



# Cumulus Linux 2.5.6

## User Guide

---

# Table of Contents

<b>Welcome to Cumulus Networks .....</b>	<b>4</b>
<b>Quick Start Guide .....</b>	<b>5</b>
Contents .....	6
What's New in Cumulus Linux 2.5.6 .....	6
Open Source Contributions .....	6
Prerequisites .....	7
Hardware Compatibility List .....	7
Installing Cumulus Linux .....	7
Upgrading Cumulus Linux .....	8
Configuring Cumulus Linux .....	8
Configuring 4x10G Port Configuration (Splitter Cables) .....	11
Testing Cable Connectivity .....	11
Configuring Switch Ports .....	12
Configuring a Loopback Interface .....	14
<b>Installation, Upgrading and Package Management .....</b>	<b>15</b>
Managing Cumulus Linux Disk Images .....	16
Adding and Updating Packages .....	46
Zero Touch Provisioning - ZTP .....	52
<b>System Management .....</b>	<b>61</b>
Setting Date and Time .....	62
Authentication, Authorization, and Accounting .....	65
Netfilter - ACLs .....	76
Configuring switchd .....	87
Power over Ethernet - PoE .....	90
<b>Configuring and Managing Network Interfaces .....</b>	<b>94</b>
Contents .....	95
Commands .....	95
Man Pages .....	96
Configuration Files .....	96
Basic Commands .....	96
Bringing All auto Interfaces Up or Down .....	97
ifupdown Behavior with Child Interfaces .....	97
ifupdown2 Interface Dependencies .....	99
Configuring IP Addresses .....	103
Specifying User Commands .....	104
Sourcing Interface File Snippets .....	105
Using Globs for Port Lists .....	105

Using Templates .....	106
Adding Descriptions to Interfaces .....	107
Caveats and Errata .....	107
Useful Links .....	108
Configuring Switch Port Attributes .....	108
Configuring Buffer and Queue Management .....	118
<b>Layer 1 and Layer 2 Features .....</b>	<b>123</b>
Spanning Tree and Rapid Spanning Tree .....	124
Link Layer Discovery Protocol .....	139
Prescriptive Topology Manager - PTM .....	145
Bonding - Link Aggregation .....	158
Ethernet Bridging - VLANs .....	162
Multi-Chassis Link Aggregation - MLAG .....	191
LACP Bypass .....	209
Virtual Router Redundancy - VRR .....	215
Network Virtualization .....	220
IGMP and MLD Snooping .....	302
<b>Layer 3 Features .....</b>	<b>308</b>
Routing .....	309
Introduction to Routing Protocols .....	314
Network Topology .....	316
Quagga Overview .....	318
Configuring Quagga .....	320
Open Shortest Path First - OSPF - Protocol .....	332
Open Shortest Path First v3 - OSPFv3 - Protocol .....	343
Configuring Border Gateway Protocol - BGP .....	345
Bidirectional Forwarding Detection - BFD .....	367
Equal Cost Multipath Load Sharing - Hardware ECMP .....	372
Management VRF .....	381
<b>Monitoring and Troubleshooting .....</b>	<b>385</b>
Contents .....	386
Commands .....	386
Using the Serial Console .....	386
Diagnostics Using cl-support .....	388
Sending Log Files to a syslog Server .....	389
Next Steps .....	391
Single User Mode - Boot Recovery .....	392
Using netshow to Troubleshoot Your Network Configuration .....	393
Monitoring Interfaces and Transceivers Using ethtool .....	399
Resource Diagnostics Using cl-resource-query .....	403
Monitoring System Hardware .....	404
Monitoring System Statistics and Network Traffic with sFlow .....	410

Monitoring Virtual Device Counters .....	412
Understanding and Decoding the cl-support Output File .....	417
Managing Application Daemons .....	433
Troubleshooting Network Interfaces .....	436
Network Troubleshooting .....	442
SNMP Monitoring .....	454
<b>Index .....</b>	<b>476</b>

# Welcome to Cumulus Networks

We are transforming networking with Cumulus Linux, the industry's first, full-featured Linux operating system for networking hardware. Cumulus Linux is a complete network operating system, based on [Debian wheezy](#). Unlike traditional embedded platforms, Cumulus Linux provides a complete environment pre-installed with scripting languages, server utilities, and monitoring tools. Management tasks are accomplished via SSH using standard Linux commands or over a serial console connection.



This documentation is current as of January 28, 2016 for version 2.5.6. Please visit the [Cumulus Networks Web site](#) for the most up to date documentation.

Read the [release notes](#) for new features and known issues in this release.

- [Release Notes for Cumulus Linux 2.5.6 \(see page 5\)](#)
- [Quick Start Guide \(see page 5\)](#)
- [Installation, Upgrading and Package Management](#)
- [System Management \(see page 61\)](#)
- [Configuring and Managing Network Interfaces](#)
- [Layer 2 Features \(see page 123\)](#)
- [Layer 3 Features \(see page 308\)](#)
- [Monitoring and Troubleshooting \(see page 385\)](#)

# Quick Start Guide

This chapter helps you get up and running with Cumulus Linux quickly and easily.

## Contents

(Click to expand)

- [Contents \(see page 6\)](#)
- [What's New in Cumulus Linux 2.5.6 \(see page 6\)](#)
- [Open Source Contributions \(see page 6\)](#)
- [Prerequisites \(see page 7\)](#)
- [Hardware Compatibility List \(see page 7\)](#)
- [Installing Cumulus Linux \(see page 7\)](#)
- [Upgrading Cumulus Linux \(see page 8\)](#)
- [Configuring Cumulus Linux \(see page 8\)](#)
  - [Login Credentials \(see page 9\)](#)
  - [Serial Console Management \(see page 9\)](#)
  - [Wired Ethernet Management \(see page 9\)](#)
  - [Configuring the Hostname and Time Zone \(see page 9\)](#)
  - [Installing the License \(see page 10\)](#)
- [Configuring 4x10G Port Configuration \(Splitter Cables\) \(see page 11\)](#)
- [Testing Cable Connectivity \(see page 11\)](#)
- [Configuring Switch Ports \(see page 12\)](#)
  - [Layer 2 Port Configuration \(see page 12\)](#)
  - [Layer 3 Port Configuration \(see page 13\)](#)
- [Configuring a Loopback Interface \(see page 14\)](#)

## What's New in Cumulus Linux 2.5.6

Cumulus Linux 2.5.6 contains bug fixes only. The [release notes](#) contain information about the release as well as the fixed and known issues.

## Open Source Contributions

Cumulus Networks has forked various software projects, like CFEEngine, Netdev and some Puppet Labs packages in order to implement various Cumulus Linux features. The forked code resides in the Cumulus Networks [GitHub repository](#).

Cumulus Networks developed and released as open source some new applications as well.

The list of open source projects is on the [open source software](#) page.

## Prerequisites

Prior intermediate Linux knowledge is assumed for this guide. You should be familiar with basic text editing, Unix file permissions, and process monitoring. A variety of text editors are pre-installed, including `vi` and `nano`.

You must have access to a Linux or UNIX shell. If you are running Windows, you should use a Linux environment like [Cygwin](#) as your command line tool for interacting with Cumulus Linux.

- ✔ If you're a networking engineer but are unfamiliar with Linux concepts, use [this reference guide](#) to see examples of the Cumulus Linux CLI and configuration options, and their equivalent Cisco Nexus 3000 NX-OS commands and settings for comparison. You can also [watch a series of short videos](#) introducing you to Linux in general and some Cumulus Linux-specific concepts in particular.

## Hardware Compatibility List

You can find the most up to date hardware compatibility list (HCL) [here](#). Use the HCL to confirm that your switch model is supported by Cumulus Networks. The HCL is updated regularly, listing products by port configuration, manufacturer, and SKU part number.

## Installing Cumulus Linux

This quick start guide walks you through the steps necessary for getting Cumulus Linux up and running on your switch, which includes:

1. Powering on the switch and entering ONIE, the Open Network Install Environment.
2. Installing Cumulus Linux on the switch via ONIE.
3. Booting into Cumulus Linux and installing the license.
4. Rebooting the switch to activate the switch ports.
5. Configuring switch ports and a loopback interface.

To install Cumulus Linux, you use [ONIE](#) (Open Network Install Environment), an extension to the traditional U-Boot software that allows for automatic discovery of a network installer image. This facilitates the ecosystem model of procuring switches, with a user's own choice of operating system loaded, such as Cumulus Linux.

- ⚠ If Cumulus Linux is already installed on your switch, and you need to upgrade the software only, you can skip to [Upgrading Cumulus Linux \(see page 8\)](#) below.

The easiest way to install Cumulus Linux with ONIE is via local HTTP discovery:

1. If your host (like a laptop or server) is IPv6-enabled, make sure it is running a Web server.  
If the host is IPv4-enabled, make sure it is running DHCP as well as a Web server.

2. **Download** the Cumulus Linux installation file to the root directory of the Web server. Rename this file `onie-installer`.
3. Connect your host via Ethernet cable to the management Ethernet port of the switch.
4. Power on the switch. The switch downloads the ONIE image installer and boots it. You can watch the progress of the install in your terminal. After the installation finishes, the Cumulus Linux login prompt appears in the terminal window.



These steps describe a flexible unattended installation method. You should not need a console cable. A fresh install via ONIE using a local Web server should generally complete in less than 10 minutes.

You have more options for installing Cumulus Linux with ONIE. Read this [knowledge base article](#) to install Cumulus Linux using ONIE in the following ways:

- DHCP/Web server with and without DHCP options
- Web server without DHCP
- FTP or TFTP without a Web server
- Local file
- USB

ONIE supports many other discovery mechanisms using USB (copy the installer to the root of the drive), DHCPv6 and DHCPv4, and image copy methods including HTTP, FTP, and TFTP. For more information on these discovery methods, refer to the [ONIE documentation](#).

After installing Cumulus Linux, you are ready to:

- Log in to Cumulus Linux on the switch.
- Install the Cumulus Linux license.
- Configure Cumulus Linux. This quick start guide provides instructions on configuring switch ports and a loopback interface.

## Upgrading Cumulus Linux

If you already have Cumulus Linux installed on your switch and are upgrading to a maintenance release (X.Y.Z, like 2.5.1) from an earlier release in the same major and minor release family **only** (like 2.2.1 to 2.2.2, or 2.5.0 to 2.5.1), you can use various methods, including `apt-get`, to upgrade to the new version instead. See [Upgrading Cumulus Linux \(see page 17\)](#) for details.

## Configuring Cumulus Linux

When bringing up Cumulus Linux for the first time, the management port makes a DHCPv4 request. To determine the IP address of the switch, you can cross reference the MAC address of the switch with your DHCP server. The MAC address should be located on the side of the switch or on the box in which the unit was shipped.

## Login Credentials

The default installation includes one system account, *root*, with full system privileges, and one user account, *cumulus*, with sudo privileges. The *root* account password is set to null by default (which prohibits login), while the *cumulus* account is configured with this default password:

```
CumulusLinux!
```

In this quick start guide, you will use the *cumulus* account to configure Cumulus Linux.



For best security, you should change the default password (using the `passwd` command) before you configure Cumulus Linux on the switch.

All accounts except *root* are permitted remote SSH login; sudo may be used to grant a non-root account root-level access. Commands which change the system configuration require this elevated level of access.

For more information about sudo, read [Using sudo to Delegate Privileges \(see page 67\)](#).

## Serial Console Management

Users are encouraged to perform management and configuration over the network, either in band or out of band. Use of the serial console is fully supported; however, many customers prefer the convenience of network-based management.

Typically, switches will ship from the manufacturer with a mating DB9 serial cable. Switches with ONIE are always set to a 115200 baud rate.

## Wired Ethernet Management

Switches supported in Cumulus Linux always contain at least one dedicated Ethernet management port, which is named *eth0*. This interface is geared specifically for out-of-band management use. The management interface uses DHCPv4 for addressing by default. You can set a static IP address in the [/etc/network/interfaces](#) file:

```
auto eth0
iface eth0
    address 192.0.2.42/24
    gateway 192.0.2.1
```

## Configuring the Hostname and Time Zone

To change the hostname, modify the `/etc/hostname` and `/etc/hosts` files with the desired hostname and reboot the switch. First, edit `/etc/hostname`:

```
cumulus@switch:~$ sudo vi /etc/hostname
```

Then replace the 127.0.1.1 IP address in `/etc/hosts` with the new hostname:

```
cumulus@switch:~$ sudo vi /etc/hosts
```

Reboot the switch:

```
cumulus@switch:~$ sudo reboot
```

To update the time zone, update the `/etc/timezone` file with the [correct timezone](#), run `dpkg-reconfigure --frontend noninteractive tzdata`, then reboot the switch:

```
cumulus@switch:~$ sudo vi /etc/timezone
cumulus@switch:~$ sudo dpkg-reconfigure --frontend noninteractive tzdata
cumulus@switch:~$ sudo reboot
```



It is possible to change the hostname without a reboot via a script available on [Cumulus Networks GitHub site](#).

## Installing the License

Cumulus Linux is licensed on a per-instance basis. Each network system is fully operational, enabling any capability to be utilized on the switch with the exception of forwarding on switch panel ports. Only eth0 and console ports are activated on an unlicensed instance of Cumulus Linux. Enabling front panel ports requires a license.

You should have received a license key from Cumulus Networks or an authorized reseller. Here is a sample license key:

```
user@company.com|thequickbrownfoxjumpsoverthelazydog312
```

There are three ways to install the license onto the switch:

- Copy it from a local server. Create a text file with the license and copy it to a server accessible from the switch. On the switch, use the following command to transfer the file directly on the switch, then install the license file:

```
cumulus@switch:~$ scp user@my_server:/home/user/my_license_file.txt .
cumulus@switch:~$ sudo cl-license -i my_license_file.txt
```

- Copy the file to an HTTP server (not HTTPS), then reference the URL when you run `cl-license`:

```
cumulus@switch:~$ sudo cl-license -i <URL>
```

- Copy and paste the license key into the `cl-license` command:

```
cumulus@switch:~$ sudo cl-license -i  
<paste license key>  
^+d
```

Once the license is installed successfully, reboot the system:

```
cumulus@switch:~$ sudo reboot
```

After the switch reboots, all front panel ports will be active. The front panel ports are identified as switch ports, and show up as `swp1`, `swp2`, and so forth.

## Configuring 4x10G Port Configuration (Splitter Cables)

If you are using 4x10G DAC or AOC cables, edit the `/etc/cumulus/ports.conf` to enable support for these cables then [restart the `switchd` service \(see page 90\)](#) using the `sudo service switchd restart` command. For more details, see [Configuring Switch Port Attributes \(see page 108\)](#).

## Testing Cable Connectivity

By default, all data plane ports (every Ethernet port except the management interface, `eth0`) are disabled. To test cable connectivity, administratively enable a port using `ip link set <interface> up`:

```
cumulus@switch:~$ sudo ip link set swp1 up
```

Run the following bash script, as root, to administratively enable all physical ports:

```
cumulus@switch:~$ sudo su -  
cumulus@switch:~$ for i in /sys/class/net/*; do iface=`basename $i`; if [[  
$iface == swp* ]]; then ip link set $iface up; fi done
```

To view link status, use `ip link show`. The following examples show the output of a port in "admin down", "down" and "up" mode, respectively:

```
# Administratively Down
swp1: <BROADCAST,MULTICAST> mtu 1500 qdisc pfifo_fast state DOWN mode
DEFAULT qlen 1000

# Administratively Up but Layer 2 protocol is Down
swp1: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc pfifo_fast state
DOWN mode DEFAULT qlen 500

# Administratively Up, Layer 2 protocol is Up
swp1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP
mode DEFAULT qlen 500
```

## Configuring Switch Ports

### ***Layer 2 Port Configuration***

To configure a front panel port or create a bridge, edit the `/etc/network/interfaces` file. After saving the file, to activate the change, use the `ifup` command.

### ***Examples***

In the following configuration example, the front panel port `swp1` is placed into a bridge called `br0`:

```
auto br0
iface br0
    bridge-ports swp1
    bridge-stp on
```

To put a range of ports into a bridge, use the `glob` keyword. For example, add `swp1` through `swp10`, `swp12`, and `swp14` through `swp20` to `br0`:

```
auto br0
iface br0
    bridge-ports glob swp1-10 swp12 glob swp14-20
    bridge-stp on
```

To activate or apply the configuration to the kernel:

```
# First, check for typos:  
cumulus@switch:~$ sudo ifquery -a  
  
# Then activate the change if no errors are found:  
cumulus@switch:~$ sudo ifup -a
```

To view the changes in the kernel, use the `brctl` command:

```
cumulus@switch:~$ brctl show  
bridge name      bridge id          STP enabled    interfaces  
br0              8000.089e01cedcc2    yes           swp1
```



A script is available to generate a configuration that [places all physical ports in a single bridge](#).

## **Layer 3 Port Configuration**

To configure a front panel port or bridge interface as a Layer 3 port, edit the `/etc/network/interfaces` file.

In the following configuration example, the front panel port `swp1` is configured a Layer 3 access port:

```
auto swp1  
iface swp1  
    address 10.1.1.1/30
```

To add an IP address to a bridge interface, include the address under the `iface` configuration in `/etc/network/interfaces`:

```
auto br0  
iface br0  
    address 10.2.2.1/24  
    bridge-ports glob swp1-10 swp12 glob swp14-20  
    bridge-stp on
```

To activate or apply the configuration to the kernel:

```
# First check for typos:  
cumulus@switch:~$ sudo ifquery -a
```

```
# Then activate the change if no errors are found:  
cumulus@switch:~$ sudo ifup -a
```

To view the changes in the kernel use the `ip addr show` command:

```
br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue state UP  
link/ether 00:02:00:00:00:28 brd ff:ff:ff:ff:ff:ff  
inet 10.2.2.1/24 scope global br0  
  
swp1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue state UP  
link/ether 44:38:39:00:6e:fe brd ff:ff:ff:ff:ff:ff  
inet 10.1.1.1/30 scope global swp1
```

## Configuring a Loopback Interface

Cumulus Linux has a loopback preconfigured in `/etc/network/interfaces`. When the switch boots up, it has a loopback interface, called `lo`, which is up and assigned an IP address of 127.0.0.1.

To see the status of the loopback interface (`lo`), use the `ip addr show lo` command:

```
cumulus@switch:~$ ip addr show lo  
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 16436 qdisc noqueue state UNKNOWN  
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00  
    inet 127.0.0.1/8 scope host lo  
        inet6 ::1/128 scope host  
            valid_lft forever preferred_lft forever
```

Note that the loopback is up and is assigned an IP address of 127.0.0.1.

To add an IP address to a loopback interface, add it directly under the `iface lo inet loopback` definition in `/etc/network/interfaces`:

```
auto lo  
iface lo inet loopback  
    address 10.1.1.1
```



If an IP address is configured without a mask, as shown above, the IP address becomes a /32. So, in the above case, 10.1.1.1 is actually 10.1.1.1/32.

Multiple loopback addresses can be configured by adding additional address lines:

```
auto lo
iface lo inet loopback
    address 10.1.1.1
    address 172.16.2.1/24
```

# Installation, Upgrading and Package Management

A Cumulus Linux switch can have up to two images of the operating system installed. This section discusses installing new and updating existing Cumulus Linux disk images, and configuring those images with additional applications (via packages) if desired.

Zero touch provisioning is a way to quickly deploy and configure new switches in a large-scale environment.

## Managing Cumulus Linux Disk Images

The Cumulus Linux operating system resides on a switch as a *disk image*. Switches running Cumulus Linux can be configured with 2 separate disk images. This section discusses how to manage them including installation and upgrading.

### Contents

(Click to expand)

- [Contents \(see page 16\)](#)
- [Commands \(see page 16\)](#)
- [Installing a New Cumulus Linux Image \(see page 16\)](#)
- [Upgrading Cumulus Linux \(see page 17\)](#)
- [Understanding Image Slots \(see page 17\)
  - \[PowerPC vs x86 vs ARM Switches \\(see page 17\\)\]\(#\)
  - \[PowerPC Image Slots \\(see page 18\\)\]\(#\)
  - \[x86 and ARM Image Slots \\(see page 19\\)\]\(#\)](#)
- [Reverting an Image to its Original Configuration \(PowerPC Only\) \(see page 21\)](#)
- [Reprovisioning the System \(Restart Installer\) \(see page 22\)](#)
- [Uninstalling All Images and Removing the Configuration \(see page 22\)](#)
- [Booting into Rescue Mode \(see page 23\)](#)
- [Inspecting Image File Contents \(see page 24\)](#)
- [Useful Links \(see page 25\)](#)

### Commands

- [apt-get](#)
- [cl-img-install](#)
- [cl-img-select](#)
- [cl-img-clear-overlay](#)
- [cl-img-pkg](#)

## ***Installing a New Cumulus Linux Image***

For details, read the chapter, [Installing a New Cumulus Linux Image \(see page 25\)](#).

## ***Upgrading Cumulus Linux***

There are two ways you can upgrade Cumulus Linux:

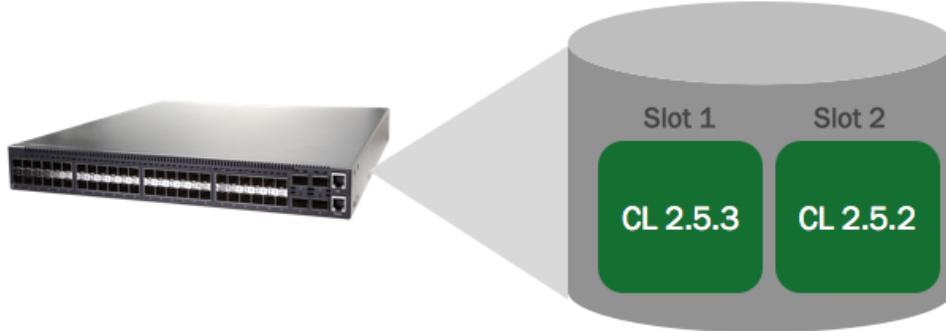
- Perform a binary (full image) install of the new version, running `cl-img-install` on the switch
- Upgrade only the changed packages, using `apt-get update` and `apt-get dist-upgrade`

The entire upgrade process is described in [Upgrading Cumulus Linux \(see page 35\)](#).

## ***Understanding Image Slots***

Cumulus Linux uses the concept of *image slots* to manage two separate Cumulus Linux images. The important terminology for the slots is as follows:

- **Active image slot:** The currently running image slot.
- **Primary image slot:** The image slot that is selected for the next boot. Often this is the same as the active image slot.
- **Alternate image slot:** The inactive image slot, **not** selected for the next boot.



To identify which slot is active, which slot is the primary, and which slot is alternate use the `cl-img-select` command:

```
cumulus@switch$ sudo cl-img-select
active => slot 1 (primary): 2.5.3-c4e83ad-201506011818-build
           slot 2 (alt     ): 2.5.2-727a0c6-201504132125-build
```

The above switch is currently running 2.5.3 as indicated by the **active**. When the switch is rebooted, it will boot into slot 1, as indicated by **primary**. The **alternate** slot is running Cumulus Linux 2.5.2 and won't be booted into unless the user selects it.

## ***PowerPC vs x86 vs ARM Switches***

The characteristics of the image slots vary, based on whether your switch is on a PowerPC, ARM or x86 platform. You can easily determine which platform the switch is on by using the `uname -m` command.

For example, on a PowerPC platform, `uname -m` outputs `ppc`:

```
cumulus@PPCswitch$ uname -m
ppc
```

While on an x86 platform, `uname -m` outputs `x86_64`:

```
cumulus@x86switch$ uname -m
x86_64
```

While on an ARM platform, `uname -m` outputs `armv7l`:

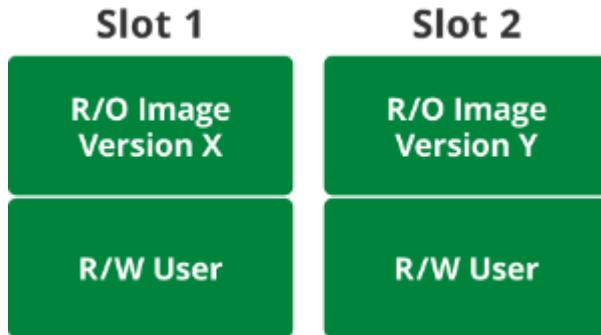
```
cumulus@ARMswitch$ uname -m
armv7l
```

You can also visit the HCL ([hardware compatibility list](#)) to look at your hardware to determine the processor type.

## PowerPC Image Slots

Read more about PowerPC image slots

On the PowerPC platform, each image slot consists of a read-only Cumulus Linux base image overlaid with a read-write user area, as shown in the following diagram:



Files you edit and create reside in the read-write user overlay. This also includes any additional software you install on top of Cumulus Linux. After an install, the user overlay is empty.

## PowerPC Image Slot Overlay Detailed Information

The root directory of an image slot on a PowerPC system is created using an `overlayfs` file system. The lower part of the overlay is a *read-only* `squashfs` file system containing the base Cumulus Linux image. The upper part of the overlay is a *read-write* directory containing all the user modifications.

The following table describes the mount points and directories used to create the overlay for image slots 1 and 2.

Slot Number	R/O squashfs device	R/O mount point	R/W block device	R/W directory
1	/dev/sysroot1	/mnt/root-ro	/dev/overlay_rw	/mnt/root-rw/config1

Slot Number	R/O squashfs device	R/O mount point	R/W block device	R/W directory
2	/dev/sysroot2	/mnt/root-ro	/dev/overlay_rw	/mnt/root-rw/config2



A single read-write partition provides separate read-write directories for the upper part of the overlay. The lower part of the overlay is a **partition**, while the upper part is a **directory**.

The following table describes all the interesting mount points.

Mount Point	File System	Purpose
/mnt/root-ro	squashfs	Contains the read-only base Cumulus Linux image.
/mnt/root-rw	ext2	Contains the read-write user directories for the overlay.
/	overlays	The union of /mnt/root-ro and /mnt/root-rw/config1 (or config2).
/mnt/persist	ext2	Contains the persistent user configuration applied to each image slot.
/mnt/initramfs	tmpfs	Contains the <code>initramfs</code> used at boot. Needed during shutdown.

## x86 and ARM Image Slots

Read more about x86 image slots

Unlike PowerPC-based switches, there is no overlay for an x86-based or ARM-based switch; instead each slot is a logical volume in the physical partition, which you can manage with [LVM](#).

When you install Cumulus Linux on an x86 or ARM switch, the following entities are created on the disk:

- A disk partition using an ext4 file system that contains three logical volumes: two logical volumes named `sysroot1` and `sysroot2`, and the `/mnt/persist` logical volume. The logical volumes represent the Cumulus Linux image slots, so `sysroot1` is slot 1 and `sysroot2` is slot 2. `/mnt/persist` is where you store your [persistent configuration](#) (see page ).
- A boot partition, shared by the logical volumes. Each volume mounts this partition as `/boot`.

## Managing Slot Sizes

As space in a slot is used, you may need to increase the size of the root filesystem by increasing the size of the corresponding logical volume. This section shows you how to check current utilization and expand the filesystem as needed.

1. Check utilization on the root filesystem with the `df -h /` command. In the following example, filesystem utilization is 16%:

```
cumulus@switch$ df -h /
Filesystem           Size  Used Avail Use% Mounted on
/dev/disk/by-uuid/64650289-cebf-4849-91ae-a34693fce2f1  4.0G
579M   3.2G  16%  /
```

2. To increase available space in the root filesystem, first use the `vgs` command to check the available space in the volume group. In this example, there is 6.34 Gigabytes of free space available in the volume group CUMULUS:

```
cumulus@switch$ sudo vgs
VG      #PV #LV #SN Attr   VSize   VFree
CUMULUS  1    3    0 wz--n- 14.36g 6.34g
```

3. Once you confirm the available space, determine the number of the currently active slot using `cl-img-select`.

```
cumulus@switch$ sudo cl-img-select | grep active
active => slot 1 (primary): 2.5.0-199c587-201501081931-build
```

`cl-img-select` indicates slot number 1 is active.

4. Resize the slot with the `lvresize` command. The following example increases slot size by 20 percent of total available space. Replace the "#" character in the example with the active slot number from the last step.

```
cumulus@switch$ sudo lvresize -l +20%FREE CUMULUS/SYSROOT#
Extending logical volume SYSROOT# to 5.27 GiB
Logical volume SYSROOT# successfully resized
```



The use of + is very important with the `lvresize` command. Issuing `lvresize` without the + results in the logical volume size being set directly to the specified size, rather than extended.

5. Once the slot has been extended, use the `resize2fs` command to expand the filesystem to fit the new space in the slot. Again, replace the "#" character in the example with the active slot number.

```
cumulus@switch$ sudo resize2fs /dev/CUMULUS/SYSROOT#
resize2fs 1.42.5 (29-Jul-2012)
```

```
Filesystem at /dev/CUMULUS/SYSROOT# is mounted on /; on-line
resizing required
old_desc_blocks = 1, new_desc_blocks = 1
Performing an on-line resize of /dev/CUMULUS/SYSROOT# to 1381376
(4k) blocks.
The filesystem on /dev/CUMULUS/SYSROOT# is now 1381376 blocks long.
```

## **Accessing the Alternate Image Slot on x86 and ARM Platforms**

It may be useful to access the content of the alternate slot to retrieve the configuration or logs.



cl-img-install fails while the alternate slot is mounted. It is important to unmount the alternate slot as shown in step 4 below when done.

1. Determine which slot is the alternate with cl-img-select:

```
cumulus@switch$ sudo cl-img-select
active => slot 1 (primary): 2.5.3-c4e83ad-201506011818-build
          slot 2 (alt      ): 2.5.2-727a0c6-201504132125-build
```

This output indicates slot 2 is the alternate slot.

2. Create a mount point for the alternate slot:

```
cumulus@switch$ sudo mkdir /mnt/alt
```

3. Mount the alternate slot to the mount point:

```
cumulus@switch$ sudo mount /dev/mapper/CUMULUS-SYSROOT# /mnt/alt
```

Where # is the number of the alternate slot.

The alternate slot is now accessible under /mnt/alt.

4. Unmount the mount point /mnt/alt when done.

```
cumulus@switch$ cd /
cumulus@switch$ sudo umount /mnt/alt/
```

## **Reverting an Image to its Original Configuration (PowerPC Only)**

On PowerPC-based systems, you may want to clear out the read-write user overlay area. Perhaps something was misconfigured, or was deleted by mistake, or some unneeded software was installed.

You can purge the read-write overlay using the `cl-img-clear-overlay` command, passing the slot number as an argument. For example, to purge the read-write overlay for image slot 2, run:

```
cumulus@switch:~$ sudo cl-img-clear-overlay 2
Success: Overlay configuration 2 will be re-initialized during the next
reboot.
```



You must reboot the switch to complete the purge.

## ***Reprovisioning the System (Restart Installer)***

You can reprovision the system, wiping out the contents of both image slots and `/mnt/persist`.

To initiate the provisioning and installation process, use `cl-img-select -i`:

```
cumulus@switch:~$ sudo cl-img-select -i
WARNING:
WARNING: Operating System install requested.
WARNING: This will wipe out all system data.
WARNING:
Are you sure (y/N)? y
Enabling install at next reboot...done.
Reboot required to take effect.
```



A reboot is required for the reinstall to begin.



If you change your mind, you can cancel a pending reinstall operation by using `cl-img-select -c`:

```
cumulus@switch:~$ sudo cl-img-select -c
 Cancelling pending install at next reboot...done.
```

## ***Uninstalling All Images and Removing the Configuration***

To remove all installed images and configurations, returning the switch to its factory defaults, use `cl-img-select -k`:

```
cumulus@switch:~$ sudo cl-img-select -k
WARNING:
WARNING: Operating System uninstall requested.
WARNING: This will wipe out all system data.
WARNING:
Are you sure (y/N)? y
Enabling uninstall at next reboot...done.
Reboot required to take effect.
```



A reboot is required for the uninstall to begin.



If you change your mind you can cancel a pending uninstall operation by using `cl-img-select -c`:

```
cumulus@switch:~$ sudo cl-img-select -c
 Cancelling pending uninstall at next reboot...done.
```

## Booting into Rescue Mode

If your system becomes broken in some way, you may be able to correct things by booting into ONIE rescue mode. In rescue mode, the file systems are unmounted and you can use various Cumulus Linux utilities to try and fix the problem.

To reboot the system into the ONIE rescue mode, use `cl-img-select -r`:

```
cumulus@switch:~$ sudo cl-img-select -r
WARNING:
WARNING: Rescue boot requested.
WARNING:
Are you sure (y/N)? y
Enabling rescue at next reboot...done.
Reboot required to take effect.
```



A reboot is required to boot into rescue mode.



If you change your mind you can cancel a pending rescue boot operation by using `cl-img-select -c`:

```
cumulus@switch:~$ sudo cl-img-select -c
Cancelling pending rescue at next reboot...done.
```

## Inspecting Image File Contents

From a running system you can display the contents of a Cumulus Linux image file using `cl-img-pkg -d`:

```
cumulus@switch:~$ sudo cl-img-pkg -d /var/lib/cumulus/installer/onie-
installer
Verifying image checksum ... OK.
Preparing image archive ... OK.
Control File Contents
=====
Description: Cumulus Linux
OS-Release: 2.1.0-0556262-201406101128-NB
Architecture: amd64
Date: Tue, 10 Jun 2014 11:44:28 -0700
Installer-Version: 1.2
Platforms: im_n29xx_t40n mlx_sx1400_i73612 dell_s6000_s1220
Homepage: http://www.cumulusnetworks.com/

Data Archive Contents
=====
  128 2014-06-10 18:44:26 file.list
    44 2014-06-10 18:44:27 file.list.sha1
  104276331 2014-06-10 18:44:27 sysroot-internal.tar.gz
    44 2014-06-10 18:44:27 sysroot-internal.tar.gz.sha1
  5391348 2014-06-10 18:44:26 vmlinuz-initrd.tar.xz
    44 2014-06-10 18:44:27 vmlinuz-initrd.tar.xz.sha1
cumulus@switch:~$
```

You can also extract the image files to the current directory with the `-e` option:

```
cumulus@switch:~$ sudo cl-img-pkg -e /var/lib/cumulus/installer/onie-
installer
Verifying image checksum ... OK.
Preparing image archive ... OK.
file.list
file.list.sha1
sysroot-internal.tar.gz
sysroot-internal.tar.gz.sha1
vmlinuz-initrd.tar.xz
vmlinuz-initrd.tar.xz.sha1
Success: Image files extracted OK.
cumulus@switch:~$ sudo ls -l
total 107120
```

```
-rw-r--r-- 1 1063 3000      128 Jun 10 18:44 file.list
-rw-r--r-- 1 1063 3000      44 Jun 10 18:44 file.list.sha1
-rw-r--r-- 1 1063 3000 104276331 Jun 10 18:44 sysroot-internal.tar.gz
-rw-r--r-- 1 1063 3000      44 Jun 10 18:44 sysroot-internal.tar.gz.
sha1
-rw-r--r-- 1 1063 3000 5391348 Jun 10 18:44 vmlinuz-initrd.tar.xz
-rw-r--r-- 1 1063 3000      44 Jun 10 18:44 vmlinuz-initrd.tar.xz.
sha1
```

## Useful Links

- Open Network Install Environment (ONIE) Home Page

## Installing a New Cumulus Linux Image

Before you install Cumulus Linux, the switch can be in two different states:

- The switch has no image on it (so the switch is only running [ONIE](#)) or you desire or require a clean installation. In this case, you can install Cumulus Linux in one of the following ways, using:
  - DHCP/a Web server with DHCP options (see page 26)
  - DHCP/a Web server without DHCP options (see page 27)
  - A Web server with no DHCP (see page 27)
  - FTP or TFTP without a Web server (see page 28)
  - Local file installation (see page 28)
  - USB (see page 28)
- The switch already has Cumulus Linux installed on it, so you only need to [upgrade it](#) (see page 35)



[ONIE](#) is an open source project, equivalent to PXE on servers, that enables the installation of network operating systems (NOS) on bare metal switches.

## Understanding these Examples

- This sections in this chapter are ordered from the most repeatable to the least repeatable methods. For instance, DHCP can scale to hundreds of switch installs with zero manual input, compared to something like USB installs. Installing via USB is fine for a single switch here and there but is not scalable.
- You can name your Cumulus Linux installer binary using any of the [ONIE naming schemes mentioned here](#).
- In the examples below, [PLATFORM] can be any supported Cumulus Linux platform, such as *x86-64*, *arm*, or *powerpc*.

## Contents

Click to expand...

- Understanding these Examples (see page 25)
- Contents (see page 25)

- Installing via a DHCP/Web Server Method with DHCP Options (see page 26)
- Installing via a DHCP/Web Server Method without DHCP Options (see page 27)
- Installing via a Web Server with no DHCP (see page 27)
- Installing via FTP or TFTP without a Web Server (see page 28)
- Installing via a Local File (see page 28)
- Installing via USB (see page 28)
  - Preparing for USB Installation (see page 28)
  - Instructions for x86 Platforms (see page 30)
  - Instructions for PowerPC and ARM Platforms (see page 33)
- Installing a New Image when Cumulus Linux Is already Installed (see page 35)

## ***Installing via a DHCP/Web Server Method with DHCP Options***

Installing Cumulus Linux in this manner is as simple as setting up a DHCP/Web server on your laptop and connecting the eth0 management port of the switch to your laptop.

Once you connect the cable, the installation proceeds as follows:

1. The bare metal switch boots up and asks for an address (DHCP request).
2. The DHCP server acknowledges and responds with DHCP option 114 and the location of the installation image.
3. ONIE downloads the Cumulus Linux binary, installs and reboots.
4. Success! You are now running Cumulus Linux.



The most common method is for you to send DHCP option 114 with the entire URL to the Web server (this could be the same system). However, there are many other ways to use DHCP even if you don't have full control over DHCP. See the [ONIE user guide](#) for help.

Here's an example DHCP configuration with an ISC DHCP server:

```
subnet 172.0.24.0 netmask 255.255.255.0 {
  range 172.0.24.20 172.0.24.200;
  option default-url = "http://172.0.24.14/onie-installer";
}
```

Here's an example DHCP configuration with dnsmasq (static address assignment):

```
dhcp-host=sw4,192.168.100.14,6c:64:1a:00:03:ba,set:sw4  
dhcp-option>tag:sw4,114,"http://roz.rtplab.test/onie-installer"
```

Don't have a Web server? There is a [free Apache example](#) you can utilize.

## **Installing via a DHCP/Web Server Method without DHCP Options**

If you have a laptop on same network and the switch can pull DHCP from the corporate network, but you cannot modify DHCP options (maybe it's controlled by another team), do the following:

1. Place the Cumulus Linux binary in a directory on the Web server.
2. Run the `onie-nos-install` command manually, since DHCP options can't be modified:

```
ONIE:/ #onie-nos-install http://10.0.1.251/path/to/cumulus-install-[PLATFORM].bin
```

## **Installing via a Web Server with no DHCP**

Use the following method if your laptop is on the same network as the switch eth0 interface but no DHCP server is available.

One thing to note is ONIE is in [\*discovery mode\*](#), so if you are setting a static IPv4 address for the eth0 management port, you need to disable discovery mode or else ONIE may get confused.

1. To disable discovery mode, run:

```
onie# onie-discovery-stop
```

or, on older ONIE versions if that command isn't supported:

```
onie# /etc/init.d/discover.sh stop
```

2. Assign a static address to eth0 via ONIE (using `ip addr add`):

```
ONIE:/ #ip addr add 10.0.1.252/24 dev eth0
```

3. Place the Cumulus Linux installer image in a directory on your Web server.
4. Run the `onie-nos-install` command manually since there are no DHCP options:

```
ONIE:/ #onie-nos-install http://10.0.1.251/path/to/cumulus-install-[PLATFORM].bin
```

## Installing via FTP or TFTP without a Web Server

1. Set up DHCP or static addressing for eth0, as in the examples above.
2. If you are utilizing static addressing, disable ONIE discovery mode.
3. Place the Cumulus Linux installer image into a TFTP or FTP directory.
4. If you are not utilizing DHCP options, run one of the following commands (`tftp` for TFTP or `ftp` for FTP):

```
ONIE# onie-nos-install ftp://local-ftp-server/cumulus-install-[PLATFORM].bin
```

```
ONIE# onie-nos-install tftp://local-tftp-server/cumulus-install-[PLATFORM].bin
```

## Installing via a Local File

1. Set up DHCP or static addressing for eth0, as in the examples above.
2. If you are utilizing static addressing, disable ONIE discovery mode.
3. Use `scp` to copy the Cumulus Linux binary to the switch.  
Note: Windows users can use [WinScp](#).
4. Run the following command:

```
ONIE# onie-nos-install /path/to/local/file/cumulus-install-[PLATFORM].bin
```

## Installing via USB

Following the steps below produces a clean installation of Cumulus Linux. This wipes out all pre-existing configuration files that may be present on the switch. Instructions are offered for x86, ARM and PowerPC platforms, and also cover the installation of a license after the software installation.



Make sure to [back up \(see page 35\)](#) any important configuration files that you may need to restore the configuration of your switch after the installation finishes.

## Preparing for USB Installation

1. Download the appropriate Cumulus Linux image for your x86, ARM or PowerPC platform from the [Cumulus Networks Downloads page](#).
2. Prepare your flash drive by formatting in one of the supported formats: FAT32, vFAT or EXT2.

Optional: Preparing a USB Drive inside Cumulus Linux



It is possible that you could severely damage your system with the following utilities, so please use caution when performing the actions below!

- a. Insert your flash drive into the USB port on the switch running Cumulus Linux and log in to the switch.
- b. Determine and note at which device your flash drive can be found by using output from `cat /proc/partitions` and `sudo fdisk -l [device]`. For example, `sudo fdisk -l /dev/sdb`.



These instructions assume your USB drive is the `/dev/sdb` device, which is typical if the USB stick was inserted after the machine was already booted. However, if the USB stick was plugged in during the boot process, it is possible the device could be `/dev/sda`. Make sure to modify the commands below to use the proper device for your USB drive!

- c. Create a new partition table on the device:

```
sudo parted /dev/sdb mklabel msdos
```



The `parted` utility should already be installed. However, if it is not, install it with:  
`sudo apt-get install parted`

- d. Create a new partition on the device:

```
sudo parted /dev/sdb -a optimal mkpart primary 0% 100%
```

- e. Format the partition to your filesystem of choice using ONE of the examples below:

```
sudo mkfs.ext2 /dev/sdb1
sudo mkfs.msdos -F 32 /dev/sdb1
sudo mkfs.vfat /dev/sdb1
```

To use `mkfs.msdos` or `mkfs.vfat`, you need to install the `dosfstools` package from the [Debian software repositories](#) (step 3 here shows you how to add repositories from Debian), as they are not included by default.

- f. To continue installing Cumulus Linux, mount the USB drive in order to move files to it.

```
sudo mkdir /mnt/usb  
sudo mount /dev/sdb1 /mnt/usb
```

3. Copy the image and license files over to the flash drive and rename the image file to:

- `onie-installer_x86-64`, if installing on an x86 platform
- `onie-installer-powerpc`, if installing on a PowerPC platform
- `onie-installer-arm`, if installing on an ARM platform



You can also use any of the [ONIE naming schemes mentioned here](#).



When using a Mac or Windows computer to rename the installation file the file extension may still be present. Make sure to remove the file extension otherwise ONIE will not be able to detect the file!

4. Insert the USB stick into the switch, then continue with the appropriate instructions below for your x86, ARM or PowerPC platform.

## ***Instructions for x86 Platforms***

Click to expand x86 instructions...

1. Prepare the switch for installation:

- If the switch is offline, connect to the console and power on the switch.
- If the switch is already online in Cumulus Linux, connect to the console and reboot the switch into the ONIE environment with the `sudo cl-img-select -i` command, followed by `sudo reboot`. Then skip to step 4 below.
- If the switch is already online in ONIE, use the `reboot` command.



SSH sessions to the switch get dropped after this step. To complete the remaining instructions, connect to the console of the switch. Cumulus Linux switches display their boot process to the console, so you need to monitor the console specifically to complete the next step.

2. Monitor the console and select the ONIE option from the first GRUB screen shown below.

```

GNU GRUB  version 1.99-27+deb7u2

+-----+
| Cumulus Linux 2.5.3a-3b46bef-201509041633-build - slot 1
| Cumulus Linux 2.5.3a-3b46bef-201509041633-build - slot 1 (recovery mode)
| Cumulus Linux 2.5.3a-3b46bef-201509041633-build - slot 2
| Cumulus Linux 2.5.3a-3b46bef-201509041633-build - slot 2 (recovery mode)
| ONIE
+-----+

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.

```

3. Cumulus Linux on x86 uses GRUB chainloading to present a second GRUB menu specific to the ONIE partition. No action is necessary in this menu to select the default option *ONIE: Install OS*.

```

GNU GRUB  version 2.02~beta2+e4a1fe391

+-----+
| *ONIE: Install OS
| ONIE: Rescue
| ONIE: Uninstall OS
| ONIE: Update ONIE
| ONIE: Embed ONIE
+-----+

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.

```

4. At this point, the USB drive should be automatically recognized and mounted. The image file should be located and automatic installation of Cumulus Linux should begin. Here is some sample output:

```

ONIE: OS Install Mode ...
Version : quanta_common_rangeley-2014.05.05-6919d98-201410171013
Build Date: 2014-10-17T10:13+0800
Info: Mounting kernel filesystems... done.
Info: Mounting LABEL=ONIE-BOOT on /mnt/onie-boot ...
initializing eth0...
scsi 6:0:0:0: Direct-Access SanDisk Cruzer Facet 1.26 PQ: 0
ANSI: 6
sd 6:0:0:0: [sdb] 31266816 512-byte logical blocks: (16.0 GB/14.9
GiB)
sd 6:0:0:0: [sdb] Write Protect is off

```

```

sd 6:0:0:0: [sdb] Write cache: disabled, read cache: enabled,
doesn't support DPO or FUA
sd 6:0:0:0: [sdb] Attached SCSI disk
<...snip...
ONIE: Executing installer: file://dev/sdb1/onie-installer-x86_64
Verifying image checksum ... OK.
Preparing image archive ... OK.
Dumping image info...
Control File Contents
=====
Description: Cumulus Linux
OS-Release: 2.5.3a-3b46bef-201509041633-build
Architecture: amd64
Date: Fri, 04 Sep 2015 17:10:30 -0700
Installer-Version: 1.2
Platforms: accton_as5712_54x accton_as6712_32x
mlx_sx1400_i73612 dell_s6000_s1220 dell_s4000_c2338
dell_s3000_c2338 cel_redstone_xp cel_smallstone_xp cel_pebble
quanta_panther quanta_ly8_rangeley quanta_ly6_rangeley
quanta_ly9_rangeley
Homepage: http://www.cumulusnetworks.com/
  
```

5. After installation completes, the switch automatically reboots into the newly installed instance of Cumulus Linux.
6. Determine and note at which device your flash drive can be found by using output from `cat /proc/partitions` and `sudo fdisk -l [device]`. For example, `sudo fdisk -l /dev/sdb`.



These instructions assume your USB drive is the `/dev/sdb` device, which is typical if the USB stick was inserted after the machine was already booted. However, if the USB stick was plugged in during the boot process, it is possible the device could be `/dev/sda`. Make sure to modify the commands below to use the proper device for your USB drive!

7. Create a mount point to mount the USB drive to:

```
sudo mkdir /mnt/mountpoint
```

8. Mount the USB drive to the newly created mount point:

```
sudo mount /dev/sdb1 /mnt/mountpoint
```

9. Install your license file with the `cl-license` command:

```
sudo cl-license -i /mnt/mountpoint/license.txt
```

10. Check that your license is installed with the `cl-license` command.

11. Reboot the switch to utilize the new license.

```
sudo reboot
```

## **Instructions for PowerPC and ARM Platforms**

Click to expand PowerPC instructions...

1. Prepare the switch for installation:

- If the switch is offline, connect to the console and power on the switch.
- If the switch is already online in Cumulus Linux, connect to the console and reboot the switch into the ONIE environment with the `sudo cl-img-select -i` command, followed by `sudo reboot`. Then skip to step 4.
- If the switch is already online in ONIE, use the `reboot` command.



SSH sessions to the switch get dropped after this step. To complete the remaining instructions, connect to the console of the switch. Cumulus Linux switches display their boot process to the console, so you need to monitor the console specifically to complete the next step.

2. Interrupt the normal boot process before the countdown (shown below) completes. Press any key to stop the autobooting.

```
U-Boot 2013.01-00016-gd6bf4a9-dirty (Feb 14 2014 - 16:30:46)
Accton: 1.4.0.5
CPU0: P2020, Version: 2.1, (0x80e20021)
Core: E500, Version: 5.1, (0x80211051)
Clock Configuration:
    CPU0:1200 MHz, CPU1:1200 MHz,
    CCB:600 MHz,
    DDR:400 MHz (800 MT/s data rate) (Asynchronous), LBC:37.500 MHz
    L1: D-cache 32 kB enabled
        I-cache 32 kB enabled
    <...snip...>
    USB: USB2513 hub OK
    Hit any key to stop autoboot: 0
```

3. A command prompt appears, so you can run commands. Execute the following command:

```
run onie_bootcmd
```

4. At this point the USB drive should be automatically recognized and mounted. The image file should be located and automatic installation of Cumulus Linux should begin. Here is some sample output:

```

Loading Open Network Install Environment ...
Platform: powerpc-as6701_32x-r0
Version : 1.6.1.3
WARNING: adjusting available memory to 30000000
## Booting kernel from Legacy Image at ec040000 ...
  Image Name: as6701_32x.1.6.1.3
  Image Type: PowerPC Linux Multi-File Image (gzip compressed)
  Data Size: 4456555 Bytes = 4.3 MiB
  Load Address: 00000000
  Entry Point: 00000000
  Contents:
    Image 0: 3738543 Bytes = 3.6 MiB
    Image 1: 706440 Bytes = 689.9 KiB
    Image 2: 11555 Bytes = 11.3 KiB
  Verifying Checksum ... OK
## Loading init Ramdisk from multi component Legacy Image at
ec040000 ...
## Flattened Device Tree from multi component Image at EC040000
  Booting using the fdt at 0xec47d388
  Uncompressing Multi-File Image ... OK
  Loading Ramdisk to 2ff53000, end 2ffff788 ... OK
  Loading Device Tree to 03ffa000, end 03ffd22 ... OK
<...snip...
ONIE: Starting ONIE Service Discovery
ONIE: Executing installer: file://dev/sdb1/onie-installer-powerpc
Verifying image checksum ... OK.
Preparing image archive ... OK.
Dumping image info...
Control File Contents
=====
Description: Cumulus Linux
OS-Release: 2.5.3a-3b46bef-201509041633-build
Architecture: powerpc
Date: Fri, 04 Sep 2015 17:08:35 -0700
Installer-Version: 1.2
Platforms: accton_as4600_54t, accton_as6701_32x, accton_5652,
accton_as5610_52x, dni_6448, dni_7448, dni_c7448n, cel_kennisis,
cel_redstone, cel_smallstone, cumulus_p2020, quanta_lb9,
quanta_ly2, quanta_ly2r, quanta_ly6_p2020
Homepage: http://www.cumulusnetworks.com/

```

5. After installation completes, the switch automatically reboots into the newly installed instance of Cumulus Linux.
6. Determine and note at which device your flash drive can be found by using output from `cat /proc/partitions` and `sudo fdisk -l [device]`. For example, `sudo fdisk -l /dev/sdb`.



These instructions assume your USB drive is the `/dev/sdb` device, which is typical if the USB stick was inserted after the machine was already booted. However, if the USB stick was plugged in during the boot process, it is possible the device could be `/dev/sda`. Make sure to modify the commands below to use the proper device for your USB drive!

7. Create a mount point to mount the USB drive to:

```
sudo mkdir /mnt/mountpoint
```

8. Mount the USB drive to the newly created mount point:

```
sudo mount /dev/sdb1 /mnt/mountpoint
```

9. Install your license file with the cl-license command:

```
sudo cl-license -i /mnt/mountpoint/license.txt
```

10. Check that your license is installed with the cl-license command.

11. Reboot the switch to utilize the new license.

```
sudo reboot
```

## ***Installing a New Image when Cumulus Linux Is already Installed***

Follow these upgrade steps for both major and minor releases, where:

- A major release upgrade is 2.X.X to 3.X.X (e.g. 1.5.1 to 2.5.0)
- A minor release upgrade is X.2.X to X.3.X (e.g. 2.2.0 to 2.5.5)

For more information, see [Upgrading Cumulus Linux \(see page 41\)](#).

## ***Upgrading Cumulus Linux***

Cumulus Networks software melds the Linux host world with the networking devices world. Each world comes with its own paradigm on how to upgrade software. Before we discuss the various ways to upgrade Cumulus Linux switches, let's review the general considerations and strategies used to upgrade network devices and Linux hosts.

## **Contents**

Click to expand...

- [Contents \(see page 35\)](#)
- [Upgrades: Comparing the Network Device Worldview vs. the Linux Host Worldview \(see page 36\)](#)
  - [Manual vs. Automated Configuration \(see page 36\)](#)
  - [Locations of Configuration Data vs. Executables \(see page 36\)](#)
  - [Pre-deployment Testing of Production Environments \(see page 37\)](#)
  - [Upgrade Procedure \(see page 37\)](#)

- Rollback Procedure (see page 37)
- Third Party Packages (see page 38)
- Upgrading Cumulus Linux Devices: Strategies and Processes (see page 38)
  - Automated Configuration Is Preferred over Manual Configuration (see page 38)
  - Out-of-Band Management Is Worth the Investment (see page 38)
  - Understanding the Locations of Configuration Data for Management, Migration, and Backup (see page 38)
  - Pre-Deployment Testing of New Releases Is Advised and Enabled (see page 40)
- Upgrading Cumulus Linux: Choosing between a Binary Install vs. Package Upgrade (see page 41)
  - Upgrading via Binary Install (cl-img-install) (see page 41)
  - Upgrading Using Package Installs (apt-get update && apt-get dist-upgrade) (see page 43)
- Rolling Back a Cumulus Linux Installation (see page 45)
  - Rolling Back after Using Binary Install (see page 45)
  - Rolling Back after Using Package Install (see page 45)
- Third Party Package Considerations (see page 45)
- Caveats while Upgrading Cumulus Linux 2.5.x (see page 46)

## ***Upgrades: Comparing the Network Device Worldview vs. the Linux Host Worldview***

### ***Manual vs. Automated Configuration***

Historically, *network devices* were configured in place, and most network devices required customized configurations, which led predominantly to configuring the hardware manually. A lack of standardization between vendors, device types, and device roles hampered the development of APIs and automation tools. However, in the case of very large data centers, configurations became uniform and repeatable, and therefore scriptable. Some larger enterprises had to develop their own custom scripts to roll out data center network configurations. Virtually no industry-standard provisioning tools existed.

In contrast to data center network devices, *Linux hosts* in the data center number in the thousands and tend to have similar configurations. This increased scale led Linux sysadmins long ago to move to common tools to automate installation and configuration, since manually installing and configuring hosts did not work at the scale of a data center. Nearly all tasks are done via commonly available provisioning and orchestration tools.

### ***Locations of Configuration Data vs. Executables***

*Network devices* generally separate configuration data from the executable code. On bootup, the executable code looks into a different file system and retrieves the configuration file or files, parses the text and uses that data to configure the software options for each software subsystem. The model is very centralized, with the executables generally being packaged together, and the configuration data following a set of rules that can be read by a centralized parser. Each vendor controls the configuration format for the entire box, since each vendor generally supports only their own software. This made sense since the platform was designed as an application-specific appliance.

Since a *Linux host* is a general purpose platform, with applications running on top of it, the locations of the files are much more distributed. Applications install and read their configuration data from text files usually stored in the /etc directory tree. Executables are generally stored in one of several *bin* directories, but the

bin and etc directories are often on the same physical device. Since each *module* (application or executable) was often developed by a different organization and shared with the world, each module was responsible for its own configuration data format. Most applications are community supported, and while there are some generally accepted guiding principles on how their configuration data is formatted, no central authority exists to control or ensure compliance.

## Pre-deployment Testing of Production Environments

Historically, the cost of *network device* testing has been hampered by the cost of a single device. Setting up an appropriately sized lab topology can be very expensive. As a result, it is difficult to do comprehensive topology-based testing of a release before deploying it. Thus, many network admins cannot or will not do comprehensive system testing of a new release before deploying it.

Alternatively, the cost of a *Linux host* is cheap (or nearly free when using virtualization), so rigorous testing of a release before deploying it is not encumbered by budgeting concerns. Most sysadmins extensively test new releases in the complete application environment.

## Upgrade Procedure

Both network admins and sysadmins generally plan upgrades only to gain new functionality or to get bug fixes when the workarounds become too onerous. The goal is to reduce the number of upgrades as much as possible.

The *network device* upgrade paradigm is to leave the configuration data in place, and *replace the executable files* either all at once from a single binary image or in large chunks (subsystems). A full release upgrade comes with risk due to unexpected behavior changes in subsystems where the admin did not anticipate or need changes.

The *Linux host* upgrade paradigm is to independently *upgrade a small list of packages* while leaving most of the OS untouched. Changing a small list of packages reduces the risk of unintended consequences. Usually upgrades are a "forward only" paradigm, where the sysadmins generally plan to move to the latest code within the same major release when needed. Every few years, when a new kernel train is released, a major upgrade is planned. A major upgrade involves wiping and replacing the entire OS and migrating configuration data.

## Rollback Procedure

Even the most well planned and tested upgrades can result in unforeseen problems, and sometimes the best solution to new problems is to roll back to the previous state.

Since *network devices* clearly separate data and executables, generally the process is to *overwrite the new release executable* with the previously running executable. If the configuration was changed by the newer release, then you either have to manually back out or repair the changes, or restore from an already backed up configuration.

The *Linux host* scenario can be more complicated. There are three main approaches:

- Back out individual packages: If the problematic package is identified, the sysadmin can downgrade the affected package directly. In rare cases the configuration files may have to be restored from backup, or edited to back out any changes that were automatically made by the upgrade package.
- Flatten and rebuild: If the OS becomes unusable, you can use orchestration tools to reinstall the previous OS release from scratch and then automatically rebuild the configuration.
- Backup and restore: Another common strategy is to restore to a previous state via a backup captured before the upgrade.

## Third Party Packages

Third party packages are rare in the *network device* world. Because the network OS is usually proprietary, third party packages are usually packaged by the network device vendor and upgrades of those packages is handled by the network device upgrade system.

Third party packages in *Linux host* world often use the same package system as the distribution into which it is to be installed (for example, Debian uses `apt-get`). Or the package may be compiled and installed by the sysadmin. Configuration and executable files generally follow the same filesystem hierarchy standards as other applications.

## Upgrading Cumulus Linux Devices: Strategies and Processes

Because Cumulus Linux is both Linux *and* a network device, it has characteristics of both paradigms. The following describes the Cumulus Linux paradigm with respect to upgrade planning and execution.

### Automated Configuration Is Preferred over Manual Configuration

Because Cumulus Linux *is* Linux, Cumulus Networks recommends that even with small networks or test labs, network admins should make the jump to deploy, provision, configure, and upgrade switches using automation from the very beginning. The small up front investment of time spent learning an orchestration tool, even to provision a small number of Cumulus Linux devices, will pay back dividends for a long time. The biggest gain is realized during the upgrade process, where the network admin can quickly upgrade dozens of devices in a repeatable manner.

Switches, like servers, should be treated like *cattle, not pets*.

### Out-of-Band Management Is Worth the Investment

Because network devices are reachable via the IP addresses on the front panel ports, many network admins of small-to-medium sized networks use *in-band* networks to manage their switches. In this design, management traffic like SSH, SNMP, and console server connections use the same networks that regular network traffic traverses — there is no separation between the *management plane* and the *data plane*. Larger data centers create a separate *out-of-band* network with a completely separate subnet and reachability path to attach to the management ports — that is accessible via eth0 and the serial console.

This is a situation where smaller companies should learn from the big companies. A separate management network isn't free, but it is relatively cheap. With an inexpensive [Cumulus RMP](#) management switch, an inexpensive console server, and a separate cable path, up to 48 devices can be completely controlled via the out-of-band network in the case of a network emergency.

There are many scenarios where in-band networking can fail and leave the network admin waiting for someone to drive to the data center or remote site to connect directly to the console of a misconfigured or failing device. The cost of one outage would usually more than pay for the investment in a separate network. For even more security, attach remote-controllable power distribution units (PDUs) in each rack to the management network, so you can have complete control to remote power cycle every device in that rack.

### Understanding the Locations of Configuration Data for Management, Migration, and Backup

As with other Linux distributions, the `/etc` directory is the primary location for all configuration data in Cumulus Linux. The following list of files is a likely set of files that should be backed up and migrated to a new release, but any file that has been changed would need to be examined:

Cumulus Networks recommends you consider making the following files and directories part of a persistent configuration.

## **Network Configuration Files**

File Name and Location	Explanation	Cumulus Linux Documentation	Debian Documentation
/etc/network/	Network configuration files, most notably /etc/network/interfaces and /etc/network/interfaces.d/	Configuring and Managing Network Interfaces (see page 94)	<a href="http://wiki.debian.org/NetworkConfiguration">wiki.debian.org/NetworkConfiguration</a>
/etc/resolv.conf	DNS resolution	Not unique to Cumulus Linux: <a href="http://wiki.debian.org/NetworkConfiguration#The_resolv.conf_configuration_file">wiki.debian.org/NetworkConfiguration#The_resolv.conf_configuration_file</a>	<a href="http://www.debian.org/doc/manuals/debian-reference/ch05.en.html">www.debian.org/doc/manuals/debian-reference/ch05.en.html</a>
/etc/quagga/	Routing application (responsible for BGP and OSPF)	Quagga Overview (see page 318)	<a href="http://packages.debian.org/wheezy/quagga">packages.debian.org/wheezy/quagga</a>
/etc/hostname	Configuration file for the hostname of the switch	Quick Start Guide#ConfiguringtheHostnameandTimeZone (see page 9)	<a href="http://wiki.debian.org/HowTo/ChangeHostname">wiki.debian.org/HowTo/ChangeHostname</a>
/etc/cumulus/ports.conf	Breakout cable configuration file	Configuring Switch Port Attributes#ConfiguringBreakoutPorts (see page 113)	N/A; please read the guide on breakout cables
/etc/cumulus/switchd.conf	Switchd configuration	Configuring switchd (see page 87)	N/A; please read the guide on switchd configuration

## Additional Commonly Used Files

File Name and Location	Explanation	Cumulus Linux Documentation	Debian Documentation
/etc/motd	Message of the day	Not unique to Cumulus Linux	<a href="http://wiki.debian.org/motd#Wheezy">wiki.debian.org/motd#Wheezy</a>
/etc/passwd	User account information	Not unique to Cumulus Linux	<a href="http://www.debian.org/doc/manuals/debian-reference/ch04.en.html">www.debian.org/doc/manuals/debian-reference/ch04.en.html</a>
/etc/shadow	Secure user account information	Not unique to Cumulus Linux	<a href="http://www.debian.org/doc/manuals/debian-reference/ch04.en.html">www.debian.org/doc/manuals/debian-reference/ch04.en.html</a>
/etc/group	Defines user groups on the switch	Not unique to Cumulus Linux	<a href="http://www.debian.org/doc/manuals/debian-reference/ch04.en.html">www.debian.org/doc/manuals/debian-reference/ch04.en.html</a>
/etc/lldpd.conf	Link Layer Discover Protocol (LLDP) daemon configuration	Link Layer Discovery Protocol (see page 139)	<a href="http://packages.debian.org/wheezy/llpd">packages.debian.org/wheezy/llpd</a>
/etc/lldpd.d/	Configuration directory for llpd	Link Layer Discovery Protocol (see page 139)	<a href="http://packages.debian.org/wheezy/llpd">packages.debian.org/wheezy/llpd</a>
/etc/nsswitch.conf	Name Service Switch (NSS) configuration file	LDAP Authentication and Authorization (see page 73)	<a href="http://wiki.debian.org/LDAP/NSS">wiki.debian.org/LDAP/NSS</a>
/etc/ssh/	SSH configuration files	SSH for Remote Access (see page 65)	<a href="http://wiki.debian.org/SSH">wiki.debian.org/SSH</a>
/etc/ldap/ldap.conf	Lightweight Directory Access Protocol configuration file	LDAP Authentication and Authorization (see page 73)	<a href="http://www.debian.org/doc/manuals/debian-reference/ch04.en.html">www.debian.org/doc/manuals/debian-reference/ch04.en.html</a>

- If you are using the root user account, consider including /root/.
- If you have custom user accounts, consider including /home/<username>/.

