10010
$ cd /var/www/html/novalink-repo10010
$ sudo tar -xzvf /mnt/pvm/repo/pvmrepo.tgz
#The real destination directory is novalink-repo without the version.
$ sudo ln -s /var/www/html/novalink-repo10010 var/www/html/novalink-repo

$ ls -al /var/www/html
total 28
drwxr-xr-x  4 root root  4096 Jun 13 17:45 .
drwxr-xr-x  3 root root  4096 Jun 12 10:35 ..
-rw-r--r--  1 root root 11321 Jun 12 10:35 index.html
lrwxrwxrwx  1 root root    12 Jun 12 15:47 novalink -> novalink10010
drwxr-xr-x 12 root root  4096 Jun 12 15:26 novalink10010
lrwxrwxrwx  1 root root    17 Jun 13 17:45 novalink-repo -> novalink-repo10010
drwxr-xr-x  6 root root  4096 Jun 13 17:45 novalink-repo10010
```

6. To test whether the services are running, run the commands that are shown in Example 2-6.

Example 2-6 Verifying that the services are running

```
$ sudo service bootp status
° bootp.service - Bootp Service
   Loaded: loaded (/etc/systemd/system/bootp.service; enabled; vendor preset: 
enabled
   Active: active (running) since Wed 2018-06-13 16:36:56 EDT; 1h 43min ago

$ sudo service apache2 status
° apache2.service - LSB: Apache2 web server
   Loaded: loaded (/etc/init.d/apache2; bad; vendor preset: enabled)
  Drop-In: /lib/systemd/system/apache2.service.d
           ··apache2-systemd.conf
   Active: active (running) since Wed 2018-06-13 16:36:56 EDT; 1h 43min ago

$ sudo service tftpd-hpa status
° tftpd-hpa.service - LSB: HPA's tftp server
   Loaded: loaded (/etc/init.d/tftpd-hpa; bad; vendor preset: enabled)
   Active: active (running) since Wed 2018-06-13 16:36:56 EDT; 1h 43min ago
```

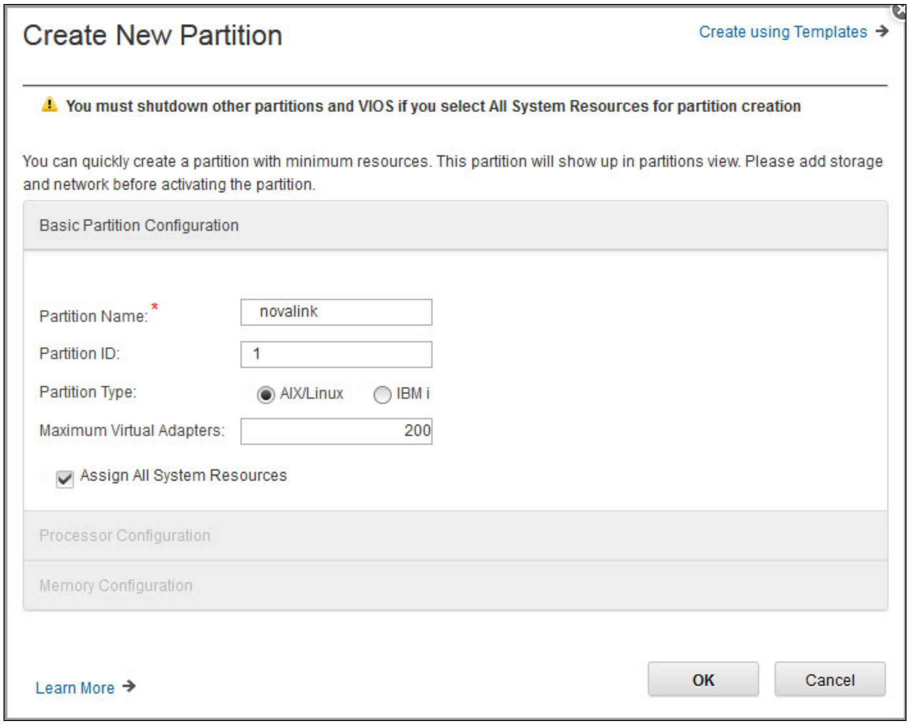
2.2.2 Installing a PowerVM NovaLink partition on an empty system that is under control of an HMC

There are several ways to install a PowerVM NovaLink partition. There are two choices for the operating system, which are Ubuntu and Red Hat. If you download the PowerVM NovaLink ISO, it includes Ubuntu. You may use Red Hat Enterprise Linux, but it is not supported in an SDI environment. In our environment, we decided to use Ubuntu.

If you want to install PowerVM NovaLink on a new system, you can use the default partition on that system because the PowerVM NovaLink installer changes the configuration of the PowerVM NovaLink LPAR.

To install the partition, complete the following steps:

1. Create a partition with all resources (like the default partition). Figure 2-1 shows the creation of a default partition.



Create New Partition

Create using Templates →

⚠ **You must shutdown other partitions and VIOS if you select All System Resources for partition creation**

You can quickly create a partition with minimum resources. This partition will show up in partitions view. Please add storage and network before activating the partition.

Basic Partition Configuration

Partition Name:* novalink

Partition ID: 1

Partition Type: ● AIX/Linux ○ IBM i

Maximum Virtual Adapters: 200

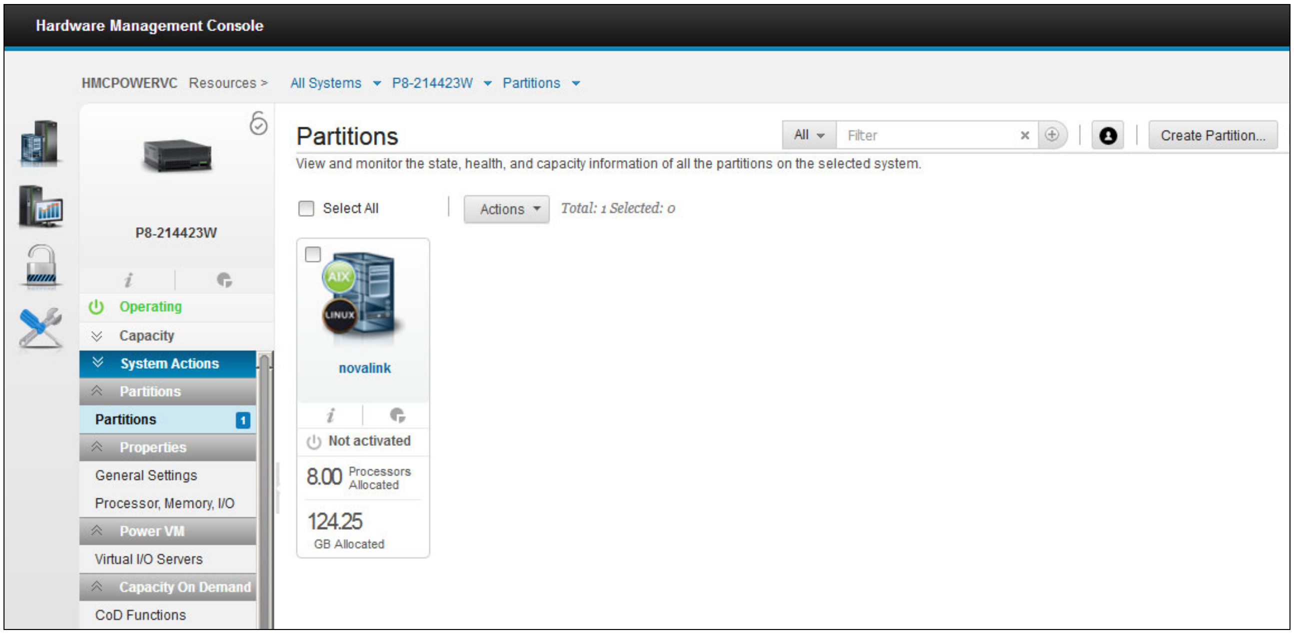
☑ Assign All System Resources

Processor Configuration

Memory Configuration

Figure 2-1 Default partition with all resources

Figure 2-2 shows that the partition appears under the Partitions pane.



Partitions

View and monitor the state, health, and capacity information of all the partitions on the selected system.

☐ Select All

Total: 1 Selected: 0

novalink

Not activated

8.00 Processors Allocated

124.25 GB Allocated

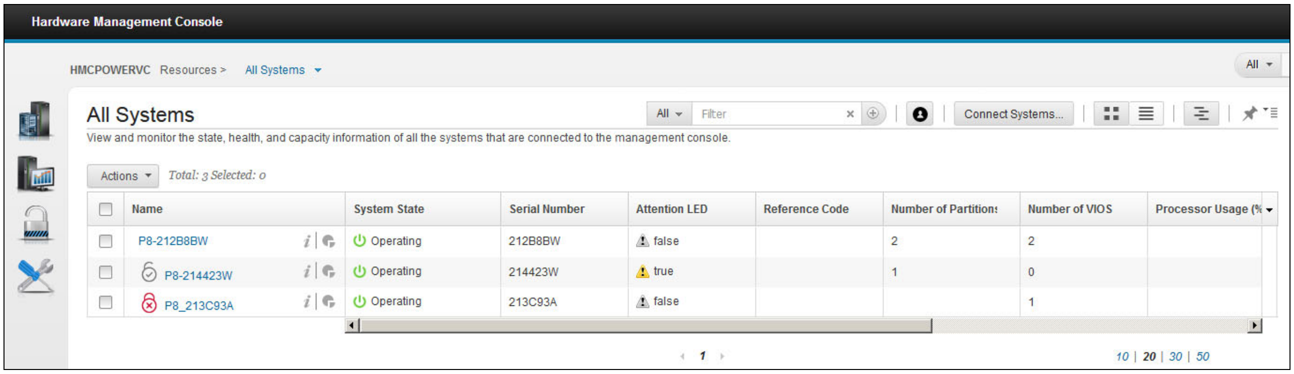
Figure 2-2 Default partition under partitions

2. Use the HMC to maker this partition a PowerVM NovaLink partition so that the HMC can take control of the virtualization by using the HMC command-line interface (CLI), as shown in Example 2-7.

Example 2-7 Making the partition a PowerVM NovaLink partition

```
$ chsyscfg -m P8-214423W -r lpar -o apply --id 1
$ chcomgmt -m P8-214423W -o setmaster -t norm
$ chsyscfg -m P8-214423W -r lpar -i lpar_id=1,powervm_mgmt_capable=1
```

The **`chcomgmt -o setmaster`** command places the HMC in control of the virtualization, but this situation will change. The HMC GUI shows a PowerVM NovaLink partition, as shown in Figure 2-3.



View and monitor the state, health, and capacity information of all the systems that are connected to the management console.

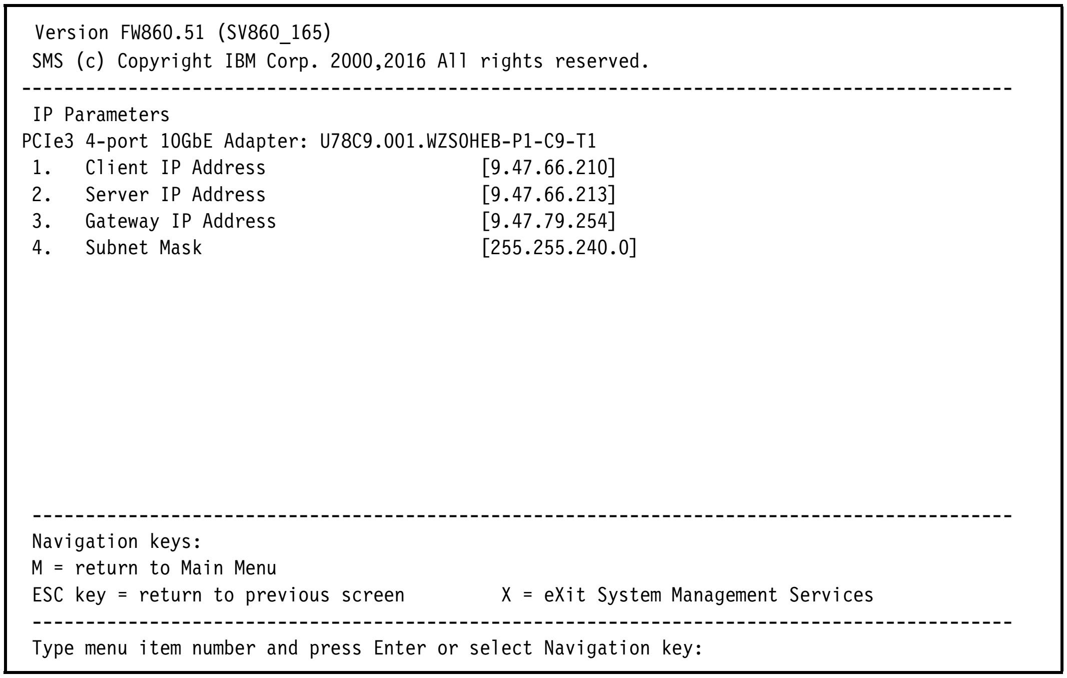
Total: 3 Selected: 0

□	Name	System State	Serial Number	Attention LED	Reference Code	Number of Partition:	Number of VIOS	Processor Usage (%
□	P8-212B8BW	Operating	212B8BW	false		2	2	
□	P8-214423W	Operating	214423W	true		1	0	
□	P8_213C93A	Operating	213C93A	false			1	

Figure 2-3 Two systems with PowerVM NovaLink

Figure 2-3 shows three systems. The two systems that have lock icons are PowerVM NovaLink partitions. The first lock is open, which means that the HMC is in control of the virtualization. The second, closed red lock shows that for this system PowerVM NovaLink has control.

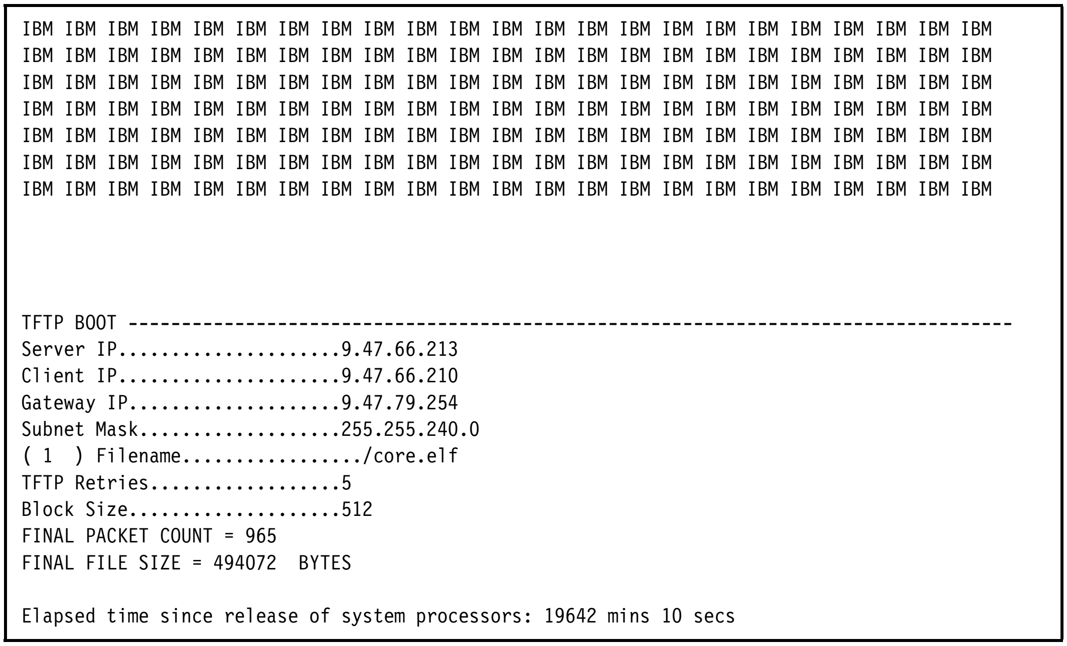
3. The partition is ready for installation. After switching it on, select one of working network adapters to boot from the prepared installation partition. Start the LPAR, start SMS, and go to the `Setup Remote IPL` → `Adapter` → `IPv4` → `BOOTP` → `IP parameters` menu, as shown in Figure 2-4. Do a ping test to see whether the installation system responds.



```
 Version FW860.51 (SV860_165)
 SMS (c) Copyright IBM Corp. 2000,2016 All rights reserved.
--------------------------------------------------------------------------------------------
 IP Parameters
PCIe3 4-port 10GbE Adapter: U78C9.001.WZS0HEB-P1-C9-T1
 1.   Client IP Address                   [9.47.66.210]
 2.   Server IP Address                   [9.47.66.213]
 3.   Gateway IP Address                  [9.47.79.254]
 4.   Subnet Mask                         [255.255.240.0]










 -------------------------------------------------------------------------------------------
 Navigation keys:
 M = return to Main Menu
 ESC key = return to previous screen         X = eXit System Management Services
 -------------------------------------------------------------------------------------------
 Type menu item number and press Enter or select Navigation key:
```

Figure 2-4   SMS IPL configuration

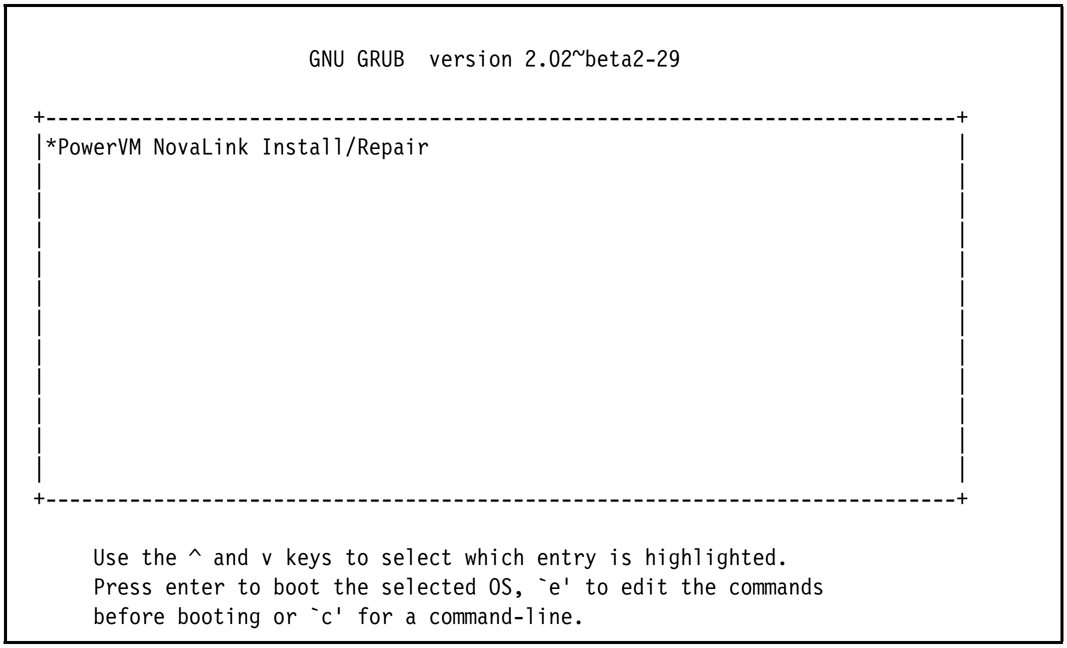
4. Start the partition over the prepared network adapter, as shown in Figure 2-5.



```
IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM
IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM
IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM
IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM
IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM
IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM
IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM




TFTP BOOT ---------------------------------------------------------------------------------
Server IP.....................9.47.66.213
Client IP.....................9.47.66.210
Gateway IP....................9.47.79.254
Subnet Mask...................255.255.240.0
( 1  ) Filename.................../core.elf
TFTP Retries..................5
Block Size....................512
FINAL PACKET COUNT = 965
FINAL FILE SIZE = 494072  BYTES

Elapsed time since release of system processors: 19642 mins 10 secs
```

Figure 2-5   Starting over the network

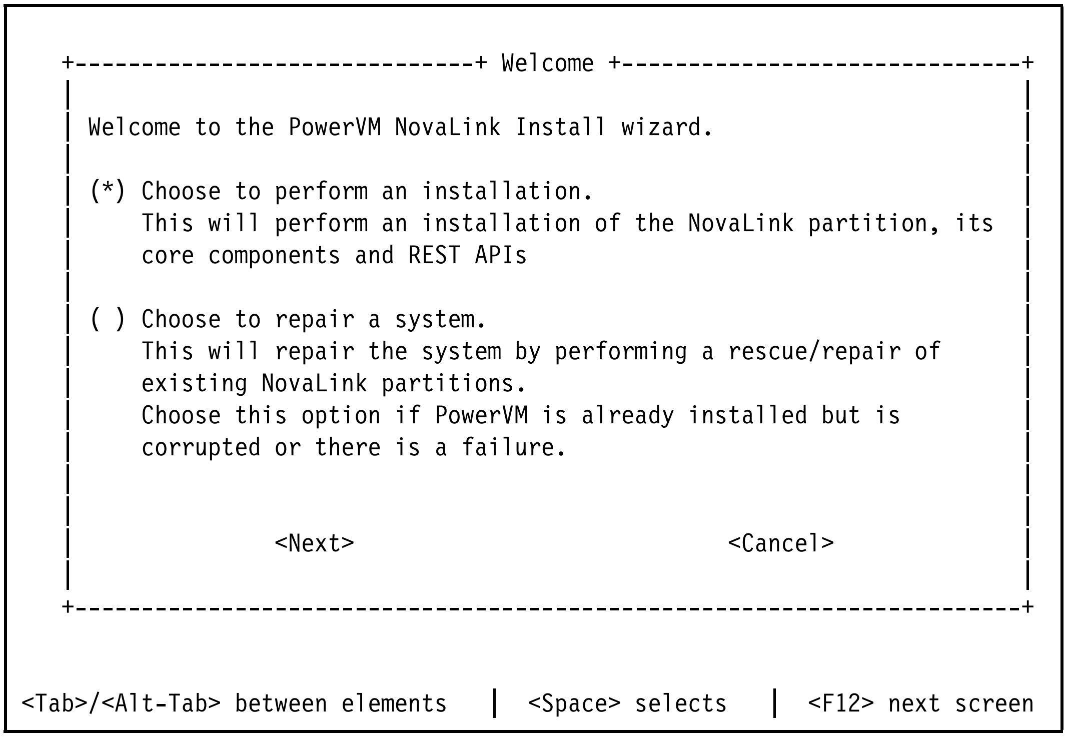
After some time, the PowerVM NovaLink network installer appears, as shown in Figure 2-6.



```
                           GNU GRUB  version 2.02~beta2-29

 +--------------------------------------------------------------------------------------+
 |*PowerVM NovaLink Install/Repair                                                      |
 |                                                                                      |
 |                                                                                      |
 |                                                                                      |
 |                                                                                      |
 |                                                                                      |
 |                                                                                      |
 |                                                                                      |
 |                                                                                      |
 |                                                                                      |
 |                                                                                      |
 |                                                                                      |
 +--------------------------------------------------------------------------------------+

        Use the ^ and v keys to select which entry is highlighted.
        Press enter to boot the selected OS, `e' to edit the commands
        before booting or `c' for a command-line.
```

Figure 2-6   The PowerVM NovaLink installer

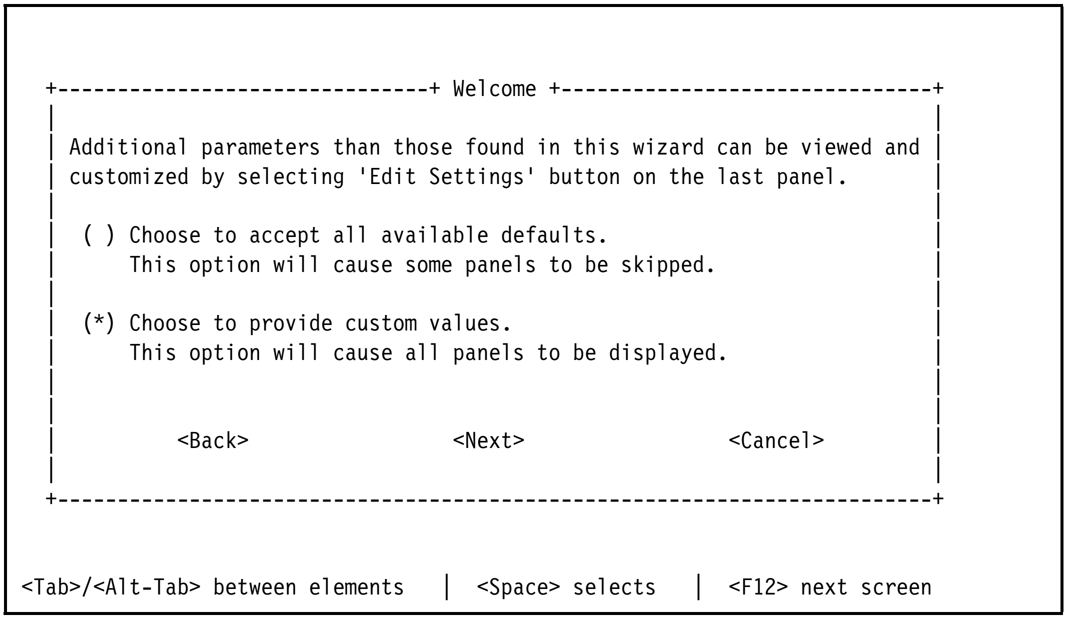
In our example, we had an issue with the PowerVM NovaLink installer configuring the wrong network adapter. Therefore, we specified the correct connected network adapter, as described in 2.2.1, “Preparing the resources for a PowerVM NovaLink network installation” on page 22. After specifying the correct interface and restarting, the system starts with the PowerVM NovaLink Install wizard, as shown in Figure 2-7.



```
   +-----------------------------+ Welcome +-----------------------------+
   |                                                                     |
   | Welcome to the PowerVM NovaLink Install wizard.                     |
   |                                                                     |
   | (*) Choose to perform an installation.                              |
   |     This will perform an installation of the NovaLink partition, its |
   |     core components and REST APIs                                   |
   |                                                                     |
   | ( ) Choose to repair a system.                                      |
   |     This will repair the system by performing a rescue/repair of    |
   |     existing NovaLink partitions.                                   |
   |     Choose this option if PowerVM is already installed but is      |
   |     corrupted or there is a failure.                                |
   |                                                                     |
   |                                                                     |
   |              <Next>                             <Cancel>            |
   |                                                                     |
   +---------------------------------------------------------------------+


<Tab>/<Alt-Tab> between elements   |  <Space> selects   |  <F12> next screen
```

Figure 2-7 PowerVM NovaLink installation wizard

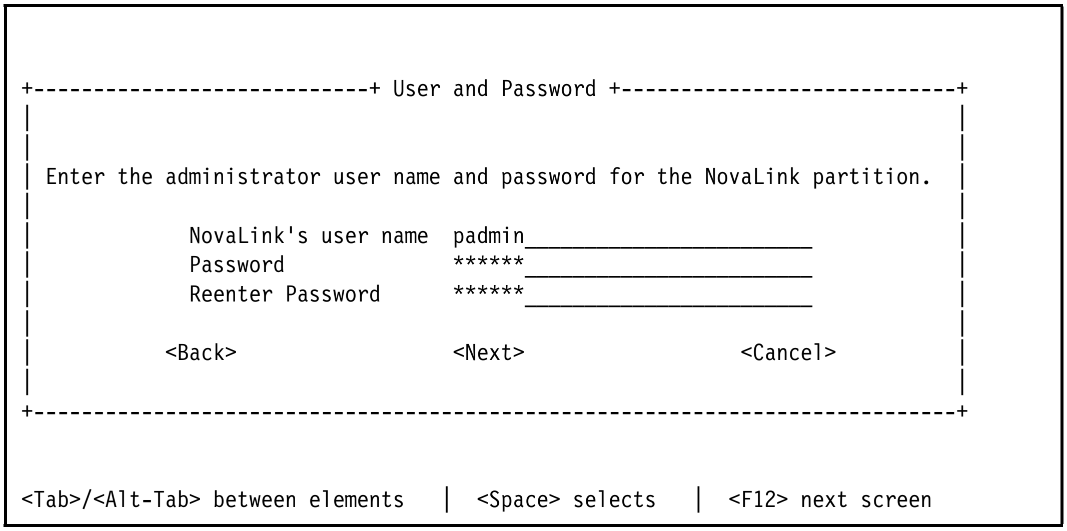
5. You may choose the default configuration or modify certain values. We selected the custom configuration, as shown in Figure 2-8.



```
  +-------------------------------+ Welcome +--------------------------------+
  |                                                                          |
  | Additional parameters than those found in this wizard can be viewed and |
  | customized by selecting 'Edit Settings' button on the last panel.       |
  |                                                                          |
  |  ( ) Choose to accept all available defaults.                           |
  |      This option will cause some panels to be skipped.                  |
  |                                                                          |
  |  (*) Choose to provide custom values.                                   |
  |      This option will cause all panels to be displayed.                 |
  |                                                                          |
  |                                                                          |
  |          <Back>                <Next>                <Cancel>           |
  |                                                                          |
  +--------------------------------------------------------------------------+


<Tab>/<Alt-Tab> between elements   |  <Space> selects   |  <F12> next screen
```

Figure 2-8   PowerVM NovaLink installation wizard

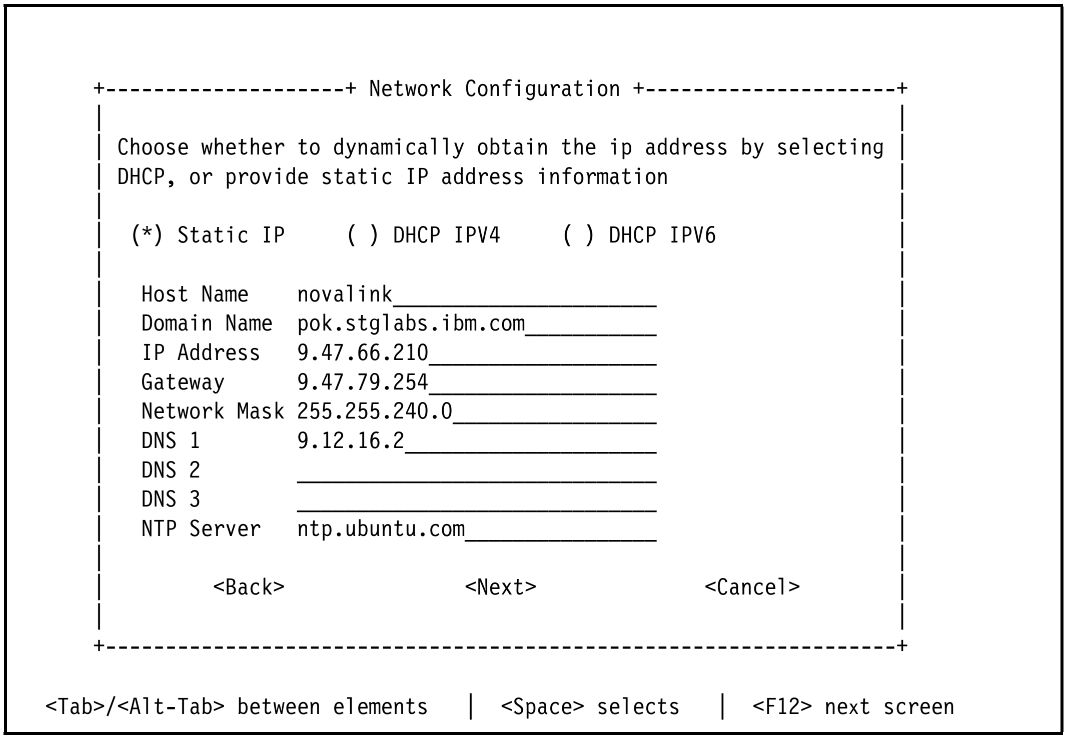
6. After you accept the license agreement, you are prompted for a user name and password, as shown in Figure 2-9.



```
+----------------------------+ User and Password +----------------------------+
|                                                                              |
|                                                                              |
| Enter the administrator user name and password for the NovaLink partition.  |
|                                                                              |
|            NovaLink's user name  padmin________________________             |
|            Password              ******________________________             |
|            Reenter Password      ******________________________             |
|                                                                              |
|          <Back>                  <Next>                  <Cancel>           |
|                                                                              |
+------------------------------------------------------------------------------+


<Tab>/<Alt-Tab> between elements   |  <Space> selects   |  <F12> next screen
```

Figure 2-9   PowerVM NovaLink user and password

7. In Figure 2-10, you can modify the network settings. The values are predefined from the `grub.cfg` file.

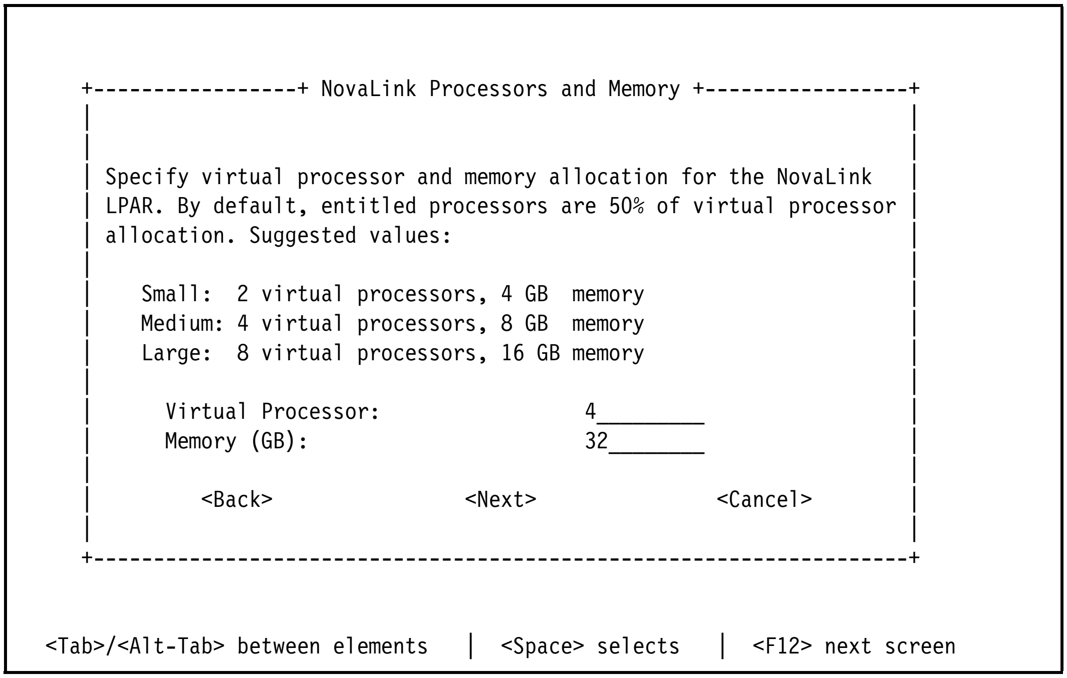


```
+--------------------+ Network Configuration +---------------------+
|                                                                  |
| Choose whether to dynamically obtain the ip address by selecting |
| DHCP, or provide static IP address information                   |
|                                                                  |
|  (*) Static IP     ( ) DHCP IPV4     ( ) DHCP IPV6               |
|                                                                  |
|   Host Name    novalink______________________                    |
|   Domain Name  pok.stglabs.ibm.com___________                    |
|   IP Address   9.47.66.210___________________                    |
|   Gateway      9.47.79.254___________________                    |
|   Network Mask 255.255.240.0_________________                    |
|   DNS 1        9.12.16.2_____________________                    |
|   DNS 2        ______________________________                    |
|   DNS 3        ______________________________                    |
|   NTP Server   ntp.ubuntu.com________________                    |
|                                                                  |
|         <Back>              <Next>             <Cancel>          |
|                                                                  |
+------------------------------------------------------------------+

<Tab>/<Alt-Tab> between elements   |  <Space> selects   |  <F12> next screen
```

Figure 2-10 PowerVM NovaLink installation wizard: Network configuration

8. In Figure 2-11, you can set the processor and memory configuration. If you use the defaults, this screen will not open and the system uses the defaults with four processors (an entitled capacity of two cores) and 32 GB of memory. The values are higher for an SDE and a PowerVM NovaLink environment without SDE. For an SDE environment, do not go below 16 GB of memory and two virtual processors; this configuration works for a small environment, but it is preferable to use the default settings.

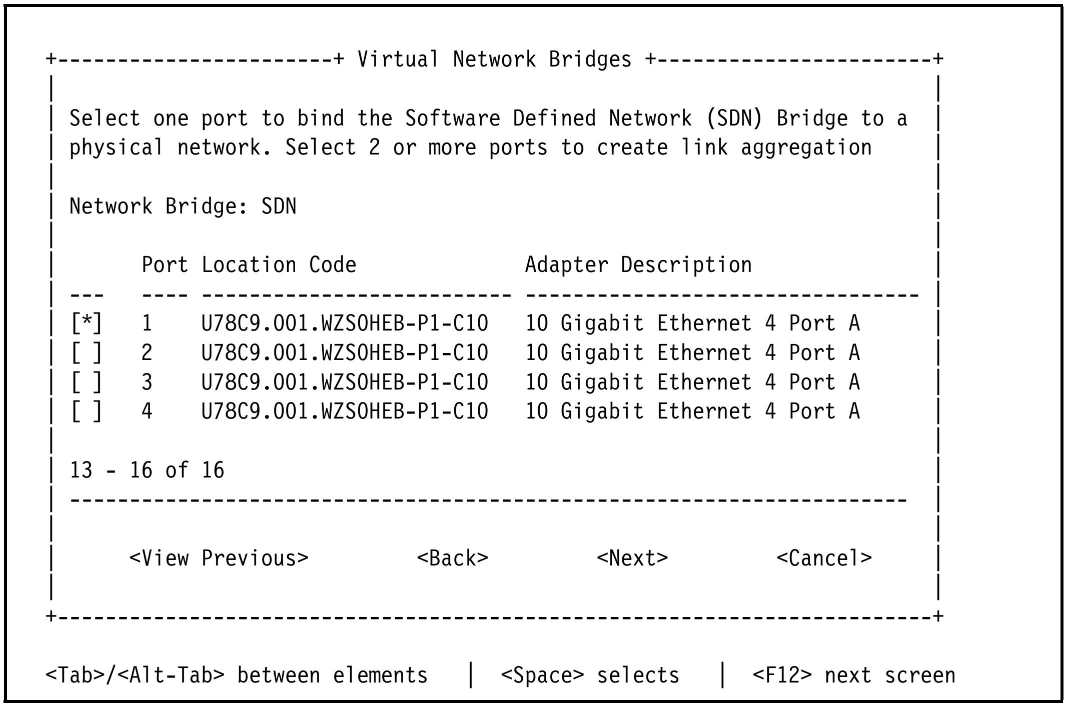
   As a preferred practice, specify about 15% of the resources of the system and no less than 16 GB of memory.