## Pre-Deployment Testing of New Releases Is Advised and Enabled

White box switches and virtualization (Cumulus VX) brings the cost of networking devices down, so network admins' testing of their own procedures, configurations, applications, and network topology in an appropriately-sized lab topology becomes extremely affordable.

## Upgrading Cumulus Linux: Choosing between a Binary Install vs. Package Upgrade

Network admins have two ways to upgrade Cumulus Linux:

- Performing a binary (full image) install of the new version, running `cl-img-install` on the switch
- Upgrading only the changed packages, using `apt-get update` and `apt-get dist-upgrade`

There are advantages and disadvantages to using these methods, which are outlined below.

### Upgrading via Binary Install (`cl-img-install`)

Pros:

- Image is installed to the [alternate disk image slot](#) (see page 17) while the switch remains operational.
- The only downtime is the reboot/init process.
- You choose the exact version that you want to upgrade to.
- Rolling back to the previous version and config is easy and quick; it requires only running `cl-img-select -s` and reboot.
- This is the only method for upgrading to a new major (X.0) or minor version (X.Y). For example, when you are upgrading from 2.5.5 to 3.0 or from 2.2.2 to 2.5.5.

Cons:

- Configuration data must be moved to the new OS via some mechanism before the new OS is booted, or soon afterwards via out-of-band management.
- Moving the configuration file can go wrong in various ways:
  - Identifying all the locations of config data is not always an easy task.
  - Config file changes in the new version may cause merge conflicts that go undetected.
- If config files aren't restored correctly, the user may be unable to attach to the switch from in-band management. Hence, out-of-band connectivity (eth0 or console) is recommended.

To upgrade the switch by running a binary install:

1. Back up the configurations off the switch.
2. Install the binary image to the [alternate slot](#) (see page 17) and select it as the new primary slot.

```
cumulus@switch$ sudo cl-img-install -s <image_url>
```



If you don't use the `-s` flag here, you will have to run `cl-img-select -s` after the installation to manually select the alternate slot.

Click to expand full output

```
cumulus@switch$ sudo cl-img-install -s CumulusLinux-2.5.3a-amd64.  
bin
```

```
Defaulting to image slot 2 for install.
Dumping image info from CumulusLinux-2.5.3a-amd64.bin ...
Verifying image checksum ... OK.
Preparing image archive ... OK.
Control File Contents
=====
Description: Cumulus Linux
OS-Release: 2.5.3a-3b46bef-201509041633-build
Architecture: amd64
Date: Fri, 04 Sep 2015 17:10:30 -0700
Installer-Version: 1.2
Platforms: accton_as5712_54x accton_as6712_32x mlx_sx1400_i73612
dell_s6000_s1220 dell_s4000_c2338 dell_s3000_c2338
cel_redstone_xp cel_smallstone_xp cel_pebble quanta_panther
quanta_ly8_rangeley quanta_ly6_rangeley quanta_ly9_rangeley
Homepage: http://www.cumulusnetworks.com/
Data Archive Contents
=====
-rw-r--r-- build/Development      131 2015-09-05 00:10:29 file.
list
-rw-r--r-- build/Development      44 2015-09-05 00:10:29 file.
list.sha1
-rw-r--r-- build/Development 140238619 2015-09-05 00:10:29
sysroot-release.tar.gz
-rw-r--r-- build/Development      44 2015-09-05 00:10:30
sysroot-release.tar.gz.sha1
-rw-r--r-- build/Development     8094220 2015-09-05 00:10:29
vmlinuz-initrd.tar.xz
-rw-r--r-- build/Development      44 2015-09-05 00:10:30
vmlinuz-initrd.tar.xz.sha1
Current image slot setup:
active => slot 1 (primary): 2.5.3-c4e83ad-201506011818-build
          slot 2 (alt    ): 2.5.2-727a0c6-201504132125-build
About to update image slot 2 using:
/home/cumulus/CumulusLinux-2.5.3a-amd64.bin
Are you sure (y/N)? y
Verifying image checksum ... OK.
Preparing image archive ... OK.
Validating sha1 for vmlinuz-initrd.tar.xz... done.
Validating sha1 for sysroot-release.tar.gz... done.
Installing OS-Release 2.5.3a-3b46bef-201509041633-build into
image slot 2 ...
Info: Copying sysroot into slot 2
Creating logical volume SYSROOT2 on volume group CUMULUS... done.
Verifying sysroot copy... OK.
Copying kernel into CLBOOT partition... done.
Verifying kernel copy... OK.
Generating grub.cfg ...
Found Cumulus Linux image: /boot/cl-vmlinuz-3.2.65-1+deb7u2+c12.5
+5-slot-1
Found Cumulus Linux image: /boot/cl-vmlinuz-3.2.65-1+deb7u2+c12.5
+5-slot-2
```

```
done
Success: /home/cumulus/CumulusLinux-2.5.3a-amd64.bin loaded into
image slot 2.
```

3. Reboot the switch.

```
cumulus@switch$ sudo reboot
```

4. Restore the configuration files to the new version — ideally via automation.
5. Verify correct operation with the old configurations on the new version.
6. Reinstall third party apps and associated configurations.

## ***Upgrading Using Package Installs (apt-get update && apt-get dist-upgrade)***

Pros:

- Configuration data stays in place while the binaries are upgraded.

Cons:

- This method works only if you are upgrading to a later maintenance release (X.Y.Z, like 2.5.5) from an earlier release in the same major and minor release family **only** (like 2.5.0 to 2.5.4, or 2.5.2 to 2.5.5).
- Rollback is quite difficult and tedious.
- You can't choose the exact release version that you want to run.
- When you upgrade, you upgrade all packages to the latest available version.
- The upgrade process takes a while to complete, and various switch functions are intermittently available during the upgrade.
- Some upgrade operations will terminate SSH sessions on the in-band (front panel) ports, leaving the user unable to monitor the upgrade process. As a workaround, use the **dtach** tool.
- Just like the binary install method, you still must reboot after the upgrade, lengthening the downtime.



Before you upgrade a PowerPC switch, run `df -m` and make sure the overlay filesystem `/mnt/root-rw` has at least 200MB of free disk space. See [this release note](#) for more details.

To upgrade the switch by updating the packages:

1. Back up the configurations off the switch.
2. Fetch the latest update meta-data from the repository.

```
cumulus@switch$ sudo apt-get update
```

3. Upgrade all the packages to the latest distribution.

```
cumulus@switch$ sudo apt-get dist-upgrade
```

4. Reboot the switch.

```
cumulus@switch$ sudo reboot
```

5. Verify correct operation with the old configurations on new version.



While this method doesn't overwrite the [target image slot \(see page 17\)](#), the disk image does occupy a lot of disk space used by both Cumulus Linux image slots.



After you successfully upgrade Cumulus Linux, you may notice some results that you may or may not have expected:

- `apt-get dist-upgrade` always updates the operating system to the most current version, so if you are currently running Cumulus Linux 2.5.2 and run `apt-get dist-upgrade` on that switch, the packages will get upgraded to their 2.5.4 versions.
- When you run `cl-img-select`, the output still shows the version of Cumulus Linux from the last binary install. So if you installed Cumulus Linux 2.5.3 as a full image install and then upgraded to 2.5.4 using `apt-get dist-upgrade`, the output from `cl-img-select` still shows version 2.5.3.

Why you should use `apt-get dist-upgrade` instead of `apt-get upgrade` (Click here to expand...)



Cumulus Networks recommends you upgrade Cumulus Linux using `apt-get dist-upgrade` instead of `apt-get upgrade`.

This ensures all the packages in the distribution get updated to the current version. `apt-get upgrade` **may** work correctly if no packages are held back by `apt`. A package can be held back if one or more of its dependencies has changed, or it can occur for other reasons. For example, if you see this message when running `apt-get upgrade`:

```
"The following packages have been kept back:  
linux-image-powerpc"
```

It means `apt-get upgrade` did not install the kernel package. However, `apt-get dist-upgrade` would have picked it up. Most applications in Cumulus Linux rely on the correct kernel version. If an application doesn't get the kernel version it expects, it may result in a non-functional system.

You can manually install a held back package by running `apt-get install` on it:

```
apt-get install linux-image-powerpc
```

If you must use `apt-get upgrade`, run it twice. For the second time, include the `-s` or `--dry-run` option to verify that all packages were picked up when you upgraded. Otherwise, you must manually install any held back packages to complete the upgrade.

```
apt-get upgrade --dry-run
```

## ***Rolling Back a Cumulus Linux Installation***

### ***Rolling Back after Using Binary Install***

1. Select the alternate slot as the new primary slot. (The primary slot will be booted at the next reboot)

```
cumulus@switch$ sudo cl-img-select -s
```

2. Reboot the switch.

```
cumulus@switch$ sudo reboot
```

### ***Rolling Back after Using Package Install***

Rolling back to an earlier release after upgrading the packages on the switch follows the same procedure as described for the Linux host OS rollback above. There are three main strategies, and all require detailed planning and execution:

- Back out individual packages: If the problematic package is identified, the network admin can downgrade the affected package directly. In rare cases the configuration files may have to be restored from backup, or edited to back out any changes that were automatically made by the upgrade package.
- Flatten and rebuild: If the OS becomes unusable, you can use orchestration tools to reinstall the previous OS release from scratch and then automatically rebuild the configuration.
- Backup and restore: Another common strategy is to restore to a previous state via a backup captured before the upgrade.

Which method you employ is specific to your deployment strategy, so providing detailed steps for each scenario is outside the scope of this document.

### ***Third Party Package Considerations***

Note that if you install any third party apps on a Cumulus Linux switch, any configuration data will likely be installed into the `/etc` directory, but it is not guaranteed. It is the responsibility of the network admin to understand the behavior and config file information of any third party packages installed on a Cumulus Linux switch.

After you upgrade the OS in the alternate image slot, you will need to reinstall any third party packages or any Cumulus Linux add-on packages, such as `c1-mgmtvrf`, or `vxsnd` and `vxrd`.

## Caveats while Upgrading Cumulus Linux 2.5.x

- **RN-287:** Copying the `/etc/passwd` file to the other slot when one version is earlier than Cumulus Linux 2.5.3 and the other version is later than Cumulus Linux 2.5.3 causes issues with LLDP not starting and Quagga logs not being created.

## Adding and Updating Packages

You use the Advanced Packaging Tool (APT) to manage additional applications (in the form of packages) and to install the latest updates.

### Contents

(Click to expand)

- Contents (see page 46)
- Commands (see page 46)
- Updating the Package Cache (see page 46)
- Listing Available Packages (see page 47)
- Adding a Package (see page 49)
- Listing Installed Packages (see page 49)
- Upgrading to Newer Versions of Installed Packages (see page 50)
  - Upgrading a Single Package (see page 50)
  - Upgrading All Packages (see page 50)
- Adding Packages from Another Repository (see page 50)
- Configuration Files (see page 52)
- Useful Links (see page 52)

### Commands

- `apt-get`
- `apt-cache`
- `dpkg`

## Updating the Package Cache

To work properly, APT relies on a local cache of the available packages. You must populate the cache initially, and then periodically update it with `apt-get update`:

```
cumulus@switch:~$ sudo apt-get update
Get:1 http://repo.cumulusnetworks.com CumulusLinux-2.5 Release.gpg [490 B]
Get:2 http://repo.cumulusnetworks.com CumulusLinux-2.5 Release [16.2 kB]
```

```

Get:3 http://repo.cumulusnetworks.com CumulusLinux-2.5/main powerpc
  Packages [181 kB]
Get:4 http://repo.cumulusnetworks.com CumulusLinux-2.5/addons powerpc
  Packages [75.1 kB]
Get:5 http://repo.cumulusnetworks.com CumulusLinux-2.5/updates powerpc
  Packages [112 kB]
Get:6 http://repo.cumulusnetworks.com CumulusLinux-2.5/security-updates
  powerpc Packages [28.5 kB]
Ign http://repo.cumulusnetworks.com CumulusLinux-2.5/addons Translation-en
Ign http://repo.cumulusnetworks.com CumulusLinux-2.5/main Translation-en
Ign http://repo.cumulusnetworks.com CumulusLinux-2.5/security-updates
  Translation-en
Ign http://repo.cumulusnetworks.com CumulusLinux-2.5/updates Translation-en
Fetched 413 kB in 3s (117 kB/s)
Reading package lists... Done

```

## ***Listing Available Packages***

Once the cache is populated, use `apt-cache` to search the cache to find the packages you are interested in or to get information about an available package. Here are examples of the `search` and `show` sub-commands:

```

cumulus@switch:~$ apt-cache search tcp
netbase - Basic TCP/IP networking system
quagga-doc - documentation files for quagga
libwrap0-dev - Wietse Venema's TCP wrappers library, development files
libwrap0 - Wietse Venema's TCP wrappers library
librelop0 - Reliable Event Logging Protocol (RELP) library
socat - multipurpose relay for bidirectional data transfer
openssh-client - secure shell (SSH) client, for secure access to remote
machines
libpq5 - PostgreSQL C client library
rsyslog - reliable system and kernel logging daemon
tcpdump - command-line network traffic analyzer
openssh-server - secure shell (SSH) server, for secure access from remote
machines
librelop-dev - Reliable Event Logging Protocol (RELP) library - development
files
fakeroot - tool for simulating superuser privileges
quagga - BGP/OSPF/RIP routing daemon
monit - utility for monitoring and managing daemons or similar programs
python-dpkt - Python packet creation / parsing module
iperf - Internet Protocol bandwidth measuring tool
nmap - The Network Mapper
tcpstat - network interface statistics reporting tool

```

```
tcpreplay - Tool to replay saved tcpdump files at arbitrary speeds
nuttcp - network performance measurement tool
collectd-core - statistics collection and monitoring daemon (core system)
tcpextract - extracts files from network traffic based on file signatures
nagios-plugins-basic - Plugins for nagios compatible monitoring systems
tcptrace - Tool for analyzing tcpdump output
jdoo - utility for monitoring and managing daemons or similar programs
hping3 - Active Network Smashing Tool
```

```
cumulus@switch:~$ apt-cache show tcpreplay
Package: tcpreplay
Priority: optional
Section: net
Installed-Size: 984
Maintainer: Noël Köthe <noel@debian.org>
Architecture: powerpc
Version: 3.4.3-2+wheezy1
Depends: libc6 (>= 2.7), libpcap0.8 (>= 0.9.8)
Filename: pool/CumulusLinux-2.5/addons/tcpreplay_3.4.3-2+wheezy1_powerpc.deb
Size: 435904
MD5sum: cf20bec7282ef77a091e79372a29fe1e
SHA1: 8ee1b9b02dacd0c48a474844f4466eb54c7e1568
SHA256: 03dc29057cb608d2ddf08207aedf18d47988ed6c23db0af69d30746768a639ae
SHA512:
a411b08e7a7bea62331c527d152533afca735b795f2118507260a5a0c3b6143500df9f6723cf
f736a1de0969a63e7a7ad0ce8a181ea7dfb36e2330a95d046fb1
Description: Tool to replay saved tcpdump files at arbitrary speeds
Tcpreplay is aimed at testing the performance of a NIDS by
replaying real background network traffic in which to hide
attacks. Tcpreplay allows you to control the speed at which the
traffic is replayed, and can replay arbitrary tcpdump traces. Unlike
programmatically-generated artificial traffic which doesn't
exercise the application/protocol inspection that a NIDS performs,
and doesn't reproduce the real-world anomalies that appear on
production networks (asymmetric routes, traffic bursts/lulls,
fragmentation, retransmissions, etc.), tcpreplay allows for exact
replication of real traffic seen on real networks.
Homepage: http://tcpreplay.synfin.net/
cumulus@switch:~$
```



The search commands look for the search terms not only in the package name but in other parts of the package information. Consequently, it will match on more packages than you would expect.

## Adding a Package

In order to add a new package, first ensure the package is not already installed in the system:

```
cumulus@switch:~$ dpkg -l | grep {name of package}
```

If the package is installed already, ensure it's the version you need. If it's an older version, then update the package from the Cumulus Linux repository:

```
cumulus@switch:~$ sudo apt-get update
```

If the package is not already on the system, add it by running `apt-get install`. This retrieves the package from the Cumulus Linux repository and installs it on your system together with any other packages that this package might depend on.

For example, the following adds the package `tcpreplay` to the system:

```
cumulus@switch:~$ sudo apt-get install tcpreplay
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following NEW packages will be installed:
tcpreplay
0 upgraded, 1 newly installed, 0 to remove and 1 not upgraded.
Need to get 436 kB of archives.
After this operation, 1008 kB of additional disk space will be used.
Get:1 https://repo.cumulusnetworks.com/ CumulusLinux-1.5/main tcpreplay
powerpc 3.4.3-2+wheezy1 [436 kB]
Fetched 436 kB in 0s (1501 kB/s)
Selecting previously unselected package tcpreplay.
(Reading database ... 15930 files and directories currently installed.)
Unpacking tcpreplay (from .../tcpreplay_3.4.3-2+wheezy1_powerpc.deb) ...
Processing triggers for man-db ...
Setting up tcpreplay (3.4.3-2+wheezy1) ...
cumulus@switch:~$
```

## Listing Installed Packages

The APT cache contains information about all the packages available on the repository. To see which packages are actually installed on your system, use `dpkg`. The following example lists all the packages on the system that have "tcp" in their package names:

```
cumulus@switch:~$ dpkg -l \*tcp\*
Desired=Unknown/Install/Remove/Purge/Hold
| Status=Not/Inst/Conf-files/Unpacked/half-conf/Half-inst/trig-aWait/Trig-
pend
| / Err?=(none)/Reinst-required (Status,Err: uppercase=bad)
|| / Name          Version       Architecture Description
+====+
=====

ii  tcpd           7.6.q-24      powerpc      Wietse Venema's TCP wrapper
utili

ii  tcpdump        4.3.0-1      powerpc      command-line network traffic
anal

ii  tcpreplay      3.4.3-2+whee powerpc      Tool to replay saved tcpdump
file

cumulus@switch:~$
```

## ***Upgrading to Newer Versions of Installed Packages***

### ***Upgrading a Single Package***

A single package can be upgraded by simply installing that package again with `apt-get install`. You should perform an update first so that the APT cache is populated with the latest information about the packages.

To see if a package needs to be upgraded, use `apt-cache show <pkgname>` to show the latest version number of the package. Use `dpkg -l <pkgname>` to show the version number of the installed package.

### ***Upgrading All Packages***

You can update all packages on the system with `apt-get update`. This upgrades all installed versions with their latest versions but will not install any new packages.

### ***Adding Packages from Another Repository***

As shipped, Cumulus Linux searches the Cumulus Linux repository for available packages. You can add additional repositories to search by adding them to the list of sources that `apt-get` consults. See `man sources.list` for more information.



For several packages, Cumulus Networks has added features or made bug fixes and these packages must not be replaced with versions from other repositories. Cumulus Linux has been configured to ensure that the packages from the Cumulus Linux repository are always preferred over packages from other repositories.

If you want to install packages that are not in the Cumulus Linux repository, the procedure is the same as above with one additional step.



-  Packages not part of the Cumulus Linux Repository have generally not been tested, and may not be supported by Cumulus Linux support.

Installing packages outside of the Cumulus Linux repository requires the use of `apt-get`, but, depending on the package, `easy-install` and other commands can also be used.

To install a new package, please complete the following steps:

1. First, ensure package is not already installed in the system. Use the `dpkg` command:

```
cumulus@switch:~$ dpkg -l | grep {name of package}
```

2. If the package is installed already, ensure it's the version you need. If it's an older version, then update the package from the Cumulus Linux repository:

```
cumulus@switch:~$ sudo apt-get update  
cumulus@switch:~$ sudo apt-get install {name of package}
```

3. If the package is not on the system, then most likely the package source location is also **not** in the `/etc/apt/sources.list` file. If the source for the new package is **not** in `sources.list`, please edit and add the appropriate source to the file. For example, add the following if you wanted a package from the Debian repository that is **not** in the Cumulus Linux repository:

```
deb http://http.us.debian.org/debian wheezy main  
deb http://security.debian.org/ wheezy/updates main
```

Otherwise, the repository may be listed in `/etc/apt/sources.list` but is commented out, as can be the case with the testing repository:

```
#deb http://repo.cumulusnetworks.com CumulusLinux-VERSION testing
```

To uncomment the repository, remove the # at the start of the line, then save the file:

```
deb http://repo.cumulusnetworks.com CumulusLinux-VERSION testing
```

4. Run `apt-get update` then install the package:

```
cumulus@switch:~$ sudo apt-get update  
cumulus@switch:~$ sudo apt-get install {name of package}
```

## Configuration Files

- /etc/apt/apt.conf
- /etc/apt/preferences
- /etc/apt/sources.list

## Useful Links

- Debian GNU/Linux FAQ, Ch 8 Package management tools
- man pages for apt-get, dpkg, sources.list, apt\_preferences

## Zero Touch Provisioning - ZTP

*Zero touch provisioning (ZTP)* allows devices to be quickly deployed in large-scale environments. Data center engineers only need to rack and stack the switch, then connect it to the management network — or alternatively, insert a USB stick and boot the switch. From here, the provisioning process can start automatically and deploy a configuration.

The provisioning framework allows for a one-time, user-provided script to be executed. This script can be used to add the switch to a configuration management (CM) platform such as [puppet](#), [Chef](#), [CFEngine](#), or even a custom, home-grown tool.

In addition, you can use the `autoprovision` command in Cumulus Linux to manually invoke your provisioning script.

ZTP in Cumulus Linux can occur automatically in one of two ways:

- Via DHCP
- Using a USB drive inserted into the switch (ZTP-USB)

The two methods for using ZTP are discussed below in greater detail.



The standard Cumulus Linux license requires you to page through the license file before accepting the terms, which can hinder an unattended installation like zero touch provisioning. To request a license without the EULA, email [licensing@cumulusnetworks.com](mailto:licensing@cumulusnetworks.com).

## Contents

(Click to expand)

- Contents (see page 52)
- Commands (see page 53)
- Zero Touch Provisioning over DHCP (see page 53)
  - Triggering ZTP over DHCP (see page 53)
  - Configuring The DHCP Server (see page 53)
  - Detailed Look at HTTP Headers (see page 54)
  - Testing and Debugging ZTP Scripts for DHCP (see page 54)
- Zero Touch Provisioning Using USB (ZTP-USB) (see page 54)

- Testing and Debugging ZTP-USB Scripts (see page 56)
- Writing ZTP Scripts (see page 57)
  - Example ZTP Scripts (see page 58)
- Manually Using the autoprovion Command (see page 60)
- Notes (see page 61)
- Configuration Files (see page 61)

## Commands

- autoprovion

## Zero Touch Provisioning over DHCP

For ZTP using DHCP, provisioning initially takes place over the management network and is initiated via a DHCP hook. A DHCP option is used to specify a configuration script. This script is then requested from the Web server and executed locally on the switch.

The zero touch provisioning process over DHCP follows these steps:

1. The first time you boot Cumulus Linux, eth0 is configured for DHCP and makes a DHCP request.
2. The DHCP server offers a lease to the switch.
3. If option 239 is present in the response, the zero touch provisioning process itself will start.
4. The zero touch provisioning process requests the contents of the script from the URL, sending additional [HTTP headers \(see page 54\)](#) containing details about the switch.
5. The script's contents are parsed to ensure it contains the `CUMULUS-AUTOPROVISIONING` flag (see [example scripts \(see page \)](#)).
6. The `autoprovision` command checks its [configuration file \(see page 61\)](#) to see if autoprovioning has already occurred and completed.
7. If `autoprovision` determines that provisioning is necessary, then the script executes locally on the switch with root privileges.
8. The return code of the script gets examined. If it is 0, then the provisioning state is marked as complete in the autoprovioning configuration file.

## Triggering ZTP over DHCP

If provisioning has not already occurred, it is possible to trigger the zero touch provisioning process over DHCP when eth0 is set to use DHCP and one of the following events occur:

- Booting the switch
- Plugging a cable into or unplugging it from the eth0 port
- Disconnecting then reconnecting the switch's power cord

## Configuring The DHCP Server

During the DHCP process over eth0, Cumulus Linux will request DHCP option 239. This option is used to specify the custom provisioning script.

For example, the `/etc/dhcp/dhcpd.conf` file for an ISC DHCP server would look like:

```
option cumulus-provision-url code 239 = text;

subnet 192.168.0.0 netmask 255.255.255.0 {
  range 192.168.0.100 192.168.0.200;
  option cumulus-provision-url "http://192.168.0.2/demo.sh";
}
```

Additionally, the hostname of the switch can be specified via the host-name option:

```
subnet 192.168.0.0 netmask 255.255.255.0 {
  range 192.168.0.100 192.168.0.200;
  option cumulus-provision-url "http://192.168.0.2/demo.sh";
  host dcl-tor-swl { hardware ethernet 44:38:39:00:1a:6b; fixed-
address 192.168.0.101; option host-name "dcl-tor-swl"; }
}
```

## Detailed Look at HTTP Headers

The following HTTP headers are sent in the request to the webserver to retrieve the provisioning script:

Header	Value	Example
User-Agent	-----	CumulusLinux-
AutoProvision/0.4	-----	
CUMULUS-ARCH	CPU architecture	powerpc
CUMULUS-BUILD	-----	1.5.1-5c6829a-2013
09251712-final	-----	
CUMULUS-LICENSE-INSTALLED	Either 0 or 1	1
CUMULUS-MANUFACTURER	-----	dni
CUMULUS-PRODUCTNAME	-----	et-7448bf
CUMULUS-SERIAL	-----	XYZ123004
CUMULUS-VERSION	-----	1.5.1
CUMULUS-PROV-COUNT	-----	0
CUMULUS-PROV-MAX	-----	32

## Testing and Debugging ZTP Scripts for DHCP

One can manually run a provisioning session at any time using --force (-f) option with the autoprovision command as shown below:

```
cumulus@switch:~$ sudo /usr/lib/cumulus/autoprovision --force --url
http://192.168.1.1/demo.sh
```

## Zero Touch Provisioning Using USB (ZTP-USB)



This feature has been tested only with "thumb" drives, not an actual external large USB hard drive.

Cumulus Linux supports the use of a FAT32, FAT16, or VFAT-formatted USB drive as an installation source for ZTP scripts. A daemon called `ztp-usb` runs by default in Cumulus Linux (you can disable it by specifying `START=no` in `/etc/default/ztp-usb`). You can plug in a USB stick at any time — when you power up a switch or even when the switch has been running for some time. This is useful for performing a full installation of the operating system for cases like fresh installs or disaster recovery.

At minimum, the script should:

- Install the Cumulus Linux operating system and license.
- Copy over a basic configuration to the switch.
- Restart the switch or the relevant server to get `switchd` up and running with that configuration.

Follow these steps to perform zero touch provisioning using USB:

1. Copy the Cumulus Linux [license and installation image \(see page \)](#) to the USB stick.
2. When Cumulus Linux boots, the `ztp-usb` daemon starts.
3. Every 30 seconds, the `ztp-usb` daemon looks for unmounted FAT32-, FAT16- or VFAT-formatted volumes.
4. Each new device detected by the kernel is mounted to `/mnt/usb`.
5. The daemon searches the root filesystem of the newly mounted device for filenames matching an [ONIE-style waterfall](#) (see the patterns and examples below), looking for the most specific name first, and ending at the most generic.
6. The script's contents are parsed to ensure it contains the `CUMULUS-AUTOPROVISIONING` flag (see [example scripts \(see page \)](#)).
7. The `autoprovision` command checks its [configuration file \(see page 61\)](#) to see if autoprovisioning has already occurred and completed.
8. If `autoprovision` determines that provisioning is necessary, then the script executes locally on the switch with root privileges.
9. The return code of the script gets examined. If it is 0, then the provisioning state is marked as complete in the autoprovisioning configuration file.

The filenames searched are as follows:

- `'cumulus-ztp-' + architecture + '-' + vendor + '_' + model + '-r' + revision`
- `'cumulus-ztp-' + architecture + '-' + vendor + '_' + model`
- `'cumulus-ztp-' + vendor + '_' + model`
- `'cumulus-ztp-' + architecture`
- `'cumulus-ztp'`

For example:

```
/mnt/usb/cumulus-ztp-powerpc-cel_smallstone-rUNKNOWN
/mnt/usb/cumulus-ztp-powerpc-cel_smallstone
/mnt/usb/cumulus-ztp-cel_smallstone
/mnt/usb/cumulus-ztp-powerpc
/mnt/usb/cumulus-ztp
```

## Testing and Debugging ZTP-USB Scripts

It is possible to test the scripts you've written for `ztp-usb` using the techniques described below. Once a script has been placed on a USB drive and is ready for testing follow the procedure below:

1. Disable the `ztp-usb` daemon.

```
cumulus@switch:~$ sudo service ztp-usb stop
cumulus@switch:~$ sudo service ztp-usb status
[FAIL] ztp-usb is not running ... failed!
```

2. Insert the USB stick into the switch.
3. Move the autoprovision configuration file to a safe location.

```
cumulus@switch:~$ sudo mv /var/lib/cumulus/autoprovision.conf
/var/lib/cumulus/autoprovision.conf.original
```

By moving the configuration file to a new location, the autoprovision framework has no record of previous provisioning successes or failures, which means any new attempt to autoprovision succeeds.

4. Use debugging mode to run the `ztp-usb` script.

```
cumulus@wan1$ sudo /usr/lib/cumulus/ztp-usb -d
ztp-usb: 2015-09-18 14:39:49,280 Initial hash value
731845549779ee9c37bd630c7d24cc1d
ztp-usb: 2015-09-18 14:39:49,280 Parsing partitions
ztp-usb: 2015-09-18 14:39:49,518 /dev/sda: unsupported partition
type =
ztp-usb: 2015-09-18 14:39:49,519 INFO: Trying to mount: "/dev
/sdal" of type: "vfat"
ztp-usb: 2015-09-18 14:39:49,519 Creating /mnt/usb mount
directory
ztp-usb: 2015-09-18 14:39:49,640 Waterfall search for /mnt/usb
/cumulus-ztp-unknown-accton_as5712_54x-rUNKNOWN
ztp-usb: 2015-09-18 14:39:49,640 Waterfall search for /mnt/usb
/cumulus-ztp-unknown-accton_as5712_54x
ztp-usb: 2015-09-18 14:39:49,640 Waterfall search for /mnt/usb
/cumulus-ztp-unknown-accton
```

```

ztp-usb: 2015-09-18 14:39:49,640 Waterfall search for /mnt/usb
/cumulus-ztp-unknown
ztp-usb: 2015-09-18 14:39:49,640 Waterfall search for /mnt/usb
/cumulus-ztp
ztp-usb: 2015-09-18 14:39:49,641 Found matching name, passing
/mnt/usb/cumulus-ztp to autoprovision wrapper
ztp-usb: 2015-09-18 14:39:49,641 Found /mnt/usb/cumulus-ztp
script, passing to autoprovision
ztp-usb: 2015-09-18 14:39:51,370 Script returned exit code 0
ztp-usb: 2015-09-18 14:39:51,370 Unmounting drive and removing
mountpoint.
ztp-usb: 2015-09-18 14:39:51,396 /dev/sdb: unsupported partition
type =
ztp-usb: 2015-09-18 14:39:51,396 /dev/sdb1: unsupported
partition type =
ztp-usb: 2015-09-18 14:39:51,396 /dev/sdb2: unsupported
partition type = ext4
ztp-usb: 2015-09-18 14:39:51,396 /dev/sdb3: unsupported
partition type = ext4
ztp-usb: 2015-09-18 14:39:51,396 /dev/sdb4: unsupported
partition type = LVM2_member
ztp-usb: 2015-09-18 14:39:51,396 /dev/CUMULUS-PERSIST:
unsupported partition type = RM=0
ztp-usb: 2015-09-18 14:39:51,396 /dev/CUMULUS-SYSROOT1:
unsupported partition type = RM=0
ztp-usb: 2015-09-18 14:39:51,397 /dev/CUMULUS-SYSROOT2:
unsupported partition type = RM=0
ztp-usb: 2015-09-18 14:39:51,397 Current hash value
731845549779ee9c37bd630c7d24cc1d
ztp-usb: 2015-09-18 14:40:21,427 Current hash value
731845549779ee9c37bd630c7d24cc1d

```

## Writing ZTP Scripts



Remember to include the following line in any of the supported scripts which are expected to be run via the autoprovisioning framework.

```
# CUMULUS-AUTOPROVISIONING
```

This line is required somewhere in the script file in order for execution to occur.

The script must contain the CUMULUS-AUTOPROVISIONING flag. This can be in a comment or remark and does not need to be echoed or written to stdout.

The script can be written in any language currently supported by Cumulus Linux, such as:

- Perl
- Python

- Ruby
- Shell

The script must return an exit code of 0 upon success, as this triggers the autoprovisioning process to be marked as complete in the autoprovisioning configuration file.

## Example ZTP Scripts

The following script install Cumulus Linux and its license from USB and applies a configuration:

```
#!/bin/bash
function error() {
    echo -e "\e[0;33mERROR: The Zero Touch Provisioning script failed
while running the command $BASH_COMMAND at line $BASH_LINENO.\e[0m" >&
2
    exit 1
}

# Log all output from this script
exec >/var/log/autoprovision 2>&1

trap error ERR

#Add Debian Repositories
echo "deb http://http.us.debian.org/debian wheezy main" >> /etc/apt
/sources.list
echo "deb http://security.debian.org/ wheezy/updates main" >> /etc/apt
/sources.list

#Update Package Cache
apt-get update -y

#Install netshow diagnostics commands
apt-get install -y netshow htop nmap

#Load interface config from usb
cp /mnt/usb/interfaces /etc/network/interfaces

#Load port config from usb
#   (if breakout cables are used for certain interfaces)
cp /mnt/usb/ports.conf /etc/cumulus/ports.conf

#Install a License from usb and restart switchd
cl-license -i /mnt/usb/license.txt && service switchd restart

#Reload interfaces to apply loaded config
ifreload -a

#Output state of interfaces
netshow interface
```

```
# CUMULUS-AUTOPROVISIONING
exit 0
```

Here is a simple script to install puppet:

```
#!/bin/bash
function error() {
    echo -e "\e[0;33mERROR: The Zero Touch Provisioning script failed
while running the command $BASH_COMMAND at line $BASH_LINENO.\e[0m" >&
2
    exit 1
}
trap error ERR
apt-get update -y
apt-get upgrade -y
apt-get install puppet -y
sed -i /etc/default/puppet -e 's/START=no/START=yes/'
sed -i /etc/puppet/puppet.conf -e 's/\\[main\\]\\/[main\\]/
\npluginsync=true'
service puppet restart
# CUMULUS-AUTOPROVISIONING
exit 0
```

This script illustrates how to specify an internal APT mirror and puppet master:

```
#!/bin/bash
function error() {
    echo -e "\e[0;33mERROR: The Zero Touch Provisioning script failed
while running the command $BASH_COMMAND at line $BASH_LINENO.\e[0m" >&
2
    exit 1
}
trap error ERR
sed -i /etc/apt/sources.list -e 's/repo.cumulusnetworks.com/labrepo.
mycompany.com/'
apt-get update -y
apt-get upgrade -y
apt-get install puppet -y
sed -i /etc/default/puppet -e 's/START=no/START=yes/'
sed -i /etc/puppet/puppet.conf -e 's/\\[main\\]\\/[main\\]/
\npluginsync=true'
sed -i /etc/puppet/puppet.conf -e 's/\\[main\\]\\/[main\\]/
\nserver=labpuppet.mycompany.com/'
service puppet restart
# CUMULUS-AUTOPROVISIONING
exit 0
```

Now puppet can take over management of the switch, configuration authentication, changing the default root password, and setting up interfaces and routing protocols.

Several ZTP example scripts are available in the [Cumulus GitHub repository](#).

## Manually Using the `autoprovision` Command



Be sure to specify the full path to the `autoprovision` command.

All forms of ZTP use the `autoprovision` command on the backend to execute a provided provisioning script, whether that script is sourced from a URL over the network or locally via a file from a USB drive. One of the benefits of using the `autoprovision` command — instead of simply scheduling a cronjob to run your script — is that `autoprovision` tracks whether or not a script has already been executed (and when) in its configuration file `/var/lib/cumulus/autoprovision.conf`, ensuring that a switch that has already been provisioned is not accidentally provisioned again at a later date.

Users with root privileges can interact with the `autoprovision` command directly using the examples below.

To enable zero touch provisioning, use the `-e` option:

```
cumulus@switch:~$ sudo /usr/lib/cumulus/autoprovision -e
```

To run the provisioning script against a script hosted on a Web server, use the `-u` option and include the URL to the script:

```
cumulus@switch:~$ sudo /usr/lib/cumulus/autoprovision -u http://192.168.0.1/ztp.sh
```

To run the provisioning script against a script hosted on the local filesystem, use the `--file` or `-i` option and include the file location of the script:

```
cumulus@switch:~$ sudo /usr/lib/cumulus/autoprovision --file /mnt/usb/cumulus-ztp.sh
```

To disable zero touch provisioning, use the `-x` option:

```
cumulus@switch:~$ sudo /usr/lib/cumulus/autoprovision -x
```

To enable startup discovery mode, without relying on DHCP when you boot the switch, use the `-s` option:

```
cumulus@switch:~$ sudo /usr/lib/cumulus/autoprovision -s
```

To force provisioning to occur and ignore the status listed in the configuration file use the `-f` option:

```
cumulus@switch:~$ sudo /usr/lib/cumulus/autoprovision -f --file /mnt  
/usb/cumulus-ztp.sh
```

## Notes

- During the development of a provisioning script, the switch may need to be reset.
- You can use the Cumulus Linux `cl-img-clear-overlay` command to revert the image to its original configuration.
- You can use the Cumulus Linux `cl-img-select -i` command to cause the switch to reprovision itself and install a network operating system again using ONIE.

## Configuration Files

- `/var/lib/cumulus/autoprovision.conf`: Stores configuration options and details for the autoprovisioning framework
- `/etc/default/ztp-usb`: Stores the enable/disable flag for the `ztp-usb` service

# System Management

## Setting Date and Time

Setting the time zone, date and time requires root privileges; use `sudo`.

### Contents

(Click to expand)

- [Contents \(see page 62\)](#)
- [Commands \(see page 62\)](#)
- [Setting the Time Zone \(see page 62\)](#)
- [Setting the Date and Time \(see page 63\)](#)
- [Setting Time Using NTP \(see page 64\)](#)
- [Configuration Files \(see page 64\)](#)
- [Useful Links \(see page 65\)](#)

### Commands

- `date`
- `dpkg-reconfigure tzdata`
- `hwclock`
- `ntpd` (daemon)
- `ntpq`

### Setting the Time Zone

To see the current time zone, list the contents of `/etc/timezone`:

```
cumulus@switch:~$ cat /etc/timezone
US/Eastern
```

To set the time zone, run `dpkg-reconfigure tzdata` as root:

```
cumulus@switch:~$ sudo dpkg-reconfigure tzdata
```

Then navigate the menus to enable the time zone you want. The following example selects the US/Pacific time zone:

```
cumulus@switch:~$ sudo dpkg-reconfigure tzdata

Configuring tzdata
-----

Please select the geographic area in which you live. Subsequent
configuration
questions will narrow this down by presenting a list of cities, representing
the time zones in which they are located.

 1. Africa      4. Australia   7. Atlantic   10. Pacific   13. Etc
 2. America     5. Arctic       8. Europe     11. SystemV
 3. Antarctica  6. Asia        9. Indian     12. US

Geographic area: 12

Please select the city or region corresponding to your time zone.

 1. Alaska      4. Central     7. Indiana-Starke 10. Pacific
 2. Aleutian    5. Eastern     8. Michigan      11. Pacific-New
 3. Arizona     6. Hawaii      9. Mountain     12. Samoa

Time zone: 10

Current default time zone: 'US/Pacific'
Local time is now:      Mon Jun 17 09:27:45 PDT 2013.
Universal Time is now:  Mon Jun 17 16:27:45 UTC 2013.
```

For more info see the Debian [System Administrator's Manual – Time](#).

## **Setting the Date and Time**

The switch contains a battery backed hardware clock that maintains the time while the switch is powered off and in between reboots. When the switch is running, the Cumulus Linux operating system maintains its own software clock.

During boot up, the time from the hardware clock is copied into the operating system's software clock. The software clock is then used for all timekeeping responsibilities. During system shutdown the software clock is copied back to the battery backed hardware clock.

You can set the date and time on the software clock using the `date` command. First, determine your current time zone:

```
cumulus@switch$ date +%Z
```



If you need to reconfigure the current time zone, refer to the instructions above.

Then, to set the system clock according to the time zone configured:

```
cumulus@switch$ sudo date -s "Tue Jan 12 00:37:13 2016"
```

See `man date(1)` for if you need more information.

You can write the current value of the system (software) clock to the hardware clock using the `hwclock` command:

```
cumulus@switch$ sudo hwclock -w
```

See `man hwclock(8)` if you need more information.

You can find a good overview of the software and hardware clocks in the Debian [System Administrator's Manual – Time](#), specifically the section [Setting and showing hardware clock](#).

## Setting Time Using NTP

The `ntpd` daemon running on the switch implements the NTP protocol. It synchronizes the system time with time servers listed in `/etc/ntp.conf`. It is started at boot by default. See `man ntpd(8)` for `ntpd` details.

By default, `/etc/ntp.conf` contains some default time servers. Edit `/etc/ntp.conf` to add or update time server information. See `man ntp.conf(5)` for details on configuring `ntpd` using `ntp.conf`.

To set the initial date and time via NTP before starting the `ntpd` daemon, use `ntpd -q` (This is same as `ntpdate`, which is to be retired and not available).



`ntpd -q` can hang if the time servers are not reachable.

To verify that `ntpd` is running on the system:

```
cumulus@switch:~$ ps -ef | grep ntp
ntp      4074      1  0 Jun20 ?          00:00:33 /usr/sbin/ntpd -p /var/run
/ntpd.pid -g -u 101:102
```

## Configuration Files

- `/etc/default/ntp` — `ntpd init.d` configuration variables
- `/etc/ntp.conf` — default NTP configuration file
- `/etc/init.d/ntp` — `ntpd` init script

## Useful Links

- Debian System Administrator's Manual – Time
- <http://www.ntp.org>
- [http://en.wikipedia.org/wiki/Network\\_Time\\_Protocol](http://en.wikipedia.org/wiki/Network_Time_Protocol)
- <http://wiki.debian.org/NTP>

## Authentication, Authorization, and Accounting

- SSH for Remote Access (see page 65)
- User Accounts (see page 66)
- Using sudo to Delegate Privileges (see page 67)
- PAM and NSS (see page 73)

### SSH for Remote Access

You use SSH to securely access a Cumulus Linux switch remotely.

### Contents

(Click to expand)

- Contents (see page 65)
- Access Using Passkey (Basic Setup) (see page 65)
  - Completely Passwordless System (see page 66)
- Useful Links (see page 66)

### Access Using Passkey (Basic Setup)

Cumulus Linux uses the openSSH package to provide SSH functionality. The standard mechanisms of generating passwordless access just applies. The example below has the cumulus user on a machine called management-station connecting to a switch called *cumulus-switch1*.

First, on management-station, generate the SSH keys:

```
cumulus@management-station:~$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/cumulus/.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/cumulus/.ssh/id_rsa.
Your public key has been saved in /home/cumulus/.ssh/id_rsa.pub.
The key fingerprint is:
8c:47:6e:00:fb:13:b5:07:b4:1e:9d:f4:49:0a:77:a9 cumulus@management-
station
```

```
The key's randomart image is:
```

```
+--[ RSA 2048]----+
|   . . = o o.   |
|   o . o * ..   |
|   . o = =.o    |
|   . O oE       |
|   + S          |
|   +             |
|               |
|               |
|               |
+-----+
```

Next, append the public key in `~/.ssh/id_rsa.pub` into `~/.ssh/authorized_keys` in the target user's home directory:

```
cumulus@management-station:~$ scp .ssh/id_rsa.pub cumulus@cumulus-switch1:.
ssh/authorized_keys
Enter passphrase for key '/home/cumulus/.ssh/id_rsa':
id_rsa.pub
```



Remember, you cannot use the root account to SSH to a switch in Cumulus Linux.

## **Completely Passwordless System**

When generating the passphrase and its associated keys, as in the first step above, do not enter a passphrase. Follow all the other instructions.

## **Useful Links**

- <http://www.debian-administration.org/articles/152>

## **User Accounts**

By default, Cumulus Linux has two user accounts: *cumulus* and *root*.

The *cumulus* account:

- Default password is *CumulusLinux!*
- Is a user account in the *sudo* group with sudo privileges
- User can log in to the system via all the usual channels like console and **SSH (see page 65)**

The *root* account:

- Default password is disabled by default

- Has the standard Linux root user access to everything on the switch
- Disabled password prohibits login to the switch by SSH, telnet, FTP, and so forth

For best security, you should change the default password (using the `passwd` command) before you configure Cumulus Linux on the switch.

You can enable a valid password for the root account using the `sudo passwd root` command and can install an SSH key for the root account if needed. Enabling a password for the root account allows the root user to log in directly to the switch. The Cumulus Linux default root account behavior is consistent with Debian.

You can add more user accounts as needed. Like the *cumulus* account, these accounts must use `sudo` to execute privileged commands (see page 67), so be sure to include them in the `sudo` group.

To access the switch without any password requires booting into a single shell/user mode. [Here are the instructions \(see page 392\)](#) on how to do this using PowerPC and x86 switches.

## Using sudo to Delegate Privileges

By default, Cumulus Linux has two user accounts: *root* and *cumulus*. The *cumulus* account is a normal user and is in the group `sudo`.

You can add more user accounts as needed. Like the *cumulus* account, these accounts must use `sudo` to execute privileged commands.

## Contents

(Click to expand)

- [Contents \(see page 67\)](#)
- [Commands \(see page 67\)](#)
- [Using sudo \(see page 67\)](#)
- [sudoers Examples \(see page 68\)](#)
- [Configuration Files \(see page 73\)](#)
- [Useful Links \(see page 73\)](#)

## Commands

- `sudo`
- `visudo`

## Using sudo

`sudo` allows you to execute a command as superuser or another user as specified by the security policy. See `man sudo(8)` for details.

The default security policy is `sudoers`, which is configured using `/etc/sudoers`. Use `/etc/sudoers.d/` to add to the default `sudoers` policy. See `man sudoers(5)` for details.



Use `visudo` only to edit the `sudoers` file; do not use another editor like `vi` or `emacs`. See `man visudo(8)` for details.

Errors in the `sudoers` file can result in losing the ability to elevate privileges to root. You can fix this issue only by power cycling the switch and booting into single user mode. Before modifying `sudoers`, enable the root user by setting a password for the root user.

By default, users in the `sudo` group can use `sudo` to execute privileged commands. To add users to the `sudo` group, use the `useradd(8)` or `usermod(8)` command. To see which users belong to the `sudo` group, see `/etc/group` (`man group(5)`).

Any command can be run as `sudo`, including `su`. A password is required.

The example below shows how to use `sudo` as a non-privileged user *cumulus* to bring up an interface:

```
cumulus@switch:~$ ip link show dev swp1
3: swp1: <BROADCAST,MULTICAST> mtu 1500 qdisc pfifo_fast master br0 state
DOWN mode DEFAULT qlen 500
link/ether 44:38:39:00:27:9f brd ff:ff:ff:ff:ff:ff

cumulus@switch:~$ ip link set dev swp1 up
RTNETLINK answers: Operation not permitted

cumulus@switch:~$ sudo ip link set dev swp1 up
Password:

cumulus@switch:~$ ip link show dev swp1
3: swp1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast master
br0 state UP mode DEFAULT qlen 500
link/ether 44:38:39:00:27:9f brd ff:ff:ff:ff:ff:ff
```

## ***sudoers Examples***

The following examples show how you grant as few privileges as necessary to a user or group of users to allow them to perform the required task. For each example, the system group *noc* is used; groups are prefixed with an %.

When executed by an unprivileged user, the example commands below must be prefixed with `sudo`.

Category	Privilege	Example Command	<code>sudoers</code> Entry
Monitoring	Switch port info	<code>ethtool -m swp1</code>	<code>%noc ALL=(ALL) NOPASSWD: /sbin/ethtool</code>
Monitoring	System diagnostics	<code>cl-support</code>	

Category	Privilege	Example Command	sudoers Entry
			<pre>%noc ALL=(ALL) NOPASSWD:/usr/cumulus/bin/cl-support</pre>
Monitoring	Routing diagnostics	<pre>cl-resource-query</pre>	<pre>%noc ALL=(ALL) NOPASSWD:/usr/cumulus/bin/cl-resource-query</pre>
Image management	Install images	<pre>cl-img-install http://lab/install.bin</pre>	<pre>%noc ALL=(ALL) NOPASSWD:/usr/cumulus/bin/cl-img-install</pre>
Image management	Swapping slots	<pre>cl-img-select 1</pre>	<pre>%noc ALL=(ALL) NOPASSWD:/usr/cumulus/bin/cl-img-select</pre>
Image management	Clearing an overlay	<pre>cl-img-clear-overlay 1</pre>	<pre>%noc ALL=(ALL) NOPASSWD:/usr/cumulus/bin/cl-img-clear-overlay</pre>
Package management	Any apt-get command	<pre>apt-get update or apt-get install</pre>	<pre>%noc ALL=(ALL) NOPASSWD:/usr/bin/apt-get</pre>
Package management	Just apt-get update		

Category	Privilege	Example Command	sudoers Entry
		<code>apt-get update</code>	%noc ALL=(ALL) NOPASSWD:/usr/bin/apt-get update
Package management	Install packages	<code>apt-get install mtr-tiny</code>	%noc ALL=(ALL) NOPASSWD:/usr/bin/apt-get install *
Package management	Upgrading	<code>apt-get upgrade</code>	%noc ALL=(ALL) NOPASSWD:/usr/bin/apt-get upgrade
Netfilter	Install ACL policies	<code>cl-acltool -i</code>	%noc ALL=(ALL) NOPASSWD:/usr/cumulus/bin/cl-acltool
Netfilter	List iptables rules	<code>iptables -L</code>	%noc ALL=(ALL) NOPASSWD:/sbin/iptables
L1 + 2 features	Any LLDP command	<code>lldpcli show neighbors / configure</code>	%noc ALL=(ALL) NOPASSWD:/usr/sbin/lldpcli
L1 + 2 features	Just show neighbors	<code>lldpcli show neighbors</code>	%noc ALL=(ALL) NOPASSWD:/usr/sbin/lldpcli show neighbours*

Category	Privilege	Example Command	sudoers Entry
Interfaces	Modify any interface	<pre>ip link set dev swp1 {up down}</pre>	<pre>%noc ALL=(ALL) NOPASSWD: /sbin/ip link set *</pre>
Interfaces	Up any interface	<pre>ifup swp1</pre>	<pre>%noc ALL=(ALL) NOPASSWD: /sbin/ifup</pre>
Interfaces	Down any interface	<pre>ifdown swp1</pre>	<pre>%noc ALL=(ALL) NOPASSWD: /sbin/ifdown</pre>
Interfaces	Up/down only swp2	<pre>ifup swp2 / ifdown swp2</pre>	<pre>%noc ALL=(ALL) NOPASSWD: /sbin/ifup swp2,/sbin /ifdown swp2</pre>
Interfaces	Any IP address chg	<pre>ip addr {add del} 192.0.2.1/30 dev swp1</pre>	<pre>%noc ALL=(ALL) NOPASSWD: /sbin/ip addr *</pre>
Interfaces	Only set IP address	<pre>ip addr add 192.0.2.1/30 dev swp1</pre>	<pre>%noc ALL=(ALL) NOPASSWD: /sbin/ip addr add *</pre>
Ethernet bridging	Any bridge command		

Category	Privilege	Example Command	sudoers Entry
		<pre>brctl addbr br0 / brctl delif br0 swp1</pre>	<pre>%noc ALL=(ALL) NOPASSWD: /sbin/brctl</pre>
Ethernet bridging	Add bridges and ints	<pre>brctl addbr br0 / brctl addif br0 swp1</pre>	<pre>%noc ALL=(ALL) NOPASSWD: /sbin/brctl addbr *,/sbin /brctl addif *</pre>
Spanning tree	Set STP properties	<pre>mstpctl setmaxage br2 20</pre>	<pre>%noc ALL=(ALL) NOPASSWD: /sbin/mstpctl</pre>
Troubleshooting	Restart switchd	<pre>service switchd restart</pre>	<pre>%noc ALL=(ALL) NOPASSWD:/usr /sbin/service switchd *</pre>
Troubleshooting	Restart any service	<pre>service switchd cron</pre>	<pre>%noc ALL=(ALL) NOPASSWD:/usr /sbin/service</pre>
Troubleshooting	Packet capture	<pre>tcpdump</pre>	<pre>%noc ALL=(ALL) NOPASSWD:/usr /sbin/tcpdump</pre>
L3	Add static routes	<pre>ip route add 10.2.0.0/16 via 10.0.0.1</pre>	<pre>%noc ALL=(ALL) NOPASSWD:/bin /ip route add *</pre>

Category	Action	Example Command	Success Entry
L3	Delete static routes	<pre>ip route del 10.2.0.0/16 via 10.0.0.1</pre>	<pre>%noc ALL=(ALL) NOPASSWD:/bin /ip route del *</pre>
L3	Any static route chg	<pre>ip route *</pre>	<pre>%noc ALL=(ALL) NOPASSWD:/bin /ip route *</pre>
L3	Any iproute command	<pre>ip *</pre>	<pre>%noc ALL=(ALL) NOPASSWD:/bin /ip</pre>
L3	Non-modal OSPF	<pre>cl-ospf area 0.0.0.1 range 10.0.0.0/24</pre>	<pre>%noc ALL=(ALL) NOPASSWD:/usr /bin/cl-ospf</pre>

## Configuration Files

- /etc/sudoers - default security policy
- /etc/sudoers.d/ - default security policy

## Useful Links

- [sudo](#)
- [Adding Yourself to sudoers](#)

## LDAP Authentication and Authorization

Cumulus Linux uses Pluggable Authentication Modules (PAM) and Name Switch Service (NSS) for user authentication.

NSS provides the lookup and mapping of users, while PAM provides login handling, authentication and session setup.

PAMs can be used with protocols like LDAP to provide user authentication for numerous services on a network.

## Contents

(Click to expand)

- [Contents \(see page 74\)](#)
- [Configuring LDAP \(see page 74\)](#)
- [Installing libnss-ldapd \(see page 74\)](#)
- [Configuring nslcd.conf \(see page 75\)](#)
- [Troubleshooting LDAP Authentication \(see page 75\)](#)
  - [Common Problems \(see page 75\)](#)
- [Configuring LDAP Authorization \(see page 75\)](#)
- [A Longer Example \(see page 76\)](#)
- [References \(see page 76\)](#)

## Configuring LDAP

There are 3 common ways of configuring LDAP authentication on Linux:

- libnss-ldap
- libnss-ldapd
- libnss-sss

This chapter covers using `libnss-ldapd` only. From internal testing, this library worked best with Cumulus Linux and was the easiest to configure, automate and troubleshoot.

## Installing libnss-ldapd

To install `libnss-ldapd`, run:

```
cumulus@switch:~$ sudo apt-get install libnss-ldapd ldap-utils
```

This brings up an interactive prompt asking questions about the LDAP URI, base domain name and so on. To pre-fill these details, run `apt-get install debconf-utils` and populate `debconf-set-selections` with the appropriate answers. Run `debconf-show <pkg>` to check the settings.

Here is an [example of how to prefill questions using debconf-set-selections](#).



For nested group support, `libnss-ldapd` must be version 0.9 or higher. For Cumulus Linux 2.x, you can get this from the [wheezy-backports](#) repo.

## Configuring `nslcd.conf`

`/etc/nslcd.conf` is the main configuration file that needs to be changed after the package is installed. The [nslcd.conf man page](#) details all the available configuration options.

Here is an [example configuration](#) using Cumulus Linux.

## Troubleshooting LDAP Authentication

By default, password and group information is cached by the `nscd` daemon. It is recommended when setting up LDAP authentication for the first time, to turn off this service using `service nscd stop`.

Stop the `nslcd` service and run it in debug mode. Debug mode works whether you are using LDAP over SSL (port 636) or an unencrypted LDAP connection (port 389).

```
cumulus@switch:~$ sudo service nslcd stop
cumulus@switch:~$ sudo nslcd -d
```

## Common Problems

- `nslcd` cannot read the SSL certificate. `nslcd` will report a “Permission denied” error in the debug during server connection negotiation. The sniffer trace output will show only a TCP handshake and then a TCP FIN from the switch. Check the permission on each directory in the path of the root SSL certificate. Ensure that is is readable by the `nslcd` user.
- The FQDN on the LDAP URI does not match the SSL FQDN exactly.
- The search filter returns wrong results. Check for typos in the search filter. Use `ldapsearch` to test your filter. For example:

```
In $HOME/.ldaprc configure basic ldapsearch parameters
-----
URI: ldaps://myadserver.rtp.example.test
BASE ou=support,dc=rtp,dc=example,dc=test
TLS_CACERT /etc/ssl/certs/rtp-example-ca.crt
-----

# ldapsearch \
-D 'CN=cumulus admin,CN=Users,DC=rtp,DC=example,DC=test' \
-w '1Q2w3e4r!' \
"(&(ObjectClass=user) \
(memberOf=cn=cumuluslnxadm,ou=groups,ou=support,dc=rtp,
dc=example, dc=test))"
```

## Configuring LDAP Authorization

In the `/etc/nslcd.conf` file, the "filter" keyword defines an LDAP search filter. Use this search filter to only show the users and or groups one desires. In the example below, only users in the `cumuluslnxadm` group are shown in the passed database:

```
# This filter says to get all users who are part of the cumuluslnxadm group.  
filter passwd (&(Objectclass=user)(!(objectClass=computer))  
(memberOf=cn=cumuluslnxadm,ou=groups,ou=support,dc=rtp,dc=example,dc=test))
```

## A Longer Example

A longer, more complete example for configuring LDAP is available on our [knowledge base](#).

## References

- <https://wiki.debian.org/LDAP/PAM>
- <https://raw.githubusercontent.com/arthurdejong/nss-pam-ldapd/master/nslcd.conf>
- <http://backports.debian.org/Instructions/>

## Netfilter - ACLs

Netfilter is the packet filtering framework in Cumulus Linux, as well as every other Linux distribution. `iptables`, `ip6tables` and `ebtables` are userspace tools in Linux to administer filtering rules for IPv4 packets, IPv6 packets and Ethernet frames respectively. `cl-acltool` is the userspace tool to administer filtering rules on Cumulus Linux, and is the only tool for configuring ACLs in Cumulus Linux.

`cl-acltool` operates on a series of configuration files, and uses `iptables`, `ip6tables` and `ebtables` to install rules into the kernel. In addition to programming rules in the kernel, `cl-acltool` programs rules in hardware for interfaces involving switch port interfaces, which `iptables`, `ip6tables` and `ebtables` do not do on their own.

## Contents

(Click to expand)

- [Contents \(see page 76\)](#)
- [Commands \(see page 77\)](#)
- [Files \(see page 77\)](#)
- [Netfilter Framework in the Cumulus Linux Kernel \(see page 77\)
  - \[Limitations on Number of Rules \\(see page 78\\)\]\(#\)
  - \[Enabling Nonatomic Updates \\(see page 78\\)\]\(#\)
  - \[ebtables and Memory Spaces \\(see page 79\\)\]\(#\)
  - \[Memory Spaces with Multiple Commands Line Options \\(see page 79\\)\]\(#\)](#)
- [Installing Packet Filtering \(ACL\) Rules using cl-acltool \(see page 80\)
  - \[Specifying which Policy Files to Install \\(see page 82\\)\]\(#\)](#)

- Managing ACL Rules with cl-acltool (see page 83)
  - Further Examples (see page 84)
- cl-acltool and Network Troubleshooting (see page 84)
- Policing Control Plane and Data Plane Traffic (see page 84)
- Useful Links (see page 85)
- Caveats and Errata (see page 86)
  - Not All Rules Supported (see page 86)
  - iptables Interactions with cl-acltool (see page 86)
  - Where to Assign Rules (see page 87)
  - Generic Error Message Displayed after ACL Rule Installation Failure (see page 87)

## **Commands**

- cl-acltool
- ebtables
- iptables
- ip6tables

## **Files**

- /etc/cumulus/acl/policy.conf
- /etc/cumulus/acl/policy.d/

## **Netfilter Framework in the Cumulus Linux Kernel**

Netfilter uses a table-based system for packet filtering. Tables are hooks in the kernel for packet filtering. Each table has a set of default *chains*, or categories of ACL rules. Each chain contains packet filter rules.

The default table in Netfilter is the *filter* table. The three chains in the filter table are:

- INPUT chain, for network traffic going to the switch
- OUTPUT chain, for traffic emanating from the switch
- FORWARD chain, for traffic being forwarded or routed through the switch

Cumulus Linux, like all Linux distributions, divides ACLs into chains. ACLs are handled both in hardware and software depending on which chain you use.

Data to Filter	iptables Chain	Hardware Accelerated?
Data plane egress	FORWARD (-o)	Yes
Data plane ingress	FORWARD (-i)	Yes
Control plane input	INPUT	Yes
Control plane output	OUTPUT	No

## ***Limitations on Number of Rules***

The maximum number of rules that can be handled in hardware is a function of the platform type (Apollo2, Firebolt2, Triumph, Trident, Trident+ or Trident II) and a mix of IPv4 and/or IPv6. See the [HCL](#) to determine which switches operate on these platforms.

### ***Apollo2 and Triumph2 Limits***

Direction	Atomic Mode IPv4 Rules	Atomic Mode IPv6 Rules	Nonatomic Mode IPv4 Rules	Nonatomic Mode IPv6 Rules
Ingress	2048	1024	4096	2048
Egress	512	256	1024	512

### ***Firebolt2 Limits***

Direction	Atomic Mode IPv4 Rules	Atomic Mode IPv6 Rules	Nonatomic Mode IPv4 Rules	Nonatomic Mode IPv6 Rules
Ingress	1024	512	2048	1024
Egress	512	256	512	256

### ***Trident/Trident+ Limits***

Direction	Atomic Mode IPv4 Rules	Atomic Mode IPv6 Rules	Nonatomic Mode IPv4 Rules	Nonatomic Mode IPv6 Rules
Ingress	384	384	1024	1024
Egress	512	256	1024	512

### ***Trident II Limits***

Direction	Atomic Mode IPv4 Rules	Atomic Mode IPv6 Rules	Nonatomic Mode IPv4 Rules	Nonatomic Mode IPv6 Rules
Ingress	1024	1024	2048	2048
Egress	512	256	1024	512

## Enabling Nonatomic Updates

You can enable nonatomic updates for `switchd`, which offer better scaling because all hardware resources are used to actively impact traffic. With atomic updates, half of the hardware resources are on standby and do not actively impact traffic.

To always start `switchd` with nonatomic updates:

1. Edit `/etc/cumulus/switchd.conf`.
2. Add the following line to the file:

```
acl.non_atomic_update_mode = TRUE
```

3. Restart `switchd` (see page 90):

```
cumulus@switch:~$ sudo service switchd restart
```



During nonatomic updates, traffic is stopped first, and enabled after the new configuration is written into the hardware completely.

## ebtables and Memory Spaces

`ebtables` rules are put into either the IPv4 or IPv6 memory space depending on whether the rule utilizes IPv4 or IPv6 to make a decision. L2-only rules, which match the MAC address, are put into the IPv4 memory space.

## Memory Spaces with Multiple Commands Line Options

INPUT and ingress (`FORWARD -i`) rules occupy the same memory space. A rule counts as ingress if the `-i` option is set. If both input and output options (`-i` and `-o`) are set, the rule is considered as ingress and shares that memory space. For example:

```
-A FORWARD -i swp1 -o swp2 -s 10.0.14.2 -d 10.0.15.8 -p tcp -j ACCEPT
```



If you set an output flag with the INPUT chain you will get an error. For example, running `cl-acltool -i` on the following rule:

```
-A FORWARD,INPUT -i swp1 -o swp2 -s 10.0.14.2 -d 10.0.15.8 -p tcp -j  
ACCEPT
```

generates the following error:

```
error: line 2 : output interface specified with INPUT chain
error processing rule '-A FORWARD,INPUT -i swp1 -o swp2 -s 10.0.14.2
-d 10.0.15.8 -p tcp -j ACCEPT'
```

However, simply removing the `-o` option and interface would make it a valid rule.

## ***Installing Packet Filtering (ACL) Rules using cl-acltool***

`cl-acltool` takes access control list (ACL) rules input in files. Each ACL policy file contains `iptables`, `ip6tables` and `ebtables` categories under the tags `[iptables]`, `[ip6tables]` and `[ebtables]` respectively.

Each rule in an ACL policy must be assigned to one of the rule categories above.

See `man cl-acltool(5)` for ACL rule details. For `iptables` rule syntax, see `man iptables(8)`. For `ip6tables` rule syntax, see `man ip6tables(8)`. For `ebtables` rule syntax, see `man ebtables(8)`.

See `man cl-acltool(5)` and `man cl-acltool(8)` for further details on using `cl-acltool`; however some examples are listed below, and more are listed in the Cumulus Networks [Help Center](#).

The default directory for ACL policy files is `/etc/cumulus/acl/policy.d`. By default, all `*.rules` files in this directory are included in `/etc/cumulus/acl/policy.conf`. And by default all files included in this `policy.conf` file are installed when the switch boots up.

Here is an example ACL policy file:

```
[iptables]
-A INPUT --in-interface swp1 -p tcp --dport 80 -j ACCEPT
-A FORWARD --in-interface swp1 -p tcp --dport 80 -j ACCEPT

[ip6tables]
-A INPUT --in-interface swp1 -p tcp --dport 80 -j ACCEPT
-A FORWARD --in-interface swp1 -p tcp --dport 80 -j ACCEPT

[ebtables]
-A INPUT -p IPv4 -j ACCEPT
-A FORWARD -p IPv4 -j ACCEPT
```

Variables can be used to specify chain and interface lists to ease administration of rules:

```
INGRESS = swp+
INPUT_PORT_CHAIN = INPUT,FORWARD
```

```
[iptables]
-A $INPUT_PORT_CHAIN --in-interface $INGRESS -p tcp --dport 80 -j ACCEPT

[ip6tables]
-A $INPUT_PORT_CHAIN --in-interface $INGRESS -p tcp --dport 80 -j ACCEPT

[ebtables]
-A INPUT -p IPv4 -j ACCEPT
```

ACL rules for the system can be written into multiple files under the default /etc/cumulus/acl/policy.d/ directory. Ordering of rules during install follow the sorted order of the files based on file names.

Use multiple files support to stack rules. The example below shows two rules files separating rules for management and datapath traffic:

```
cumulus@switch:~$ ls /etc/cumulus/acl/policy.d/
00sample_mgmt.rules
01sample_datapath.rules

cumulus@switch:~$ cat /etc/cumulus/acl/policy.d/00sample_mgmt.rules
INGRESS_INTF = swp+
INGRESS_CHAIN = INPUT

[iptables]
# protect the switch management
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -s 10.0.14.2 -d 10.0.15.8 -p
tcp -j ACCEPT
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -s 10.0.11.2 -d 10.0.12.8 -p
tcp -j ACCEPT
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -d 10.0.16.8 -p udp -j DROP

cumulus@switch:~$ cat /etc/cumulus/acl/policy.d/01sample_datapath.rules
INGRESS_INTF = swp+
INGRESS_CHAIN = INPUT, FORWARD

[iptables]
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -s 192.0.2.5 -p icmp -j
ACCEPT
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -s 192.0.2.6 -d 192.0.2.4 -j
DROP
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -s 192.0.2.2 -d 192.0.2.8 -j
DROP
```

Install all ACL policies under a directory:

```
cumulus@switch:~$ sudo cl-acltool -i -P ./rules
Reading files under rules
Reading rule file ./rules/01_http_rules.txt ...
Processing rules in file ./rules/01_http_rules.txt ...
Installing acl policy ...
Done.
```

Install all rules and policies included in /etc/cumulus/acl/policy.conf:

```
cumulus@switch:~$ sudo cl-acltool -i
```

## ***Specifying which Policy Files to Install***

By default, any .rules file you configure in /etc/cumulus/acl/policy.d/ will be installed by Cumulus Linux. To add other policy files to an ACL, you need to include them in /etc/cumulus/acl/policy.conf. For example, in order for Cumulus Linux to install a rule in a policy file called 01\_new.acl, you would add include /etc/cumulus/acl/policy.d/01\_new.acl to policy.conf, as in this example:

```
cumulus@switch:~$ sudo vi /etc/cumulus/acl/policy.conf

#
# This file is a master file for acl policy file inclusion
#
# Note: This is not a file where you list acl rules.
#
# This file can contain:
# - include lines with acl policy files
#   example:
#       include <filepath>
#
#   see manpage cl-acltool(5) and cl-acltool(8) for how to write policy
files
#


include /etc/cumulus/acl/policy.d/*.rules
include /etc/cumulus/acl/policy.d/01_new.acl
```

## Managing ACL Rules with *cl-acltool*

You manage Cumulus Linux ACLs with *cl-acltool*. Rules are first written to the *iptables* chains, as described above, and then synced to hardware via *switchd*.

To examine the current state of chains and list all installed rules, run:

```
cumulus@switch:~$ sudo cl-acltool -L all
----- Listing rules of type iptables:
-----
TABLE filter :
Chain INPUT (policy ACCEPT 90 packets, 14456 bytes)
pkts bytes target prot opt in out source destination
0 0 DROP all -- swp+ any 240.0.0.0/5 anywhere
0 0 DROP all -- swp+ any loopback/8 anywhere
0 0 DROP all -- swp+ any base-address.mcast.net/8 anywhere
0 0 DROP all -- swp+ any 255.255.255.255 anywhere
...
...
```

To list installed rules using native *iptables*, *ip6tables* and *ebtables*, run these commands:

```
cumulus@switch:~$ sudo iptables -L
cumulus@switch:~$ sudo ip6tables -L
cumulus@switch:~$ sudo ebttables -L
```

To flush all installed rules, run:

```
cumulus@switch:~$ sudo cl-acltool -F all
```

To flush only the IPv4 *iptables* rules, run:

```
cumulus@switch:~$ sudo cl-acltool -F ip
```

If the install fails, ACL rules in the kernel and hardware are rolled back to previous state. Errors from programming rules in kernel or BCM hardware are reported appropriately.

## Further Examples

More examples demonstrating how to use `cl-acltool` are available in the [Help Center](#).

### ***cl-acltool and Network Troubleshooting***

You use `cl-acltool` for both system diagnostics and troubleshooting the whole network. See [Network Troubleshooting](#) (see page 442) for information on using ACLs for [counting rules](#) (see page 445) as well as monitoring packets via [SPAN](#) and [ERSPAN](#) (see page 446).

### ***Policing Control Plane and Data Plane Traffic***

You can configure quality of service for traffic on both the control plane and the data plane. By using QoS policers, you can rate limit traffic so incoming packets get dropped if they exceed specified thresholds.



Counters on POLICE ACL rules in iptables do not currently show the packets that are dropped due to those rules.

Use the `POLICE` target with `iptables`. `POLICE` takes these arguments:

- `--set-class value`: Sets the system internal class of service queue configuration to *value*.
- `--set-rate value`: Specifies the maximum rate in kilobytes (KB) or packets.
- `--set-burst value`: Specifies the number of packets or kilobytes (KB) allowed to arrive sequentially.
- `--set-mode string`: Sets the mode in KB (kilobytes) or *pkt* (packets) for rate and burst size.

For example, to rate limit the incoming traffic on `swp1` to 400 packets/second with a burst of 100 packets/second and set the class of the queue for the policed traffic as 0, set this rule in your appropriate `.rules` file:

```
-A INPUT --in-interface swp1 -j POLICE --set-mode pkt --set-rate 400 --
set-burst 100 --set-class 0
```

Here is another example of control plane ACL rules to lock down the switch. This is specified in `/etc/cumulus/acl/policy.d/00control_plane.rules`:

```
INGRESS_INTF = swp+
INGRESS_CHAIN = INPUT
INNFWD_CHAIN = INPUT,FORWARD
MARTIAN_SOURCES_4 = "240.0.0.0/5,127.0.0.0/8,224.0.0.0/8,255.255.255.255/32"
MARTIAN_SOURCES_6 = "ff00::/8,::128,::ffff:0.0.0.0/96,::1/128"

#Custom Policy Section
SSH_SOURCES_4 = "192.168.0.0/24"
```

```

NTP_SERVERS_4 = "192.168.0.1/32,192.168.0.4/32"
DNS_SERVERS_4 = "192.168.0.1/32,192.168.0.4/32"
SNMP_SERVERS_4 = "192.168.0.1/32"

[iptables]
-A $INNFWD_CHAIN --in-interface $INGRESS_INTF -s $MARTIAN_SOURCES_4 -j DROP
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -p ospf -j POLICE --set-mode
pkt --set-rate 2000 --set-burst 2000 --set-class 7
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -p tcp --dport bgp -j POLICE
--set-mode pkt --set-rate 2000 --set-burst 2000 --set-class 7
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -p tcp --sport bgp -j POLICE
--set-mode pkt --set-rate 2000 --set-burst 2000 --set-class 7
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -p icmp -j POLICE --set-mode
pkt --set-rate 100 --set-burst 40 --set-class 2
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -p udp --dport bootps:bootpc
-j POLICE --set-mode pkt --set-rate 100 --set-burst 100 --set-class 2
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -p tcp --dport bootps:bootpc
-j POLICE --set-mode pkt --set-rate 100 --set-burst 100 --set-class 2
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -p igmp -j POLICE --set-mode
pkt --set-rate 300 --set-burst 100 --set-class 6

# Custom policy
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -p tcp --dport 22 -s
$SSH_SOURCES_4 -j ACCEPT
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -p udp --sport 123 -s
$NTP_SERVERS_4 -j ACCEPT
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -p udp --sport 53 -s
$DNS_SERVERS_4 -j ACCEPT
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -p udp --dport 161 -s
$SNMP_SERVERS_4 -j ACCEPT
# Allow UDP traceroute when we are the current TTL expired hop
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -p udp --dport 1024:65535 -m
ttl --ttl-eq 1 -j ACCEPT
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -j DROP

```

## Useful Links

- <http://www.netfilter.org/>
- <http://www.netfilter.org/documentation/HOWTO//packet-filtering-HOWTO-6.html>

## Caveats and Errata

### Not All Rules Supported

Please note that not all `iptables` and `ebtables` rules are fully supported. See `man cl-acltool(5)` for more information.

Further, there is no way to implement or extend transit filtering in software, and there is no way to hardware accelerate the OUTPUT chain. If the maximum number of rules for a particular table is exceeded, `cl-acltool -i` generates the following error:

```
error: hw sync failed (sync_acl hardware installation failed)
Rolling back ..
failed.
```

### *iptables* Interactions with `cl-acltool`

Since Cumulus Linux is a Linux operating system, the `iptables` commands can be used directly and will work. However, you should consider using `cl-acltool` instead because:

- Without using `cl-acltool`, rules are not installed into hardware.
- Running `cl-acltool -i` (the installation command) will reset all rules and delete anything that is not stored in `/etc/cumulus/acl/policy.conf`.

For example performing:

```
cumulus@switch:~$ sudo iptables -A INPUT -p icmp --icmp-type
echo-request -j DROP
```

Does work, and the rules appear when you run `cl-acltool -L`:

```
cumulus@switch:~$ sudo cl-acltool -L ip
-----
Listing rules of type iptables:
-----
TABLE filter :
Chain INPUT (policy ACCEPT 72 packets, 5236 bytes)
 pkts bytes target     prot opt in      out
 source          destination
               icmp --  any      any
               anywhere      anywhere      icmp echo-request
```

However, running `cl-acltool -i` or `reboot` will remove them. To ensure all rules that can be in hardware are hardware accelerated, place them in `/etc/cumulus/acl/policy.conf` and run `cl-acltool -i`.

## Where to Assign Rules

- If a switch port is assigned to a bond, any egress rules must be assigned to the bond.
- When using the OUTPUT chain, rules must be assigned to the source. For example, if a rule is assigned to the switch port in the direction of traffic but the source is a bridge (VLAN), the traffic won't be affected by the rule and must be applied to the bridge.
- If all transit traffic needs to have a rule applied, use the FORWARD chain, not the OUTPUT chain.

## Generic Error Message Displayed after ACL Rule Installation Failure

After an ACL rule installation failure, a generic error message like the following is displayed:

```
cumulus@switch:$ sudo cl-acltool -i -p 00control_plane.rules
Using user provided rule file 00control_plane.rules
Reading rule file 00control_plane.rules ...
Processing rules in file 00control_plane.rules ...
error: hw sync failed (sync_acl hardware installation failed)
Installing acl policy... Rolling back ..
failed.
```

## Configuring switchd

`switchd` is the daemon at the heart of Cumulus Linux. It communicates between the switch and Cumulus Linux, and all the applications running on Cumulus Linux.

The `switchd` configuration is stored in `/etc/cumulus/switchd.conf`.



Versions of Cumulus Linux prior to 2.1 stored the `switchd` configuration at `/etc/default/switchd`.

## Contents

(Click to expand)

- [Contents \(see page 87\)](#)
- [The switchd File System \(see page 88\)](#)
- [Configuring switchd Parameters \(see page 89\)](#)
- [Restarting switchd \(see page 90\)](#)
- [Commands \(see page 90\)](#)

- Configuration Files (see page 90)

## The switchd File System

switchd also exports a file system, mounted on `/cumulus/switchd`, that presents all the switchd configuration options as a series of files arranged in a tree structure. You can see the contents by parsing the switchd tree; run `tree /cumulus/switchd`. The output below is for a switch with one switch port configured:

```
cumulus@cumulus:~# sudo tree /cumulus/switchd/
/cumulus/switchd/
|-- config
|   |-- acl
|   |   |-- non_atomic_update_mode
|   |   `-- optimize_hw
|   |-- arp
|   |   |-- next_hops
|   |-- buf_util
|   |   |-- measure_interval
|   |   `-- poll_interval
|   |-- coalesce
|   |   |-- reducer
|   |   `-- timeout
|   |-- disable_internal_restart
|   |-- ignore_non_swps
|   |-- interface
|   |   |-- swp1
|   |   |   |-- storm_control
|   |   |   |-- broadcast
|   |   |   |-- multicast
|   |   |   `-- unknown_unicast
|   |-- logging
|   |-- route
|   |   |-- host_max_percent
|   |   |-- max_routes
|   |   `-- table
`-- stats
    `-- poll_interval
-- ctrl
    |-- acl
    |-- hal
    |   `-- resync
    |-- logger
    |-- netlink
    |   `-- resync
```

```

|   | -- resync
|   '-- sample
|       '-- ulog_channel
|-- run
|   '-- route_info
|       |-- ecmp_nh
|           |-- count
|           |-- max
|           '-- max_per_route
|       '-- host
|           |-- count
|           |-- count_v4
|           |-- count_v6
|           '-- max
|       '-- mac
|           |-- count
|           '-- max
`-- route
    |-- count_0
    |-- count_1
    |-- count_total
    |-- count_v4
    |-- count_v6
    |-- mask_limit
    |-- max_0
    |-- max_1
    '-- max_total
`-- version

```

## **Configuring switchd Parameters**

You can use cl-cfg to configure many switchd parameters at runtime (like ACLs, interfaces, and route table utilization), which minimizes disruption to your running switch. However, some options are read only and cannot be configured at runtime.

For example, to see data related to routes, run:

```

cumulus@cumulus:~$ sudo cl-cfg -a switchd | grep route
route.table = 254
route.max_routes = 32768
route.host_max_percent = 50
cumulus@cumulus:~$
```

To modify the configuration, run `cl-cfg -w`. For example, to set the buffer utilization measurement interval to 1 minute, run:

```
cumulus@cumulus:~$ sudo cl-cfg -w switchd buf_util.measure_interval=1
```

To verify that the value changed, use `grep`:

```
cumulus@cumulus:~# cl-cfg -a switchd | grep buf
buf_util.poll_interval = 0
buf_util.measure_interval = 1
```



You can get some of this information by running `cl-resource-query`; though you cannot update the `switchd` configuration with it.

## **Restarting switchd**

Whenever you modify any `switchd` hardware configuration file (typically changing any `*.conf` file that requires making a change to the switching hardware, like `/etc/cumulus/datapath/traffic.conf`), you must restart `switchd` for the change to take effect:

```
cumulus@switch:~$ sudo service switchd restart
```



You do not have to restart the `switchd` service when you update a network interface configuration (that is, edit `/etc/network/interfaces`).



Restarting `switchd` causes all network ports to reset in addition to resetting the switch hardware configuration.

## **Commands**

- `cl-cfg`

## **Configuration Files**

- `/etc/cumulus/switchd.conf`

## Power over Ethernet - PoE

Cumulus Linux supports Power over Ethernet (PoE), so certain Cumulus Linux switches can supply power from Ethernet switch ports to enabled devices over the Ethernet cables that connect them.

The [currently supported platforms](#) include:

- Accton AS4610-54P, a newly supported switch with an ARM processor



PoE+ and uPoE are not supported at this time.

### How It Works

When a powered device is connected to the switch via an Ethernet cable:

- If the available power is greater than the power required by the connected device, power is supplied to the switch port, and the device powers on
- If available power is less than the power required by the connected device and the switch port's priority is less than the port priority set on all powered ports, power is **not** supplied to the port
- If available power is less than the power required by the connected device and the switch port's priority is greater than the priority of a currently powered port, power is removed from lower priority port(s) and power is supplied to the port
- If the total consumed power exceeds the configured power limit of the power source, low priority ports are turned off. In the case of a tie, the port with the lower port number gets priority

For the Accton AS4610-54P switch, power is available as follows:

PSU 1	PSU 2	PoE Power Budget
920W	x	750W
x	920W	750W
920W	920W	1650W

The AS4610-54P has an LED on the front panel to indicate PoE status:

- Green: The `poed` daemon is running and no errors are detected
- Yellow: One or more errors are detected or the `poed` daemon is not running

### About Link State and PoE State

Link state and PoE state are completely independent of each other. When a link is brought down on particular port using `ip link <port> down`, power on that port is not turned off.

### LLDP with POE Attributes not Supported

Cumulus Linux does not support LLDP auto discovery and negotiation of PoE attributes via LLDP between the powered device and the switch.

## Configuring PoE

You use the `poectl` command utility to configure PoE on a [switch that supports](#) the feature. You can:

- Enable or disable PoE for a given switch port
- Set a switch port's PoE priority to one of three values: *low*, *high* or *critical*

By default, PoE is enabled on all Ethernet/1G switch ports, and these ports are set with a low priority. Switch ports can have low, high or critical priority.

To change the priority for one or more switch ports, run `poectl -p swp# [low|high|critical]`. For example:

```
cumulus@switch:~$ sudo poectl -p swp1-swp5,swp7 high
```

To disable PoE for one or more ports, run `poectl -d [port_numbers]`:

```
cumulus@switch:~$ sudo poectl -d swp1-swp5,swp7
```

To display PoE information for a set of switch ports, run `poectl -i [port_numbers]`:

```
cumulus@switch:~$ sudo poectl -i swp1-swp5,swp7
Port      Status          Priority   PD type   PD class Voltage Current
Power
-----
-----
swp1      searching      low        none       none      0.00 V     0
mA       0.00 W
swp2      searching      low        none       none      0.00 V     0
mA       0.00 W
swp3      searching      low        none       none      0.00 V     0
mA       0.00 W
swp4      disabled       low        none       none      0.00 V     0
mA       0.00 W
swp5      delivering power low        802.3af   1        53.94 V    39
mA       2.10 W
swp7      searching      high       none       none      0.00 V     0
mA       0.00 W
```

Or to see all the PoE information for a switch, run `poectl -s`:

```
cumulus@switch:~$ poectl -s
System power:
```

```
Total:      730.0 W
Used:       11.0 W
Available:  719.0 W
Connected ports:
    swp11, swp24, swp27, swp48
```

The set commands (priority, enable, disable) either succeed silently or display an error message if the command fails.

## ***poectl Arguments***

The **poectl** command takes the following arguments:

Argument	Description
-h, --help	Show this help message and exit
-i, --port-info PORT_LIST	Returns detailed information for the specified ports. For example: -i swp1-swp5,swp10
-p, --priority PORT_LIST PRIORITY	Sets priority for the specified ports: low, high, critical.
-d, --disable-ports PORT_LIST	Disables PoE operation on the specified ports.
-e, --enable-ports PORT_LIST	Enables PoE operation on the specified ports.
-s, --system	Returns PoE status for the entire switch.
-r, --reset PORT_LIST	Performs a hardware reset on the specified ports. Use this if one or more ports are stuck in an error state. This does not reset any configuration settings for the specified ports.
-v, --version	Displays version information.
--save	Saves the current configuration. The saved configuration is automatically loaded on system boot.
--load	Loads and applies the saved configuration.

## ***Logging poed Events***

The poed service logs the following events to syslog:

- When a switch provides power to a powered device
- When a device that was receiving power is removed
- When the power available to the switch changes
- Errors

## ***Man Pages***

man poectl

# Configuring and Managing Network Interfaces

`ifupdown` is the network interface manager for Cumulus Linux. Cumulus Linux 2.1 and later uses an updated version of this tool, `ifupdown2`.

For more information on network interfaces, see [Configuring Switch Port Attributes](#) (see page 108).



By default, `ifupdown` is quiet; use the verbose option `-v` when you want to know what is going on when bringing an interface down or up.

## Contents

(Click to expand)

- [Contents \(see page 95\)](#)
- [Commands \(see page 95\)](#)
- [Man Pages \(see page 96\)](#)
- [Configuration Files \(see page 96\)](#)
- [Basic Commands \(see page 96\)](#)
- [Bringing All auto Interfaces Up or Down \(see page 97\)](#)
- [ifupdown Behavior with Child Interfaces \(see page 97\)](#)
- [ifupdown2 Interface Dependencies \(see page 99\)](#)
  - [ifup Handling of Upper \(Parent\) Interfaces \(see page 102\)](#)
- [Configuring IP Addresses \(see page 103\)](#)
  - [Purging Existing IP Addresses on an Interface \(see page 104\)](#)
- [Specifying User Commands \(see page 104\)](#)
- [Sourcing Interface File Snippets \(see page 105\)](#)
- [Using Globs for Port Lists \(see page 105\)](#)
- [Using Templates \(see page 106\)](#)
- [Adding Descriptions to Interfaces \(see page 107\)](#)
- [Caveats and Errata \(see page 107\)](#)
- [Useful Links \(see page 108\)](#)

## Commands

- [ifdown](#)
- [ifquery](#)

- ifreload
- ifup
- mako-render

## Man Pages

The following man pages have been updated for `ifupdown2`:

- man ifdown(8)
- man ifquery(8)
- man ifreload
- man ifup(8)
- man ifupdown-addons-interfaces(5)
- man interfaces(5)

## Configuration Files

- /etc/network/interfaces

## Basic Commands

To bring up an interface or apply changes to an existing interface, run:

```
cumulus@switch:~$ sudo ifup <ifname>
```

To bring down a single interface, run:

```
cumulus@switch:~$ sudo ifdown <ifname>
```

Runtime Configuration (Advanced)



A runtime configuration is non-persistent, which means the configuration you create here does not persist after you reboot the switch.

To administratively bring an interface up or down, run:

```
cumulus@switch:~$ sudo ip link set dev swp1 {up|down}
```

If you specified *manual* as the address family, you must bring up that interface manually using `ifconfig`. For example, if you configured a bridge like this:

```
auto bridge01
iface bridge01 inet manual
```

You can only bring it up by running `ifconfig bridge01 up`.



`ifdown` always deletes logical interfaces after bringing them down. Use the `--admin-state` option if you only want to administratively bring the interface up or down.

To see the link and administrative state, use the `ip link show` command:

```
cumulus@switch:~$ ip link show dev swp13: swp1: <BROADCAST,MULTICAST,
UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP mode DEFAULT qlen 500
link/ether 44:38:39:00:03:c1 brd ff:ff:ff:ff:ff:ff
```

In this example, `swp1` is administratively UP and the physical link is UP (`LOWER_UP` flag). More information on interface administrative state and physical state can be found in [this knowledge base article](#).

## Bringing All auto Interfaces Up or Down

You can easily bring up or down all interfaces marked `auto` in `/etc/network/interfaces`. Use the `-a` option. For further details, see individual man pages for `ifup(8)`, `ifdown(8)`, `ifreload(8)`.

To administratively bring up all interfaces marked `auto`, run:

```
cumulus@switch:~$ sudo ifup -a
```

To administratively bring down all interfaces marked `auto`, run:

```
cumulus@switch:~$ sudo ifdown -a
```

To reload all network interfaces marked `auto`, use the `ifreload` command, which is equivalent to running `ifdown` then `ifup`, the one difference being that `ifreload` skips any configurations that didn't change):

```
cumulus@switch:~$ sudo ifreload -a
```

## ifupdown Behavior with Child Interfaces

By default, ifupdown recognizes and uses any interface present on the system — whether a VLAN, bond or physical interface — that is listed as a dependent of an interface. You are not required to list them in the `interfaces` file unless they need a specific configuration, for MTU, link speed, and so forth (see page 108). And if you need to delete a child interface, you should delete all references to that interface from the `interfaces` file.

For this example, `swp1` and `swp2` below do not need an entry in the `interfaces` file. The following stanzas defined in `/etc/network/interfaces` provide the exact same configuration:

With Child Interfaces Defined	Without Child Interfaces Defined
<pre>auto swp1 iface swp1  auto swp2 iface swp2  auto bridge iface bridge   bridge-vlan-aware yes   bridge-ports swp1 swp2   bridge-vids 1-100   bridge-pvid 1   bridge-stp on</pre>	<pre>auto bridge iface bridge   bridge-vlan-aware yes   bridge-ports swp1 swp2   bridge-vids 1-100   bridge-pvid 1   bridge-stp on</pre>

### Bridge in Traditional Mode - Example

For this example, `swp1.100` and `swp2.100` below do not need an entry in the `interfaces` file. The following stanzas defined in `/etc/network/interfaces` provide the exact same configuration:

With Child Interfaces Defined	Without Child Interfaces Defined
<pre>auto swp1.100 iface swp1.100 auto swp2.100 iface swp2.100 auto br-100 iface br-100   address 10.0.12.2/2 /24   address 2001:dad: beef::3/64   bridge-ports swp1.100 swp2.100   bridge-stp on</pre>	<pre>auto br-100 iface br-100   address 10.0.12.2/2 4   address 2001:dad: beef::3/64   bridge-ports swp1.1 00 swp2.100   bridge-stp on</pre>

For more information on the bridge in traditional mode vs the bridge in VLAN-aware mode, please read [this knowledge base article](#).

## ifupdown2 Interface Dependencies

`ifupdown2` understands interface dependency relationships. When `ifup` and `ifdown` are run with all interfaces, they always run with all interfaces in dependency order. When run with the interface list on the command line, the default behavior is to not run with dependents. But if there are any built-in dependents, they will be brought up or down.

To run with dependents when you specify the interface list, use the `--with-dependents` option. `--with-dependents` walks through all dependents in the dependency tree rooted at the interface you specify. Consider the following example configuration:

```
auto bond1
iface bond1
    address 100.0.0.2/16
    bond-slaves swp29 swp30
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lACP-rate 1
    bond-min-links 1
    bond-xmit-hash-policy layer3+4

auto bond2
iface bond2
    address 100.0.0.5/16
    bond-slaves swp31 swp32
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lACP-rate 1
    bond-min-links 1
    bond-xmit-hash-policy layer3+4

auto br2001
iface br2001
    address 12.0.1.3/24
    bridge-ports bond1.2001 bond2.2001
    bridge-stp on
```

Using `ifup --with-dependents br2001` brings up all dependents of `br2001`: `bond1.2001`, `bond2.2001`, `bond1`, `bond2`, `bond1.2001`, `bond2.2001`, `swp29`, `swp30`, `swp31`, `swp32`.

```
cumulus@switch:~$ sudo ifup --with-dependents br2001
```

Similarly, specifying `ifdown --with-dependents br2001` brings down all dependents of br2001: bond1.2001, bond2.2001, bond1, bond2, bond1.2001, bond2.2001, swp29, swp30, swp31, swp32.

```
cumulus@switch:~$ sudo ifdown --with-dependents br2001
```

- !** As mentioned earlier, `ifdown2` always deletes logical interfaces after bringing them down. Use the `--admin-state` option if you only want to administratively bring the interface up or down. In terms of the above example, `ifdown br2001` deletes `br2001`.

To guide you through which interfaces will be brought down and up, use the `--print-dependency` option to get the list of dependents.

Use `ifquery --print-dependency=list -a` to get the dependency list of all interfaces:

```
cumulus@switch:~$ sudo ifquery --print-dependency=list -a
lo : None
eth0 : None
bond0 : ['swp25', 'swp26']
bond1 : ['swp29', 'swp30']
bond2 : ['swp31', 'swp32']
br0 : ['bond1', 'bond2']
bond1.2000 : ['bond1']
bond2.2000 : ['bond2']
br2000 : ['bond1.2000', 'bond2.2000']
bond1.2001 : ['bond1']
bond2.2001 : ['bond2']
br2001 : ['bond1.2001', 'bond2.2001']
swp40 : None
swp25 : None
swp26 : None
swp29 : None
swp30 : None
swp31 : None
swp32 : None
```

To print the dependency list of a single interface, use:

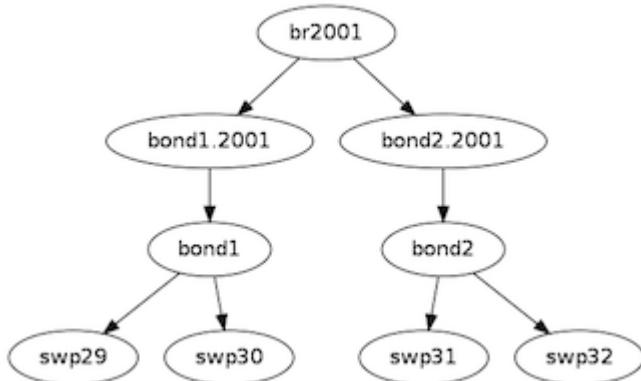
```
cumulus@switch:~$ sudo ifquery --print-dependency=list br2001
br2001 : ['bond1.2001', 'bond2.2001']
bond1.2001 : ['bond1']
bond2.2001 : ['bond2']
bond1 : ['swp29', 'swp30']
bond2 : ['swp31', 'swp32']
swp29 : None
swp30 : None
swp31 : None
```

```
swp32 : None
```

To print the dependency information of an interface in dot format:

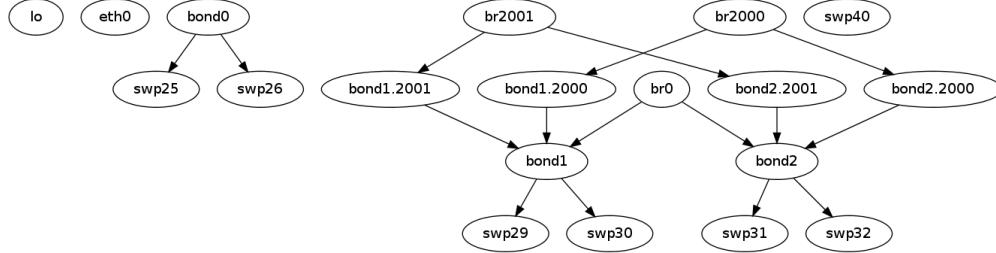
```
cumulus@switch:~$ sudo ifquery --print-dependency=dot br2001
/* Generated by GvGen v.0.9 (http://software.inl.fr/trac/wiki/GvGen)
*/
digraph G {
    compound=true;
    node1 [label="br2001"];
    node2 [label="bond1.2001"];
    node3 [label="bond2.2001"];
    node4 [label="bond1"];
    node5 [label="bond2"];
    node6 [label="swp29"];
    node7 [label="swp30"];
    node8 [label="swp31"];
    node9 [label="swp32"];
    node1->node2;
    node1->node3;
    node2->node4;
    node3->node5;
    node4->node6;
    node4->node7;
    node5->node8;
    node5->node9;
}
}
```

You can use dot to render the graph on an external system where dot is installed.



To print the dependency information of the entire interfaces file:

```
cumulus@switch:~$ sudo ifquery --print-dependency=dot -a >interfaces_all.dot
```



## ***ifup Handling of Upper (Parent) Interfaces***

When you run `ifup` on a logical interface (like a bridge, bond or VLAN interface), if the `ifup` resulted in the creation of the logical interface, by default it implicitly tries to execute on the interface's upper (or parent) interfaces as well. This helps in most cases, especially when a bond is brought down and up, as in the example below. This section describes the behavior of bringing up the upper interfaces.

Consider this example configuration:

```

auto br100
iface br100
    bridge-ports bond1.100 bond2.100

auto bond1
iface bond1
    bond-slaves swp1 swp2
  
```

If you run `ifdown bond1`, `ifdown` deletes bond1 and the VLAN interface on bond1 (bond1.100); it also removes bond1 from the bridge br100. Next, when you run `ifup bond1`, it creates bond1 and the VLAN interface on bond1 (bond1.100); it also executes `ifup br100` to add the bond VLAN interface (bond1.100) to the bridge br100.

As you can see above, implicitly bringing up the upper interface helps, but there can be cases where an upper interface (like br100) is not in the right state, which can result in warnings. The warnings are mostly harmless.

If you want to disable these warnings, you can disable the implicit upper interface handling by setting `skip_upperinterfaces=1` in `/etc/network/ifupdown2/ifupdown2.conf`.

With `skip_upperinterfaces=1`, you will have to explicitly execute `ifup` on the upper interfaces. In this case, you will have to run `ifup br100` after an `ifup bond1` to add bond1 back to bridge br100.



Although specifying a subinterface like `swp1.100` and then running `ifup swp1.100` will also result in the automatic creation of the `swp1` interface in the kernel, Cumulus Networks recommends you specify the parent interface `swp1` as well. A parent interface is one where any physical layer configuration can reside, such as `link-speed 1000` or `link-duplex full`.

It's important to note that if you only create `swp1.100` and not `swp1`, then you cannot run `ifup swp1` since you did not specify it.

## Configuring IP Addresses

In `/etc/network/interfaces`, list all IP addresses as shown below under the `iface` section (see `man interfaces` for more information):

```
auto swp1
iface swp1
    address 12.0.0.1/30
    address 12.0.0.2/30
```

The address method and address family are not mandatory. They default to `inet/inet6` and `static` by default, but `inet/inet6` **must** be specified if you need to specify `dhcp` or `loopback`:

```
auto lo
iface lo inet loopback
```

You can specify both IPv4 and IPv6 addresses in the same `iface` stanza:

```
auto swp1
iface swp1
    address 192.0.2.1/30
    address 192.0.2.2/30
    address 2001:DB8::1/126
```

### Runtime Configuration (Advanced)



A runtime configuration is non-persistent, which means the configuration you create here does not persist after you reboot the switch.

To make non-persistent changes to interfaces at runtime, use `ip addr add`:

```
cumulus@switch:~$ sudo ip addr add 192.0.2.1/30 dev swp1
cumulus@switch:~$ sudo ip addr add 2001:DB8::1/126 dev swp1
```

To remove an addresses from an interface, use `ip addr del`:

```
cumulus@switch:~$ sudo ip addr del 192.0.2.1/30 dev swp1
cumulus@switch:~$ sudo ip addr del 2001:DB8::1/126 dev swp1
```

See `man ip` for more details on the options available to manage and query interfaces.

To show the assigned address on an interface, use `ip addr show`:

```
cumulus@switch:~$ ip addr show dev swp1
3: swp1: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
    state UP qlen 500
        link/ether 44:38:39:00:03:c1 brd ff:ff:ff:ff:ff:ff
        inet 192.0.2.1/30 scope global swp1
            inet 192.0.2.2/30 scope global swp1
            inet6 2001:DB8::1/126 scope global tentative
                valid_lft forever preferred_lft forever
```

## Purging Existing IP Addresses on an Interface

By default, `ifupdown2` purges existing IP addresses on an interface. If you have other processes that manage IP addresses for an interface, you can disable this feature including the `address-purge` setting in the interface's configuration. For example, add the following to the interface configuration in `/etc/network/interfaces`:

```
auto swp1
iface swp1
    address-purge no
```



Purging existing addresses on interfaces with multiple `iface` stanzas is not supported. Doing so can result in the configuration of multiple addresses for an interface after you change an interface address and reload the configuration with `ifreload -a`. If this happens, you must shut down and restart the interface with `ifup` and `ifdown`, or manually delete superfluous addresses with `ip address delete specify.ip.address.here/mask dev DEVICE`. See also the [Caveats and Errata \(see page 107\)](#) section below for some cautions about using multiple `iface` stanzas for the same interface.

## Specifying User Commands

You can specify additional user commands in the `interfaces` file. As shown in the example below, the interface stanzas in `/etc/network/interfaces` can have a command that runs at pre-up, up, post-up, pre-down, down, and post-down:

```
auto swp1
iface swp1
    address 12.0.0.1/30
    up /sbin/foo bar
```

Any valid command can be hooked in the sequencing of bringing an interface up or down, although commands should be limited in scope to network-related commands associated with the particular interface.

For example, it wouldn't make sense to install some Debian package on `ifup` of `swp1`, even though that is technically possible. See `man interfaces` for more details.

## Sourcing Interface File Snippets

Sourcing interface files helps organize and manage the `interfaces(5)` file. For example:

```
cumulus@switch:~$ cat /etc/network/interfaces
# The loopback network interface
auto lo
iface lo inet loopback

# The primary network interface
auto eth0
iface eth0 inet dhcp

source /etc/network/interfaces.d/bond0
```

The contents of the sourced file used above are:

```
cumulus@switch:~$ cat /etc/network/interfaces.d/bond0
auto bond0
iface bond0
    address 14.0.0.9/30
    address 2001:ded:beef:2::1/64
    bond-slaves swp25 swp26
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lACP-rate 1
    bond-min-links 1
    bond-xmit-hash-policy layer3+4
```

## Using Globs for Port Lists

Some modules support globs to define port lists (that is, a range of ports). You can use the `glob` keyword to specify bridge ports and bond slaves:

```
auto br0
iface br0
    bridge-ports glob swp1-6.100
```

```
auto br1
iface br1
    bridge-ports glob swp7-9.100  swp11.100 glob swp15-18.100
```

## Using Templates

ifupdown2 supports Mako-style templates. The Mako template engine is run over the `interfaces` file before parsing.

Use the template to declare cookie-cutter bridges in the `interfaces` file:

```
%for v in [11,12]:
auto vlan${v}
iface vlan${v}
    address 10.20.${v}.3/24
    bridge-ports glob swp19-20.${v}
    bridge-stp on
%endfor
```

And use it to declare addresses in the `interfaces` file:

```
%for i in [1,12]:
auto swp${i}
iface swp${i}
    address 10.20.${i}.3/24
```



Regarding Mako syntax, use square brackets ([1,12]) to specify a list of individual numbers (in this case, 1 and 12). Use `range(1,12)` to specify a range of interfaces.



You can test your template and confirm it evaluates correctly by running `mako-render /etc/network/interfaces`.



For more examples of configuring Mako templates, read this [knowledge base article](#).

## Adding Descriptions to Interfaces

You can add descriptions to the interfaces configured in /etc/network/interfaces by using the alias keyword. For example:

```
auto swp1
iface swp1
    alias swp1 hypervisor_port_1
```

You can query interface descriptions by running ip link show. The alias appears on the alias line:

```
cumulus@switch$ ip link show swp1
3: swp1: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc pfifo_fast
state DOWN mode DEFAULT qlen 500
    link/ether aa:aa:aa:aa:aa:bc brd ff:ff:ff:ff:ff:ff
    alias hypervisor_port_1
```

Interface descriptions also appear in the SNMP OID (see page 408) IF-MIB::ifAlias.

## Caveats and Errata

While ifupdown2 supports the inclusion of multiple iface stanzas for the same interface, Cumulus Networks recommends you use a single iface stanza for each interface, if possible.

There are cases where you must specify more than one iface stanza for the same interface. For example, the configuration for a single interface can come from many places, like a template or a sourced file.

If you do specify multiple iface stanzas for the same interface, make sure the stanzas do not specify the same interface attributes. Otherwise, unexpected behavior can result.

For example, swp1 is configured in two places:

```
cumulus@switch:~$ cat /etc/network/interfaces
source /etc/interfaces.d/speed_settings

auto swp1
iface swp1
    address 10.0.14.2/24
```

As well as /etc/interfaces.d/speed\_settings

```
cumulus@switch:~$ cat /etc/interfaces.d/speed_settings  
  
auto swp1  
iface swp1  
    link-speed 1000  
    link-duplex full
```

`ifupdown2` correctly parses a configuration like this because the same attributes are not specified in multiple `iface` stanzas.

And, as stated in the note above, you cannot purge existing addresses on interfaces with multiple `iface` stanzas.

## Useful Links

- <http://wiki.debian.org/NetworkConfiguration>
- <http://www.linuxfoundation.org/collaborate/workgroups/networking/bonding>
- <http://www.linuxfoundation.org/collaborate/workgroups/networking/bridge>
- <http://www.linuxfoundation.org/collaborate/workgroups/networking/vlan>

## Configuring Switch Port Attributes

This chapter discusses the various network interfaces on a switch running Cumulus Linux.

### Contents

(Click to expand)

- Contents (see page 108)
- Commands (see page 109)
- Man Pages (see page 109)
- Configuration Files (see page 109)
- Interface Types (see page 109)
- Settings (see page 109)
  - Port Speed and Duplexing (see page 110)
  - Auto-negotiation (see page 111)
  - MTU (see page 111)
- Configuring Breakout Ports (see page 113)
  - Breaking out a 40G port into 4x10G Ports (see page 114)
  - Combining Four 10G Ports into One 40G Port (see page 114)
- Logical Switch Port Limitations (see page 115)
- Verification and Troubleshooting Commands (see page 116)
  - Statistics (see page 116)
  - Querying SFP Port Information (see page 117)

- Useful Links (see page 117)

## Commands

- ethtool
- ip

## Man Pages

- man ethtool
- man interfaces
- man ip
- man ip addr
- man ip link

## Configuration Files

- /etc/network/interfaces

## Interface Types

Cumulus Linux exposes network interfaces for several types of physical and logical devices:

- lo, network loopback device
- ethN, switch management port(s), for out of band management only
- swpN, switch front panel ports
- (optional) brN, bridges (IEEE 802.1Q VLANs)
- (optional) bondN, bonds (IEEE 802.3ad link aggregation trunks, or port channels)

## Settings

You can set the MTU, speed, duplex and auto-negotiation settings under a physical or logical interface stanza:

```
auto swp1
iface swp1
    address 10.1.1.1/24
    mtu 9000
    link-speed 10000
    link-duplex full
    link-autoneg off
```

To load the updated configuration, run the `ifreload -a` command:

```
cumulus@switch:~$ sudo ifreload -a
```

## Port Speed and Duplexing

Cumulus Linux supports both half- and **full-duplex** configurations. Supported port speeds include 1G, 10G and 40G. Set the speeds in terms of Mbps, where the setting for 1G is 1000, 10G is 10000 and 40G is 40000.

You can create a persistent configuration for port speeds in `/etc/network/interfaces`. Add the appropriate lines for each switch port stanza. For example:

```
auto swp1
iface swp1
    address 10.1.1.1/24
    link-speed 10000
    link-duplex full
```



If you specify the port speed in `/etc/network/interfaces`, you must also specify the duplex mode setting along with it; otherwise, `ethtool` defaults to half duplex.

You can also configure these settings at run time, using `ethtool`.

### Runtime Configuration (Advanced)



A runtime configuration is non-persistent, which means the configuration you create here does not persist after you reboot the switch.

You can use `ethtool` to configure duplexing and the speed for your switch ports. You must specify both port speed and duplexing in the `ethtool` command; auto-negotiation is optional. The following examples use `sdp1`.

- To set the port speed to 1G, run:

```
ethtool -s sdp1 speed 1000 duplex full
```

- To set the port speed to 10G, run:

```
ethtool -s sdp1 speed 10000 duplex full
```

- To enable duplexing, run:

```
ethtool -s sdp1 speed 10000 duplex full|half
```

## Port Speed Limitations

Ports can be configured to one speed less than their maximum speed.

Switch port Type	Lowest Configurable Speed
1G	100 Mb
10G	1 Gigabit (1000 Mb)
40G	10G*

\*Requires the port to be converted into a breakout port.

## Auto-negotiation

You can enable or disable **auto-negotiation** (that is, set it *on* or *off*) on a switch port.

```
auto swp1
iface swp1
    link-autoneg off
```

Runtime Configuration (Advanced)



A runtime configuration is non-persistent, which means the configuration you create here does not persist after you reboot the switch.

You can use `ethtool` to configure auto-negotiation for your switch ports. The following example use `swp1`:

- To enable or disable auto-negotiation, run:

```
ethtool -s swp1 speed 10000 duplex full autoneg on|off
```

## MTU

Interface MTU applies to the management port, front panel port, bridge, VLAN subinterfaces and bonds.

```
auto swp1
iface swp1
    mtu 9000
```

## Runtime Configuration (Advanced)



A runtime configuration is non-persistent, which means the configuration you create here does not persist after you reboot the switch.

To set swp1 to Jumbo Frame MTU=9000, use `ip link set`:

```
cumulus@switch:~$ sudo ip link set dev swp1 mtu 9000
cumulus@switch:~$ ip link show dev swp1
3: swp1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9000 qdisc pfifo_fast
    state UP mode DEFAULT qlen 500
        link/ether 44:38:39:00:03:c1 brd ff:ff:ff:ff:ff:ff
```



You must take care to ensure there are no MTU mismatches in the conversation path. MTU mismatches will result in dropped or truncated packets, degrading or blocking network performance.

When you are configuring MTU for a bridge, don't set MTU on the bridge itself; set it on the individual members of the bridge. The MTU setting is the lowest MTU setting of any interface that is a member of that bridge (that is, every interface specified in `bridge-ports` in the bridge configuration in the `interfaces` file), even if another bridge member has a higher MTU value. Consider this bridge configuration:

```
auto br0
iface br0
    bridge-ports bond1 bond2 bond3 bond4 peer5
    bridge-vlan-aware yes
    bridge-vids 100-110
    bridge-stp on
```

In order for br0 to have an MTU of 9000, set the MTU for each of the member interfaces (bond1 to bond 4, and peer5), to 9000 at minimum.

```
auto peer5
iface peer5
    bond-slaves swp3 swp4
    bond-mode 802.3ad
    bond-miimon 100
    bond-lacp-rate 1
```

```
bond-min-links 1
bond-xmit_hash_policy layer3+4
mtu 9000
```

When configuring MTU for a bond, configure the MTU value directly under the bond interface; the configured value is inherited by member links.

To show MTU, use `ip link show`:

```
cumulus@switch:~$ ip link show dev swp1
3: swp1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
state UP mode DEFAULT qlen 500
    link/ether 44:38:39:00:03:c1 brd ff:ff:ff:ff:ff:ff
```

## Configuring Breakout Ports

Cumulus Linux has the ability to:

- Break out 40G switch ports into four separate 10G ports for use with a breakout cable.
- Combine (also called *aggregating* or *ganging*) four 10G switch ports into one 40G port for use with a breakout cable ([not to be confused with a bond](#) (see page 158)).

A typical DAC (directly-attached copper) 40G 1xQSFP to 10G 4xSFP+ looks like this:



You configure breakout ports with the `/etc/cumulus/ports.conf` file. After you modify the configuration, restart `switchd` to push the new configuration (run `sudo service switchd restart`; [this interrupts network services](#) (see page 90)).

## Breaking out a 40G port into 4x10G Ports



/etc/cumulus/ports.conf varies across different hardware platforms. Check the current list of supported platforms on the [hardware compatibility list](#).

A snippet from the /etc/cumulus/ports.conf looks like this:

```
# QSFP+ ports
#
# <port label 49-52> = [4x10G|40G]
49=40G
50=40G
51=40G
52=40G
```

To change a 40G port to 4x10G ports, edit the /etc/cumulus/ports.conf file with a text editor (nano, vi, zile). Change 40G to 4x10G.

In the following example, switch port 49 is changed to a breakout port:

```
# QSFP+ ports
#
# <port label 49-52> = [4x10G|40G]
49=4x10G
50=40G
51=40G
52=40G
```

To load the change restart switchd:

```
cumulus@switch:~$ sudo service switchd restart
```

Many services depend on switchd. It is highly recommended to restart Cumulus Linux if possible in this situation.

## Combining Four 10G Ports into One 40G Port

To gang (aggregate) four 10G ports into one 40G port for use with a breakout cable, you must edit /etc/cumulus/ports.conf.



/etc/cumulus/ports.conf varies across different hardware platforms. Check the current list of supported platforms on the [hardware compatibility list](#).

A snippet from the `/etc/cumulus/ports.conf` looks like this:

```
# SFP+ ports#
# <port label 1-48> = [10G|40G/4]
1=10G
2=10G
3=10G
4=10G
5=10G
```

To change four 10G ports into one 40G port, edit the `/etc/cumulus/ports.conf` file with a text editor (nano, vi, zile). Change `10G` to `40G/4` for every port being ganged.

In the following example, switch ports `swp1-4` are changed to a ganged port:

```
# SFP+ ports#
# <port label 1-48> = [10G|40G/4]
1=40G/4
2=40G/4
3=40G/4
4=40G/4
5=10G
```

To load the change, restart `switchd`.

```
cumulus@switch:~$ sudo service switchd restart
```

Many services depend on `switchd`. It is highly recommended to restart Cumulus Linux if possible in this situation.



- You must gang four 10G ports in sequential order. For example, you cannot gang `swp1`, `swp10`, `swp20` and `swp40` together.
- The ports must be in increments of four, with the starting port being `swp1` (or `swp5`, `swp9`, or so forth); so you cannot gang `swp2`, `swp3`, `swp4` and `swp5` together.

## **Logical Switch Port Limitations**

40G switches with Trident II chipsets (check the *40G Portfolio* section of the [HCL](#)) can support a certain number of logical ports, depending upon the manufacturer.

Before you configure any logical/unganged ports on a switch, check the limitations listed in `/etc/cumulus/ports.conf`; this file is specific to each manufacturer.

For example, the Dell S6000 `ports.conf` file indicates the logical port limitation like this:

```
# ports.conf --
```

```

#
# This file controls port aggregation and subdivision. For example,
QSFP+
# ports are typically configurable as either one 40G interface or four
# 10G/1000/100 interfaces. This file sets the number of interfaces
per port
# while /etc/network/interfaces and ethtool configure the link speed f
or each
# interface.
#
# You must restart switchd for changes to take effect.
#
# The DELL S6000 has:
#   32 QSFP ports numbered 1-32
# These ports are configurable as 40G, split into 4x10G ports or
# disabled.
#
# The X pipeline covers QSFP ports 1 through 16 and the Y pipeline
# covers QSFP ports 17 through 32.
#
# The Trident2 chip can only handle 52 logical ports per pipeline.
#
# This means 13 is the maximum number of 40G ports you can ungang
# per pipeline, with the remaining three 40G ports set to
# "disabled". The 13 40G ports become 52 unganged 10G ports, which
# totals 52 logical ports for that pipeline.
#

```

The means the maximum number of ports for this Dell S6000 is 104.

## **Verification and Troubleshooting Commands**

### **Statistics**

High-level interface statistics are available with the `ip -s link` command:

```
cumulus@switch:~$ ip -s link show dev swp1
3: swp1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
state UP mode DEFAULT qlen 500
    link/ether 44:38:39:00:03:c1 brd ff:ff:ff:ff:ff:ff
    RX: bytes packets errors dropped overrun mcast
      21780      242       0       0       0      242
    TX: bytes packets errors dropped carrier collsns
      1145554     11325       0       0       0       0
```

Low-level interface statistics are available with `ethtool`:

```
cumulus@switch:~$ sudo ethtool -S swp1
```

```
NIC statistics:
HwIfInOctets: 21870
HwIfInUcastPkts: 0
HwIfInBcastPkts: 0
HwIfInMcastPkts: 243
HwIfOutOctets: 1148217
HwIfOutUcastPkts: 0
HwIfOutMcastPkts: 11353
HwIfOutBcastPkts: 0
HwIfInDiscards: 0
HwIfInL3Drops: 0
HwIfInBufferDrops: 0
HwIfInAclDrops: 0
HwIfInBlackholeDrops: 0
HwIfInDot3LengthErrors: 0
HwIfInErrors: 0
SoftInErrors: 0
SoftInDrops: 0
SoftInFrameErrors: 0
HwIfOutDiscards: 0
HwIfOutErrors: 0
HwIfOutQDrops: 0
HwIfOutNonQDrops: 0
SoftOutErrors: 0
SoftOutDrops: 0
SoftOutTxFifoFull: 0
HwIfOutQLen: 0
```

## ***Querying SFP Port Information***

You can verify SFP settings using `ethtool -m`. The following example shows the output for 1G and 10G modules:

```
cumulus@switch:~# sudo ethtool -m | egrep '(swp|RXPower :|TXPower :|EthernetComplianceCode)'

swp1: SFP detected
    EthernetComplianceCodes : 1000BASE-LX
    RXPower : -10.4479dBm
    TXPower : 18.0409dBm

swp3: SFP detected
    10GEthernetComplianceCode : 10G Base-LR
    RXPower : -3.2532dBm
    TXPower : -2.0817dBm
```

## ***Useful Links***

- <http://wiki.debian.org/NetworkConfiguration>

- <http://www.linuxfoundation.org/collaborate/workgroups/networking/vlan>
- <http://www.linuxfoundation.org/collaborate/workgroups/networking/bridge>
- <http://www.linuxfoundation.org/collaborate/workgroups/networking/bonding>

## Configuring Buffer and Queue Management

Hardware datapath configuration manages packet buffering, queueing, and scheduling in hardware. There are two configuration input files:

- `/etc/cumulus/datapath/traffic.conf`, which describes priority groups and assigns the scheduling algorithm and weights
- `/etc/bcm.d/datapath/datapath.conf`, which assigns buffer space and egress queues



Versions of these files prior to Cumulus Linux 2.1 are incompatible with Cumulus Linux 2.1 and later; using older files will cause `switchd` to fail to start and return an error that it cannot find the `/var/lib/cumulus/rc.datapath` file.

Each packet is assigned to an ASIC Class of Service (CoS) value based on the packet's priority value stored in the 802.1p (Class of Service) or DSCP (Differentiated Services Code Point) header field. The packet is assigned to a priority group based on the CoS value.

Priority groups include:

- *Control*: Highest priority traffic
- *Service*: Second-highest priority traffic
- *Lossless*: Traffic protected by priority flow control
- *Bulk*: All remaining traffic

A lossless traffic group is protected from packet drops by configuring the datapath to use priority pause. A lossless priority group requires a port group configuration, which specifies the ports configured for priority flow control and the additional buffer space assigned to each port for packets in the lossless priority group.

The scheduler is configured to use a hybrid scheduling algorithm. It applies strict priority to control traffic queues and a weighted round robin selection from the remaining queues. Unicast packets and multicast packets with the same priority value are assigned to separate queues, which are assigned equal scheduling weights.

Datapath configuration takes effect when you initialize `switchd`. Changes to the `traffic.conf` file require you to [restart `switchd` \(see page 90\)](#).

## Contents

(Click to expand)

- [Contents \(see page 118\)](#)
- [Commands \(see page 119\)](#)
- [Configuration Files \(see page 119\)](#)
- [Configuring Traffic Marking through ACL Rules \(see page 121\)](#)
- [Configuring Link Pause \(see page 122\)](#)
- [Useful Links \(see page 123\)](#)

- Caveats and Errata (see page 123)

## Commands

If you modify the configuration in the `/etc/cumulus/datapath/traffic.conf` file, you must restart `switchd` (see page 90) for the changes to take effect:

```
cumulus@switch:~$ sudo service switchd restart
```

## Configuration Files

The following configuration applies to 10G and 40G switches only (any switch on the Trident, Trident+, or Trident II platform).

- `/etc/cumulus/datapath/traffic.conf`: The default datapath configuration file.
- `/etc/cumulus/datapath/custom_traffic.conf`: An optional customized configuration file.

Sample traffic.conf file (Click to expand)

```
cumulus@switch:~$ cat /etc/cumulus/datapath/traffic.conf
#
# /etc/cumulus/datapath/traffic.conf
#
# packet header field used to determine the packet priority
level
# fields include {802.1p,
dscp}
traffic.packet_priority_source = 802.
1p
# remark packet priority
value
# fields include {802.1p,
none}
traffic.remark_packet_priority = none
# packet priority values assigned to each internal cos
value
# internal cos values {cos_0..
cos_7}
# (internal cos 3 has been reserved for CPU-generated
traffic)
# 802.1p values = {0..7}, dscp values = {0..63}
traffic.cos_0.packet_priorities = [0]
traffic.cos_1.packet_priorities = [1]
traffic.cos_2.packet_priorities = [2]
traffic.cos_3.packet_priorities = []
traffic.cos_4.packet_priorities = [3,4]
traffic.cos_5.packet_priorities = [5]
traffic.cos_6.packet_priorities = [6]
traffic.cos_7.packet_priorities = [7]
```

```
# priority groups
traffic.priority_group_list = [control, service, bulk]
# internal cos values assigned to each priority group
# each cos value should be assigned exactly once
# internal cos values {0..7}
priority_group.control.cos_list = [7]
priority_group.service.cos_list = [2]
priority_group.bulk.cos_list = [0,1,3,4,5,6]
# to configure a lossless priority group:
# -- uncomment the cos list config and assign cos value(s)
# -- uncomment port_group_0 configurations and set the lossless flag,
buffer si
ze,
ports
# -- (currently only one traffic group is allowed, with port range 'allports')
# priority_group.lossless.cos_list = []
# lossless port group
# -- lossless flag
arranging in: tiled
arranging in: tiled
# -- buffer size in bytes for each port
# -- port group
# priority_group.lossless.lossless_flag = true
# priority_group.lossless.port_group_0.port_buffer_bytes = 4096
# priority_group.lossless.port_group_0.port_range = allports
# to configure pause on a group of ports:
# uncomment the link pause port group list
# add or replace a port group name to the list
# populate the port set, e.g.
#     swp1-swp4,swp8,swp50s0-swp50s3
# enable pause frame transmit and/or pause frame receive
# link pause
# link_pause.port_group_list = [port_group_0]
# link_pause.port_group_0.port_set = swp1-swp4,swp6
# link_pause.port_group_0.rx_enable = true
# link_pause.port_group_0.tx_enable = true
# scheduling algorithm: algorithm values = {dwrr}
scheduling.algorithm = dwrr
# traffic group scheduling weight
# weight values = {0..127}
# '0' indicates strict priority
priority_group.control.weight = 0
priority_group.service.weight = 32
priority_group.bulk.weight = 16
priority_group.lossless.weight = 16
# To turn on/off Denial of service (DOS) prevention checks
dos_enable = false
# To enable cut-through forwarding
cut_through_enable = true
# Enable resilient hashing
#resilient_hash_enable = FALSE
```

```
# Resilient hashing flowset entries per ECMP group
# Valid values - 64, 128, 256, 512, 1024
#resilient_hash_entries_ecmp = 128
# Enable symmetric hashing
#symmetric_hash_enable = TRUE
# Set sflow/sample ingress cpu packet rate and burst in packets/sec
# Values: {0..16384}
#sflow.rate = 16384
#sflow.burst = 16384
#Specify the maximum number of paths per route entry.
# Maximum paths supported is 200.
# Default value 0 takes the number of physical ports as the max path
size.
#ecmp_max_paths = 0
```

## Configuring Traffic Marking through ACL Rules

You can mark traffic for egress packets through `iptables` or `ip6tables` rule classifications. To enable these rules, you do one of the following:

- Mark DSCP values in egress packets.
- Mark 802.1p CoS values in egress packets.

To enable traffic marking, use `cl-acltool`. Add the `-p` option to specify the location of the policy file. By default, if you don't include the `-p` option, `cl-acltool` looks for the policy file in `/etc/cumulus/acl/policy.d/`.

The `iptables`-`ip6tables`-based marking is supported via the following action extension:

```
-j SETQOS --set-dscp 10 --set-cos 5
```

You can specify one of the following targets for `SETQOS`:

Option	Description
<code>-set-cos INT</code>	Sets the datapath resource/queuing class value. Values are defined in <a href="#">IEEE_P802.1p</a> .
<code>-set-dscp value</code>	Sets the DSCP field in packet header to a value, which can be either a decimal or hex value.
<code>-set-dscp-class class</code>	Sets the DSCP field in the packet header to the value represented by the DiffServ class value. This class can be EF, BE or any of the CSxx or AFxx classes.



You can specify either `--set-dscp` or `--set-dscp-class`, but not both.

Here are two example rules:

```
[iptables]
-t mangle -A FORWARD -i --in-interface swp+ -p tcp --dport bgp -j SETQOS --
set-dscp 10 --set-cos 5

[ip6tables]
-t mangle -A FORWARD -i --in-interface swp+ -j SETQOS --set-dscp 10
```

You can put the rule in either the *mangle* table or the default *filter* table; the mangle table and filter table are put into separate TCAM slices in the hardware.

To put the rule in the mangle table, include `-t mangle`; to put the rule in the filter table, omit `-t mangle`.

## Configuring Link Pause

The PAUSE frame is a flow control mechanism that halts the transmission of the transmitter for a specified period of time. A server or other network node within the data center may be receiving traffic faster than it can handle it, thus the PAUSE frame. In Cumulus Linux, individual ports can be configured to execute link pause by:

- Transmitting pause frames when its ingress buffers become congested (TX pause enable) and/or
- Responding to received pause frames (RX pause enable).

Just like configuring buffer and queue management link pause is configured by editing `/etc/cumulus/datapath/traffic.conf`.

Here is an example configuration which turns of both types of link pause for swp2 and swp3:

```
# to configure pause on a group of ports:
# uncomment the link pause port group list
# add or replace a port group name to the list
# populate the port set, e.g.
# swp1-swp4,swp8,swp50s0-swp50s3
# enable pause frame transmit and/or pause frame receive
# link pause
link_pause.port_group_list = [port_group_0]
link_pause.port_group_0.port_set = swp2-swp3
link_pause.port_group_0.rx_enable = true
link_pause.port_group_0.tx_enable = true
```

A *port group* refers to one or more sequences of contiguous ports. Multiple port groups can be defined by:

- Adding a comma-separated list of port group names to the `port_group_list`.
- Adding the `port_set`, `rx_enable`, and `tx_enable` configuration lines for each port group.

You can specify the set of ports in a port group in comma-separated sequences of contiguous ports; you can see which ports are contiguous in `/var/lib/cumulus/porttab`. The syntax supports:

- A single port (swp1s0 or swp5)

- A sequence of regular swp ports (swp2-swp5)
- A sequence within a breakout swp port (swp6s0-swp6s3)
- A sequence of regular and breakout ports, provided they are all in a contiguous range. For example:

```
...
swp2
swp3
swp4
swp5
swp6s0
swp6s1
swp6s2
swp6s3
swp7
...
```

Restart `switchd` (see page 90) to allow link pause configuration changes to take effect:

```
cumulus@switch:~$ sudo service switchd restart
```

## ***Useful Links***

- [iptables-extensions man page](#)

## ***Caveats and Errata***

- You can configure Quality of Service (QoS) for 10G and 40G switches only; that is, any switch on the Trident, Trident+, or Trident II platform.

# Layer 1 and Layer 2 Features

## Spanning Tree and Rapid Spanning Tree

Spanning tree protocol (STP) is always recommended in layer 2 topologies, as it prevents bridge loops and broadcast radiation on a bridged network.

`mstpd` is a daemon that implements IEEE802.1D 2004 and IEEE802.1Q 2011. Currently, STP is disabled by default on the bridge in Cumulus Linux.

To enable STP, configure `brctl stp <bridge> on`.



The STP modes Cumulus Linux supports vary depending upon which **bridge driver mode** (see page 162) is in use. For a bridge configured in *traditional* mode, STP, RSTP, PVST and PVRST are supported; with the default set to PVRST. **VLAN-aware** (see page 182) bridges only operate in RSTP mode.

If a bridge running RSTP (802.1w) receives a common STP (802.1D) BPDU, it will automatically fall back to 802.1D operation.