```
+-----------------+ NovaLink Processors and Memory +-----------------+
|                                                                   |
|                                                                   |
| Specify virtual processor and memory allocation for the NovaLink  |
| LPAR. By default, entitled processors are 50% of virtual processor |
| allocation. Suggested values:                                     |
|                                                                   |
|    Small:  2 virtual processors, 4 GB  memory                     |
|    Medium: 4 virtual processors, 8 GB  memory                     |
|    Large:  8 virtual processors, 16 GB memory                     |
|                                                                   |
|      Virtual Processor:                4_________                 |
|      Memory (GB):                      32________                 |
|                                                                   |
|         <Back>               <Next>               <Cancel>        |
|                                                                   |
+-------------------------------------------------------------------+


<Tab>/<Alt-Tab> between elements   |  <Space> selects   |  <F12> next screen
```

Figure 2-11 PowerVM NovaLink installation wizard: Processor and memory configuration

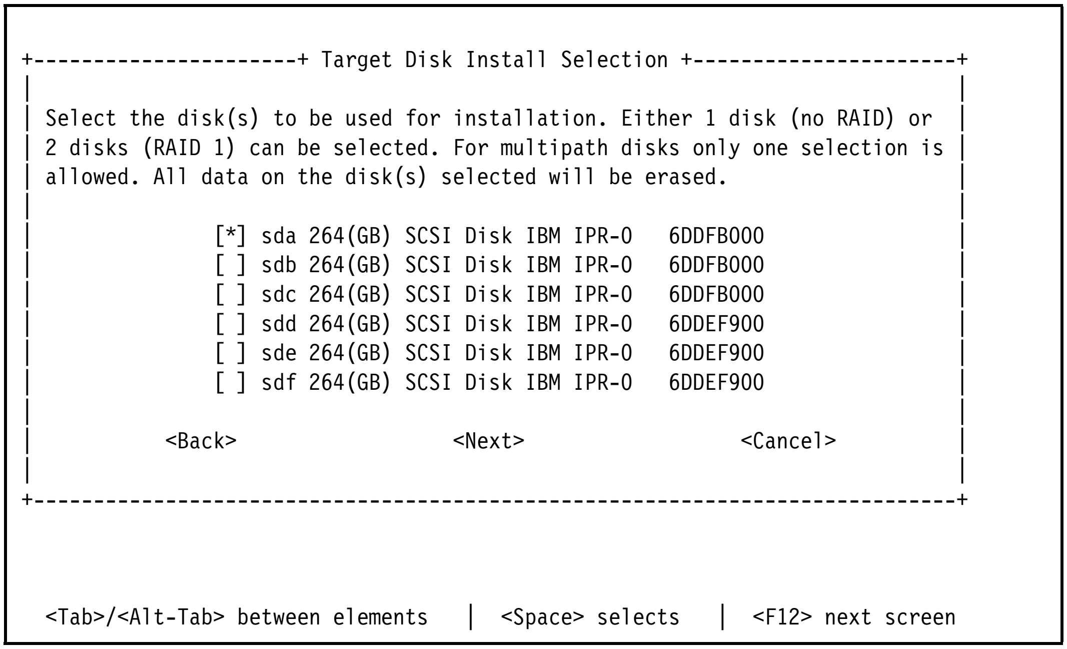
9. Select the adapter for the Open vSwitch (OVS) bridge, as shown in Figure 2-12.



```
+-----------------------+ Virtual Network Bridges +-----------------------+
|                                                                          |
| Select one port to bind the Software Defined Network (SDN) Bridge to a  |
| physical network. Select 2 or more ports to create link aggregation     |
|                                                                          |
| Network Bridge: SDN                                                      |
|                                                                          |
|       Port Location Code             Adapter Description                |
| ---   ---- -------------------------- --------------------------------- |
| [*]   1    U78C9.001.WZS0HEB-P1-C10   10 Gigabit Ethernet 4 Port A       |
| [ ]   2    U78C9.001.WZS0HEB-P1-C10   10 Gigabit Ethernet 4 Port A       |
| [ ]   3    U78C9.001.WZS0HEB-P1-C10   10 Gigabit Ethernet 4 Port A       |
| [ ]   4    U78C9.001.WZS0HEB-P1-C10   10 Gigabit Ethernet 4 Port A       |
|                                                                          |
| 13 - 16 of 16                                                            |
| ----------------------------------------------------------------------  |
|                                                                          |
|      <View Previous>        <Back>        <Next>        <Cancel>         |
|                                                                          |
+--------------------------------------------------------------------------+

<Tab>/<Alt-Tab> between elements   |  <Space> selects   |  <F12> next screen
```

Figure 2-12 PowerVM NovaLink installation wizard: Network adapter selection

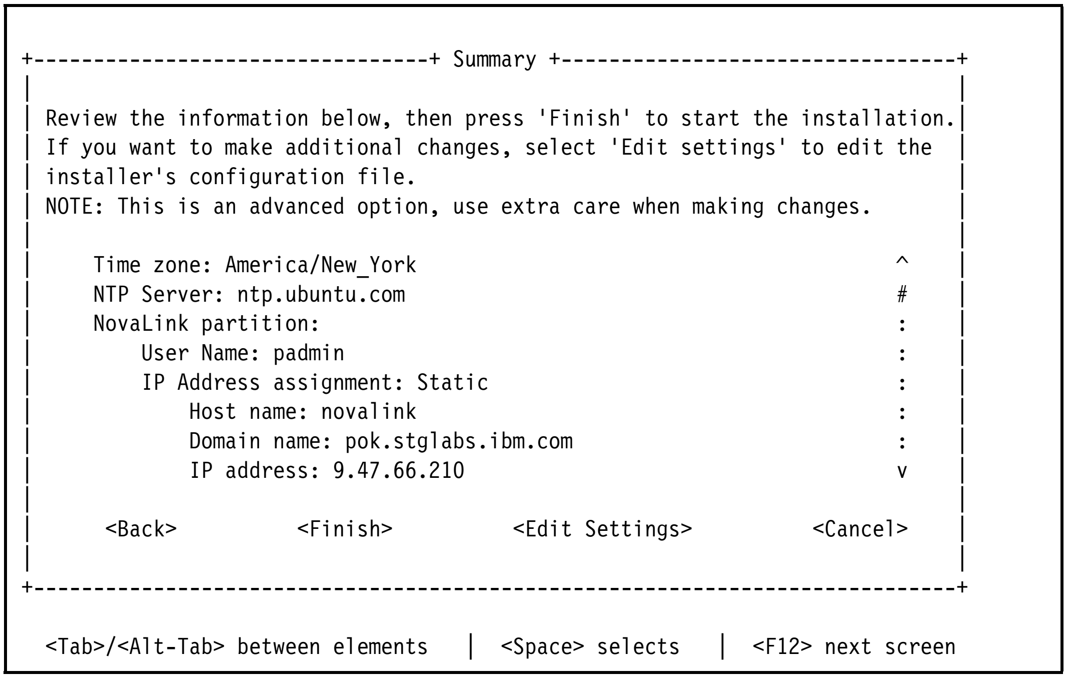
10. Select the destination disk for the operating system, which can be either one disk or two (mirrored), as shown in Figure 2-13.



```
+----------------------+ Target Disk Install Selection +----------------------+
|                                                                              |
| Select the disk(s) to be used for installation. Either 1 disk (no RAID) or  |
| 2 disks (RAID 1) can be selected. For multipath disks only one selection is |
| allowed. All data on the disk(s) selected will be erased.                   |
|                                                                              |
|              [*] sda 264(GB) SCSI Disk IBM IPR-0   6DDFB000                |
|              [ ] sdb 264(GB) SCSI Disk IBM IPR-0   6DDFB000                |
|              [ ] sdc 264(GB) SCSI Disk IBM IPR-0   6DDFB000                |
|              [ ] sdd 264(GB) SCSI Disk IBM IPR-0   6DDEF900                |
|              [ ] sde 264(GB) SCSI Disk IBM IPR-0   6DDEF900                |
|              [ ] sdf 264(GB) SCSI Disk IBM IPR-0   6DDEF900                |
|                                                                              |
|          <Back>                 <Next>                 <Cancel>             |
|                                                                              |
+------------------------------------------------------------------------------+



  <Tab>/<Alt-Tab> between elements   |  <Space> selects   |  <F12> next screen
```

Figure 2-13 PowerVM NovaLink installation wizard: Operating system disk selection

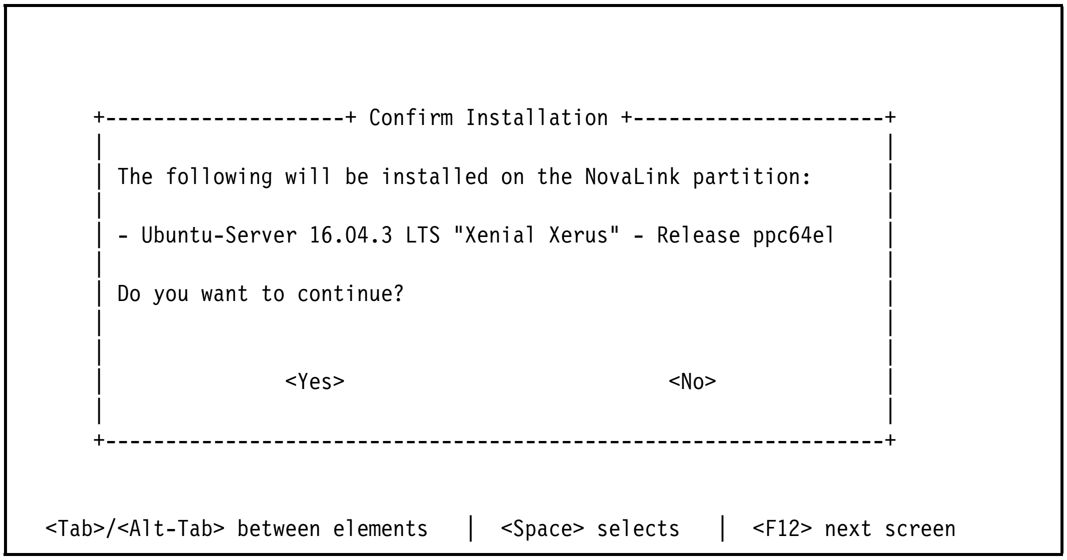
11.Figure 2-14 shows a summary with all the settings. If you want to, you can still edit some parts of the installation here.



```
+--------------------------------+ Summary +--------------------------------+
|                                                                            |
| Review the information below, then press 'Finish' to start the installation.|
| If you want to make additional changes, select 'Edit settings' to edit the |
| installer's configuration file.                                            |
| NOTE: This is an advanced option, use extra care when making changes.      |
|                                                                            |
|     Time zone: America/New_York                                      ^     |
|     NTP Server: ntp.ubuntu.com                                       #     |
|     NovaLink partition:                                              :     |
|         User Name: padmin                                            :     |
|         IP Address assignment: Static                                :     |
|             Host name: novalink                                      :     |
|             Domain name: pok.stglabs.ibm.com                         :     |
|             IP address: 9.47.66.210                                  v     |
|                                                                            |
|      <Back>         <Finish>         <Edit Settings>         <Cancel>      |
|                                                                            |
+----------------------------------------------------------------------------+

  <Tab>/<Alt-Tab> between elements   |  <Space> selects   |  <F12> next screen
```

Figure 2-14 PowerVM NovaLink installation wizard: Summary

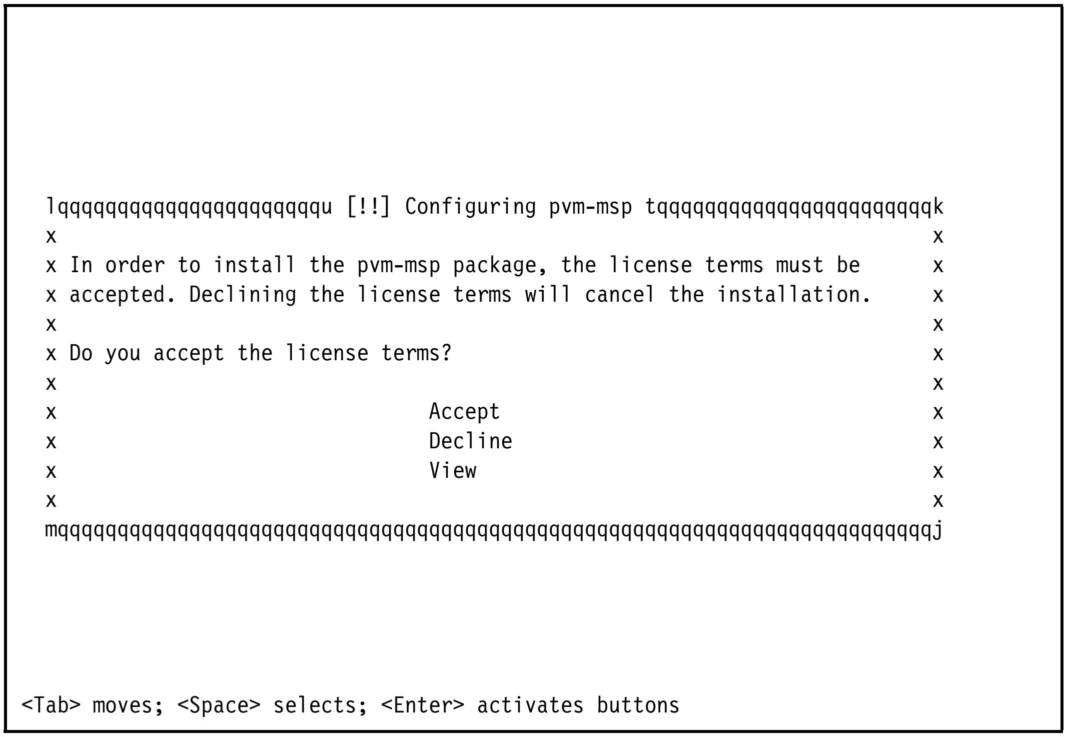
12.Figure 2-15 informs you that PowerVM NovaLink will install Ubuntu 16.04.3 and prompts you for confirmation.



```
      +--------------------+ Confirm Installation +---------------------+
      |                                                                 |
      | The following will be installed on the NovaLink partition:      |
      |                                                                 |
      | - Ubuntu-Server 16.04.3 LTS "Xenial Xerus" - Release ppc64el    |
      |                                                                 |
      | Do you want to continue?                                        |
      |                                                                 |
      |                                                                 |
      |               <Yes>                          <No>               |
      |                                                                 |
      +-----------------------------------------------------------------+


  <Tab>/<Alt-Tab> between elements   |  <Space> selects   |  <F12> next screen
```

Figure 2-15 PowerVM NovaLink installation wizard: Confirmation

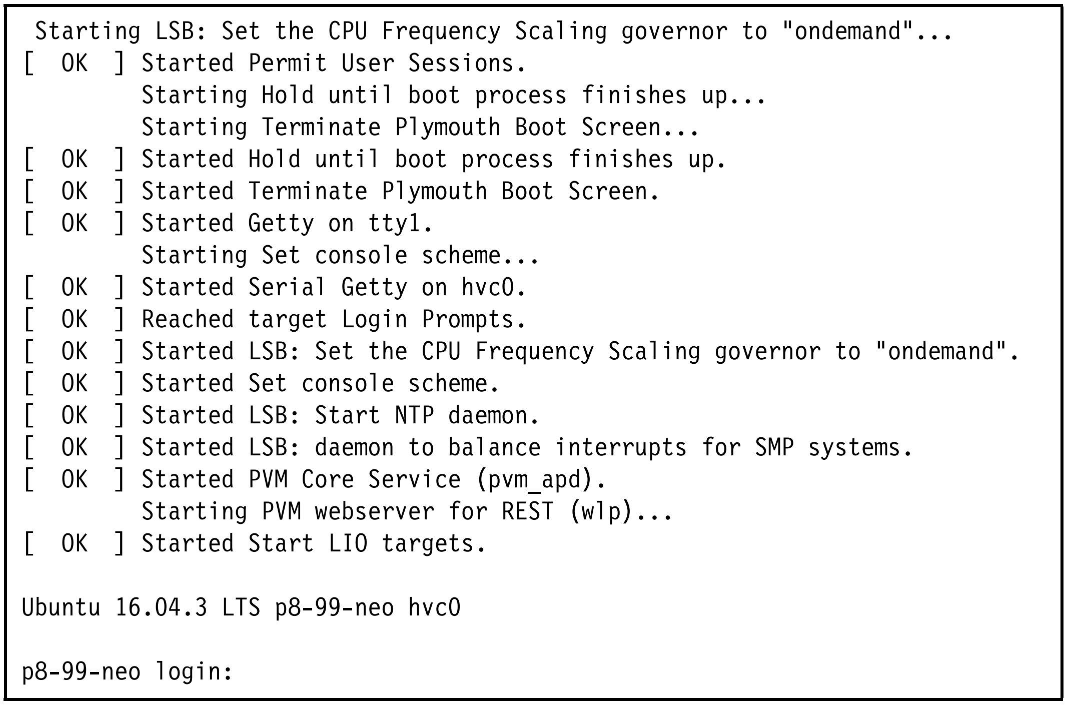
13. During the installation, you are promoted to accept the license agreement for the pvm-msp package, as shown in Figure 2-16. Accept the agreement.



```
  lqqqqqqqqqqqqqqqqqqqqqqu [!!] Configuring pvm-msp tqqqqqqqqqqqqqqqqqqqqqqqk
  x                                                                         x
  x In order to install the pvm-msp package, the license terms must be      x
  x accepted. Declining the license terms will cancel the installation.     x
  x                                                                         x
  x Do you accept the license terms?                                        x
  x                                                                         x
  x                               Accept                                    x
  x                               Decline                                   x
  x                               View                                      x
  x                                                                         x
  mqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqj



<Tab> moves; <Space> selects; <Enter> activates buttons
```

Figure 2-16   Accepting the license agreement

After some time, the installation finishes, as shown in Figure 2-17.

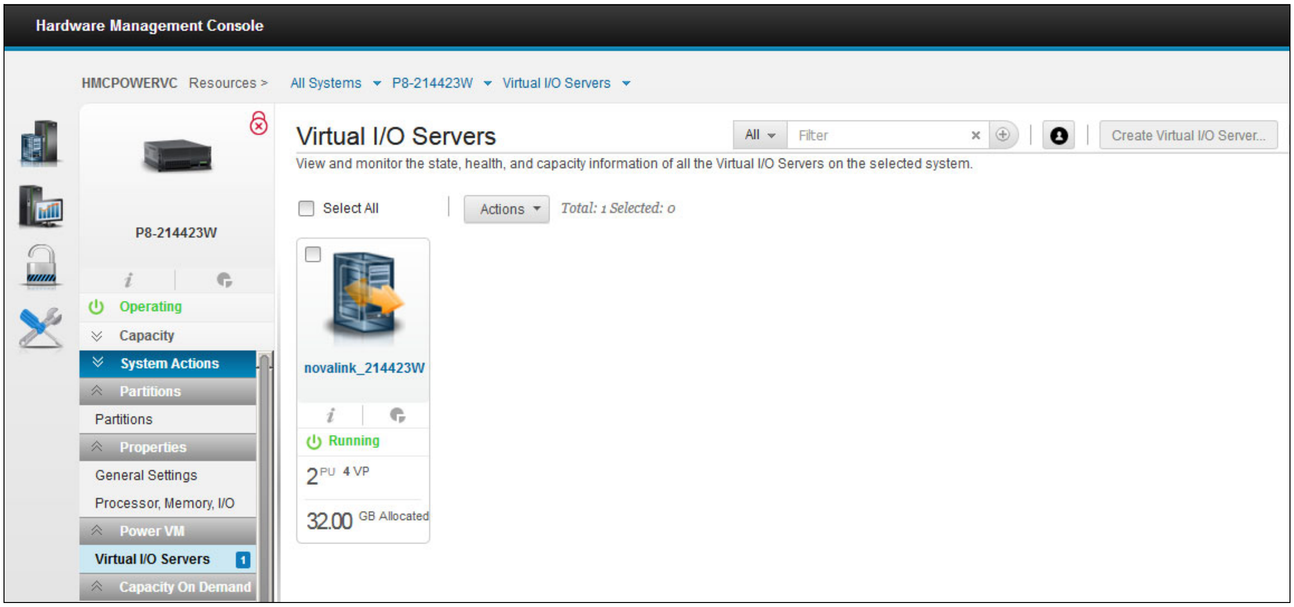


```
 Starting LSB: Set the CPU Frequency Scaling governor to "ondemand"...
[  OK  ] Started Permit User Sessions.
         Starting Hold until boot process finishes up...
         Starting Terminate Plymouth Boot Screen...
[  OK  ] Started Hold until boot process finishes up.
[  OK  ] Started Terminate Plymouth Boot Screen.
[  OK  ] Started Getty on tty1.
         Starting Set console scheme...
[  OK  ] Started Serial Getty on hvc0.
[  OK  ] Reached target Login Prompts.
[  OK  ] Started LSB: Set the CPU Frequency Scaling governor to "ondemand".
[  OK  ] Started Set console scheme.
[  OK  ] Started LSB: Start NTP daemon.
[  OK  ] Started LSB: daemon to balance interrupts for SMP systems.
[  OK  ] Started PVM Core Service (pvm_apd).
         Starting PVM webserver for REST (wlp)...
[  OK  ] Started Start LIO targets.

Ubuntu 16.04.3 LTS p8-99-neo hvc0

p8-99-neo login:
```

Figure 2-17   PowerVM NovaLink installation wizard: Finished

After the installation, the PowerVM NovaLink partition shows up under VIOS and uses only the resources that were defined during the installation.

Figure 2-18 shows that the PowerVM NovaLink partition moved to VIOSes.



View and monitor the state, health, and capacity information of all the Virtual I/O Servers on the selected system.

Figure 2-18   The PowerVM NovaLink partition under Virtual I/O Servers

14.From a networking point of view, it is preferable to separate the storage traffic from the data traffic. Therefore, add another network bridge for the storage. Figure 2-19 shows how to configure a second storage bridge by using an extra adapter in `/etc/network/interfaces`.

```
# This file describes the network interfaces available on your system
# and how to activate them. For more information, see interfaces(5).

source /etc/network/interfaces.d/*

# The loopback network interface
auto lo
iface lo inet loopback

# Data Network Configuration
auto br-ex
allow-ovs br-ex
iface br-ex inet static
        address 9.47.66.218
        netmask 255.255.240.0
        gateway 9.47.79.254
        dns-nameservers 9.12.16.2
        dns-search pok.stglabs.ibm.com
        ovs_type OVSBridge
        ovs_ports br_ex_port0

allow-br-ex br_ex_port0
iface br_ex_port0 inet manual
        ovs_bridge br-ex
        ovs_type OVSPort

# Storage & Management Network Configuration
auto br-st
allow-ovs br-st

iface br-st inet static
      address 10.10.10.1
      netmask 255.255.255.0
      ovs_type OVSBridge
      ovs_ports enP24p1s0f2

allow-br-st enP24p1s0f2
iface enP24p1s0f2 inet manual
    ovs_bridge br-st
    ovs_type OVSPort

# Internal NovaLink

iface ibmveth3 inet static
        address 192.168.128.1
        netmask 255.255.128.0

iface ibmveth3 inet6 auto
```

Figure 2-19 Example of an extra storage bridge by using one adapter

15. If you want to define a storage bridge by using a network bond, use the following example, as shown in Figure 2-20.

```
...
# Bonded Ethernet Ports for the data network.
allow-br-ex bonded_data_port
iface bonded_data_port inet manual
  ovs_bridge br-ex
  ovs_type OVSBond
  ovs_bonds enP30p128s0f0 enP40p1s0f0
  ovs_options other_config:bond_mode=active-backup

# Storage & Management Network Configuration
auto br-st
allow-ovs br-st

# IP configuration of the Storage & Management network
iface br-st inet static
        address 10.10.10.1
        netmask 255.255.255.0
        ovs_type OVSBridge
        ovs_ports bonded_storage_port

# Bonded Ethernet Ports for the storage network
allow-br-st bonded_storage_port
iface bonded_storage_port inet manual
  ovs_bridge br-st
  ovs_type OVSBond
  ovs_bonds enP30p128s0f1 enP40p1s0f1
  ovs_options other_config:bond_mode=active-backup
...
```

Figure 2-20 Example of an extra storage bridge by using a bond of two adapters

The system is now ready to be added to a cluster by IBM PowerVC, which is described n in 3.3, “Adding compute nodes to IBM PowerVC” on page 66.

Note: When the system is added to the cluster by using IBM PowerVC, it decides which bridge (br-ex or br-st) is used by the IP address that is entered in the host add dialog. In the example that is shown in Figure 2-19 on page 38, if you use the IP 9.47.66.218, IBM PowerVC uses br-ex for the storage traffic. If you use 10.10.10.1, it uses br-st for the storage traffic.

2.3 Installing Ubuntu and PowerVM NovaLink for a system running KVM on Open Power Abstraction Layer

If you choose to use KVM with Ubuntu on Power, then the server must have Ubuntu 16.04 LTS installed. If the server has a Flexible Service Processor (FSP), you must set the firmware mode to Open Power Abstraction Layer (OPAL), as described in 2.5, “Preparing your Red Hat Enterprise Linux KVM on an IBM POWER8 server” on page 43. This setting ensures that when Ubuntu installs, it can run KVM virtual machines (VMs). If the Power System server has a baseboard management controller (BMC), this step does not need to be done.

Download the installation ISO and follow the standard Ubuntu installation instructions.

For a smaller cloud, the easiest path might be to burn the ISO to a DVD and put it in the server. Use the Intelligent Platform Management Interface (IPMI) console to walk through the server installation.

For larger, scale- out clouds, set up a network installation server.

During installation, be sure to pick the `Virtual Machine Host` and `OpenSSH server` options as part of the software selection. The SSH connections are used by IBM PowerVC to install the agents on the host.

For more information about how to use IPMI or use an installation server for a network installation, see 2.6, “Installing Red Hat Enterprise Linux KVM on the IBM Power System server” on page 50.

PowerVM NovaLink software repository configuration

After the Ubuntu installation is complete, add the PowerVM NovaLink software repository to the KVM host. The PowerVM NovaLink software provides an API compatibility layer between IBM PowerVC and the KVM host.

Use the Ubuntu Advanced Packaging Tool (APT) to install the PowerVM NovaLink repository by editing `/etc/apt/sources.list.d/pvm.list` and adding the following line:

```
ftp://public.dhe.ibm.com/systems/virtualization/Novalink/debian
```

Then, run **`apt-get update`** to make the repository available to the system.

Configuring the network

You must define the storage/management network and the data network. The y can run atop a logical network or be separated into two networks.

The data network must be running on top of an OVS. To configure this, run **`apt-get install openvswitch-switch`**. Then, install `kvm-novalink` by running **`apt-get install kvm-novalink`**.

After the installation is done, define a new virtual switch that is named br-ex and attach the port or create a link aggregation atop it. This configuration must persist across restarts of the server as well.

Example 2-8 shows a sample configuration of the `/etc/network/interfaces` file for an Ubuntu 16.04 server. Ensure that the operating system's documentation is referenced for the correct configuration in your environment.

Example 2-8 Sample configuration of the /etc/network/interfaces file

```
# This file describes the network interfaces available on your system
# and how to activate them. For more information, see interfaces(5).

source /etc/network/interfaces.d/*

# The loopback network interface
auto lo
iface lo inet loopback

# Data Network Configuration
# All data traffic flows over the br-ex network
auto br-ex
allow-ovs br-ex

# IP configuration of the OVS Bridge
iface br-ex inet static
        address 10.10.58.10
        netmask 255.255.240.0
        gateway 10.10.58.1
        dns-nameservers 10.10.10.10
        dns-search my.testenv
        ovs_type OVSBridge
        ovs_ports bonded_data_port

# Bonded Ethernet Ports for the data network.
allow-br-ex bonded_data_port
iface bonded_data_port inet manual
  ovs_bridge br-ex
  ovs_type OVSBond
  ovs_bonds enP30p128s0f0 enP40p1s0f0
  ovs_options other_config:bond_mode=active-backup

# Storage & Management Network Configuration
auto br-st
allow-ovs br-st

# IP configuration of the Storage & Management network
iface br-st inet static
        address 192.168.10.23
        netmask 255.255.255.0
        ovs_type OVSBridge
        ovs_ports bonded_storage_port

# Bonded Ethernet Ports for the storage network
allow-br-st bonded_storage_port
iface bonded_storage_port inet manual
  ovs_bridge br-st
  ovs_type OVSBond
  ovs_bonds enP30p128s0f1 enP40p1s0f1
  ovs_options other_config:bond_mode=active-backup
```

Note: For an evaluation environment, drop the br-st and make the configuration of br-ex simpler by using only one interface. For more information, see 2.2.2, “Installing a PowerVM NovaLink partition on an empty system that is under control of an HMC” on page 25.

2.4 Installing Red Hat Enterprise Linux KVM on an IBM Power System server

This section describes all the requirements and steps that are needed for installing and configuring a Red Hat Enterprise Linux KVM on IBM Power Systems servers.

2.4.1 Hardware requirements

Here is a list of hardware requirements for running Red Hat Enterprise Linux KVM on a bare metal IBM Power Systems server:

- The system firmware must be set up as OPAL firmware.
- The console must be set to IPMI mode to enable the remote installation.
- The compute nodes require that two physical network ports are available (preferably on different physical adapters): One for the data network and one for the storage network.
- Whether you are reusing Power Systems servers that previously ran IBM PowerVM or acted as a full partition Linux server, it is important to ensure that no SAN disk logical unit numbers (LUNs) are mapped to the host.

  You must have an array of disks of the same vendor, model, and size. Contact your SAN storage administrator and request that they remove all the LUNs. This is a preferred practice for File Placement Optimizer (FPO) topologies where a local/internal disk array is used because you do not want to be mixing different disk types within the same IBM Spectrum Scale cluster.

2.4.2 Administrator’s workstation requirements

You must use the IPMI tool to perform the Red Hat Enterprise Linux installation through Petitboot, which is covered in 2.6, “Installing Red Hat Enterprise Linux KVM on the IBM Power System server” on page 50.

Linux-based workstations should download the ipmitool package from GitHub.

Windows-based workstations should download the ipmiutil package from Sourceforge.

Note: For a Windows workstation, download the Microsoft Windows installer (MSI) version because it automatically adds ipmiutil to the Windows PATH environment variable so that you can run it on a Windows command prompt without having to type the full path of the program.

2.5 Preparing your Red Hat Enterprise Linux KVM on an IBM POWER8 server

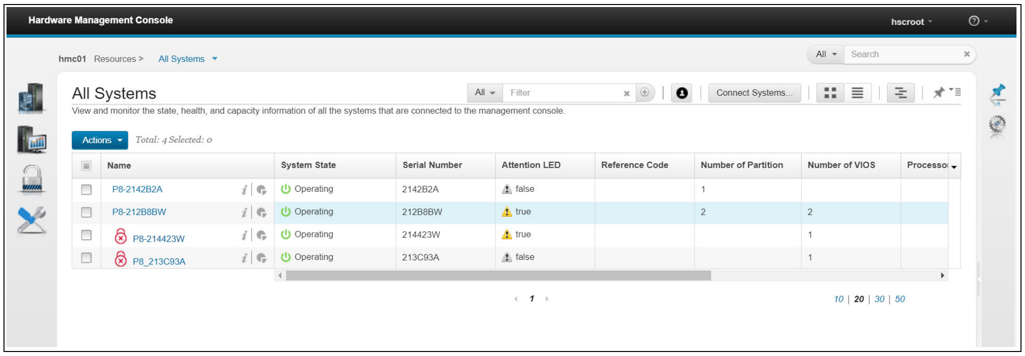
When you acquire a new IBM Power Systems server, at the time of ordering, you specify that you are installing a bare metal Linux server, which means that it should come pre-configured by IBM with the appropriate firmware and console settings.

However, in many cases you reuse an existing server that previously had one of the PowerVM offerings. To convert the server to an OPAL / BMC firmware stack so that it can run KVM on Power, complete the following steps:

1. Open your preferred internet browser to log in to the system:

   a. If the Hardware Management Console (HMC) network topology is based on DHCP (meaning that the managed systems obtain an IP from the HMC DHCP service interface), then log in to the Advanced System Management Interface (ASMI) through the HMC.

   Figure 2-21 shows the HMC home page that lists all the systems that are connected.



View and monitor the state, health, and capacity information of all the systems that are connected to the management console.

Total: 4 Selected: 0

Name	System State	Serial Number	Attention LED	Reference Code	Number of Partition	Number of VIOS	Processor
P8-2142B2A	Operating	2142B2A	false		1		
P8-212B8BW	Operating	212B8BW	true		2	2	
P8-214423W	Operating	214423W	true			1	
P8_213C93A	Operating	213C93A	false			1	

Figure 2-21 All managed frames that are currently connected to the HMC

Select the managed system and then select **Actions** → **View all actions** → **Launch Advanced System Management (ASM)**.

Figure 2-22 shows the HMC actions menu that you use to start the ASMI interface.

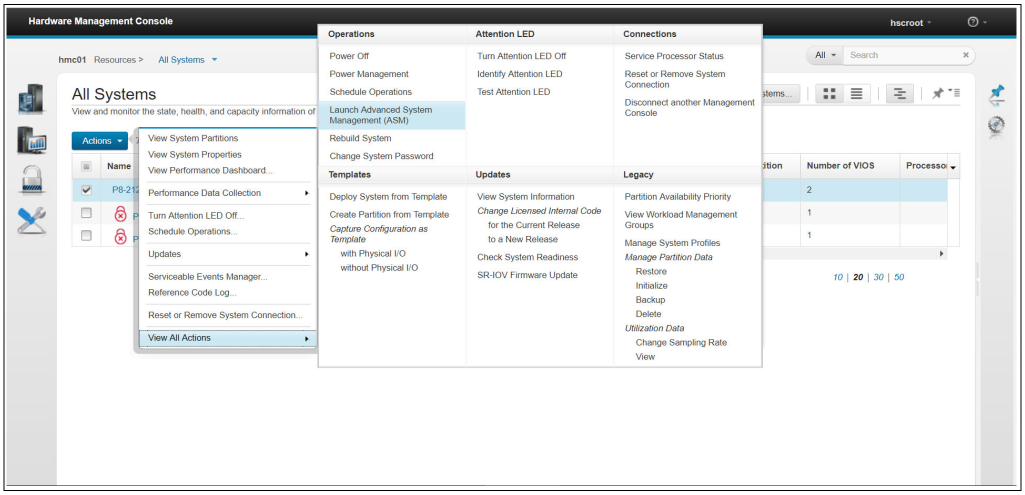


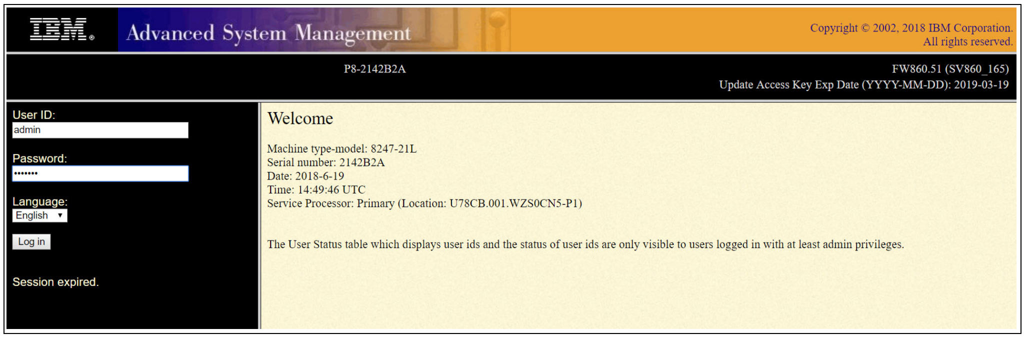
Figure 2-22 HMC actions menu

b. If the HMC network topology is based on a static IP, log in to the ASMI directly by using the SSL protocol, the IP address of the FSP, and admin as the User ID. Here is an example URL:

```
https://9.47.66.224
```

> **Note:** If the access credentials have not changed, then the default ASMI user ID/password is always admin/admin.

Figure 2-23 shows the ASMI login welcome window.



Advanced System Management

Copyright © 2002, 2018 IBM Corporation. All rights reserved.

P8-2142B2A

FW860.51 (SV860_165)
Update Access Key Exp Date (YYYY-MM-DD): 2019-03-19

User ID: admin

Password: •••••••

Language: English

Session expired.

Welcome

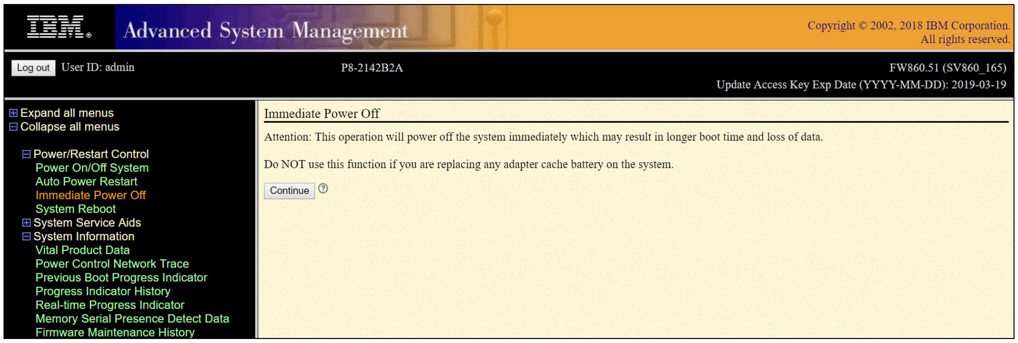
Machine type-model: 8247-21L
Serial number: 2142B2A
Date: 2018-6-19
Time: 14:49:46 UTC
Service Processor: Primary (Location: U78CB.001.WZS0CN5-P1)

The User Status table which displays user ids and the status of user ids are only visible to users logged in with at least admin privileges.

Figure 2-23 ASMI welcome login window

2. Go to the menu at the left, and select **Power/Restart Control** → **Immediate Power Off** → **Continue**.

Figure 2-24 shows the ASMI power-off menu.



Advanced System Management

Copyright © 2002, 2018 IBM Corporation. All rights reserved.

User ID: admin

P8-2142B2A

FW860.51 (SV860_165)
Update Access Key Exp Date (YYYY-MM-DD): 2019-03-19

Immediate Power Off

Attention: This operation will power off the system immediately which may result in longer boot time and loss of data.

Do NOT use this function if you are replacing any adapter cache battery on the system.

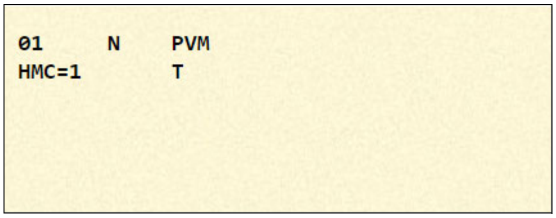
Figure 2-24 Power off a managed system

The system shuts down quickly. You can run the shutdown safely because no one is using the system and you are not performing any hardware changes.

You can monitor the system shutdown process in real time by going to the left menu and select **System Information** → **Real-time Progress Indicator**.

After the system is in the power-off state, the front panel should look like Figure 2-25 on page 45. Notice the **PVM** string, which reflects that it is running IBM PowerVM firmware, and **HMC=1**, which shows that it is connected to an HMC.

Figure 2-25 shows the state of the front panel after the system is in a shutdown state.

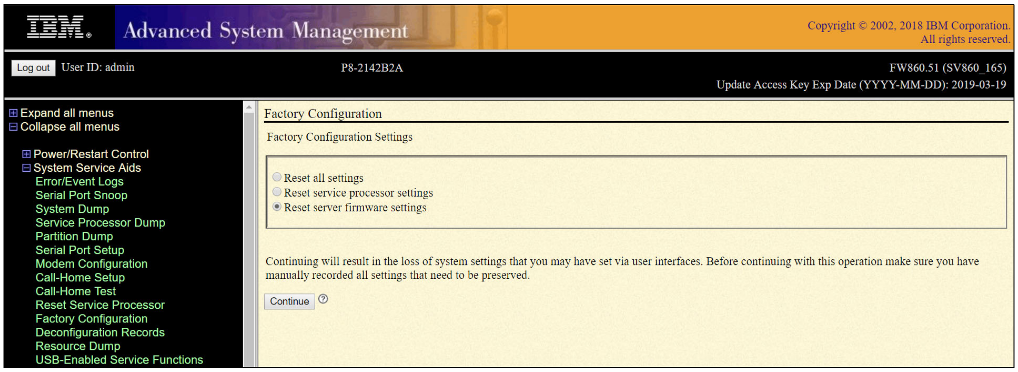


```
01     N    PVM
HMC=1       T
```

Figure 2-25   Actual state of the physical front panel in an IBM Power System server

3. Go to the menu at the left and select **System Service Aids** → **Factory Configuration** → **Reset server firmware settings**.

   Figure 2-26 shows the ASMI factory reset menu.



User ID: admin   P8-2142B2A   FW860.51 (SV860_165)
Update Access Key Exp Date (YYYY-MM-DD): 2019-03-19

Factory Configuration

Factory Configuration Settings

○ Reset all settings
○ Reset service processor settings
● Reset server firmware settings

Continuing will result in the loss of system settings that you may have set via user interfaces. Before continuing with this operation make sure you have manually recorded all settings that need to be preserved.

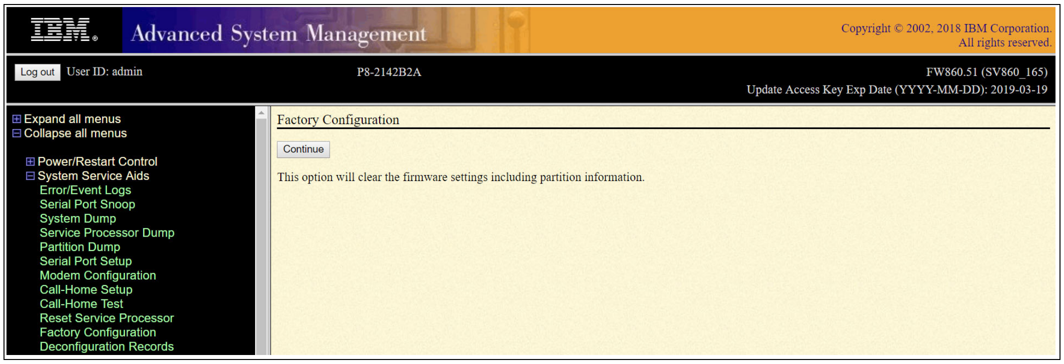
Continue

Figure 2-26   Factory reset configuration menu

> **Important:** Ensure that you are resetting the firmware settings and not the service processor, or you lose remote access to the ASMI (the network settings are reset).

A confirmation window appears. Click **Continue**.

Figure 2-27 shows the factory reset confirmation window.



User ID: admin   P8-2142B2A   FW860.51 (SV860_165)
Update Access Key Exp Date (YYYY-MM-DD): 2019-03-19

Factory Configuration

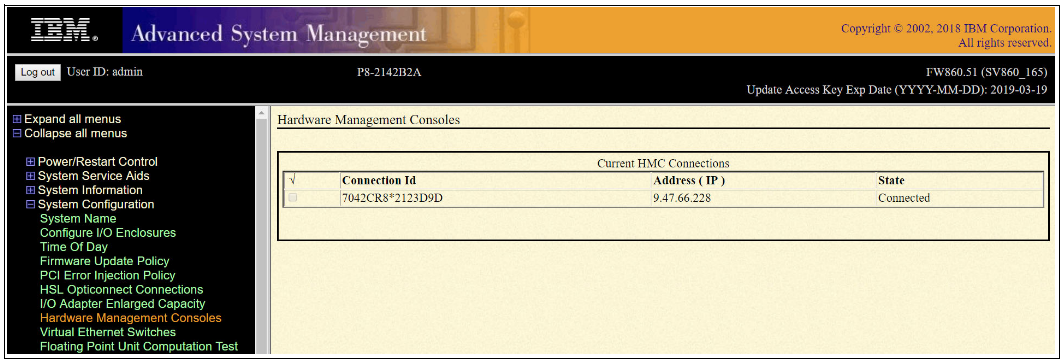
Continue

This option will clear the firmware settings including partition information.

Figure 2-27   Factory reset confirmation window

4. Go to the menu at the left, and select **System Configuration** → **Hardware Management Consoles**. If the HMC connection cannot be selected, it means it cannot be removed. An example of this scenario is shown in Figure 2-28.

   Figure 2-28 shows an unconfigurable HMC connection.


IBM
Advanced System Management
Log out
Expand all menus
Collapse all menus
Power/Restart Control
System Service Aids
System Information
System Configuration
System Name
Configure I/O Enclosures
Time Of Day
Firmware Update Policy
PCI Error Injection Policy
HSL Opticonnect Connections
I/O Adapter Enlarged Capacity
Hardware Management Consoles
Virtual Ethernet Switches
Floating Point Unit Computation Test

Copyright © 2002, 2018 IBM Corporation. All rights reserved.

User ID: admin

P8-2142B2A

FW860.51 (SV860_165)

Update Access Key Exp Date (YYYY-MM-DD): 2019-03-19

Hardware Management Consoles

Current HMC Connections

√	Connection Id	Address (IP)	State
□	7042CR8*2123D9D	9.47.66.228	Connected

Figure 2-28 Unremovable managed system's FSP connection

If you are experiencing this scenario, it means that the removal cannot be performed through the ASMI, so it must be done manually from the HMC:

a. To remove the FSP connection from the HMC GUI, select your system and then select **Go to actions** → **View all actions** → **Reset or Remove**.

   Figure 2-29 shows how to manually remove the FSP connection from the HMC GUI.

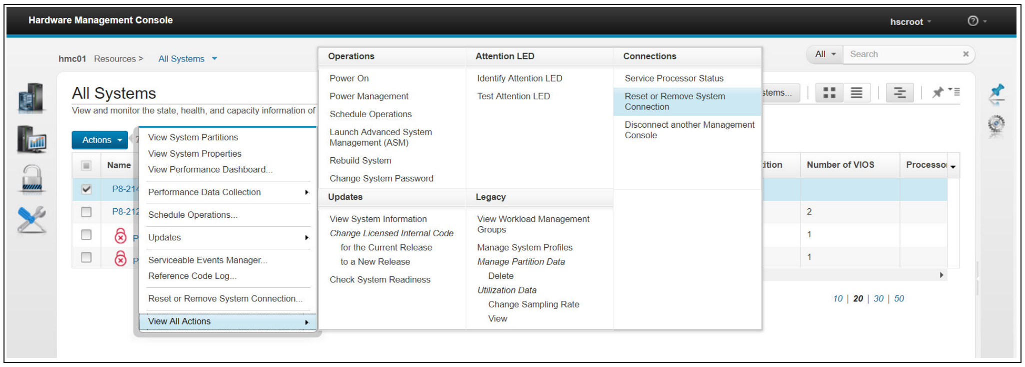
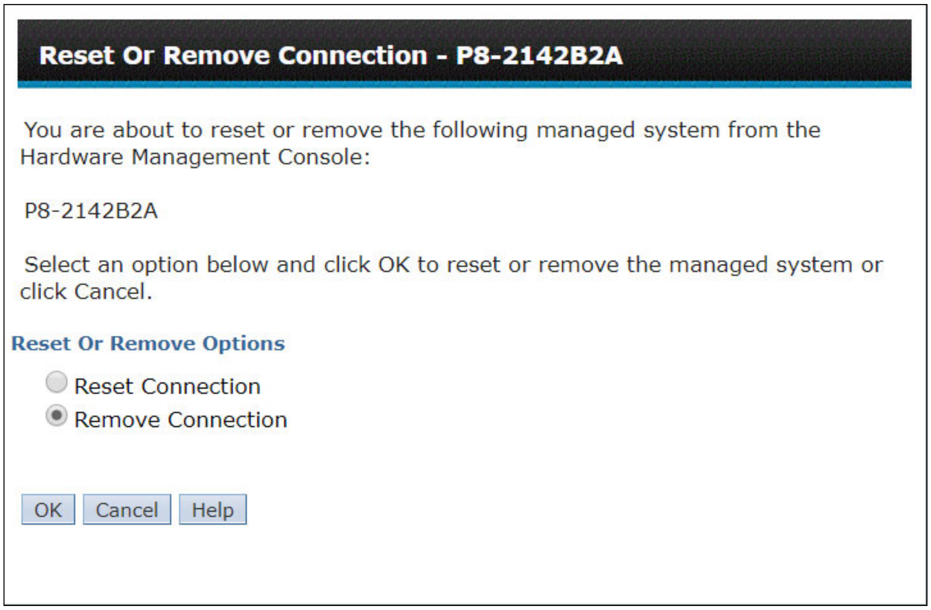


Figure 2-29 Removing the managed system's FSP connection

Figure 2-30 shows the removal confirmation dialog when removing the FSP connection.


Reset Or Remove Connection - P8-2142B2A

You are about to reset or remove the following managed system from the Hardware Management Console:

P8-2142B2A

Select an option below and click OK to reset or remove the managed system or click Cancel.

Reset Or Remove Options

○ Reset Connection
● Remove Connection

Figure 2-30 Managed system removal confirmation dialog

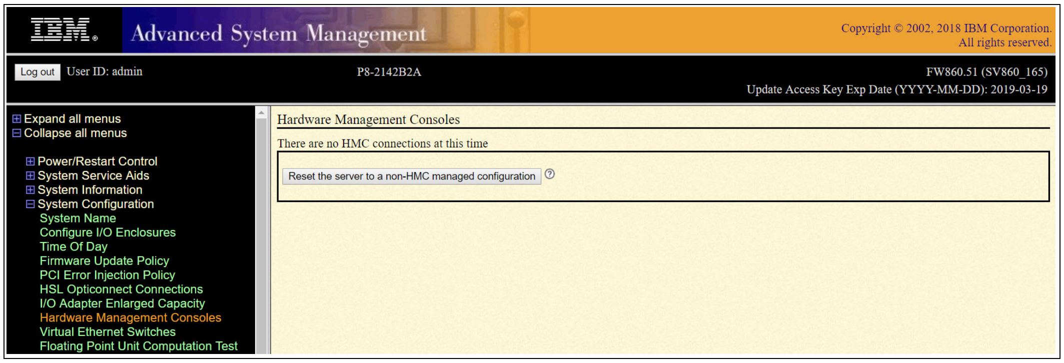
b. Alternatively, you can remove the managed system manually through the HMC CLI by running the following command:

```
rmsysconn --ip <FSP IP ADDRESS> -o remove
```

5. Go back to the ASMI window. At the menu on the left, select **System Configuration** → **Hardware Management Consoles**. You should now have the option available to reset the server to a non-HMC managed configuration. Click **Reset the server to a non-HMC managed configuration**.

> **Note:** If the FSP connection still shows up as HMC-managed, or the connection cannot be selected, restart the HMC, restart the managed system, and perform the firmware reset again. If you still cannot switch to non-HMC mode, proceed to open a problem record at IBM.

Figure 2-31 shows the menu option to reset the managed system to a non-HMC managed configuration.


IBM.
Advanced System Management
Copyright © 2002, 2018 IBM Corporation. All rights reserved.
Log out

User ID: admin

P8-2142B2A

FW860.51 (SV860_165)

Update Access Key Exp Date (YYYY-MM-DD): 2019-03-19

Expand all menus
Collapse all menus
Power/Restart Control
System Service Aids
System Information
System Configuration
System Name
Configure I/O Enclosures
Time Of Day
Firmware Update Policy
PCI Error Injection Policy
HSL Opticonnect Connections
I/O Adapter Enlarged Capacity
Hardware Management Consoles
Virtual Ethernet Switches
Floating Point Unit Computation Test

Hardware Management Consoles

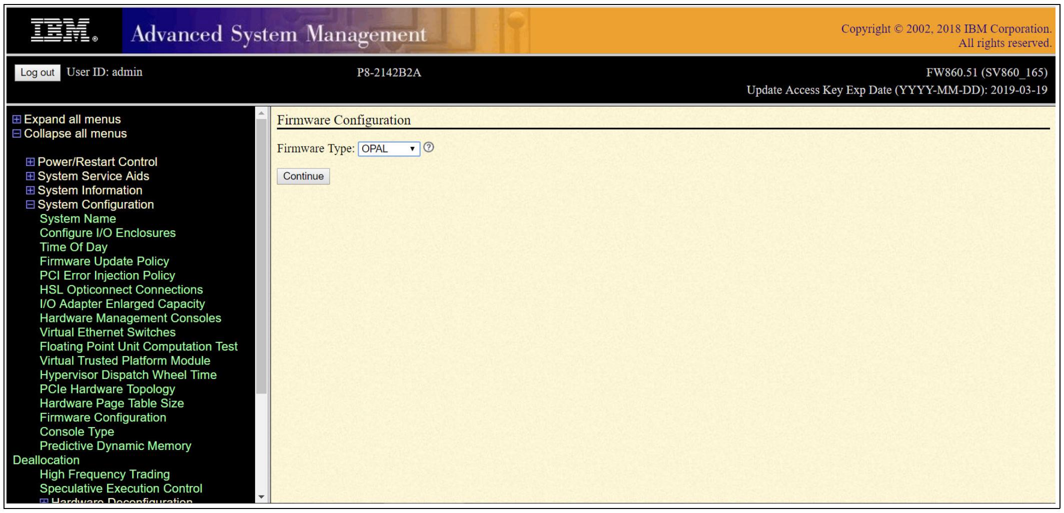
There are no HMC connections at this time

Reset the server to a non-HMC managed configuration

Figure 2-31 Resetting the server to a non-HMC managed configuration on the ASMI window

6. In the left menu, select **System Configuration** → **Firmware Configuration** → **Firmware Type: OPAL**.

   Figure 2-32 shows ASMI menu section that you use to set OPAL as the firmware type.

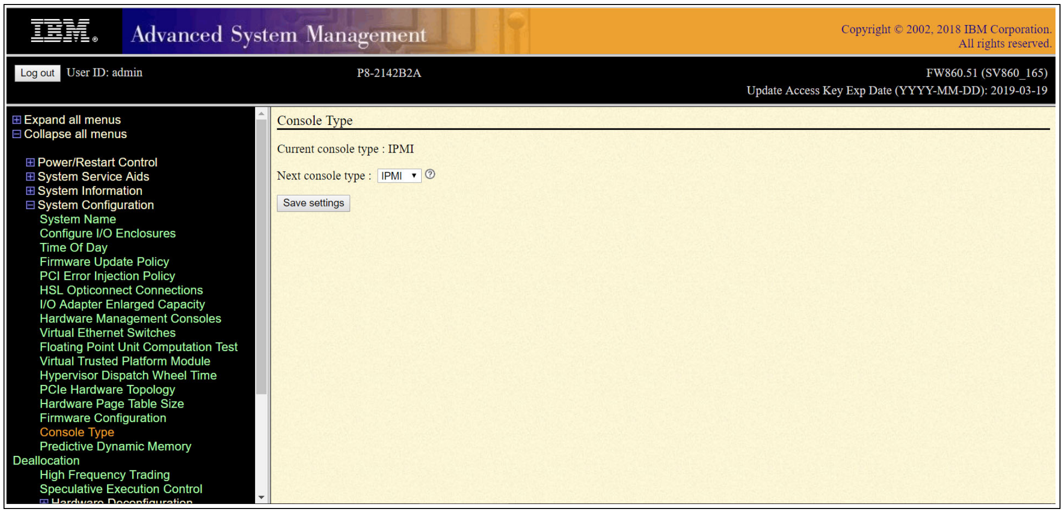

IBM.
Advanced System Management
Copyright © 2002, 2018 IBM Corporation.
All rights reserved.
Log out
User ID: admin
P8-2142B2A
FW860.51 (SV860_165)
Update Access Key Exp Date (YYYY-MM-DD): 2019-03-19
⊞ Expand all menus
⊟ Collapse all menus
⊞ Power/Restart Control
⊞ System Service Aids
⊞ System Information
⊟ System Configuration
System Name
Configure I/O Enclosures
Time Of Day
Firmware Update Policy
PCI Error Injection Policy
HSL Opticonnect Connections
I/O Adapter Enlarged Capacity
Hardware Management Consoles
Virtual Ethernet Switches
Floating Point Unit Computation Test
Virtual Trusted Platform Module
Hypervisor Dispatch Wheel Time
PCIe Hardware Topology
Hardware Page Table Size
Firmware Configuration
Console Type
Predictive Dynamic Memory Deallocation
High Frequency Trading
Speculative Execution Control
⊞ Hardware Deconfiguration
Firmware Configuration
Continue

Firmware Type: OPAL

Figure 2-32   Setting the managed system's firmware type to OPAL

7. In the left menu, select **System Configuration** → **Console Type** → **Next Console Type: IPMI**.

   Figure 2-33 shows the ASMI menu section that is used to set the console type as IPMI.


IBM.
Advanced System Management
Copyright © 2002, 2018 IBM Corporation.
All rights reserved.
Log out
User ID: admin
P8-2142B2A
FW860.51 (SV860_165)
Update Access Key Exp Date (YYYY-MM-DD): 2019-03-19
⊞ Expand all menus
⊟ Collapse all menus
⊞ Power/Restart Control
⊞ System Service Aids
⊞ System Information
⊟ System Configuration
System Name
Configure I/O Enclosures
Time Of Day
Firmware Update Policy
PCI Error Injection Policy
HSL Opticonnect Connections
I/O Adapter Enlarged Capacity
Hardware Management Consoles
Virtual Ethernet Switches
Floating Point Unit Computation Test
Virtual Trusted Platform Module
Hypervisor Dispatch Wheel Time
PCIe Hardware Topology
Hardware Page Table Size
Firmware Configuration
Console Type
Predictive Dynamic Memory Deallocation
High Frequency Trading
Speculative Execution Control
⊞ Hardware Deconfiguration
Console Type
Save settings

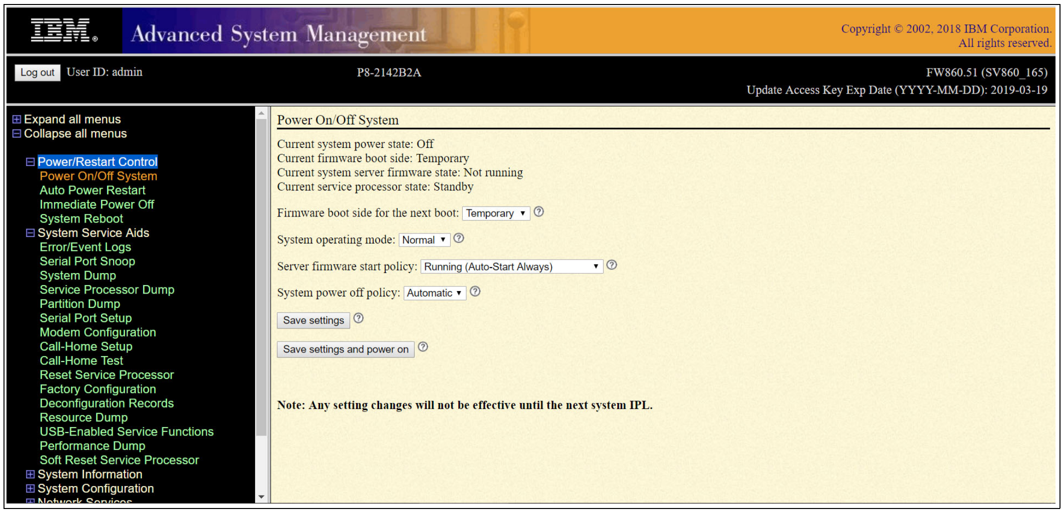
Current console type : IPMI

Next console type : IPMI

Figure 2-33   Setting the managed system's console type to IPMI

8. In the left menu, select **Power/Restart Control** → **Save settings and power-on**.

   Figure 2-34 shows the ASMI menu section that is used to power-on/restart the managed system.



Power On/Off System

Current system power state: Off
Current firmware boot side: Temporary
Current system server firmware state: Not running
Current service processor state: Standby

Firmware boot side for the next boot: Temporary

System operating mode: Normal

Server firmware start policy: Running (Auto-Start Always)

System power off policy: Automatic

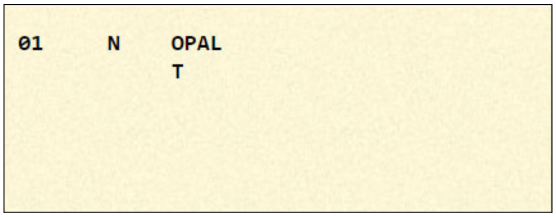
Note: Any setting changes will not be effective until the next system IPL.

Figure 2-34 Power-on/restart the managed system

9. In the left menu, select **System Information** → **Real-time Progress Indicator**. You should see codes cycling.

   After the system performs the Power-On-Self-Test (POST), the front panel should look as shown in Figure 2-35. Notice the OPAL string, which reflects that it is running the IBM OPAL / BMC firmware stack, and the HMC string is no longer visible, which reflects that the system is now in bare metal mode.

   Figure 2-35 shows the expected front panel display after the system has been successfully converted to OPAL and completed the start process.



```
01      N     OPAL
              T
```

Figure 2-35 Managed system's front panel showing that it is started and configured with OPAL / BMC firmware stack

> **Note:** The system can take some time to start, even after the front panel shows OPAL. Allow for 5 - 10 minutes until you get into the bootloader.

2.6 Installing Red Hat Enterprise Linux KVM on the IBM Power System server

You have two methods to install Red Hat Enterprise Linux on an IBM Power OPAL-based System server:

- By using a USB drive on site.
- By using remote installation through a network boot, which is the procedure that is outlined in this section.

Performing a remote Red Hat Enterprise Linux installation through a network boot

As described in 2.4.2, “Administrator’s workstation requirements” on page 42, you perform a remote Red Hat Enterprise Linux installation by using the IPMI tool for Windows, which is named ipmiutil.

Complete the following steps:

1. On your Windows workstation, open a Windows command prompt by running **`cmd.exe`**.
2. Because the system was turned on through the ASMI, you can now connect to the BMC by using the IPMI protocol and activating and simulating a remote Serial-Over-LAN (SOL) connection. To do this task, run the following command:

   ```
   ipmiutil sol -a -r -N <ip address> -P <asmi password>
   ```

   Example 2-9 shows the return output after a successful activation of the SOL interface.

   Example 2-9 Successful activation of the SOL interface

   ```
   C:\Windows\System32>ipmiutil sol -a -r -N 9.47.66.224 -P mypass
   ipmiutil sol ver 3.11
   Connecting to node  9.47.66.224
   ipmilan_open_session error, rv = -15
   ipmilan BMC only supports lan v2
   Opening lanplus connection to node 9.47.66.224 ...
   -- BMC version 8.61, IPMI version 2.0
   [SOL session is running, use '~.' to end, '~?' for help.]

   Red Hat Enterprise Linux
   Kernel 3.10.0-862.3.3.el7.ppc64le on an ppc64le

   p8-2142b2a-rhel-kvm login: root
   Password:
   Last login: Mon Jun 25 13:03:39 from 9.24.18.117
   [root@p8-2142b2a-rhel-kvm ~]#
   ```

   If you receive an error while trying to log in through the SOL interface, as shown in Example 2-10, it is because there is an existing SOL connection to the BMC (only one session at a time is allowed).

   Example 2-10 Failed activation of the SOL interface

   ```
   C:\Windows\System32>ipmiutil sol -a -r -N 9.47.66.224 -P ibmitso
   ipmiutil sol ver 3.11
   Connecting to node  9.47.66.224
   ipmilan_open_session error, rv = -15
   ```

```
ipmilan BMC only supports lan v2
Opening lanplus connection to node 9.47.66.224 ...
-- BMC version 8.61, IPMI version 2.0
SOL payload already active on another session

ipmiutil sol, Invalid Session Handle or Empty Buffer

C:\Windows\System32>
```

To fix the issue, run the following command:

```
ipmiutil sol -d -r -N <ip address> -P <ASMI password>
```

Example 2-11 shows the return output after the de-activating SOL interface through ipmiutil.

Example 2-11 De-activating the SOL interface through ipmiutil

```
C:\Windows\System32>ipmiutil sol -d -r -N 9.47.66.224 -P mypass
ipmiutil sol ver 3.11
Connecting to node  9.47.66.224
ipmilan_open_session error, rv = -15
ipmilan BMC only supports lan v2
Opening lanplus connection to node 9.47.66.224 ...
-- BMC version 8.61, IPMI version 2.0
ipmiutil sol, completed successfully
```

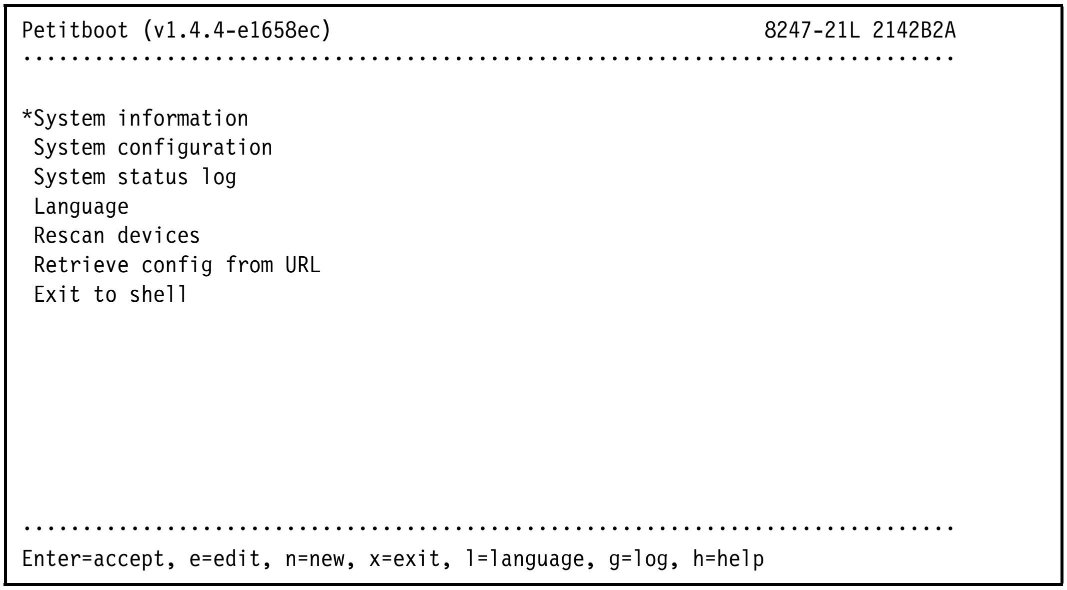
3. You should see the Petitboot bootloader welcome screen after the SOL connection is established because no OS is installed. Otherwise, the system starts the installed operating system quickly. You must interrupt the boot process by pressing any key.

   Petitboot is a platform-independent bootloader that is based on the Linux kexec warm restart mechanism. Petitboot supports loading kernel, initrd, and device tree files from any Linux mountable file system or through the network by using the FTP, SFTP, TFTP, NFS, HTTP, and HTTPS protocols.

   To move around the menu, use either the Tab or arrows keys. You select menus by pressing Enter, and select items within a menu by pressing the spacebar. Going back to a previous menu or welcome screen is done by pressing the Esc key.

   For more information about the Petitboot bootloader, see IBM Knowledge Center.

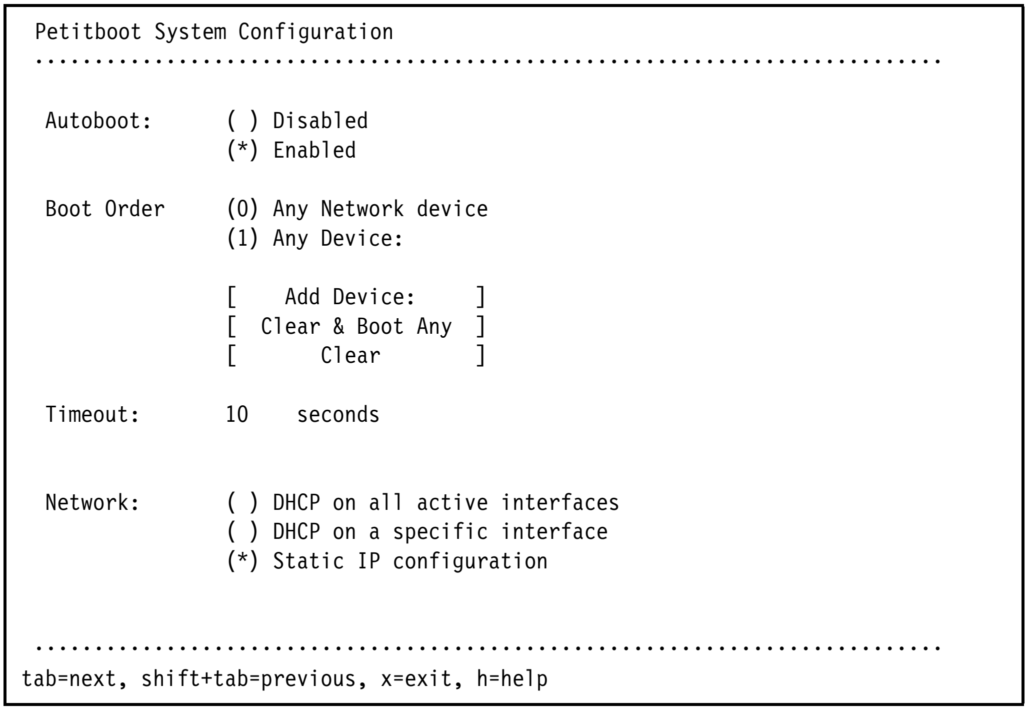
4. At the Petitboot welcome screen, select `System configuration`, as shown in Figure 2-36.



```
Petitboot (v1.4.4-e1658ec)                                  8247-21L 2142B2A
............................................................................