You can configure `mstpd` to be in common STP mode only, by setting `setforcevers` to *STP*.

## Contents

(Click to expand)

- [Contents \(see page 124\)](#)
- [Commands \(see page 124\)](#)
- [PVST/PVRST \(see page 125\)](#)
- [Creating a Bridge and Configuring STP \(see page 125\)](#)
- [Configuring Spanning Tree Parameters \(see page 127\)](#)
  - [Understanding the Spanning Tree Parameters \(see page 127\)](#)
- [Bridge Assurance \(see page 135\)](#)
- [BPDU Guard \(see page 135\)](#)
  - [Configuring BPDU Guard \(see page 136\)](#)
  - [Recovering a Port Disabled by BPDU Guard \(see page 136\)](#)
- [BPDU Filter \(see page 138\)](#)
- [Configuration Files \(see page 139\)](#)
- [Man Pages \(see page 139\)](#)
- [Useful Links \(see page 139\)](#)
- [Caveats and Errata \(see page 139\)](#)

## Commands

- `brctl`

- `mstptcl`

`mstptcl` is a utility to configure STP. `mstp` is started by default on bootup. `mstp` logs and errors are located in `/var/log/syslog`.

## PVST/PVRST

Per VLAN Spanning Tree (PVST) creates a spanning tree instance for a bridge. Rapid PVST (PVRST) supports RSTP enhancements for each spanning tree instance. You must create a bridge corresponding to the untagged native/access VLAN, and all the physical switch ports must be part of the same VLAN. When connected to a switch that has a native VLAN configuration, the native VLAN **must** be configured to be VLAN 1 only.

Cumulus Linux supports the RSTP/PVRST/PVST modes of STP natively when the bridge is configured in traditional mode (see page 162).

## ***Creating a Bridge and Configuring STP***

To create a bridge, configure the bridge stanza under `/etc/network/interfaces`. More information on configuring bridges can be found here. (see page 162) To enable STP on the bridge, include the keyword `bridge-stp on`.

```
auto br2
iface br2
    bridge-ports swp1.101 swp4.101 swp5.101
    bridge-stp on
```

To enable the bridge, run `ifreload -a`:

```
cumulus@switch:~$ sudo ifreload -a
```

Runtime Configuration (Advanced)



A runtime configuration is non-persistent, which means the configuration you create here does not persist after you reboot the switch.

You use `brctl` to create the bridge, add bridge ports in the bridge and configure STP on the bridge. `mstptcl` is used only when an admin needs to change the default configuration parameters for STP:

```
cumulus@switch:~$ sudo brctl addbr br2
cumulus@switch:~$ sudo brctl addif br2 swp1.101 swp4.101 swp5.101
cumulus@switch:~$ sudo brctl stp br2 on
cumulus@switch:~$ sudo ifconfig br2 up
```

To get the bridge state, use:

```
cumulus@switch:~$ sudo brctl show
bridge name      bridge id          STP enabled    interfaces
br2              8000.001401010100  yes           swp1.101
                                         swp4.101
                                         swp5.101
```

To get the mstpd bridge state, use:

```
cumulus@switch:~$ sudo mstpcctl showbridge br2
br2 CIST info
enabled        yes
bridge id      F.000.00:14:01:01:01:00
designated root F.000.00:14:01:01:01:00
regional root  F.000.00:14:01:01:01:00
root port      none
path cost      0          internal path cost  0
max age        20         bridge max age     20
forward delay  15         bridge forward delay 15
tx hold count 6          max hops            20
hello time    2          ageing time       200
force protocol version   rstp
time since topology change 90843s
topology change count    4
topology change          no
topology change port     swp4.101
last topology change port swp5.101
```

To get the mstpd bridge port state, use:

```
cumulus@switch:~$ sudo mstpcctl showport br2
E swp1.101 8.001 forw F.000.00:14:01:01:01:00 F.000.00:14:01:01:01:00
8.001 Desg
      swp4.101 8.002 forw F.000.00:14:01:01:01:00 F.000.00:14:01:01:01:00
8.002 Desg
      E swp5.101 8.003 forw F.000.00:14:01:01:01:00 F.000.00:14:01:01:01:00
8.003 Desg

cumulus@switch:~$ sudo mstpcctl showportdetail br2 swp1.101
br2:swp1.101 CIST info
  enabled        yes          role          Designated
  port id       8.001        state         forwarding
  external port cost 2000    admin external cost 0
```

internal port cost	2000	admin internal cost	0
designated root	F.000.00:14:01:01:01:00	dsgn external cost	0
dsgn regional root	F.000.00:14:01:01:01:00	dsgn internal cost	0
designated bridge	F.000.00:14:01:01:01:00	designated port	8.001
admin edge port	no	auto edge port	yes
oper edge port	yes	topology change ack	no
point-to-point	yes	admin point-to-point	auto
restricted role	no	restricted TCN	no
port hello time	2	disputed	no
bpdu guard port	no	bpdu guard error	no
network port	no	BA inconsistent	no
Num TX BPDU	45772	Num TX TCN	4
Num RX BPDU	0	Num RX TCN	0
Num Transition FWD	2	Num Transition BLK	2

## Configuring Spanning Tree Parameters

The persistent configuration for a bridge is set in `/etc/network/interfaces`. The configuration below shows every possible option configured. There is no requirement to configure any of these options:

```
auto br2
iface br2 inet static
    bridge-ports swp1 swp2 swp3 swp4
    bridge-stp on
    mstptctl-maxage 20
    mstptctl-ageing 300
    mstptctl-fdelay 15
    mstptctl-maxhops 20
    mstptctl-txholdcount 6
    mstptctl-forcevers rstp
    mstptctl-treepriority 32768
    mstptctl-treeportpriority swp3=128
    mstptctl-hello 2
    mstptctl-portpathcost swp1=0 swp2=0
    mstptctl-portadminedge swp1=no swp2=no
    mstptctl-portautoedge swp1=yes swp2=yes
    mstptctl-portp2p swp1=no swp2=no
    mstptctl-portrestrole swp1=no swp2=no
    mstptctl-portresttcn swp1=no swp2=no
    mstptctl-portnetwork swp1=no
    mstptctl-bpduguard swp1=no swp2=no
    mstptctl-bpdufilter swp4=yes
```

## Understanding the Spanning Tree Parameters

The spanning tree parameters are defined in the IEEE [802.1D](#), [802.1Q](#) specifications and in the table below.

While configuring spanning tree in a persistent configuration, as described above, is the preferred method, you can also use `mstptcl` to configure spanning tree protocol parameters at runtime.



A runtime configuration is non-persistent, which means the configuration you create here does not persist after you reboot the switch.

The `mstp` daemon is an open source project that some network engineers may be unfamiliar with. For example, many incumbent vendors use the keyword `portfast` to describe a port that is automatically set to forwarding when the port is brought up. The `mstpd` equivalent is `mstptcl-portadminedge`. For more comparison [please read this knowledge base article](#).

Examples are included below:

Parameter	Description
<b>maxage</b>	<p>Sets the bridge's <i>maximum age</i> to &lt;max_age&gt; seconds. The default is 20. The maximum age must meet the condition <math>2 * (\text{Bridge Forward Delay} - 1 \text{ second}) \geq \text{Bridge Max Age}</math>.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre>mstptcl-maxage 24</pre> <p>To set this parameter at runtime, use:</p> <pre>mstptcl setmaxage &lt;bridge&gt; &lt;max_age&gt;</pre> <pre>cumulus@switch:~\$ sudo mstptcl setmaxage br2 24</pre>
<b>ageing</b>	<p>Sets the Ethernet (MAC) address <i>ageing time</i> in &lt;time&gt; seconds for the bridge when the running version is STP, but not RSTP/MSTP. The default is 300.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre>mstptcl-ageing 240</pre> <p>To set this parameter at runtime, use:</p> <pre>mstptcl setageing &lt;bridge&gt; &lt;time&gt;</pre> <pre>cumulus@switch:~\$ sudo mstptcl setageing br2 240</pre>

Parameter	Description
<b>fdelay</b>	<p>Sets the bridge's <i>bridge forward delay</i> to &lt;time&gt; seconds. The default is 15.</p> <p>The bridge forward delay must meet the condition <math>2 * (\text{Bridge Forward Delay} - 1 \text{ second}) \geq \text{Bridge Max Age}</math>.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre data-bbox="453 530 763 566">mstpctl -fdelay 15</pre> <p>To set this parameter at runtime, use:</p> <pre data-bbox="453 720 992 756">mstpctl setfdelay &lt;bridge&gt; &lt;time&gt;</pre> <pre data-bbox="453 861 1286 897">cumulus@switch:~\$ sudo mstpctl setfdelay br2 15</pre>
<b>maxhops</b>	<p>Sets the bridge's <i>maximum hops</i> to &lt;max_hops&gt;. The default is 20.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre data-bbox="453 1127 780 1163">mstpctl -maxhops 24</pre> <p>To set this parameter at runtime, use:</p> <pre data-bbox="453 1317 1073 1353">mstpctl setmaxhops &lt;bridge&gt; &lt;max_hops&gt;</pre> <pre data-bbox="453 1459 1310 1495">cumulus@switch:~\$ sudo mstpctl setmaxhops br2 24</pre>
<b>txholdcount</b>	<p>Sets the bridge's <i>bridge transmit hold count</i> to &lt;tx_hold_count&gt;. The default is 6.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre data-bbox="453 1719 829 1755">mstpctl -txholdcount 6</pre> <p>To set this parameter at runtime, use:</p> <pre data-bbox="421 1854 1188 1890">mstpctl settxholdcount &lt;bridge&gt; &lt;tx_hold_count&gt;</pre>

Parameter	Description
	<pre>cumulus@switch:~\$ sudo mstpcctl settxholdcount br2 5</pre>
<b>forcevers</b>	<p>Sets the bridge's <i>force STP</i> version to either RSTP/STP. MSTP is not supported currently. The default is <i>RSTP</i>.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre>mstpctl-forcevers rstp</pre> <p>To set this parameter at runtime, use:</p> <pre>mstpctl setforcevers &lt;bridge&gt; {mstp rstp stp}</pre> <pre>cumulus@switch:~\$ sudo mstpctl setforcevers br2 rstp</pre>
<b>treeprion</b>	<p>Sets the bridge's <i>tree priority</i> to &lt;priority&gt; for an MSTI instance. The priority value is a number between 0 and 65535 and must be a multiple of 4096. The bridge with the lowest priority is elected the <i>root bridge</i>. The default is 32768.</p> <div style="border: 2px solid red; padding: 5px; margin-top: 10px;"> <span style="color: red;">!</span> For <code>msti</code>, only 0 is supported currently.     </div>
	<p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre>mstpctl-treeprion 8192</pre> <p>To set this parameter at runtime, use:</p> <pre>mstpctl settreeprion &lt;bridge&gt; &lt;mstid&gt; &lt;priority&gt;</pre> <pre>cumulus@switch:~\$ sudo mstpctl settreeprion br2 0 8192</pre>
<b>treportprion</b>	<p>Sets the <i>priority</i> of port &lt;port&gt; to &lt;priority&gt; for the MSTI instance. The priority value is a number between 0 and 240 and must be a multiple of 16. The default is 128.</p>

Parameter	Description
	<p>For <code>msti</code>, only 0 is supported currently.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre data-bbox="453 481 1024 513"><code>mstpctl-treeportpri swp4.101 64</code></pre> <p>To set this parameter at runtime, use:</p> <pre data-bbox="453 661 1388 692"><code>mstpctl settreeportpri &lt;bridge&gt; &lt;port&gt; &lt;mstid&gt; &lt;priority&gt;</code></pre> <pre data-bbox="453 808 1344 872"><code>cumulus@switch:~\$ sudo mstpctl settreeportpri br2 swp4.101 0 64</code></pre>
<b>hello</b>	<p>Sets the bridge's <i>bridge hello time</i> to <code>&lt;time&gt;</code> seconds. The default is 2.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre data-bbox="453 1115 747 1146"><code>mstpctl-hello 20</code></pre> <p>To set this parameter at runtime, use:</p> <pre data-bbox="453 1294 975 1326"><code>mstpctl sethello &lt;bridge&gt; &lt;time&gt;</code></pre> <pre data-bbox="453 1442 1269 1474"><code>cumulus@switch:~\$ sudo mstpctl sethello br2 20</code></pre>
<b>portpathcost</b>	<p>Sets the <i>port cost</i> of the port <code>&lt;port&gt;</code> in bridge <code>&lt;bridge&gt;</code> to <code>&lt;cost&gt;</code>. The default is 0. <code>mstp</code> supports only long mode; that is, 32 bits for the path cost.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre data-bbox="453 1759 1024 1790"><code>mstpctl-portpathcost swp1.101=10</code></pre> <p>To set this parameter at runtime, use:</p>

Parameter	Description
	<pre>mstpctl setportpathcost &lt;bridge&gt; &lt;port&gt; &lt;cost&gt;</pre> <pre>cumulus@switch:~\$ sudo mstpctl setportpathcost br2 swp1.101 10</pre>
<b>portadminedge</b>	<p>Enables/disables the <i>initial edge state</i> of the port &lt;port&gt; in bridge &lt;bridge&gt;. The default is <i>no</i>.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre>mstpctl -portadminedge swp1.101=yes</pre> <p>To set this parameter at runtime, use:</p> <pre>mstpctl setportadminedge &lt;bridge&gt; &lt;port&gt; {yes no}</pre> <pre>cumulus@switch:~\$ sudo mstpctl setportadminedge br2 swp1.101 yes</pre>
<b>portautoedge</b>	<p>Enables/disables the <i>auto transition</i> to/from the edge state of the port &lt;port&gt; in bridge &lt;bridge&gt;. The default is <i>yes</i>.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre>mstpctl -portautoedge swp1.101=no</pre> <p>To set this parameter at runtime, use:</p> <pre>mstpctl setportautoedge &lt;bridge&gt; &lt;port&gt; {yes no}</pre> <pre>cumulus@switch:~\$ sudo mstpctl setportautoedge br2 swp1.101 no</pre>
<b>portp2p</b>	

Parameter	Description
	<p>Enables/disables the <i>point-to-point detection mode</i> of the port &lt;port&gt; in bridge &lt;bridge&gt;. The default is <i>auto</i>.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre data-bbox="453 487 943 519">mstpctl -portp2p swp1.101=no</pre> <p>To set this parameter at runtime, use:</p> <pre data-bbox="453 677 1230 709">mstpctl setportp2p &lt;bridge&gt; &lt;port&gt; {yes no auto}</pre> <pre data-bbox="453 819 1410 872">cumulus@switch:~\$ sudo mstpctl setportp2p br2 swp1.101 no</pre>
<b>portrestrrole</b>	<p>Enables/disables the ability of the port &lt;port&gt; in bridge &lt;bridge&gt; to take the <i>root role</i>. The default is <i>no</i>.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre data-bbox="453 1163 1046 1195">mstpctl -portrestrrole swp1.101=no</pre> <p>To set this parameter at runtime, use:</p> <pre data-bbox="453 1343 1246 1374">mstpctl setportrestrrole &lt;bridge&gt; &lt;port&gt; {yes no}</pre> <pre data-bbox="453 1484 1361 1548">cumulus@switch:~\$ sudo mstpctl setportrestrrole br2 swp1.101 yes</pre>
<b>portrestrtcn</b>	<p>Enables/disables the ability of the port &lt;port&gt; in bridge &lt;bridge&gt; to propagate <i>received topology change notifications</i>. The default is <i>no</i>.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre data-bbox="453 1818 1046 1850">mstpctl -portrestrtcn swp1.101=yes</pre> <p>To set this parameter at runtime, use:</p>

Parameter	Description
	<pre>mstpctl setportrestrtcn &lt;bridge&gt; &lt;port&gt; {yes no}</pre> <pre>cumulus@switch:~\$ sudo mstpctl setportrestrtcn br2 swp1.101 yes</pre>
<b>portnetwork</b>	<p>Enables/disables the <i>bridge assurance capability</i> for a network port <code>&lt;port&gt;</code> in bridge <code>&lt;bridge&gt;</code>. The default is <i>no</i>.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre>mstpctl -portnetwork swp4.101=yes</pre> <p>To set this parameter at runtime, use:</p> <pre>mstpctl setportnetwork &lt;bridge&gt; &lt;port&gt; {yes no}</pre> <pre>cumulus@switch:~\$ sudo mstpctl setportnetwork br2 swp4. 101 yes</pre>
<b>bpduguard</b>	<p>Enables/disables the <i>BPDU guard configuration</i> of the port <code>&lt;port&gt;</code> in bridge <code>&lt;bridge&gt;</code>. The default is <i>no</i>.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre>mstpctl -bpduguard swp1=no</pre> <p>To set this parameter at runtime, use:</p> <pre>mstpctl setbpduguard &lt;bridge&gt; &lt;port&gt; {yes no}</pre> <pre>cumulus@switch:~\$ sudo mstpctl setbpduguard br2 swp1. 101 yes</pre>
<b>portbpdufilter</b>	

Parameter	Description
	<p>Enables/disables the <i>BPDU filter</i> functionality for a port &lt;port&gt; in bridge &lt;bridge&gt;. The default is <i>no</i>.</p> <p>To set this parameter persistently, configure it under the bridge stanza:</p> <pre data-bbox="453 481 1008 517">mstpctl-bpdulfiler swp4.101=yes</pre> <p>To set this parameter at runtime, use:</p> <pre data-bbox="453 671 1258 707">mstpctl setportbpdufilter &lt;bridge&gt; &lt;port&gt; {yes no}</pre> <pre data-bbox="453 813 1372 876">cumulus@switch:~\$ sudo mstpctl setportbpdufilter br2 swp4.101 yes</pre>

## Bridge Assurance

On a point-to-point link where RSTP is running, if you want to detect unidirectional links and put the port in a discarding state (in error), you can enable bridge assurance on the port by enabling port type network. The port would be in a bridge assurance inconsistent state until a BPDU is received from the peer. You need to configure the port type network on both the ends of the link:

```
cumulus@switch:~$ sudo mstpctl setportnetwork br1007 swp1.1007 yes

cumulus@switch:~$ sudo mstpctl showportdetail br1007 swp1.1007 | grep
network
    network port          yes           BA inconsistent      yes

cumulus@switch:~$ sudo grep -in assurance /var/log/syslog | grep mstp
1365:Jun 25 18:03:17 mstpd: br1007:swp1.1007 Bridge assurance inconsistent
```

## BPDU Guard

To protect the spanning tree topology from unauthorized switches affecting the forwarding path, you can configure *BPDU guard* (Bridge Protocol Data Unit). One very common example is when someone hooks up a new switch to an access port off of a leaf switch. If this new switch is configured with a low priority, it could become the new root switch and affect the forwarding path for the entire Layer 2 topology.

## Configuring BPDU Guard

You configure BPDU guard under the bridge stanza in `/etc/network/interfaces`:

```
auto br2
iface br2 inet static
    bridge-ports swp1 swp2 swp3 swp4 swp5 swp6
    bridge-stp on
    mstpcctl-bpduguard swp1=yes swp2=yes swp3=yes swp4=yes
```

To load the new configuration, run `ifreload -a`:

```
cumulus@switch:~$ sudo ifreload -a
```

## Non-Persistent Configuration

You can also configure BPDU guard on an individual port using a runtime configuration.

Runtime Configuration (Advanced)



A runtime configuration is non-persistent, which means the configuration you create here does not persist after you reboot the switch.

```
cumulus@switch:~$ sudo mstpcctl setbpduguard br2 swp1 yes
cumulus@switch:~$ sudo mstpcctl setbpduguard br2 swp2 yes
cumulus@switch:~$ sudo mstpcctl setbpduguard br2 swp3 yes
cumulus@switch:~$ sudo mstpcctl setbpduguard br2 swp4 yes
```

## Recovering a Port Disabled by BPDU Guard

If a BPDU is received on the port, STP will bring down the port and log an error in `/var/log/syslog`. The following is a sample error:

```
mstpd: error, MSTP_IN_rx_bpdu: bridge:bond0 Recvd BPDU on BPDU Guard
Port - Port Down
```

To determine whether BPDU guard is configured, or if a BPDU has been received, run `mstpcctl showportdetail <bridge name>`:

```
cumulus@switch:~$ sudo mstpcctl showportdetail br2 swp1 | grep guard
bpdu guard port      yes                      bpdu guard error      yes
```

The only way to recover a port that has been placed in the disabled state is to manually un-shut or bring up the port with `sudo ifup [port]`, as shown in the example below:



Bringing up the disabled port does not fix the problem if the configuration on the connected end-station has not been rectified.

```
cumulus@leaf2$ mstptcl showportdetail bridge bond0
bridge:bond0 CIST info
  enabled          no           role
Disabled
  port id         8.001        state
discarding
  external port cost 305      admin external cost 0
  internal port cost 305     admin internal cost 0
  designated root    8.000.6C:64:1A:00:4F:9C dsgn external cost 0
  dsgn regional root 8.000.6C:64:1A:00:4F:9C dsgn internal cost 0
  designated bridge 8.000.6C:64:1A:00:4F:9C designated port 8.00
1
  admin edge port   no          auto edge port      yes
  oper edge port   no          topology change ack no
  point-to-point   yes         admin point-to-point auto
  restricted role  no          restricted TCN       no
  port hello time  10          disputed            no
  bpdu guard port  yes         bpdu guard error  yes
  network port     no          BA inconsistent    no
  Num TX BPDU      3           Num TX TCN        2
  Num RX BPDU      488         Num RX TCN        2
  Num Transition FWD 1          Num Transition BLK 2
  bpdulfILTER port no          clag ISL Oper UP no
  clag ISL          no          clag dual conn mac 0:0:
  clag role        unknown
  0:0:0:0
  clag remote portID F.FFF
  0:0:0:0
```

```
cumulus@leaf2$ sudo ifup bond0
```

```
cumulus@leaf2$ mstptcl showportdetail bridge bond0
bridge:bond0 CIST info
  enabled          yes          role          Root
  port id         8.001        state
forwarding
  external port cost 305      admin external cost 0
  internal port cost 305     admin internal cost 0
  designated root    8.000.6C:64:1A:00:4F:9C dsgn external cost 0
  dsgn regional root 8.000.6C:64:1A:00:4F:9C dsgn internal cost 0
```

designated bridge	8.000.6C:64:1A:00:4F:9C	designated port	8.00
1			
admin edge port	no	auto edge port	yes
oper edge port	no	topology change ack	no
point-to-point	yes	admin point-to-point	auto
restricted role	no	restricted TCN	no
port hello time	2	disputed	no
bpdu guard port	no	bpdu guard error	no
network port	no	BA inconsistent	no
Num TX BPDU	3	Num TX TCN	2
Num RX BPDU	43	Num RX TCN	1
Num Transition FWD	1	Num Transition BLK	0
bpdufilter port	no		
clag ISL	no	clag ISL Oper UP	no
clag role	unknown	clag dual conn mac	0:0:
0:0:0:0			
clag remote portID F.FFF		clag system mac	0:0:
0:0:0:0			

## BPDUs Filter

You can enable `bpdufilter` on a switch port, which filters BPDUs in both directions. This effectively disables STP on the port.

To enable it, add the following to `/etc/network/interfaces` under the `bridge port iface` section example:

```
auto br100
iface br100
  bridge-ports swp1.100 swp2.100
  mstptctl-portbpdufilter swp1=yes swp2=yes
```

To load the new configuration from `/etc/network/interfaces`, run `ifreload -a`:

```
cumulus@switch:~$ sudo ifreload -a
```

For more information, see `man(5) ifupdown-addons-interfaces`.

Runtime Configuration (Advanced)



A runtime configuration is non-persistent, which means the configuration you create here does not persist after you reboot the switch.

To enable BPDU filter at runtime, run `mstptctl`:

```
cumulus@switch:~$ sudo mstpctl setportbpdufilter br100 swp1.100=yes swp2.  
100=yes
```

## Configuration Files

- /etc/network/interfaces

## Man Pages

- brctl(8)
- bridge-utils-interfaces(5)
- ifupdown-addons-interfaces(5)
- mstpctl(8)
- mstpctl-utils-interfaces(5)

## Useful Links

The source code for `mstpd/mstpctl` was written by [Vitalii Demianets](#) and is hosted at the sourceforge URL below.

- <https://sourceforge.net/projects/mstpd/>
- [http://en.wikipedia.org/wiki/Spanning\\_Tree\\_Protocol](http://en.wikipedia.org/wiki/Spanning_Tree_Protocol)

## Caveats and Errata

- MSTP is not supported currently. However, interoperability with MSTP networks can be accomplished using PVRSTP or PVSTP.

## Link Layer Discovery Protocol

The `lldpd` daemon implements the IEEE802.1AB (Link Layer Discovery Protocol, or LLDP) standard. LLDP allows you to know which ports are neighbors of a given port. By default, `lldpd` runs as a daemon and is started at system boot. `lldpd` command line arguments are placed in `/etc/default/lldpd`. `lldpd` configuration options are placed in `/etc/lldpd.conf` or under `/etc/lldpd.d/`.

For more details on the command line arguments and config options, please see `man lldpd(8)`.

`lldpd` supports CDP (Cisco Discovery Protocol, v1 and v2). `lldpd` logs by default into `/var/log/daemon.log` with an `lldpd` prefix.

`lldpccli` is the CLI tool to query the `lldpd` daemon for neighbors, statistics and other running configuration information. See `man lldpccli(8)` for details.

## Contents

(Click to expand)

- [Contents \(see page 139\)](#)
- [Commands \(see page 140\)](#)
- [Man Pages \(see page 140\)](#)
- [Configuring LLDP \(see page 140\)](#)
- [Example lldpcli Commands \(see page 140\)](#)
- [Enabling the SNMP Subagent in LLDP \(see page 144\)](#)
- [Configuration Files \(see page 145\)](#)
- [Useful Links \(see page 145\)](#)
- [Caveats and Errata \(see page 145\)](#)

## Commands

- [lldpd \(daemon\)](#)
- [lldpcli \(interactive CLI\)](#)

## Man Pages

- [man lldpd](#)
- [man lldpcli](#)

## Configuring LLDP

You configure lldpd settings in /etc/lldpd.conf or /etc/lldpd.d/.

Here is an example persistent configuration:

```
cumulus@switch:~$ sudo cat /etc/lldpd.conf
configure lldp tx-interval 40
configure lldp tx-hold 3
configure system interface pattern-blacklist "eth0"
```

lldpd logs to /var/log/daemon.log with the *lldpd* prefix:

```
cumulus@switch:~$ sudo tail -f /var/log/daemon.log | grep lldp
Aug  7 17:26:17 switch lldpd[1712]: unable to get system name
Aug  7 17:26:17 switch lldpd[1712]: unable to get system name
Aug  7 17:26:17 switch lldpcli[1711]: lldpd should resume operations
Aug  7 17:26:32 switch lldpd[1805]: NET-SNMP version 5.4.3 AgentX subagent
connected
```

## Example lldpcli Commands

To see all neighbors on all ports/interfaces:

```
cumulus@switch:~$ sudo lldpcli show neighbors
-----
LLDP neighbors:
-----
Interface: eth0, via: CDPv1, RID: 72, Time: 0 day, 00:33:40
Chassis:
    ChassisID: local test-server-1
    SysName: test-server-1
    SysDescr: Linux running on
Linux 3.2.2+ #1 SMP Mon Jun 10 16:21:22 PDT 2013 ppc
    MgmtIP: 192.0.2.72
    Capability: Router, on
Port:
    PortID: ifname eth1
-----
Interface: swp1, via: CDPv1, RID: 87, Time: 0 day, 00:36:27
nChassis:
    ChassisID: local T1
    SysName: T1
    SysDescr: Linux running on
Cumulus Linux
    MgmtIP: 192.0.2.15
    Capability: Router, on
Port:
    PortID: ifname swp1
    PortDescr: swp1
-----
... and more (output truncated to fit this doc)
```

To see neighbors on specific ports:

```
cumulus@switch:~$ sudo lldpcli show neighbors ports swp1,swp2
-----
Interface: swp1, via: CDPv1, RID: 87, Time: 0 day, 00:36:27
Chassis:
    ChassisID: local T1
    SysName: T1
    SysDescr: Linux running on
Cumulus Linux
    MgmtIP: 192.0.2.15
    Capability: Router, on
Port:
```

```
PortID:        ifname swp1
PortDescr:     swp1
-----
Interface:    swp2, via: CDPv1, RID: 123, Time: 0 day, 00:36:27
Chassis:
  ChassisID:   local T2
  SysName:     T2
  SysDescr:    Linux running on
Cumulus Linux
  MgmtIP:      192.0.2.15
  Capability:  Router, on
Port:
  PortID:      ifname swp1
  PortDescr:   swp1
```

To see lldpd statistics for all ports:

```
cumulus@switch:~$ sudo lldpccli show statistics
```

```
-----  
LLDP statistics:
```

```
-----  
Interface:    eth0
  Transmitted: 9423
  Received:    17634
  Discarded:   0
  Unrecognized: 0
  Ageout:      10
  Inserted:    20
  Deleted:     10
```

```
-----  
Interface:    swp1
  Transmitted: 9423
  Received:    6264
  Discarded:   0
  Unrecognized: 0
  Ageout:      0
  Inserted:    2
  Deleted:     0
```

```
-----  
Interface:    swp2
  Transmitted: 9423
  Received:    6264
  Discarded:   0
```

```

Unrecognized: 0
Ageout: 0
Inserted: 2
Deleted: 0
-----
Interface: swp3
Transmitted: 9423
Received: 6265
Discarded: 0
Unrecognized: 0
Ageout: 0
Inserted: 2
Deleted: 0
-----
... and more (output truncated to fit this document)

```

To see lldpd statistics summary for all ports:

```

cumulus@switch:~$ sudo lldpcli show statistics summary
-----
LLDP Global statistics:
-----
Summary of stats:
Transmitted: 648186
Received: 437557
Discarded: 0
Unrecognized: 0
Ageout: 10
Inserted: 38
Deleted: 10

```

To see the lldpd running configuration:

```

cumulus@switch:~$ sudo lldpcli show running-configuration
-----
Global configuration:
-----
Configuration:
Transmit delay: 1
Transmit hold: 4
Receive mode: no
Pattern for management addresses: (none)

```

```
Interface pattern: (none)
Interface pattern for chassis ID: (none)
Override description with: (none)
Override platform with: (none)
Advertise version: yes
Disable LLDP-MED inventory: yes
LLDP-MED fast start mechanism: yes
LLDP-MED fast start interval: 1
```

---

## Runtime Configuration (Advanced)



A runtime configuration does not persist when you reboot the switch — all changes are lost.

To configure active interfaces:

```
lldpcli configure system interface pattern "swp*"
```

To configure inactive interfaces:

```
lldpcli configure system interface pattern-blacklist "eth0"
```



The active interface list always overrides the inactive interface list.

To reset any interface list to none:

```
lldpcli configure system interface pattern-blacklist ""
```

## **Enabling the SNMP Subagent in LLDP**

LLDP does not enable the SNMP subagent by default. You need to edit /etc/default/lldpd and enable the -x option.

```
cumulus@switch:~$ sudo nano /etc/default/lldpd
# Uncomment to start SNMP subagent and enable CD
P, SONMP and EDP protocol
#DAEMON_OPTS="-x -c -s -e"
```

```
# Enable CDP by default
DAEMON_ARGS="-c"
DAEMON_ARGS="-x"
```

## Configuration Files

- /etc/lldpd.conf
- /etc/lldpd.d
- /etc/default/lldpd

## Useful Links

- <http://vincentbernat.github.io/lldpd/>
- [http://en.wikipedia.org/wiki/Link\\_Layer\\_Discovery\\_Protocol](http://en.wikipedia.org/wiki/Link_Layer_Discovery_Protocol)

## Caveats and Errata

- Annex E (and hence Annex D) of IEEE802.1AB (lldp) is not supported.

## Prescriptive Topology Manager - PTM

In data center topologies, right cabling is a time-consuming endeavor and is error prone. Prescriptive Topology Manager (PTM) is a dynamic cabling verification tool to help detect and eliminate such errors. It takes a graphviz-DOT specified network cabling plan (something many operators already generate), stored in a `topology.dot` file, and couples it with runtime information derived from LLDP to verify that the cabling matches the specification. The check is performed on every link transition on each node in the network.

You can customize the `topology.dot` file to control `ptmd` at both the global/network level and the node /port level.

PTM runs as a daemon, named `ptmd`.

For more information, see `man ptmd(8)`.

## Contents

(Click to expand)

- Contents (see page 145)
- Supported Features (see page 146)
- Configuring PTM (see page 146)
- Basic Topology Example (see page 147)
- Advanced PTM Configuration (see page 148)
  - Scripts (see page 148)
  - Configuration Parameters (see page 148)
    - Host-only Parameters (see page 149)

- Global Parameters (see page 149)
- Per-port Parameters (see page 150)
- Templates (see page 150)
- Supported BFD and LLDP Parameters (see page 150)
- Bidirectional Forwarding Detection (BFD) (see page 152)
  - Configuring BFD (see page 152)
  - Echo Function (see page 152)
- Enabling Quagga to Check Link State (see page 153)
- Using ptmd Service Commands (see page 154)
- Using ptmctl Commands (see page 155)
  - ptmctl Examples (see page 155)
  - ptmctl Error Outputs (see page 157)
- Configuration Files (see page 157)
- Useful Links (see page 158)
- Caveats and Errata (see page 158)

## Supported Features

- Topology verification using LLDP. `ptmd` creates a client connection to the LLDP daemon, `lldpd`, and retrieves the neighbor relationship between the nodes/ports in the network and compares them against the prescribed topology specified in the `topology.dot` file.
- Only physical interfaces, like `swp1` or `eth0`, are currently supported. Cumulus Linux does not support specifying virtual interfaces like bonds or subinterfaces like `eth0.200` in the topology file.
- Forwarding path failure detection using **Bidirectional Forwarding Detection** (BFD); however, demand mode is not supported. For more information on how BFD operates in Cumulus Linux, [see below \(see page 152\)](#) and see `man ptmd(8)`.
- Integration with Quagga (PTM to Quagga notification).
- Client management: `ptmd` creates an abstract named socket `/var/run/ptmd.socket` on startup. Other applications can connect to this socket to receive notifications and send commands.
- Event notifications: see Scripts below.
- User configuration via a `topology.dot` file; [see below \(see page 146\)](#).

## Configuring PTM

`ptmd` verifies the physical network topology against a DOT-specified network graph file, `/etc/ptm.d/topology.dot`.



This file must be present or else `ptmd` will not start. You can specify an alternate file using the `-c` option.

PTM supports undirected graphs.

At startup, `ptmd` connects to `lldpd`, the LLDP daemon, over a Unix socket and retrieves the neighbor name and port information. It then compares the retrieved port information with the configuration information that it read from the topology file. If there is a match, then it is a PASS, else it is a FAIL.

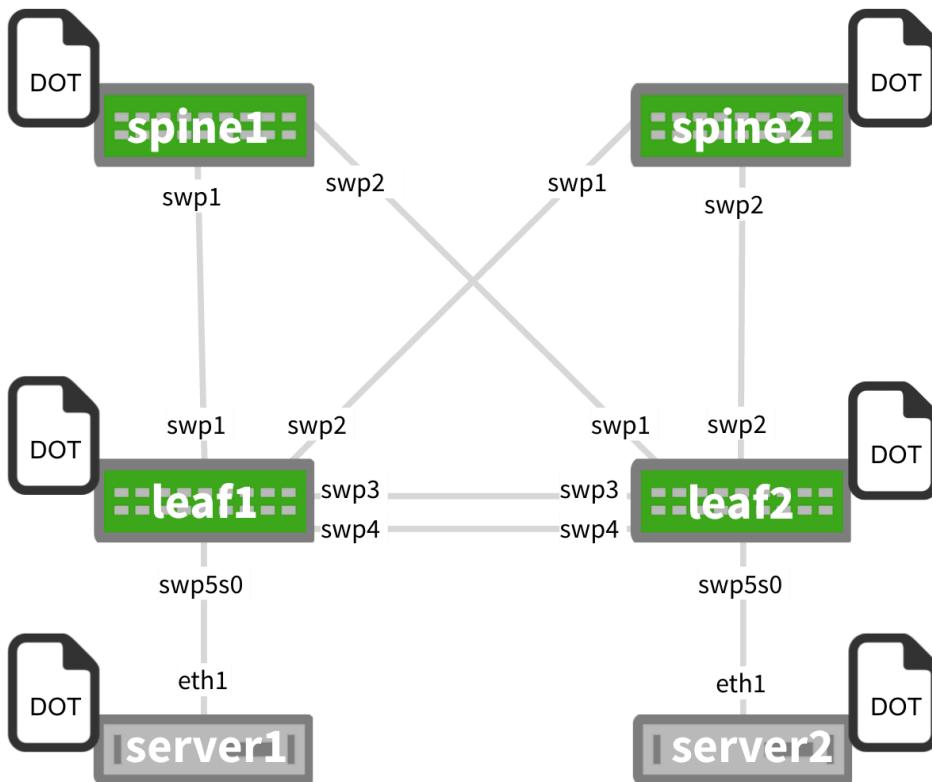


PTM performs its LLDP neighbor check using the PortID ifname TLV information. Previously, it used the PortID port description TLV information.

## Basic Topology Example

This is a basic example DOT file and its corresponding topology diagram. You should use the same `topology.dot` file on all switches, and don't split the file per device; this allows for easy automation by pushing/pulling the same exact file on each device!

```
graph G {
    "spine1":"swp1" -- "leaf1":"swp1";
    "spine1":"swp2" -- "leaf2":"swp1";
    "spine2":"swp1" -- "leaf1":"swp2";
    "spine2":"swp2" -- "leaf2":"swp2";
    "leaf1":"swp3" -- "leaf2":"swp3";
    "leaf1":"swp4" -- "leaf2":"swp4";
    "leaf1":"swp5s0" -- "server1":"eth1";
    "leaf2":"swp5s0" -- "server2":"eth1";
}
```



## Advanced PTM Configuration

PTM allows for more advanced configuration of the topology file using parameters you specify in the topology file.

### Scripts

`ptmd` executes scripts at `/etc/ptm.d/if-topo-pass` and `/etc/ptm.d/if-topo-fail` for each interface that goes through a change, running `if-topo-pass` when an LLDP or BFD check passes and running `if-topo-fails` when the check fails. The scripts receive an argument string that is the result of the `ptmctl` command, described in the [ptm commands section below](#) (see page 154).

You should modify these default scripts as needed.

### Configuration Parameters

You can configure `ptmd` parameters in the topology file. The parameters are classified as host-only, global, per-port/node and templates.

## Host-only Parameters

*Host-only parameters* apply to the entire host on which PTM is running. You can include the `hostnametype` host-only parameter, which specifies whether PTM should use only the host name (`hostname`) or the fully-qualified domain name (`fqdn`) while looking for the self-node in the graph file. For example, in the graph file below, PTM will ignore the FQDN and only look for `switch04`, since that is the host name of the switch it's running on:

- ✔ It's a good idea to always wrap the hostname in double quotes, like "www.example.com". Otherwise, `ptmd` can fail if you specify a fully-qualified domain name as the hostname and do not wrap it in double quotes.

```
graph G {
    hostnametype="hostname"
    BFD="upMinTx=150,requiredMinRx=250"
    "cumulus": "swp44" -- "switch04.cumulusnetworks.com": "swp20"
    "cumulus": "swp46" -- "switch04.cumulusnetworks.com": "swp22"
}
```

However, in this next example, PTM will compare using the FQDN and look for `switch05.cumulusnetworks.com`, which is the FQDN of the switch it's running on:

```
graph G {
    hostnametype="fqdn"
    "cumulus": "swp44" -- "switch05.cumulusnetworks.com": "swp20"
    "cumulus": "swp46" -- "switch05.cumulusnetworks.com": "swp22"
}
```

## Global Parameters

*Global parameters* apply to every port listed in the topology file. There are two global parameters: LLDP and BFD. LLDP is enabled by default; if no keyword is present, default values are used for all ports. However, BFD is disabled if no keyword is present, unless there is a per-port override configured. For example:

```
graph G {
    LLDP=""
    BFD="upMinTx=150,requiredMinRx=250,afi=both"
    "cumulus": "swp44" -- "qct-ly2-04": "swp20"
    "cumulus": "swp46" -- "qct-ly2-04": "swp22"
}
```

## Per-port Parameters

*Per-port parameters* provide finer-grained control at the port level. These parameters override any global or compiled defaults. For example:

```
graph G {
    LLDP= ""
    BFD= "upMinTx=300,requiredMinRx=100"
    "cumulus": "swp44" -- "qct-ly2-04": "swp20" [BFD="upMinTx=150,
    requiredMinRx=250,afi=both"]
    "cumulus": "swp46" -- "qct-ly2-04": "swp22"
}
```

## Templates

*Templates* provide flexibility in choosing different parameter combinations and applying them to a given port. A template instructs `ptmd` to reference a named parameter string instead of a default one. There are two parameter strings `ptmd` supports:

- `bfdtmp1`, which specifies a custom parameter tuple for BFD.
- `lldptmp1`, which specifies a custom parameter tuple for LLDP.

For example:

```
graph G {
    LLDP= ""
    BFD= "upMinTx=300,requiredMinRx=100"
    BFD1= "upMinTx=200,requiredMinRx=200"
    BFD2= "upMinTx=100,requiredMinRx=300"
    LLDP1="match_type=ifname"
    LLDP2="match_type=portdescr"
    "cumulus": "swp44" -- "qct-ly2-04": "swp20" [BFD="bfdtmp1=BFD1",
    LLDP="lldptmp1=LLDP1"]
    "cumulus": "swp46" -- "qct-ly2-04": "swp22" [BFD="bfdtmp1=BFD2",
    LLDP="lldptmp1=LLDP2"]
    "cumulus": "swp46" -- "qct-ly2-04": "swp22"
}
```

In this template, LLDP1 and LLDP2 are templates for LLDP parameters while BFD1 and BFD2 are templates for BFD parameters.

## Supported BFD and LLDP Parameters

`ptmd` supports the following BFD parameters:

- `upMinTx`: the minimum transmit interval, which defaults to `300ms`, specified in milliseconds.
- `requiredMinRx`: the minimum interval between received BFD packets, which defaults to `300ms`, specified in milliseconds.
- `detectMult`: the detect multiplier, which defaults to `3`, and can be any non-zero value.
- `afi`: the address family to be supported for the edge. The address family must be one of the following:
  - `v4`: BFD sessions will be built for only IPv4 connected peer. This is the default value.
  - `v6`: BFD sessions will be built for only IPv6 connected peer.
  - `both`: BFD sessions will be built for both IPv4 and IPv6 connected peers.

The following is an example of a topology with BFD applied at the port level:

```
graph G {
    "cumulus-1": "swp44" -- "cumulus-2": "swp20" [BFD="upMinTx=300,
requiredMinRx=100,afi=v6"]
    "cumulus-1": "swp46" -- "cumulus-2": "swp22" [BFD="detectMult=4"]
}
```

`ptmd` supports the following LLDP parameters:

- `match_type`, which defaults to the interface name (`ifname`), but can accept a port description (`portdescr`) instead if you want `lldpd` to compare the topology against the port description instead of the interface name. You can set this parameter globally or at the per-port level.
- `match_hostname`, which defaults to the host name (`hostname`), but enables PTM to match the topology using the fully-qualified domain name (`fqdn`) supplied by LLDP.

The following is an example of a topology with LLDP applied at the port level:

```
graph G {
    "cumulus-1": "swp44" -- "cumulus-2": "swp20" [LLDP=
match_hostname=fqdn]
    "cumulus-1": "swp46" -- "cumulus-2": "swp22" [LLDP=
match_type=portdescr]
}
```



When you specify `match_hostname=fqdn`, `ptmd` will match the entire FQDN, like `cumulus-2.domain.com` in the example below. If you do not specify anything for `match_hostname`, `ptmd` will match based on hostname only, like `cumulus-3` below, and ignore the rest of the URL:

```
graph G {
    "cumulus-1": "swp44" -- "cumulus-2.domain.com": "swp20"
[LLDP="match_hostname=fqdn"]
```

```
"cumulus-1" : "swp46" -- "cumulus-3" : "swp22" [ LLDP= "match_type=portdescr" ] }
```

## Bidirectional Forwarding Detection (BFD)

BFD provides low overhead and rapid detection of failures in the paths between two network devices. It provides a unified mechanism for link detection over all media and protocol layers. Use BFD to detect failures for IPv4 and IPv6 single or multihop paths between any two network devices, including unidirectional path failure detection. For more information, see the [BFD chapter \(see page 367\)](#).



BFD requires an IP address for any interface on which it is configured. The neighbor IP address for a single hop BFD session must be in the ARP table before BFD can start sending control packets.



You cannot specify BFD multihop sessions in the `topology.dot` file since you cannot specify the source and destination IP address pairs in that file. Use [Quagga \(see page 320\)](#) to configure multihop sessions.

## Configuring BFD

You configure BFD one of two ways: by specifying the configuration in the `topology.dot` file, or using [Quagga \(see page 367\)](#). However, the topology file has some limitations:

- The `topology.dot` file supports creating BFD IPv4 and IPv6 single hop sessions only; you cannot specify IPv4 or IPv6 multihop sessions in the topology file.
- The topology file supports BFD sessions for only link-local IPv6 peers; BFD sessions for global IPv6 peers discovered on the link will not be created.

## Echo Function

Cumulus Linux supports the *echo function* for IPv4 single hops only, and with the a synchronous operating mode only (Cumulus Linux does not support demand mode).

You use the echo function primarily to test the forwarding path on a remote system. To enable the echo function, set `echoSupport` to 1 in the topology file.

Once the echo packets are looped by the remote system, the BFD control packets can be sent at a much lower rate. You configure this lower rate by setting the `slowMinTx` parameter in the topology file to a non-zero value of milliseconds.

You can use more aggressive detection times for echo packets since the round-trip time is reduced because they are accessing the forwarding path. You configure the detection interval by setting the `echoMinRx` parameter in the topology file to a non-zero value of milliseconds; the minimum setting is 50 milliseconds. Once configured, BFD control packets are sent out at this required minimum echo Rx interval. This indicates to the peer that the local system can loop back the echo packets. Echo packets are transmitted if the peer supports receiving echo packets.

## About the Echo Packet

BFD echo packets are encapsulated into UDP packets over destination and source UDP port number 3785. The BFD echo packet format is vendor-specific and has not been defined in the RFC. BFD echo packets that originate from Cumulus Linux are 8 bytes long and have the following format:

0	1	2	3
Version	Length	Reserved	
My Discriminator			

Where:

- **Version** is the version of the BFD echo packet.
- **Length** is the length of the BFD echo packet.
- **My Discriminator** is a non-zero value that uniquely identifies a BFD session on the transmitting side. When the originating node receives the packet after being looped back by the receiving system, this value uniquely identifies the BFD session.

## Transmitting and Receiving Echo Packets

BFD echo packets are transmitted for a BFD session only when the peer has advertised a non-zero value for the required minimum echo Rx interval (the `echoMinRx` setting) in the BFD control packet when the BFD session starts. The transmit rate of the echo packets is based on the peer advertised echo receive value in the control packet.

BFD echo packets are looped back to the originating node for a BFD session only if locally the `echoMinRx` and `echoSupport` are configured to a non-zero values.

## Using Echo Function Parameters

You configure the echo function by setting the following parameters in the topology file at the global template and port level:

- **echoSupport:** Enables and disables echo mode. Set to 1 to enable the echo function. It defaults to 0 (disable).
- **echoMinRx:** The minimum interval between echo packets the local system is capable of receiving. This is advertised in the BFD control packet. When the echo function is enabled, it defaults to 50. If you disable the echo function, this parameter is automatically set to 0, which indicates the port or the node cannot process or receive echo packets.
- **slowMinTx:** The minimum interval between transmitting BFD control packets when the echo packets are being exchanged.

## Enabling Quagga to Check Link State

The Quagga routing suite enables additional checks to ensure that routing adjacencies are formed only on links that have connectivity conformant to the specification, as determined by `ptmd`.



You only need to do this to check link state; you don't need to enable PTM to determine BFD status.

To enable the check:

```
quagga# conf t
quagga(config)# ptm-enable
quagga(config)#+
```

To disable the checks:

```
quagga# conf t
quagga(config)# no ptm-enable
quagga(config)#+
```

When the `ptm-enable` flag is configured by the user, the `zebra` daemon connects to `ptmd` over a Unix socket. Any time there is a change of status for an interface, `ptmd` sends notifications to `zebra`. `zebra` maintains a `ptm-status` flag per interface and evaluates routing adjacency based on this flag. To check the per-interface `ptm-status`:

```
quagga# show interface swp1
Interface swp1 is up, line protocol is up
  PTM status: pass
  Description: T1
  index 3 metric 1 mtu 1500
  flags: <UP,BROADCAST,RUNNING,MULTICAST>
  HWaddr: 44:38:39:00:27:1d
  inet 192.0.2.1/31 broadcast 255.255.255.255
  inet6 2001:DB8::271d/64
quagga#
```

## Using `ptmd` Service Commands

PTM sends client notifications in CSV format.

`cumulus@switch:~$ sudo service ptmd start|restart|force-reload`: Starts or restarts the `ptmd` service. The `topology.dot` file must be present in order for the service to start.

cumulus@switch:~\$ sudo service ptmd reconfig: Instructs `ptmd` to read the `topology.dot` file again without restarting, applying the new configuration to the running state.

cumulus@switch:~\$ sudo service ptmd stop: Stops the `ptmd` service.

cumulus@switch:~\$ sudo service ptmd status: Retrieves the current running state of `ptmd`.

## ***Using ptmctl Commands***

`ptmctl` is a client of `ptmd`; it retrieves the daemon's operational state. It connects to `ptmd` over a Unix socket and listens for notifications. `ptmctl` parses the CSV notifications sent by `ptmd`.

See `man ptmctl` for more information.

## ***ptmctl Examples***

For basic output, use `ptmctl` without any options:

```
cumulus@switch:~$ sudo ptmctl

-----
port  cb1      BFD      BFD
      status    status   peer
                  local    type
-----
swp1  pass     pass     11.0.0.2
      N/A       N/A
      N/A       N/A
      N/A       N/A
      N/A       N/A
      N/A       N/A
```

For more detailed output, use the `-d` option:

```
cumulus@switch:~$ sudo ptmctl -d

-----
port  cb1      exp      act      sysname  portID  portDescr  match  last
BFD   BFD      BFD      BFD      det_mult tx_timeout rx_timeout
echo_tx_timeout echo_rx_timeout max_hop_cnt
      status  nbr      nbr
      Type   state  peer  DownDiag
-----
swp45 pass     h1:swp1 h1:swp1  h1      swp1      swp1      IfName 5m: 5s  N
/A     N/A      N/A     N/A      N/A      N/A      N/A      N
/A           N/A
      N/A
```

```

swp46 fail    h2:swp1 h2:swp1  h2          swp1      swp1      IfName 5m: 5s  N
/A     N/A      N/A      N/A          N/A      N/A      N/A
/A           N/A          N/A
  
```

To return information on active BFD sessions `ptmd` is tracking, use the `-b` option:

```

cumulus@switch:~$ sudo ptmctl -b

-----
port  peer          state  local          type      diag
-----

swp1  11.0.0.2    Up     N/A          singlehop  N/A
N/A   12.12.12.1  Up     12.12.12.4  multihop   N/A
  
```

To return LLDP information, use the `-l` option. It returns only the active neighbors currently being tracked by `ptmd`.

```

cumulus@switch:~$ sudo ptmctl -l

-----
port  sysname  portID  port      match  last
              descr   on       upd
-----

swp45 h1        swp1    swp1    IfName 5m:59s
swp46 h2        swp1    swp1    IfName 5m:59s
  
```

To return detailed information on active BFD sessions `ptmd` is tracking, use the `-b` and `-d` options (results are for an IPv6-connected peer):

```

cumulus@switch:~$ sudo ptmctl -b -d

-----
-----
-----
port  peer          state  local  type      diag  det  tx_timeout
rx_timeout echo        echo    max    rx_ctrl  tx_ctrl  rx_echo
tx_echo

mult          tx_timeout  rx_timeout
hop_cnt
  
```

```
-----
-----
-----
swp1  fe80::202:ff:fe00:1  Up      N/A     singlehop  N/A   3    300
900      0          0          N/A       187172  185986  0
0
swp1  3101:abc:bcad::2  Up      N/A     singlehop  N/A   3    300
900      0          0          N/A       501      533     0
0
```

## ***ptmctl Error Outputs***

If there are errors in the topology file or there isn't a session, PTM will return appropriate outputs. Typical error strings are:

```
Topology file error [/etc/ptm.d/topology.dot] [cannot find node cumulus] -
please check /var/log/ptmd.log for more info

Topology file error [/etc/ptm.d/topology.dot] [cannot open file (errno 2)] -
please check /var/log/ptmd.log for more info

No Hostname/MgmtIP found [Check LLDPD daemon status] -
please check /var/log/ptmd.log for more info

No BFD sessions . Check connections

No LLDP ports detected. Check connections

Unsupported command
```

For example:

```
cumulus@switch:~$ sudo ptmctl
-----
cmd      error
-----
get-status Topology file error [/etc/ptm.d/topology.dot] [cannot open file
(errno 2)] - please check /var/log/ptmd.log for more info
```

If you encounter errors with the `topology.dot` file, you can use `dot` (included in the Graphviz package) to validate the syntax of the topology file.

## Configuration Files

- /etc/ptm.d/topology.dot
- /etc/ptm.d/if-topo-pass
- /etc/ptm.d/if-topo-fail

## Useful Links

- Bidirectional Forwarding Detection (BFD)
- Graphviz
- LLDP on Wikipedia
- PTMd GitHub repo

## Caveats and Errata

- Prior to version 2.1, Cumulus Linux stored the `ptmd` configuration files in `/etc/cumulus/ptm.d`. When you upgrade to version 2.1 or later, all the existing `ptmd` files are copied from their original location to `/etc/ptm.d` with a `.dpkg-old` extension, except for `topology.dot`, which gets copied to `/etc/ptm.d`.

If you customized the `if-topo-pass` and `if-topo-fail` scripts, they are also copied to `.dpkg-old`, and you must modify them so they can parse the CSV output correctly.

Sample `if-topo-pass` and `if-topo-fail` scripts are available in `/etc/ptm.d`. A sample `topology.dot` file is available in `/usr/share/doc/ptmd/examples`.

## Bonding - Link Aggregation

Linux bonding provides a method for aggregating multiple network interfaces (the slaves) into a single logical bonded interface (the bond). Cumulus Linux bonding supports the IEEE 802.3ad link aggregation mode. Link aggregation allows one or more links to be aggregated together to form a *link aggregation group* (LAG), such that a media access control (MAC) client can treat the link aggregation group as if it were a single link. The benefits of link aggregation are:

- Linear scaling of bandwidth as links are added to LAG
- Load balancing
- Failover protection

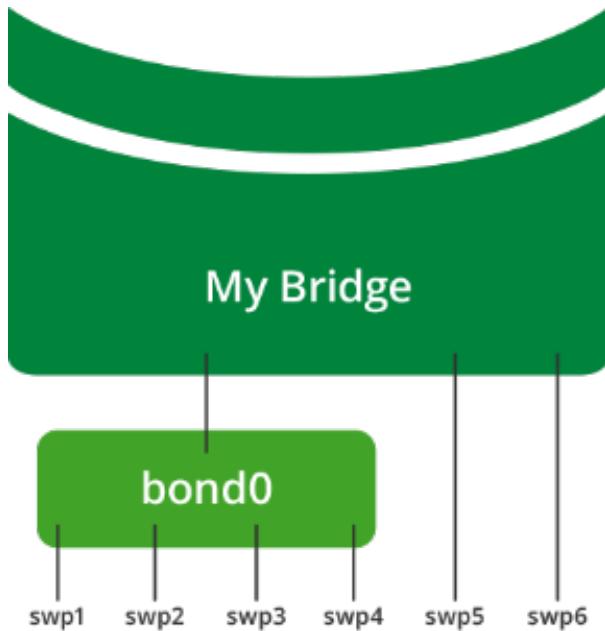
Cumulus Linux LAG control protocol is LACP version 1.

## Contents

(Click to expand)

- Contents (see page 158)
- Example: Bonding 4 Slaves (see page 159)
- Hash Distribution (see page 161)
- Configuration Files (see page 161)
- Useful Links (see page 161)
- Caveats and Errata (see page 161)

### Example: Bonding 4 Slaves



In this example, front panel port interfaces `swp1`-`swp4` are slaves in `bond0` (`swp5` and `swp6` are not part of `bond0`). The name of the bond is arbitrary as long as it follows Linux interface naming guidelines, and is unique within the switch. The only bonding mode supported in Cumulus Linux is `802.3ad`. There are several `802.3ad` settings that can be applied to each bond:

- `bond-slave`: The list of slaves in bond.
- `bond-mode`: **Must** be set to `802.3ad`.
- `bond-miimon`: How often the link state of each slave is inspected for link failures. It defaults to `0`, but `100` is the recommended value.



`bond-miimon` **must** be defined in `/etc/network/interfaces`.

- `bond-use-carrier`: How to determine link state.
- `bond-xmit-hash-policy`: Hash method used to select the slave for a given packet; **must** be set to `layer3+4`.
- `bond-lacp-rate`: Rate to ask link partner to transmit LACP control packets.
- `bond-min-links`: Specifies the minimum number of links that must be active before asserting carrier on the bond. Minimum value is `1`, but a value greater than `1` is useful if higher level services need to ensure a minimum of aggregate bandwidth before putting the bond in service.



`bond-min-links` **must** be defined in `/etc/network/interfaces` and it cannot be set to `0`. See also [this release note](#).

See Useful Links below for more details on settings.

To configure the bond, edit `/etc/network/interfaces` and add a stanza for bond0:

```
auto bond0
iface bond0
    address 10.0.0.1/30
    bond-slaves swp1 swp2 swp3 swp4
    bond-mode 802.3ad
    bond-mimon 100
    bond-use-carrier 1
    bond-lACP-rate 1
    bond-min-links 1
    bond-xmit-hash-policy layer3+4
```

However, if you are intending that the bond become part of a bridge, you don't need to specify an IP address. The configuration would look like this:

```
auto bond0
iface bond0
    bond-slaves glob swp1-4
    bond-mode 802.3ad
    bond-mimon 100
    bond-use-carrier 1
    bond-lACP-rate 1
    bond-min-links 1
    bond-xmit-hash-policy layer3+4
```

See `man interfaces` for more information on `/etc/network/interfaces`.

Here the link state sampling rate is 1/10 sec, and the LACP transmit rate is set to high. `bond-min-links` is set to 1 to indicate the bond must have at least one active member for bond to assert carrier. If the number of active members drops below the `bond-min-links` setting, the bond will appear to upper-level protocols as *link-down*. When the number of active links returns to greater than or equal to `bond-min-links`, the bond will become *link-up*.

When networking is started on switch, bond0 is created as MASTER and interfaces swp1-swp4 come up in SLAVE mode, as seen in the `ip link show` command:

```
3: swp1: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
   master bond0 state UP mode DEFAULT qlen 500
     link/ether 44:38:39:00:03:c1 brd ff:ff:ff:ff:ff:ff
4: swp2: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
   master bond0 state UP mode DEFAULT qlen 500
     link/ether 44:38:39:00:03:c1 brd ff:ff:ff:ff:ff:ff
5: swp3: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
```

```

master bond0 state UP mode DEFAULT qlen 500
    link/ether 44:38:39:00:03:c1 brd ff:ff:ff:ff:ff:ff
6: swp4: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
master bond0 state UP mode DEFAULT qlen 500
    link/ether 44:38:39:00:03:c1 brd ff:ff:ff:ff:ff:ff

```

And

```

55: bond0: <BROADCAST,MULTICAST,MASTER,UP,LOWER_UP> mtu 1500 qdisc noqueue
state UP mode DEFAULT
    link/ether 44:38:39:00:03:c1 brd ff:ff:ff:ff:ff:ff

```



All slave interfaces within a bond will have the same MAC address as the bond. Typically, the first slave added to the bond donates its MAC address for the bond. The other slaves' MAC addresses are set to the bond MAC address. The bond MAC address is used as source MAC address for all traffic leaving the bond, and provides a single destination MAC address to address traffic to the bond.

## **Hash Distribution**

Egress traffic through a bond is distributed to a slave based on a packet hash calculation. This distribution provides load balancing over the slaves. The hash calculation uses packet header data to pick which slave to transmit the packet. For IP traffic, IP header source and destination fields are used in the calculation. For IP + TCP/UDP traffic, source and destination ports are included in the hash calculation. Traffic for a given conversation flow will always hash to the same slave. Many flows will be distributed over all the slaves to load balance the total traffic. In a failover event, the hash calculation is adjusted to steer traffic over available slaves.

## **Configuration Files**

- /etc/network/interfaces

## **Useful Links**

- <http://www.linuxfoundation.org/collaborate/workgroups/networking/bonding>
- [802.3ad \(Accessible writeup\)](#)
- [Link aggregation from Wikipedia](#)

## **Caveats and Errata**

- An interface cannot belong to multiple bonds.
- Slave ports within a bond should all be set to the same speed/duplex, and should match the link partner's slave ports.

- A bond cannot enslave VLAN subinterfaces. A bond can have subinterfaces, but not the other way around.

## Ethernet Bridging - VLANs

Ethernet bridges provide a means for hosts to communicate at layer 2. Bridge members can be individual physical interfaces, bonds or logical interfaces that traverse an 802.1Q VLAN trunk.

Cumulus Linux 2.5.0 introduced a new method for configuring bridges that are [VLAN-aware \(see page 182\)](#). The bridge driver in Cumulus Linux 2.5.x is capable of VLAN filtering, which allows for configurations that are similar to incumbent network devices. While Cumulus Linux supports Ethernet bridges in traditional mode Cumulus Networks recommends using [VLAN-aware](#) mode unless you are using VXLANS in your network.

For a comparison of traditional and VLAN-aware modes, read [this knowledge base article](#).



You can configure both VLAN-aware and traditional mode bridges on the same network in Cumulus Linux; however you should not have more than one VLAN-aware bridge on a given switch. If you are implementing [VXLANS \(see page 287\)](#), you **must** use traditional bridge mode.

## Contents

(Click to expand)

- [Contents \(see page 162\)](#)
- [Configuration Files \(see page 162\)](#)
- [Commands \(see page 163\)](#)
- [Creating a Bridge between Physical Interfaces \(see page 163\)](#)
  - [Creating the Bridge and Adding Interfaces \(see page 163\)](#)
  - [Showing and Verifying the Bridge Configuration \(see page 165\)](#)
- [Examining MAC Addresses \(see page 165\)](#)
- [Multiple Bridges \(see page 166\)](#)
- [Configuring an SVI \(Switch VLAN Interface\) \(see page 169\)](#)
  - [Showing and Verifying the Bridge Configuration \(see page 170\)](#)
- [Using Trunks in Traditional Bridging Mode \(see page 171\)](#)
  - [Trunk Example \(see page 172\)](#)
  - [Showing and Verifying the Trunk \(see page 173\)](#)
  - [Additional Examples \(see page 173\)](#)
- [Configuration Files \(see page 173\)](#)
- [Useful Links \(see page 174\)](#)
- [Caveats and Errata \(see page 174\)](#)

## Configuration Files

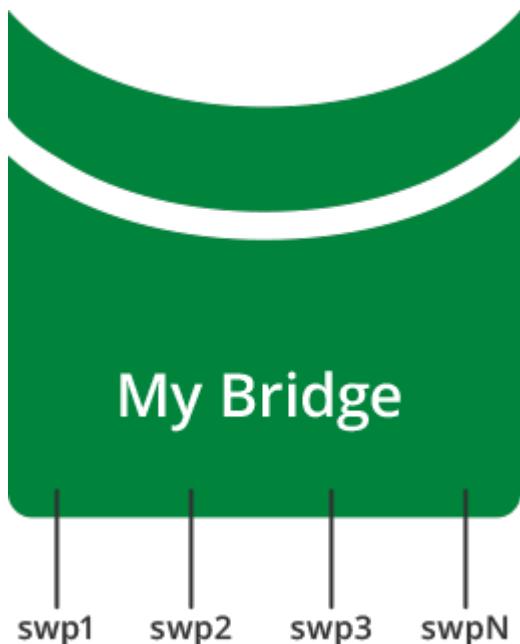
- [/etc/network/interfaces](#)

## Commands

- brctl
- bridge
- ip addr
- ip link

## ***Creating a Bridge between Physical Interfaces***

The basic use of bridging is to connect all of the physical and logical interfaces in the system into a single layer 2 domain.