*System information
 System configuration
 System status log
 Language
 Rescan devices
 Retrieve config from URL
 Exit to shell








............................................................................
Enter=accept, e=edit, n=new, x=exit, l=language, g=log, h=help
```

Figure 2-36 Petitboot bootloader welcome screen

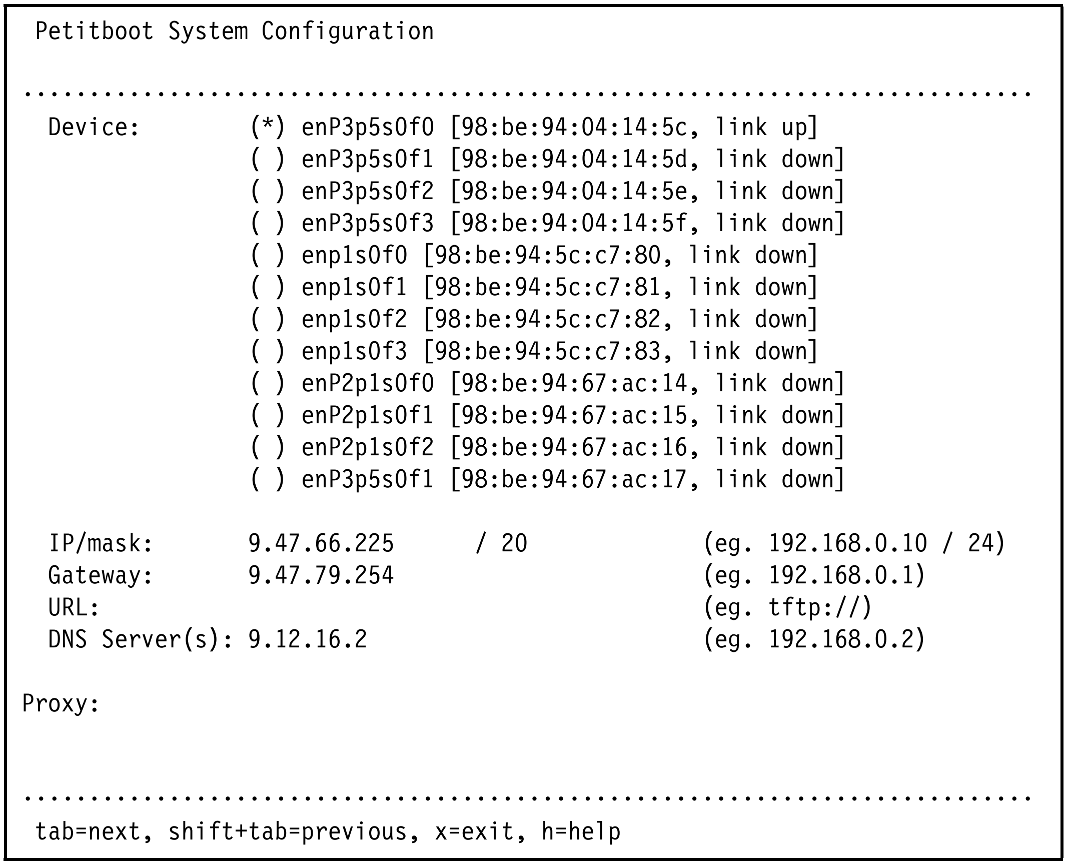
5. In Figure 2-37, leave the default values in the following fields:
   - Autoboot: `Enabled`.
   - Boot Order: `Any Device`.
   - Network: `Static IP Configuration`.



```
 Petitboot System Configuration
 ............................................................................

  Autoboot:       ( ) Disabled
                  (*) Enabled

  Boot Order      (0) Any Network device
                  (1) Any Device:

                  [    Add Device:     ]
                  [  Clear & Boot Any  ]
                  [       Clear        ]

  Timeout:        10    seconds


  Network:        ( ) DHCP on all active interfaces
                  ( ) DHCP on a specific interface
                  (*) Static IP configuration


 ............................................................................
tab=next, shift+tab=previous, x=exit, h=help
```

Figure 2-37 Petitboot bootloader system configuration menu

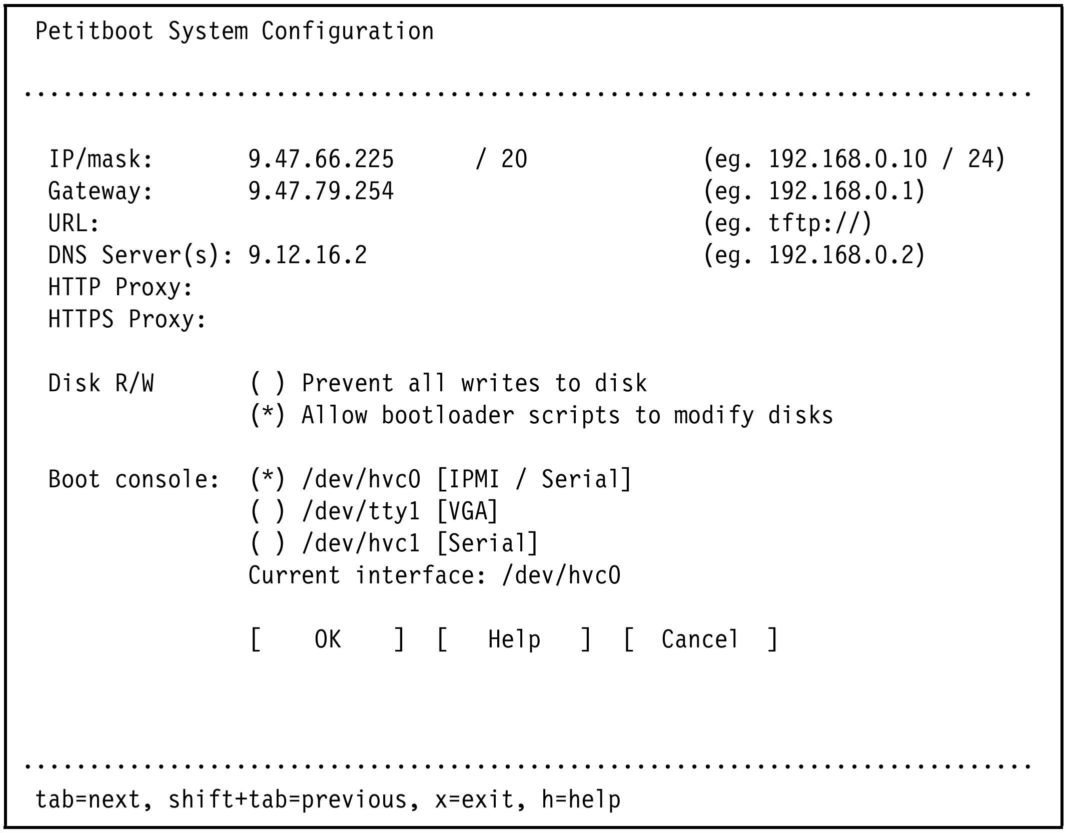
6. Move to the next section by using the Tab or arrow keys and select the physical network adapter with an active Ethernet link by using the spacebar key. Configure all the standard network adapter settings for your server, such as the IP address, netmask, gateway, and DNS, as shown in Figure 2-38.



```
 Petitboot System Configuration

............................................................................
  Device:        (*) enP3p5s0f0 [98:be:94:04:14:5c, link up]
                 ( ) enP3p5s0f1 [98:be:94:04:14:5d, link down]
                 ( ) enP3p5s0f2 [98:be:94:04:14:5e, link down]
                 ( ) enP3p5s0f3 [98:be:94:04:14:5f, link down]
                 ( ) enp1s0f0 [98:be:94:5c:c7:80, link down]
                 ( ) enp1s0f1 [98:be:94:5c:c7:81, link down]
                 ( ) enp1s0f2 [98:be:94:5c:c7:82, link down]
                 ( ) enp1s0f3 [98:be:94:5c:c7:83, link down]
                 ( ) enP2p1s0f0 [98:be:94:67:ac:14, link down]
                 ( ) enP2p1s0f1 [98:be:94:67:ac:15, link down]
                 ( ) enP2p1s0f2 [98:be:94:67:ac:16, link down]
                 ( ) enP3p5s0f1 [98:be:94:67:ac:17, link down]

  IP/mask:       9.47.66.225      / 20            (eg. 192.168.0.10 / 24)
  Gateway:       9.47.79.254                      (eg. 192.168.0.1)
  URL:                                            (eg. tftp://)
  DNS Server(s): 9.12.16.2                        (eg. 192.168.0.2)

Proxy:


............................................................................
 tab=next, shift+tab=previous, x=exit, h=help
```

Figure 2-38 Physical network port configuration within the Petitboot System Configuration menu

7. When you have entered all the network settings, go to the bottom of the screen by using the Tab or arrow keys and select `OK`, as shown in Figure 2-39.



```
 Petitboot System Configuration
.......................................................................

  IP/mask:       9.47.66.225      / 20           (eg. 192.168.0.10 / 24)
  Gateway:       9.47.79.254                     (eg. 192.168.0.1)
  URL:                                           (eg. tftp://)
  DNS Server(s): 9.12.16.2                       (eg. 192.168.0.2)
  HTTP Proxy:
  HTTPS Proxy:

  Disk R/W       ( ) Prevent all writes to disk
                 (*) Allow bootloader scripts to modify disks

  Boot console:  (*) /dev/hvc0 [IPMI / Serial]
                 ( ) /dev/tty1 [VGA]
                 ( ) /dev/hvc1 [Serial]
                 Current interface: /dev/hvc0

                 [    OK    ]  [   Help   ]  [  Cancel  ]



.......................................................................
 tab=next, shift+tab=previous, x=exit, h=help
```

Figure 2-39 Petitboot bootloader System Configuration submenu

8. Go back to the main menu by pressing the Esc key, then press the g key to view the logs. The output that is shown in Example 2-12 should be visible at the top of the screen.

Example 2-12 Successful network adapter configuration log entry on the Petitboot menu

```
[enP3p5s0f0] Configuring with static address (ip: 9.47.66.225/20)
```

The Petitboot bootloader has successfully configured the network interface.

Ping the configured IP address by opening, on your workstation, another Windows cmd session and pinging the IP. If you receive a response to your ping, you are ready for the installation; otherwise, review your network settings or contact your network administrator.

As an alternative, you can skip configuring the interface network settings through the Petitboot menu and do it directly through the CLI. When you are on the main menu, press the Esc key. You see the classic root command prompt (#), as shown in Example 2-13.

Example 2-13 Manual configuration of the network interface under the Petitboot bootloader command prompt

```
# /sbin/ip link set lo up
# /sbin/ip link set enP3p5s0f0 up
# /sbin/ip address add 9.47.66.225/20 dev enP3p5s0f0
# /sbin/ip link set enP3p5s0f0 up
# /sbin/ip route add default via 9.47.79.254
```

9. Go back to the Petitboot bootloader welcome screen by running the **`exit`** command. Press the n key to create a boot profile.

   In this case, we use a network boot through FTP, but you may use HTTP, NFS, and SFTP, depending on how your network is set up.

   On the boot profile, complete the fields as described below:

   – Device: Specify paths/URLs manually.

   – Kernel: `ftp://<your FTP server/<path to your RHEL vmzlinux image>`.

   – Initrd: `ftp://<your ftp server>/<path to your RHEL initrd.img>`.

   – Device tree: Leave this blank.

   – Boot Arguments: `inst.text inst.repo=ftp://<your ftp server>/<path to your rhel os directory> ro ifname=<interface name>:<mac address> ip=<ip address>::<gateway>:<netmask>:<hostname>:<interface name>:none nameserver=<dns server>`.

   Example 2-14 through Example 2-16 show how the boot arguments should be completed in an FTP network boot scenario, with user and password authentication requirements.

Important: When you use the FTP protocol as your network boot source and your FTP server requires authentication, you must ensure that the “@” characters the are used for login are expressed as “%40” or the download fails.

Note: When using ipmiutil on Windows, there is a potential limitation when pasting content from the clipboard into the Windows cmd.exe: The cmd terminal truncates the text. You must be careful and copy the long kernel, initrd, and boot syntax lines in portions. This situation might not be the case when using the ipmitool on Linux.

Example 2-14 The vmlinuz syntax for a successful network boot

```
ftp://bazan%40ar.ibm.com:mypass@ftp3.linux.ibm.com/redhat/release_cds/RHEL-7
.5-GA/Server/ppc64le/os/ppc/ppc64/vmlinuz
```

Example 2-15 The initrd syntax for a successful network boot

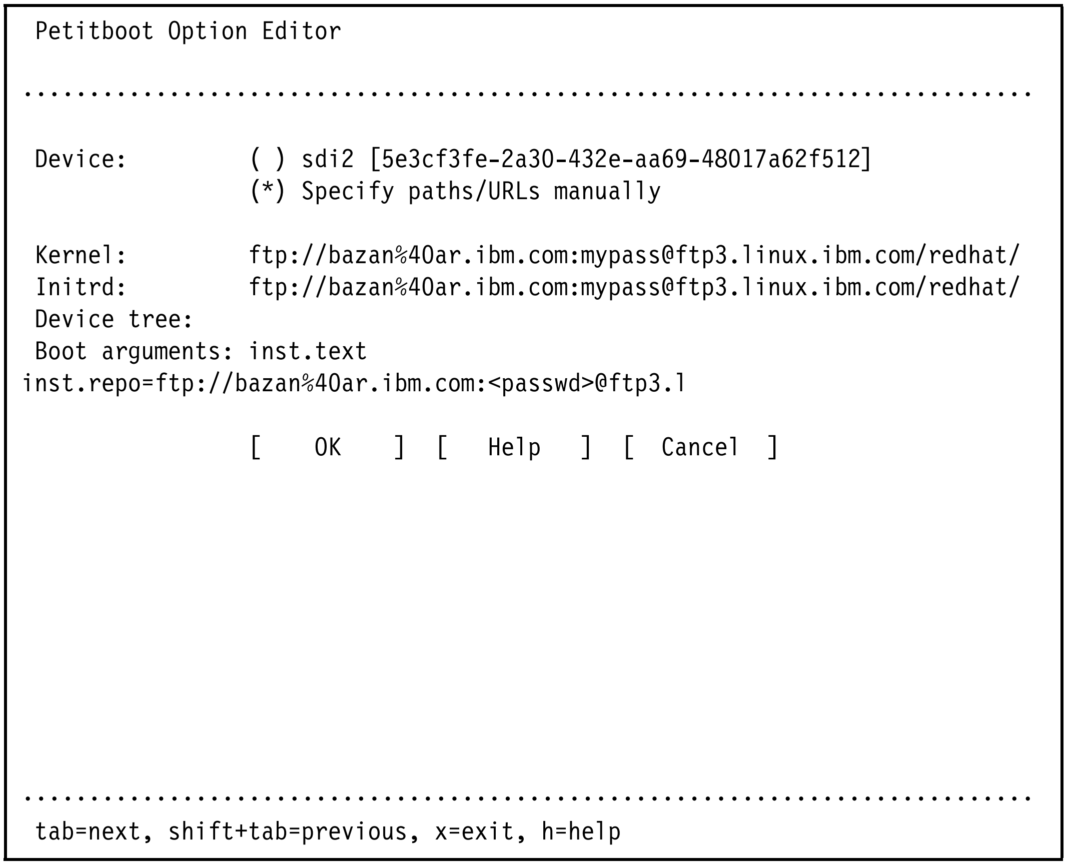
```
ftp://bazan%40ar.ibm.com:mypass@ftp3.linux.ibm.com/redhat/release_cds/RHEL-7
.5-GA/Server/ppc64le/os/ppc/ppc64/initrd.img
```

Example 2-16 The boot arguments syntax for a successful network boot

```
inst.text
inst.repo=ftp://bazan%40ar.ibm.com:mypass@ftp3.linux.ibm.com/redhat/release_
cds/RHEL-7.5-GA/Server/ppc64le/os/ ro ifname=enP3p5s0f0:98:be:94:04:14:5c
ip=9.47.66.225::9.47.79.254:255.255.240.0:p8-2142b2a-rhel-kvm:enP3p5s0f0:non
e nameserver=9.12.16.2
```

Note: Example 2-16 shows how to start the Red Hat Enterprise Linux installation in text mode by using `inst.text` as the first argument. You can also start the installation in VNC mode by placing `inst.vnc` as the first argument, and then connect to the server IP address at port 1.

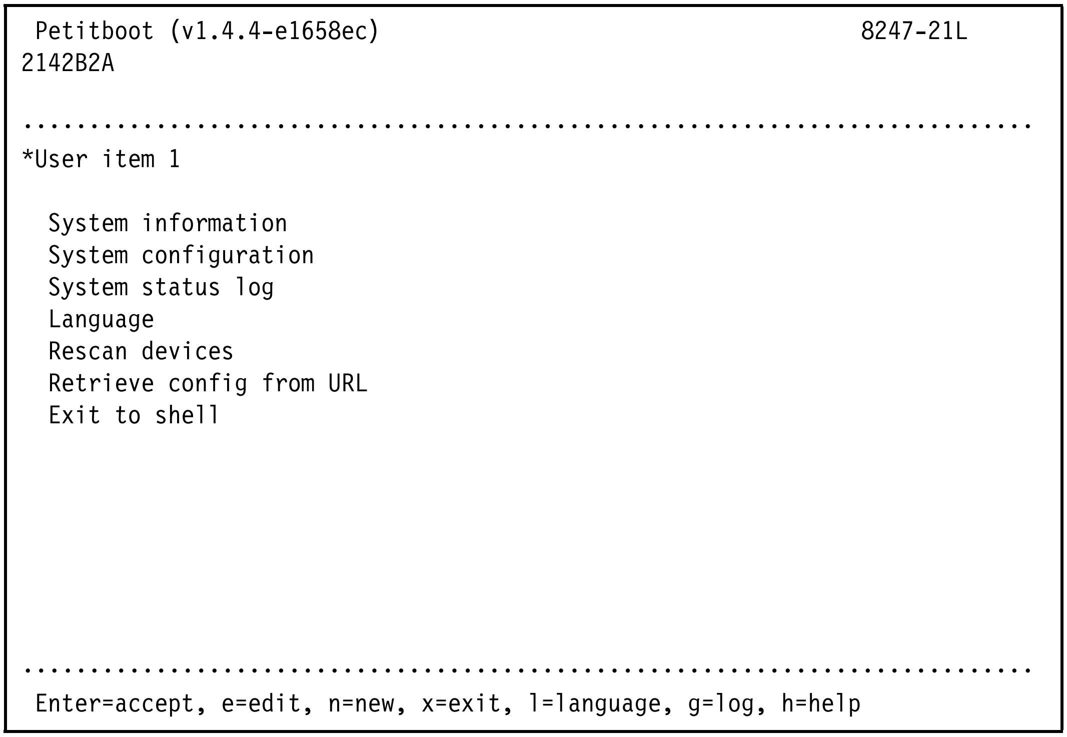
10. When all the fields are complete, create the network boot profile by selecting **OK**, as shown in Figure 2-40.



```
 Petitboot Option Editor

............................................................................

 Device:         ( ) sdi2 [5e3cf3fe-2a30-432e-aa69-48017a62f512]
                 (*) Specify paths/URLs manually

 Kernel:         ftp://bazan%40ar.ibm.com:mypass@ftp3.linux.ibm.com/redhat/
 Initrd:         ftp://bazan%40ar.ibm.com:mypass@ftp3.linux.ibm.com/redhat/
 Device tree:
 Boot arguments: inst.text
inst.repo=ftp://bazan%40ar.ibm.com:<passwd>@ftp3.l

                 [    OK    ]  [   Help   ]  [  Cancel  ]









............................................................................
 tab=next, shift+tab=previous, x=exit, h=help
```

Figure 2-40 Petitboot bootloader boot profile creation

11.Back at the Petitboot bootloader welcome menu, the new boot profile should be visible as `User item 1`, as shown in Figure 2-41. Select it and press Enter to start the network boot.



```
 Petitboot (v1.4.4-e1658ec)                                    8247-21L
2142B2A

.........................................................................
*User item 1

  System information
  System configuration
  System status log
  Language
  Rescan devices
  Retrieve config from URL
  Exit to shell










.........................................................................
 Enter=accept, e=edit, n=new, x=exit, l=language, g=log, h=help
```

Figure 2-41 User item 1 boot profile

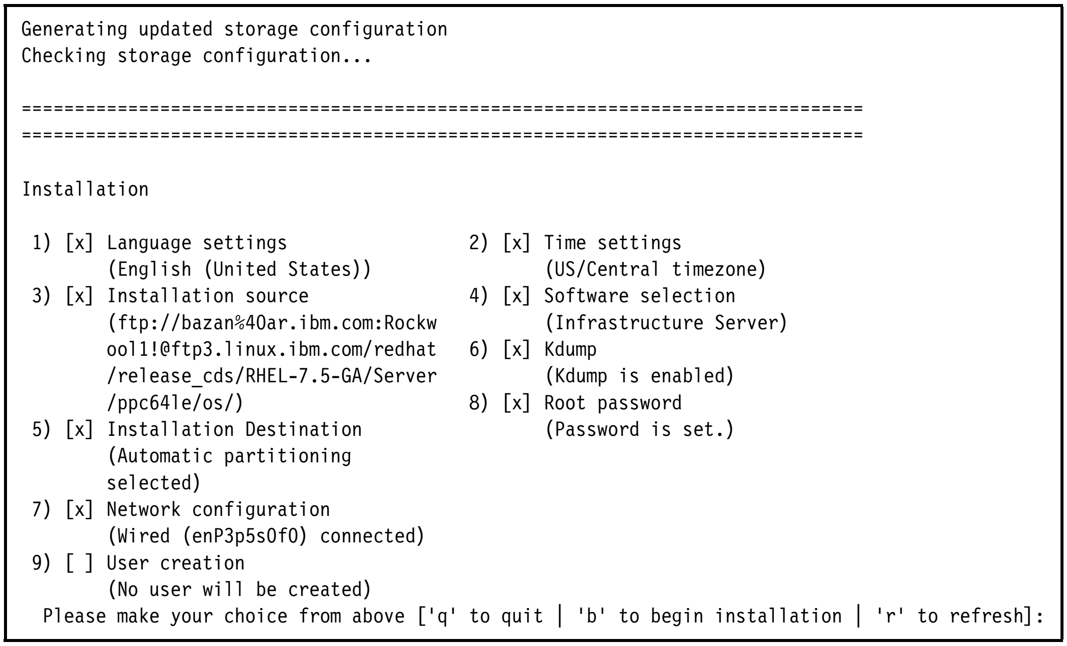
The system should now boot to the Red Hat Enterprise Linux installation wizard by using **dracut** and **curl**, as shown in Example 2-17.

Example 2-17 The Petiboot bootloader downloads the vmlinuz and initrd images by using dracut

```
2 downloads in progress...
The system is going down NOW!
Sent SIGTERM to all processes
Sent SIGKILL to all processes
[72295.388713] kexec_core: Starting new kernel
[    0.000000] OPAL V3 detected !
[    0.000000] Using PowerNV machine description
[    0.000000] Page sizes from device-tree:
[    0.000000] base_shift=12: shift=12, sllp=0x0000, avpnm=0x00000000,
[   17.982604] dracut-initqueue[1222]: % Total    % Received % Xferd  Average
Speed   Time    Time     Time  Current
[   17.983291] dracut-initqueue[1222]: Dload  Upload   Total   Spent    Left
Speed
100  386M  100  386M    0     0  30.7M      0  0:00:12  0:00:12 --:--:-- 41.9M
0 --:--:-- --:--:-- --:--:--     0
```

The Red Hat Enterprise Linux installation wizard starts. For more information about how to use the wizard, see the *Red Hat Enterprise Linux 7 Installation Guide*.

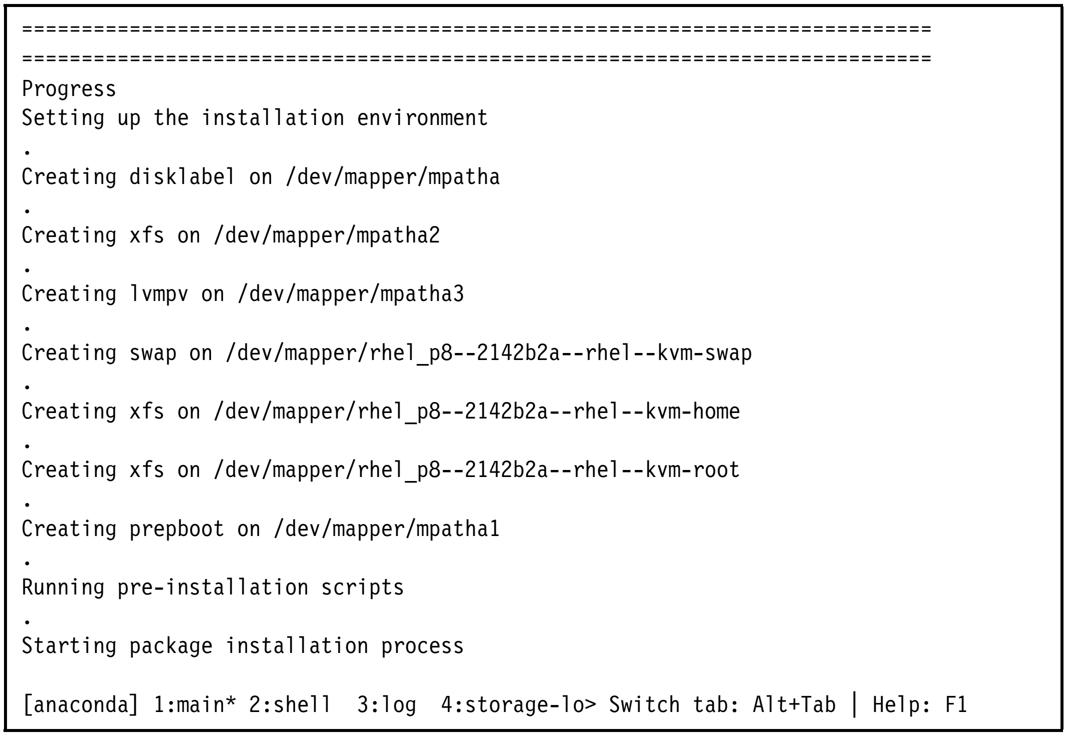
12. After you set all your Red Hat Enterprise Linux installation settings, your installation menu should look like Figure 2-42.



```
Generating updated storage configuration
Checking storage configuration...

===============================================================================
===============================================================================

Installation

 1) [x] Language settings                 2) [x] Time settings
        (English (United States))                (US/Central timezone)
 3) [x] Installation source               4) [x] Software selection
        (ftp://bazan%40ar.ibm.com:Rockw          (Infrastructure Server)
        ool1!@ftp3.linux.ibm.com/redhat   6) [x] Kdump
        /release_cds/RHEL-7.5-GA/Server          (Kdump is enabled)
        /ppc64le/os/)                     8) [x] Root password
 5) [x] Installation Destination                 (Password is set.)
        (Automatic partitioning
        selected)
 7) [x] Network configuration
        (Wired (enP3p5s0f0) connected)
 9) [ ] User creation
        (No user will be created)
  Please make your choice from above ['q' to quit | 'b' to begin installation | 'r' to refresh]:
```

Figure 2-42 Red Hat Enterprise Linux V7.5 installation menu

Press the b key to begin the installation, as shown in Figure 2-43.



```
============================================================================
============================================================================
Progress
Setting up the installation environment
.
Creating disklabel on /dev/mapper/mpatha
.
Creating xfs on /dev/mapper/mpatha2
.
Creating lvmpv on /dev/mapper/mpatha3
.
Creating swap on /dev/mapper/rhel_p8--2142b2a--rhel--kvm-swap
.
Creating xfs on /dev/mapper/rhel_p8--2142b2a--rhel--kvm-home
.
Creating xfs on /dev/mapper/rhel_p8--2142b2a--rhel--kvm-root
.
Creating prepboot on /dev/mapper/mpatha1
.
Running pre-installation scripts
.
Starting package installation process

[anaconda] 1:main* 2:shell  3:log  4:storage-lo> Switch tab: Alt+Tab | Help: F1
```

Figure 2-43 Red Hat Enterprise Linux installation begins

Note: Depending on your local network environment, the installer might take some time to obtain the Red Hat Enterprise Linux packages from your FTP server. The installation might display "`Starting Package installation...`" for a while, but that does not mean that the installation failed. Allow for some time for this step to complete.

13. After the operating system starts for the first time, you must register it with the Red Hat Subscription Manager Services (RHSM) to your local Red Hat Satellite server.

    You must have access to the Extended Update Support (EUS) and Red Hat Virtualization (RHV).

    For more information about EUS repositories, see Extended Update Support (EUS) Standard Operating Environment (SOE) Guides and What is the Red Hat Enterprise Linux Extended Update Support Subscription?.

    For more information about RHV licensing, see Red Hat Virtualization.

    After you have addressed all the Red Hat subscription requirements, you might want to enable all ppc64 repositories that are available to your subscription because they are needed by the IBM PowerVC drivers at the time your Red Hat Enterprise Linux KVM server is added as a compute node. Running the bash loop that is shown in Example 2-18 can accomplish this objective.

    Example 2-18 Bash loop to enable all available repositories from your current configured subscriptions

```
subscription-manager repos --list | grep "Repo ID:" | awk '{print $3}' | sort |
uniq | while read REPO_ID
do
subscription-manager repos --enable=${REPO_ID}
done
```

    Extra packages are needed so that the IBM PowerVC drivers can be successfully added to the node, which cannot be obtained through the standard Red Hat ppc64 repositories. Proceed to manually create Extra Packages for Enterprise Linux (EPEL), RHV, and PowerVM NovaLink repositories because they contain all the required kvm-novalink package dependencies. To simplify the process, you can put all the repositories in a single repository file. Create `/etc/yum.repos.d/kvm.repo` and append the content that is listed in Example 2-19.

    Example 2-19 Content that must be placed in the kvm.repository file

```
[ftp3-rhv42ga]
name=FTP3 RHV 4.2 GA
baseurl=ftp://bazan%40ar.ibm.com:mypass@ftp3.linux.ibm.com/redhat/release_cds/R
HV/4.2-GA/Management-Agent-Power-7/ppc64le/os/
gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-redhat-release

[novalink]
name=NovaLink
baseurl=http://public.dhe.ibm.com/systems/virtualization/Novalink/rhel/73/noval
ink_1.0.0
failovermethod=priority
enabled=1
gpgcheck=0

[epel]
name=Extra Packages for Enterprise Linux 7 - $basearch
```

```
metalink=https://mirrors.fedoraproject.org/metalink?repo=epel-7&arch=$basearch
failovermethod=priority
enabled=1
gpgcheck=0

[epel-debuginfo]
name=Extra Packages for Enterprise Linux 7 - $basearch - Debug
metalink=https://mirrors.fedoraproject.org/metalink?repo=epel-debug-7&arch=$bas
earch
failovermethod=priority
enabled=0
gpgcheck=0

[epel-source]
name=Extra Packages for Enterprise Linux 7 - $basearch - Source
metalink=https://mirrors.fedoraproject.org/metalink?repo=epel-source-7&arch=$ba
search
failovermethod=priority
enabled=0
gpgcheck=0
```

14. Update the operating system to the current version and perform a system restart so that the updates take effect immediately by running the following commands:

```
yum -y update
reboot
```

2.7 Installing and configuring IBM PowerVC

If you already have IBM PowerVC, ensure that it is upgraded to IBM PowerVC V1.4.1 or later.

If you are performing a new installation, IBM Knowledge Center for IBM PowerVC has a wealth of information about this topic. See IBM Knowledge Center: Installing and uninstalling and IBM Knowledge Center: Hardware and software requirements.

Note: A minimum of 16 GB of memory RAM is required for the IBM PowerVC partition to manage the IBM Spectrum Scale cluster.

2.7.1 Adding IBM Spectrum Scale to IBM PowerVC

SDI can be used with either IBM PowerVC Standard or IBM Cloud PowerVC Manager Edition. However, servers that use SDS must use IBM Spectrum Scale.

There is an orderable bundle that is called IBM Cloud PowerVC Manager for SDI, which includes entitlement to both IBM Cloud PowerVC Manager and IBM Spectrum Scale Data Management Edition.

After your IBM PowerVC installation is done, you must copy the IBM Spectrum Scale Data Management V5.0.x installers to IBM PowerVC. Download the installation packages, without decompressing them, and put them in `/opt/ibm/powervc/images/spectrum-scale/`.

Download all the available installers for all architectures. For example, if your IBM PowerVC installation is running on Red Hat Enterprise Linux BE and your cloud compute nodes are on Ubuntu 16.04 LE, then you must download the ppc64 BE and ppc64 LE installers to that location.

Example 2-20 shows the copied installation binary files in IBM PowerVC.

Example 2-20 IBM Spectrum Scale installation binary files that are copied into the IBM PowerVC installation path

```
[root@powervc01 ~]# ls -l  /opt/ibm/powervc/images/spectrum-scale/
total 2176476
-rw-r--r--. 1 root root      9116 Jun 11 18:47 ibmpowervc-powervm-gpfs-1.4.1.0-1.noarch.rpm
-rw-r--r--. 1 root root      5936 Jun 11 18:47 ibmpowervc-powervm-gpfs_1.4.1.0-ubuntu1_all.deb
-r-xr-xr-x. 1 root root     16915 Jun 11 18:47 install
-rwxr-xr-x. 1 root root 861305698 Jun 16 18:53
Spectrum_Scale_Data_Management-5.0.1.1-ppc64LE-Linux-install
-rwxr-xr-x. 1 root root 461442400 Jun 16 18:53
Spectrum_Scale_Data_Management-5.0.1.1-ppc64-Linux-install
-rwxr-xr-x. 1 root root 905916774 Jun 16 18:53
Spectrum_Scale_Data_Management-5.0.1.1-x86_64-Linux-install
```

3

Implementing software-defined storage in IBM PowerVC

This chapter describes how to implement software-defined storage (SDS) in IBM PowerVC.

3.1 Preparing your compute nodes

So that IBM Spectrum Scale can communicate with IBM PowerVC and vice versa without any issues, you must open ports in the firewall for both the Red Hat Enterprise Linux kernel-based virtual machine (KVM) and the Ubuntu PowerVM NovaLink nodes.

Complete the following steps:

1. In a new Red Hat Enterprise Linux v7.x installation, `firewalld` is the firewall service that comes active by default, which means that your default firewall zone is always labeled as `public`. You can still verify this on your compute node by running the following command:

   ```
   firewall-cmd --get-active-zones
   ```

   Example 3-1 shows the output of this command.

   Example 3-1 Current active firewall zone attached to the primary network interface of the node

   ```
   [root@p8-2142b2a-rhel-kvm ~]# firewall-cmd --get-active-zones
   public
     interfaces: enP3p5s0f0
   ```

2. Using the firewall zone name that you obtained in Example 3-1, run the following commands to open the required network ports:

   ```
   firewall-cmd --permanent --zone=public --add-port=1191/tcp
   firewall-cmd --permanent --zone=public --add-port=4789/tcp
   firewall-cmd --permanent --zone=public --add-port=4789/udp
   firewall-cmd --permanent --zone=public --add-port=5000/tcp
   firewall-cmd --permanent --zone=public --add-port=5671/tcp
   firewall-cmd --permanent --zone=public --add-port=5901/tcp
   firewall-cmd --permanent --zone=public --add-port=8080/tcp
   firewall-cmd --permanent --zone=public --add-port=8472/tcp
   firewall-cmd --permanent --zone=public --add-port=8472/udp
   firewall-cmd --permanent --zone=public --add-port=8774/tcp
   firewall-cmd --permanent --zone=public --add-port=9000/tcp
   firewall-cmd --permanent --zone=public --add-port=9292/tcp
   firewall-cmd --permanent --zone=public --add-port=9696/tcp
   firewall-cmd --permanent --zone=public --add-port=31000/tcp
   firewall-cmd --permanent --zone=public --add-port=32047/tcp
   firewall-cmd --permanent --zone=public --add-port=49152/tcp
   firewall-cmd --permanent --zone=public --add-port=49153/tcp
   firewall-cmd --permanent --zone=public --add-port=49154/tcp
   firewall-cmd --permanent --zone=public --add-port=49155/tcp
   firewall-cmd --permanent --zone=public --add-port=49156/tcp
   systemctl restart firewalld.service
   ```

3. If you are not using a firewall, you may permanently disable `firewalld` now and on system startup by running the following commands:

   ```
   systemctl stop firewalld.service
   systemctl disable firewalld.service
   ```

4. If you replaced `firewalld.service` with `iptables` or are using Ubuntu 16.xx, then run the following commands to disable this service now and on system startup:

   ```
   service iptables save
   service iptables stop
   iptables -flush
   service iptables disable
   ```

For more information about IBM PowerVC and software-defined infrastructure (SDI) network ports, see IBM Knowledge Center.

3.2 Using multiple bridges in a software-defined storage environment

After you install the PowerVM NovaLink software, the system has just one bridge, which is called br-ex. This single bridge is sufficient to run an SDS environment. However, as described in 2.2.2, “Installing a PowerVM NovaLink partition on an empty system that is under control of an HMC” on page 25, if you use internal disks in File Placement Optimizer (FPO) mode (where data is tripled in the cluster), it is a good idea to separate that traffic from the other network traffic. To do so, create a separate bridge for the storage traffic that is named, for example, br-st. The implementation of this second bridge is described in detail in 2.2, “Installing PowerVM NovaLink for a system that uses PowerVM” on page 22.

Figure 3-1 shows a system with two bridges inside the PowerVM NovaLink virtual machine (VM).

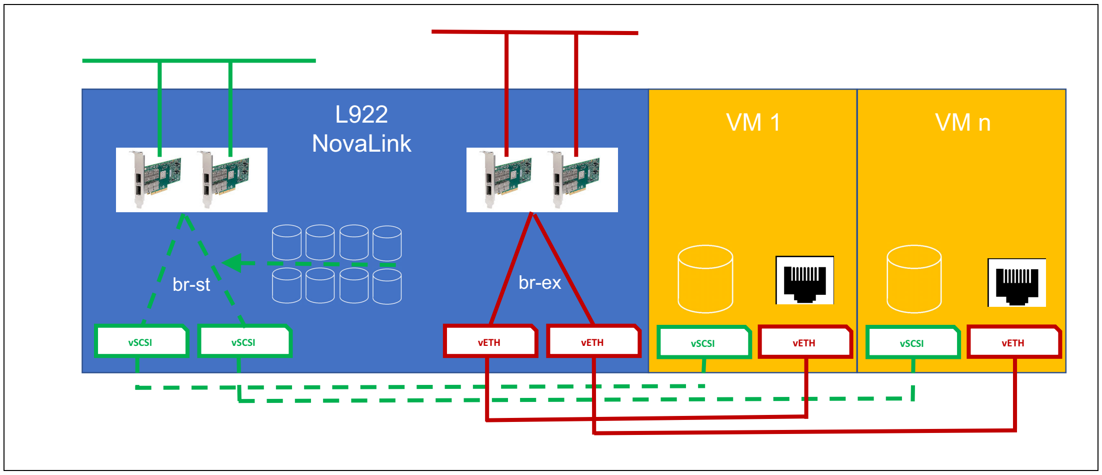


Figure 3-1 System with multiple bridges

Every bridge should also have an IP address that is assigned to it. For the storage bridge, this address might be, for example, a private network that is accessible only by all the compute nodes and the IBM PowerVC management server. IBM PowerVC is the IBM Spectrum Scale cluster manager, so it needs access as well.

When you add the first node to IBM PowerVC, which creates the IBM Spectrum Scale cluster, select the IP address of the bridge that you want to use for the storage traffic. This address cannot be changed afterward except by a complete removal of the cluster.

Note: IBM PowerVC manages the IBM Spectrum Scale cluster for you. Do not try to modify the cluster yourself by using IBM Spectrum Scale commands because IBM PowerVC created a special cluster configuration that should not be changed.

3.3 Adding compute nodes to IBM PowerVC

Your bare metal Ubuntu PowerVM NovaLink and Red Hat Enterprise Linux KVM hosts are ready to be added and converted into compute nodes in IBM PowerVC.

IBM PowerVC has three basic resource types: compute, storage, and network. In the SDS mode, when you register a compute node, it automatically creates the associated storage.

You can still add non-software-defined nodes to IBM PowerVC. You are not locked into one type of resource. Mixing servers that are software-defined with those that have traditional Virtual I/O Server (VIOS) is supported and encouraged.

After you add a host, the system examines whether the server is installed in software-defined environment (SDE) mode or is a KVM host. If either is true, it installs the IBM Spectrum Scale software on the IBM PowerVC server and the compute node, and then builds a cluster. The registration also automatically creates a storage provider behind the scenes.

The system uses analytics to determine the type of storage that is available and creates the appropriate cluster. It identifies the best disk for metadata, determines the redundancy levels that are needed (local versus remote storage), sets up the appropriate storage rules, and so on.

Note: Adding the host can take several minutes. With IBM PowerVC, you may register multiple systems at once, but in SDS mode, the registrations are queued. As hosts are added (or removed), the storage pool grows (or shrinks) dynamically.

If you have one host with local storage, no redundancy is available. As you add a second host with local storage, the data is duplicated across the hosts. The third host adds triplication. Beyond that, storage is added.

When using local disks, add at least three hosts before too much data is in the system. If there is a large amount of data in the system, the host registration must copy it all to the second and third system during the registration, which might take a long time depending on your native disk and network speed.

When registration is complete, the IBM PowerVC server is mounted into the IBM Spectrum Scale cluster. The path for it is `/gpfs/powervc_gpfs/infrastructure/`, which is where volumes are created. The IBM PowerVC server does not provide any disks; it is using the storage that is provided by the server. All accesses for that storage run over the network.

If you happen to be familiar with IBM Spectrum Scale concepts such as quorum, High Availability Write Cache (HAWC) or Local Read-Only Cache (LROC), IBM PowerVC automatically manages them for you. The system administrator should not be interacting with the IBM Spectrum Scale management functions directly. IBM PowerVC supports a specific configuration.

3.3.1 Considerations before adding a node

As with any cluster configuration, the host names and IP addresses must be consistent across the environment. The preferred practice is to use a Domain Name Server (DNS). All the nodes in the cluster (including your IBM PowerVC server) must be properly registered in the DNS domain zone with both forward (A records) and reverse (PTR records), and the IP address that is configured on the node's network interface must also match the DNS. Host name miss-matches result in the denial of a host add/remove task.

Ensure that the hosts keys match what the IBM PowerVC server and the compute nodes have recorded in the `known_hosts` files. If you have reinstalled a node or if you have renamed it, you must remove the old host key entry to avoid any conflicts during the add node and remove node routines. To accomplish this task, delete the offending entries in the following files:

```
/opt/ibm/powervc/data/known_hosts
/root/.ssh/known_hosts
```

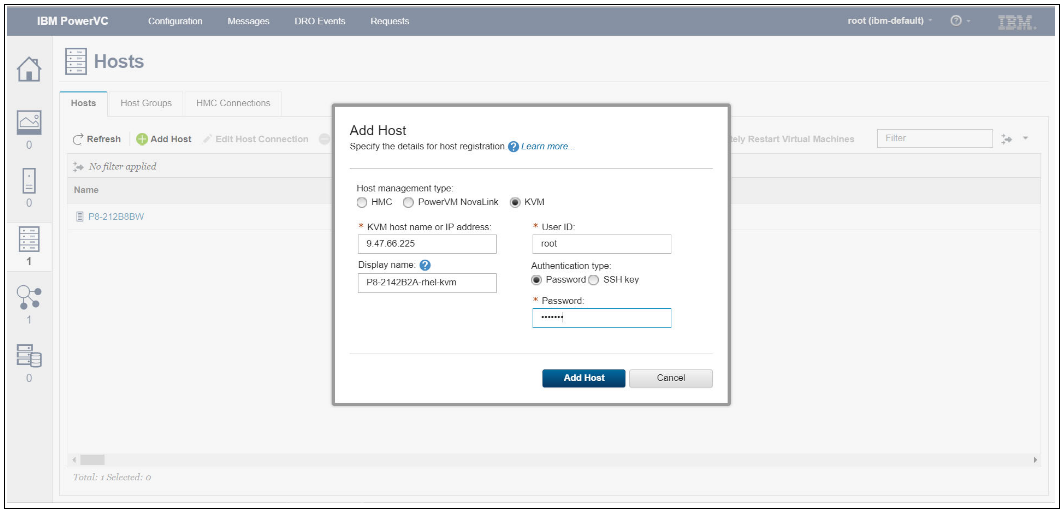
> **Note:** The host key files must be reviewed, and if applicable, cleaned on the IBM PowerVC server, but also on any node that is added to the IBM Spectrum Scale cluster.

3.3.2 Adding a node step-by-step

To connect a compute node, complete the following steps:

1. Sign in to IBM PowerVC.
2. Click **Add** next to the hosts.

   Figure 3-2 shows the IBM PowerVC Add Host dialog.



Add Host

Specify the details for host registration. Learn more...

Host management type:
○ HMC ○ PowerVM NovaLink ● KVM

* KVM host name or IP address: 9.47.66.225

Display name: P8-2142B2A-rhel-kvm

* User ID: root

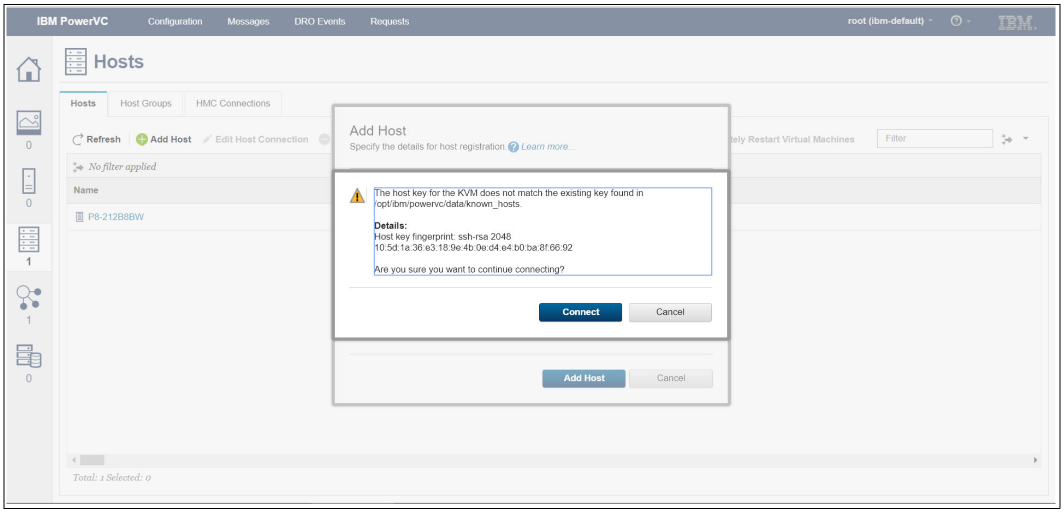
Authentication type:
● Password ○ SSH key

* Password: •••••••

Figure 3-2 IBM PowerVC Add Host dialog

3. Click **Connect** when prompted to accept/validate the host key.

   Figure 3-3 shows the IBM PowerVC Add Host dialog for accepting the host key.



The host key for the KVM does not match the existing key found in /opt/ibm/powervc/data/known_hosts.

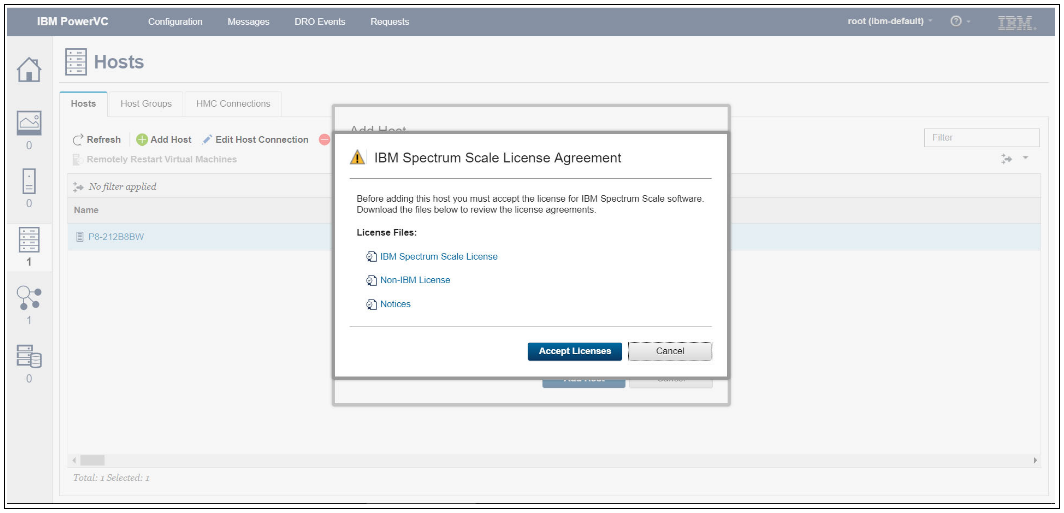
Details:
Host key fingerprint: ssh-rsa 2048
10:5d:1a:36:e3:18:9e:4b:0e:d4:e4:b0:ba:8f:66:92

Are you sure you want to continue connecting?

Figure 3-3 Add Host dialog for accepting the host key

4. Click **Accept Licenses** when prompted to review and accept the IBM Spectrum Scale license.

   Figure 3-4 shows IBM Spectrum Scale License Agreement review and acceptance dialog.



Before adding this host you must accept the license for IBM Spectrum Scale software. Download the files below to review the license agreements.

Figure 3-4 IBM Spectrum Scale License Agreement acceptance dialog

5. Here is a list of high-level tasks that occur in the background when IBM PowerVC starts the host addition for IBM Spectrum Scale:

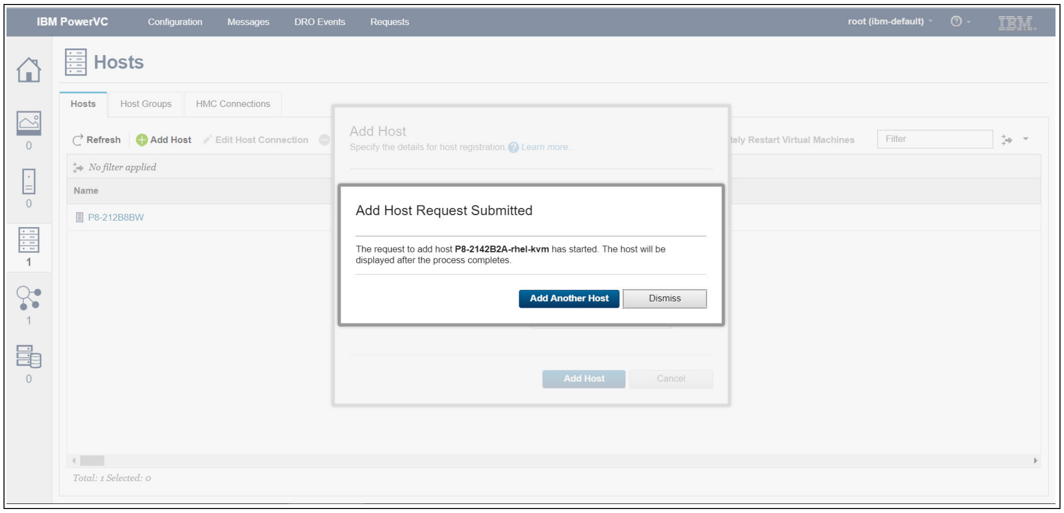
   a. If IBM PowerVC is creating a IBM Spectrum Scale cluster for the first time, then the IBM Spectrum Scale (IBM General Parallel File System (IBM GPFS))) binary files are installed on it. If the registration detects that a GPFS cluster exists, everything is reverted and no action is taken (no impact to the existing cluster).

   b. Install IBM PowerVC drivers on the compute node.

   c. Install IBM Spectrum Scale on the compute node.

   d. Create the cluster and make it available.

Figure 3-6 shows the Add Host dialog.



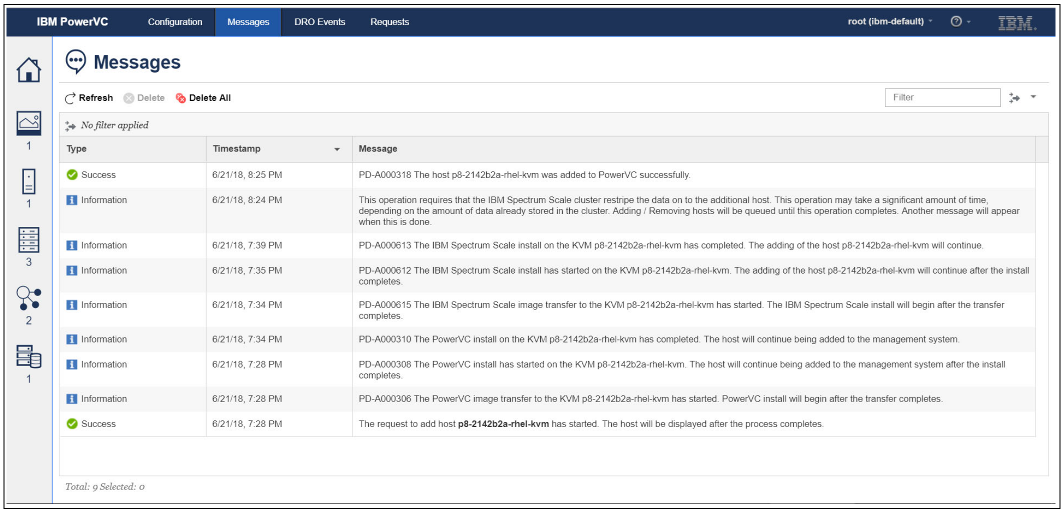
Add Host Request Submitted

The request to add host **P8-2142B2A-rhel-kvm** has started. The host will be displayed after the process completes.

Add Another Host | Dismiss

Figure 3-5 IBM PowerVC Add Host dialog

6. After the add node routine completes, you can go into the IBM PowerVC message center and review the log entries. A successful node registration should look like Figure 3-6.

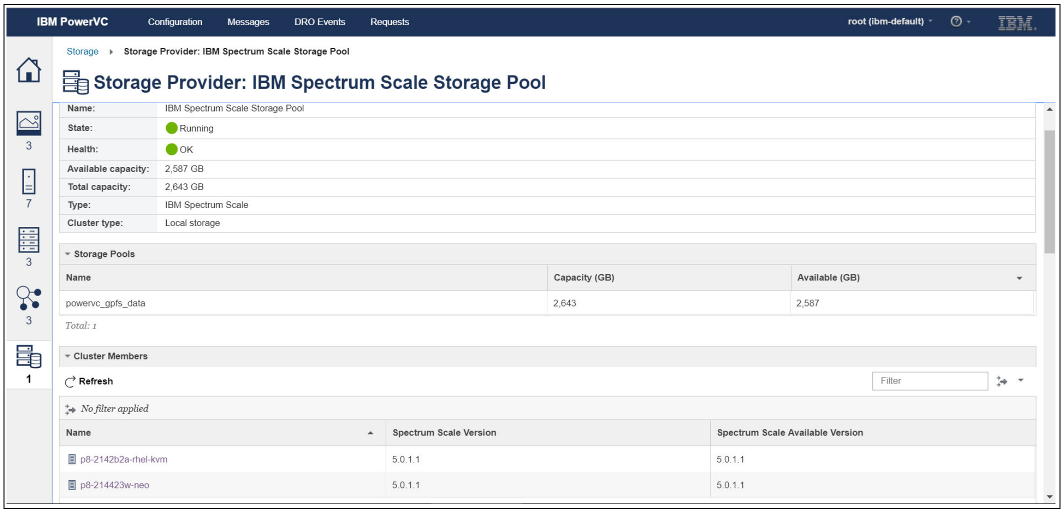


Type	Timestamp	Message
✓ Success	6/21/18, 8:25 PM	PD-A000318 The host p8-2142b2a-rhel-kvm was added to PowerVC successfully.
Information	6/21/18, 8:24 PM	This operation requires that the IBM Spectrum Scale cluster restripe the data on to the additional host. This operation may take a significant amount of time, depending on the amount of data already stored in the cluster. Adding / Removing hosts will be queued until this operation completes. Another message will appear when this is done.
Information	6/21/18, 7:39 PM	PD-A000613 The IBM Spectrum Scale install on the KVM p8-2142b2a-rhel-kvm has completed. The adding of the host p8-2142b2a-rhel-kvm will continue.
Information	6/21/18, 7:35 PM	PD-A000612 The IBM Spectrum Scale install has started on the KVM p8-2142b2a-rhel-kvm. The adding of the host p8-2142b2a-rhel-kvm will continue after the install completes.
Information	6/21/18, 7:34 PM	PD-A000615 The IBM Spectrum Scale image transfer to the KVM p8-2142b2a-rhel-kvm has started. The IBM Spectrum Scale install will begin after the transfer completes.
Information	6/21/18, 7:34 PM	PD-A000310 The PowerVC install on the KVM p8-2142b2a-rhel-kvm has completed. The host will continue being added to the management system.
Information	6/21/18, 7:28 PM	PD-A000308 The PowerVC install has started on the KVM p8-2142b2a-rhel-kvm. The host will continue being added to the management system after the install completes.
Information	6/21/18, 7:28 PM	PD-A000306 The PowerVC image transfer to the KVM p8-2142b2a-rhel-kvm has started. PowerVC install will begin after the transfer completes.
✓ Success	6/21/18, 7:28 PM	The request to add host **p8-2142b2a-rhel-kvm** has started. The host will be displayed after the process completes.

Figure 3-6 IBM PowerVC message center

7. Click the storage providers icon at the left to see the storage pool. By default, IBM PowerVC names the pool IBM Spectrum Scale Storage Pool. Click it to get detailed information, such as available capacity, total capacity, cluster members (the nodes), and volumes. You can select a host and view the disks that are part of the cluster, as seen by the node's operating system.

Figure 3-7 shows the IBM Spectrum Scale Storage Pool in detail, including state, health, available capacity, total capacity, and cluster members.



Storage Provider: IBM Spectrum Scale Storage Pool

Name:	IBM Spectrum Scale Storage Pool
State:	● Running
Health:	● OK
Available capacity:	2,587 GB
Total capacity:	2,643 GB
Type:	IBM Spectrum Scale
Cluster type:	Local storage

Storage Pools

Name	Capacity (GB)	Available (GB)
powervc_gpfs_data	2,643	2,587

Total: 1

Cluster Members

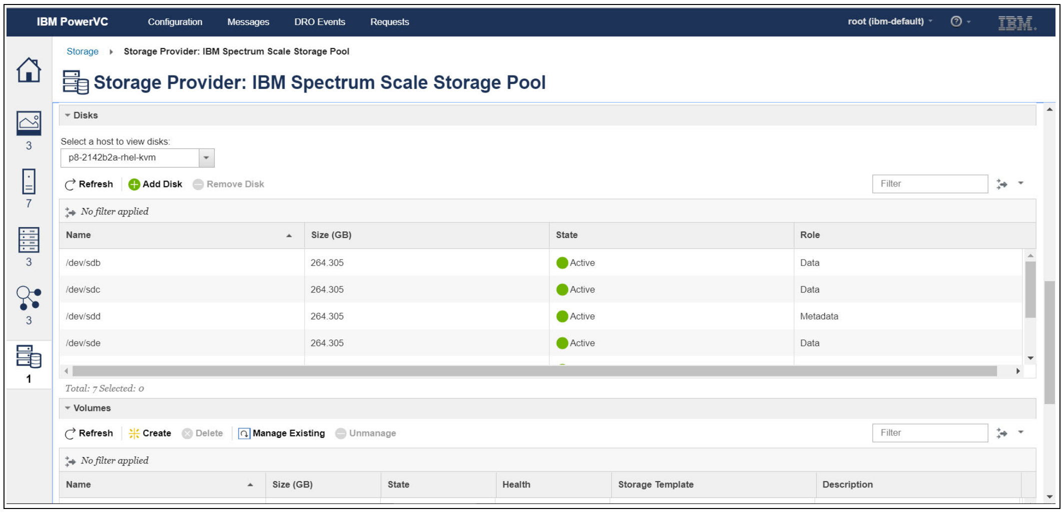
Refresh Filter

No filter applied

Name	Spectrum Scale Version	Spectrum Scale Available Version
p8-2142b2a-rhel-kvm	5.0.1.1	5.0.1.1
p8-214423w-neo	5.0.1.1	5.0.1.1

Figure 3-7 IBM Spectrum Scale Storage Pool in detail: Part one

Figure 3-8 shows the IBM Spectrum Scale Storage Pool in detail, including the node's local disks. Since IBM PowerVC version V1.4.1, it is possible to add and remove disks from a node after adding the host to the cluster by clicking **Add Disk** or **Remove Disk**. For more information about adding or removing disks, see IBM Knowledge Center.



Storage Provider: IBM Spectrum Scale Storage Pool

Disks

Select a host to view disks: p8-2142b2a-rhel-kvm

Refresh Add Disk Remove Disk Filter

No filter applied

Name	Size (GB)	State	Role
/dev/sdb	264.305	● Active	Data
/dev/sdc	264.305	● Active	Data
/dev/sdd	264.305	● Active	Metadata
/dev/sde	264.305	● Active	Data

Total: 7 Selected: 0

Volumes

Refresh Create Delete Manage Existing Unmanage Filter

No filter applied

Name	Size (GB)	State	Health	Storage Template	Description

Figure 3-8 BM Spectrum Scale Storage Pool in detail: Part two

8. Log in to the Red Hat Enterprise Linux KVM node and configure the Open vSwitch (OVS) bridge interface by attaching it to the second physical network port that is available. Configure the IP address and netmask, activate its Ethernet link, and start the OVS agent service now and during system start. Run the following commands:

```
ovs-vsctl add-port br-ex <interface name>
ifconfig br-ex <ip address> netmask <netmask> up
ovs-ofctl show br-ex
service neutron-openvswitch-agent restart
```

Example 3-2 shows an example of these commands.

Example 3-2 Configuring the Open vSwitch bridge interface for a storage node

```
ovs-vsctl add-port br-ex enP2p1s0f2
ifconfig br-ex 10.10.10.3 netmask 255.255.255.0 up
service neutron-openvswitch-agent enable
service neutron-openvswitch-agent restart
```

9. To ensure that this configuration persists, create an `ifcfg` file in the following directory:

   `/etc/sysconfig/network-scripts/icfg-<interface name>`

   When the operating system starts, all `ifcfg-*` files are picked up automatically by systemd.

   Example 3-3 shows an example of this process.

Example 3-3 Creating the ifcfg file for the br-ex interface and modifying the physical port interface for the Open vSwitch configuration

```
[root@p8-2142b2a-rhel-kvm network-scripts]# vi ifcfg-enP2p1s0f2
DEVICE=enP2p1s0f2
TYPE=OVSPort
DEVICETYPE=ovs
OVS_BRIDGE=br-ex
ONBOOT=yes
NM_CONTROLLED=no
BOOTPROTO=none
UUID=6068555c-34b1-4fe0-a373-e1b19e99a29b
:wq!

[root@p8-2142b2a-rhel-kvm network-scripts]# vi ifcfg-br-ex
DEVICE=br-ex
TYPE=OVSBridge
DEVICETYPE=ovs
ONBOOT=yes
NM_CONTROLLED=no
BOOTPROTO=none
IPADDR="10.10.10.3"
NETMASK="255.255.255.0"
:wq!
```

3.4 Removing a compute node from IBM PowerVC

Before removing a compute node, ensure that there are no VMs running on it. If you are supporting a production environment, then you can take advantage of the *migration* feature within IBM PowerVC and move the affected VM into a different node. For more information about how to perform the migration, see IBM Knowledge Center.

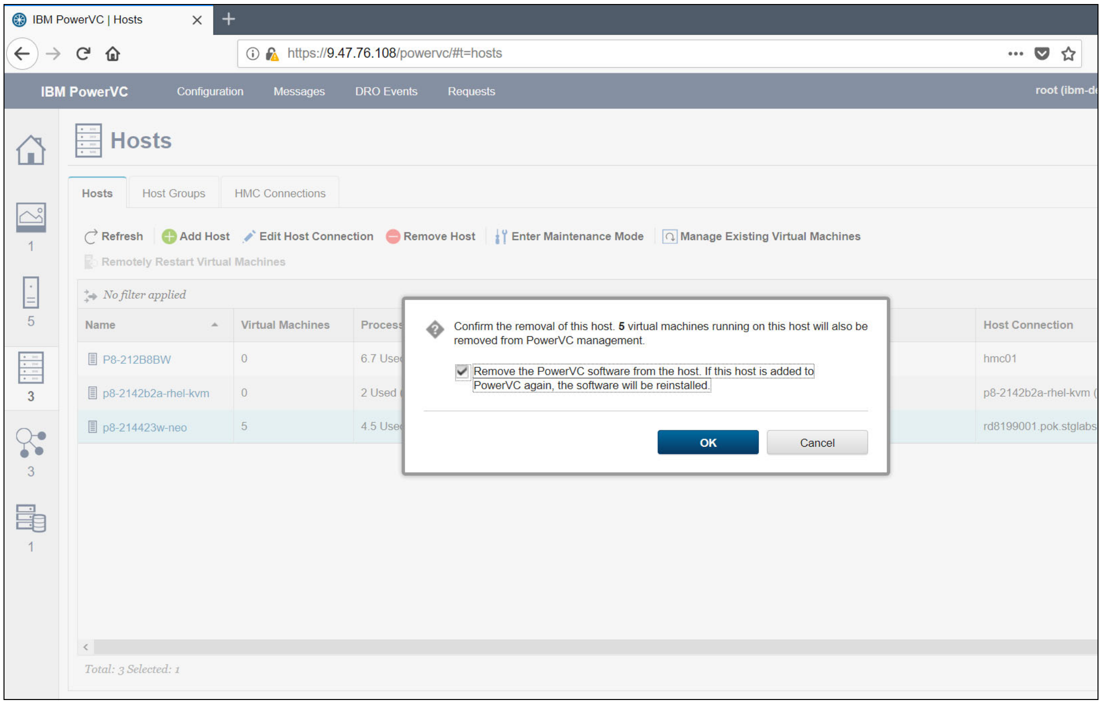
You should also ensure that the remaining storage in the IBM Spectrum Scale cluster has enough space for the remaining VMs. Additionally, you must verify that no processes are currently using the `/gpfs/powervc_gpfs` file system. If either of these conditions are not met, IBM PowerVC blocks the node removal.

When the last compute node is removed, the IBM Spectrum Scale storage is removed as well.

3.4.1 Removing the compute node

In IBM PowerVC, click the **Hosts** icon and select the node that you want to remove, and then click **Remove Node**. A confirmation dialog opens. Check **Remove PowerVC** and click **OK**. The host removal starts.

Figure 3-9 shows the IBM PowerVC Remove Host dialog.



Name	Virtual Machines	Process	Host Connection
P8-212B8BW	0	6.7 Use	hmc01
p8-2142b2a-rhel-kvm	0	2 Used	p8-2142b2a-rhel-kvm (
p8-214423w-neo	5	4.5 Use	rd8199001.pok.stglabs

Confirm the removal of this host. **5** virtual machines running on this host will also be removed from PowerVC management.

☑ Remove the PowerVC software from the host. If this host is added to PowerVC again, the software will be reinstalled.

Figure 3-9 IBM PowerVC Remove Host dialog

3.4.2 Monitoring the registration and removal of compute nodes

If you have any issue with the registration or removal of a compute node, or if you want to see how the tasks are progressing in detail, there are IBM PowerVC and IBM Spectrum Scale logs that you can examine.

To monitor the add node routine, you can view the addnode log in the IBM PowerVC server in the following directory:

```
/var/log/cinder/powervc_spectrum_scale_addnode.log
```

You can also monitor the nova API log in the IBM PowerVC server to view the API transactions details at the time of registration:

```
/var/log/nova/nova-api.log
```

To debug any issues that are related to IBM Spectrum Scale, or either IBM PowerVC or the compute node, look at the installation log files:

```
/var/log/cinder/powervc_spectrum_scale_install_*.log.
```

Specifically for the compute node, you can live monitor the Red Hat yum log to observe which specific packages and dependencies are installed at the time of registration:

```
/var/log/yum.log
```

Example 3-4 shows an example of the yum log.

Example 3-4 Open vSwitch, IBM PowerVC, and IBM Spectrum Scale packages installed on the compute node at the time of registration