## ***Creating the Bridge and Adding Interfaces***

You statically manage bridge configurations in `/etc/network/interfaces`. The following configuration snippet details an example bridge used throughout this chapter, explicitly enabling [spanning tree \(see page 124\)](#) and setting the bridge MAC address ageing timer. First, create a bridge with a descriptive name of 15 characters or fewer. Then add the logical interfaces (`bond0`) and physical interfaces (`swp5`, `swp6`) to assign to that bridge.

```
auto my_bridge
iface my_bridge
    bridge-ports bond0 swp5 swp6
    bridge-ageing 150
    bridge-stp on
```

Keyword	Explanation
bridge-ports	List of logical and physical ports belonging to the logical bridge.
bridge-ageing	Maximum amount of time before a MAC addresses learned on the bridge expires from the bridge MAC cache. The default value is 300 seconds.
bridge-stp	Enables spanning tree protocol on this bridge. The default spanning tree mode is Per VLAN Rapid Spanning Tree Protocol (PVRST).  For more information on spanning-tree configurations see the configuration section: <a href="#">Spanning Tree and Rapid Spanning Tree (see page 124)</a> .

To bring up the bridge `my_bridge`, use the `ifreload` command:

```
cumulus@switch:~$ sudo ifreload -a
```

#### Runtime Configuration (Advanced)



A runtime configuration is non-persistent, which means the configuration you create here does not persist after you reboot the switch.

To create the bridge and interfaces on the bridge, run:

```
cumulus@switch:~$ sudo brctl addbr my_bridge

cumulus@switch:~$ sudo brctl addif my_bridge bond0 swp5 swp6

cumulus@switch:~$ sudo brctl show
bridge name          bridge id      STP enabled     interfaces
my_bridge            8000.44383900129b  yes           bond0
                                         swp5
                                         swp6
```

```
cumulus@switch:~$ sudo ip link set up dev my_bridge
```

```
cumulus@switch:~$ sudo ip link set up dev bond0
```

```
cumulus@switch:~$ sudo for I in {5..6}; do ip link set up dev swp$I; done
```

## Showing and Verifying the Bridge Configuration

```
cumulus@switch:~$ ip link show my_bridge
56: my_bridge: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue
state UP mode DEFAULT
    link/ether 44:38:39:00:12:9b brd ff:ff:ff:ff:ff:ff
```



Do not try to bridge the management port, eth0, with any switch ports (like swp0, swp1, and so forth). For example, if you created a bridge with eth0 and swp1, it will **not** work.

### Using netshow to Display Bridge Information

`netshow` is a Cumulus Linux tool for retrieving information about your network configuration.

```
cumulus@switch$ netshow interface bridge
      Name      Speed      Mtu      Mode      Summary
      --  -----  -----  -----  -----
UP   my_bridge  N/A      1500  Bridge/L2  Untagged: bond0, swp5-6
                                Root Port: bond0
                                VlanID: Untagged
```

## Bridge Interface MAC Address and MTU

A bridge is a logical interface with a MAC address and an [MTU \(see page 111\)](#) (maximum transmission unit). The bridge MTU is the minimum MTU among all its members. The bridge's MAC address is inherited from the first interface that is added to the bridge as a member. The bridge MAC address remains unchanged until the member interface is removed from the bridge, at which point the bridge will inherit from the next member interface, if any. The bridge can also be assigned an IP address, as discussed later in this section.

## Examining MAC Addresses

A bridge forwards frames by looking up the destination MAC address. A bridge learns the source MAC address of a frame when the frame enters the bridge on an interface. After the MAC address is learned, the bridge maintains an age for the MAC entry in the bridge table. The age is refreshed when a frame is seen again with the same source MAC address. When a MAC is not seen for greater than the MAC ageing time, the MAC address is deleted from the bridge table.

The following shows the MAC address table of the example bridge. Notice that the `is_local?` column indicates if the MAC address is the interface's own MAC address (`is_local` is *yes*), or if it is learned on the interface from a packet's source MAC (where `is_local` is *no*):

```
cumulus@switch:~$ sudo brctl showmacs my_bridge
port name mac addr      is local? ageing timer
swp4      06:90:70:22:a6:2e    no        19.47
swp1      12:12:36:43:6f:9d    no        40.50
bond0     2a:95:22:94:d1:f0    no        1.98
swp1      44:38:39:00:12:9b    yes       0.00
swp2      44:38:39:00:12:9c    yes       0.00
swp3      44:38:39:00:12:9d    yes       0.00
swp4      44:38:39:00:12:9e    yes       0.00
bond0     44:38:39:00:12:9f    yes       0.00
swp2      90:e2:ba:2c:b1:94    no        12.84
swp2      a2:84:fe:fc:bf:cd    no        9.43
```

You can use the `bridge fdb` command to display the MAC address table as well:

```
cumulus@switch:~$ bridge fdb show
70:72:cf:9d:4e:36 dev swp2 VLAN 0 master bridge-A permanent
70:72:cf:9d:4e:35 dev swp1 VLAN 0 master bridge-A permanent
70:72:cf:9d:4e:38 dev swp4 VLAN 0 master bridge-B permanent
70:72:cf:9d:4e:37 dev swp3 VLAN 0 master bridge-B permanent
```

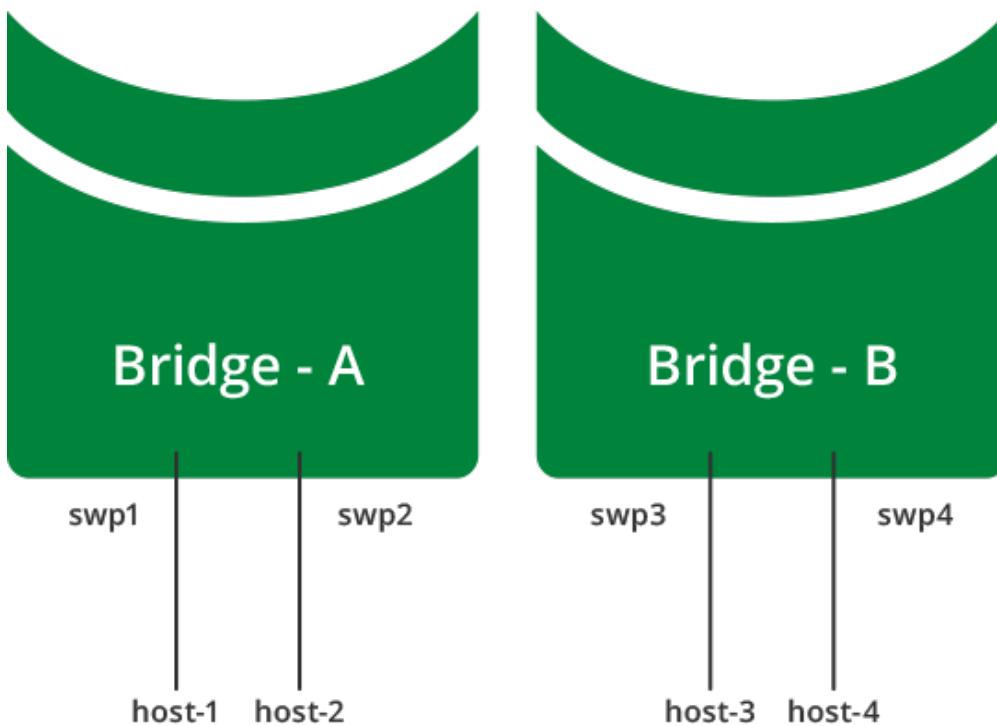


You can clear a MAC address from the table using the `bridge fdb` command:

```
cumulus@switch:~$ sudo bridge fdb del 90:e2:ba:2c:b1:94 dev swp2
```

## Multiple Bridges

Sometimes it is useful to logically divide a switch into multiple layer 2 domains, so that hosts in one domain can communicate with other hosts in the same domain but not in other domains. You can achieve this by configuring multiple bridges and putting different sets of interfaces in the different bridges. In the following example, host-1 and host-2 are connected to the same bridge (bridge-A), while host-3 and host-4 are connected to another bridge (bridge-B). host-1 and host-2 can communicate with each other, so can host-3 and host-4, but host-1 and host-2 cannot communicate with host-3 and host-4.



To configure multiple bridges, edit `/etc/network/interfaces`:

```

auto bridge-A
iface bridge-A
    bridge-ports swp1 swp2
    bridge-stp on

auto my_bridge
iface my_bridge
    bridge-ports swp3 swp4
    bridge-stp on

```

To bring up the bridges bridge-A and bridge-B, use the `ifreload` command:

```
cumulus@switch:~$ sudo ifreload -a
```

#### Runtime Configuration (Advanced)



A runtime configuration is non-persistent, which means the configuration you create here does not persist after you reboot the switch.

```

cumulus@switch:~$ sudo brctl addbr bridge-A

cumulus@switch:~$ sudo brctl addif bridge-A swp1 swp2

cumulus@switch:~$ sudo brctl addbr bridge-B

cumulus@switch:~$ sudo brctl addif bridge-B swp3 swp4

cumulus@switch:~$ sudo for I in {1..4}; do ip link set up dev swp$I; done

cumulus@switch:~$ sudo ip link set up dev bridge-A

cumulus@switch:~$ sudo ip link set up dev bridge-B

cumulus@switch:~$ sudo brctl show
bridge name      bridge id          STP enabled    interfaces
bridge-A        8000.44383900129b    yes           swp1
                                         swp2
bridge-B        8000.44383900129d    yes           swp3
                                         swp4

cumulus@switch:~$ ip link show bridge-A
97: bridge-A: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue
state UP mode DEFAULT
    link/ether 70:72:cf:9d:4e:35 brd ff:ff:ff:ff:ff:ff
cumulus@switch:~$ ip link show bridge-B
98: bridge-B: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue
state UP mode DEFAULT
    link/ether 70:72:cf:9d:4e:37 brd ff:ff:ff:ff:ff:ff

```

### Using netshow to Display the Bridges

netshow is a Cumulus Linux tool for retrieving information about your network configuration.

```

cumulus@switch$ netshow interface bridge
      Name      Speed      Mtu      Mode      Summary
      --  -----  -----  -----  -----
UP  bridge-A   N/A      1500  Bridge/L2  Untagged: swp1-2
                                         Root Port: swp2
                                         VlanID: Untagged
UP  bridge-B   N/A      1500  Bridge/L2  Untagged: swp3-4
                                         Root Port: swp3
                                         VlanID: Untagged

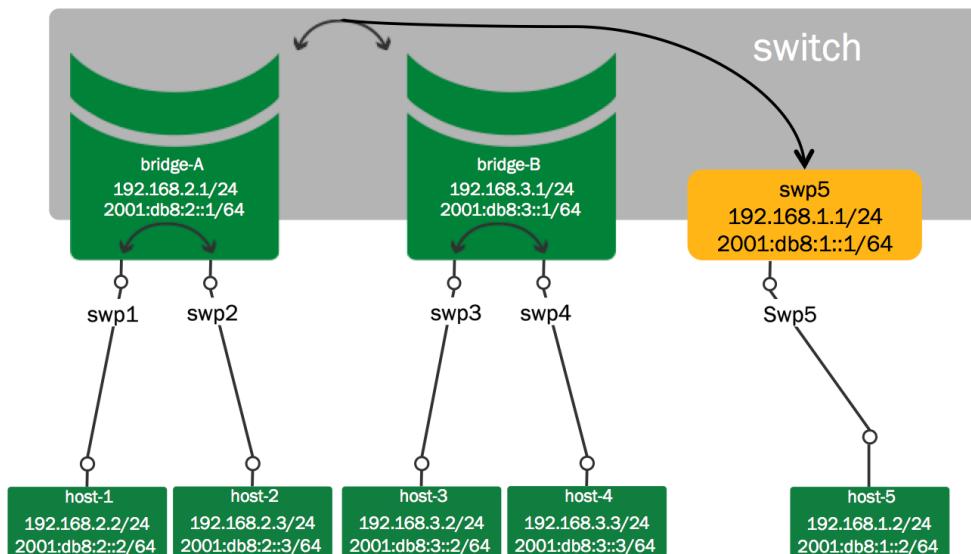
```

## Configuring an SVI (Switch VLAN Interface)

A bridge creates a layer 2 forwarding domain for hosts to communicate. A bridge can be assigned an IP address — typically of the same subnet as the hosts that are members of the bridge — and participate in routing topologies. This enables hosts within a bridge to communicate with other hosts outside the bridge through layer 3 routing.



When an interface is added to a bridge, it ceases to function as a router interface, and the IP address on the interface, if any, becomes unreachable.



The configuration for the two bridges example looks like the following:

```

auto swp5
iface swp5
    address 192.168.1.2/24
    address 2001:DB8:1::2/64
auto bridge-A
iface bridge-A
    address 192.168.2.1/24
    address 2001:DB8:2::1/64
    bridge-ports swp1 swp2
    bridge-stp on
auto bridge-B
iface bridge-B
    address 192.168.3.1/24
    address 2001:DB8:3::1/64
    bridge-ports swp3 swp4
    bridge-stp on

```

To bring up swp5 and bridges bridge-A and bridge-B, use the `ifreload` command:

```
cumulus@switch:~$ sudo ifreload -a
```

## ***Showing and Verifying the Bridge Configuration***

```
cumulus@switch$ ip addr show bridge-A
106: bridge-A: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue
state UP
link/ether 70:72:cf:9d:4e:35 brd ff:ff:ff:ff:ff:ff
inet 192.168.2.1/24 scope global bridge-A
    inet6 2001:db8:2::1/64 scope global
        valid_lft forever preferred_lft forever
    inet6 fe80::7272:ffff:fe9d:4e35/64 scope link
        valid_lft forever preferred_lft forever
```

```
cumulus@switch$ ip addr show bridge-B
107: bridge-B: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue
state UP
link/ether 70:72:cf:9d:4e:37 brd ff:ff:ff:ff:ff:ff
inet 192.168.3.1/24 scope global bridge-B
    inet6 2001:db8:3::1/64 scope global
        valid_lft forever preferred_lft forever
    inet6 fe80::7272:ffff:fe9d:4e37/64 scope link
        valid_lft forever preferred_lft forever
```

To see all the routes on the switch use the `ip route show` command:

```
cumulus@switch$ ip route show
192.168.1.0/24 dev swp5 proto kernel scope link src 192.168.1.2 dead
192.168.2.0/24 dev bridge-A proto kernel scope link src 192.168.2.1
192.168.3.0/24 dev bridge-B proto kernel scope link src 192.168.3.1
```

### Runtime Configuration (Advanced)



A runtime configuration is non-persistent, which means the configuration you create here does not persist after you reboot the switch.

To add an IP address to a bridge:

```
cumulus@switch:~$ sudo ip addr add 192.0.2.101/24 dev bridge-A
cumulus@switch:~$ sudo ip addr add 192.0.2.102/24 dev bridge-B
```

Using netshow to Display the SVI

`netshow` is a Cumulus Linux tool for retrieving information about your network configuration.

```
cumulus@switch$ netshow interface bridge
      Name      Speed      Mtu      Mode      Summary
      --  -----  -----  -----
-----  

UP   bridge-A  N/A       1500    Bridge/L3  IP: 192.168.2.1/24, 2001:db8:2::1
/64                                         Untagged: swp1-2
                                         Root Port: swp2
                                         VlanID: Untagged
UP   bridge-B  N/A       1500    Bridge/L3  IP: 192.168.3.1/24, 2001:db8:3::1
/64                                         Untagged: swp3-4
                                         Root Port: swp3
                                         VlanID: Untagged
```

## Using Trunks in Traditional Bridging Mode

The [IEEE standard](#) for trunking is 802.1Q. The 802.1Q specification adds a 4 byte header within the Ethernet frame that identifies the VLAN of which the frame is a member.

802.1Q also identifies an *untagged* frame as belonging to the *native* VLAN (most network devices default their native VLAN to 1). The concept of native, non-native, tagged or untagged has generated confusion due to mixed terminology and vendor-specific implementations. Some clarification is in order:

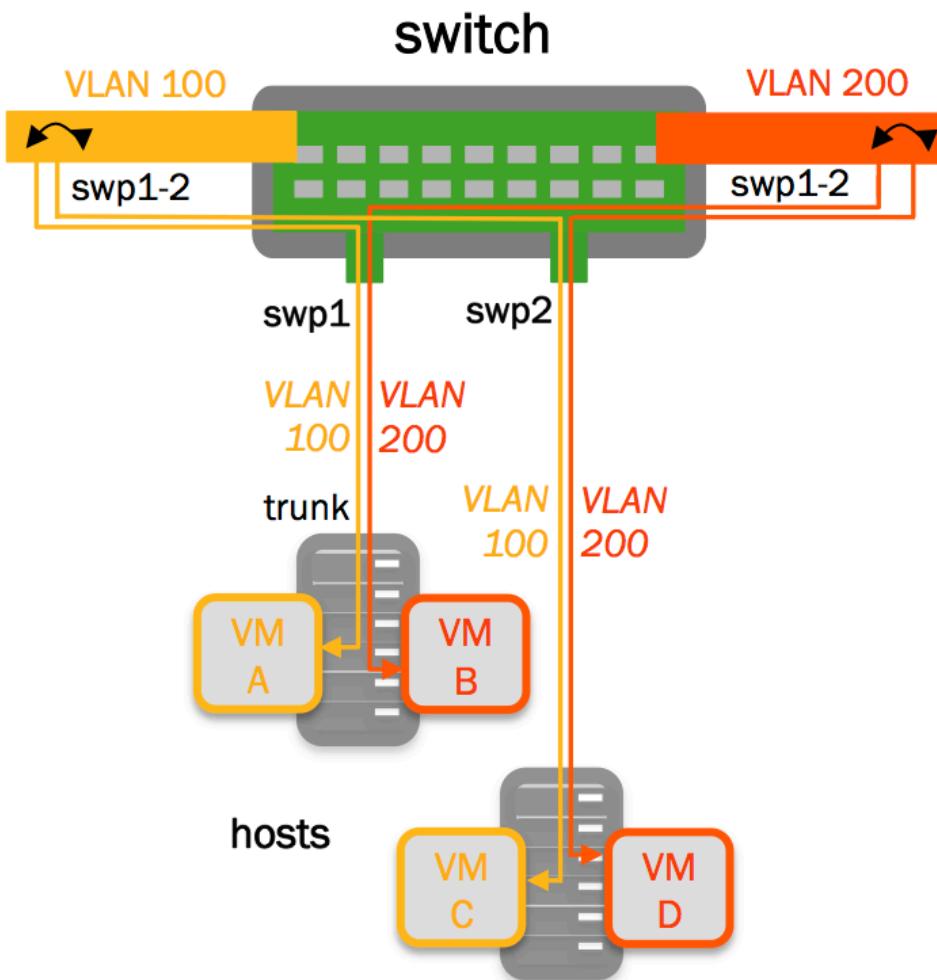
- A *trunk port* is a switch port configured to send and receive 802.1Q tagged frames.
- A switch sending an untagged (bare Ethernet) frame on a trunk port is sending from the native VLAN defined on the trunk port.
- A switch sending a tagged frame on a trunk port is sending to the VLAN identified by the 802.1Q tag.
- A switch receiving an untagged (bare Ethernet) frame on a trunk port places that frame in the native VLAN defined on the trunk port.
- A switch receiving a tagged frame on a trunk port places that frame in the VLAN identified by the 802.1Q tag.

A bridge in traditional mode has no concept of trunks, just tagged or untagged frames. With a trunk of 200 VLANs, there would need to be 199 bridges, each containing a tagged physical interface, and one bridge containing the native untagged VLAN. See the examples below for more information.



The interaction of tagged and un-tagged frames on the same trunk often leads to undesired and unexpected behavior. A switch that uses VLAN 1 for the native VLAN may send frames to a switch that uses VLAN 2 for the native VLAN, thus merging those two VLANs and their spanning tree state.

## Trunk Example



Configure the following in `/etc/network/interfaces`:

```
auto br-VLAN100
iface br-VLAN100
    bridge-ports swp1.100 swp2.100
    bridge-stp on
```

```
auto br-VLAN200
iface br-VLAN200
    bridge-ports swp1.200 swp2.200
    bridge-stp on
```

To bring up br-VLAN100 and br-VLAN200, use the `ifreload` command:

```
cumulus@switch:~$ sudo ifreload -a
```

## ***Showing and Verifying the Trunk***

```
cumulus@switch:~$ brctl show
bridge name bridge id          STP enabled interfaces
br-VLAN100  8000.7072cf9d4e35 no        swp1.100
                                         swp2.100
br-VLAN200  8000.7072cf9d4e35 no        swp1.200
                                         swp2.200
```

Using `netshow` to Display the Trunk

`netshow` is a Cumulus Linux tool for retrieving information about your network configuration.

```
cumulus@switch$ netshow interface bridge
      Name      Speed      Mtu      Mode      Summary
      --  -----  -----  -----  -----
UP   br-VLAN100  N/A       1500     Bridge/L2  Tagged: swp1-2
                                         STP: rootSwitch(32768)
                                         VlanID: 100
UP   br-VLAN200  N/A       1500     Bridge/L2  Tagged: swp1-2
                                         STP: rootSwitch(32768)
                                         VlanID: 200
```

## ***Additional Examples***

You can find additional examples of VLAN tagging in [this chapter](#) (see page 174).

## ***Configuration Files***

- `/etc/network/interfaces`
- `/etc/network/interfaces.d/`
- `/etc/network/if-down.d/`

- /etc/network/if-post-down.d/
- /etc/network/if-pre-up.d/
- /etc/network/if-up.d/

## Useful Links

- <http://www.linuxfoundation.org/collaborate/workgroups/networking/bridge>
- <http://www.linuxfoundation.org/collaborate/workgroups/networking/vlan>
- <http://www.linuxjournal.com/article/8172>

## Caveats and Errata

- The same bridge cannot contain multiple subinterfaces of the **same** port as members. Attempting to apply such a configuration will result in an error.

## VLAN Tagging

This article shows two examples of **VLAN tagging** (see page 174), one basic and one more advanced. They both demonstrate the streamlined interface configuration from `ifupdown2`. For more information, see **Configuring and Managing Network Interfaces** (see page 94).

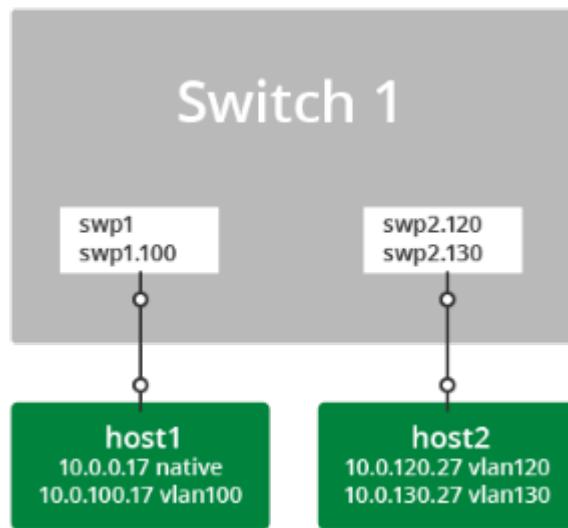
## Contents

(Click to expand)

- [Contents \(see page 174\)](#)
- [VLAN Tagging, a Basic Example \(see page 174\)
  - \[Persistent Configuration \\(see page 175\\)\]\(#\)](#)
- [VLAN Tagging, an Advanced Example \(see page 175\)
  - \[Persistent Configuration \\(see page 176\\)\]\(#\)
  - \[VLAN Translation \\(see page 181\\)\]\(#\)](#)

## VLAN Tagging, a Basic Example

A simple configuration demonstrating VLAN tagging involves two hosts connected to a switch.



- *host1* connects to *swp1* with both untagged frames and with 802.1Q frames tagged for *vlan100*.
- *host2* connects to *swp2* with 802.1Q frames tagged for *vlan120* and *vlan130*.

## Persistent Configuration

To configure the above example persistently, configure `/etc/network/interfaces` like this:

```

# Config for host1

auto swp1
iface swp1

auto swp1.100
iface swp1.100

# Config for host2
# swp2 must exist to create the .1Q subinterfaces, but it is not assigned
# an address

auto swp2
iface swp2

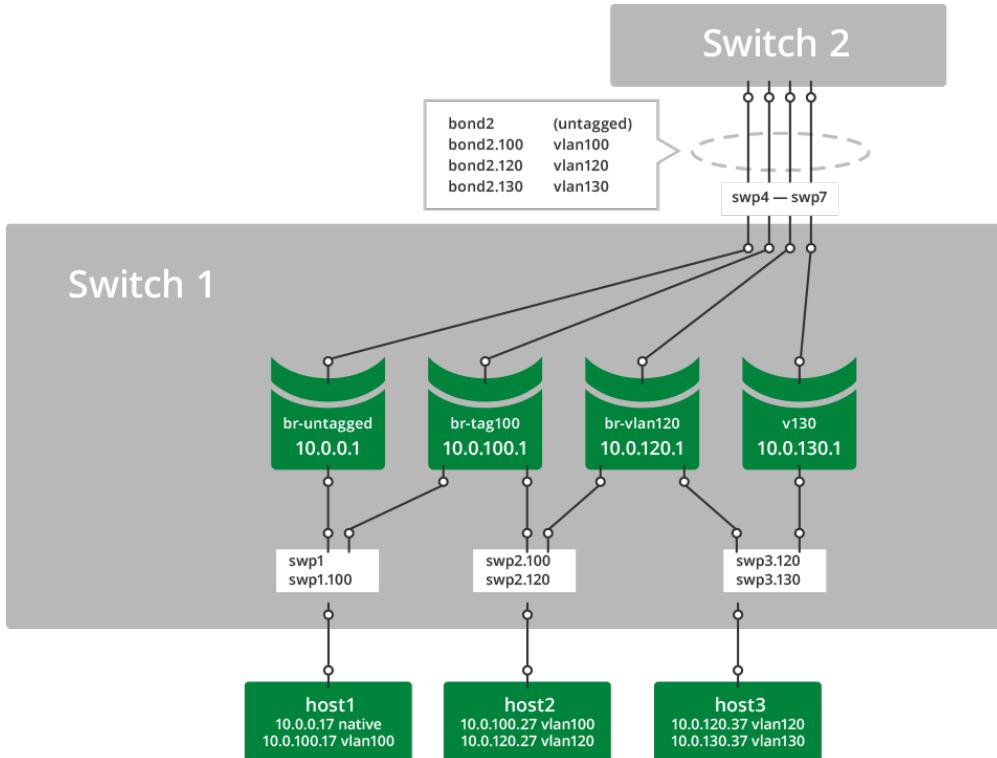
auto swp2.120
iface swp2.120

auto swp2.130
iface swp2.130

```

## VLAN Tagging, an Advanced Example

This example of VLAN tagging is more complex, involving three hosts and two switches, with a number of bridges and a bond connecting them all.



- *host1* connects to bridge *br-untagged* with bare Ethernet frames and to bridge *br-tag100* with 802.1q frames tagged for *vlan100*.
- *host2* connects to bridge *br-tag100* with 802.1q frames tagged for *vlan100* and to bridge *br-vlan120* with 802.1q frames tagged for *vlan120*.
- *host3* connects to bridge *br-vlan120* with 802.1q frames tagged for *vlan120* and to bridge *v130* with 802.1q frames tagged for *vlan130*.
- *bond2* carries tagged and untagged frames in this example.

Although not explicitly designated, the bridge member ports function as 802.1Q access ports and *trunk ports*. In the example above, comparing Cumulus Linux with a traditional Cisco device:

- *swp1* is equivalent to a trunk port with untagged and *vlan100*.
- *swp2* is equivalent to a trunk port with *vlan100* and *vlan120*.
- *swp3* is equivalent to a trunk port with *vlan120* and *vlan130*.
- *bond2* is equivalent to an EtherChannel in trunk mode with untagged, *vlan100*, *vlan120*, and *vlan130*.
- Bridges *br-untagged*, *br-tag100*, *br-vlan120*, and *v130* are equivalent to SVIs (switched virtual interfaces).

## Persistent Configuration

From /etc/network/interfaces :

```
# Config for host1 -----
```

```

- - - - -
# swp1 does not need an iface section unless it has a specific setting,
# it will be picked up as a dependent of swp1.100.
# And swp1 must exist in the system to create the .1q subinterfaces..
# but it is not applied to any bridge..or assigned an address.

auto swp1.100
iface swp1.100

# Config for host2
# swp2 does not need an iface section unless it has a specific setting,
# it will be picked up as a dependent of swp2.100 and swp2.120.
# And swp2 must exist in the system to create the .1q subinterfaces..
# but it is not applied to any bridge..or assigned an address.

auto swp2.100
iface swp2.100

auto swp2.120
iface swp2.120

# Config for host3
# swp3 does not need an iface section unless it has a specific setting,
# it will be picked up as a dependent of swp3.120 and swp3.130.
# And swp3 must exist in the system to create the .1q subinterfaces..
# but it is not applied to any bridge..or assigned an address.

auto swp3.120
iface swp3.120

auto swp3.130
iface swp3.130

# Configure the bond - - - - -
- - - - -
auto bond2
iface bond2
    bond-slaves glob swp4-7
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lacp-rate 1
    bond-min-links 1

```

```

bond-xmit-hash-policy layer3+4

# configure the bridges  -----
-----

auto br-untagged
iface br-untagged
    address 10.0.0.1/24
    bridge-ports swp1 bond2
    bridge-stp on

auto br-tag100
iface br-tag100
    address 10.0.100.1/24
    bridge-ports swp1.100 swp2.100 bond2.100
    bridge-stp on

auto br-vlan120
iface br-vlan120
    address 10.0.120.1/24
    bridge-ports swp2.120 swp3.120 bond2.120
    bridge-stp on

auto v130
iface v130
    address 10.0.130.1/24
    bridge-ports swp2.130 swp3.130 bond2.130
    bridge-stp on

# -----

```

To verify:

```

cumulus@switch:~$ sudo mstptctl showbridge br-tag100
br-tag100 CIST info
  enabled          yes
  bridge id        8.000.44:38:39:00:32:8B
  designated root  8.000.44:38:39:00:32:8B
  regional root   8.000.44:38:39:00:32:8B
  root port        none
  path cost        0          internal path cost  0
  max age          20         bridge max age      20
  forward delay   15         bridge forward delay 15

```

```

tx hold count 6          max hops           20
hello time   2          ageing time        300
force protocol version   rstp
time since topology change 333040s
topology change count    1
topology change          no
topology change port     swp2.100
last topology change port None

cumulus@switch:~$ sudo mstpcctl showportdetail br-tag100 | grep -B 2 state
br-tag100:bond2.100 CIST info
  enabled      yes             role      Designated
  port id     8.003           state    forwarding
--
br-tag100:swp1.100 CIST info
  enabled      yes             role      Designated
  port id     8.001           state    forwarding
--
br-tag100:swp2.100 CIST info
  enabled      yes             role      Designated
  port id     8.002           state    forwarding

cumulus@switch:~$ cat /proc/net/vlan/config
VLAN Dev name    | VLAN ID
Name-Type: VLAN_NAME_TYPE_RAW_PLUS_VID_NO_PAD
bond2.100         | 100   | bond2
bond2.120         | 120   | bond2
bond2.130         | 130   | bond2
swp1.100          | 100   | swp1
swp2.100          | 100   | swp2
swp2.120          | 120   | swp2
swp3.120          | 120   | swp3
swp3.130          | 130   | swp3

cumulus@switch:~$ cat /proc/net/bonding/bond2
Ethernet Channel Bonding Driver: v3.7.1 (April 27, 2011)

Bonding Mode: IEEE 802.3ad Dynamic link aggregation
Transmit Hash Policy: layer3+4 (1)
MII Status: up
MII Polling Interval (ms): 100
Up Delay (ms): 0
Down Delay (ms): 0

```

```
802.3ad info
LACP rate: fast
Min links: 0
Aggregator selection policy (ad_select): stable
Active Aggregator Info:
    Aggregator ID: 3
    Number of ports: 4
    Actor Key: 33
    Partner Key: 33
    Partner Mac Address: 44:38:39:00:32:cf

Slave Interface: swp4
MII Status: up
Speed: 10000 Mbps
Duplex: full
Link Failure Count: 0
Permanent HW addr: 44:38:39:00:32:8e
Aggregator ID: 3
Slave queue ID: 0

Slave Interface: swp5
MII Status: up
Speed: 10000 Mbps
Duplex: full
Link Failure Count: 0
Permanent HW addr: 44:38:39:00:32:8f
Aggregator ID: 3
Slave queue ID: 0

Slave Interface: swp6
MII Status: up
Speed: 10000 Mbps
Duplex: full
Link Failure Count: 0
Permanent HW addr: 44:38:39:00:32:90
Aggregator ID: 3
Slave queue ID: 0

Slave Interface: swp7
MII Status: up
Speed: 10000 Mbps
Duplex: full
Link Failure Count: 0
Permanent HW addr: 44:38:39:00:32:91
Aggregator ID: 3
```

Slave queue ID: 0



A single bridge cannot contain multiple subinterfaces of the **same** port as members. Attempting to apply such a configuration will result in an error:

```
cumulus@switch:~$ sudo brctl addbr another_bridge
cumulus@switch:~$ sudo brctl addif another_bridge swp9 swp9.100
bridge cannot contain multiple subinterfaces of the same port: swp9,
swp9.100
```

## VLAN Translation

By default, Cumulus Linux does not allow VLAN subinterfaces associated with different VLAN IDs to be part of the same bridge. Base interfaces are not explicitly associated with any VLAN IDs and are exempt from this restriction:

```
cumulus@switch:~$ sudo brctl addbr br_mix

cumulus@switch:~$ sudo ip link add link swp10 name swp10.100 type vlan id
100
cumulus@switch:~$ sudo ip link add link swp11 name swp11.200 type vlan id
200

cumulus@switch:~$ sudo brctl addif br_mix swp10.100 swp11.200
can't add swp11.200 to bridge br_mix: Invalid argument
```

In some cases, it may be useful to relax this restriction. For example, two servers may be connected to the switch using VLAN trunks, but the VLAN numbering provisioned on the two servers are not consistent. You can choose to just bridge two VLAN subinterfaces of different VLAN IDs from the servers. You do this by enabling the `sysctl net.bridge.bridge-allow-multiple-vlans`. Packets entering a bridge from a member VLAN subinterface will egress another member VLAN subinterface with the VLAN ID translated.



A bridge in [VLAN-aware mode](#) (see page 182) cannot have VLAN translation enabled for it; only bridges configured in traditional mode can utilize VLAN translation.

The following example enables the VLAN translation `sysctl`:

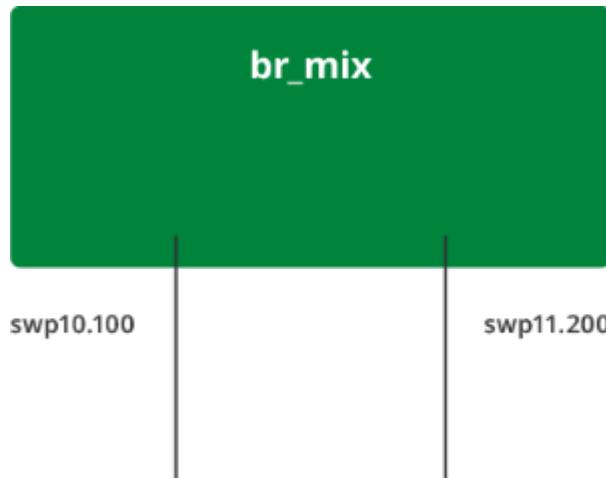
```
cumulus@switch:~$ echo net.bridge.bridge-allow-multiple-vlans = 1 | sudo
tee /etc/sysctl.d/multiple_vlans.conf
net.bridge.bridge-allow-multiple-vlans = 1
cumulus@switch:~$ sudo sysctl -p /etc/sysctl.d/multiple_vlans.conf
net.bridge.bridge-allow-multiple-vlans = 1
```

If the `sysctl` is enabled and you want to disable it, run the above example, setting the `sysctl net.bridge.bridge-allow-multiple-vlans` to 0.

Once the `sysctl` is enabled, ports with different VLAN IDs can be added to the same bridge. In the following example, packets entering the bridge `br_mix` from `swp10.100` will be bridged to `swp11.200` with the VLAN ID translated from 100 to 200:

```
cumulus@switch:~$ sudo brctl addif br_mix swp10.100 swp11.200

cumulus@switch:~$ sudo brctl show br_mix
bridge name      bridge id          STP enabled     interfaces
br_mix          8000.4438390032bd    yes           swp10.100
                                         swp11.200
```



## VLAN-aware Bridge Mode for Large-scale Layer 2 Environments

Cumulus Linux bridge driver supports two configuration modes, one that is VLAN-aware, and one that follows a more traditional Linux bridge model.

For traditional Linux bridges, the kernel supports VLANs in the form of VLAN subinterfaces. Enabling bridging on multiple VLANs means configuring a bridge for each VLAN and, for each member port on a bridge, creating one or more VLAN subinterfaces out of that port. This mode poses scalability challenges in terms of configuration size as well as boot time and run time state management, when the number of ports times the number of VLANs becomes large.

The VLAN-aware mode in Cumulus Linux implements a configuration model for large-scale L2 environments, with **one single instance** of [Spanning Tree \(see page 124\)](#). Each physical bridge member port is configured with the list of allowed VLANs as well as its port VLAN ID (either PVID or native VLAN — see below). MAC address learning, filtering and forwarding are *VLAN-aware*. This significantly reduces the configuration size, and eliminates the large overhead of managing the port/VLAN instances as subinterfaces, replacing them with lightweight VLAN bitmaps and state updates.



You can configure both VLAN-aware and traditional mode bridges on the same network in Cumulus Linux; however you should not have more than one VLAN-aware bridge on a given switch. If you are implementing [VXLANs \(see page 287\)](#), you **must** use non-aware bridges.

## Contents

(Click to expand)

- [Contents \(see page 183\)](#)
- [Defining VLAN-aware Bridge Attributes \(see page 183\)](#)
- [Basic Trunking \(see page 183\)](#)
- [VLAN Filtering/VLAN Pruning \(see page 184\)](#)
- [Untagged/Access Ports \(see page 185\)
  - \[Dropping Untagged Frames \\(see page 185\\)\]\(#\)](#)
- [VLAN Layer 3 Addressing/Switch Virtual Interfaces and other VLAN Attributes \(see page 186\)](#)
- [Using the glob Keyword to Configure Multiple Ports in a Range \(see page 187\)](#)
- [Example Configuration with Access Ports and Pruned VLANs \(see page 188\)](#)
- [Example Configuration with Bonds \(see page 188\)](#)
- [Converting a Traditional Bridge to VLAN-aware or Vice Versa \(see page 191\)](#)
- [Caveats and Errata \(see page 191\)](#)

## Defining VLAN-aware Bridge Attributes

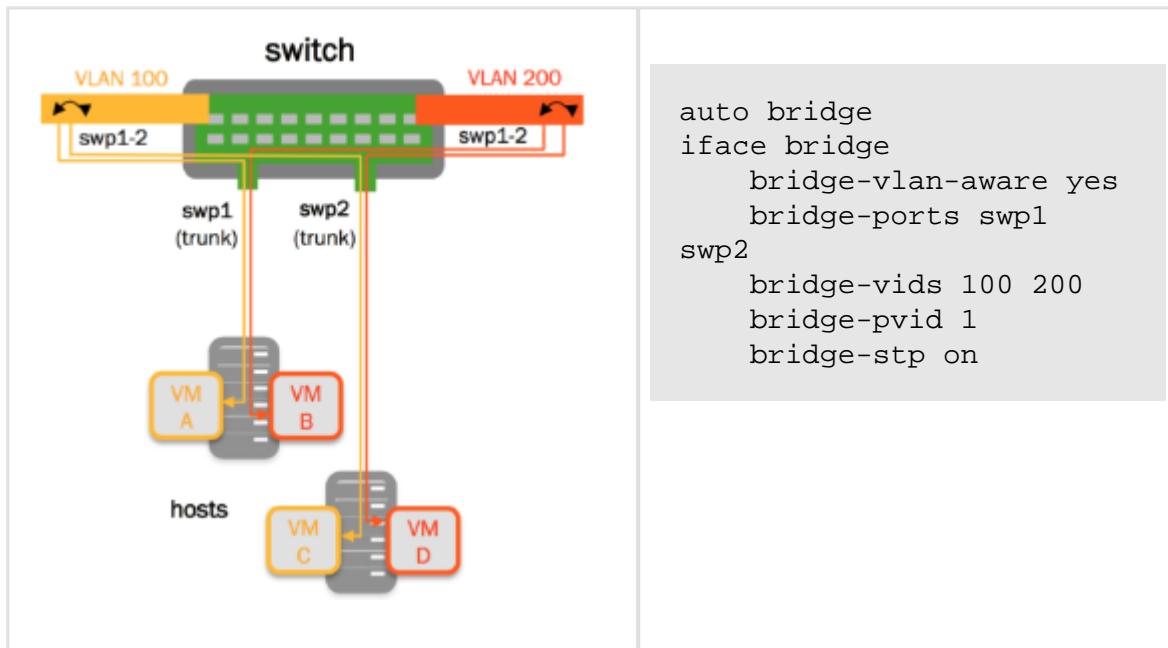
To configure a VLAN-aware bridge, include the `bridge-vlan-aware` attribute, setting it to `yes`. Name the bridge `bridge` to help ensure it is the only VLAN-aware bridge on the switch. The following attributes are useful for configuring VLAN-aware bridges:

- `bridge-vlan-aware`: Set to `yes` to indicate that the bridge is in VLAN-aware mode.
- `bridge-pvid`: A PVID is the bridge's *Primary VLAN Identifier*. The PVID defaults to 1; specifying the PVID identifies that VLAN as the native VLAN.
- `bridge-vids`: A VID is the *VLAN Identifier*, which declares the VLANs associated with this bridge.
- `bridge-access`: Declares the physical switch port as an *access port*. Access ports ignore all tagged packets; put all untagged packets into the `bridge-pvid`.
- `bridge-allow-untagged`: When set to `no`, it drops any untagged frames for a given switch port.

For a definitive list of bridge attributes, run `ifquery --syntax-help` and look for the entries under **bridge**, **bridgevlan** and **mstpclt**.

## Basic Trunking

A basic configuration for a VLAN-aware bridge configured for STP that contains two switch ports looks like this:



The above configuration actually includes 3 VLANs: the tagged VLANs 100 and 200 and the untagged (native) VLAN of 1.



The `bridge-pvid 1` is implied by default. You do not have to specify `bridge-pvid`. And while it does not hurt the configuration, it helps other users for readability.

The following configurations are identical to each other and the configuration above:

```

auto bridge
iface bridge
    bridge-vlan-
    aware yes
    bridge-ports
    swp1 swp2
    bridge-vids
    1 100 200
    bridge-stp on

```

```

auto bridge
iface bridge
    bridge-vlan-
    aware yes
    bridge-ports
    swp1 swp2
    bridge-vids
    1 100 200
    bridge-pvid 1
    bridge-stp on

```

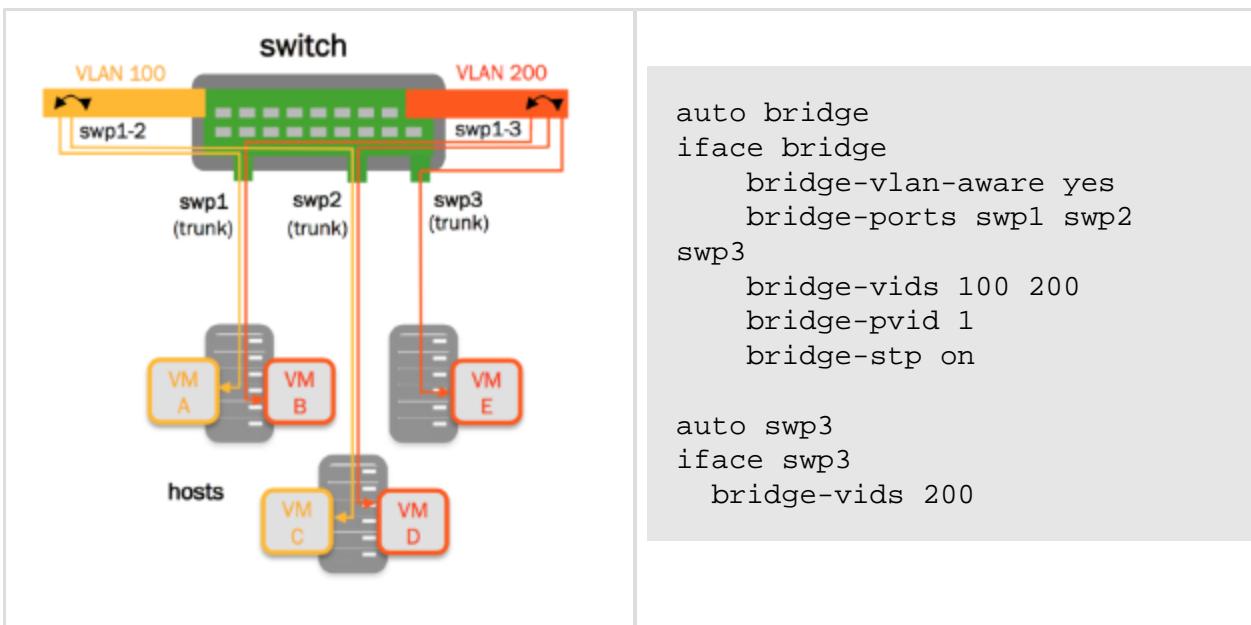
```

auto bridge
iface bridge
    bridge-vlan-
    aware yes
    bridge-ports
    swp1 swp2
    bridge-vids
    100 200
    bridge-stp on

```

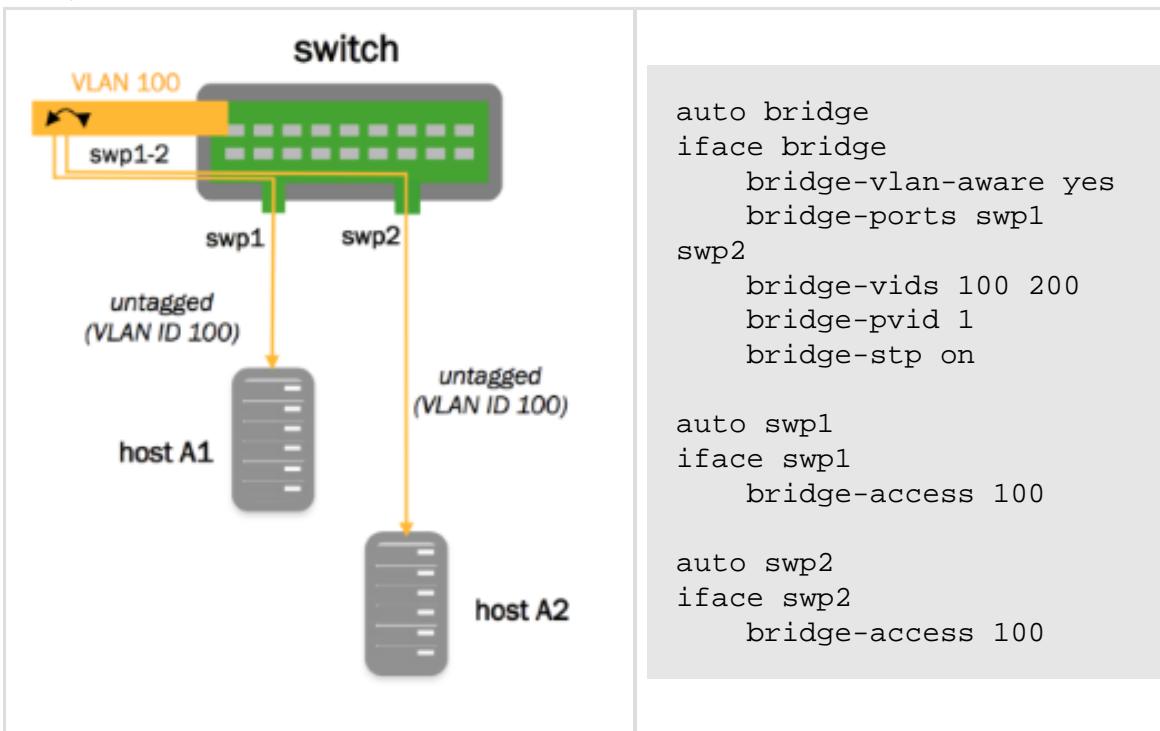
## VLAN Filtering/VLAN Pruning

By default, the bridge port inherits the bridge VIDs. A port's configuration can override the bridge VIDs. Do this by specifying port-specific VIDs using the `bridge-vids` attribute.



## Untagged/Access Ports

As described above, access ports ignore all tagged packets. In the configuration below, swp1 and swp2 are configured as access ports. All untagged traffic goes to the specified VLAN, which is VLAN 100 in the example below.



## Dropping Untagged Frames

With VLAN-aware bridge mode, it's possible to configure a switch port so it drops any untagged frames. To do this, add `bridge-allow-untagged no` under the switch port stanza in `/etc/network/interfaces`. This leaves the bridge port without a PVID and drops untagged packets.

Consider the following example bridge:

```
auto bridge
iface bridge
    bridge-vlan-aware yes
    bridge-ports swp1 swp9
    bridge-vids 2-100
    bridge-pvid 101
    bridge-stp on
```

Here is the VLAN membership for that configuration:

```
cumulus@switch$ bridge vlan show
portvlan ids
swp1 101 PVID Egress Untagged
    2-100

swp9 101 PVID Egress Untagged
    2-100

bridge 101
```

To configure swp9 to drop untagged frames, add `bridge-allow-untagged no`:

```
auto swp9
iface swp9
    bridge-allow-untagged no
```

When you check VLAN membership for that port, it shows that there is **no** untagged VLAN.

```
cumulus@switch$ bridge vlan show
portvlan ids
swp1 101 PVID Egress Untagged
    2-100

swp9 2-100

bridge 101
```

## VLAN Layer 3 Addressing/Switch Virtual Interfaces and other VLAN Attributes

When configuring the VLAN attributes for the bridge, put the attributes in a separate stanza for each VLAN interface: <bridge>.<vlanid>. If you are configuring the SVI for the native VLAN, you must declare the native VLAN in its own stanza and specify its IP address. Specifying the IP address in the bridge stanza itself returns an error.

```
auto bridge.100
iface bridge.100
    address 192.168.10.1/24
    address 2001:db8::1/32
    hwaddress 44:38:39:ff:00:00

# 12 attributes
auto bridge.100
vlan bridge.100
    bridge-igmp-querier-src 172.16.101.1
```



The `vlan` object type in the l2 attributes section above is used to specify layer 2 VLAN attributes only. Currently, the only supported layer 2 VLAN attribute is `bridge-igmp-querier-src`.

However, if your switch is configured for multicast routing, then you do not need to specify `bridge-igmp-querier-src`, as there is no need for a static IGMP querier configuration on the switch. Otherwise, the static IGMP querier configuration helps to probe the hosts to refresh their IGMP reports.

You can specify a range of VLANs as well. For example:

```
auto bridge.[1-2000]
vlan bridge.[1-2000]
    ATTRIBUTE VALUE
```

## Using the `glob` Keyword to Configure Multiple Ports in a Range

The `glob` keyword referenced in the `bridge-ports` attribute indicates that swp1 through swp52 are part of the bridge, which is a short cut that saves you from enumerating each port individually:

```
auto bridge
iface bridge
    bridge-vlan-aware yes
    bridge-ports glob swp1-52
    bridge-stp on
    bridge-vids 310 700 707 712 850 910
```

## Example Configuration with Access Ports and Pruned VLANs

The following example contains an access port and a switch port that is *pruned*; that is, it only sends and receives traffic tagged to and from a specific set of VLANs declared by the `bridge-vids` attribute. It also contains other switch ports that send and receive traffic from all the defined VLANs.

```
# ports swp3-swp48 are trunk ports which inherit vlans from the
'bridge'
# ie vlans 310,700,707,712,850,910
#
auto bridge
iface bridge
    bridge-vlan-aware yes
    bridge-ports glob swp1-52
    bridge-stp on
    bridge-vids 310 700 707 712 850 910

auto swp1
iface swp1
    mstpctl-portadminedge yes
    mstpctl-bpduguard yes
    bridge-access 310

# The following is a trunk port that is "pruned".
# native vlan is 1, but only .1q tags of 707, 712, 850 are
# sent and received
#
auto swp2
iface swp2
    mstpctl-portadminedge yes
    mstpctl-bpduguard yes
    bridge-vids 707 712 850

# The following port is the trunk uplink and inherits all vlans
# from 'bridge'; bridge assurance is enabled using 'portnetwork'
attribute
auto swp49
iface swp49
    mstpctl-portpathcost 10
    mstpctl-portnetwork yes

# The following port is the trunk uplink and inherits all vlans
# from 'bridge'; bridge assurance is enabled using 'portnetwork'
attribute
auto swp50
iface swp50
    mstpctl-portpathcost 0
    mstpctl-portnetwork yes
```

## ***Example Configuration with Bonds***

This configuration demonstrates a VLAN-aware bridge with a large set of bonds. The bond configurations are generated from a [Mako](#) template.

```

#
# vlan-aware bridge with bonds example
#
# uplink1, peerlink and downlink are bond interfaces.
# 'bridge' is a vlan aware bridge with ports uplink1, peerlink
# and downlink (swp2-20).
#
# native vlan is by default 1
#
# 'bridge-vids' attribute is used to declare vlans.
# 'bridge-pvid' attribute is used to specify native vlans if other
than 1
# 'bridge-access' attribute is used to declare access port
#
auto lo
iface lo

auto eth0
iface eth0 inet dhcp

# bond interface
auto uplink1
iface uplink1
    bond-slaves swp32
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lacp-rate 1
    bond-min-links 1
    bond-xmit-hash-policy layer3+4
    bridge-vids 2000-2079

# bond interface
auto peerlink
iface peerlink
    bond-slaves swp30 swp31
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lacp-rate 1
    bond-min-links 1
    bond-xmit-hash-policy layer3+4
    bridge-vids 2000-2079 4094

# bond interface
auto downlink

```

```
iface downlink
    bond-slaves swp1
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lacp-rate 1
    bond-min-links 1
    bond-xmit-hash-policy layer3+4
    bridge-vids 2000-2079

#
# Declare vlans for all swp ports
# swp2-20 get vlans from 2004 to 2022.
# The below uses mako templates to generate iface sections
# with vlans for swp ports
#
%for port, vlanid in zip(range(2, 20), range(2004, 2022)) :
    auto swp${port}
    iface swp${port}
        bridge-vids ${vlanid}

%endfor

# svi vlan 4094
auto bridge.4094
iface bridge.4094
    address 11.100.1.252/24

# 12 attributes for vlan 4094
auto bridge.4094
vlan bridge.4094
    bridge-igmp-querier-src 172.16.101.1

#
# vlan-aware bridge
#
auto bridge
iface bridge
    bridge-vlan-aware yes
    bridge-ports uplink1 peerlink downlink glob swp2-20
    bridge-stp on

# svi peerlink vlan
auto peerlink.4094
iface peerlink.4094
    address 192.168.10.1/30
    broadcast 192.168.10.3
```

## Converting a Traditional Bridge to VLAN-aware or Vice Versa

You cannot automatically convert a traditional bridge to/from a VLAN-aware bridge simply by changing the configuration in the `/etc/network/interfaces` file. If you need to change the mode for a bridge, do the following:

1. Delete the traditional mode bridge from the configuration and bring down all its member switch port interfaces.
2. Create a new VLAN-aware bridge, as described above.
3. Bring up the bridge.

These steps assume you are converting a traditional mode bridge to a VLAN-aware one. To do the opposite, delete the VLAN-aware bridge in step 1, and create a new traditional mode bridge in step 2.

## Caveats and Errata

- **STP:** Because [Spanning Tree and Rapid Spanning Tree \(see page 124\)](#) (STP) are enabled on a per-bridge basis, VLAN-aware mode essentially supports a single instance of STP across all VLANs. A common practice when using a single STP instance for all VLANs is to define all every VLAN on each switch in the spanning tree instance. `mstpd` continues to be the user space protocol daemon, and Cumulus Linux supports RSTP.
- **IGMP snooping:** IGMP snooping and group membership are supported on a per-VLAN basis, though the IGMP snooping configuration (including enable/disable, mrouter port and so forth) are defined on a per-bridge port basis.
- **VXLANS:** Use the traditional configuration mode for [VXLAN configuration \(see page 287\)](#).
- **Reserved VLAN range:** For hardware data plane internal operations, the switching silicon requires VLANs for every physical port, Linux bridge, and layer 3 subinterface. Cumulus Linux reserves a range of 700 VLANs by default; this range is 3300-3999. In case any of your user-defined VLANs conflict with the default reserved range, you can modify the range, as long as the new range is a contiguous set of VLANs with IDs anywhere between 2 and 4094, and the minimum size of the range is 300 VLANs:

1. Edit `/etc/cumulus/switchd.conf`, uncomment `resv_vlan_range` and specify the new range.
2. [Restart `switchd` \(see page 90\)](#) (`sudo service switchd restart`) for the new range to take effect.



While restarting `switchd`, all running ports will flap and forwarding will be interrupted ([see page 90](#)).

- **VLAN translation:** A bridge in VLAN-aware mode cannot have VLAN translation enabled for it; only bridges configured in [traditional mode \(see page 162\)](#) can utilize VLAN translation.

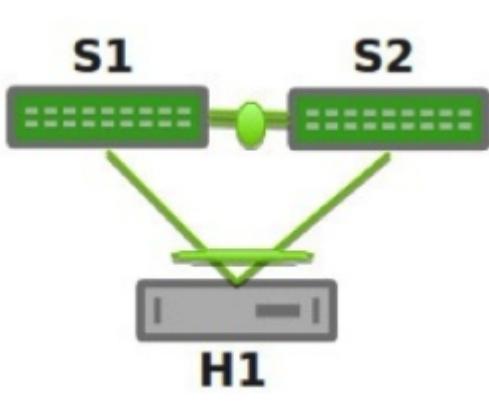
## Multi-Chassis Link Aggregation - MLAG

*Host HA* is a set of L2 and L3 features supporting high availability for hosts, including multi-Chassis Link Aggregation (MLAG) for L2 and [redistribute neighbor](#) (an experimental L3 feature).

Multi-Chassis Link Aggregation, or MLAG, enables a server or switch with a two-port bond (such as a link aggregation group/LAG, EtherChannel, port group, or trunk) to connect those ports to different switches and operate as if they are connected to a single, logical switch. This provides greater redundancy and greater system throughput.

Dual-connected devices can create LACP bonds that contain links to each physical switch. Thus, active-active links from the dual-connected devices are supported even though they are connected to two different physical switches.

A basic setup looks like this:



The two switches, S1 and S2, known as *peer switches*, cooperate so that they appear as a single device to host H1's bond. H1 distributes traffic between the two links to S1 and S2 in any manner that you configure on the host. Similarly, traffic inbound to H1 can traverse S1 or S2 and arrive at H1.

## Contents

(Click to expand)

- Contents (see page 192)
- MLAG Requirements (see page 193)
- LACP and Dual-Connectedness (see page 194)
- Understanding Switch Roles (see page 195)
- Configuring MLAG (see page 195)
  - Reserved MAC Address Range (see page 196)
  - Configuring the Host or Switch (see page 196)
  - Configuring the Interfaces (see page 196)
  - Example MLAG Configuration (see page 197)
  - Configuring MLAG with a Traditional Mode Bridge (see page 201)
  - Using the clagd Command Line Interface (see page 202)
- Peer Link Interfaces and the protodown State (see page 202)
  - Specifying a Backup Link (see page 203)
- Monitoring Dual-Connected Peers (see page 204)
- Configuring Layer 3 Routed Uplinks (see page 205)
- IGMP Snooping with MLAG (see page 205)

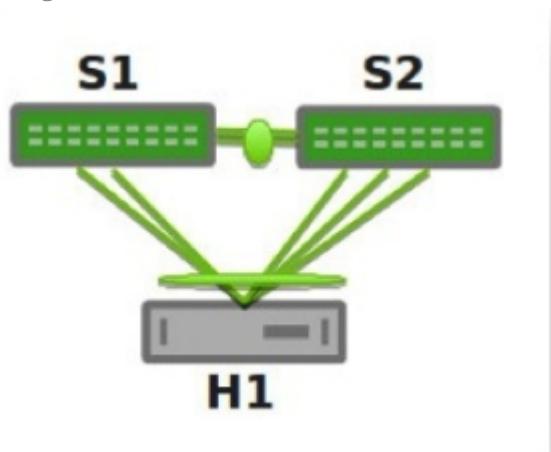
- Monitoring the Status of the clagd Service (see page 206)
- MLAG Best Practices (see page 207)
  - Understanding MTU in an MLAG Configuration (see page 207)
- STP Interoperability with MLAG (see page 207)
  - Debugging STP with MLAG (see page 208)
  - Best Practices for STP with MLAG (see page 208)
- Troubleshooting MLAG (see page 209)
- Caveats and Errata (see page 209)
- Configuration Files (see page 209)

## **MLAG Requirements**

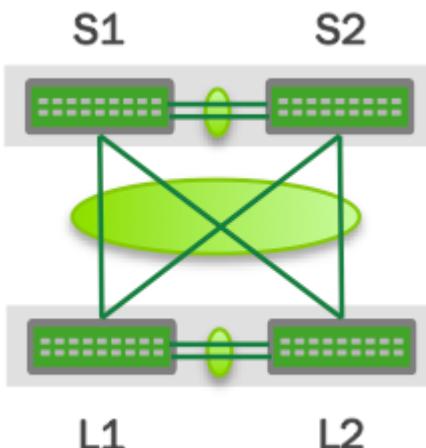
MLAG has these requirements:

- There must be a direct connection between the two peer switches implementing MLAG (S1 and S2). This is typically a bond for increased reliability and bandwidth.
- There must be only two peer switches in one MLAG configuration, but you can have multiple configurations in a network for *switch-to-switch MLAG* (see below).
- The peer switches implementing MLAG must be running Cumulus Linux version 2.5 or later.
- You must specify a unique `c1ag-id` for every dual-connected bond on each peer switch; the value must be between 1 and 65535 and must be the same on both peer switches in order for the bond to be considered *dual-connected*.
- The dual-connected devices (hosts or switches) must use LACP (IEEE 802.3ad/802.1ax) to form the bond. The peer switches must also use LACP.

More elaborate configurations are also possible. The number of links between the host and the switches can be greater than two, and does not have to be symmetrical:



Additionally, since S1 and S2 appear as a single switch to other bonding devices, pairs of MLAG switches can also be connected to each other in a switch-to-switch MLAG setup:

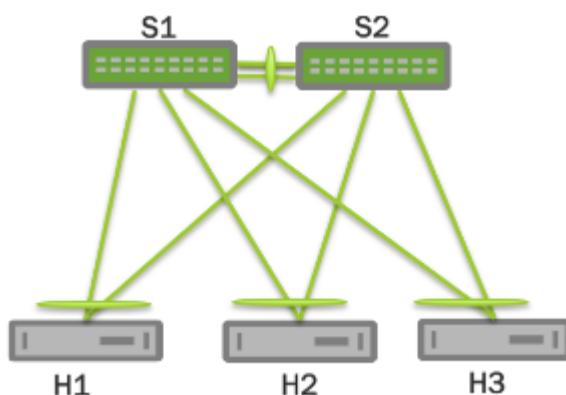


In this case, L1 and L2 are also MLAG peer switches, and thus present a two-port bond from a single logical system to S1 and S2. S1 and S2 do the same as far as L1 and L2 are concerned. For a switch-to-switch MLAG configuration, each switch pair must have a unique system MAC address. In the above example, switches L1 and L2 each have the same system MAC address configured. Switch pair S1 and S2 each have the same system MAC address configured; however, it is a different system MAC address than the one used by the switch pair L1 and L2.

## LACP and Dual-Connectedness

In order for MLAG to operate correctly, the peer switches must know which links are *dual-connected*, or are connected to the same host or switch. To do this, specify a `c1ag-id` for every dual-connected bond on each peer switch; the `c1ag-id` must be the same for the corresponding bonds on both peer switches. [Link Aggregation Control Protocol \(LACP\)](#), the IEEE standard protocol for managing bonds, is used for verifying dual-connectedness. LACP runs on the dual-connected device and on each of the peer switches. On the dual-connected device, the only configuration requirement is to create a bond that will be managed by LACP.