```
[root@p8-2142b2a-rhel-kvm ~]# tail -f /var/log/yum.log
Jun 21 19:27:38 Installed: openvswitch-2.9.0-19.el7fdp.ppc64le
Jun 21 19:31:15 Installed: ibmpowervc-novalink-kvm-1.4.1.0-1.noarch
Jun 21 19:33:09 Installed: python-openvswitch-2.9.0-19.el7fdp.noarch
Jun 21 19:33:10 Installed: ibmpowervc-powervm-ras-1.4.1.0-1.noarch
Jun 21 19:33:17 Installed: ibmpowervc-powervm-oslo-1.4.1.0-1.noarch
Jun 21 19:33:17 Installed: ibmpowervc-powervm-monitor-1.4.1.0-1.noarch
Jun 21 19:33:28 Installed: ibmpowervc-powervm-block-1.4.1.0-1.noarch
Jun 21 19:33:29 Installed:
1:openstack-neutron-openvswitch-12.0.0-201805301537.ibm.el7.129.noarch
Jun 21 19:34:26 Installed: ibmpowervc-powervm-network-1.4.1.0-1.noarch
Jun 21 19:34:30 Installed: ibmpowervc-powervm-compute-1.4.1.0-1.noarch
Jun 21 19:34:30 Installed: ibmpowervc-powervm-1.4.1.0-1.noarch
[...]
Jun 21 19:38:34 Installed: gpfs.base-5.0.1-1.180605.104530.ppc64le
Jun 21 19:38:34 Installed: gpfs.ext-5.0.1-1.180605.104530.ppc64le
Jun 21 19:38:34 Installed: gpfs.crypto-5.0.1-1.180605.104530.ppc64le
Jun 21 19:38:34 Installed: gpfs.adv-5.0.1-1.180605.104530.ppc64le
Jun 21 19:38:34 Installed: gpfs.gpl-5.0.1-1.180605.104530.noarch
Jun 21 19:38:34 Installed: gpfs.license.dm-5.0.1-1.180605.104530.ppc64le
Jun 21 19:38:34 Installed: gpfs.docs-5.0.1-1.180605.104530.noarch
Jun 21 19:38:35 Installed: gpfs.gskit-8.0.50-86.ppc64le
Jun 21 19:38:35 Installed: gpfs.msg.en_US-5.0.1-1.180605.104530.noarch
Jun 21 19:38:53 Installed: ibmpowervc-powervm-gpfs-1.4.1.0-1.noarch
Jun 21 20:06:57 Installed: gpfs.base-5.0.1-1.180605.104530.ppc64le
Jun 21 20:06:57 Installed: gpfs.ext-5.0.1-1.180605.104530.ppc64le
Jun 21 20:06:57 Installed: gpfs.crypto-5.0.1-1.180605.104530.ppc64le
Jun 21 20:06:57 Installed: gpfs.adv-5.0.1-1.180605.104530.ppc64le
Jun 21 20:06:57 Installed: gpfs.gpl-5.0.1-1.180605.104530.noarch
Jun 21 20:06:57 Installed: gpfs.license.dm-5.0.1-1.180605.104530.ppc64le
[...]
```

3.5 Working with volumes and capturing of images

You can create, delete, and manage volumes in SDI with IBM PowerVC, which is the basis for creating images.

3.5.1 Creating an image from an installable ISO

Because you are creating a SDI environment from the beginning, you must build your first boot image so that you can later deploy your workloads.

Complete the following steps:

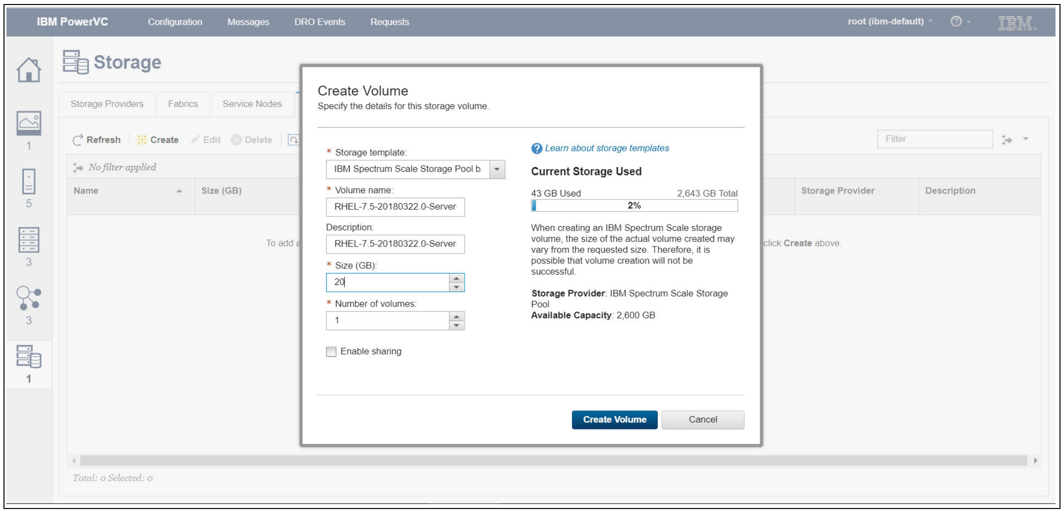
1. Download the required base ISO images from your preferred operating system:
   - For AIX and IBM i, use IBM Entitled Systems Support (ESS).
   - For Red Hat Enterprise Linux, use Red Hat Customer Portal.

Note: You must have a license or support agreement in place before downloading both the Red Hat and IBM operating system base ISO images.

2. Open IBM PowerVC, click the **Storage providers** icon on the left, select **Data Volumes**, select **Create**, and complete the Create volume dialog. Next, click **Create Volume**.

   Note: Ensure that the specified size is large enough to serve as the boot volume for the VM that is selected to deploy later. The value must support the total size of the ISO image, but also the size that you want your root volume group to have. For example, creating a volume for an AIX image with 100 GB as the specified size results in a rootvg of a 100 GB after the system starts.

   Figure 3-10 shows the IBM PowerVC Create Volume dialog.



Create Volume
Specify the details for this storage volume.

* Storage template: IBM Spectrum Scale Storage Pool b
* Volume name: RHEL-7.5-20180322.0-Server
Description: RHEL-7.5-20180322.0-Server
* Size (GB): 20
* Number of volumes: 1
□ Enable sharing

Current Storage Used
43 GB Used 2,643 GB Total 2%

When creating an IBM Spectrum Scale storage volume, the size of the actual volume created may vary from the requested size. Therefore, it is possible that volume creation will not be successful.

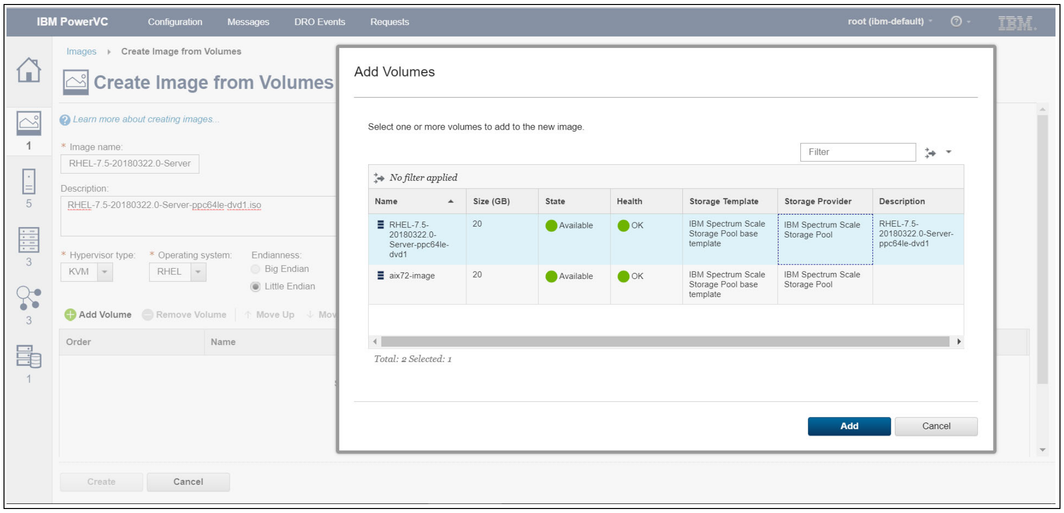
Storage Provider: IBM Spectrum Scale Storage Pool
Available Capacity: 2,600 GB

Figure 3-10 IBM PowerVC Create Volume dialog

3. Click the **Images** icon at the left, select **Create**, and provide the required information. Ensure that you specify the operating system type and endianness (if applicable) that matches the ISO that you use for the installation. In this case, Red Hat Enterprise Linux 7.5 under KVM is deployed, so the only option is ppc64le.

   Select **Add Volume**, select the volume that is used in step 2, and mark the volume as part of the boot set. Click **Create** to finish the image creation.

Figure 3-11 shows the IBM PowerVC Add Volume to the image dialog.



Add Volumes

Select one or more volumes to add to the new image.

Filter

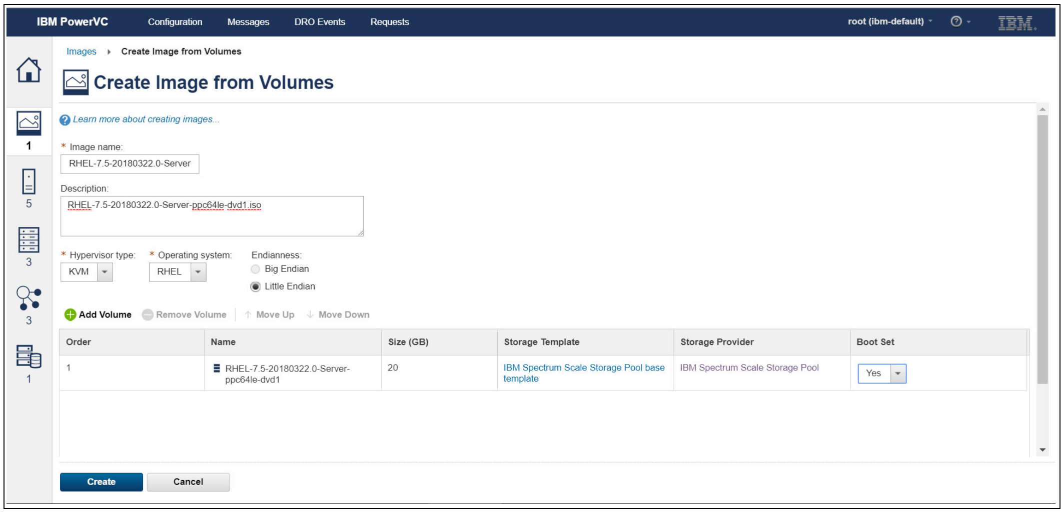
No filter applied

Name	Size (GB)	State	Health	Storage Template	Storage Provider	Description
RHEL-7.5-20180322.0-Server-ppc64le-dvd1	20	● Available	● OK	IBM Spectrum Scale Storage Pool base template	IBM Spectrum Scale Storage Pool	RHEL-7.5-20180322.0-Server-ppc64le-dvd1
aix72-image	20	● Available	● OK	IBM Spectrum Scale Storage Pool base template	IBM Spectrum Scale Storage Pool	

Total: 2 Selected: 1

Figure 3-11 IBM PowerVC Add Volume to the Image dialog

Figure 3-12 shows the IBM PowerVC image creation window after the volume is attached to the image.



Create Image from Volumes

Learn more about creating images...

* Image name: RHEL-7.5-20180322.0-Server

Description: RHEL-7.5-20180322.0-Server-ppc64le-dvd1.iso

* Hypervisor type: KVM

* Operating system: RHEL

Endianness:
○ Big Endian
● Little Endian

Order	Name	Size (GB)	Storage Template	Storage Provider	Boot Set
1	RHEL-7.5-20180322.0-Server-ppc64le-dvd1	20	IBM Spectrum Scale Storage Pool base template	IBM Spectrum Scale Storage Pool	Yes

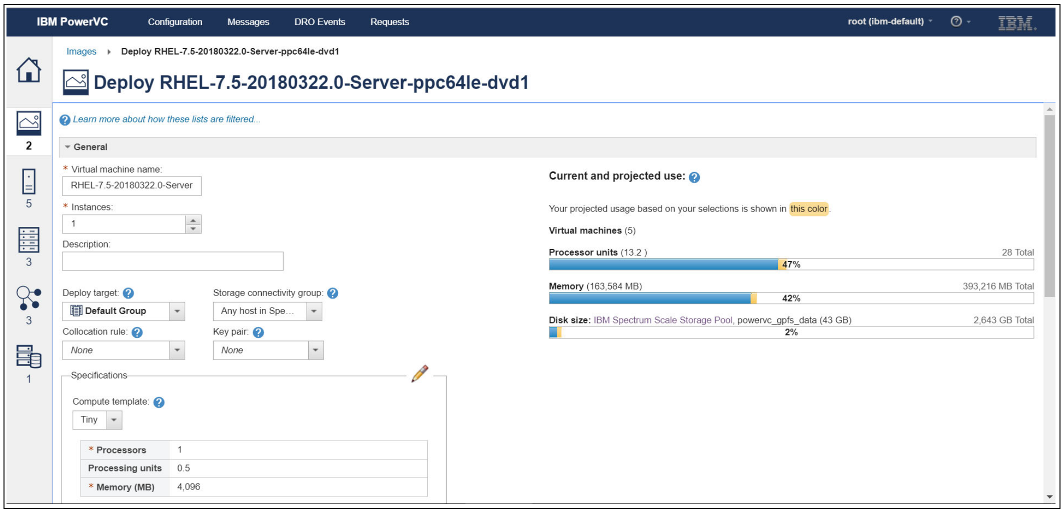
Figure 3-12 IBM PowerVC image creation window

4. Deploy the image. As seen in Figure 3-13 and Figure 3-14, the deployment dialog sets the storage connectivity group to the default because there is only one storage connectivity group in an SDS environment. Add an appropriate network, change any other settings that are wanted, and click **Deploy**.

   What happens behind the scenes during a deployment is that IBM Spectrum Scale creates a linked clone of the original file. A virtual SCSI connection is created between the VM and that file.

   Operations such as live migration, accessing the VM console, and VM resizing are supported in this environment as well.

   Figure 3-13 shows part one of the IBM PowerVC image deployment window, including the size and the destination of the new VM.



Images ▸ Deploy RHEL-7.5-20180322.0-Server-ppc64le-dvd1

Deploy RHEL-7.5-20180322.0-Server-ppc64le-dvd1

Learn more about how these lists are filtered...

General

* Virtual machine name: RHEL-7.5-20180322.0-Server

* Instances: 1

Description:

Deploy target: Default Group

Storage connectivity group: Any host in Spe...

Collocation rule: None

Key pair: None

Specifications

Compute template: Tiny

* Processors	1
Processing units	0.5
* Memory (MB)	4,096

Current and projected use:

Your projected usage based on your selections is shown in this color.

Virtual machines (5)

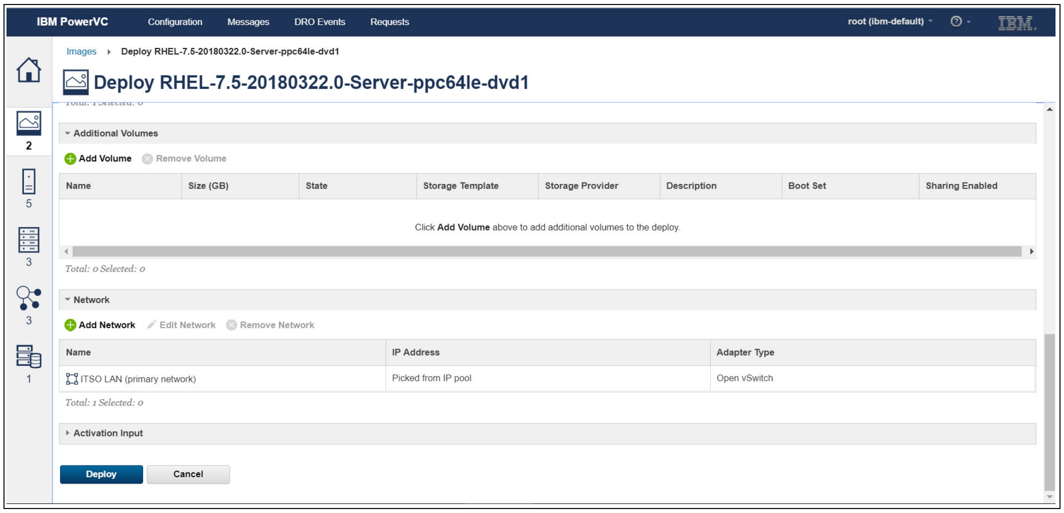
Processor units (13.2) 47% 28 Total

Memory (163,584 MB) 42% 393,216 MB Total

Disk size: IBM Spectrum Scale Storage Pool, powervc_gpfs_data (43 GB) 2% 2,643 GB Total

Figure 3-13 IBM PowerVC image deployment: Part one

Figure 3-14 shows part two of the IBM PowerVC image deployment window.



Images ▸ Deploy RHEL-7.5-20180322.0-Server-ppc64le-dvd1

Deploy RHEL-7.5-20180322.0-Server-ppc64le-dvd1

Additional Volumes

Name	Size (GB)	State	Storage Template	Storage Provider	Description	Boot Set	Sharing Enabled

Click **Add Volume** above to add additional volumes to the deploy.

Total: 0 Selected: 0

Network

Name	IP Address	Adapter Type
ITSO LAN (primary network)	Picked from IP pool	Open vSwitch

Total: 1 Selected: 0

Activation Input

Figure 3-14 IBM PowerVC image deployment: Part two

5. After the deployment completes and the VM shows the **Active** status, stop the VM. The VM does not start because the boot disk has no data on it yet.

6. Log in to the Red Hat Enterprise Linux KVM storage node and upload a copy of the installation ISO that you obtained in step 1 on page 73 into the virtual optical media repository by running the following command:

   ```
   pvmctl vom upload --vios NovaLink --data <full path to iso file> --name <a name
   for the iso file>
   ```

   Example 3-5 shows the command syntax that is used to upload the virtual optical media.

   Example 3-5 Command syntax that is used to upload the virtual optical media

   ```
   # pvmctl vom upload --vios NovaLink --data
   /gpfs/powervc_gpfs/infrastructure/RHEL-7.5-20180322.0-Server-ppc64le-dvd1.iso
   --name rhel75
   ```

   > **Note:** When running **`vom upload`**, ensure that the ISO file has world-readable permissions (777). Also, the name that is given to the ISO file must not contain any special characters such as "." or "-".

7. List and identify the name of the VM that was deployed in step 4 on page 76, as seen by the KVM host, by running the following command:

   ```
   pvmctl LogicalPartition list
   ```

   Example 3-6 shows the command that is used to list the VMs that are running on the storage node.

   Example 3-6 Command that is used to list the VMs that are running on the storage node

   ```
   [root@p8-2142b2a-rhel-kvm ~]# pvmctl LogicalPartition list
   Logical Partitions
   +--------------------------------+----+---------------+----------+-----------+
   |              Name              | ID |     State     |   RMC    |    Env    |
   +--------------------------------+----+---------------+----------+-----------+
   | RHEL-7.5-2018-55a7e82b-0000000b | 2  | not activated | inactive | AIX/Linux |
   +--------------------------------+----+---------------+----------+-----------+
   ```

8. Run **`pvmctl scsi create`** to attach the newly uploaded virtual optical repository to the VM:

   ```
   pvmctl scsi create --type vopt --lpar name=<your VM name> --stor-id <name of
   the vom you used in previous step> -p name=NovaLink
   ```

   Example 3-7 shows the command that is used to attach the ISO image into the VM.

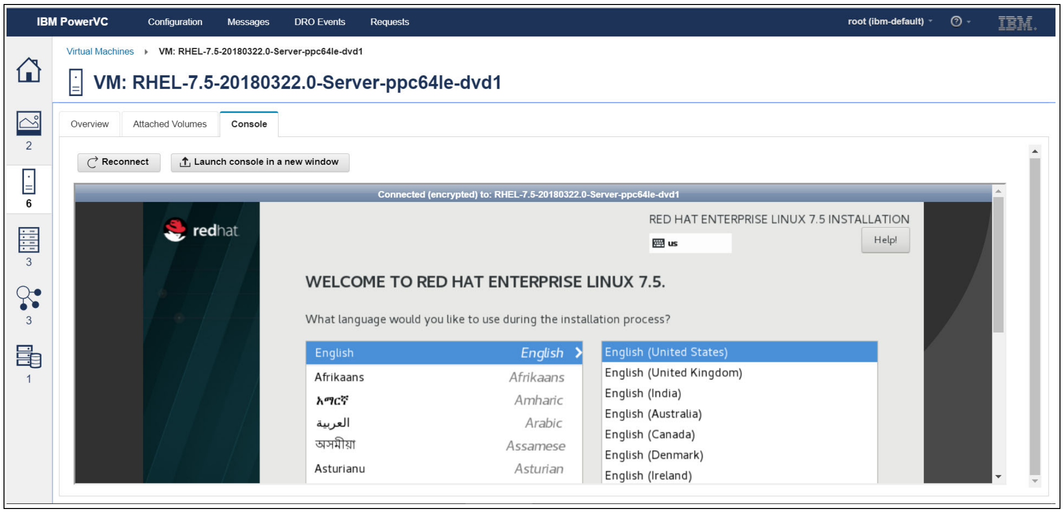
   Example 3-7 Command that is used to attach the ISO image into the VM

   ```
   [root@p8-2142b2a-rhel-kvm ~]# pvmctl scsi create --type vopt --lpar
   name=RHEL-7.5-2018-55a7e82b-0000000b --stor-id rhel75 -p name=NovaLink
   ```

9. Restart the VM.

10. Start the virtual console by going to the VM details page and selecting the **Console** tab. You should see how the system boots into the Red Hat Enterprise Linux installation wizard. Define your preferred settings and begin the installation.

Figure 3-15 shows the VM’s console, which displays the Red Hat Enterprise Linux installation wizard after the ISO image is attached.



VM: RHEL-7.5-20180322.0-Server-ppc64le-dvd1

Connected (encrypted) to: RHEL-7.5-20180322.0-Server-ppc64le-dvd1

RED HAT ENTERPRISE LINUX 7.5 INSTALLATION

WELCOME TO RED HAT ENTERPRISE LINUX 7.5.

What language would you like to use during the installation process?

English	English	English (United States)
Afrikaans	Afrikaans	English (United Kingdom)
አማርኛ	Amharic	English (India)
العربية	Arabic	English (Australia)
অসমীয়া	Assamese	English (Canada)
Asturianu	Asturian	English (Denmark)
		English (Ireland)

Figure 3-15 Red Hat Enterprise Linux installation wizard as seen by the VM’s console

11. When installation is finished and the operating system is started, install the Resource Monitoring and Control (RMC) packages on the VM.

    Log in to the VM by using the root user and the password that was defined during installation and run the following commands to download and install the required packages, in this specific order:

```
wget
http://public.dhe.ibm.com/software/server/POWER/Linux/yum/OSS/RHEL/7/ppc64le/rs
ct.core-3.2.3.2-18040.ppc64le.rpm
wget
http://public.dhe.ibm.com/software/server/POWER/Linux/yum/OSS/RHEL/7/ppc64le/rs
ct.core.utils-3.2.3.2-18040.ppc64le.rpm
wget
http://public.dhe.ibm.com/software/server/POWER/Linux/yum/OSS/RHEL/7/ppc64le/sr
c-3.2.3.2-18040.ppc64le.rpm

yum -y install librtas*
yum -y localinstall src-3.2.3.2-18040.ppc64le.rpm
yum -y localinstall rsct.core.utils-3.2.3.2-18040.ppc64le.rpm
yum -y localinstall rsct.core-3.2.3.2-18040.ppc64le.rpm
```

    For more information regarding RMC installation in Linux, see IBM Knowledge Center.

12. The Cloud-init utility must be installed and configured on the VM so that basic configuration routines, such as setting up the network interface and defining a customized host name, can be performed automatically at the time of deployment. For more information about Cloud-init, see the Cloud-init documentation.

    a. Follow the steps that are described at the end of 2.6, “Installing Red Hat Enterprise Linux KVM on the IBM Power System server” on page 50 to enable all the Red Hat Enterprise Linux ppc64le repositories, which are needed to obtain the required package dependencies.

b. If Cloud-init is installed, remove it and its dependencies by running the following commands:

```
yum -y install yum-plugin-remove-with-leaves
yum -y remove cloud-init --remove-leaves
yum autoremove
```

c. Copy the IBM PowerVC Cloud-init version into the VM. From the IBM PowerVC server, copy the required RPM files by running the following command:

```
scp
/opt/ibm/powervc/images/cloud-init/rhel/cloud-init-0.7.4-10.el7.noarch.rpm
root@9.47.76.99:/root/
```

d. Download and install the Extra Packages for Enterprise Linux (EPEL) repositories to provide the required dependencies during the Cloud-init installation:

```
wget
http://dl.fedoraproject.org/pub/epel/7Server/ppc64/Packages/e/epel-release-7
-11.noarch.rpm
rpm -Uvh epel-release-7*.rpm
```

e. Install the IBM PowerVC cloud-init package locally and use the yum repositories to search, identify, download, and install any required dependencies. Run the following command:

```
yum localinstall cloud-init-0.7.4-10.el7.noarch.rpm
```

f. Modify the `cloud.cfg` file in /etc/cloud/cloud.cfg. Change these values if they exist, as shown below:

```
disable_root:      0

ssh_pwauth:        1

ssh_deletekeys:    1
```

g. Add these values after `ssh_deletekeys`:

```
disable_ec2_metadata:true

datasource_list:    ['ConfigDrive']
```

h. Disable selinux now and during system startup. Run the following commands:

```
setenforce permissive
sed -i 's/enforcing/disabled/g' /etc/selinux/config
sestatus
```

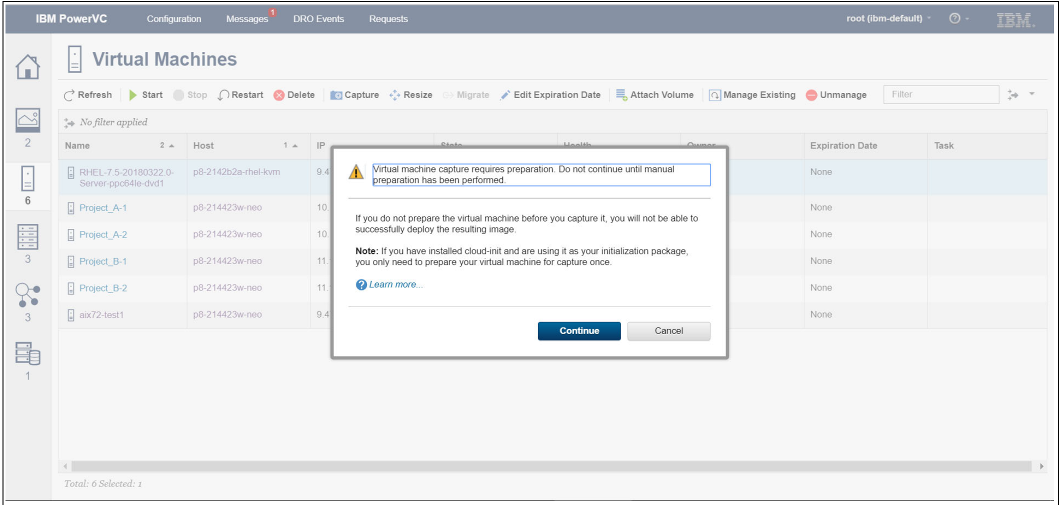
i. Set all network interface configuration files to start automatically during system startup and disable the network manager controller. Run the following commands:

```
echo "NM_CONTROLLED=no" >> /etc/sysconfig/network-scripts/ifcfg-eth*
echo "NM_CONTROLLED=no" >> /etc/sysconfig/network-scripts/ifcfg-en*
sed -i 's/ONBOOT=no/ONBOOT=yes/g' /etc/sysconfig/network-scripts/ifcfg-eth*
sed -i 's/ONBOOT=no/ONBOOT=yes/g' /etc/sysconfig/network-scripts/ifcfg-en*
```

For more information and guidance about how to configure cloud-init on Red Hat Enterprise Linux for IBM Power System Servers, see IBM Knowledge Center.

13. The image is ready for capture. Stop the VM and initiate the capture process. When prompted, click **Continue**.

Figure 3-16 shows the IBM PowerVC capture dialog.



Virtual machine capture requires preparation. Do not continue until manual preparation has been performed.

If you do not prepare the virtual machine before you capture it, you will not be able to successfully deploy the resulting image.

Note: If you have installed cloud-init and are using it as your initialization package, you only need to prepare your virtual machine for capture once.

Figure 3-16 IBM PowerVC capture dialog

14. Now you have a deployable image that can automatically activate and configure itself. Click the **Images** icon at the left and deploy a new VM.

3.5.2 Creating an image with an existing raw image

If you have a raw image (not OVS) that should be used in your IBM PowerVC SDI environment, it is easy to import it. A raw image can be created by using the command **dd** to copy the content of the disk into a file, which is then used for import. It is also possible to use several files for a multivolume image.

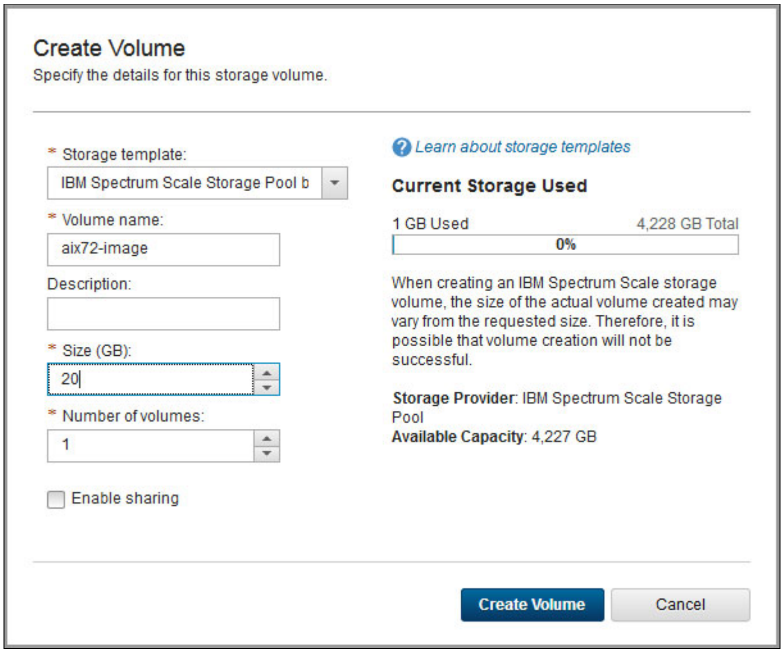
AIX raw images that are prepared by IBM

IBM provides prepared raw images for AIX 7.1 and AIX 7.2. These images contain cloud-init and can be used for deployment without any further changes. A customer with a software maintenance agreement (SWMA) may download these images from the ESS website.

Importing a raw image into an IBM PowerVC SDI environment

To import a prepared AIX image into IBM PowerVC, complete the following steps:

1. Create a data volume in the storage section of IBM PowerVC with the size of the raw image, as shown in Figure 3-17.



Create Volume

Specify the details for this storage volume.

* Storage template: IBM Spectrum Scale Storage Pool b

* Volume name: aix72-image

Description:

* Size (GB): 20

* Number of volumes: 1

□ Enable sharing

Learn about storage templates

Current Storage Used

1 GB Used 4,228 GB Total

0%

When creating an IBM Spectrum Scale storage volume, the size of the actual volume created may vary from the requested size. Therefore, it is possible that volume creation will not be successful.

Storage Provider: IBM Spectrum Scale Storage Pool
Available Capacity: 4,227 GB

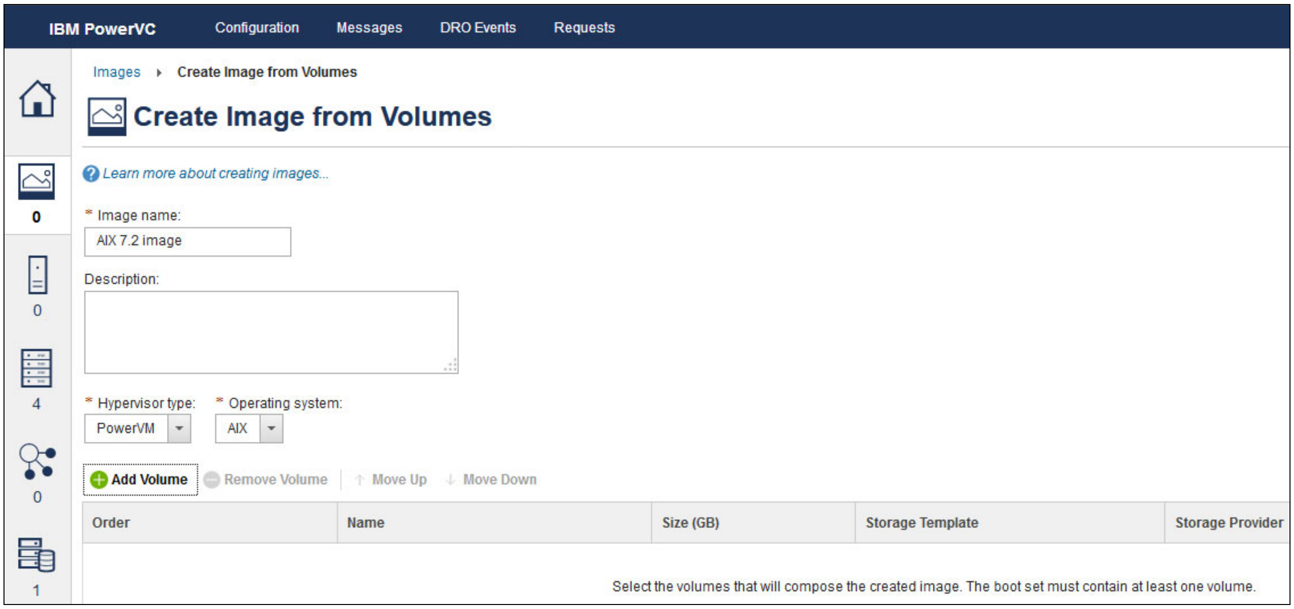
Figure 3-17 Creating a volume for the image

2. Enter any cluster node or the IBM PowerVC server and copy the image into the IBM Spectrum Scale cluster, as shown in Figure 3-18.

```
$ (sudo) su -

# cd /gpfs/powervc_gpfs/infrastructure

# scp <source>:../AIX_v7.2_Virt_Mach_Image_7200-02-00_102017.gz .
# gunzip AIX_v7.2_Virt_Mach_Image_7200-02-00_102017.gz
# ls -al
total 62915330
drwxr-xr-x 2 root root        4096 Jun 23 15:31 .
drwxr-xr-x 3 root root      262144 Jun 22 18:40 ..
-rw-r--r-- 1 root root 21474836480 Jun 23 15:38
AIX_v7.2_Virt_Mach_Image_7200-02-00_1020
-rw-rw---- 1 root root 21474836480 Jun 23 15:28 volume-aix72-image-ccec961b-a88c