On each of the peer switches the links connected to the dual-connected host or switch must be placed in the bond. This is true even if the links are a single port on each peer switch, where each port is placed into a bond, as shown below:



All of the dual-connected bonds on the peer switches have their system ID set to the MLAG system ID. Therefore, from the point of view of the hosts, each of the links in its bond is connected to the same system, and so the host will use both links.

Each peer switch periodically makes a list of the LACP partner MAC addresses of all of their bonds and sends that list to its peer (using the `c1agd` service; see below). The LACP partner MAC address is the MAC address of the system at the other end of a bond, which in the figure above would be hosts H1, H2 and H3. When a switch receives this list from its peer, it compares the list to the LACP partner MAC addresses on its switch. If any matches are found and the `c1ag-id` for those bonds match, then that bond is a dual-connected bond. You can also find the LACP partner MAC address in the `/sys/class/net/<bondname>/bonding/ad_partner_mac sysfs` file for each bond.

## **Understanding Switch Roles**

Each MLAG-enabled switch in the pair has a role. When the peering relationship is established between the two switches, one switch will be in *primary* role, and the other one will be in *secondary* role. When an MLAG-enabled switch is in the secondary role, it does not send STP BPDUs on dual-connected links; it only sends BPDUs on single-connected links. The switch in the primary role sends STP BPDUs on all single- and dual-connected links.

<b>Send BPDUs</b>	<b>Primary</b>	<b>Secondary</b>
Single-connected links	Yes	Yes
Dual-connected links	Yes	No

By default, the role is determined by comparing the MAC addresses of the two sides of the peering link; the switch with the lower MAC address assumes the primary role. You can override this by setting the priority configuration, either by specifying the `c1agd-priority` option in `/etc/network/interfaces`, or by using `c1agctl`. The switch with the lower priority value is given the primary role; the default value is 32768, and the range is 0 to 65535. Read the `c1agd(8)` and `c1agctl(8)` man pages for more information.

When the `c1agd` service is exited during switch reboot or the service is stopped in the primary switch, the peer switch that is in the secondary role will become primary. If the primary switch goes down without stopping the `c1agd` service for any reason or the peer link goes down, the secondary switch will **not** change its role. In case the peer switch is determined to be not alive, the switch in the secondary role will roll back the LACP system ID to be the bond interface MAC address instead of the `c1agd-sys-mac` and the switch in primary role uses the `c1agd-sys-mac` as the LACP system ID on the bonds.

## **Configuring MLAG**

Configuring MLAG involves:

- On the dual-connected devices, create a bond that uses LACP.
- On each peer switch, configure the interfaces, including bonds, VLANs, bridges and peer links.



MLAG synchronizes the dynamic state between the two peer switches, but it does not synchronize the switch configurations. After modifying the configuration of one peer switch, you must make the same changes to the configuration on the other peer switch. This applies to all configuration changes, including:

- Port configuration: For example, VLAN membership, [MTU \(see page 207\)](#), and bonding parameters.
- Bridge configuration: For example, spanning tree parameters or bridge properties.

- Static address entries: For example, static FDB entries and static IGMP entries.
- QoS configuration: For example, ACL entries.

You can verify the configuration of VLAN membership using the `clagctl -v verifyvlans` command.

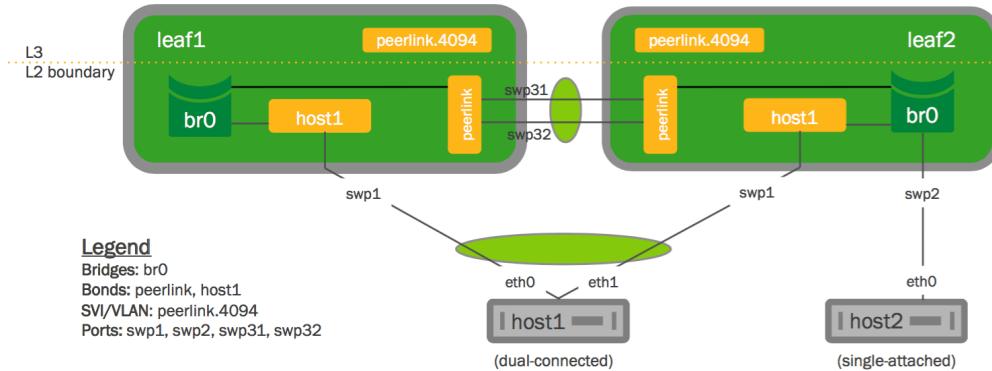
## Reserved MAC Address Range

In order to prevent MAC address conflicts with other interfaces in the same bridged network, Cumulus Networks has reserved a range of MAC addresses specifically to use with MLAG. This range of MAC addresses is 44:39:39:ff:00:00 to 44:39:39:ff:ff:ff.

Cumulus Networks recommends you use this range of MAC addresses when configuring MLAG.

## Configuring the Host or Switch

On your dual-connected device, create a bond that uses LACP. The method you use varies with the type of device you are configuring. The following image is a basic MLAG configuration, showing all the essential elements; a more detailed two-leaf/two-spine configuration is [below \(see page \)](#).



## Configuring the Interfaces

Every interface that connects to the MLAG pair from a dual-connected device should be placed into a **bond** ([see page 158](#)), even if the bond contains only a single link on a single physical switch (since the MLAG pair contains two or more links). Layer 2 data travels over this bond. In the examples throughout this chapter, **peerlink** is the name of the bond.

Single-attached hosts, also known as *orphan ports*, can be just a member of the bridge.

Additionally, the fast mode of LACP should be configured on the bond to allow more timely updates of the LACP state. These bonds will then be placed in a bridge, which will include the peer link between the switches.

In order to enable communication between the `clagd` services on the peer switches, you should choose an unused VLAN (also known as a *switched virtual interface* or *SVI* here) and assign an unrouteable link-local address to give the peer switches layer 3 connectivity between each other. To ensure that the VLAN is completely independent of the bridge and spanning tree forwarding decisions, configure the VLAN as a VLAN subinterface on the peer link bond rather than the VLAN-aware bridge. Cumulus Networks recommends you use 4094 for the peerlink VLAN (**peerlink.4094** below) if possible.

You can also specify a backup interface, which is any layer 3 backup interface for your peer links in the event that the peer link goes down. [See below \(see page 203\)](#) for more information about the backup link.

For example, if peerlink is the inter-chassis bond, and VLAN 4094 is the peerlink VLAN, configure peerlink 4094 using:

```
auto peerlink.4094
iface peerlink.4094
    address 169.254.1.1/30
    clagd-enable yes
    clagd-peer-ip 169.254.1.2
    clagd-backup-ip 10.0.1.50
    clagd-sys-mac 44:39:39:FF:40:94
```

Then run `ifup` on the peerlink VLAN interface. In this example, the command would be `sudo ifup peerlink.4094`.

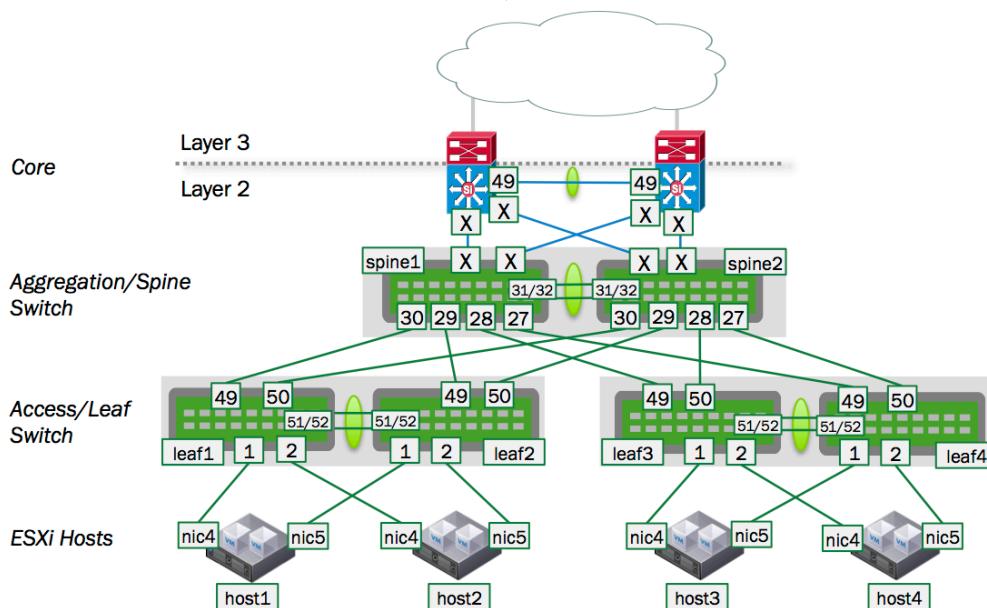
There is no need to add VLAN 4094 to the bridge VLAN list, as it is unnecessary there.



Keep in mind that when you change the MLAG configuration in the `interfaces` file, the changes take effect when you bring the peerlink interface up with `ifup`. Do **not** use `service clagd restart` to apply the new configuration.

## Example MLAG Configuration

An example configuration is included below. It configures two bonds for MLAG, each with a single port, a peer link that is a bond with two member ports, and three VLANs on each port. You store the configuration in `/etc/network/interfaces` on each peer switch.



Configuring these interfaces uses syntax from `ifupdown2` and the [VLAN-aware bridge driver mode](#) (see [page 182](#)). The bridges use these Cumulus Linux-specific keywords:

- `bridge-vids`, which defines the allowed list of tagged 802.1q VLAN IDs for all bridge member interfaces. You can specify non-contiguous ranges with a space-separated list, like `bridge-vids 100-200 300 400-500`.
- `bridge-pvid`, which defines the untagged VLAN ID for each port. This is commonly referred to as the *native VLAN*.

The bridge configurations below indicate that each bond carries tagged frames on VLANs 1000 to 3000 but untagged frames on VLAN 1. Also, take note on how you configure the VLAN subinterface used for `clagd` communication (`peerlink.4094` in the sample configuration below).



At minimum, this VLAN subinterface should not be in your Layer 2 domain, and you should give it a very high VLAN ID (up to 4094). Read more about the [range of VLAN IDs you can use \(see page\)](#).

The configuration for the spines should look like the following (note that the `clag-id` and `clagd-sys-mac` must be the same for the corresponding bonds on spine1 and spine2):

spine1

```
# The loopback network
interface auto lo
iface lo
inet loopback

# The primary network
interface
auto eth0
iface eth0
    address 10.0.0.1
    netmask 255.255.255.0

auto peerlink
iface peerlink
    bond-slaves swp31 swp32
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lACP-rate 1
    bond-min-links 1
    bond-xmit-hash-policy
layer3+4

auto peerlink.4094
iface peerlink.4094
    address 169.254.255.1
    netmask 255.255.255.0
    clagd-priority 4096
    clagd-peer-ip 169.254.255.2
```

spine2

```
# The loopback network
interface auto lo
iface lo
inet loopback

# The primary network
interface
auto eth0
iface eth0
    address 10.0.0.2
    netmask 255.255.255.0

auto peerlink
iface peerlink
    bond-slaves swp31 swp32
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lACP-rate 1
    bond-min-links 1
    bond-xmit-hash-policy
layer3+4

auto peerlink.4094
iface peerlink.4094
    address 169.254.255.2
    netmask 255.255.255.0
    clagd-priority 8192
    clagd-peer-ip 169.254.255.1
```

```

clagd-backup-ip 10.0.0.2
clagd-sys-mac 44:38:39:ff:
00:01

# ToR pair #1
auto downlink1
iface downlink1
    bond-slaves swp29 swp30
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lacp-rate 1
    bond-min-links 1
    bond-xmit-hash-policy
layer3+4
    clag-id 1

# ToR pair #2
auto downlink2
iface downlink2
    bond-slaves swp27 swp28
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lacp-rate 1
    bond-min-links 1
    bond-xmit-hash-policy
layer3+4
    clag-id 2

auto br
iface br
    bridge-vlan-aware yes
    bridge-ports uplinkA
peerlink downlink1 downlink2
    bridge-stp on
    bridge-vids 1000-3000
    bridge-pvid 1
    bridge-mcsnoop 1

```

```

clagd-backup-ip 10.0.0.1
clagd-sys-mac 44:38:39:ff:
00:01

# ToR pair #1
auto downlink1
iface downlink1
    bond-slaves swp29 swp30
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lacp-rate 1
    bond-min-links 1
    bond-xmit-hash-policy
layer3+4
    clag-id 1

# ToR pair #2
auto downlink2
iface downlink2
    bond-slaves swp27 swp28
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lacp-rate 1
    bond-min-links 1
    bond-xmit-hash-policy
layer3+4
    clag-id 2

auto br
iface br
    bridge-vlan-aware yes
    bridge-ports uplinkA
peerlink downlink1 downlink2
    bridge-stp on
    bridge-vids 1000-3000
    bridge-pvid 1
    bridge-mcsnoop 1

```

Here is an example configuration file for the switches leaf1 and leaf2. Note that the `clag-id` and `clagd-sys-mac` must be the same for the corresponding bonds on leaf1 and leaf2:

leaf1

leaf2

```

# The loopback network
interface
auto lo
iface lo inet loopback

# The primary network interface
auto eth0
iface eth0
    address 10.0.0.3
    netmask 255.255.255.0

auto spine1-2
iface spine1-2
    bond-slaves swp49 swp50
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lACP-rate 1
    bond-min-links 1
    bond-xmit-hash-policy
layer3+4
    clag-id 1

auto peerlink
iface peerlink
    bond-slaves swp51 swp52
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lACP-rate 1
    bond-min-links 1
    bond-xmit-hash-policy
layer3+4

auto peerlink.4094
iface peerlink.4094
    address 169.254.255.3
    netmask 255.255.255.0
    clagd-priority 4096
    clagd-peer-ip 169.254.255.4
    clagd-backup-ip 10.0.0.4
    clagd-sys-mac 44:38:39:ff:
01:02

auto host1
iface host1
    bond-slaves swp1
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lACP-rate 1
  
```

```

# The loopback network
interface
auto lo
iface lo inet loopback

# The primary network interface
auto eth0
iface eth0
    address 10.0.0.4
    netmask 255.255.255.0

auto spine1-2
iface spine1-2
    bond-slaves swp49 swp50
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lACP-rate 1
    bond-min-links 1
    bond-xmit-hash-policy
layer3+4
    clag-id 1

auto peerlink
iface peerlink
    bond-slaves swp51 swp52
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lACP-rate 1
    bond-min-links 1
    bond-xmit-hash-policy
layer3+4

auto peerlink.4094
iface peerlink.4094
    address 169.254.255.4
    netmask 255.255.255.0
    clagd-priority 8192
    clagd-peer-ip 169.254.255.3
    clagd-backup-ip 10.0.0.3
    clagd-sys-mac 44:38:39:ff:
01:02

auto host1
iface host1
    bond-slaves swp1
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lACP-rate 1
  
```

```

bond-min-links 1
bond-xmit-hash-policy
layer3+4
  clag-id 2
  mstpcctl-portadmindedge yes
  mstpcctl-bpduguard yes

auto host2
iface host2
  bond-slaves swp2
  bond-mode 802.3ad
  bond-miimon 100
  bond-use-carrier 1
  bond-lacp-rate 1
  bond-min-links 1
  bond-xmit-hash-policy
layer3+4
  clag-id 3
  mstpcctl-portadmindedge yes
  mstpcctl-bpduguard yes

auto br0
iface br0
  bridge-vlan-aware yes
  bridge-ports spinel-2
peerlink host1 host2
  bridge-stp on
  bridge-vids 1000-3000
  bridge-pvid 1

```

```

bond-min-links 1
bond-xmit-hash-policy
layer3+4
  clag-id 2
  mstpcctl-portadmindedge yes
  mstpcctl-bpduguard yes

auto host2
iface host2
  bond-slaves swp2
  bond-mode 802.3ad
  bond-miimon 100
  bond-use-carrier 1
  bond-lacp-rate 1
  bond-min-links 1
  bond-xmit-hash-policy
layer3+4
  clag-id 3
  mstpcctl-portadmindedge yes
  mstpcctl-bpduguard yes

auto br0
iface br0
  bridge-vlan-aware yes
  bridge-ports spinel-2
peerlink host1 host2
  bridge-stp on
  bridge-vids 1000-3000
  bridge-pvid 1

```

The configuration is almost identical, except for the IP addresses used for managing the `clagd` service.



In the configurations above, the `clagd-peer-ip` and `clagd-sys-mac` parameters are mandatory, while the rest are optional. When mandatory `clagd` commands are present under a peer link subinterface, by default `clagd-enable` is treated as `yes`; to disable `clagd` on the subinterface, set `clagd-enable` to `no`. Use `clagd-priority` to set the role of the MLAG peer switch to primary or secondary. Each peer switch in an MLAG pair must have the same `clagd-sys-mac` setting. Each `clagd-sys-mac` setting should be unique to each MLAG pair in the network. For more details refer to `man clagd`.

## Configuring MLAG with a Traditional Mode Bridge

It's possible to configure MLAG with a bridge in [traditional mode](#) (see page 162) instead of [VLAN-aware mode](#) (see page 182). In order to do so, the peer link and all dual-connected links must be configured as [untagged/native](#) (see page 171) ports on a bridge (note the absence of any VLANs in the `bridge-ports` line and the lack of the `bridge-vlan-aware` parameter below):

```
auto br
iface br
  bridge-ports peerlink spine1-2 host1 host2
```



For a deeper comparison of traditional versus VLAN-aware bridge modes, read this [knowledge base article](#).

## ***Using the clagd Command Line Interface***

A command line utility called `clagctl` is available for interacting with a running `clagd` service to get status or alter operational behavior. For detailed explanation of the utility, please refer to the `clagctl(8)` man page. The following is a sample output of the MLAG operational status displayed by the utility:

```
cumulus@switch$ clagctl
The peer is alive
  Our Priority, ID, and Role: 8192 00:e0:ec:26:50:89 primary
  Peer Priority, ID, and Role: 8192 00:e0:ec:27:49:f6 secondary
  Peer Interface and IP: peerlink.4094 169.254.255.2
  System MAC: 44:38:39:ff:00:01

          Dual Attached Ports
Our Interface    Peer Interface    CLAG Id
-----  -----
downlink1        downlink1        1
downlink2        downlink2        2
```

## ***Peer Link Interfaces and the protodown State***

In addition to the standard UP and DOWN administrative states, an interface that is a member of an MLAG bond can also be in a `protodown` state. When MLAG detects a problem that could result in connectivity issues such as traffic black-holing or a network meltdown if the link carrier was left in an UP state, it can put that interface into `protodown` state. Such connectivity issues include:

- When the peer link goes down but the peer switch is up (that is, the backup link is active).
- When the bond is configured with an MLAG ID, but the `clagd` service is not running (whether it was deliberately stopped or simply died).
- When an MLAG-enabled node is booted or rebooted, the MLAG bonds are placed in a `protodown` state until the node establishes a connection to its peer switch, or five minutes have elapsed.

When an interface goes into a `protodown` state, it results in a local OPER DOWN (carrier down) on the interface. As of Cumulus Linux 2.5.5, the `protodown` state can be manipulated with the `ip link set` command. Given its use in preventing network meltdowns, manually manipulating `protodown` is not recommended outside the scope of interaction with the Cumulus Networks support team.

The following `ip link show` command output shows an interface in `protodown` state. Notice that the link carrier is down (NO-CARRIER):

```
cumulus@switch:~$ ip link show swp1
3: swp1: <NO-CARRIER,BROADCAST,MULTICAST,SLAVE,UP> mtu 1500 qdisc
pfifo_fast master host-bond1 state DOWN mode DEFAULT qlen 500 protodown on
link/ether 44:38:39:00:69:84 brd ff:ff:ff:ff:ff:ff
```

## Specifying a Backup Link

You can specify a backup link for your peer links in the event that the peer link goes down. When this happens, the `clagd` service uses the backup link to check the health of the peer switch. To configure this, edit `/etc/network/interfaces` and add `clag-backup-ip <ADDRESS>` to the peer link configuration. Here's an example:

```
auto peerlink.4094
iface peerlink.4094
    address 169.254.255.1
    netmask 255.255.255.0
    clagd-enable yes
    clagd-priority 8192
    clagd-peer-ip 169.254.255.2
    clagd-backup-ip 10.0.1.50
    clagd-sys-mac 44:38:39:ff:00:01
    clagd-args --priority 1000
```



The backup IP address must be different than the peer link IP address (`clagd-peer-ip` above). It must be reachable by a route that doesn't use the peer link and it must be in the same network namespace as the peer link IP address.

Cumulus Networks recommends you use the switch's management IP address for this purpose.

You can also specify the backup UDP port. The port defaults to 5342, but you can configure it as an argument in `clagd-args` using `--backupPort <PORT>`.

```
auto peerlink.4094
iface peerlink.4094
    address 169.254.255.1
    netmask 255.255.255.0
    clagd-enable yes
    clagd-priority 8192
    clagd-peer-ip 169.254.255.2
```

```
clagd-backup-ip 10.0.1.50
clagd-sys-mac 44:38:39:ff:00:01
clagd-args --backupPort 5400
```

You can see the backup IP address if you run `clagctl`:

```
cumulus@switch$ clagctl
The peer is alive
  Our Priority, ID, and Role: 8192 00:e0:ec:26:50:89 primary
  Peer Priority, ID, and Role: 8192 00:e0:ec:27:49:f6 secondary
    Peer Interface and IP: peerlink.4094 169.254.255.2
      Backup IP: 10.0.1.50
      System MAC: 44:38:39:ff:00:01

      Dual Attached Ports
  Our Interface      Peer Interface      CLAG Id
  -----            -----
  downlink1        downlink1          1
  downlink2        downlink2          2
```

## ***Monitoring Dual-Connected Peers***

Upon receipt of a valid message from its peer, the switch knows that `clagd` is alive and executing on that peer. This causes `clagd` to change the system ID of each bond that was assigned a `clag-id` from the default value (the MAC address of the bond) to the system ID assigned to both peer switches. This makes the hosts connected to each switch act as if they are connected to the same system so that they will use all ports within their bond. Additionally, `clagd` determines which bonds are dual-connected and modifies the forwarding and learning behavior to accommodate these dual-connected bonds.

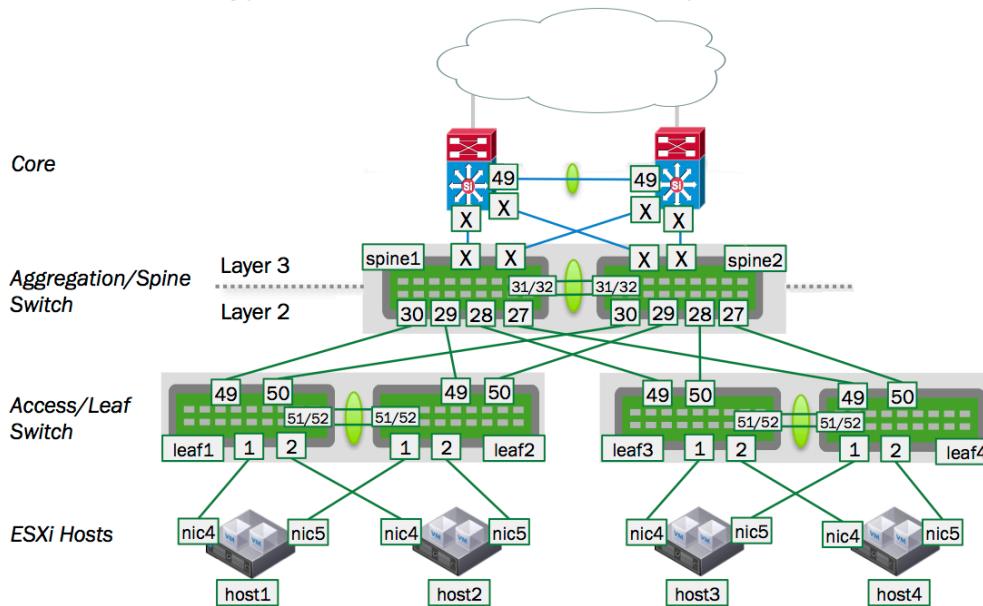
If the peer does not receive any messages for three update intervals, then that peer switch is assumed to no longer be acting as an MLAG peer. In this case, the switch reverts all configuration changes so that it operates as a standard non-MLAG switch. This includes removing all statically assigned MAC addresses, clearing the egress forwarding mask, and allowing addresses to move from any port to the peer port. Once a message is again received from the peer, MLAG operation starts again as described earlier. You can configure a custom timeout setting by adding `--peerTimeout <VALUE>` to `clagd-args` in `/etc/network/interfaces`.

Once bonds are identified as dual-connected, `clagd` sends more information to the peer switch for those bonds. The MAC addresses (and VLANs) that have been dynamically learned on those ports are sent along with the LACP partner MAC address for each bond. When a switch receives MAC address information from its peer, it adds MAC address entries on the corresponding ports. As the switch learns and ages out MAC addresses, it informs the peer switch of these changes to its MAC address table so that the peer can keep its table synchronized. Periodically, at 45% of the bridge ageing time, a switch will send its entire MAC address table to the peer, so that peer switch can verify that its MAC address table is properly synchronized.

The switch sends an update frequency value in the messages to its peer, which tells `clagd` how often the peer will send these messages. You can configure a different frequency by adding `--lacpPoll <SECONDS>` to `clagd-args` in `/etc/network/interfaces`.

## Configuring Layer 3 Routed Uplinks

In this scenario, the spine switches connect at layer 3, as shown in the image below. Alternatively, the spine switches can be singly connected to each core switch at layer 3 (not shown below).



In this design, the spine switches route traffic between the server hosts in the layer 2 domains and the core. The servers (host1 - host4) each have a layer 2 connection up to the spine layer where the default gateway for the host subnets resides. However, since the spine switches as gateway devices communicate at layer 3, you need to configure a protocol such as [VRR \(see page 215\)](#) (Virtual Router Redundancy) between the spine switch pair to support active/active forwarding.

Then, to connect the spine switches to the core switches, you need to determine whether the routing is static or dynamic. If it's dynamic, you must choose which protocol — [OSPF \(see page 332\)](#) or [BGP \(see page 345\)](#) — to use. When enabling a routing protocol in an MLAG environment it is also necessary to manage the uplinks, because by default MLAG is not aware of layer 3 uplink interfaces. In the event of a peerlink failure MLAG does not remove static routes or bring down a BGP or OSPF adjacency unless a separate link state daemon such as `ifplugd` is used.

## IGMP Snooping with MLAG

IGMP snooping processes IGMP reports received on a bridge port in a bridge to identify hosts that are configured to receive multicast traffic destined to that group. An IGMP query message received on a port is used to identify the port that is connected to a router and configured to receive multicast traffic.

IGMP snooping is enabled by default on the bridge. IGMP snooping multicast database entries and router port entries are synced to the peer MLAG switch. If there is no multicast router in the VLAN, the IGMP querier can be configured on the switch to generate IGMP query messages by adding a configuration like the following to `/etc/network/interfaces`:

```
auto br.100
vlan br.100
    #igmp snooping is enabled by default, but is shown here for completeness
```

```
bridge-mcsnoop 1
# If you need to specify the querier IP address
bridge-igmp-querier-source 123.1.1.1
```

To display multicast group and router port information, use the `bridge -d mdb show` command:

```
cumulus@switch:~# sudo bridge -d mdb show
dev br port bond0 vlan 100 grp 234.1.1.1 temp
router ports on br: bond0
```

#### Runtime Configuration (Advanced)

```
cumulus@switch:~# sudo brctl setmcqv4src br 100 123.1.1.1
cumulus@switch:~# sudo brctl setmcquerier br 1
cumulus@switch:~# sudo brctl showmcqv4src br

vlan      querier address
100       123.1.1.1
```

## **Monitoring the Status of the clagd Service**

Due to the critical nature of the `clagd` service, an external process, called `jdoe`, continuously monitors the status of `clagd`. If the `clagd` service dies or becomes unresponsive for any reason, the `jdoe` process will get `clagd` up and running again. This monitoring is automatically configured and enabled as long as `clagd` is enabled (that is, `clagd-peer-ip` and `clagd-sys-mac` are configured in `/etc/network/interfaces`) and `clagd` been started. When `clagd` is explicitly stopped, for example with the `service clagd stop` command, monitoring of `clagd` is also stopped.

The `jdoe` process checks two things to make sure the `clagd` service is operating properly:

- The result of the `service clagd status` command. If the command returns that `clagd` is running, or that `clagd` is not configured to run, then `jdoe` does nothing. If `service clagd status` returns that `clagd` is *not* running but was configured to run, `jdoe` will start the `clagd` service. This check is performed every 30 seconds. Due to the way the `jdoe` process implements this check, it may start the `clagd` process twice. This is harmless, since `clagd` checks to make sure another instance is not already running when it begins executing. This is indicated with a message in the `clagd` log file, `/var/log/clagd.log`.
- The modification time of the `/var/run/clagd.alive` file. As `clagd` runs, it periodically updates the modification time of the `/var/run/clagd.alive` file (by default, every 4 seconds). If `jdoe` notices that this file's modification time has not been updated within the last 4 minutes, it will assume `clagd` is alive, but hung, and will restart `clagd`. If `clagd` is not enabled to run, this check still occurs and `jdoe` will start `clagd`. But since `clagd` is not configured to run, nothing will happen except that a message is written to the `jdoe` log file that it tried to start `clagd`.

You can check the status of `clagd` monitoring by using the `jdoe summary` command:

```
cumulus@switch:~$ sudo jdoe summary
The jdoe daemon 5.4 uptime: 15m
...
Program 'clagd'                      Status ok
File 'clagd.alive'                    Waiting
...
```

## **MLAG Best Practices**

For MLAG to function properly, the dual-connected hosts' interfaces should be configured identically on the pair of peering switches. See the note above in the [Configuring MLAG \(see page 195\)](#) section.

## ***Understanding MTU in an MLAG Configuration***

Note that the [MTU \(see page 111\)](#) in MLAG traffic is determined by the bridge MTU. Bridge MTU is determined by the lowest MTU setting of an interface that is a member of the bridge. If an MTU other than the default of 1500 bytes is desired, you must configure the MTU on each physical interface and bond interface that are members of the MLAG bridges in the entire bridged domain.

For example, if an MTU of 9216 is desired through the MLAG domain in the example shown above:

On the leaf switches, [configure mtu 9216 \(see page 111\)](#) for each of following interfaces, since they are members of bridge *br0*: spine1-2, peerlink, host1, host2.

```
auto br0
iface br0
  bridge-vlan-aware yes
  bridge-ports spine1-2 peerlink host1 host2    <- List of bridge member
  interfaces
...
...
```

Likewise, to ensure the MTU 9216 path is respected through the spine switches above, also change the MTU setting for bridge *br* by configuring [mtu 9216](#) for each of the following members of bridge *br* on spine1 and spine2: uplinkA, peerlink, downlink1, downlink2.

```
auto br
iface br
  bridge-vlan-aware yes
  bridge-ports uplinkA peerlink downlink1 downlink2
...
...
```

## ***STP Interoperability with MLAG***

Cumulus Networks recommends that you always enable STP in your layer 2 network.

Further, with MLAG, Cumulus Networks recommends you enable BPDU guard on the host-facing bond interfaces. (For more information about BPDU guard, see [BPDU Guard and Bridge Assurance \(see page 135\)](#).)

## Debugging STP with MLAG

/var/log/daemon.log has mstpd logs.

Run `mstpctl debuglevel 3` to see MLAG-related logs in /var/log/daemon.log:

```
cumulus@switch:~$ sudo mstpctl showportdetail br peer-bond
br:peer-bond CIST info
  enabled      yes          role      Designated
  port id     8.008        state    forwarding
  .....
  bpdufilter port no
  clag ISL      yes          clag ISL Oper UP   yes
  clag role     primary      clag dual conn mac 0:0:0:0:0:0:
  0
  clag remote portID F.FFF           clag system mac 44:38:39:
ff:0:1
cumulus@switch:~$

cumulus@switch:~$ sudo mstpctl showportdetail br downlink-1
br:downlink-1 CIST info
  enabled      yes          role      Designated
  port id     8.006        state    forwarding
  .....
  bpdufilter port no
  clag ISL      no          clag ISL Oper UP   no
  clag role     primary      clag dual conn mac 0:0:0:0:3:
  11:1
  clag remote portID F.FFF           clag system mac 44:38:39:
ff:0:1
cumulus@switch:~$
```

## Best Practices for STP with MLAG

- The STP global configuration must be the same on both the switches.
- The STP configuration for dual-connected ports should be the same on both peer switches.
- Use `mstpctl` commands for all spanning tree configurations, including bridge priority, path cost and so forth. Do not use `brctl` commands for spanning tree, except for `brctl stp on/off`, as changes are not reflected to `mstpd` and can create conflicts.

## Troubleshooting MLAG

By default, when clagd is running, it logs its status to the `/var/log/clagd.log` file and syslog. Example log file output is below:

```
Jan 14 23:45:10 switch clagd[3704]: Beginning execution of clagd version
1.0.0
Jan 14 23:45:10 switch clagd[3704]: Invoked with: /usr/sbin/clagd --daemon
169.254.2.2 peer-bond.4000 44:38:39:ff:00:01 --priority 8192
Jan 14 23:45:11 switch clagd[3995]: Role is now secondary
Jan 14 23:45:31 switch clagd[3995]: Role is now primary
Jan 14 23:45:32 switch clagd[3995]: The peer switch is active.
Jan 14 23:45:35 switch clagd[3995]: downlink-1 is now dual connected.
```

## Caveats and Errata

If both the backup and peer connectivity are lost within a 30-second window, the switch in the secondary role misinterprets the event sequence, believing the peer switch is down, so it takes over as the primary.

## Configuration Files

- `/etc/network/interfaces`

## LACP Bypass

On Cumulus Linux, *LACP Bypass* is a feature that allows a [bond](#) (see page 158) configured in 802.3ad mode to become active and forward traffic even when there is no LACP partner. A typical use case for this feature is to enable a host, without the capability to run LACP, to PXE boot while connected to a switch on a bond configured in 802.3ad mode. Once the pre-boot process finishes and the host is capable of running LACP, the normal 802.3ad link aggregation operation takes over.

## Contents

(Click to expand)

- [Contents \(see page 209\)](#)
- [Understanding LACP Bypass Modes \(see page 210\)](#)
  - [LACP Bypass Timeout \(see page 210\)](#)
  - [LACP Bypass and MLAG Deployments \(see page 210\)](#)
  - [Configuring LACP Bypass \(see page 210\)](#)
  - [Configuration Examples \(see page 211\)](#)
    - [Default Configuration with Priority Mode and Optional Timeout Period \(see page 211\)](#)
    - [All-active Mode Configuration with Multiple Simultaneous Active Interfaces \(see page 212\)](#)

## Understanding LACP Bypass Modes

When a bond has multiple slave interfaces, you can control which of them should go into LACP bypass using one of two modes:

- *Priority mode*: This is the default mode. On a switch, if a bond has multiple slave interfaces, you can configure a bypass priority value (default is 0) for each interface in the bond; the one with higher numerical priority value wins. A string comparison of the interface names serves as a tiebreaker in case the priority values are equal; the string with the lower ASCII values wins. Note that the priority value is significant within a switch; there is no coordination between two switches in an [MLAG](#) (see page 191) peering relationship.
- *All-active mode*: In this mode, each bond slave interface operates as an active link while the bond is in bypass mode. This mode is useful during PXE boot of a server with multiple NICs, when the user cannot determine beforehand which port needs to be active. By default, all-active mode is disabled.



All-active mode is not supported on bonds that are not specified as bridge ports on the switch.



STP does not run on the individual bond slave interfaces, when the LACP bond is in all-active mode. Therefore, only use all-active mode on host-facing LACP bonds. Cumulus Networks highly recommends you configure [STP BPDU guard](#) along with all-active mode.

## LACP Bypass Timeout

As a safeguard, you can configure a timeout period to limit the duration in which bypass is enabled. The timeout period works with both modes. The valid range of timeout period is 0 to 900 seconds; the default is 0 seconds, which indicates no timeout. If no LACP partner is detected before the timeout period expires, the bond becomes inactive and stops forwarding traffic. The timer is restarted when all slave interfaces are enabled; which can be achieved by setting the interface down and then up. At any point in time, receiving LACP PDU on any slave interface aborts the bypass, and normal LACP protocol negotiation takes over. Enabling or disabling bypass during LACP exchange does not affect link aggregation.

## LACP Bypass and MLAG Deployments

In an [MLAG deployment](#) (see page 191) where bond slaves of a host are connected to two switches and the bond is in **priority mode**, the bypass priority is determined using the MLAG switch role. The bond on the switch with the primary role has a **higher** bypass priority than the bond on the switch with the secondary role. When multiple slave interfaces of a bond are connected to each switch, the slave with the highest priority on the primary MLAG switch will be the active interface. All other slaves on the same device will not be active during bypass mode.

When a dual-connected (MLAG) bond is in **all-active mode**, all the slaves of bond are active on both the primary and secondary MLAG nodes.

## Configuring LACP Bypass

You configure LACP bypass in the `/etc/network/interfaces` file.

To enable LACP bypass on the host-facing bond, under the bond interface stanza, set `bond-lacp-bypass-allow` to 1. Then optionally configure one of the following:

- To configure **priority mode**, which is the *default* mode, set `bond-lacp-bypass-priority` to a value, with the priority values for each slave interface. The default priority value is 0.
- To configure **all-active mode** for multiple active interfaces, set `bond-lacp-bypass-all-active` to 1. This enables all interfaces to pass traffic (become active) until the server can form an LACP bond.

**(Optional):** To configure a timeout period for either mode, set `bond-lacp-bypass-period` to a valid value (0-900); however, it is recommended to not configure this, and use the default value of 0.

## Configuration Examples

### Default Configuration with Priority Mode and Optional Timeout Period

The following configuration shows LACP bypass enabled in the default priority mode, with a timeout period set. Here there are two slave interfaces, and `swp2` will be preferred as the active bypass interface:

```
auto bond0
iface bond0
    bond-mode 802.3ad
    bond-lacp-rate 1
    bond-min-links 1
    bond-lacp-bypass-allow 1
    bond-slaves swp4 swp5
    bond-lacp-bypass-period 300
    bond-lacp-bypass-priority swp4=2 swp5=1
```

The following command shows that `swp4` bypass timeout has expired and the bond is operationally down:

```
cumulus@switch$ ip link show bond0
7: bond0: <NO-CARRIER,BROADCAST,MULTICAST,MASTER,UP> mtu 1500 qdisc noqueue
state DOWN mode DEFAULT
    link/ether 00:02:00:00:00:02 brd ff:ff:ff:ff:ff:ff
cumulus@switch$ cat /proc/net/bonding/bond0
Ethernet Channel Bonding Driver: v3.7.1 (April 27, 2011)
Bonding Mode: IEEE 802.3ad Dynamic link aggregation
Transmit Hash Policy: layer2 (0)
MII Status: up
MII Polling Interval (ms): 0
Up Delay (ms): 0
Down Delay (ms): 0
802.3ad info
LACP rate: fast
Min links: 1
Aggregator selection policy (ad_select): stable
```

```
System Identification: 65535 00:02:00:00:00:02
Active Aggregator Info:
    Aggregator ID: 1
    Number of ports: 1
    Actor Key: 33
    Partner Key: 1
    Partner Mac Address: 00:00:00:00:00:00
Fall back Info:
    Allowed: 1
    Timeout: 300
Slave Interface: swp4
MII Status: up
Speed: 10000 Mbps
Duplex: full
Link Failure Count: 0
Permanent HW addr: 00:02:00:00:00:02
Aggregator ID: 1
LACP bypass priority: 2
LACP bypass: expired
Slave queue ID: 0
Slave Interface: swp5
MII Status: up
Speed: 10000 Mbps
Duplex: full
Link Failure Count: 0
Permanent HW addr: 00:02:00:00:00:01
Aggregator ID: 2
Bypass priority: 1
Slave queue ID: 0
```

## All-active Mode Configuration with Multiple Simultaneous Active Interfaces

The following configuration shows LACP bypass enabled for multiple active interfaces (all-active mode) with a bridge in VLAN-aware mode (see page 182):

```
auto bond1
iface bond1 inet static
    bond-slaves swp3 swp4
    bond-mode 802.3ad
    bond-lacp-rate 1
    bond-min-links 1
    bond-lacp-bypass-allow 1
    bond-lacp-bypass-all-active 1
```

```

auto br0
iface br0 inet static
    bridge-vlan-aware yes
    bridge-ports bond1 bond2 bond3 bond4 peer5
    bridge-stp on
    bridge-vids 100-105
    mstpctl-bpduguard bond1=yes

cumulus@switch:~$ ip link show bond1
58: bond1: <BROADCAST,MULTICAST,MASTER,UP,LOWER_UP> mtu 1500 qdisc noqueue
    master br0 state UP mode DORMANT
        link/ether 44:38:39:00:38:44 brd ff:ff:ff:ff:ff:ff
cumulus@switch:~$ ip link show swp3
5: swp3: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
    master bond1 state UP mode DEFAULT qlen 500
        link/ether 44:38:39:00:38:44 brd ff:ff:ff:ff:ff:ff
cumulus@switch:~$ ip link show swp4
6: swp4: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
    master bond1 state UP mode DEFAULT qlen 500
        link/ether 44:38:39:00:38:44 brd ff:ff:ff:ff:ff:ff

cumulus@switch:~$ cat /proc/net/bonding/bond1
Ethernet Channel Bonding Driver: v3.7.1 (April 27, 2011)

Bonding Mode: IEEE 802.3ad Dynamic link aggregation
Transmit Hash Policy: layer3+4 (1)
MII Status: up
MII Polling Interval (ms): 100
Up Delay (ms): 0
Down Delay (ms): 0

802.3ad info
LACP rate: fast
Min links: 1
Aggregator selection policy (ad_select): stable
System Identification: 65535 00:00:00:aa:bb:01
Active Aggregator Info:
    Aggregator ID: 1
    Number of ports: 1
    Actor Key: 33
    Partner Key: 33
    Partner Mac Address: 00:02:00:00:00:05
LACP Bypass Info:

```

```
    Allowed: 1
    Timeout: 0
    All-active: 1

Slave Interface: swp3
MII Status: up
Speed: 10000 Mbps
Duplex: full
Link Failure Count: 1
Permanent HW addr: 44:38:39:00:38:44
Aggregator ID: 1
LACP bypass priority: 0
LACP bypass: on
Slave queue ID: 0

Slave Interface: swp4
MII Status: up
Speed: 10000 Mbps
Duplex: full
Link Failure Count: 1
Permanent HW addr: 44:38:39:00:38:45
Aggregator ID: 2
LACP bypass priority: 0
LACP bypass: on
Slave queue ID: 0
```

The following configuration shows LACP bypass enabled for multiple active interfaces (all-active mode) with a bridge in [traditional bridge mode \(see page 162\)](#):

```
auto bond1
iface bond1 inet static
    bond-slaves swp3 swp4
    bond-mode 802.3ad
    bond-lacp-rate 1
    bond-min-links 1
    bond-lacp-bypass-allow 1
    bond-lacp-bypass-all-active 1

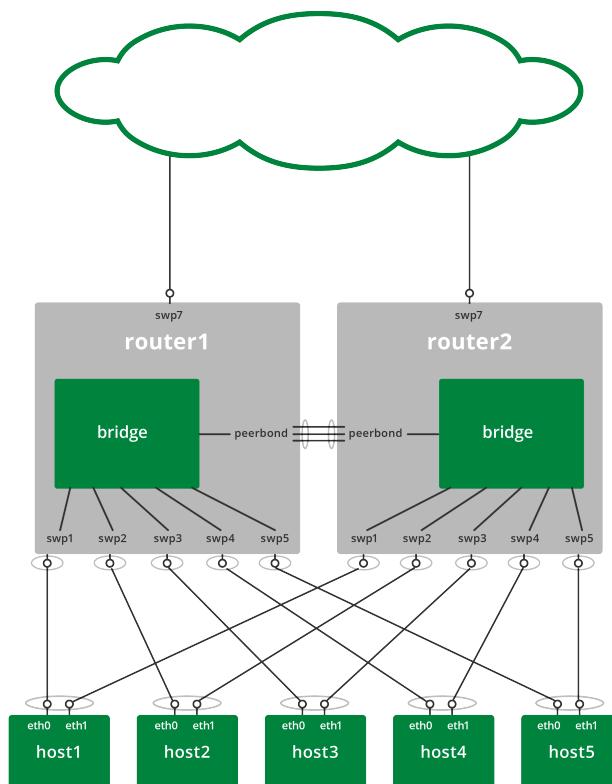
auto br0
iface br0 inet static
    bridge-ports bond1 bond2 bond3 bond4 peer5
    bridge-stp on
    mstpctl-bpduguard bond1=yes
```

## Virtual Router Redundancy - VRR

VRR provides virtualized router redundancy in network configurations, which enables the hosts to communicate with any redundant router without:

- Needing to be reconfigured
- Having to run dynamic router protocols
- Having to run router redundancy protocols

A basic VRR-enabled network configuration is shown below. The network consists of several hosts, two routers running Cumulus Linux and configured with [MLAG \(see page 191\)](#), and the rest of the network:



An actual implementation will have many more server hosts and network connections than are shown here. But this basic configuration provides a complete description of the important aspects of the VRR setup.

## Contents

(Click to expand)

- [Contents \(see page 215\)](#)
- [Configuring the Network \(see page 216\)](#)
  - [Reserved MAC Address Range \(see page 217\)](#)

- Configuring the Hosts (see page 217)
- Configuring the Routers (see page 217)
- Other Network Connections (see page 218)
- Handling ARP Requests (see page 218)
- Monitoring Peer Links and Uplinks (see page 218)
- Using ifplugd (see page 218)
- Notes (see page 220)

## Configuring the Network

Configuring this network is fairly straightforward. First create the bridge subinterface, then create the secondary address for the virtual router. Configure each router with a bridge; edit each router's `/etc/network/interfaces` file and add a configuration similar to the following:

```
auto bridge.500
iface bridge.500
    address 192.168.0.252/24
    address-virtual 00:00:5e:00:01:01 192.168.0.254/24
```



Notice the simpler configuration of the bridge with `ifupdown2`. For more information, see [Configuring and Managing Network Interfaces \(see page 94\)](#).

You should always use `ifupdown2` to configure VRR, because it ensures correct ordering when bringing up the virtual and physical interfaces and it works best with [VLAN-aware bridges \(see page 182\)](#).

If you are using the `traditional mode` bridge driver, the configuration would look like this:

```
auto bridge500
iface bridge500
    address 192.168.0.252/24
    address-virtual 00:00:5e:00:01:01 192.168.0.254/24
    bridge-ports bond1.500 bond2.500 bond3.500
```

The IP address assigned to the bridge is the unique address for the bridge. The parameters of this configuration are:

- `bridge.500`: 500 represents a VLAN subinterface of the bridge, sometimes called a switched virtual interface, or SVI.
- `192.168.0.252/24`: The unique IP address assigned to this bridge. It is unique because, unlike the 192.168.0.254 address, it is assigned only to this bridge, not the bridge on the other router.
- `00:00:5e:00:01:01`: The MAC address of the virtual router. This must be the same on all virtual routers. Cumulus Linux has a reserved range for VRR MAC addresses. See below for details.

- `192.168.0.254/24`: The IP address of the virtual router, including the routing prefix. This must be the same on all the virtual routers and must match the default gateway address configured on the servers as well as the size of the subnet.
- `address-virtual`: This keyword enables and configures VRR.

The above bridge configuration enables VRR by creating a *MAC VLAN interface* on the SVI. This MAC VLAN interface is:

- Named bridge-500-v0, which is the name of the SVI with dots changed to dashes and "-v0" appended to the end.
- Assigned a MAC address of `00:00:5e:00:01:01`.
- Assigned an IP address of `192.168.0.254`.

## Reserved MAC Address Range

In order to prevent MAC address conflicts with other interfaces in the same bridged network, Cumulus Networks has **reserved a range of MAC addresses** specifically to use with VRR. This range of MAC addresses is `00:00:5E:00:01:00` to `00:00:5E:00:01:ff`.

You may notice that this is the same range reserved for VRRP, since VRR serves a similar function. Cumulus Networks recommends you use this range of MAC addresses when configuring VRR.

## Configuring the Hosts

Each host should have two network interfaces. The routers configure the interfaces as bonds running LACP; the hosts should also configure its two interfaces using teaming, port aggregation, port group, or EtherChannel running LACP. Configure the hosts, either statically or via DHCP, with a gateway address that is the IP address of the virtual router; this default gateway address never changes.

Configure the links between the hosts and the routers in *active-active* mode for First Hop Redundancy Protocol.



If you are configuring VRR without MLAG (see page 191), use *active-standby* mode instead.

## Configuring the Routers

The routers implement the layer 2 network interconnecting the hosts, as well as the redundant routers. If you are using **MLAG** (see page 191), configure each router with a bridge interface, named *bridge* in our example, with these different types of interfaces:

- One bond interface to each host (`swp1-swp5` in the image above).
- One or more interfaces to each peer router (`peerbond` in the image above). Multiple inter-peer links are typically bonded interfaces in order to accommodate higher bandwidth between the routers and to offer link redundancy.



If you are not using MLAG, then the bridge should have one switch port interface to each host instead of a bond.

## Other Network Connections

Other interfaces on the router can connect to other subnets and are accessed through layer 3 forwarding (swp7 in the image above).

## Handling ARP Requests

The entire purpose of this configuration is to have all the redundant routers respond to ARP requests from hosts for the virtual router IP address (192.168.0.254 in the example above) with the virtual router MAC address (00:00:5e:00:01:01 in the example above). All of the routers should respond in an identical manner, but if one router fails, the other redundant routers will continue to respond in an identical manner, leaving the hosts with the impression that nothing has changed.

Since the bridges in each of the redundant routers are connected, they will each receive and reply to ARP requests for the virtual router IP address. Each ARP request made by a host will receive multiple replies (typically two). But these replies will be identical and so the host that receives these replies will not get confused over which response is "correct" and will either ignore replies after the first, or accept them and overwrite the previous reply with identical information.

## Monitoring Peer Links and Uplinks

When an uplink on a switch in active-active mode goes down, the peer link may get congested. When this occurs, you should monitor the uplink and shut down all host-facing ports using `ifplugd` (or another script).

When the peer link goes down in a MLAG environment, one of the switches becomes secondary and all host-facing dual-connected bonds go down. The host side bond sees two different system MAC addresses, so the link to primary is active on host. If any traffic from outside this environment goes to the secondary MLAG switch, traffic will be black-holed. To avoid this, shut down all the uplinks when the peer link goes down using `ifplugd`.

## Using `ifplugd`

`ifplugd` is a link state monitoring daemon that can execute user-specified scripts on link transitions (not admin-triggered transitions, but transitions when a cable is plugged in or removed).

Run the following commands to install the `ifplugd` service:

```
cumulus@switch:$ sudo apt-get update
cumulus@switch:$ sudo apt-get install ifplugd
```

Next, configure `ifplugd`. The example below indicates that when the `peerbond` goes down in a MLAG environment, `ifplugd` brings down all the uplinks. Run the following `ifplugd` script on both the primary and secondary MLAG (see page 191) switches.

To configure `ifplugd`, modify `/etc/default/ifplugd` and add the appropriate `peerbond` interface name. `/etc/default/ifplugd` will look like this:

```
INTERFACES="peerbond"
HOTPLUG_INTERFACES=" "
```

```
ARGS="-q -f -u0 -d1 -w -I"
SUSPEND_ACTION="stop"
```

Next, modify the `/etc/ifplugd/action.d/ifupdown` script.

```
#!/bin/sh
set -e
case "$2" in
up)
    clagrole=$(clagctl | grep "Our Priority" | awk '{print $8}')
    if [ "$clagrole" = "secondary" ]
    then
        #List all the interfaces below to bring up when clag
        peerbond comes up.
        for interface in swp1 bond1 bond3 bond4
        do
            echo "bringing up : $interface"
            ip link set $interface up
        done
    fi
;;
down)
    clagrole=$(clagctl | grep "Our Priority" | awk '{print $8}')
    if [ "$clagrole" = "secondary" ]
    then
        #List all the interfaces below to bring down when clag
        peerbond goes down.
        for interface in swp1 bond1 bond3 bond4
        do
            echo "bringing down : $interface"
            ip link set $interface down
        done
    fi
;;
esac
```

Finally, restart `ifplugd` for your changes to take effect:

```
cumulus@switch:$ sudo service ifplugd restart
```

## Notes

- The default shell is `/bin/sh`, which is `dash` and not `bash`. This makes for faster execution of the script since `dash` is small and quick, but consequently less featureful than `bash`. For example, it doesn't handle multiple uplinks.

## Network Virtualization

Cumulus Linux supports these forms of [network virtualization](#):

VXLAN (Virtual Extensible LAN), is a standard overlay protocol that abstracts logical virtual networks from the physical network underneath. You can deploy simple and scalable layer 3 Clos architectures while extending layer 2 segments over that layer 3 network.

VXLAN uses a VLAN-like encapsulation technique to encapsulate MAC-based layer 2 Ethernet frames within layer 3 UDP packets. Each virtual network is a VXLAN logical L2 segment. VXLAN scales to 16 million segments – a 24-bit VXLAN network identifier (VNI ID) in the VXLAN header – for multi-tenancy.

Hosts on a given virtual network are joined together through an overlay protocol that initiates and terminates tunnels at the edge of the multi-tenant network, typically the hypervisor vSwitch or top of rack. These edge points are the VXLAN tunnel end points (VTEP).

Cumulus Linux can initiate and terminate VTEPs in hardware and supports wire-rate VXLAN with Trident II platforms. VXLAN provides an efficient hashing scheme across IP fabric during the encapsulation process; the source UDP port is unique, with the hash based on L2-L4 information from the original frame. The UDP destination port is the standard port 4789.

Cumulus Linux includes the native Linux VXLAN kernel support and integrates with controller-based overlay solutions like VMware NSX and Midokura MidoNet.

VXLAN is supported only on switches in the [Cumulus Linux HCL](#) using Trident II chipsets.



VXLAN encapsulation over layer 3 subinterfaces is not supported. Therefore, VXLAN uplinks should be only configured as layer 3 interfaces without any subinterfaces.

## Commands

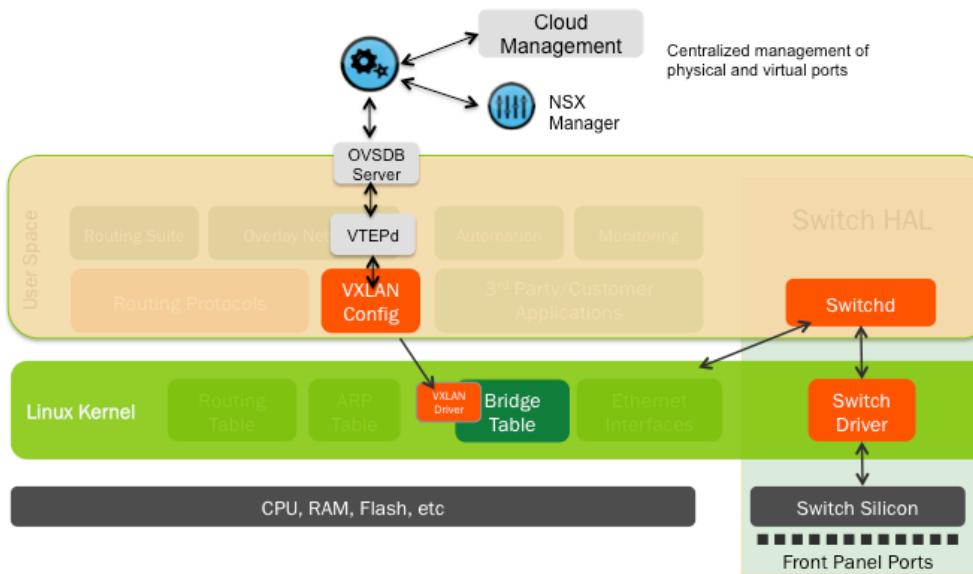
- `brctl`
- `bridge fdb`
- `ip link`
- `ovs-pki`
- `ovsdb-client`
- `vtep-ctl`

## Useful Links

- [VXLAN IETF draft](#)
- [ovsdb-server](#)

## Integrating with VMware NSX

Switches running Cumulus Linux can integrate with VMware NSX to act as VTEP gateways. The VMware NSX controller provides consistent provisioning across virtual and physical server infrastructures.



## Contents

(Click to expand)

- [Contents \(see page 221\)](#)
- [Getting Started \(see page 221\)](#)
  - [Caveats and Errata \(see page 222\)](#)
- [Bootstrapping the NSX Integration \(see page 222\)](#)
  - [Enabling the openvswitch-vtep Package \(see page 222\)](#)
  - [Using the Bootstrapping Script \(see page 223\)](#)
  - [Manually Bootstrapping the NSX Integration \(see page 224\)](#)
  - [Generating the Credentials Certificate \(see page 224\)](#)
  - [Configuring the Switch as a VTEP Gateway \(see page 225\)](#)
- [Configuring the Transport Layer \(see page 228\)](#)
- [Configuring the Logical Layer \(see page 229\)](#)
  - [Defining Logical Switches \(see page 229\)](#)
  - [Defining Logical Switch Ports \(see page 231\)](#)
- [Verifying the VXLAN Configuration \(see page 233\)](#)
- [Persistent VXLAN Configuration in NSX \(see page 234\)](#)
- [Troubleshooting VXLANs in NSX \(see page 234\)](#)

## Getting Started

Before you integrate VXLANs with NSX, make sure you have the following components:

- A switch (L2 gateway) with a Trident II chipset running Cumulus Linux 2.0 or later;
- OVSDB server (ovsdb-server), included in Cumulus Linux 2.0 and later
- VTEPd (ovs-vtep), included in Cumulus Linux 2.0 and later

Integrating a VXLAN with NSX involves:

- Bootstrapping the NSX Integration
- Configuring the Transport Layer
- Configuring the Logical Layer
- Verifying the VXLAN Configuration

Once you finish the integration, you can make the configuration persistent across upgrades (see [Persistent VXLAN Configuration in NSX](#) (see page 234) below).

## **Caveats and Errata**

- The switch with the sourcing VTEP must connect to a router.
- There is no support for VXLAN routing in the Trident II chip; use a loopback interface or external router.
- Do not use 0 or 16777215 as the VNI ID, as they are reserved values under Cumulus Linux.
- For more information about NSX, see the VMware NSX User Guide, version 4.0.0 or later.

## **Bootstrapping the NSX Integration**

Before you start configuring the gateway service and logical switches and ports that comprise the VXLAN, you need to complete some steps to bootstrap the process. You need to do the bootstrapping just once, before you begin the integration.

## **Enabling the `openvswitch-vtep` Package**

Before you start bootstrapping the integration, you need to enable the `openvswitch-vtep` package, as it is disabled by default in Cumulus Linux.

1. In `/etc/default/openvswitch-vtep`, change the `START` option from `no` to `yes`:

```
cumulus@switch$ cat /etc/default/openvswitch-vtep
# This is a POSIX shell fragment          -*- sh -*-

# Start openvswitch at boot ? yes/no
START=yes

# FORCE_COREFILES: If 'yes' then core files will be enabled.
# FORCE_COREFILES=yes

# BRCOMPAT: If 'yes' and the openvswitch-brcompat package is
#           installed, then
#           Linux bridge compatibility will be enabled.
# BRCOMPAT=no
```

2. Start the daemon:

```
cumulus@switch$ sudo service openvswitch-vtep start
```

Make sure to include this file in your persistent configuration (see [Persistent VXLAN Configuration in NSX \(see page 234\)](#) below) so it's available after you upgrade Cumulus Linux.

## ***Using the Bootstrapping Script***

A script is available so you can do the bootstrapping automatically. For information, read `man vtep-bootstrap`. The output of the script is displayed here:

```
cumulus@vtep7: ~
cumulus@vtep7$ sudo vtep-bootstrap --credentials-path /var/lib/openvswitch vtep7 192.168.100.17 172.1
6.20.157 192.168.100.157
Executed:
  create certificate on a switch, to be used for authentication with controller
  ().

Executed:
  sign certificate
  (vtep7-req.pem      Fri Jan 17 18:04:33 UTC 2014
   fingerprint 6f443eb8445317b545d8564c2ae9638ea0a9184a).

Executed:
  define physical switch
  ().

Executed:
  define NSX controller IP address in OVSDB
  ().

Executed:
  define local tunnel IP address on the switch
  ().

Executed:
  define management IP address on the switch
  ().

Executed:
  restart a service
  (Killing ovs-vtep (4973).
Killing ovsdb-server (4969).
Starting ovsdb-server.
Starting ovs-vtep.).
cumulus@vtep7$ 
cumulus@vtep7$ ps xa | grep ovsdb-server
 5184 pts/0    S+    0:00 ovsdb-server: monitoring pid 5185 (healthy)

5185 ?      S<    0:00 ovsdb-server /var/lib/openvswitch/conf.db -vANY:CONSOLE;EMER -vANY:SYSLOG:ERR -vANY:FILE:INFO --remote=unix:/var/run/openvswitch/db.sock --remote=db:Global_managers --remote=ptcp:6633: --private-key=/var/lib/openvswitch/vtep7-privkey.pem --certificate=/var/lib/openvswitch/vt
ep7-cert.pem --bootstrap-ca-cert=/var/lib/openvswitch/controller.cacert --no-chdir --log-file=/var/lo
g/openvswitch/ovsdb-server.log --pidfile=/var/run/openvswitch/ovsdb-server.pid --detach --monitor
 5436 pts/0    S+    0:00 grep ovsdb-server
cumulus@vtep7$
```

In the above example, the following information was passed to the `vtep-bootstrap` script:

- `--credentials-path /var/lib/openvswitch`: Is the path to where the certificate and key pairs for authenticating with the NSX controller are stored.
- `vtep7`: is the ID for the VTEP.
- `192.168.100.17`: is the IP address of the NSX controller.

- 172.16.20.157: is the datapath IP address of the VTEP.
- 192.168.100.157: is the IP address of the management interface on the switch.

These IP addresses will be used throughout the rest of the examples below.

## **Manually Bootstrapping the NSX Integration**

If you don't use the script, then you must:

- Initialize the OVS database instance
- Generate a certificate and key pair for authentication by NSX
- Configure a switch as a VTEP gateway

These steps are described next.

## **Generating the Credentials Certificate**

First, in Cumulus Linux, you must generate a certificate that the NSX controller uses for authentication.

1. In a terminal session connected to the switch, run the following commands:

```
cumulus@switch:~$ sudo ovs-pki init
Creating controllerca...
Creating switchca...
cumulus@switch:~$ sudo ovs-pki req+sign cumulus

cumulus-req.pem Wed Oct 23 05:32:49 UTC 2013
    fingerprint b587c9fe36f09fb371750ab50c430485d33a174a
cumulus@switch:~$
cumulus@switch:~$ ls -l
total 12
-rw-r--r-- 1 root root 4028 Oct 23 05:32 cumulus-cert.pem
-rw----- 1 root root 1679 Oct 23 05:32 cumulus-privkey.pem
-rw-r--r-- 1 root root 3585 Oct 23 05:32 cumulus-req.pem
```

2. In /usr/share/openvswitch/scripts/ovs-ctl-vtep, make sure the lines containing **private-key**, **certificate** and **bootstrap-ca-cert** point to the correct files; **bootstrap-ca-cert** is obtained dynamically the first time the switch talks to the controller:

```
# Start ovsdb-server.
set ovsdb-server "$DB_FILE"
set "$@" -vANY:CONSOLE:EMER -vANY:SYSLOG:ERR -vANY:FILE:INFO
set "$@" --remote=unix:"$DB_SOCK"
set "$@" --remote=db:Global,managers
set "$@" --remote=ptcp:6633:$LOCALIP
set "$@" --private-key=/root/cumulus-privkey.pem
```

```
set "$@" --certificate=/root/cumulus-cert.pem  
set "$@" --bootstrap-ca-cert=/root/controller.cacert
```

If files have been moved or regenerated, restart the OVSDB server and vtepd:

```
cumulus@switch:~$ sudo service openvswitch-vtep restart
```

3. Define the NSX controller cluster IP address in OVSDB. This causes the OVSDB server to start contacting the NSX controller:

```
cumulus@switch:~$ sudo vtep-ctl set-manager ssl:192.168.100.17:6632
```

4. Define the local IP address on the VTEP for VXLAN tunnel termination. First, find the physical switch name as recorded in OVSDB:

```
cumulus@switch:~$ sudo vtep-ctl list-ps  
vtep7
```

Then set the tunnel source IP address of the VTEP. This is the datapath address of the VTEP, which is typically an address on a loopback interface on the switch that is reachable from the underlying L3 network:

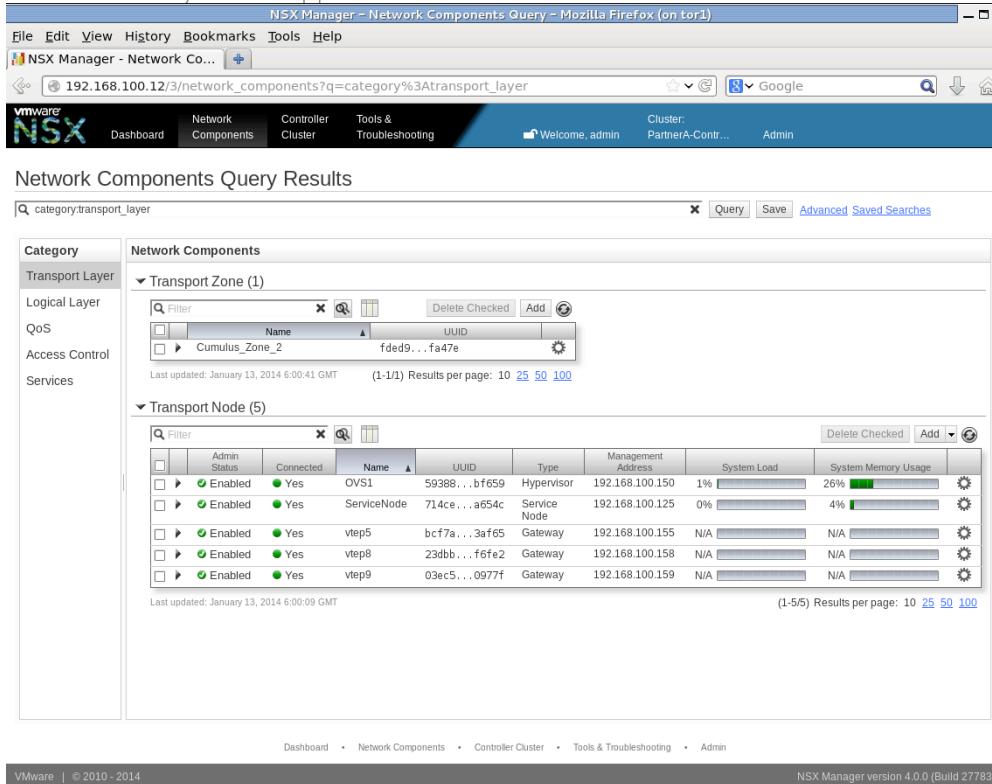
```
cumulus@switch:~$ sudo vtep-ctl set Physical_Switch vtep7  
tunnel_ips=172.16.20.157
```

Once you finish generating the certificate, keep the terminal session active, as you need to paste the certificate into NSX Manager when you configure the VTEP gateway.

## ***Configuring the Switch as a VTEP Gateway***

After you create a certificate, connect to NSX Manager in a browser to configure a Cumulus Linux switch as a VTEP gateway. In this example, the IP address of the NSX manager is 192.168.100.12.

- In NSX Manager, add a new gateway. Click the **Network Components** tab, then the **Transport Layer** category. Under **Transport Node**, click **Add**, then select **Manually Enter All Fields**. The Create Gateway wizard appears.



The screenshot shows the NSX Manager interface with the following details:

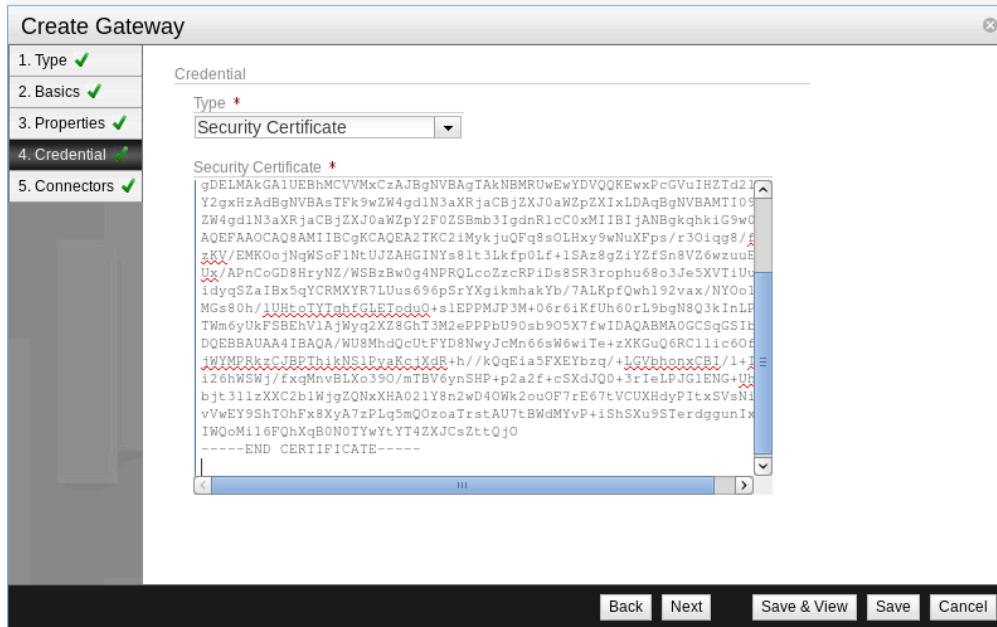
- Network Components Query Results**
- Category:** Transport Layer
- Transport Zone (1):**
  - Filter: Name (Cumulus\_Zone\_2), UUID (fded9...fa47e)
  - Last updated: January 13, 2014 6:00:41 GMT
  - (1-1) Results per page: 10, 25, 50, 100
- Transport Node (5):**
  - Filter: Admin Status (Enabled), Connected (Yes), Name (OVS1, ServiceNode, vtep5, vtep8, vtep9), UUID (5938..., 714c..., bcf7..., 23db..., 03ec...), Type (Hypervisor, Service Node, Gateway, Gateway, Gateway), Management Address (192.168.100.150, 192.168.100.125, 192.168.100.155, 192.168.100.158, 192.168.100.159), System Load (1%, 0%, N/A, N/A, N/A), System Memory Usage (26%, 4%, N/A, N/A, N/A).
  - Last updated: January 13, 2014 6:00:09 GMT
  - (1-5) Results per page: 10, 25, 50, 100

Bottom navigation: Dashboard, Network Components, Controller Cluster, Tools & Troubleshooting, Admin. Footer: VMware | © 2010 - 2014, NSX Manager version 4.0.0 (Build 27783).

- In the Create Gateway dialog, select **Gateway** for the **Transport Node Type**, then click **Next**.
- In the **Display Name** field, give the gateway a name, then click **Next**.
- Enable the VTEP service. Select the **VTEP Enabled** checkbox, then click **Next**.
- From the terminal session connected to the switch where you generated the certificate, copy the certificate and paste it into the **Security Certificate** text field. Copy only the bottom portion, including the `BEGIN CERTIFICATE` and `END CERTIFICATE` lines. For example, copy all the highlighted text in the terminal:

```
ubuntu@tor1: ~
87:3f:ea;76:8b:67:fe:71:25:dd:25:0d:3e:de:b2:1e:2c:f2;
46:94:43:46:f9:48:43:6e:3b:77:96:5e:d7:5c:2d:9b:95:68;
e0:65:03:71:5c:70:34:da:56:3c:9f:6e:03:e0:f5:a4:da:8b;
8e:17:ba:c4:eb:bb:59:09:45:c7:77:23:c8:b7:14:95:b0:d8;
ba:bd:5c:04:63:4d:a1:4c:e8:45:c7:c5:f2:03:bc:cf:2e:ae;
66:40:ec:e8:69:3a:ec:b4:05:3b:b4:15:9d:31:b8:cf:fa:24;
a1:49:7b:bd:49:37:ab:76:08:2e:9c:8c:44:21:64:28:32:2d;
7a:19:08:57:a8:1d:0d:0d:36:30:02:db:13:e1:95:c9:0a:c6;
6d:b5:08:ce
-----BEGIN CERTIFICATE-----
MIIBdCCA18CAQYwDQYJKoZIhvNA0gEBOAwgExCzA1BgNVBAYTA1VTM0swCQD
M0Q1EwJJOtEVMBIGA1UECHMT3B1b1B2U3dpdG0NMRExDwYDVQQLEuhzd210V2hj
Y1E7MDkgA1UEAxM1ZT1HN3aXRjaGhNIEBNEN1cnRpZnlyXR11CsgyIDEzER1
YgAuMyAxNzowOBYNCkuUhNMJThxkjIzTcxMzAxJhJhMTM0:MjIzMtCxJxzAxJjCB
gDELMAgCA1UEBhMCVWhCcRAjBqgWVAqgTkkBNBRUwEYDVQKExwPcGVwIHZTd210
Y2gxHzAdBgNVBAsTFK9uZl44gdIN3aXkjaC8jZXJ0alZpZXIxLxDqBgwVBANT109w
Zl44gdIN3aXkjaC8jZXJ0alZpZl2RmB31gdR1cOwM1B1JNBqkphkIGw0B
9QEFA0DCAQ8M1IBCaQkCAQEA2TKC2iMyjQFq8s0LHxy3wNuXfps/r30iq98/FP
zKV/EMK0oJNdsf1htUJZAHGINy81t3LkfplF+1Sa28g21YzFSn8VZ6zuuE8
Jx/APnCoGdBhNz/WSBzBw0g4NPwQlco2zcRP1DsS93r0phu683jeXVtIUuZ
jdqg52a1bx5qYCRMXYR7LUsos96p5YQgikmhakYb/7HLkpQ0uh192vax/YNy0li
MG80h1UHt0TYTghGLETodu0+sLEPPNJP3M+06:61kFuhs6:L9bgwB03kInLPN
Tlw6UkFSEhV1AjJyqg2XZ8GhT3M2ePPBbJ9Osbs05X7fuID9QABMAGCgSg51b3
DOEBBaUa44IBa09/wUJ8MhddeUfYD8NwJcm6s4lwjTe+zXGu06RC11:ic60fq
jUyYMPRkzCJBPTth1NSIPyaKcjUxRdr//hKQeia5FXEYbzq/+LGVbhonxB1i/1+IC
i26hWsuJ/fxmnvBLko390/WTBw6unSHp+2a2f+c5xdJ00+3r1eLPJG1ENG+UnD
bjc311zXc2b1wJg2UNxHA021Y8w2u040lk2uoF7rB7t+VCUxHdyItxvSvN16
vUvEY9ShTOHfx3uA7zPLq5m02ozaTrstAU7tBldMyvP+ShSXu95STerDggunIxE
lWQoM16FQhkaBONOTyYtYt4ZJC0s2ztQj0
-----END CERTIFICATE-----
root@vtep-1:~#
```

And paste it into NSX Manager:



Then click **Next**.

6. In the Connectors dialog, click **Add Connector** to add a transport connector. This defines the tunnel endpoint that terminates the VXLAN tunnel and connects NSX to the physical gateway. You must choose a tunnel **Transport Type** of **VXLAN**. Choose an existing transport zone for the connector, or click **Create** to create a new transport zone.
7. Define the connector's IP address (that is, the underlay IP address on the switch for tunnel termination).
8. Click **OK** to save the connector, then click **Save** to save the gateway.

Once communication is established between the switch and the controller, a `controller.cacert` file will be downloaded onto the switch.

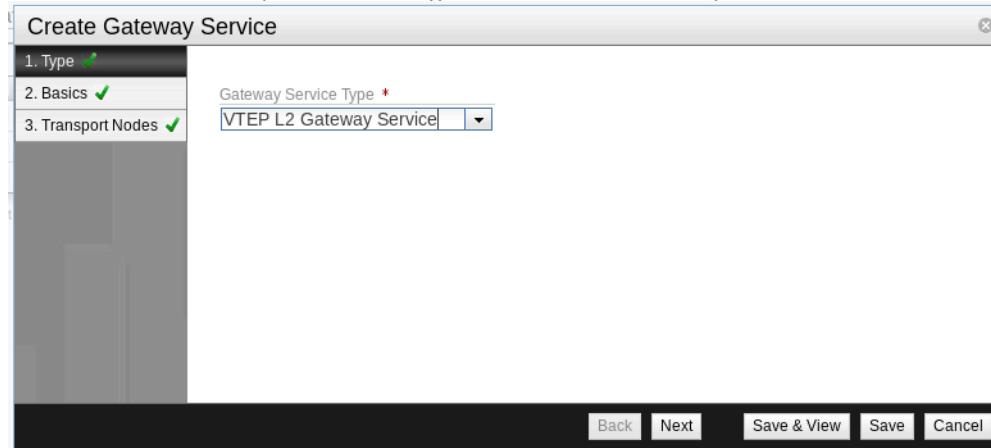
Verify the controller and switch handshake is successful. In a terminal connected to the switch, run this command:

```
cumulus@switch:~$ sudo ovsdb-client dump -f list | grep -A 7 "Manager"
Manager table
_uuid : 505f32af-9acb-4182-a315-022e405aa479
inactivity_probe : 30000
is_connected : true
max_backoff : []
other_config : {}
status : {sec_since_connect="18223", sec_since_disconnect="18225", state=ACTIVE}
target : "ssl:192.168.100.17:6632"
```

## Configuring the Transport Layer

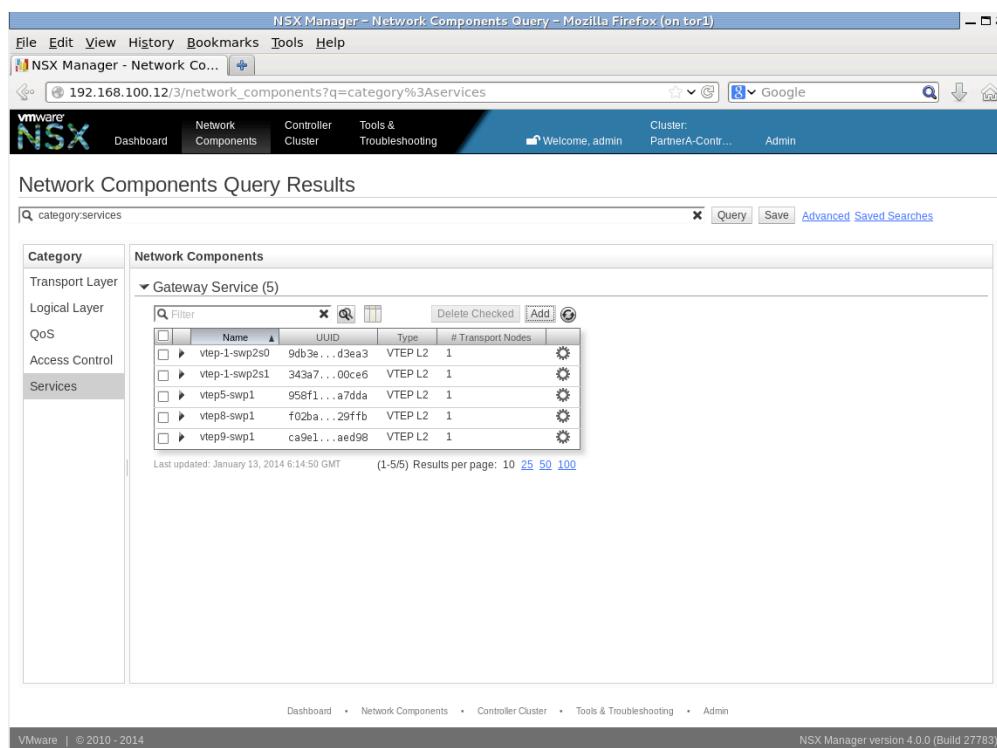
After you finish bootstrapping the NSX integration, you need to configure the transport layer. For each host-facing switch port that is to be associated with a VXLAN instance, define a **Gateway Service** for the port.

1. In NSX Manager, add a new gateway service. Click the **Network Components** tab, then the **Services** category. Under **Gateway Service**, click **Add**. The Create Gateway Service wizard appears.
2. In the Create Gateway Service dialog, select *VTEP L2 Gateway Service* as the **Gateway Service Type**.



3. Give the service a **Display Name** to represent the VTEP in NSX.
4. Click **Add Gateway** to associate the service with the gateway you created earlier.
5. In the **Transport Node** field, choose the name of the gateway you created earlier.
6. In the **Port ID** field, choose the physical port on the gateway (for example, swp10) that will connect to a logical L2 segment and carry data traffic.
7. Click **OK** to save this gateway in the service, then click **Save** to save the gateway service.

The gateway service shows up as type *VTEP L2* in NSX.



The screenshot shows the NSX Manager interface for querying network components. The left sidebar has categories: Transport Layer, Logical Layer, QoS, Access Control, and Services (which is selected). The main area displays a table titled 'Network Components' under the 'Gateway Service (5)' section. The table columns are Name, UUID, Type, and # Transport Nodes. The data shows five entries, each with a gear icon for configuration. At the bottom of the table, it says 'Last updated: January 13, 2014 6:14:50 GMT' and '(1-5/5) Results per page: 10 25 50 100'. The footer includes links for Dashboard, Network Components, Controller Cluster, Tools & Troubleshooting, Admin, and the text 'VMware | © 2010 - 2014' and 'NSX Manager version 4.0.0 (Build 27783)'.

Next, you will configure the logical layer on NSX.

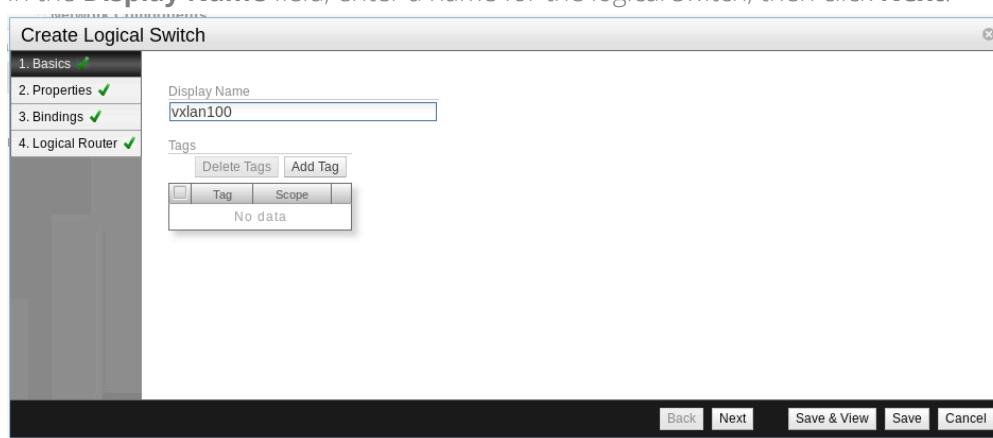
## Configuring the Logical Layer

To complete the integration with NSX, you need to configure the logical layer, which requires defining a logical switch (the VXLAN instance) and all the logical ports needed.

### Defining Logical Switches

To define the logical switch, do the following:

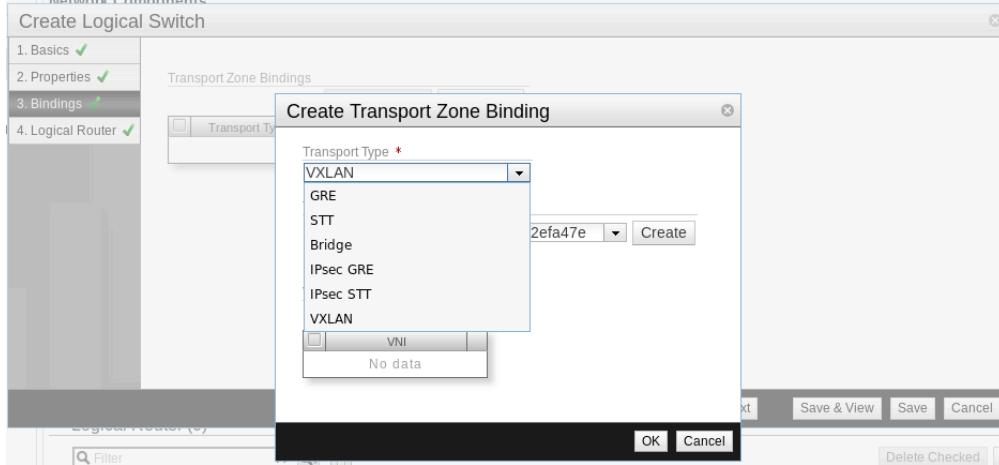
1. In NSX Manager, add a new logical switch. Click the **Network Components** tab, then the **Logical Layer** category. Under **Logical Switch**, click **Add**. The Create Logical Switch wizard appears.
2. In the **Display Name** field, enter a name for the logical switch, then click **Next**.



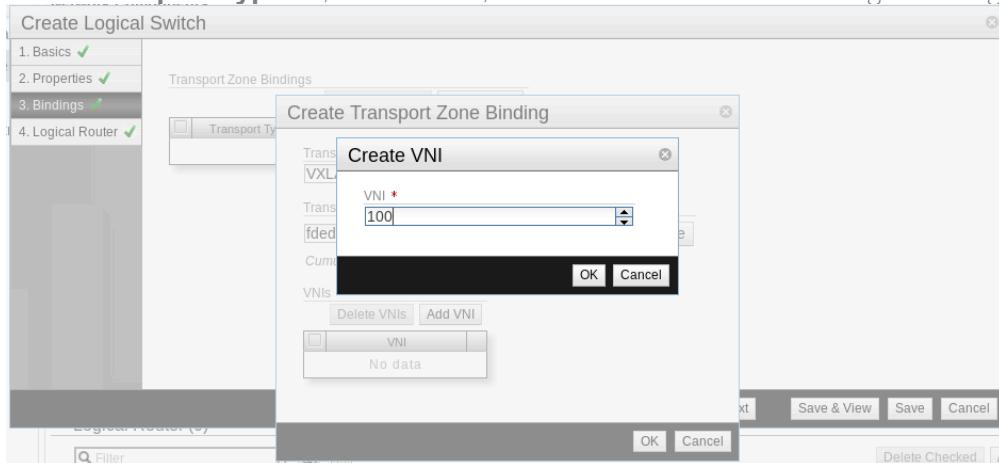
The screenshot shows the 'Create Logical Switch' wizard, step 1: Basics. The 'Display Name' field is populated with 'vxlan100'. Below it is a 'Tags' section with 'Delete Tags' and 'Add Tag' buttons, and a table for 'Tag' and 'Scope' with the message 'No data'. At the bottom are 'Back', 'Next', 'Save & View', 'Save', and 'Cancel' buttons.

3. Under **Replication Mode**, select **Service Nodes**, then click **Next**.

4. Specify the transport zone bindings for the logical switch. Click **Add Binding**. The Create Transport Zone Binding dialog appears.



5. In the **Transport Type** list, select **VXLAN**, then click **OK** to add the binding to the logical switch.

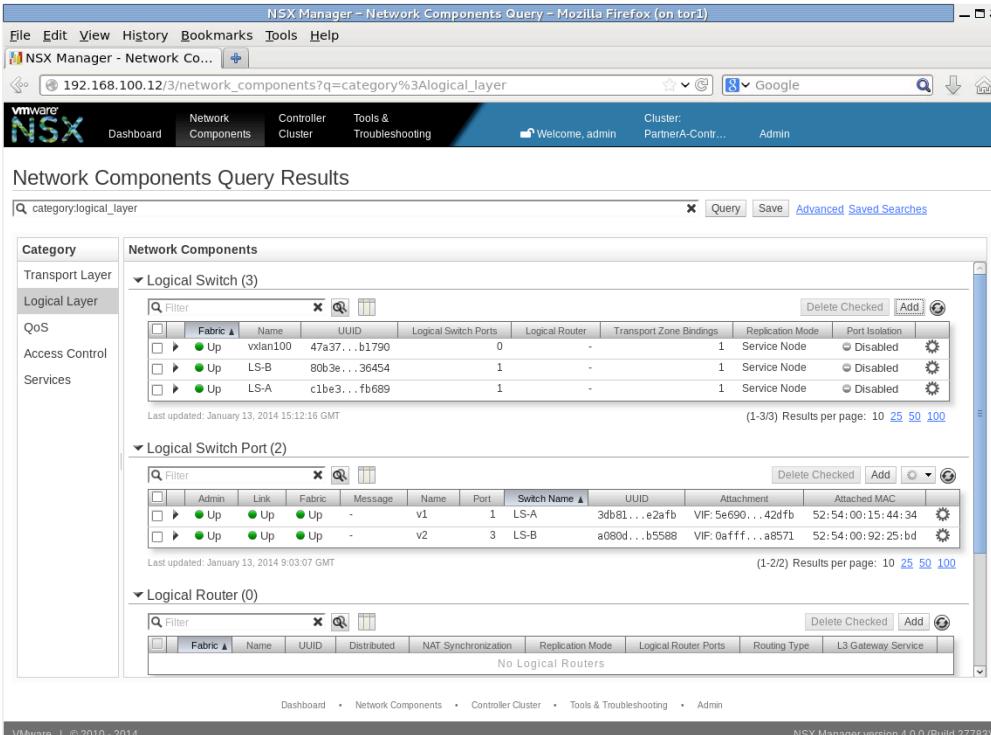


6. In the **VNI** field, assign the switch a VNI ID, then click **OK**.



Do not use 0 or 16777215 as the VNI ID, as they are reserved values under Cumulus Linux.

7. Click **Save** to save the logical switch configuration.



The screenshot shows the NSX Manager interface for querying network components. The main title is "NSX Manager - Network Components Query - Mozilla Firefox (on tor1)". The URL is "192.168.100.12:3/network\_components?q=category%3Alogical\_layer". The navigation bar includes File, Edit, View, History, Bookmarks, Tools, Help, and a search bar for "category:logical\_layer". The left sidebar has categories: Transport Layer, Logical Layer, QoS, Access Control, and Services. The main content area displays two tables:

- Logical Switch (3)**: Shows three entries:
 

Fabric	Name	UUID	Logical Switch Ports	Logical Router	Transport Zone Bindings	Replication Mode	Port Isolation
Up	vxlan100	47a37...b1790	0	-	1	Service Node	Disabled
Up	LS-B	80b3e...36454	1	-	1	Service Node	Disabled
Up	LS-A	c1be3...fb689	1	-	1	Service Node	Disabled

 Last updated: January 13, 2014 15:12:16 GMT. Results per page: 10, 25, 50, 100.
- Logical Switch Port (2)**: Shows two entries:
 

Admin	Link	Fabric	Message	Name	Port	Switch Name	UUID	Attachment	Attached MAC
Up	Up	Up	-	v1	1	LS-A	3db81...e2afb	VIF: 5e690...42dfb	52:54:00:15:44:34
Up	Up	Up	-	v2	3	LS-B	a080d...b5588	VIF: 0afff...a8571	52:54:00:92:25:bd

 Last updated: January 13, 2014 9:03:07 GMT. Results per page: 10, 25, 50, 100.

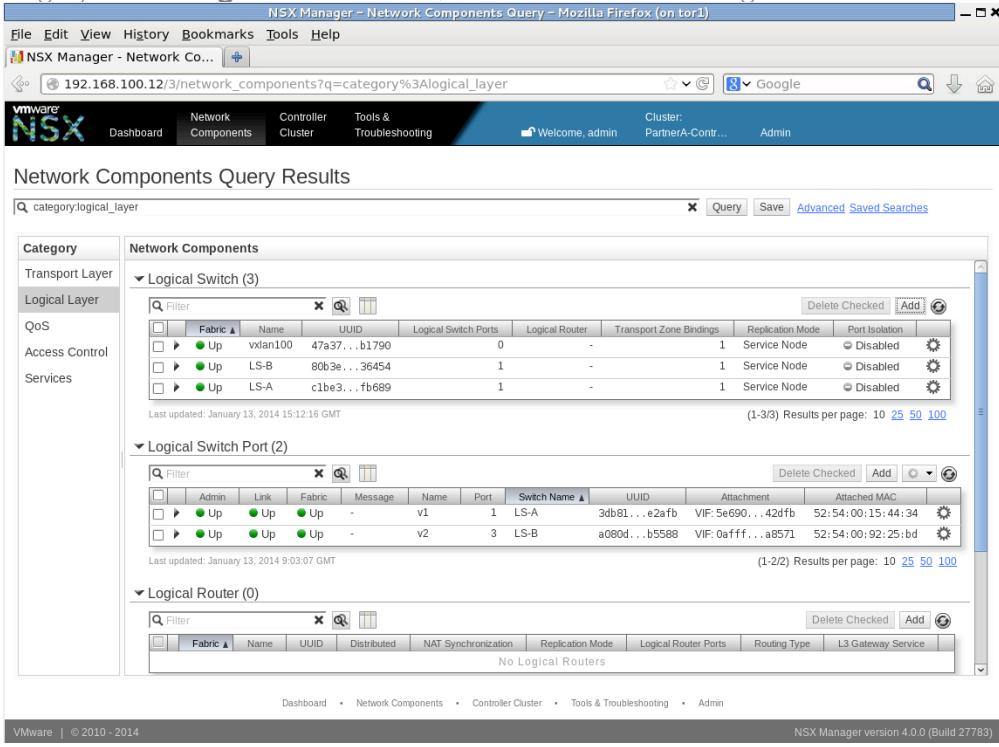
At the bottom, there are links for Dashboard, Network Components, Controller Cluster, Tools & Troubleshooting, Admin, and the footer text "VMware | © 2010 - 2014" and "NSX Manager version 4.0.0 (Build 27783)".

## Defining Logical Switch Ports

As the final step, define the logical switch ports. They can be virtual machine VIF interfaces from a registered OVS, or a VTEP gateway service instance on this switch, as defined above in the Configuring the Transport Layer. A VLAN binding can be defined for each VTEP gateway service associated with the particular logical switch.

To define the logical switch ports, do the following:

- In NSX Manager, add a new logical switch port. Click the **Network Components** tab, then the **Logical Layer** category. Under **Logical Switch Port**, click **Add**. The Create Logical Switch Port



The screenshot shows the NSX Manager interface with the 'Network Components' tab selected. In the left sidebar, 'Logical Layer' is chosen. The main area displays 'Network Components Query Results' with two tables:

- Logical Switch (3)**: Shows three entries: vxlan100 (Fabric: Up, Name: vxlan100, UUID: 47a37...b1790, Logical Switch Ports: 0, Logical Router: -, Transport Zone Bindings: 1, Replication Mode: Service Node, Port Isolation: Disabled), LS-B (Fabric: Up, Name: LS-B, UUID: 80b3e...36454, Logical Switch Ports: 1, Logical Router: -, Transport Zone Bindings: 1, Replication Mode: Service Node, Port Isolation: Disabled), and LS-A (Fabric: Up, Name: LS-A, UUID: c1be3...fb689, Logical Switch Ports: 1, Logical Router: -, Transport Zone Bindings: 1, Replication Mode: Service Node, Port Isolation: Disabled).
- Logical Switch Port (2)**: Shows two entries: v1 (Admin: Up, Link: Up, Fabric: Up, Message: -, Name: v1, Port: 1, Switch Name: LS-A, UUID: 3db81...e2afb, VIF: 5e690...42dfb, Attachment: 52:54:00:15:44:34, Attached MAC: 00:0c:29:44:34:b3) and v2 (Admin: Up, Link: Up, Fabric: Up, Message: -, Name: v2, Port: 3, Switch Name: LS-B, UUID: a080d...b5588, VIF: 0affff...a8571, Attachment: 52:54:00:92:25:bd, Attached MAC: 00:0c:29:44:34:b3).

At the bottom, there are links for Dashboard, Network Components, Controller Cluster, Tools & Troubleshooting, Admin, and the NSX Manager version (4.0.0 Build 27783).

wizard appears.

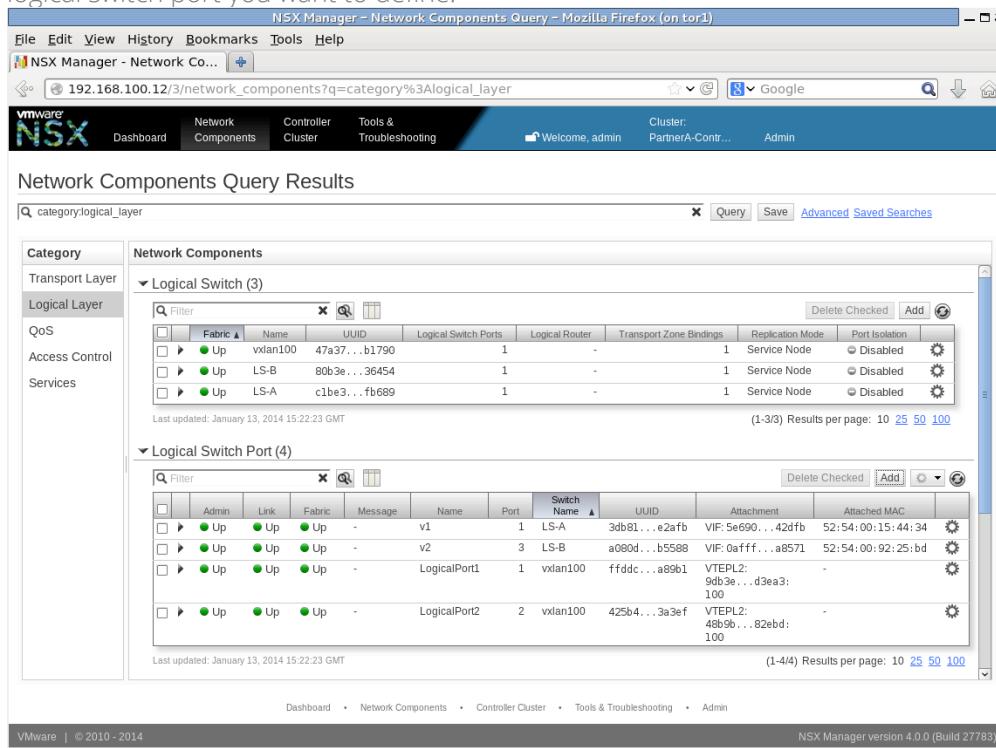
- In the **Logical Switch UUID** list, select the logical switch you created above, then click **Create**.



The screenshot shows the 'Create Logical Switch Port' wizard. On the left, a vertical list of steps is shown: 1. Logical Switch (checked), 2. Basics, 3. Properties, 4. Mirror Targets, 5. Attachment, 6. Port Security, 7. Access Control. The main panel has a field 'Logical Switch UUID \*' containing 'b35d5166-97a8-4ea1-a5c2-64db4adb4af8'. At the bottom are buttons: Back, Next, Save & View, Save, and Cancel.

- In the **Display Name** field, give the port a name that indicates it is the port that connects the gateway, then click **Next**.
- In the **Attachment Type** list, select *VTEP L2 Gateway*.
- In the **VTEP L2 Gateway Service UUID** list, choose the name of the gateway service you created earlier.
- In the **VLAN** list, you can optionally choose a VLAN if you wish to connect only traffic on a specific VLAN of the physical network. Leave it blank to handle all traffic.

- Click **Save** to save the logical switch port. Connectivity is established. Repeat this procedure for each logical switch port you want to define.



The screenshot shows the NSX Manager interface for Network Components Query. On the left, a sidebar lists categories: Transport Layer, Logical Layer (selected), QoS, Access Control, and Services. The main area displays two tables:

- Logical Switch (3)**: Shows three entries. The first entry is vxlan100, which is up, has a Fabric name of v1, and is associated with port 1 on LS-A. The second entry is LS-B, and the third is LS-A. All three have replication mode set to Service Node and Port Isolation disabled.
- Logical Switch Port (4)**: Shows four entries. The first three are logical ports (v1, v2, LogicalPort1) connected to vxlan100 at ports 1, 3, and 1 respectively. The fourth entry is LogicalPort2 connected to vxlan100 at port 2. All ports are up and connected to LS-A.

Both tables include filters, search, and pagination controls (25, 50, 100 results per page).

## Verifying the VXLAN Configuration

Once configured, you can verify the VXLAN configuration using these Cumulus Linux commands in a terminal connected to the switch:

```
cumulus@switch1:~$ sudo ip -d link show vxln100
71: vxln100: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue
    master br-vxln100 state UNKNOWN mode DEFAULT
        link/ether d2:ca:78:bb:7c:9b brd ff:ff:ff:ff:ff:ff
        vxlan id 100 local 172.16.20.157 port 32768 61000 nolearning ageing 300
        svcnode 172.16.21.125
```

OR

```
cumulus@switch1:~$ sudo bridge fdb show
52:54:00:ae:2a:e0 dev vxln100 dst 172.16.21.150 self permanent
d2:ca:78:bb:7c:9b dev vxln100 permanent
90:e2:ba:3f:ce:34 dev swp2s1.100
90:e2:ba:3f:ce:35 dev swp2s0.100
44:38:39:00:48:0e dev swp2s1.100 permanent
44:38:39:00:48:0d dev swp2s0.100 permanent
```

## Persistent VXLAN Configuration in NSX

If you want your VXLAN configuration to persist across upgrades of Cumulus Linux (see [Making Configurations Persist across Upgrades \(see page \)](#)), you need to include the following items in the persistent configuration. Use `scp` to copy the files to `/mnt/persist`:

- `/usr/share/openvswitch/ovs-ctl-vtep`
- Certificates and key pairs, as above
- `/etc/default/openvswitch-vtep`
- The `ovsdb` database file; the default is `/var/lib/openvswitch/conf.db`



Copying the `ovsdb` database file is optional; the persistent database file helps to speed up convergence on a system upgrade. NSX Manager pushes any configuration created or changed in NSX Manager when the connection with the VTEP is reestablished, which overwrites the database file.

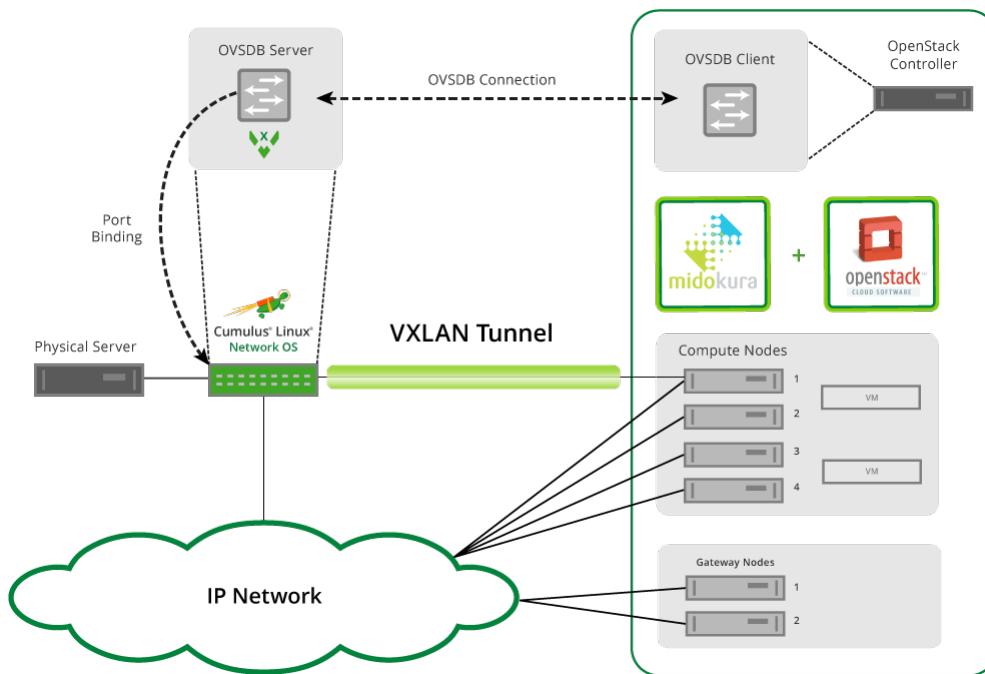
## Troubleshooting VXLANS in NSX

Use `ovsdb-client dump` to troubleshoot issues on the switch. It verifies that the controller and switch handshake is successful. This command works only for VXLANS integrated with NSX:

```
cumulus@switch:~$ sudo ovsdb-client dump -f list | grep -A 7 "Manager"
Manager table
_uuid              : 505f32af-9acb-4182-a315-022e405aa479
_inactivity_probe   : 30000
_is_connected       : true
_max_backoff        : []
_other_config        : {}
_status             : {sec_since_connect="18223", sec_since_disconnect="
18225", state=ACTIVE}
_target              : "ssl:192.168.100.17:6632"
```

## Integrating Hardware VTEPs with Midokura MidoNet and OpenStack

Cumulus Linux seamlessly integrates with the MidoNet OpenStack infrastructure, where the switches provide the VTEP gateway for terminating VXLAN tunnels from within MidoNet. MidoNet connects to the OVSDB server running on the Cumulus Linux switch, and exchanges information about the VTEPs and MAC addresses associated with the OpenStack Neutron networks. This provides seamless Ethernet connectivity between virtual and physical server infrastructures.



## Contents

- Contents (see page 235)
- Getting Started (see page 236)
  - Caveats and Errata (see page 236)
  - Preparing for the MidoNet Integration (see page 236)
    - Enabling the `openvswitch-vtep` Package (see page 236)
  - Bootstrapping the OVSDDB Server and VTEP (see page 237)
    - Automating with the Bootstrap Script (see page 237)
    - Manually Bootstrapping (see page 238)
  - Configuring MidoNet VTEP and Port Bindings (see page 239)
    - Using the MidoNet Manager GUI (see page 239)
      - Creating a Tunnel Zone (see page 239)
      - Adding Hosts to a Tunnel Zone (see page 239)
      - Creating the VTEP (see page 240)
      - Binding Ports to the VTEP (see page 241)
    - Using the MidoNet CLI (see page 242)
- Troubleshooting MidoNet and Cumulus VTEPs (see page 244)
  - Troubleshooting the Control Plane (see page 245)
    - Verifying VTEP and OVSDDB Services (see page 245)
    - Verifying OVSDDB-server Connections (see page 245)
    - Verifying the VXLAN Bridge and VTEP Interfaces (see page 245)
  - Datapath Troubleshooting (see page 246)

- Verifying IP Reachability (see page 247)
- MidoNet VXLAN Encapsulation (see page 247)
- Inspecting the OVSDB (see page 248)
  - Using VTEP-CTL (see page 248)
    - Listing the Physical Switch (see page 248)
    - Listing the Logical Switch (see page 248)
    - Listing Local or Remote MAC Addresses (see page 248)
  - Getting Open Vswitch Database (OVSDB) Data (see page 249)

## Getting Started

Before you create VXLANs with MidoNet, make sure you have the following components:

- A switch (L2 gateway) with a Trident 2 chipset running Cumulus Linux 2.0 and later
- OVSDB server (`ovsdb-server`), included in Cumulus Linux 2.0 and later
- VTEPd (`ovs-vtep`), included in Cumulus Linux 2.0 and later

Integrating a VXLAN with MidoNet involves:

- Preparing for the MidoNet integration
- Bootstrapping the OVS and VTEP
- Configuring the MidoNet VTEP binding
- Verifying the VXLAN configuration

## Caveats and Errata

- There is no support for VXLAN routing in the Trident 2 chipset; use a loopback interface or external router.
- For more information about MidoNet, see the MidoNet Operations Guide, version 1.8 or later.

## Preparing for the MidoNet Integration

Before you start configuring the MidoNet tunnel zones, VTEP binding and connecting virtual ports to the VXLAN, you need to complete the bootstrap process on each switch to which you plan to build VXLAN tunnels. This creates the VTEP gateway and initializes the OVS database server. You only need to do the bootstrapping once, before you begin the MidoNet integration.

## Enabling the `openvswitch-vtep` Package

Before you start bootstrapping the integration, you need to enable the `openvswitch-vtep` package, since it is disabled by default in Cumulus Linux.

1. Edit the `/etc/default/openvswitch-vtep` file, changing the `START` option from `no` to `yes`. This simple `sed` command does this, and creates a backup as well:

```
sudo sed -i.bak s/START=no/START=yes/g /etc/default/openvswitch-vtep
```



Make sure to include this file in persistent storage prior to Cumulus Linux upgrades.

2. Start the daemon:

```
cumulus@switch$ sudo service openvswitch-vtep start
```



Prior to Cumulus Linux 2.5.1, you must edit the control file `/usr/share/openvswitch/scripts/ovs-ctl-vtep`, by adding a parameter for the remote OVS connection. The OVS server connection is not encrypted, so it is necessary to change the PTCP port in this file. The following `sed` command makes the proper change:

```
sed -i.bak "/remote=db/a \\\tset \"\$@\\" --remote=ptcp:6632"
/usr/share/openvswitch/scripts/ovs-ctl-vtep
```

## ***Bootstrapping the OVSDB Server and VTEP***

### ***Automating with the Bootstrap Script***

The `vtep-bootstrap` script is available so you can do the bootstrapping automatically. For information, read `man vtep-bootstrap`. This script requires three parameters, in this order:

- Switch name: The name of the switch that is the VTEP gateway.
- Tunnel IP address: The datapath IP address of the VTEP.
- Management IP address: The IP address of the switch's management interface.

For example, click here ...

```
root@sw11:~# vtep-bootstrap sw11 10.111.1.1 10.50.20.21 --no_encryption
```

```
Executed:
define physical switch
().
Executed:
define local tunnel IP address on the switch
().
Executed:
define management IP address on the switch
().
Executed:
restart a service
```

```
(Killing ovs-vtep (28170).  
Killing ovsdb-server (28146).  
Starting ovsdb-server.  
Starting ovs-vtep.).
```



Prior to Cumulus Linux 2.5.1, the `vtep-bootstrap` command required 4 parameters: `<switch_name> <controller_ip> <tunnel_ip> <management_ip>`. Since MidoNet does not have a controller, you need to use a dummy IP address (for example, 1.1.1.1) for the controller parameter in the bootstrap script. After the script completes, delete the VTEP manager, since it is not needed and will otherwise fill the logs with inconsequential error messages:

```
vtep-ctl del-manager
```

## Manually Bootstrapping

If you don't use the bootstrap script, then you must initialize the OVS database instance manually, and create the VTEP.

Perform the following commands in order (see the automated bootstrapping example above for values):

1. Define the switch in OVSDB:

```
sudo vtep-ctl add-ps <switch_name>
```

2. Define the VTEP tunnel IP address:

```
sudo vtep-ctl set Physical_switch <switch_name> tunnel_ips=<tunnel_ip>
```

3. Define the management interface IP address:

```
sudo vtep-ctl set Physical_switch <switch_name>  
management_ips=<management_ip>
```

4. Restart the OVSDB server and `vtep`:

```
sudo service openvswitch-vtep restart
```

At this point, the switch is ready to connect to MidoNet. The rest of the configuration is performed in the MidoNet Manager GUI, or using the MidoNet API.

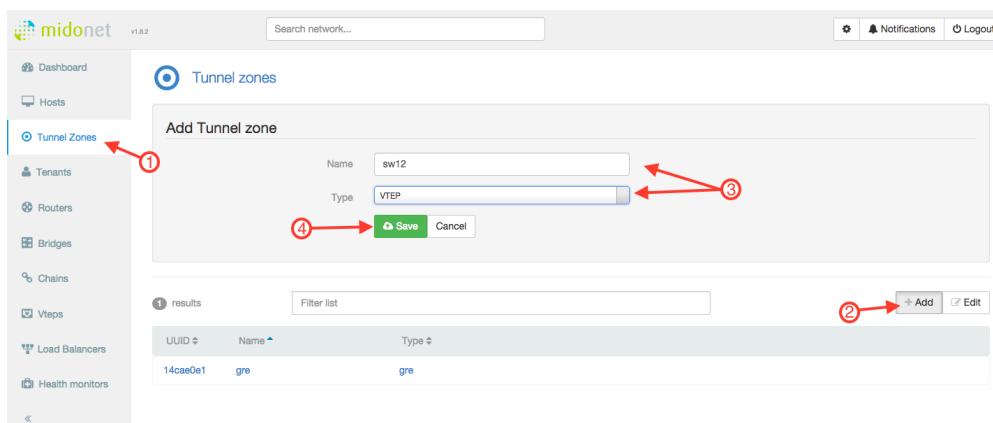
## Configuring MidoNet VTEP and Port Bindings

This part of the configuration sets up MidoNet and OpenStack to connect the virtualization environment to the Cumulus Linux switch. The `midonet-agent` is the networking component that manages the VXLAN, while the Open Virtual Switch (OVS) client on the OpenStack controller node communicates MAC address information between the `midonet-agent` and the Cumulus Linux OVS database (OVSDB) server.

## Using the MidoNet Manager GUI

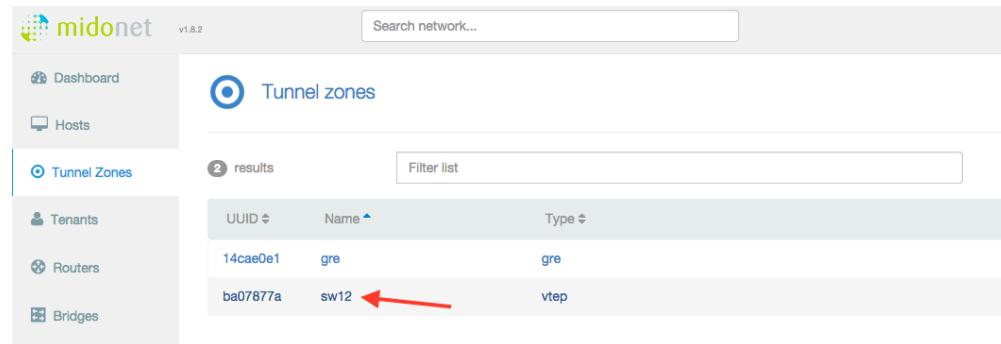
Creating a Tunnel Zone

1. Click **Tunnel Zones** in the menu on the left side.
2. Click **Add**.
3. Give the tunnel zone a **Name** and select "**VTEP**" for the **Type**.
4. Click **Save**.



Adding Hosts to a Tunnel Zone

Once the tunnel zone is created, click the name of the tunnel zone to view the hosts table.

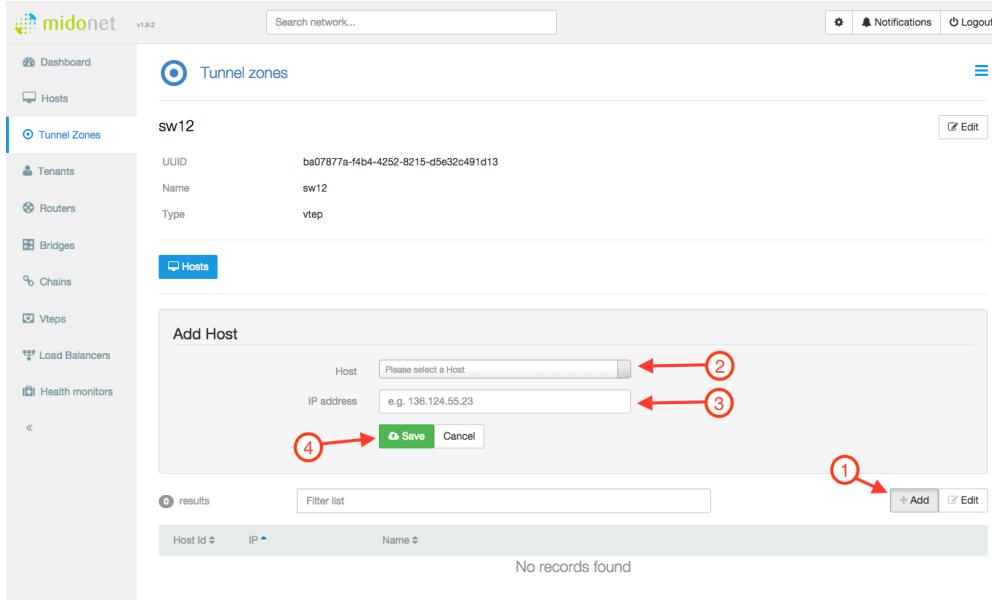


UUID	Name	Type
14cae0e1	gre	gre
ba07877a	sw12	vtep

The tunnel zone is a construct used to define the VXLAN source address used for the tunnel. This host's address is used for the source of the VXLAN encapsulation, and traffic will transit into the routing domain from this point. Thus, the host must have layer 3 reachability to the Cumulus Linux switch tunnel IP.

Next, add a host entry to the tunnel zone:

1. Click **Add**.
2. Select a host from the **Host** list.
3. Provide the tunnel source **IP Address** to use on the selected host.
4. Click **Save**.



The screenshot shows the midonet web interface. On the left sidebar, under the 'Tunnel Zones' section, there is a 'Hosts' tab which is currently selected. In the main content area, there is a 'Tunnel zones' section with a table for 'sw12'. Below it is a 'Hosts' section with a 'Add Host' dialog. The dialog has fields for 'Host' (with a dropdown menu) and 'IP address' (with a value 'e.g. 136.124.55.23'). At the bottom of the dialog are 'Save' and 'Cancel' buttons. A red arrow labeled '1' points to the '+ Add' button in the host list table. Red arrows labeled '2', '3', and '4' point to the 'Host' dropdown, the 'IP address' field, and the 'Save' button respectively.

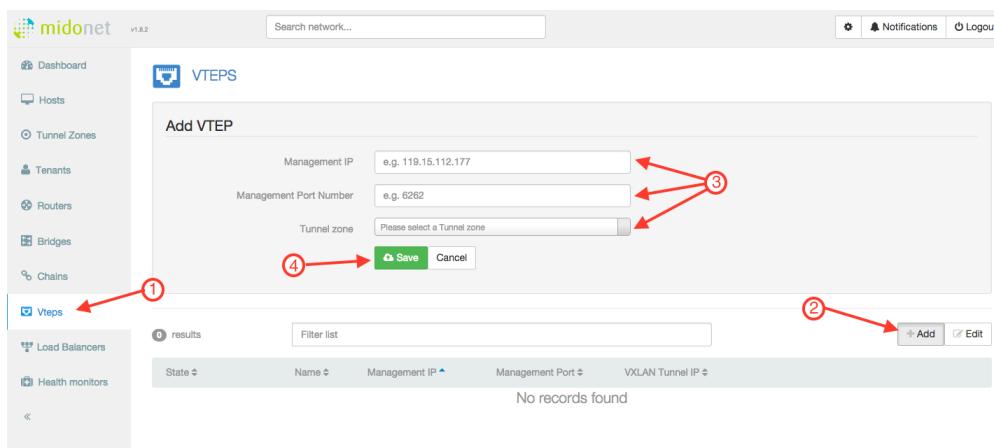
The host list now displays the new entry:



The screenshot shows the 'Hosts' list. The table has columns: Host Id, IP, and Name. There is one row with values: 4d03509b, 10.50.21.182, and os-compute1. At the top right of the table, there is a '+ Add' button.

## Creating the VTEP

1. Click the **Vtaps** menu on the left side.
2. Click **Add**.
3. Fill out the fields using the same information you used earlier on the switch for the bootstrap procedure:
  - **Management IP** is typically the eth0 address of the switch. This tells the OVS-client to connect to the OVSDB-server on the Cumulus Linux switch.
  - **Management Port Number** is the PTCP port you configured in the `ovs-ctl-vtep` script earlier (the example uses 6632).
  - **Tunnel Zone** is the name of the zone you created in the previous procedure.
4. Click **Save**.



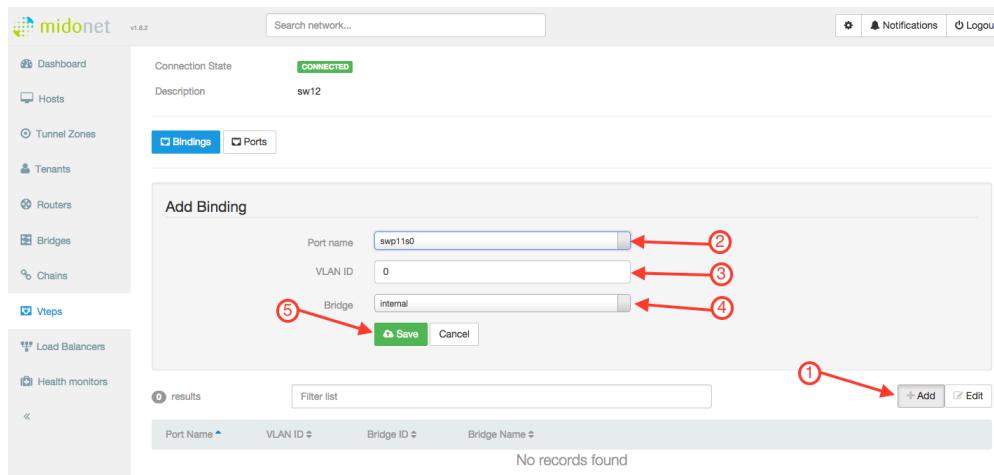
The new VTEP appears in the list below. MidoNet then initiates a connection between the OpenStack Controller and the Cumulus Linux switch. If the OVS client is successfully connected to the OVSDB server, the VTEP entry should display the switch name and VXLAN tunnel IP address, which you specified during the bootstrapping process.

VTEPs				
State	Name	Management IP	Management Port	VXLAN Tunnel IP
CONNECTED	sw11	10.50.20.21	6632	10.111.1.1
CONNECTED	sw12	10.50.20.22	6632	10.111.1.2

### Binding Ports to the VTEP

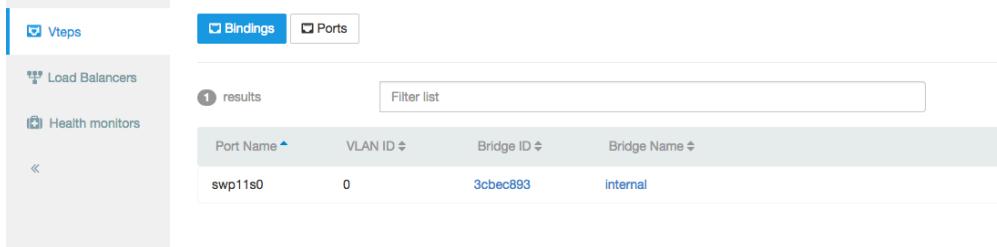
Now that connectivity is established to the switch, you need to add a physical port binding to the VTEP on the Cumulus Linux switch:

1. Click **Add**.
2. In the **Port Name** list, select the port on the Cumulus Linux switch that you are using to connect to the VXLAN segment.
3. Specify the **VLAN ID** (enter 0 for untagged).
4. In the **Bridge** list, select the MidoNet bridge that the instances (VMs) are using in OpenStack.
5. Click **Save**.



The screenshot shows the MidoNet Manager interface. On the left, there's a sidebar with various navigation options like Dashboard, Hosts, Tunnel Zones, Tenants, Routers, Bridges, Chains, VTEPs, Load Balancers, and Health monitors. The main area shows a 'Connected' status for 'Connection State' and a 'Description' of 'sw12'. Below this, there are tabs for 'Bindings' (which is selected) and 'Ports'. A modal window titled 'Add Binding' is open, containing fields for 'Port name' (set to 'swp11s0'), 'VLAN ID' (set to '0'), and 'Bridge' (set to 'internal'). At the bottom of the modal are 'Save' and 'Cancel' buttons. A red arrow labeled '1' points to the 'Add' button in the bottom right corner of the modal. Red arrows labeled '2', '3', '4', and '5' point to the 'Port name', 'VLAN ID', 'Bridge', and 'Save' buttons respectively.

You should see the port binding displayed in the binding table under the VTEP.



The screenshot shows the 'Bindings' table under the 'VTEPs' tab. The table has four columns: 'Port Name', 'VLAN ID', 'Bridge ID', and 'Bridge Name'. There is one entry: 'Port Name' is 'swp11s0', 'VLAN ID' is '0', 'Bridge ID' is '3bec893', and 'Bridge Name' is 'internal'. A red arrow labeled '1' points to the 'Add' button located at the top right of the table header.

Once the port is bound, this automatically configures a VXLAN bridge interface, and includes the VTEP interface and the port bound to the bridge. Now the OpenStack instances (VMs) should be able to ping the hosts connected to the bound port on the Cumulus switch. The Troubleshooting section below demonstrates the verification of the VXLAN data and control planes.

## Using the MidoNet CLI

To get started with the MidoNet CLI, you can access the CLI prompt on the OpenStack Controller:

```
root@os-controller:~# midonet-cli
midonet>
```

Now from the MidoNet CLI, the commands explained in this section perform the same operations depicted in the previous section with the MidoNet Manager GUI.

1. Create a tunnel zone with a name and type vtep:

```
midonet> tunnel-zone create name sw12 type vtep
tzone1
```

2. The tunnel zone is a construct used to define the VXLAN source address used for the tunnel. This host's address is used for the source of the VXLAN encapsulation, and traffic will transit into the routing domain from this point. Thus, the host must have layer 3 reachability to the Cumulus Linux switch tunnel IP.

- First, get the list of available hosts connected to the Neutron network and the MidoNet bridge.
- Next, get a listing of all the interfaces.
- Finally, add a host entry to the tunnel zone ID returned in the previous step, and specify which interface address to use.

```

midonet> list host
host host0 name os-compute1 alive true
host host1 name os-network alive true
midonet> host host0 list interface
iface midonet host_id host0 status 0 addresses [] mac 02:4b:
38:92:dd:ce mtu 1500 type Virtual endpoint DATAPATH
iface lo host_id host0 status 3 addresses [u'127.0.0.1',
u'169.254.169.254', u'0:0:0:0:0:0:0:1'] mac 00:00:00:00:00:
00 mtu 65536 type Virtual endpoint LOCALHOST
iface virbr0 host_id host0 status 1 addresses
[u'192.168.122.1'] mac 22:6e:63:90:1f:69 mtu 1500 type
Virtual endpoint UNKNOWN
iface tap7cf84c-26 host_id host0 status 3 addresses
[u'fe80:0:0:0:e822:94ff:fee2:d41b'] mac ea:22:94:e2:d4:1b
mtu 65000 type Virtual endpoint DATAPATH
iface eth1 host_id host0 status 3 addresses
[u'10.111.0.182', u'fe80:0:0:0:5054:ff:fe85:acd6'] mac 52:
54:00:85:ac:d6 mtu 1500 type Physical endpoint PHYSICAL
iface tapfd4abcea-df host_id host0 status 3 addresses
[u'fe80:0:0:0:14b3:45ff:fe94:5b07'] mac 16:b3:45:94:5b:07
mtu 65000 type Virtual endpoint DATAPATH
iface eth0 host_id host0 status 3 addresses
[u'10.50.21.182', u'fe80:0:0:0:5054:ff:feef:c5dc'] mac 52:
54:00:ef:c5:dc mtu 1500 type Physical endpoint PHYSICAL
midonet> tunnel-zone tzone0 add member host host0 address
10.111.0.182
zone tzone0 host host0 address 10.111.0.182

```

Repeat this procedure for each OpenStack host connected to the Neutron network and the MidoNet bridge.

3. Create a VTEP and assign it to the tunnel zone ID returned in the previous step. The management IP address (the destination address for the VXLAN/remote VTEP) and the port must be the same ones you configured in the `vtep-bootstrap` script or the manual bootstrapping:

```
midonet> vtep add management-ip 10.50.20.22 management-port 6632
tunnel-zone tzone0
name sw12 description sw12 management-ip 10.50.20.22 management-port
6632 tunnel-zone tzone0 connection-state CONNECTED
```

In this step, MidoNet initiates a connection between the OpenStack Controller and the Cumulus Linux switch. If the OVS client is successfully connected to the OVSDB server, the returned values should show the name and description matching the `switch-name` parameter specified in the bootstrap process.



Verify the connection-state as CONNECTED, otherwise if ERROR is returned, you must debug. Typically this only fails if the `management-ip` and/or `management-port` settings are wrong.

4. The VTEP binding uses the information provided to MidoNet from the OVSDB server, providing a list of ports that the hardware VTEP can use for layer 2 attachment. This binding virtually connects the physical interface to the overlay switch, and joins it to the Neutron bridged network.