# dd if=AIX_v7.2_Virt_Mach_Image_7200-02-00_1020 of=volume-aix72-image-ccec961b-a88c
# rm AIX_v7.2_Virt_Mach_Image_7200-02-00_1020
```

Figure 3-18 Running dd to add a raw image file into a volume

3. Create the image by using the prepared volume. In the images section of IBM PowerVC, click **Create** and enter the name of the image (for AIX, select **PowerVM and AIX**, as shown in Figure 3-19).



Create Image from Volumes

Learn more about creating images...

* Image name: AIX 7.2 image

Description:

* Hypervisor type: PowerVM

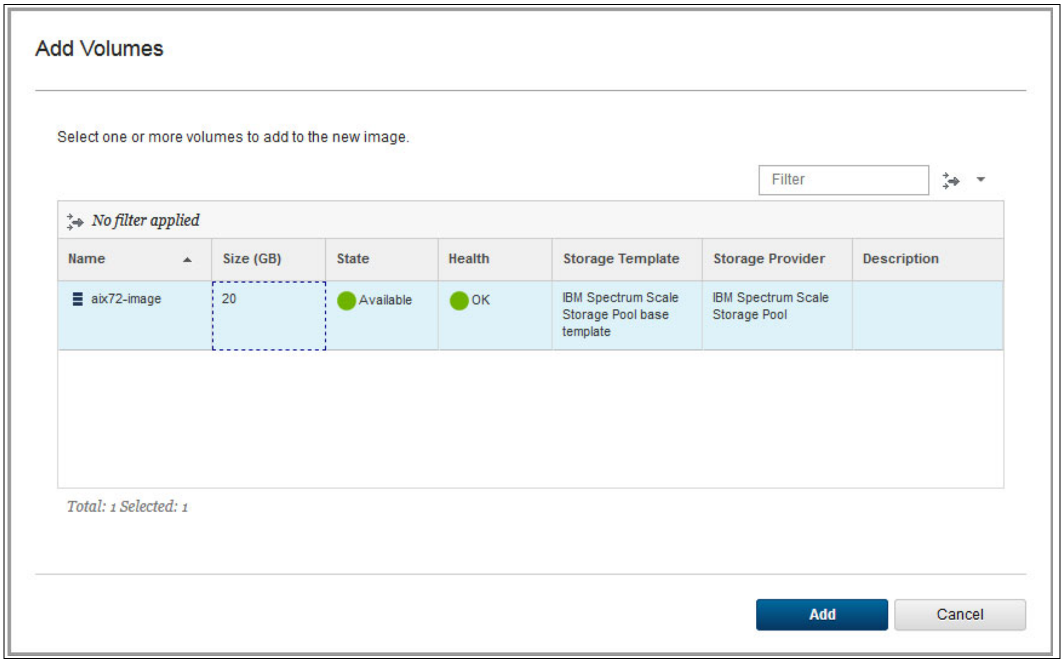
* Operating system: AIX

Order	Name	Size (GB)	Storage Template	Storage Provider

Select the volumes that will compose the created image. The boot set must contain at least one volume.

Figure 3-19 Creating the image

4. Add all the volumes of the image. In this example, select the aix72 image volume, as shown in Figure 3-20.



Add Volumes

Select one or more volumes to add to the new image.

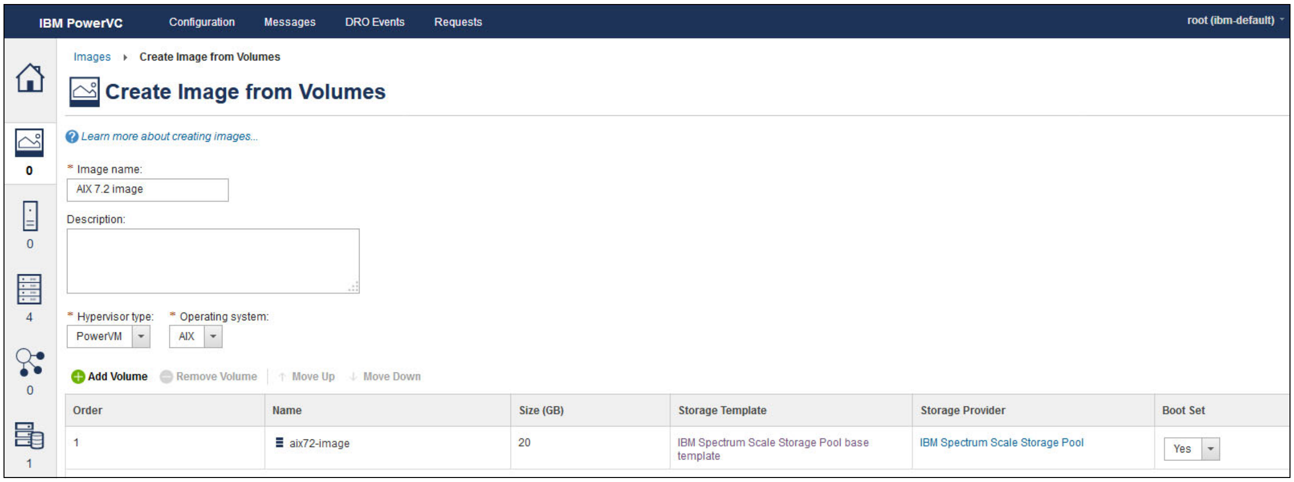
No filter applied

Name	Size (GB)	State	Health	Storage Template	Storage Provider	Description
aix72-image	20	Available	OK	IBM Spectrum Scale Storage Pool base template	IBM Spectrum Scale Storage Pool	

Total: 1 Selected: 1

Figure 3-20 Selecting the volume

5. Change the **Boot Set** flag on the right of the volume to **Yes**, as shown in Figure 3-21.



Create Image from Volumes

Learn more about creating images...

* Image name: AIX 7.2 image

Description:

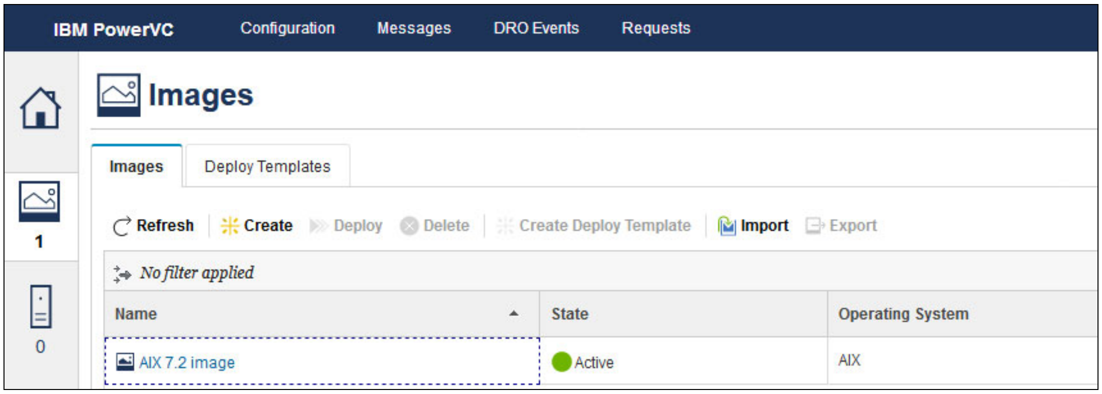
* Hypervisor type: PowerVM

* Operating system: AIX

Order	Name	Size (GB)	Storage Template	Storage Provider	Boot Set
1	aix72-image	20	IBM Spectrum Scale Storage Pool base template	IBM Spectrum Scale Storage Pool	Yes

Figure 3-21 Switch Boot Set to Yes

The image appears in the Images window, as shown in Figure 3-22. In the storage section, you can see that the volume moved from Data Volumes to Boot Volumes.



Images

No filter applied

Name	State	Operating System
AIX 7.2 image	Active	AIX

Figure 3-22 First image is ready

4

Software-defined networking

This chapter describes the usage of software-defined networking (SDN) within an IBM PowerVC software-defined infrastructure (SDI) solution.

4.1 Software-defined networking with IBM PowerVC

In an IBM PowerVC SDI solution, the SDN is provided by Open vSwitch (OVS) technology. OVS is an open source multilayer virtual switch with tremendous possibilities. It is part of many Linux distributions, such as Ubuntu or Red Hat Enterprise Linux.

For more information about OVS, see Open vSwitch.

In IBM PowerVC, SDN may be used without software-defined storage (SDS), as described in Chapter 3, “Implementing software-defined storage in IBM PowerVC” on page 63. SDN has the same requirement as SDS, that is, it runs in a virtual machine (VM) with PowerVM NovaLink technology, but it is possible to also use that technology in a classic environment with Virtual I/O Servers (VIOSes) (with PowerVM NovaLink as the control instance). In that case, specify `pvm-installmode=SDN` (instead of `SDE`) in `grub.cfg` when installing PowerVM NovaLink. For more information about installing PowerVM NovaLink, see 2.2.1, “Preparing the resources for a PowerVM NovaLink network installation” on page 22.

4.1.1 Optional usage of Virtual Extensible LANs

In addition to traditional network concepts, such as a flat LAN or virtual local area network (VLAN), OVS can work with Virtual Extensible LANs (VXLANs). Using VXLANs is optional and might not be needed in many cases if VLANs or a flat network is sufficient for your environment.

A VXLAN is basically an encapsulation protocol running on top of a typical Layer 3 network infrastructure. A VXLAN package is 1450 bytes in size and is encapsulated in a standard 1500 bytes (MTU size) packet on the network infrastructure. VXLANs are private, so only members of a VXLAN may communicate with each other.

Figure 4-1 shows a VXLAN environment, where two VMs (1.n, 2.n, and 3.n in pairs) can communicate with each other. VMs in different VXLANs may use the same IP addresses and not interfere with each other. The OVS bridges in PowerVM NovaLink have all the routing rules to ensure that the packets are delivered to correct place.

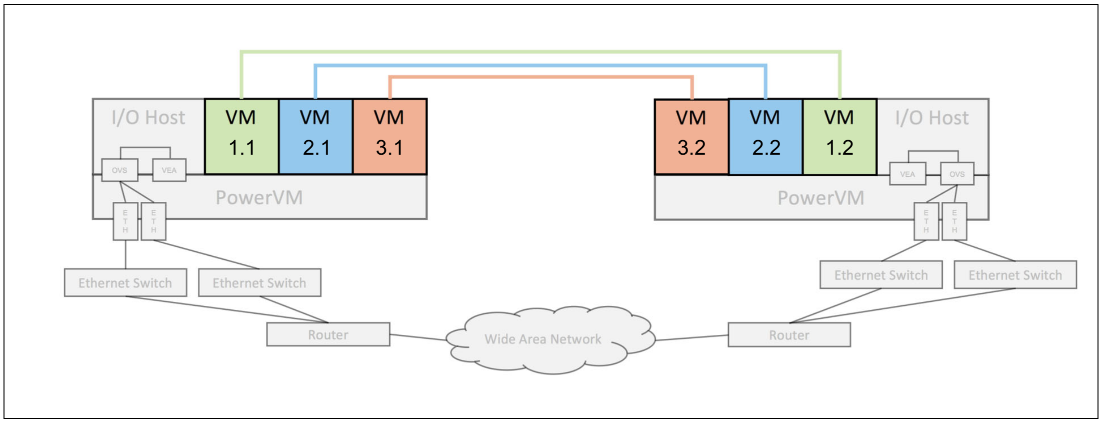


Figure 4-1 Typical VXLAN environment

Usually, VXLANs are not enough because some of the VMs need access to the wide area network (WAN). In this case, a network node is required. A network node is basically a network address translation (NAT) router.

> **Note:** With IBM PowerVC, you can create VXLANs *only* if a network node is present in the cluster. In theory, it is possible to use only internal VXLANs without access to the outside world, but this is not practical in a production environment, so this configuration is prohibited by IBM PowerVC.

Figure 4-2 shows an environment with VXLANs and a network node.

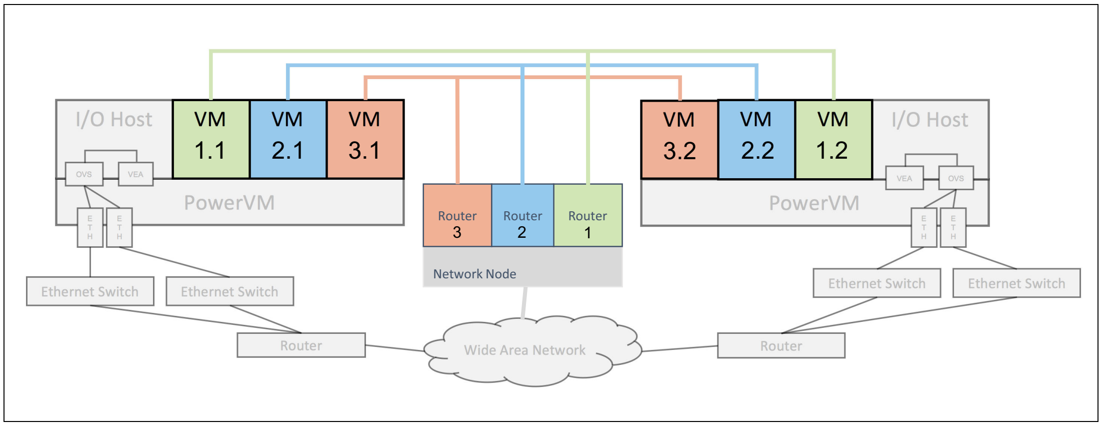


Figure 4-2 VXLAN environment with a network node

4.2 Network nodes

The following sections describe the architecture of network nodes and how they might be implemented.

4.2.1 The architecture of network nodes

Network nodes act as the bridge between your overlay network (for example, VXLAN) and your WAN. If you do not plan to take advantage of overlay networking, then you do not need a network node. Instead, your compute nodes should be plugged directly into the WAN.

A basic topology that uses network nodes is shown in Figure 4-3.

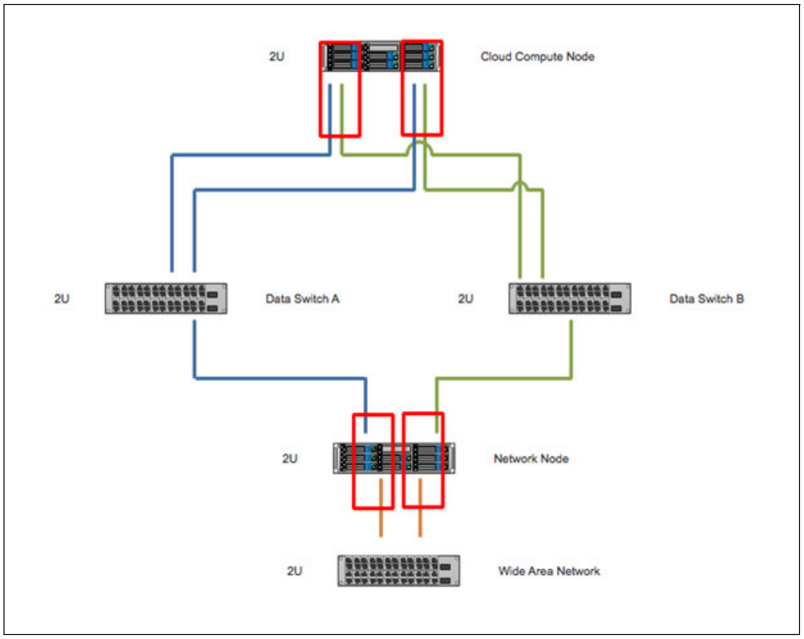


Figure 4-3 Network node topology

As you can see, the network nodes connect to two physical networks, although it can be one physical network with logical segmentations. Here are some considerations:

- The cloud compute node connects to the network node through its `tunneling-ip`, which is generally the IP address on br-ex unless the administrator updated it.
- The network node connects to the WAN through its local br-ex.
- The network node connects to the cloud compute node through its `tunneling-ip`, which is generally the IP address on br-ex, unless the administrator updated it.

The network node is an Ubuntu 16.04 LTS server (either Power System server or x86 server). It should have sufficient bandwidth (10 GbE or higher) and the server must be able to access it by using SSH. Multiple network nodes can be created for redundancy and load-balancing.

The basic connectivity requirements are that an IP address must be on a port (or bridge) that can communicate with the cloud compute nodes. Then, an OVS bridge that is named br-ex must have ports that are connected to the WAN.

For more information about network nodes, see IBM Knowledge Center: IBM PowerVC, which also includes information about how to update the `tunneling-ip`.

The network nodes also must have firewall ports open to communicate properly. For more information about these ports, see IBM Knowledge Center: IBM PowerVC.

4.2.2 Installation of a network node

To add a network node to your IBM PowerVC SDI environment, complete the following steps:

1. Install an Ubuntu 16.04 system and update it with the latest fixes. In our lab environment, we defined padmin as a user for Ubuntu to make it consistent with the PowerVM NovaLink compute nodes. The Ubuntu system may be installed in a VM in a Power System server or in an x86-based system. The VM must have a physical Ethernet adapter; a virtual adapter does not work.
2. Install the following extra packages:
   - `ssh`
   - `openvswitch-common`
   - `openvswitch-switch`
3. With the OVS packages installed, you can define the br-ex bridge, which is a requirement for the network node, as shown in Figure 4-4.

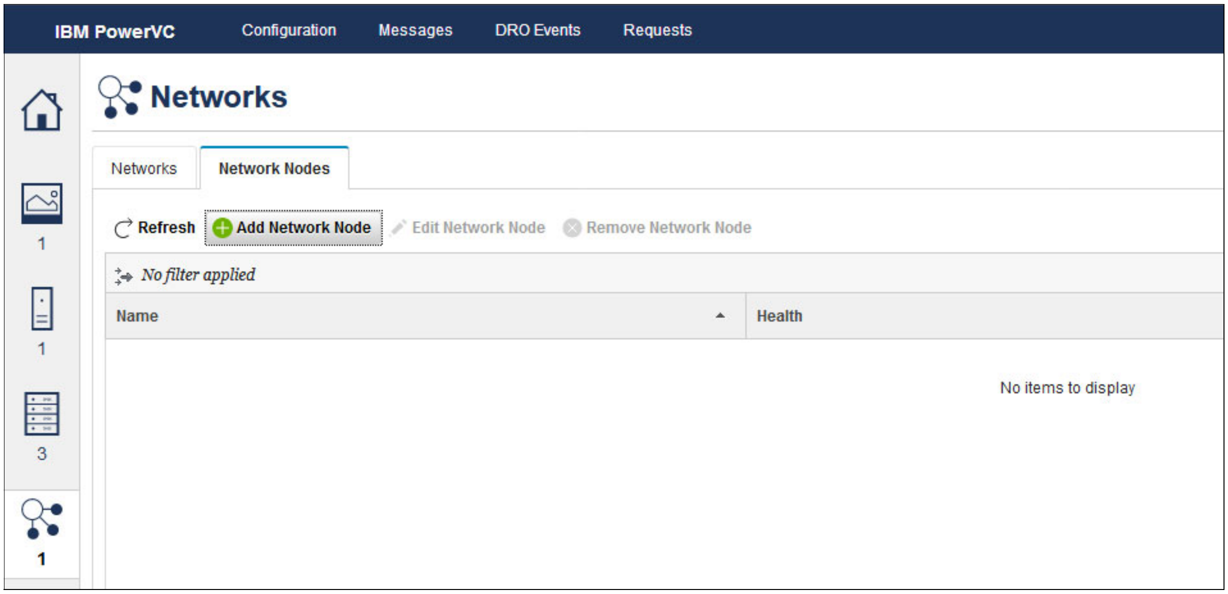
```
$ sudo vi /etc/network/interfaces

# This file describes the network interfaces available on your system
# and how to activate them. For more information, see interfaces(5).

source /etc/network/interfaces.d/*

# The loopback network interface
auto lo
iface lo inet loopback

auto br-ex
allow-ovs br-ex
iface br-ex inet static
      address 9.47.76.117
      netmask 255.255.240.0
      gateway 9.47.79.254
      dns-nameservers 9.12.16.2
      dns-search pok.stglabs.ibm.com
      ovs_type OVSBridge
      ovs_ports enP30p128s0f0

allow-br-ex enP30p128s0f0
iface enP30p128s0f0 inet manual
    ovs_bridge br-ex
    ovs_type OVSPort
```

Figure 4-4 The br-ex configuration of a network node

In our lab, we connected the bridge to one port of an adapter. It is possible to create more sophisticated configurations with port bonds.

> **Note:** To verify the bridge configuration, run **`ovs-vsctl show`** to show the configuration of the bridge.
>
> **Hint:** If you do too many changes and the configuration in memory is incorrect, a restart might help.

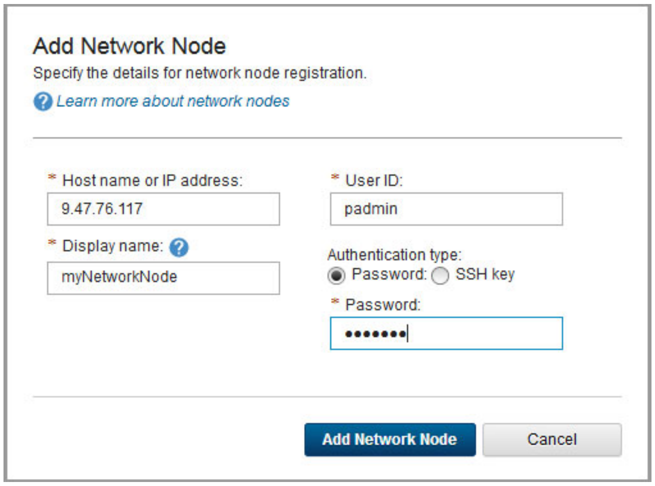
4. The node is ready to be added into IBM PowerVC. In IBM PowerVC, open the Networks window and select the **Network Nodes** tab, as shown in Figure 4-5.



No items to display

Figure 4-5 Add Network Node

5. Click **Add Network Node**. In the window that opens, enter the IP address and the credentials for accessing the prepared network node, as shown in Figure 4-6.


Add Network Node

Specify the details for network node registration.

Learn more about network nodes

* Host name or IP address: 9.47.76.117

* Display name: myNetworkNode

* User ID: padmin

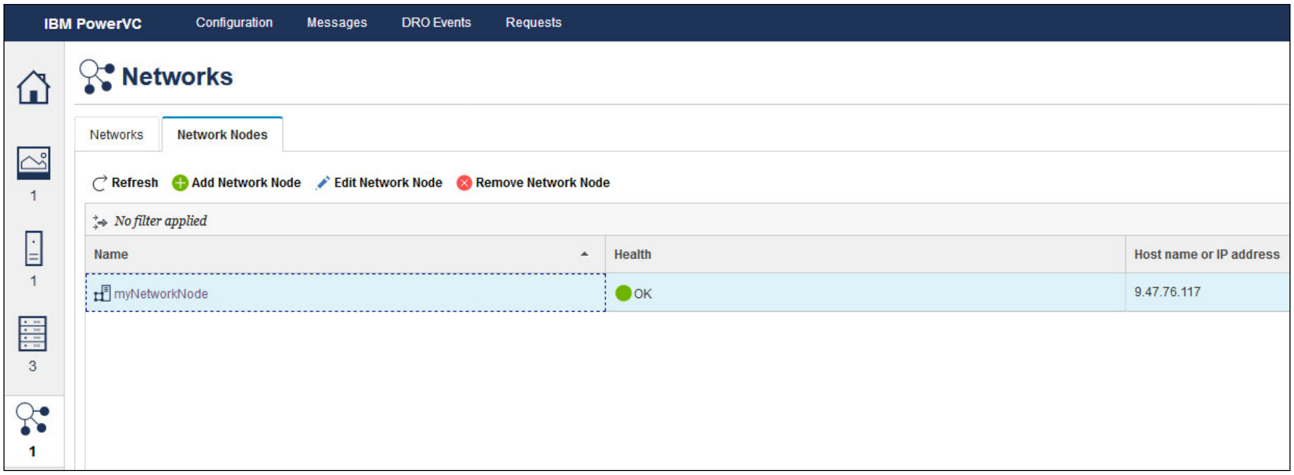
Authentication type: ● Password: ○ SSH key

* Password: •••••••

Figure 4-6 Add Network Node details

In our scenario, we used our created padmin user for access because root access is not necessary.

After some time, the network node appears, as shown in Figure 4-7.



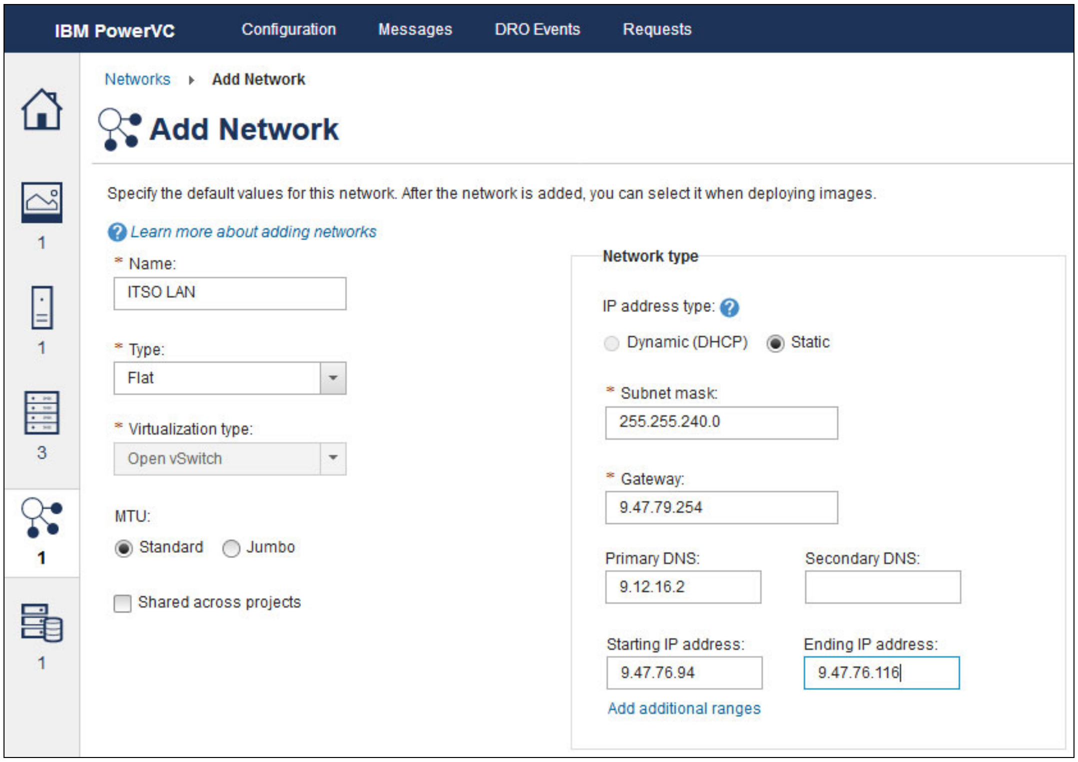
Name	Health	Host name or IP address
myNetworkNode	OK	9.47.76.117

Figure 4-7 Network node is available

4.3 The configuration and usage of Virtual Extensible LANs

Before you can define a VXLAN, you must set up a network node, as described in 4.2.2, “Installation of a network node” on page 89. Then, you must define a LAN with external access, which may be either a flat network or a VLAN. In our lab, we used a flat network with a range of available IPs that may be used for direct deployments or for accessing VMs with VXLANs from the WAN.

Figure 4-8 shows our lab definition.



Specify the default values for this network. After the network is added, you can select it when deploying images.

* Name: ITSO LAN

* Type: Flat

* Virtualization type: Open vSwitch

MTU: ● Standard ○ Jumbo

□ Shared across projects

Network type

IP address type: ○ Dynamic (DHCP) ● Static

* Subnet mask: 255.255.240.0

* Gateway: 9.47.79.254

Primary DNS: 9.12.16.2

Secondary DNS:

Starting IP address: 9.47.76.94

Ending IP address: 9.47.76.116

Figure 4-8 Definition of a flat network

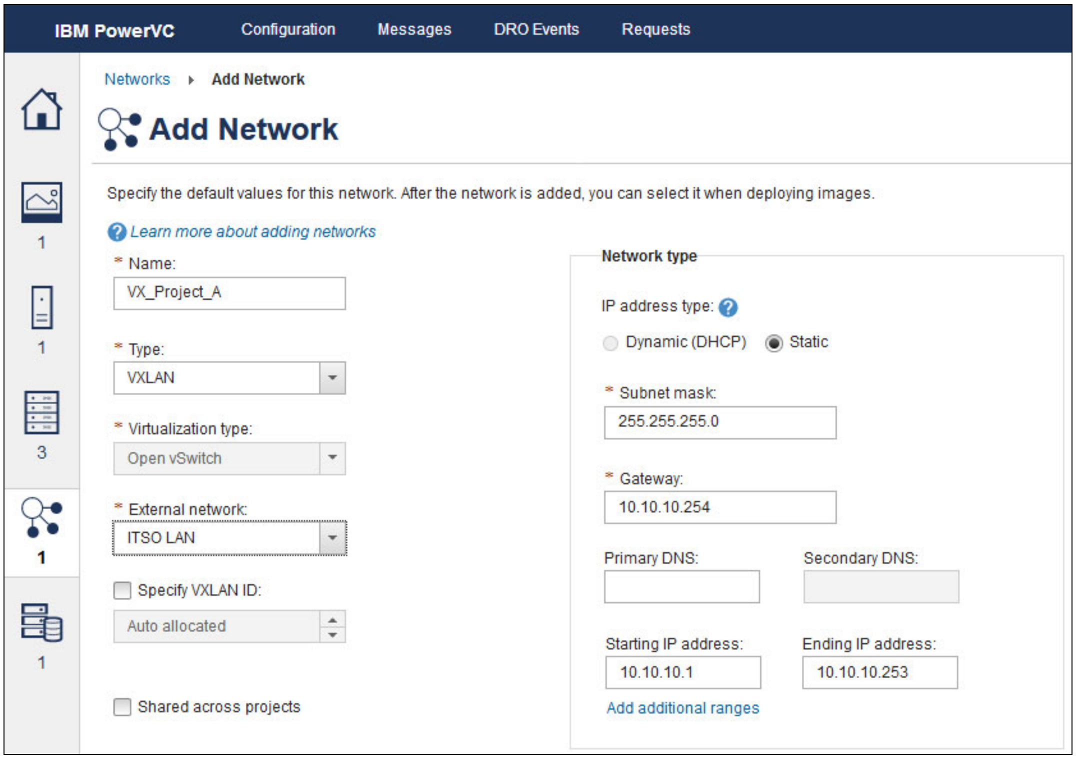
Note: As shown in Figure 4-8, the virtualization type is predefined with OVS because you may have only an OVS environment. All traffic is handled from br-ex.

Now, you can create multiple VXLANs on top of that environment. The VXLANs are private, but can communicate with the outside network or even the internet if activated.

You can think of this solution as a home network, where you have several devices in a private network behind a router. The router sends the packets into the internet and receives packets and forwards them to the device by using NAT. In IBM PowerVC SDI, the network node is the router with NAT functions.

As an example, imagine that you have two projects with several VMs that should be able to communicate inside their project, but not to the other project. Also, the VMs should be able to access the internet. This configuration can be easily accomplished by creating two individual VXLANs. The connection to the internet is done automatically because the flat network can access the internet.

Figure 4-9 shows the definition of the first VXLAN.



Add Network

Specify the default values for this network. After the network is added, you can select it when deploying images.

* Name: VX_Project_A

* Type: VXLAN

* Virtualization type: Open vSwitch

* External network: ITSO LAN

□ Specify VXLAN ID: Auto allocated

□ Shared across projects

Network type

IP address type: ○ Dynamic (DHCP) ● Static

* Subnet mask: 255.255.255.0

* Gateway: 10.10.10.254

Primary DNS:

Secondary DNS:

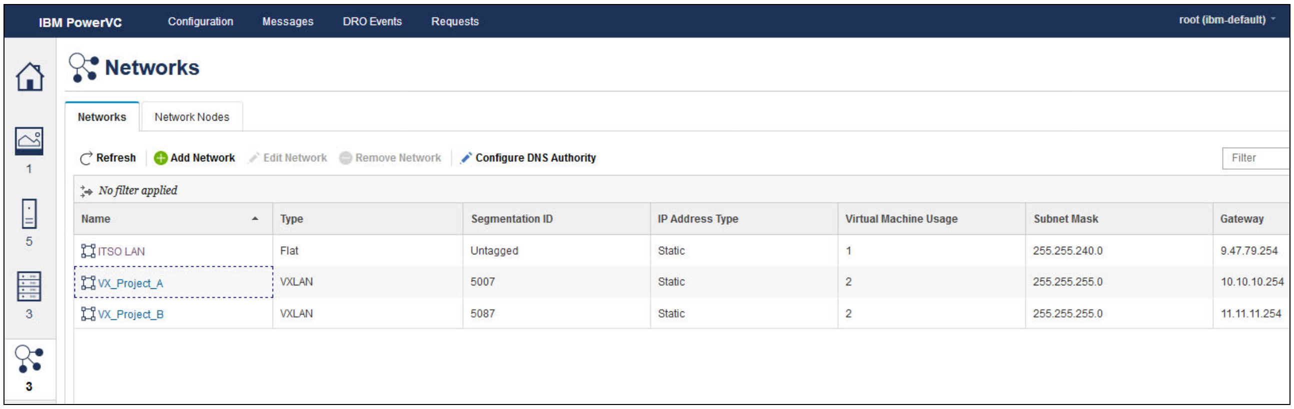
Starting IP address: 10.10.10.1

Ending IP address: 10.10.10.253

Figure 4-9 Definition of a VXLAN

In the definition, you must select the external network (ITSO LAN in our case), and in the right column enter the IP values. If you want, you can specify the ID of the VXLAN, or leave that task to IBM PowerVC.

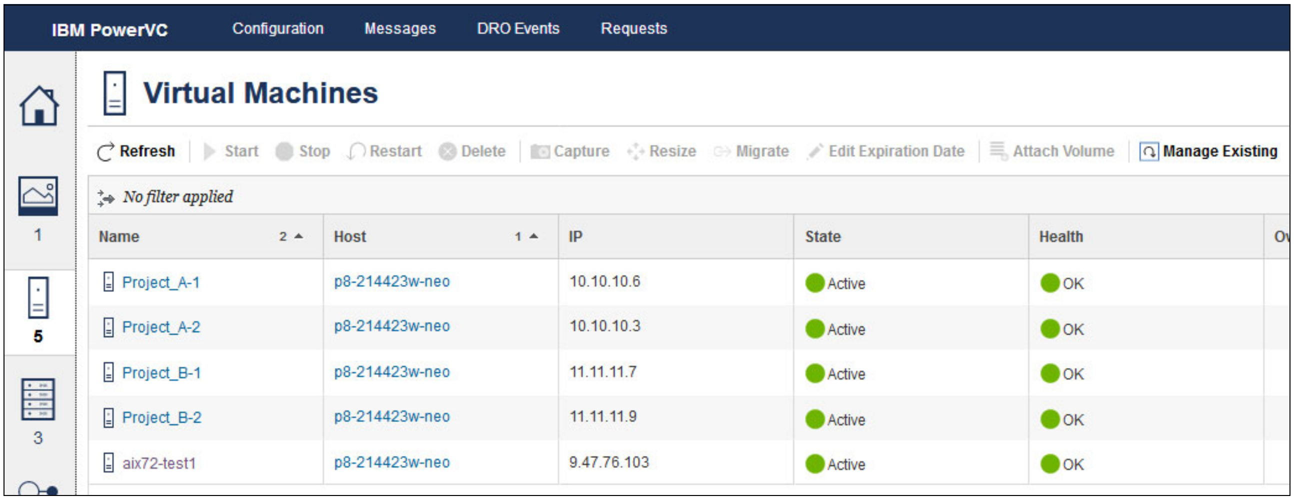
Using this example, create a second VXLAN for Project B with the IP range 11.11.11.1 - 11.11.11.253. Figure 4-10 shows the result of the definitions in IBM PowerVC.



Name	Type	Segmentation ID	IP Address Type	Virtual Machine Usage	Subnet Mask	Gateway
ITSO LAN	Flat	Untagged	Static	1	255.255.240.0	9.47.79.254
VX_Project_A	VXLAN	5007	Static	2	255.255.255.0	10.10.10.254
VX_Project_B	VXLAN	5087	Static	2	255.255.255.0	11.11.11.254

Figure 4-10 Two VXLANs for two projects

After you define the networks, deploy two VMs for each project by using the VXLAN of the project, as shown in Figure 4-11. Each VM receives an IP from the IP range of the selected VXLAN.



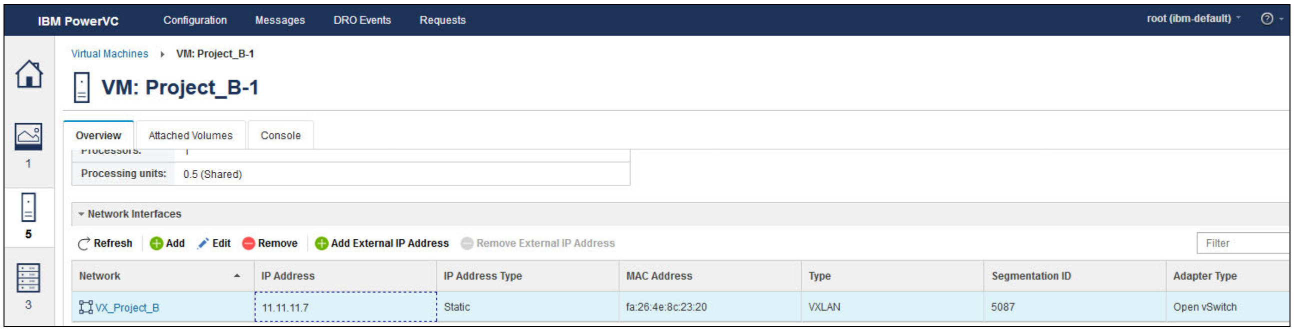
Name 2 ▲	Host 1 ▲	IP	State	Health	O
Project_A-1	p8-214423w-neo	10.10.10.6	● Active	● OK	
Project_A-2	p8-214423w-neo	10.10.10.3	● Active	● OK	
Project_B-1	p8-214423w-neo	11.11.11.7	● Active	● OK	
Project_B-2	p8-214423w-neo	11.11.11.9	● Active	● OK	
aix72-test1	p8-214423w-neo	9.47.76.103	● Active	● OK	

Figure 4-11 Two projects with two VMs with VXLANs

If you want to log in to one of the VMs, this is possible only over a terminal connection because there are only private VXLANs in this example at the moment. To open a terminal connection, either use IBM PowerVC, or run **mkvterm** in PowerVM NovaLink or the HMC. Inside the VM, you can ping all the other VMs in the same VXLAN or to the outside world, such as the internet.

In a production environment, VMs with only private VXLANs are impractical. At least one VM must be accessible from outside, for example, the web server of a multitier application. The rest of the application may be in private VMs behind the web server.

Coming back to our example, imagine that you want to access VM Project_B-1 with its private IP 11.11.11.7 from outside. This task can be easily done in IBM PowerVC. Access the details of the VM in IBM PowerVC as shown in Figure 4-12.



VM: Project_B-1

Processors: 1

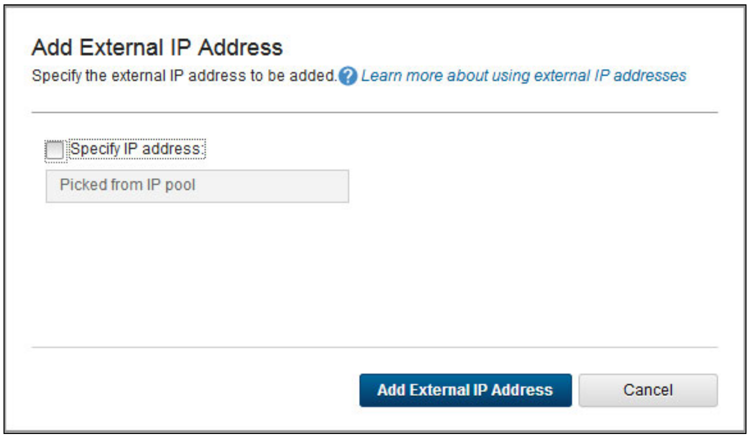
Processing units: 0.5 (Shared)

Network Interfaces

Network	IP Address	IP Address Type	MAC Address	Type	Segmentation ID	Adapter Type
VX_Project_B	11.11.11.7	Static	fa:26:4e:8c:23:20	VXLAN	5087	Open vSwitch

Figure 4-12 VXLAN details of a VM

Now, select the VXLAN and click **Add External IP Address**. In the window that opens, enter a specific IP address from the pool of the external LAN (in our case, ITSO LAN), or let IBM PowerVC pick an address, as shown in Figure 4-13.



Add External IP Address

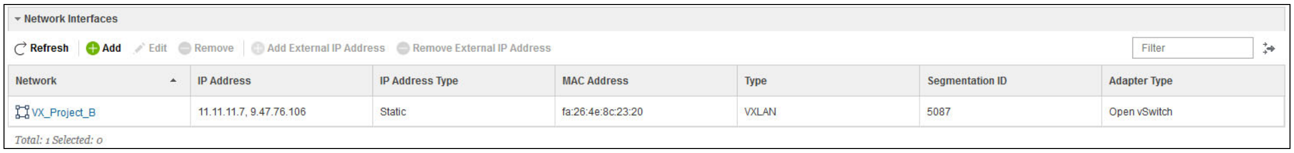
Specify the external IP address to be added. Learn more about using external IP addresses

□ Specify IP address:

Picked from IP pool

Figure 4-13 Add External IP Address

The network entry in the VM details changed, as shown in Figure 4-14. The additional external IP address 9.47.76.106 is present.



Network	IP Address	IP Address Type	MAC Address	Type	Segmentation ID	Adapter Type
VX_Project_B	11.11.11.7, 9.47.76.106	Static	fa:26:4e:8c:23:20	VXLAN	5087	Open vSwitch

Total: 1 Selected: 0

Figure 4-14 VXLAN interface with external IP

You can log in by using the external IP, which is known only by the network node. If you show the IP details of the network, as shown in Figure 4-15, you find only the internal VXLAN IP.

```
# ssh 9.47.76.106
root@9.47.76.106's password:
Last unsuccessful login: Tue Jun 26 20:41:33 CDT 2018 on ssh from 9.47.76.108
Last login: Tue Jun 26 20:41:47 CDT 2018 on /dev/pts/0 from 9.47.76.108

# ifconfig -a
en0:
flags=1e084863,814c0<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,
64BIT,CHECKSUM_OFFLOAD(ACTIVE),LARGESEND,CHAIN>
        inet 11.11.11.7 netmask 0xffffff00 broadcast 11.11.11.255
         tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
sit0: flags=8100041<UP,RUNNING,LINK0>
        inet6 ::11.11.11.7/96
lo0:
flags=e08084b,c0<UP,BROADCAST,LOOPBACK,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT,
LARGESEND,CHAIN>
        inet 127.0.0.1 netmask 0xff000000 broadcast 127.255.255.255
        inet6 ::1%1/128
         tcp_sendspace 131072 tcp_recvspace 131072 rfc1323 1
```

Figure 4-15 Logging in with the external IP address

4.4 Security groups

Security groups is a new technology that is introduced with IBM PowerVC V1.4.1. By using them, you may restrict network traffic that goes into a VM or outside a VM. You can think of it like a firewall in the network.

Security groups work only in an OVS environment (not with VIO SEA networks or SR-IOV). However, this technology works with VLANs, either a flat network or VXLANs. It is not tied to VXLANs only.

In IBM PowerVC V1.4.1, security groups are available only through the OpenStack interface on the command-line interface (CLI).

A security group consists of several rules that are tied to it. A security group can be attached to a port of a VM. There is a default security group that is defined during the IBM PowerVC installation. This security group is tied to a port automatically so that the VM can talk on every port to the outside and be accessible on all ports from the outside (everything is allowed). To change that configuration, the default security group must be removed from the port and new one (or more) must be attached.

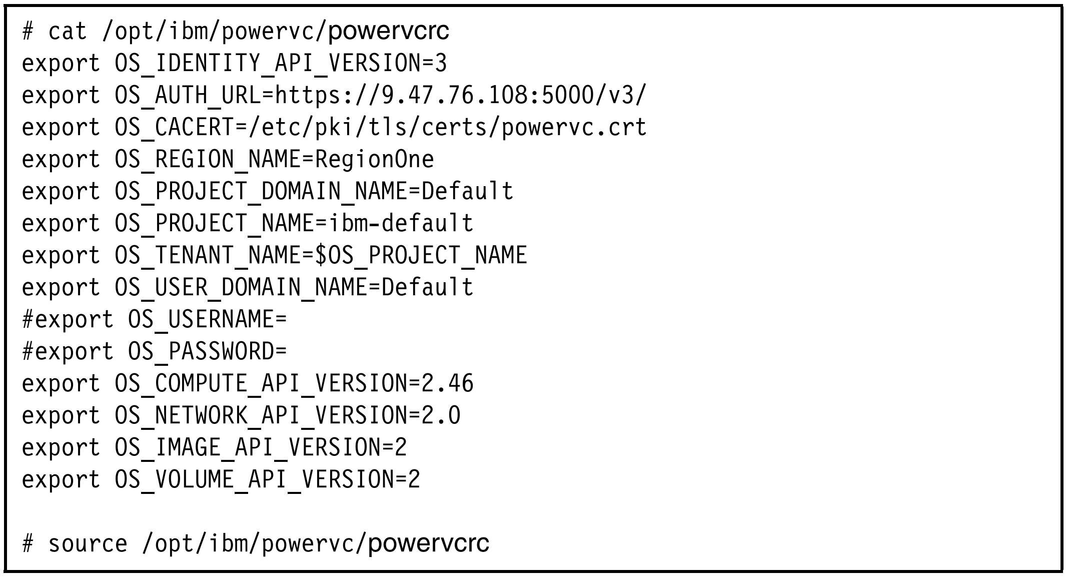
4.4.1 Creating and modifying security groups

Before you can work with the security groups, you must prepare the OpenStack environment. For a default environment without projects, use the example file `powervcrc` from the `/opt/ibm/powervc` directory, as shown in Figure 4-16.

You may enter the user name and password for every OpenStack command you enter, or you can store the credentials in `powervcrc`. If you want to work in a different project, replace `ibm-default` with the name of your project.

Complete the following steps:

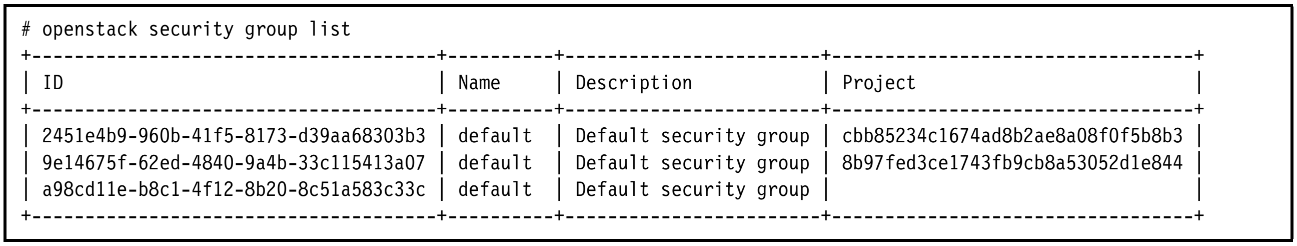
1. To make your environment active, source it, as shown in Figure 4-16.



```
# cat /opt/ibm/powervc/powervcrc
export OS_IDENTITY_API_VERSION=3
export OS_AUTH_URL=https://9.47.76.108:5000/v3/
export OS_CACERT=/etc/pki/tls/certs/powervc.crt
export OS_REGION_NAME=RegionOne
export OS_PROJECT_DOMAIN_NAME=Default
export OS_PROJECT_NAME=ibm-default
export OS_TENANT_NAME=$OS_PROJECT_NAME
export OS_USER_DOMAIN_NAME=Default
#export OS_USERNAME=
#export OS_PASSWORD=
export OS_COMPUTE_API_VERSION=2.46
export OS_NETWORK_API_VERSION=2.0
export OS_IMAGE_API_VERSION=2
export OS_VOLUME_API_VERSION=2

# source /opt/ibm/powervc/powervcrc
```

Figure 4-16 The powervcrc file

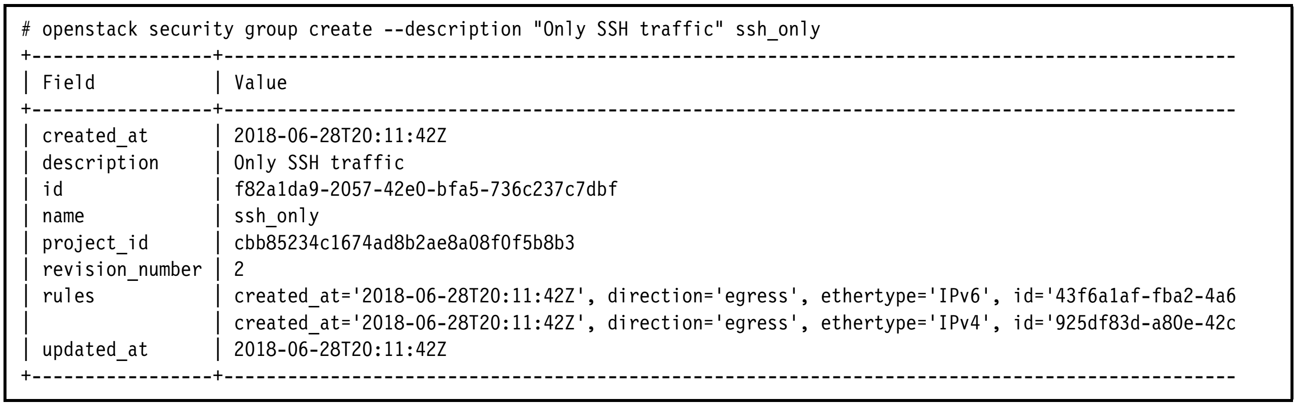
2. Now, you may run OpenStack commands. To view the security groups that exist, run the command that is shown in Figure 4-17.



```
# openstack security group list
+--------------------------------------+----------+------------------------+----------------------------------+
| ID                                   | Name     | Description            | Project                          |
+--------------------------------------+----------+------------------------+----------------------------------+
| 2451e4b9-960b-41f5-8173-d39aa68303b3 | default  | Default security group | cbb85234c1674ad8b2ae8a08f0f5b8b3 |
| 9e14675f-62ed-4840-9a4b-33c115413a07 | default  | Default security group | 8b97fed3ce1743fb9cb8a53052d1e844 |
| a98cd11e-b8c1-4f12-8b20-8c51a583c33c | default  | Default security group |                                  |
+--------------------------------------+----------+------------------------+----------------------------------+
```

Figure 4-17 Default security groups

3. To define a security group, for example, that is called `ssh_only`, run the command that is shown in Figure 4-18.



```
# openstack security group create --description "Only SSH traffic" ssh_only
+-----------------+-----------------------------------------------------------------------------------------------
| Field           | Value
+-----------------+-----------------------------------------------------------------------------------------------
| created_at      | 2018-06-28T20:11:42Z
| description     | Only SSH traffic
| id              | f82a1da9-2057-42e0-bfa5-736c237c7dbf
| name            | ssh_only
| project_id      | cbb85234c1674ad8b2ae8a08f0f5b8b3
| revision_number | 2
| rules           | created_at='2018-06-28T20:11:42Z', direction='egress', ethertype='IPv6', id='43f6a1af-fba2-4a6
|                 | created_at='2018-06-28T20:11:42Z', direction='egress', ethertype='IPv4', id='925df83d-a80e-42c
| updated_at      | 2018-06-28T20:11:42Z
+-----------------+-----------------------------------------------------------------------------------------------
```

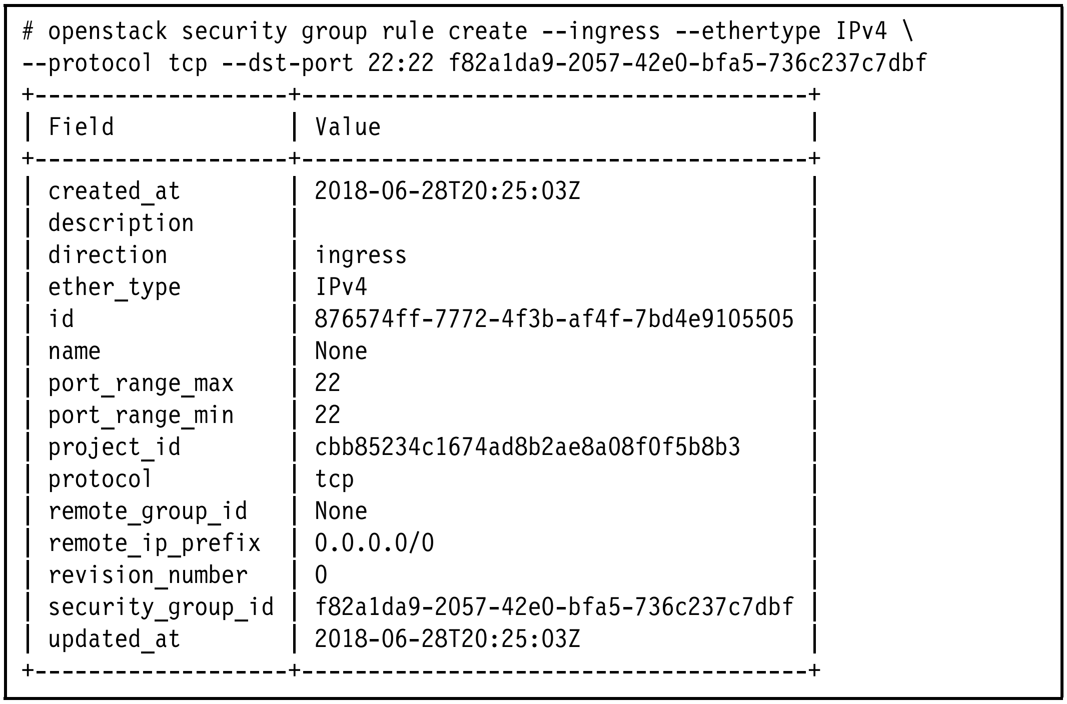
Figure 4-18 Creating a security group

There are two types of traffic directions: egress means out-bound traffic from the VM into the LAN/WAN, and ingress means in-bound traffic from the outside into the VM.

When you create a security group, it has as two default egress rules that allow any traffic for IPv4 and IPv6 to go outside the VM.

As you probably assume from the name `ssh_onlyp`, the goal is to restrict the in-bound traffic to SSH.

4. Add a rule that allows only in-bound SSH traffic, as shown in Figure 4-19.



```
# openstack security group rule create --ingress --ethertype IPv4 \
--protocol tcp --dst-port 22:22 f82a1da9-2057-42e0-bfa5-736c237c7dbf
+-------------------+--------------------------------------+
| Field             | Value                                |
+-------------------+--------------------------------------+
| created_at        | 2018-06-28T20:25:03Z                 |
| description       |                                      |
| direction         | ingress                              |
| ether_type        | IPv4                                 |
| id                | 876574ff-7772-4f3b-af4f-7bd4e9105505 |
| name              | None                                 |
| port_range_max    | 22                                   |
| port_range_min    | 22                                   |
| project_id        | cbb85234c1674ad8b2ae8a08f0f5b8b3     |
| protocol          | tcp                                  |
| remote_group_id   | None                                 |
| remote_ip_prefix  | 0.0.0.0/0                            |
| revision_number   | 0                                    |
| security_group_id | f82a1da9-2057-42e0-bfa5-736c237c7dbf |
| updated_at        | 2018-06-28T20:25:03Z                 |
+-------------------+--------------------------------------+
```

Figure 4-19 Adding a rule to a security group

5. To show the content of the security group after adding the rule, run **openstack security group show**, as shown in Figure 4-20.

```
# openstack security group show f82a1da9-2057-42e0-bfa5-736c237c7dbf
+-----------------+-------------------------------------------------------------------------------------------------
| Field           | Value
+-----------------+-------------------------------------------------------------------------------------------------
| created_at      | 2018-06-27T21:59:28Z
| description     | Only SSH
| id              | f82a1da9-2057-42e0-bfa5-736c237c7dbf
| name            | ssh_only
| project_id      | cbb85234c1674ad8b2ae8a08f0f5b8b3
| revision_number | 5
| rules           | created_at='2018-06-27T21:59:28Z', direction='egress', ethertype='IPv6', id='6ac0af42-080c-4e5
|                 | created_at='2018-06-27T22:07:23Z', direction='ingress', ethertype='IPv4', id='876574ff-7772-
|                 |   4f3b-af4f-7bd4e9105505', port_range_max='22', port_range_min='22', protocol='tcp',
|                 |   remote_ip_prefix='0.0.0.0/0', updated_at='2018-06-27T22:07:23Z'
|                 | created_at='2018-06-27T21:59:28Z', direction='egress', ethertype='IPv4', id='e96e1b7b-ec69-4f6
| updated_at      | 2018-06-28T20:29:59Z
+-----------------+-------------------------------------------------------------------------------------------------
```

Figure 4-20 Showing the rules of a security group

4.4.2 Attaching a security group to a port

Before you can attach a security group to a port, you must discover the ID of the port where you want to attach the security group by running **openstack port list**, as shown in Figure 4-21.

```
# openstack port list
+--------------------------------------+------+-------------------+---------------------------------------------
| ID                                   | Name | MAC Address       | Fixed IP Addresses
+--------------------------------------+------+-------------------+---------------------------------------------
| 0a8ef313-34dd-430c-9aa9-01c0701cb033 |      | fa:20:f5:3f:37:20 | ip_address='9.47.76.99', subnet_id='ce7d4a4b-
| 2dfef4b9-38f1-43f4-abaf-f5c0cfc28df0 |      | fa:16:3e:6e:17:df | ip_address='9.47.76.104', subnet_id='ce7d4a4b-
| 32a57ef1-c4db-47cc-9dcd-adb901532ce5 |      | fa:16:3e:cf:e7:04 | ip_address='11.11.11.254', subnet_id='ffd7038b
| 4dc16450-54a4-4c09-9b70-ffe70e42380c |      | fa:16:3e:79:fa:4b | ip_address='10.10.10.254', subnet_id='c07f84b3
| 61141592-9b70-418c-b97a-fbe02db45ffe |      | fa:92:d6:09:88:20 | ip_address='10.10.10.6', subnet_id='c07f84b3-
| 64370c63-99bf-431c-a70a-74eb529fd4aa | HA p | fa:16:3e:e8:ec:53 | ip_address='169.254.192.11', subnet_id='b0e511
| 72c393c2-997f-4b0e-864f-c55aa2a1a2f9 |      | fa:54:1b:10:ce:20 | ip_address='9.47.76.102', subnet_id='ce7d4a4b-
| 7642ef66-17d3-4ef9-9cf6-e1fc6b7dfa35 |      | fa:16:3e:84:f6:af | ip_address='9.47.76.106', subnet_id='ce7d4a4b-
| 966f11f1-24df-4f2a-b24d-d5a4ef9a11ed |      | fa:fd:ca:32:12:20 | ip_address='9.47.76.103', subnet_id='ce7d4a4b-
| b297ca42-390d-48ad-bf19-5df0c6378bc4 |      | fa:4b:12:1c:df:20 | ip_address='10.10.10.3', subnet_id='c07f84b3-
| b6634580-6a17-45e5-8ad0-7fb261c13419 | HA p | fa:16:3e:01:64:79 | ip_address='169.254.192.10', subnet_id='b0e511
| bb6b08de-dcfb-4deb-a630-4d9416e8ea44 |      | fa:21:01:90:23:20 | ip_address='11.11.11.9', subnet_id='ffd7038b-
| cf7a43d6-7093-4fb7-9ed5-2244dfcf8ebe |      | fa:26:4e:8c:23:20 | ip_address='11.11.11.7', subnet_id='ffd7038b-
| e05dd91b-78e5-4852-bf14-9c19f9f45f26 |      | fa:16:3e:fd:85:98 | ip_address='9.47.76.98', subnet_id='ce7d4a4b-
+--------------------------------------+------+-------------------+---------------------------------------------
```

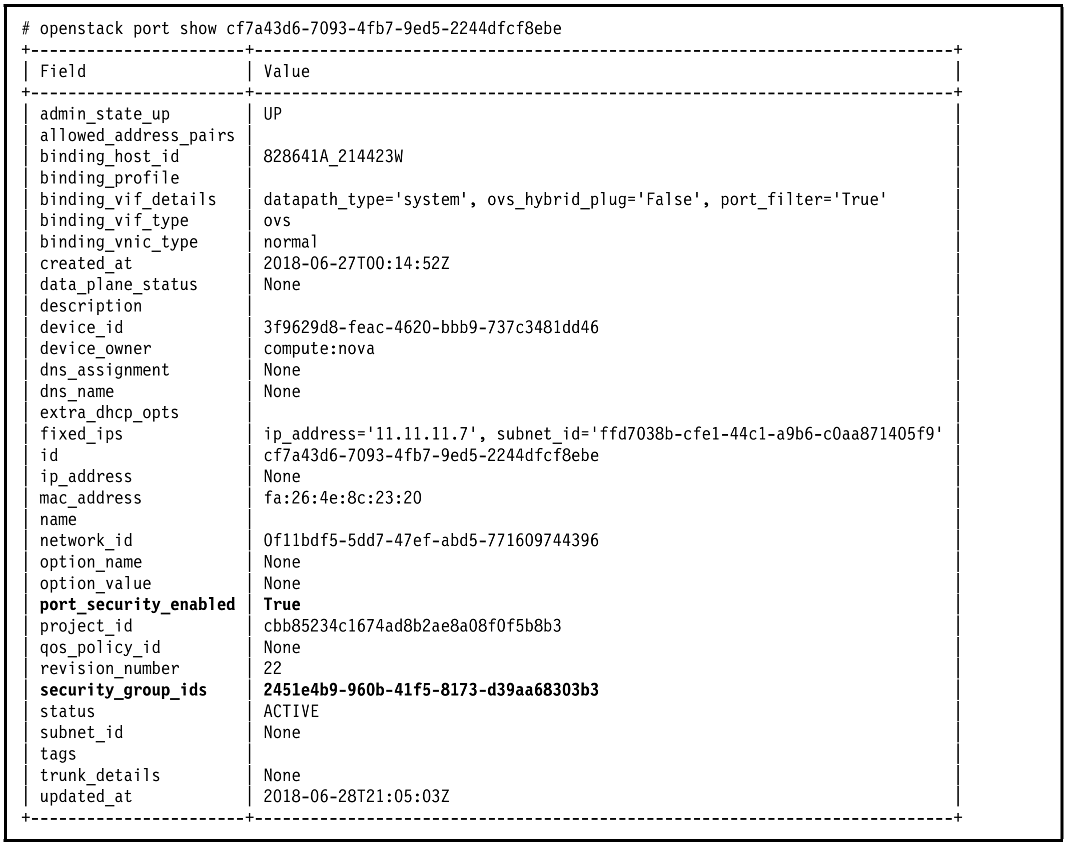
Figure 4-21 Listing all the network ports of the cluster

On the right side of Figure 4-21, there is a column cut that shows the status for the port. In our case, the status is Active for all ports, except for the external attached IP address 9.47.76.106 that we attached in 4.3, “The configuration and usage of Virtual Extensible LANs” on page 91. This external IP has as status “N/A”.

The status of N/A is important because you should not attach a security group to an external IP on top of a VXLAN IP. Instead, attach it to the VXLAN IP, which in this case is 11.11.11.7. The ID for that IP is cf7a43d6-7093-4fb7-9ed5-2244dfcf8ebe.

To attach the security group to a port, complete the following steps:

1. Remove the default security group. Figure 4-22 shows the default entries of that port.

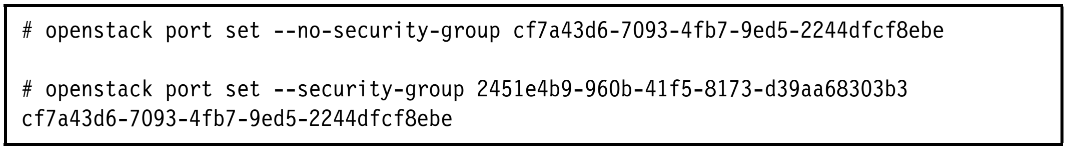


```
# openstack port show cf7a43d6-7093-4fb7-9ed5-2244dfcf8ebe
+-----------------------+----------------------------------------------------------------------------+
| Field                 | Value                                                                      |
+-----------------------+----------------------------------------------------------------------------+
| admin_state_up        | UP                                                                         |
| allowed_address_pairs |                                                                            |
| binding_host_id       | 828641A_214423W                                                            |
| binding_profile       |                                                                            |
| binding_vif_details   | datapath_type='system', ovs_hybrid_plug='False', port_filter='True'        |
| binding_vif_type      | ovs                                                                        |
| binding_vnic_type     | normal                                                                     |
| created_at            | 2018-06-27T00:14:52Z                                                       |
| data_plane_status     | None                                                                       |
| description           |                                                                            |
| device_id             | 3f9629d8-feac-4620-bbb9-737c3481dd46                                       |
| device_owner          | compute:nova                                                               |
| dns_assignment        | None                                                                       |
| dns_name              | None                                                                       |
| extra_dhcp_opts       |                                                                            |
| fixed_ips             | ip_address='11.11.11.7', subnet_id='ffd7038b-cfe1-44c1-a9b6-c0aa871405f9'  |
| id                    | cf7a43d6-7093-4fb7-9ed5-2244dfcf8ebe                                       |
| ip_address            | None                                                                       |
| mac_address           | fa:26:4e:8c:23:20                                                          |
| name                  |                                                                            |
| network_id            | 0f11bdf5-5dd7-47ef-abd5-771609744396                                       |
| option_name           | None                                                                       |
| option_value          | None                                                                       |
| port_security_enabled | True                                                                       |
| project_id            | cbb85234c1674ad8b2ae8a08f0f5b8b3                                           |
| qos_policy_id         | None                                                                       |
| revision_number       | 22                                                                         |
| security_group_ids    | 2451e4b9-960b-41f5-8173-d39aa68303b3                                       |
| status                | ACTIVE                                                                     |
| subnet_id             | None                                                                       |
| tags                  |                                                                            |
| trunk_details         | None                                                                       |
| updated_at            | 2018-06-28T21:05:03Z                                                       |
+-----------------------+----------------------------------------------------------------------------+
```

Figure 4-22 Default security group of a port

   If you look for ID 2451e4b9-960b-41f5-8173-d39aa68303b3 in Figure 4-17 on page 96, you see that this ID belongs to the default security group.

2. To remove the default security group and attach the one that we created (`ssh_only`), run the commands that are shown in Figure 4-23.



```
# openstack port set --no-security-group cf7a43d6-7093-4fb7-9ed5-2244dfcf8ebe

# openstack port set --security-group 2451e4b9-960b-41f5-8173-d39aa68303b3
cf7a43d6-7093-4fb7-9ed5-2244dfcf8ebe
```

Figure 4-23 Replacing the default security group

The rule is in place, and access from outside into VXLAN VM 11.11.11.7 from the VXLAN or from outside by using the attached floating IP 9.47.76.106 is possible only with SSH. All other ports are no longer accessible.

5

Maintenance in a software-defined infrastructure environment

An IBM PowerVC software-defined infrastructure (SDI) environment also needs maintenance periodically. This chapter describes some maintenance cases and how they might be accomplished.

5.1 Updating IBM Spectrum Scale

IBM PowerVC deploys and manages IBM Spectrum Scale, but because IBM Spectrum Scale is a piece of software, its updates can come independently of IBM PowerVC updates. You can upgrade to new IBM Spectrum Scale releases by completing the following steps:

1. Downloading the new levels to `/opt/ibm/powervc/images/spectrum-scale`.
2. Open the IBM PowerVC Hosts window. You see that the hosts have an IBM Spectrum Scale update. IBM PowerVC helps manage the update across the hosts.
3. IBM PowerVC prompts you to put the host into maintenance mode. Click **Update Spectrum Scale** to have IBM PowerVC automatically update the IBM Spectrum Scale software on that host. When this is complete, you can take the server out of host maintenance mode.
4. Repeat this process until all hosts in the cluster have the upgraded levels of IBM Spectrum Scale.

For more information, see IBM Knowledge Center.

5.2 Checking the cluster health and resolving issues

In the event of a failure, the first step is to check the health of the cluster from the IBM PowerVC server.

> **Important:** IBM PowerVC manages the IBM Spectrum Scale cluster for you. *Do not try to modify the cluster yourself by using IBM Spectrum Scale commands* because IBM PowerVC created a special cluster configuration that should not be changed.

Here are some helpful commands to check the health of the cluster:

- **`/usr/lpp/mmfs/bin/mmgetstate -a`**

  Shows the state of all of the nodes in the cluster. If a node is down, check the network connectivity between the servers. Make sure that IP addresses or host names have not changed. Try to start the node.

- **`/usr/lpp/mmfs/bin/mmstartup -N <SERVER>`**

  Starts the IBM Spectrum Scale processes on the remote server. Use **`mmgetstate`** to check the status.

- **`/usr/lpp/mmfs/bin/mmlsmount powervc_gpfs -L`**

  Shows which servers the file system is mounted on. After the node starts, it should mount automatically. If a mount is missing but the node is active, use the corresponding **`mmmount`** command to attempt to mount the file system.

- **`/usr/lpp/mmfs/bin/mmlsdisk powervc_gpfs`**

  Lists the disks and their states within the cluster. If enough disks are in an invalid state, the cluster might stop.

  IBM PowerVC tries to restart failed disks that are in a limbo state. However, if a disk fails, the replace disk procedure must be run.

  Instructions to replace a failed disk can be found in IBM Knowledge Center.

5.3 Updating PowerVM NovaLink or the KVM host operating system

Servers must have operating system updates applied. Whether it is the PowerVM NovaLink partition on PowerVM or your kernel-based virtual machine (KVM) Linux host, updates must be applied periodically.

It is a good idea to do updates of PowerVM NovaLink or the KVM host operating system because if the network connection drops, you do not lose your SSH connection, which is critical because the network components receive periodic operating system updates that take the server off the network.

If the system is co-managed by an HMC, you may open a terminal from the HMC even if PowerVM NovaLink has the control over the system.

If your system is not co-managed by an HMC, especially for KVM-based system, use an **ipmi** connection to your host to apply the updates. This process connects you to the server through a service processor (such as Flexible Service Processor (FSP) or baseboard management controller (BMC)) and provides a terminal for the updates.

A sample **ipmi** command that provides a terminal connection to the host might look like the following one:

```
ipmitool -I lanplus -H <FSP IP ADDRESS> -P <PASSWORD> sol activate
```

5.4 Switching a host between PowerVM NovaLink managed and HMC-managed

If your systems are co-managed by an HMC, sometimes it is a good idea to switch the control to the HMC. Examples are when you do a firmware update of a whole node or exchange hardware parts. This switch can be accomplished by setting the node into maintenance mode by completing the following steps:

1. In the Hosts section of IBM PowerVC, enter the HMC and enter the following command:

   ```
   $ chcomgmt -o setmaster -t norm -m <system>
   ```

   The red closed lock in the HMC GUI for that system should be replaced by a grey open lock. The HMC now controls that system and you may do your maintenance activity.

2. After finishing the maintenance activity, give back control to PowerVM NovaLink by running the following command:

   ```
   $ chcomgmt -o relmaster -m <system>
   ```

3. Exit the maintenance mode in IBM PowerVC.

5.5 Backing up and restoring

Backing up and restoring your management server is important, but data is the most important resource. It is always wise to have a backup/recovery solution for your data.

5.5.1 Approaches to backing up cluster data

Several solutions are available for backing up and restoring the data for a server. Here are two of them:

- Manual backup jobs
- IBM Spectrum Protect

Manual backup jobs

IBM Spectrum Scale is effectively a scale-out file system across all the servers. A simple backup/restore operation, such as copying the files to an independent storage subsystem, might suffice for some smaller clouds.

For considerations about how to make the backup operation consistent, see “Volume consistency: Using IBM Spectrum Scale snapshots” on page 106.

IBM Spectrum Protect

Administrators should take advantage of IBM Spectrum Protect™ to back up their data.

IBM Knowledge Center documents how to integrate IBM Spectrum Protect, including considerations when doing the backup. You run **mmbackup** to start backing up to a set of IBM Spectrum Protect servers.

Consider the following items when pairing IBM Spectrum Protect with IBM PowerVC SDI deployments:

- IBM PowerVC stores raw virtual machine (VM) block devices as files. These files can be large and are binary in nature.
- A significant amount of network bandwidth might be required to drive the backup/restore process. Careful planning is necessary, and you might need a separate, backup-specific network.

When deciding how frequently to back up the data, consider how long each backup takes. This frequency might change as your cloud scales out.

IBM Spectrum Protect uses a server/agent model. The servers participating in the backup process provide the storage for the actual backups. The agents are used as a vehicle to move the data into the servers. Therefore, the IBM Spectrum Protect Client agents must be installed on the servers that participate in the backup process.

The **mmbackup** command identifies any servers that do not have the agent installed and excludes those servers from the backup. If a server is excluded from the backup process, it does not mean that its cluster data will not be backed up. Instead, another server in the cluster sends the data to the IBM Spectrum Protect server. For example, if there are three servers in the cluster and all three servers have the agent installed, each server sends one third of the data. However, if only two of the servers have the agent installed, each server sends half of the data.

After you have a IBM Spectrum Protect server (or set of servers) set up, you run a command similar to the following one to start the backup process:

```
/usr/lpp/mmfs/bin/mmbackup powervc_gpfs -N pvc_compute_node --tsm-servers
TSMServer[,TSMServer...] -S SNAPSHOT
```

In the example, **`pvc_compute_node`** is a built-in alias that indicates that it should run on the compute nodes instead of the IBM PowerVC server. This setting generally makes more sense because the compute nodes have direct access to the storage device.

The **`-S`** references a file system snapshot that should be used for the backup process. For information about why this is needed, see “Volume consistency: Using IBM Spectrum Scale snapshots” on page 106.

While a backup is running, it takes bandwidth away from the workloads. You can reduce the hosts that are used for the backup operation by completing the following steps:

1. Find the cluster member server names by running the following command:

   ```
   /usr/lpp/mmfs/bin/mmlscluster
   ```

   Look for the nodes that do not have the designation of `manager` (that is the IBM PowerVC server). The admin node name is the server name.
2. Choose the hosts to run the backup. You might, for example, select hosts with fewer workloads or higher bandwidth.
3. When running the **`mmbackup`** command, replace **`-N pvc_compute_node`** with **`-N server1, server2`**, and so on. Doing so limits the backup operation to just those hosts.

For more information about this task, see 5.5.2, “Considerations when backing up data” on page 105.

5.5.2 Considerations when backing up data

When backing up storage data, there are aspects that can affect your backup performance and file stability:

- High-bandwidth usage: The files that are used for the disks are generally large, requiring multiple gigabytes. Large volumes might be terabytes in size. Moving that much data requires a significant data connection.
- Data consistency: While VMs are running on top of the volumes, the backing file contents might change. Determining the correct strategy or backup windows is critical to ensure data consistency.

Bandwidth consideration: Limiting the hosts that are used for the backup process

Every server in your cloud is connected to the IBM Spectrum Scale storage cluster. If you have 100 servers, you might not want all of them spending bandwidth on backup operations. Then again, having 100 servers participating in the backup might also reduce the burden on each individual server.

Some environments maintain *hot-spare* servers that do not run any VMs unless one of the other hosts has an error or needs maintenance. Therefore, they might be running a workload only when a maintenance situation arises. Those “quiet” servers can be prime candidates for backup operations.

If you do not have *spare* servers, it is wiser for the backup operation to be spread out across the hosts.

With IBM Spectrum Protect and IBM Spectrum Scale, you can specify specific hosts. If you are using another mechanism for backup/restore operations, then use an approach that is appropriate to that backup solution.

Bandwidth consideration: Separating the backup network

The storage and management network might be busy if there is significant storage I/O. Adding backup operations to it might lead to congestion. An alternative is to use a separate network interface controller (NIC) or port that has a route to the backup servers, such as IBM Spectrum Protect.

This technique does not eliminate the impact on your storage network. Read operations from the compute node for a file might affect the network. However, it does eliminate the traffic from the compute node to the backup server.

Because IBM PowerVC does not manage such a network, the network must be set up manually across the compute hosts.

Volume consistency: Using IBM Spectrum Scale snapshots

A simple approach to solving the volume consistency at backup time is to use snapshots. A *snapshot* takes a point-in-time capture of all of the data in the IBM Spectrum Scale file system, but the VMs continue to run. Think of it as a live capture of all the data in the file system.

This data is a linked, so it does not take up significant extra space when the snapshot is taken.

The basic flow involves these steps:

1. Create snapshot.
2. Back up the snapshot.
3. Delete the snapshot.

During this entire flow, the VMs continue to run.

To create the snapshot, run the **mmcrsnapshot** command from the IBM PowerVC server:

```
/usr/lpp/mmfs/bin/mmcrsnapshot powervc_gpfs backup_snap
```

The parameter backup_snap is an identifier for the snapshot. The powervc_gpfs parameter identifies the IBM Spectrum Scale file system that IBM PowerVC creates and maintains.

The command creates a snapshot in /gpfs/powervc_gpfs/.snapshots/backup_snap/. The files can be copied from there to the backup device. If you use IBM Spectrum Protect, this process can be specified by the **mmbackup -S** command.

Although the **mmcrsnapshot** and **mmdelsnapshot** commands must be run from the IBM PowerVC server, it is better if the compute hosts copy the data. These hosts have more bandwidth and can scale out the copy operations across the nodes.

For more information about creating the snapshot, see IBM Knowledge Center.

After the backup is completed, the snapshot can be deleted by using the **mmdelsnapshot** command. The basic syntax for this command is:

```
/usr/lpp/mmfs/bin/mmdelsnapshot powervc_gpfs backup_snap
```

5.5.3 Restoring data

Backing up data is hopefully the only operation you need. However, in the event that data is lost, recovery is possible because you have been diligent in taking backups.

When your cluster is healthy, restoration of data might be necessary. This process involves a few steps:

1. Restore the volume by using your preferred method. In IBM Spectrum Protect, run a **dsmc restore** command:

   ```
   /opt/tivoli/tsm/client/ba/bin/dsmc restore
   /gpfs/powervc_gpfs/infrastructure/<file> -latest
   ```

   In this example, we restore a single file. However, you might find yourself needing to restore multiple files, depending on the scope of the outage. Restart any VMs after their image files are restored.

   If you find yourself in a situation where the VM was deleted (and thus its data is gone), you might still be able to restore it. In that scenario, restore the files as done in step 1.

2. Use IBM PowerVC to import that volume.
3. After the volume (or set of volumes) has been imported, create an image by going to the IBM PowerVC user interface, opening the Images window, and clicking **Create**.
4. Specify the volumes that are needed for this VM and other data for the image that is necessary.
5. Deploy the VM from that new image.

   The VM has a new NIC and some other new hardware devices, but the data within it should be restored.

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 publication provides more information about the topics in this document. It might be available in softcopy only.

- *IBM PowerVC Version 1.3.2 Introduction and Configuration*, SG24-8199

You can search for, view, download, or order this document 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:

- IBM Community: Power Systems

  https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Power%20Systems/page/Welcome

- IBM Knowledge Center: Installing PowerVM NovaLink

  https://www.ibm.com/support/knowledgecenter/POWER8/p8eig/p8eig_installing.htm

- IBM Knowledge Center: Software-defined networking

  https://www.ibm.com/support/knowledgecenter/en/SSVSPA_1.4.1/com.ibm.powervc.cloud.help.doc/powervc_sdn_cloud.html

- IBM Knowledge Center: Software-defined storage

  https://www.ibm.com/support/knowledgecenter/en/SSVSPA_1.4.1/com.ibm.powervc.cloud.help.doc/powervc_sds_cloud.html

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

REDP-5455-02

ISBN 0738457035

Printed in U.S.A.