First, get the UUID of the Neutron network behind the MidoNet bridge:

```
midonet> list bridge
bridge bridge0 name internal state up
bridge bridgel name internal2 state up
midonet> show bridge bridgel id
6c9826da-6655-4fe3-a826-4dcba6477d2d
```

Next, create the VTEP binding, using the UUID and the switch port being bound to the VTEP on the remote end. If there is no VLAN ID, set `vlan` to 0:

```
midonet> vtep name sw12 binding add network-id 6c9826da-6655-4fe3-a826-
4dcba6477d2d physical-port swp11s0 vlan 0
management-ip 10.50.20.22 physical-port swp11s0 vlan 0 network-id
6c9826da-6655-4fe3-a826-4dcba6477d2d
```

At this point, the VTEP should be connected, and the layer 2 overlay should be operational. From the openstack instance (VM), you should be able to ping a physical server connected to the port bound to the hardware switch VTEP.

## Troubleshooting MidoNet and Cumulus VTEPs

As with any complex system, there is a control plane and data plane.

### Troubleshooting the Control Plane

In this solution, the control plane consists of the connection between the OpenStack Controller, and each Cumulus Linux switch running the `ovsdb-server` and `vtep` daemons.

### Verifying VTEP and OVSDB Services

First, it is important that the OVSDB server and `ovs-vtep` daemon are running. Verify this is the case:

```
cumulus@switch12:~$ service openvswitch-vtep status
ovsdb-server is running with pid 17440
ovs-vtep is running with pid 17444
```

### Verifying OVSDB-server Connections

From the OpenStack Controller host, verify that it can connect to the `ovsdb-server`. Telnet to the switch IP address on port 6632:

```
root@os-controller:~# telnet 10.50.20.22 6632
Trying 10.50.20.22...
Connected to 10.50.20.22.
Escape character is '^]'.
<Ctrl+c>
Connection closed by foreign host.
```

If the connection fails, verify IP reachability from the host to the switch. If that succeeds, it is likely the bootstrap process did not set up port 6632. Redo the bootstrapping procedures above.

```
root@os-controller:~# ping -c1 10.50.20.22
PING 10.50.20.22 (10.50.20.22) 56(84) bytes of data.
64 bytes from 10.50.20.22: icmp_seq=1 ttl=63 time=0.315 ms
--- 10.50.20.22 ping statistics ---
1 packets transmitted, 1 received, 0% packet loss, time 0ms
rtt min/avg/max/mdev = 0.315/0.315/0.315/0.000 ms
```

### Verifying the VXLAN Bridge and VTEP Interfaces

After creating the VTEP in MidoNet and adding an interface binding, you should see `br-vxln` and `vxln` interfaces on the switch. You can verify that the VXLAN bridge and VTEP interface are created and UP:

```
cumulus@switch12:~$ sudo brctl show
bridge name  bridge id          STP      enabled interfaces
br-vxln10006 8000.00e0ec2749a2    no       swp11s0
                                         vxln10006
cumulus@switch12:~$ sudo ip -d link show vxln10006
55: vxln10006: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue
master br-vxln10006 state UNKNOWN mode DEFAULT
  link/ether 72:94:eb:b6:6c:c3 brd ff:ff:ff:ff:ff:ff
    vxlan id 10006 local 10.111.1.2 port 32768 61000 nolearning ageing 300
    svcnode 10.111.0.182
      bridge_slave
```

Next, look at the bridging table for the VTEP and the forwarding entries. The bound interface and the VTEP should be listed along with the MAC addresses of those interfaces. When the hosts attached to the bound port send data, those MACs are learned, and entered into the bridging table, as well as the OVSDB.

```
cumulus@switch12:~$ brctl showmacs br-vxln10006
port name      mac addr           vlan      is
local?        ageing timer
swp11s0        00:e0:ec:27:49:a2   0
yes            0.00
swp11s0        64:ae:0c:32:f1:41   0
no              0.01
vxln10006     72:94:eb:b6:6c:c3   0
yes            0.00

cumulus@switch12:~$ sudo bridge fdb show br-vxln10006
fa:16:3e:14:04:2e dev vxln10004 dst 10.111.0.182 vlan 65535 self permanent
00:e0:ec:27:49:a2 dev swp11s0 vlan 0 master br-vxln10004 permanent
b6:71:33:3b:a7:83 dev vxln10004 vlan 0 master br-vxln10004 permanent
64:ae:0c:32:f1:41 dev swp11s0 vlan 0 master br-vxln10004
```

## Datapath Troubleshooting

If you have verified the control plane is correct, and you still cannot get data between the OpenStack instances and the physical nodes on the switch, there may be something wrong with the data plane. The data plane consists of the actual VXLAN encapsulated path, between one of the OpenStack nodes running the midolman service. This is typically the compute nodes, but can include the MidoNet gateway nodes. If the OpenStack instances can ping the tenant router address but cannot ping the physical device connected to the switch (or vice versa), then something is wrong in the data plane.

## Verifying IP Reachability

First, there must be IP reachability between the encapsulating node, and the address you bootstrapped as the tunnel IP on the switch. Verify the OpenStack host can ping the tunnel IP. If this doesn't work, check the routing design, and fix the layer 3 problem first.

```
root@os-compute1:~# ping -c1 10.111.1.2
PING 10.111.1.2 (10.111.1.2) 56(84) bytes of data.
64 bytes from 10.111.1.2: icmp_seq=1 ttl=62 time=0.649 ms
--- 10.111.1.2 ping statistics ---
1 packets transmitted, 1 received, 0% packet loss, time 0ms
rtt min/avg/max/mdev = 0.649/0.649/0.649/0.000 ms
```

## MidoNet VXLAN Encapsulation

If the instance (VM) cannot ping the physical server, or the reply is not returning, look at the packets on the OpenStack node. Initiate a ping from the OpenStack instance, then using `tcpdump`, hopefully you can see the VXLAN data. This example displays what it looks like when it is working.

```
root@os-compute1:~# tcpdump -i eth1 -l -nnn -vvv -X -e port 4789
52:54:00:85:ac:d6 > 00:e0:ec:26:50:36, ethertype IPv4 (0x0800), length 148:
(tos 0x0, ttl 255, id 7583, offset 0, flags [none], proto UDP (17), length
134)
  10.111.0.182.41568 > 10.111.1.2.4789: [no cksum] VXLAN, flags [I] (0x08),
vni 10008
  fa:16:3e:14:04:2e > 64:ae:0c:32:f1:41, ethertype IPv4 (0x0800), length 98:
(tos 0x0, ttl 64, id 64058, offset 0, flags [DF], proto ICMP (1), length 84)
  10.111.102.104 > 10.111.102.2: ICMP echo request, id 15873, seq 0, length
64
  0x0000: 4500 0086 1d9f 0000 ff11 8732 0a6f 00b6 E.....2.o..
  0x0010: 0a6f 0102 a260 12b5 0072 0000 0800 0000 .o...`....r....
  0x0020: 0027 1800 64ae 0c32 f141 fa16 3e14 042e .'..d..2.A..>...
  0x0030: 0800 4500 0054 fa3a 4000 4001 5f26 0a6f ..E..T.:@._&.
  0x0040: 6668 0a6f 6602 0800 f9de 3e01 0000 4233 fh.of.....>...B3
  0x0050: 7dec 0000 0000 0000 0000 0000 0000 0000 }.....
  0x0060: 0000 0000 0000 0000 0000 0000 0000 0000 .....
  0x0070: 0000 0000 0000 0000 0000 0000 0000 0000 .....
  0x0080: 0000 0000 0000 .....
00:e0:ec:26:50:36 > 52:54:00:85:ac:d6, ethertype IPv4 (0x0800), length 148:
(tos 0x0, ttl 62, id 2689, offset 0, flags [none], proto UDP (17), length
134)
  10.111.1.2.63385 > 10.111.0.182.4789: [no cksum] VXLAN, flags [I] (0x08),
```

```
vni 10008
64:ae:0c:32:f1:41 > fa:16:3e:14:04:2e, ethertype IPv4 (0x0800), length 98:
(tos 0x0, ttl 255, id 64058, offset 0, flags [DF], proto ICMP (1), length
84)
10.111.102.2 > 10.111.102.104: ICMP echo reply, id 15873, seq 0, length 64
0x0000: 4500 0086 0a81 0000 3e11 5b51 0a6f 0102 E.....>. [Q.o..
0x0010: 0a6f 00b6 f799 12b5 0072 0000 0800 0000 .o.....r.....
0x0020: 0027 1800 fa16 3e14 042e 64ae 0c32 f141 .'....>...d..2.A
0x0030: 0800 4500 0054 fa3a 4000 ff01 a025 0a6f ..E..T.:@....%..o
0x0040: 6602 0a6f 6668 0000 01df 3e01 0000 4233 f..ofh....>...B3
0x0050: 7dec 0000 0000 0000 0000 0000 0000 0000 }.....'.
0x0060: 0000 0000 0000 0000 0000 0000 0000 0000 .....'.
0x0070: 0000 0000 0000 0000 0000 0000 0000 0000 .....'.
0x0080: 0000 0000 0000 .....
```

## Inspecting the OVSDB

### Using VTEP-CTL

These commands show you the information installed in the OVSDB. This database is structured using the *physical switch* ID, with one or more *logical switch* IDs associated with it. The bootstrap process creates the physical switch, and MidoNet creates the logical switch after the control session is established.

#### Listing the Physical Switch

```
cumulus@switch12:~$ vtep-ctl list-ps
sw12
```

#### Listing the Logical Switch

```
cumulus@switch12:~$ vtep-ctl list-ls
mn-6c9826da-6655-4fe3-a826-4dcba6477d2d
```

#### Listing Local or Remote MAC Addresses

These commands show the MAC addresses learned from the connected port bound to the logical switch, or the MAC addresses advertised from MidoNet. The *unknown-dst* entries are installed to satisfy the ethernet flooding of unknown unicast, and important for learning.

```
cumulus@switch12:~$ vtep-ctl list-local-macs mn-6c9826da-6655-4fe3-a826-
4dcba6477d2d
ucast-mac-local
64:ae:0c:32:f1:41 -> vxlan_over_ipv4/10.111.1.2
```

```

mcast-mac-local
    unknown-dst -> vxlan_over_ipv4/10.111.1.2

cumulus@switch12:~$ vtep-ctl list-remote-macs mn-6c9826da-6655-4fe3-a826-
4dcba6477d2d
ucast-mac-remote
    fa:16:3e:14:04:2e -> vxlan_over_ipv4/10.111.0.182
mcast-mac-remote
    unknown-dst -> vxlan_over_ipv4/10.111.0.182oh

```

## Getting Open Vswitch Database (OVSDB) Data

The `ovsdb-client dump` command is large, but shows all of the information and tables that are used in communication between the OVS client and server.

```

cumulus@switch12:~$ ovsdb-client dump
Arp_Sources_Local table
_uuid locator src_mac
-----
Arp_Sources_Remote table
_uuid locator src_mac
-----
Global table
_uuid managers switches
-----

76672d6a-2740-4c8d-9618-9e8dfb4b0bd7 [ ] [6d459554-0c75-4170-bb3d-
117eb4ce1f4d]
Logical_Binding_Stats table
_uuid bytes_from_local bytes_to_local packets_from_local packets_to_local
-----
d2e378b4-61c1-4daf-9aec-a7fd352d3193 5782569 1658250 21687 14589
Logical_Router table
_uuid description name static_routes switch_binding
-----
Logical_Switch table
_uuid description name tunnel_key
-----
44d162dc-0372-4749-a802-5b153c7120ec "" "mn-6c9826da-6655-4fe3-a826-
4dcba6477d2d" 10006
Manager table
_uuid inactivity_probe is_connected max_backoff other_config status target

```

```
-----  
Mcast_Macs_Local table  
MAC _uuid ipaddr locator_set logical_switch  
-----  
-----  
unknown-dst 25eaf29a-c540-46e3-8806-3892070a2de5 "" 7a4c000a-244e-4b37-8f25-  
fd816c1a80dc 44d162dc-0372-4749-a802-5b153c7120ec  
Mcast_Macs_Remote table  
MAC _uuid ipaddr locator_set logical_switch  
-----  
-----  
unknown-dst b122b897-5746-449e-83ba-fa571a64b374 "" 6c04d477-18d0-41df-8d52-  
dc7b17845ebe 44d162dc-0372-4749-a802-5b153c7120ec  
Physical_Locator table  
_uuid dst_ip encapsulation_type  
-----  
2fcf8b7e-e084-4bcb-b668-755ae7ac0bfb "10.111.0.182" "vxlan_over_ipv4"  
3f78dbb0-9695-42ef-a31f-aaaf525147f1 "10.111.1.2" "vxlan_over_ipv4"  
Physical_Locator_Set table  
_uuid locators  
-----  
6c04d477-18d0-41df-8d52-dc7b17845ebe [2fcf8b7e-e084-4bcb-b668-755ae7ac0bfb]  
7a4c000a-244e-4b37-8f25-fd816c1a80dc [3f78dbb0-9695-42ef-a31f-aaaf525147f1]  
Physical_Port table  
_uuid description name port_fault_status vlan_bindings vlan_stats  
-----  
-----  
bf69fcbb-36b3-4dbc-a90d-fc7412e57076 "swp1" "swp1" [] {} {}  
bf38137d-3a14-454e-8df0-9c56e4b4e640 "swp10" "swp10" [] {} {}  
69585fff-4360-4177-901d-8360ade5391b "swp11s0" "swp11s0" [] {0=44d162dc-  
0372-4749-a802-5b153c7120ec} {0=d2e378b4-61c1-4daf-9aec-a7fd352d3193}  
2a2d04fa-7190-41fe-8cee-318fcbafb2ea "swp11s1" "swp11s1" [] {} {}  
684f99d5-426c-45c8-b964-211489f45599 "swp11s2" "swp11s2" [] {} {}  
47cc66fb-eef8a-4a9b-a497-1844b89f7d32 "swp11s3" "swp11s3" [] {} {}  
5be3a052-be0f-4258-94cb-5e8be9afb896 "swp12" "swp12" [] {} {}  
631b19bd-3022-4353-bb2d-f498b0c1cb17 "swp13" "swp13" [] {} {}  
3001c904-b152-4dc4-9d8e-718f24ffa439 "swp14" "swp14" [] {} {}  
a6f8a88a-3877-4f81-b9b4-d75394a09d2c "swp15" "swp15" [] {} {}  
7cb681f4-2206-4c70-85b7-23b60963cd21 "swp16" "swp16" [] {} {}  
3943fb6a-0b49-4806-a014-2bcd4d469537 "swp17" "swp17" [] {} {}  
109a9911-d6c7-4142-b6c9-7c985506abb4 "swp18" "swp18" [] {} {}  
93b85c31-be38-4384-8b7a-9696764f9ba9 "swp19" "swp19" [] {} {}  
bcfb2920-6676-494c-9dcb-b474123b7e59 "swp2" "swp2" [] {} {}  
4223559a-dalc-4c34-b8bf-bff7ced376ad "swp20" "swp20" [] {} {}
```

6bbccda8-d7e5-4b19-b978-4ec7f5b868e0 "swp21" "swp21" [] {} {}  
c6876886-8386-4e34-a307-931909fca58f "swp22" "swp22" [] {} {}  
c5a88dd6-d931-4b2c-9baa-a0abfb9d41f5 "swp23" "swp23" [] {} {}  
124d1e01-a187-4427-819f-21de66e76f13 "swp24" "swp24" [] {} {}  
55b49814-b5c5-405e-8e9f-898f3df4f872 "swp25" "swp25" [] {} {}  
b2b2cd14-662d-45a5-87c1-277acbccdfdf "swp26" "swp26" [] {} {}  
c35f55f5-8ec6-4fed-bef4-49801cd0934c "swp27" "swp27" [] {} {}  
a44c5402-6218-4f09-bf1e-518f41a5546e "swp28" "swp28" [] {} {}  
a9294152-2b32-4058-8796-23520fffb7379 "swp29" "swp29" [] {} {}  
e0ee993a-8383-4701-a766-d425654dbb7f "swp3" "swp3" [] {} {}  
d9db91a6-1c10-4154-9269-84877faa79b4 "swp30" "swp30" [] {} {}  
b26ce4dd-b771-4d7b-8647-41fa97aa40e3 "swp31" "swp31" [] {} {}  
652c6cd1-0823-4585-bb78-658e6ca2abfc "swp32" "swp32" [] {} {}  
5b15372b-89f0-4e14-a50b-b6c6f937d33d "swp4" "swp4" [] {} {}  
e00741f1-ba34-47c5-ae23-9269c5d1a871 "swp5" "swp5" [] {} {}  
7096abaf-eebf-4ee3-b0cc-276224bc3e71 "swp6" "swp6" [] {} {}  
439afb62-067e-4bbe-a0d9-ee33a23d2a9c "swp7" "swp7" [] {} {}  
54f6c9df-01a1-4d96-9dcf-3035a33ffb3e "swp8" "swp8" [] {} {}  
c85ed6cd-a7d4-4016-b3e9-34df592072eb "swp9s0" "swp9s0" [] {} {}  
cf382ed6-60d3-43f5-8586-81f4f0f2fb28 "swp9s1" "swp9s1" [] {} {}  
c32a9ff9-fd11-4399-815f-806322f26ff5 "swp9s2" "swp9s2" [] {} {}  
9a7e42c4-228f-4b55-b972-7c3b8352c27d "swp9s3" "swp9s3" [] {} {}  
  
Physical\_Switch table  

_uuid	description	management_ips	name	ports	switch_fault_status	tunnel_ips
tunnels						

```
-----
6d459554-0c75-4170-bb3d-117eb4ce1f4d "sw12" ["10.50.20.22"] "sw12"
[109a9911-d6c7-4142-b6c9-7c985506abb4, 124d1e01-a187-4427-819f-
21de66e76f13, 2a2d04fa-7190-41fe-8cee-318fcbafb2ea, 3001c904-b152-4dc4-9d8e-
718f24ffa439, 3943fb6a-0b49-4806-a014-2bcd4d469537, 4223559a-dalc-4c34-b8bf-
bff7ced376ad, 439afb62-067e-4bbe-a0d9-ee33a23d2a9c, 47cc66fb-ef8a-4a9b-a497-
1844b89f7d32, 54f6c9df-01a1-4d96-9dcf-3035a33ffb3e, 55b49814-b5c5-405e-8e9f-
898f3df4f872, 5b15372b-89f0-4e14-a50b-b6c6f937d33d, 5be3a052-be0f-4258-94cb-
5e8be9afb896, 631b19bd-3022-4353-bb2d-f498b0c1cb17, 652c6cd1-0823-4585-bb78-
658e6ca2abfc, 684f99d5-426c-45c8-b964-211489f45599, 69585fff-4360-4177-901d-
8360ade5391b, 6bcccd8-d7e5-4b19-b978-4ec7f5b868e0, 7096abaf-eebf-4ee3-b0cc-
276224bc3e71, 7cb681f4-2206-4c70-85b7-23b60963cd21, 93b85c31-be38-4384-8b7a-
9696764f9ba9, 9a7e42c4-228f-4b55-b972-7c3b8352c27d, a44c5402-6218-4f09-bf1e-
518f41a5546e, a6f8a88a-3877-4f81-b9b4-d75394a09d2c, a9294152-2b32-4058-8796-
23520fffb7379, b26ce4dd-b771-4d7b-8647-41fa97aa40e3, b2b2cd14-662d-45a5-87c1-
277acbccdfdf, bcfb2920-6676-494c-9dcbb474123b7e59, bf38137d-3a14-454e-8df0-
9c56e4b4e640, bf69fcbb-36b3-4dbc-a90d-fc7412e57076, c32a9ff9-fd11-4399-815f-
806322f26ff5, c35f55f5-8ec6-4fed-bef4-49801cd0934c, c5a88dd6-d931-4b2c-9baa-
a0abfb9d41f5, c6876886-8386-4e34-a307-931909fca58f, c85ed6cd-a7d4-4016-b3e9-
34df592072eb, cf382ed6-60d3-43f5-8586-81f4f0f2fb28, d9db91a6-1c10-4154-9269-
84877faa79b4, e00741f1-ba34-47c5-ae23-9269c5d1a871, e0ee993a-8383-4701-a766-
d425654dbb7f] [] ["10.111.1.2"] [062eaf89-9bd5-4132-8b6b-09db254325af]
```

Tunnel table

```
_uuid bfd_config_local bfd_config_remote bfd_params bfd_status local remote
```

```
-----
062eaf89-9bd5-4132-8b6b-09db254325af {bfd_dst_ip="169.254.1.0",
bfd_dst_mac="00:23:20:00:00:01"} {} {} {} 3f78dbb0-9695-42ef-a31f-
aaaf525147f1 2fcf8b7e-e084-4bcb-b668-755ae7ac0bfb
```

Ucast\_Macs\_Local table

```
MAC _uuid ipaddr locator logical_switch
```

```
-----
"64:ae:0c:32:f1:41" 47a83a7c-bd2d-4c02-9814-8222229c592f "" 3f78dbb0-9695-
42ef-a31f-aaaf525147f1 44d162dc-0372-4749-a802-5b153c7120ec
```

Ucast\_Macs\_Remote table

```
MAC _uuid ipaddr locator logical_switch
```

```
-----
"fa:16:3e:14:04:2e" 65605488-9ee5-4c8e-93e5-7b1cc15cfcc7 "" 2fcf8b7e-e084-
4bcb-b668-755ae7ac0bfb 44d162dc-0372-4749-a802-5b153c7120ec
```

## ***Lightweight Network Virtualization - LNV***

Lightweight Network Virtualization (LNV) is a technique for deploying VXLANS (see page 220) without a central controller on bare metal switches. This solution requires no external controller or software suite; it runs the VXLAN service and registration daemons on Cumulus Linux itself. The data path between bridge entities is established on top of a layer 3 fabric by means of a simple service node coupled with traditional MAC address learning.

To see an example of a full solution before reading the following background information, [please read this chapter](#) (see page 280).



LNV is a lightweight controller option. Please [contact Cumulus Networks](#) with your scale requirements so we can make sure this is the right fit for you. There are also other controller options that can work on Cumulus Linux.

## **Contents**

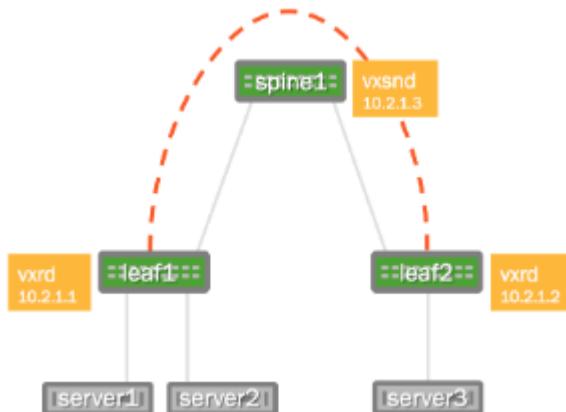
(Click to expand)

- [Contents \(see page 253\)](#)
- [Understanding LNV Concepts \(see page 254\)](#)
  - [Acquiring the Forwarding Database at the Service Node \(see page 254\)](#)
  - [MAC Learning and Flooding \(see page 254\)](#)
  - [Handling BUM Traffic \(see page 255\)](#)
- [Requirements \(see page 256\)](#)
  - [Hardware Requirements \(see page 256\)](#)
  - [Configuration Requirements \(see page 256\)](#)
  - [Installing the LNV Packages \(see page 256\)](#)
- [Sample LNV Configuration \(see page 256\)](#)
  - [Network Connectivity \(see page 257\)](#)
  - [Layer 3 IP Addressing \(see page 257\)](#)
  - [Layer 3 Fabric \(see page 258\)](#)
  - [Host Configuration \(see page 260\)](#)
- [Configuring the VLAN to VXLAN Mapping \(see page 261\)](#)
- [Verifying the VLAN to VXLAN Mapping \(see page 262\)](#)
- [Enabling and Managing Service Node and Registration Daemons \(see page 263\)](#)
  - [Enabling the Service Node Daemon \(see page 263\)](#)
  - [Enabling the Registration Daemon \(see page 264\)](#)
  - [Checking the Daemon Status \(see page 264\)](#)
- [Configuring the Registration Node \(see page 265\)](#)
- [Configuring the Service Node \(see page 267\)](#)
- [Verification and Troubleshooting \(see page 268\)](#)
  - [Verifying the Registration Node Daemon \(see page 268\)](#)

- Verifying the Service Node Daemon (see page 269)
- Verifying Traffic Flow and Checking Counters (see page 269)
- Pinging to Test Connectivity (see page 270)
- Troubleshooting with MAC Addresses (see page 272)
- Checking the Service Node Configuration (see page 272)
- Creating a Layer 3 Gateway (see page 273)
- Advanced LNV Usage (see page 273)
  - Scaling LNV by Load Balancing with Anycast (see page 273)
- Additional Resources (see page 279)
- See Also (see page 280)

## ***Understanding LNV Concepts***

To best describe this feature, consider the following example deployment:



The two switches running Cumulus Linux, called leaf1 and leaf2, each have a bridge configured. These two bridges contain the physical switch port interfaces connecting to the servers as well as the logical VXLAN interface associated with the bridge. By creating a logical VXLAN interface on both leaf switches, the switches become VTEPs (virtual tunnel end points). The IP address associated with this VTEP is most commonly configured as its loopback address — in the image above, the loopback address is 10.2.1.1 for leaf1 and 10.2.1.2 for leaf2.

## ***Acquiring the Forwarding Database at the Service Node***

In order to connect these two VXLANs together and forward BUM (Broadcast, Unknown-unicast, Multicast) packets to members of a VXLAN, the service node needs to acquire the addresses of all the VTEPs for every VXLAN it serves. The service node daemon does this through a registration daemon running on each leaf switch that contains a VTEP participating in LNV. The registration process informs the service node of all the VXLANs to which the switch belongs.

## ***MAC Learning and Flooding***

With LNV, as with traditional bridging of physical LANs or VLANs, a bridge automatically learns the location of hosts as a side effect of receiving packets on a port.

For example, when server1 sends an L2 packet to server3, leaf2 learns that server1's MAC address is located on that particular VXLAN, and the VXLAN interface learns that the IP address of the VTEP for server1 is 10.2.1.1. So when server3 sends a packet to server1, the bridge on leaf2 forwards the packet out of the port to the VXLAN interface and the VXLAN interface sends it, encapsulated in a UDP packet, to the address 10.2.1.1.

But what if server3 sends a packet to some address that has yet to send it a packet (server2, for example)? In this case, the VXLAN interface sends the packet to the service node, which sends a copy to every other VTEP that belongs to the same VXLAN.

## Handling BUM Traffic

Cumulus Linux has two ways of handling BUM (Broadcast Unknown-Unicast and Multicast) traffic:

- Head end replication
- Service node replication

Head end replication is enabled by default in Cumulus Linux.



You cannot have both service node and head end replication configured simultaneously, as this causes the BUM traffic to be duplicated — both the source VTEP and the service node sending their own copy of each packet to every remote VTEP.

## Head End Replication

The Trident II chipset is capable of head end replication — the ability to generate all the BUM (Broadcast Unknown-Unicast and Multicast) traffic in hardware. The most scalable solution available with LNV is to have each VTEP (top of rack switch) generate all of its own BUM traffic rather than relying on an external service node.

Cumulus Linux supports up to 64 VTEPs with head end replication.

To disable head end replication, edit `/etc/vxrd.conf` and set `head_rep` to `False`.

## Service Node Replication

Cumulus Linux also supports service node replication for VXLAN BUM packets. This is useful with LNV if you have more than 64 VTEPs. However, it is not recommended because it forces the spine switches running the `vxsnd` (service node daemon) to replicate the packets in software instead of in hardware, unlike head end replication. If you're not using a controller but have more than 64 VTEPs, contact a [Cumulus Networks consultant](#).

To enable service node replication:

1. Disable head end replication; set `head_rep` to `False` in `/etc/vxrd.conf`.
2. Edit `/etc/network/interfaces` and configure a service node IP address for VXLAN interfaces using `vxrd-svcnode-ip <>`.
3. Edit `/etc/vxsnd.conf`, and configure the following:
  - Set the same service node IP address that you did in the previous step:  
`svcnode_ip = <>`

- To forward VXLAN data traffic, set the following variable to *True*:  
`enable_vxlan_listen = true`

## Requirements

### Hardware Requirements

- Switches with a Trident II chipset running Cumulus Linux 2.5.4 or later. Please refer to the Cumulus Networks [hardware compatibility list](#) for a list of supported switch models.

### Configuration Requirements

- The VXLAN has an associated **VXLAN Network Identifier** (VNI), also interchangeably called a VXLAN ID.
- The VNI should not be 0 or 16777215, as these two numbers are reserved values under Cumulus Linux.
- The VXLAN link and physical interfaces are added to the bridge to create the association between the port, VLAN and VXLAN instance.
- Each bridge on the switch has only one VXLAN interface. Cumulus Linux does not support more than one VXLAN link in a bridge; however, a switch can have multiple bridges.
- Only use bridges in [traditional mode \(see page 162\)](#); [VLAN-aware bridges \(see page 182\)](#) are not supported with VXLAN at this time.
- An SVI (Switch VLAN Interface) or L3 address on the bridge is not supported. For example, you can't ping from the leaf1 SVI to the leaf2 SVI via the VXLAN tunnel; you would need to use server1 and server2 to verify. See [Creating a Layer 3 Gateway \(see page 272\)](#) below for more information.

### Installing the LNV Packages

The LNV packages are not installed automatically if you upgrade Cumulus Linux. You can install LNV in one of two ways:

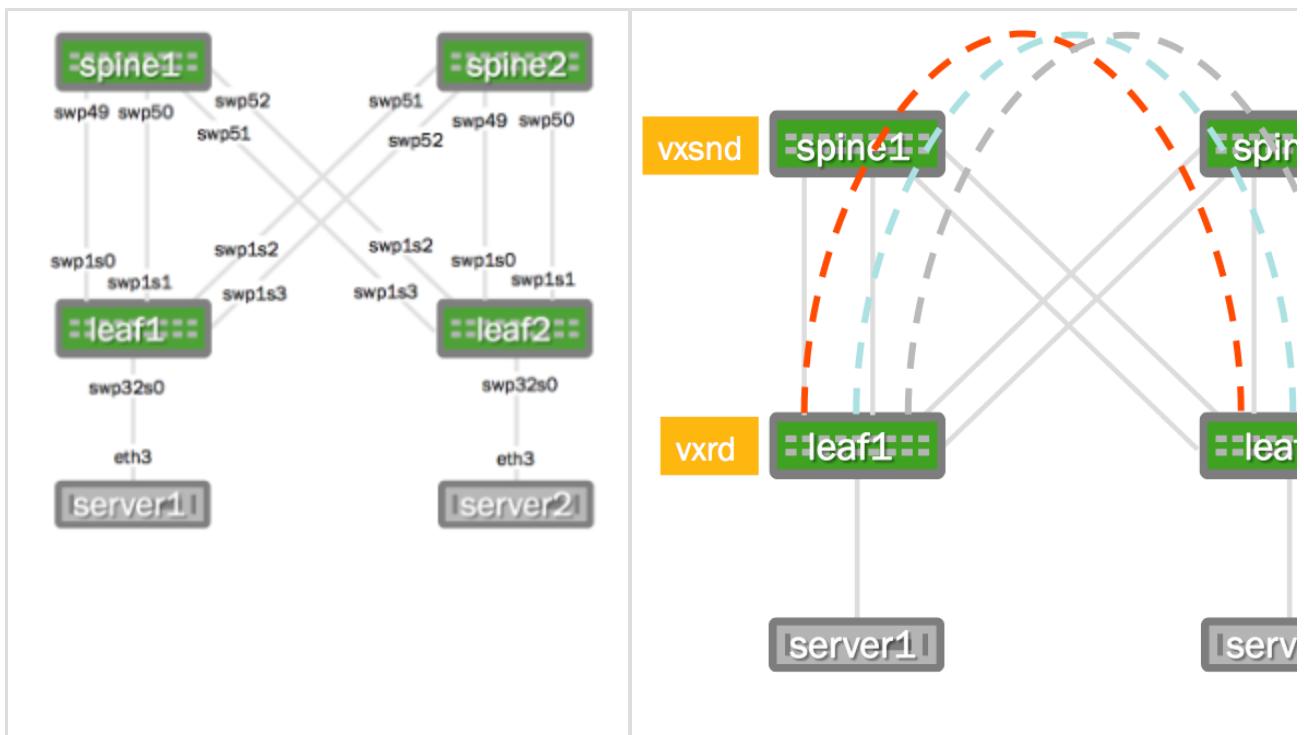
- Do a [binary image install \(see page 16\)](#) of Cumulus Linux, using `cl-img-install`
- Install the LNV packages for the registration and service node daemons using `apt-get install vxflid-vxrd` and/or `apt-get install vxflid-vxsnd`, depending upon how you intend to use LNV

### Sample LNV Configuration

The following images illustrate the configuration that is referenced throughout this chapter.

Physical Cabling Diagram

Network Virtualization Diagram



Want to try out configuring LNV and don't have a Cumulus Linux switch? Sign up to use the [Cumulus Workbench](#), which has this exact topology.

## Network Connectivity

There must be full network connectivity before you can configure LNV. The layer 3 IP addressing information as well as the OSPF configuration (`/etc/quagga/Quagga.conf`) below is provided to make the LNV example easier to understand.



OSPF is not a requirement for LNV, LNV just requires L3 connectivity. With Cumulus Linux this can be achieved with static routes, OSPF or BGP.

## Layer 3 IP Addressing

Here is the configuration for the IP addressing information used in this example.

**spine1:** `/etc/network/interfaces`

```
auto lo
iface lo inet loopback
    address 10.2.1.3/32

auto eth0
iface eth0 inet dhcp
```

**spine2:** `/etc/network/interfaces`

```
auto lo
iface lo inet loopback
    address 10.2.1.4/32

auto eth0
iface eth0 inet dhcp
```

```

auto swp49
iface swp49
  address 10.1.1.2/30

auto swp50
iface swp50
  address 10.1.1.6/30

auto swp51
iface swp51
  address 10.1.1.50/30

auto swp52
iface swp52
  address 10.1.1.54/30
  
```

```

auto swp49
iface swp49
  address 10.1.1.18/30

auto swp50
iface swp50
  address 10.1.1.22/30

auto swp51
iface swp51
  address 10.1.1.34/30

auto swp52
iface swp52
  address 10.1.1.38/30
  
```

**leaf1:** /etc/network/interfaces

```

auto lo
iface lo inet loopback
  address 10.2.1.1/32

auto eth0
iface eth0 inet dhcp

auto swp1s0
iface swp1s0
  address 10.1.1.1/30

auto swp1s1
iface swp1s1
  address 10.1.1.5/30

auto swp1s2
iface swp1s2
  address 10.1.1.33/30

auto swp1s3
iface swp1s3
  address 10.1.1.37/30
  
```

**leaf2:** /etc/network/interfaces

```

auto lo
iface lo inet loopback
  address 10.2.1.2/32

auto eth0
iface eth0 inet dhcp

auto swp1s0
iface swp1s0
  address 10.1.1.17/30

auto swp1s1
iface swp1s1
  address 10.1.1.21/30

auto swp1s2
iface swp1s2
  address 10.1.1.49/30

auto swp1s3
iface swp1s3
  address 10.1.1.53/30
  
```

## Layer 3 Fabric

The service nodes and registration nodes must all be routable between each other. The L3 fabric on Cumulus Linux can either be [BGP](#) (see page 345) or [OSPF](#) (see page 332). In this example, OSPF is used to demonstrate full reachability. Expand the Quagga configurations below.

Quagga configuration using OSPF:

### spine1

```

interface lo
 ip ospf area 0.0.0.0
interface swp49
 ip ospf network point-to-point
 ip ospf area 0.0.0.0
!
interface swp50
 ip ospf network point-to-point
 ip ospf area 0.0.0.0
!
interface swp51
 ip ospf network point-to-point
 ip ospf area 0.0.0.0
!
interface swp52
 ip ospf network point-to-point
 ip ospf area 0.0.0.0
!
!
!
!
!
router-id 10.2.1.3
router ospf
 ospf router-id 10.2.1.3

```

### spine2

```

interface lo
 ip ospf area 0.0.0.0
interface swp49
 ip ospf network point-to-point
 ip ospf area 0.0.0.0
!
interface swp50
 ip ospf network point-to-point
 ip ospf area 0.0.0.0
!
interface swp51
 ip ospf network point-to-point
 ip ospf area 0.0.0.0
!
interface swp52
 ip ospf network point-to-point
 ip ospf area 0.0.0.0
!
!
!
!
!
router-id 10.2.1.4
router ospf
 ospf router-id 10.2.1.4

```

### leaf1

```

interface lo
 ip ospf area 0.0.0.0
interface swp1s0
 ip ospf network point-to-
point
 ip ospf area 0.0.0.0
!
interface swp1s1
 ip ospf network point-to-
point
 ip ospf area 0.0.0.0

```

### leaf2

```

interface lo
 ip ospf area 0.0.0.0
interface swp1s0
 ip ospf network point-to-
point
 ip ospf area 0.0.0.0
!
interface swp1s1
 ip ospf network point-to-
point
 ip ospf area 0.0.0.0

```

```
!
interface swp1s2
  ip ospf network point-to-
  point
  ip ospf area 0.0.0.0
!
interface swp1s3
  ip ospf network point-to-
  point
  ip ospf area 0.0.0.0
!
!
!
!
!
router-id 10.2.1.1
router ospf
  ospf router-id 10.2.1.1
```

```
!
interface swp1s2
  ip ospf network point-to-
  point
  ip ospf area 0.0.0.0
!
interface swp1s3
  ip ospf network point-to-
  point
  ip ospf area 0.0.0.0
!
!
!
!
!
router-id 10.2.1.2
router ospf
  ospf router-id 10.2.1.2
```

## Host Configuration

In this example, the servers are running Ubuntu 14.04. There needs to be a trunk mapped from server1 and server2 to the respective switch. In Ubuntu this is done with subinterfaces. You can expand the configurations below.

server1

```
auto eth3.10
iface eth3.10 inet
static
  address 10.10.10.1/24

auto eth3.20
iface eth3.20 inet
static
  address 10.10.20.1/24

auto eth3.30
iface eth3.30 inet
static
  address 10.10.30.1/24
```

server2

```
auto eth3.10
iface eth3.10 inet
static
  address 10.10.10.2/24

auto eth3.20
iface eth3.20 inet
static
  address 10.10.20.2/24

auto eth3.30
iface eth3.30 inet
static
  address 10.10.30.2/24
```

On Ubuntu it is more reliable to use `ifup` and `if down` to bring the interfaces up and down individually, rather than restarting networking entirely (that is, there is no equivalent to `if reload` like there is in Cumulus Linux):

```
cumulus@server1:~$ sudo ifup eth3.10
Set name-type for VLAN subsystem. Should be visible in /proc/net/vlan
/config
Added VLAN with VID == 10 to IF -:eth3:-
cumulus@server1:~$ sudo ifup eth3.20
Set name-type for VLAN subsystem. Should be visible in /proc/net/vlan
/config
Added VLAN with VID == 20 to IF -:eth3:-
cumulus@server1:~$ sudo ifup eth3.30
Set name-type for VLAN subsystem. Should be visible in /proc/net/vlan
/config
Added VLAN with VID == 30 to IF -:eth3:-
```

## Configuring the VLAN to VXLAN Mapping

Configure the VLANS and associated VXLANs. In this example, there are 3 VLANS and 3 VXLAN IDs (VNIs). VLANS 10, 20 and 30 are used and associated with VNIs 10, 2000 and 30 respectively. The loopback address, used as the vxlan-local-tunnelip, is the only difference between leaf1 and leaf2 for this demonstration.

For leaf1:

```
cumulus@leaf1$ sudo nano /etc
/network/interfaces
```

Add the following to the loopback stanza

```
auto lo
iface lo
    vxrd-src-ip 10.2.1.1
    vxrd-svcnode-ip 10.2.1.3
```

Now append the following for the VXLAN configuration itself:

```
leaf1: /etc/network/interfaces

auto vni-10
iface vni-10
    vxlan-id 10
    vxlan-local-tunnelip
    10.2.1.1

auto vni-2000
iface vni-2000
    vxlan-id 2000
    vxlan-local-tunnelip 10.2.1.1
```

For leaf2:

```
cumulus@leaf2$ sudo nano /etc
/network/interfaces
```

Add the following to the loopback stanza

```
auto lo
iface lo
    vxrd-src-ip 10.2.1.2
    vxrd-svcnode-ip 10.2.1.3
```

Now append the following for the VXLAN configuration itself:

```
leaf2: /etc/network/interfaces

auto vni-10
iface vni-10
    vxlan-id 10
    vxlan-local-tunnelip
    10.2.1.2

auto vni-2000
iface vni-2000
    vxlan-id 2000
    vxlan-local-tunnelip 10.2.1.2
```

```

auto vni-30
iface vni-30
  vxlan-id 30
  vxlan-local-tunnelip 10.2.1.1

auto br-10
iface br-10
  bridge-ports swp32s0.10 vni-10

auto br-20
iface br-20
  bridge-ports swp32s0.20 vni-2000

auto br-30
iface br-30
  bridge-ports swp32s0.30 vni-30
  
```

To bring up the bridges and VNIs, use the `ifreload` command:

```
cumulus@leaf1$ sudo ifreload -a
```

```

auto vni-30
iface vni-30
  vxlan-id 30
  vxlan-local-tunnelip 10.2.1.2

auto br-10
iface br-10
  bridge-ports swp32s0.10 vni-10

auto br-20
iface br-20
  bridge-ports swp32s0.20 vni-2000

auto br-30
iface br-30
  bridge-ports swp32s0.30 vni-30
  
```

To bring up the bridges and VNIs, use the `ifreload` command:

```
cumulus@leaf2$ sudo ifreload -a
```

- i** Why is br-20 not vni-20? For example, why not tie VLAN 20 to VNI 20, or why was 2000 used? VXLANs and VLANs do not need to be the same number. This was done on purpose to highlight this fact. However if you are using fewer than 4096 VLANs, there is no reason not to make it easy and correlate VLANs to VXLANs. It is completely up to you.

## Verifying the VLAN to VXLAN Mapping

Use the `brctl show` command to see the physical and logical interfaces associated with that bridge:

```

cumulus@leaf1:~$ brctl show
bridge name      bridge id      STP enabled      interfaces
br-10           8000.443839008404    no            swp32s0.10
                                         vni-10
br-20           8000.443839008404    no            swp32s0.20
                                         vni-2000
br-30           8000.443839008404    no            swp32s0.30
                                         vni-30
  
```

As with any logical interfaces on Linux, the name does not matter (other than a 15-character limit). To verify the associated VNI for the logical name, use the `ip -d link show` command:

```
cumulus@leaf1$ ip -d link show vni-10
43: vni-10: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue
    master br-10 state UNKNOWN mode DEFAULT
        link/ether 02:ec:ec:bd:7f:c6 brd ff:ff:ff:ff:ff:ff
        vxlan id 10 srcport 32768 61000 dstport 4789 ageing 300
        bridge_slave
```

The `vxlan id 10` indicates the VXLAN ID/VNI is indeed 10 as the logical name suggests.

## Enabling and Managing Service Node and Registration Daemons

Every VTEP must run the registration daemon (`vxrld`). Typically, every leaf switch acts as a VTEP. A minimum of 1 switch (a switch not already acting as a VTEP) must run the service node daemon (`vxsnd`). The instructions for enabling these daemons follows.

### Enabling the Service Node Daemon

The service node daemon (`vxsnd`) is included in the Cumulus Linux repository as `vxf1d-vxsnd`. The service node daemon can run on any switch running Cumulus Linux as long as that switch is not also a VXLAN VTEP. In this example, enable the service node only on the spine1 switch.



Do not run `vxsnd` on a switch that is already acting as a VTEP.

Edit the `/etc/default/vxsnd` configuration file:

```
cumulus@spine1$ sudo nano /etc/default/vxsnd
```

Change the `vxsnd` file by changing `no` to `yes`:

```
START=yes
```

Save and quit the text editor and reboot the `vxsnd` daemon:

```
cumulus@spine1$ sudo service vxsnd restart
[ ok ] Starting /usr/bin/vxsnd ....
```

## Enabling the Registration Daemon

The registration daemon (`vxrd`) is included in the Cumulus Linux package as `vxf1d-vxrd`. The registration daemon must run on each VTEP participating in LNV, so you must enable it on every TOR (leaf) switch acting as a VTEP.

Edit the `/etc/default/vxrd` configuration file on leaf1:

```
cumulus@leaf1$ sudo nano /etc/default/vxrd
```

Change the `vxrd` file by changing `no` to `yes`:

```
START=yes
```

Save and quit the text editor and reboot the `vxrd` daemon:

```
cumulus@leaf1$ sudo service vxrd restart  
[ ok ] Starting /usr/bin/vxrd ....
```

Open the `vxrd` configuration file on leaf2 with the following commands:

```
cumulus@leaf2$ sudo nano /etc/default/vxrd
```

Change the `vxsnd` file by changing `no` to `yes`:

```
START=yes
```

Save and quit the text editor and reboot the `vxrd` daemon:

```
cumulus@leaf1$ sudo service vxrd restart  
[ ok ] Starting /usr/bin/vxrd ....
```

## Checking the Daemon Status

To determine if the daemon is running, use the `service <daemon name> status` command.

For the service node daemon:

```
cumulus@spine1$ sudo service vxsnd status  
[ ok ] vxsnd is running.
```

For the registration daemon:

```
cumulus@leaf1$ sudo service vxrd status  
[ ok ] vxrd is running.
```

## Configuring the Registration Node

The registration node was configured earlier in /etc/network/interfaces in the [VXLAN mapping](#) (see [page 261](#)) section above; no additional configuration is typically needed. However, if you need to modify the registration node configuration, edit /etc/vxrd.conf.

Alternate location for configuration and additional knobs for the registration node are found in /etc/vxrd.conf

```
cumulus@leaf1$ sudo nano /etc/vxrd.conf
```

Then edit the svcnode\_ip variable:

```
svcnode_ip = 10.2.1.3
```

Then perform the same on leaf2:

```
cumulus@leaf2$ sudo nano /etc/vxrd.conf
```

And again edit the svcnode\_ip variable:

```
svcnode_ip = 10.2.1.3
```

Restart the registration node daemon for the change to take effect:

```
cumulus@leaf1$ sudo service vxrd restart  
[ ok ] Starting /usr/bin/vxrd ....
```

Restart the daemon on leaf2:

```
cumulus@leaf2$ sudo service vxrd restart  
[ ok ] Starting /usr/bin/vxrd ....
```

The complete list of options you can configure is listed below:

Name	Description	Default
loglevel	The log level, which can be DEBUG, INFO, WARNING, ERROR, CRITICAL.	INFO
logdest	The destination for log messages. It can be a file name, <code>stdout</code> or <code>syslog</code> .	syslog
logfilesize	Log file size in bytes. Used when <code>logdest</code> is a file name.	512000
logbackupcount	Maximum number of log files stored on the disk. Used when <code>logdest</code> is a file name.	14
pidfile	The PIF file location for the <code>vxrd</code> daemon.	/var/run/vxrd.pid
udsfile	The file name for the Unix domain socket used for management.	/var/run/vxrd.sock
vxfld_port	The UDP port used for VXLAN control messages.	10001
svcnode_ip	The address to which registration daemons send control messages for registration and/or BUM packets for replication. This can also be configured under <code>/etc/network/interfaces</code> with the <code>vxrd-svcnode-ip</code> keyword.	
holdtime	Hold time (in seconds) for soft state, which is how long the service node waits before ageing out an IP address for a VNI. The <code>vxrd</code> includes this in the register messages it sends to a <code>vxsnd</code> .	90 seconds
src_ip	Local IP address to bind to for receiving control traffic from the service node daemon.	
refresh_rate	Number of times to refresh within the hold time. The higher this number, the more lost UDP refresh messages can be tolerated.	3 seconds
config_check_rate	The number of seconds to poll the system for current VXLAN membership.	5 seconds
head_rep	Enables self replication. Instead of using the service node to replicate BUM packets, it will be done in hardware on the VTEP switch.	true



Use `1, yes, true` or `on` for True for each relevant option. Use `0, no, false` or `off` for False.

## Configuring the Service Node

To configure the service node daemon, edit the `/etc/vxsnd.conf` configuration file.



For the example configuration, default values are used, except for the `svcnod_ip` field.

```
cumulus@spine1$ sudo nano /etc/vxsnd.conf
```

The address field is set to the loopback address of the switch running the `vxsnd` dameon.

```
svcnod_ip = 10.2.1.3
```

Restart the service node daemon for the change to take effect:

```
cumulus@spine1$ sudo service vxsnd restart
[ ok ] Starting /usr/bin/vxsnd ....
```

The complete list of options you can configure is listed below:

Name	Description	Default
loglevel	The log level, which can be DEBUG, INFO, WARNING, ERROR, CRITICAL.	INFO
logdest	Destination for log messages. It can be a file name, <code>stdout</code> or <code>syslog</code> .	syslog
logfilesize	The log file size in bytes. Used when <code>logdest</code> is a file name.	512000
logbackupcount	Maximum number of log files stored on disk. Used when <code>logdest</code> is a file name.	14
pidfile	The PID file location for the <code>vxrd</code> daemon.	/var/run/vxrd.pid
udsfile	The file name for the Unix domain socket used for management.	/var/run/vxrd.sock
vxld_port	The UDP port used for VXLAN control messages.	10001
svcnod_ip	This is the address to which registration daemons send control messages for registration and/or BUM packets for replication.	0.0.0.0

Name	Description	Default
holdtime	Holddate (in seconds) for soft state. It is used when sending a register message to peers in response to learning a <vni, addr> from a VXLAN data packet.	90
src_ip	Local IP address to bind to for receiving inter-vxsnnd control traffic.	0.0.0.0
svcnodes_peers	Space-separated list of IP addresses with which the vxsnnd shares its state.	
enable_vxlan_listen	When set to true, the service node listens for VXLAN data traffic.	true
install_svcnode_ip	When set to true, the <code>snd_peer_address</code> gets installed on the loopback interface. It gets withdrawn when the <code>vxsnnd</code> is not in service. If set to true, you must define the <code>snd_peer_address</code> configuration variable.	false
age_check	Number of seconds to wait before checking the database to age out stale entries.	90 seconds



Use `1, yes, true` or `on` for True for each relevant option. Use `0, no, false` or `off` for False.

## Verification and Troubleshooting

### Verifying the Registration Node Daemon

Use the `vxrdctl vxlans` command to see the configured VNIs, the local address being used to source the VXLAN tunnel and the service node being used.

```
cumulus@leaf1$ vxrdctl vxlans
VNI      Local Addr      Svc
Node
===
=====
10        10.2.1.1
10.2.1.3
30        10.2.1.1
10.2.1.3
2000      10.2.1.1
10.2.1.3
```

```
cumulus@leaf2$ vxrdctl vxlans
VNI      Local Addr      Svc
Node
===
=====
10        10.2.1.2
10.2.1.3
30        10.2.1.2
10.2.1.3
2000      10.2.1.2
10.2.1.3
```

Use the `vxrdctl peers` command to see configured VNIs and all VTEPs (leaf switches) within the network that have them configured.

```
cumulus@leaf1$ vxrdctl peers
VNI      Peer Addrs
==        =====
10       10.2.1.1,
10.2.1.2
30       10.2.1.1,
10.2.1.2
2000     10.2.1.1,
10.2.1.2
```

```
cumulus@leaf2$ vxrdctl peers
VNI      Peer Addrs
==        =====
10       10.2.1.1,
10.2.1.2
30       10.2.1.1,
10.2.1.2
2000     10.2.1.1,
10.2.1.2
```



When head end replication mode is disabled, the command won't work.

Use the `vxrdctl peers` command to see the other VTEPs (leaf switches) and what VNIs are associated with them. This does not show anything unless you enabled head end replication mode by setting the `head_rep` option to *True*. Otherwise, replication is done by the service node.

```
cumulus@leaf2$ vxrdctl peers
Head-end replication is turned off on this device.
This command will not provide any output
```

## Verifying the Service Node Daemon

Use the `vxsndctl fdb` command to verify which VNIs belong to which VTEP (leaf switches).

```
cumulus@spine1$ vxsndctl fdb
VNI      Address      Ageout
==        =====      =====
10       10.2.1.1    82
10       10.2.1.2    77
30       10.2.1.1    82
30       10.2.1.2    77
2000     10.2.1.1    82
2000     10.2.1.2    77
```

## Verifying Traffic Flow and Checking Counters

VXLAN transit traffic information is stored in a flat file located at `/cumulus/switchd/run/stats/vxlan/all`.

```
cumulus@leaf1$ cat /cumulus/switchd/run/stats/vxlan/all
VNI                               : 10
Network In Octets                 : 1090
Network In Packets                : 8
Network Out Octets                : 1798
Network Out Packets               : 13
Total In Octets                   : 2818
Total In Packets                  : 27
Total Out Octets                  : 3144
Total Out Packets                 : 39
VN Interface                      : vni: 10, swp32s0.10
Total In Octets                   : 1728
Total In Packets                  : 19
Total Out Octets                  : 552
Total Out Packets                 : 18
VNI                               : 30
Network In Octets                 : 828
Network In Packets                : 6
Network Out Octets                : 1224
Network Out Packets               : 9
Total In Octets                   : 2374
Total In Packets                  : 23
Total Out Octets                  : 2300
Total Out Packets                 : 32
VN Interface                      : vni: 30, swp32s0.30
Total In Octets                   : 1546
Total In Packets                  : 17
Total Out Octets                  : 552
Total Out Packets                 : 17
VNI                               : 2000
Network In Octets                 : 676
Network In Packets                : 5
Network Out Octets                : 1072
Network Out Packets               : 8
Total In Octets                   : 2030
Total In Packets                  : 20
Total Out Octets                  : 2042
Total Out Packets                 : 30
VN Interface                      : vni: 2000, swp32s0.20
Total In Octets                   : 1354
Total In Packets                  : 15
Total Out Octets                  : 446
```

## Pinging to Test Connectivity

To test the connectivity across the VXLAN tunnel with an ICMP echo request (ping), make sure to ping from the server rather than the switch itself.



- As mentioned above, SVIs (switch VLAN interfaces) are not supported when using VXLAN. That is, there cannot be an IP address on the bridge that also contains a VXLAN.

Following is the IP address information used in this example configuration.

VNI	server1	server2
10	10.10.10.1	10.10.10.2
2000	10.10.20.1	10.10.20.2
30	10.10.30.1	10.10.30.2

To test connectivity between VNI 10 connected servers by pinging from server1:

```
cumulus@server1:~$ ping 10.10.10.2
PING 10.10.10.2 (10.10.10.2) 56(84) bytes of data.
64 bytes from 10.10.10.2: icmp_seq=1 ttl=64 time=3.90 ms
64 bytes from 10.10.10.2: icmp_seq=2 ttl=64 time=0.202 ms
64 bytes from 10.10.10.2: icmp_seq=3 ttl=64 time=0.195 ms
^C
--- 10.10.10.2 ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2002ms
rtt min/avg/max/mdev = 0.195/1.432/3.900/1.745 ms
cumulus@server1:~$
```

The other VNIs were also tested and can be viewed in the expanded output below.

Test connectivity between VNI-2000 connected servers by pinging from server1:

```
cumulus@server1:~$ ping 10.10.20.2
PING 10.10.20.2 (10.10.20.2) 56(84) bytes of data.
64 bytes from 10.10.20.2: icmp_seq=1 ttl=64 time=1.81 ms
64 bytes from 10.10.20.2: icmp_seq=2 ttl=64 time=0.194 ms
64 bytes from 10.10.20.2: icmp_seq=3 ttl=64 time=0.206 ms
^C
--- 10.10.20.2 ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2000ms
rtt min/avg/max/mdev = 0.194/0.739/1.819/0.763 ms
```

Test connectivity between VNI-30 connected servers by pinging from server1:

```
cumulus@server1:~$ ping 10.10.30.2
PING 10.10.30.2 (10.10.30.2) 56(84) bytes of data.
64 bytes from 10.10.30.2: icmp_seq=1 ttl=64 time=1.85 ms
64 bytes from 10.10.30.2: icmp_seq=2 ttl=64 time=0.239 ms
64 bytes from 10.10.30.2: icmp_seq=3 ttl=64 time=0.185 ms
```

```
64 bytes from 10.10.30.2: icmp_seq=4 ttl=64 time=0.212 ms
^C
--- 10.10.30.2 ping statistics ---
4 packets transmitted, 4 received, 0% packet loss, time 3000ms
rtt min/avg/max/mdev = 0.185/0.622/1.853/0.711 ms
```

## Troubleshooting with MAC Addresses

Since there is no SVI, there is no way to ping from the server to the directly attached leaf (top of rack) switch without cabling the switch to itself (see [Creating a Layer 3 Gateway](#) (see page 272) below). The easiest way to see if the server can reach the leaf switch is to check the MAC address table of the leaf switch.

First, get the MAC address of the server:

```
cumulus@server1:~$ ip addr show eth3.10 | grep ether
link/ether 90:e2:ba:55:f0:85 brd ff:ff:ff:ff:ff:ff
```

Next, check the MAC address table of the leaf switch:

```
cumulus@leaf1$ brctl showmacs br-10
port name mac addr      vlan   is local?    ageing timer
vni-10  46:c6:57:fc:1f:54  0     yes          0.00
swp32s0.10 90:e2:ba:55:f0:85  0     no           75.87
vni-10  90:e2:ba:7e:a9:c1  0     no           75.87
swp32s0.10 ec:f4:bb:fc:67:a1  0     yes          0.00
```

90:e2:ba:55:f0:85 appears in the MAC address table, which indicates that connectivity is occurring between leaf1 and server1.

## Checking the Service Node Configuration

Use `ip -d link show` to verify the service node, VNI and administrative state of a particular logical VNI interface:

```
cumulus@leaf1$ ip -d link show vni-10
35: vni-10: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue
  master br-10 state UNKNOWN mode DEFAULT
    link/ether 46:c6:57:fc:1f:54 brd ff:ff:ff:ff:ff:ff
    vxlan id 10 remote 10.2.1.3 local 10.2.1.1 srcport 32768 dstport 4789 ageing 300 svcnode 10.2.1.3
      bridge_slave
```

## ***Creating a Layer 3 Gateway***

The Trident II ASIC has a limitation because of a restriction in the hardware, where an IP address cannot be configured on the same bridge of which a VXLAN is also a part. This limitation will not exist in future ASICs. For example, the Trident II+ has the [RIOT \(Routing In/Out of Tunnels\) feature](#).

For the Trident II, this limitation means a physical cable must be attached from one port on leaf1 to another port on leaf1. One port is an L3 port while the other is a member of the bridge. For example, following the configuration above, in order for a layer 3 address to be used as the gateway for vni-10, you could configure the following on leaf1:

```
auto swp47
iface swp47
alias 12 port connected to swp48

auto swp48
iface swp48
alias gateway
address 10.10.10.3/24

auto vni-10
iface vni-10
vxlan-id 10
vxlan-local-tunnelip 10.2.1.1

auto br-10
iface br-10
bridge-ports swp47 swp32s0.10 vni-10
```

A loopback cable must be connected between swp47 and swp48 for this to work. This will be addressed in a future version of Cumulus Linux so a physical port does not need to be used for this purpose.

## ***Advanced LNV Usage***

### ***Scaling LNV by Load Balancing with Anycast***

The above configuration assumes a single service node. A single service node can quickly be overwhelmed by BUM traffic. To load balance BUM traffic across multiple service nodes, use [Anycast](#). Anycast enables BUM traffic to reach the topologically nearest service node rather than overwhelming a single service node.

### ***Enabling the Service Node Daemon on Additional Spine Switches***

In this example, spine1 already has the service node daemon enabled. Enable it on the spine2 switch with the following commands:

Edit the /etc/default/vxsnd configuration file:

```
cumulus@spine2$ sudo nano /etc/default/vxsnd
```

Change the vxsnd file by changing *no* to *yes*:

```
START=yes
```

Save and quit the text editor and reboot the vxsnd daemon:

```
cumulus@spine2$ sudo service vxsnd restart
[ ok ] Starting /usr/bin/vxsnd ....
```

## **Configuring the AnyCast Address on All Participating Service Nodes**

### **spine1**

Use a text editor to edit the network configuration:

```
cumulus@spine1$ sudo nano /etc
/network/interfaces
```

Add the 10.10.10.10/32 address to the loopback address:

```
auto lo
iface lo inet loopback
    address 10.2.1.3/32
    address 10.10.10.10/32
```

Run ifreload -a:

```
cumulus@spine1$ sudo ifreload -
a
```

Verify the IP address is configured:

```
cumulus@spine1$ ip addr show lo
1: lo: <LOOPBACK,UP,LOWER_UP>
mtu 16436 qdisc noqueue state
UNKNOWN
    link/loopback 00:00:00:00:
00:00 brd 00:00:00:00:00:00
        inet 127.0.0.1/8 scope
host lo
    inet 10.2.1.3/32 scope
global lo
    inet 10.10.10.10/32 scope
global lo
    inet6 ::1/128 scope host
        valid_lft forever
preferred_lft forever
```

### **spine2**

Use a text editor to edit the network configuration:

```
cumulus@spine2$ sudo nano /etc
/network/interfaces
```

Add the 10.10.10.10/32 address to the loopback address:

```
auto lo
iface lo inet loopback
    address 10.2.1.4/32
    address 10.10.10.10/32
```

Run ifreload -a:

```
cumulus@spine2$ sudo ifreload -
a
```

Verify the IP address is configured:

```
cumulus@spine2$ ip addr show lo
1: lo: <LOOPBACK,UP,LOWER_UP>
mtu 16436 qdisc noqueue state
UNKNOWN
    link/loopback 00:00:00:00:
00:00 brd 00:00:00:00:00:00
        inet 127.0.0.1/8 scope
host lo
    inet 10.2.1.4/32 scope
global lo
    inet 10.10.10.10/32 scope
global lo
    inet6 ::1/128 scope host
        valid_lft forever
preferred_lft forever
```

## Configuring the Service Node vxsnd.conf File

### spine1

Use a text editor to edit the network configuration:

```
cumulus@spine1$ sudo nano /etc/vxsnd.conf
```

Change the following values:

```
svcnod_ip = 10.10.10.10  
svcnod_peers = 10.2.1.4  
src_ip = 10.2.1.3
```

- i** This sets the address on which the service node listens to VXLAN messages to the configured Anycast address and sets it to sync with spine2.

Restart the vxsnd daemon:

```
cumulus@spine1$ service vxsnd  
restart  
[ ok ] Starting /usr/bin/vxsnd  
....
```

### spine2

Use a text editor to edit the network configuration:

```
cumulus@spine2$ sudo nano /etc/vxsnd.conf
```

Change the following values:

```
svcnod_ip = 10.10.10.10  
svcnod_peers = 10.2.1.3  
src_ip = 10.2.1.4
```

- i** This sets the address on which the service node listens to VXLAN messages to the configured Anycast address and sets it to sync with spine1.

Restart the vxsnd daemon:

```
cumulus@spine1$ service vxsnd  
restart  
[ ok ] Starting /usr/bin/vxsnd  
....
```

## Reconfiguring the VTEPs (Leafs) to Use the Anycast Address

### leaf1

Use a text editor to edit the network configuration:

```
cumulus@leaf1$ sudo nano /etc/network/interfaces
```

Change the `vxrd-svcnode-ip` field to the Anycast address:

```
auto lo
iface lo inet loopback
    address 10.2.1.1
    vxrd-svcnode-ip 10.10.10.10
```

Run `ifreload -a`:

```
cumulus@leaf1$ sudo ifreload -a
```

Verify the new service node is configured:

```
cumulus@leaf1$ ip -d link show vni-10
35: vni-10: <BROADCAST, MULTICAST, UP, LOWER_UP> mtu 1500 qdisc noqueue master br-10 state UNKNOWN mode DEFAULT
    link/ether 46:c6:57:fc:1f:54 brd ff:ff:ff:ff:ff:ff
        vxlan id 10 remote 10.10.10.10 local 10.2.1.1 srcport 32768 dstport 4789 ageing 300 svcnode 10.10.10.10 bridge_slave
```

```
cumulus@leaf1$ ip -d link show vni-2000
39: vni-2000: <BROADCAST, MULTICAST, UP, LOWER_UP> mtu 1500 qdisc noqueue master br-20 state UNKNOWN mode DEFAULT
```

### leaf2

Use a text editor to edit the network configuration:

```
cumulus@leaf2$ sudo nano /etc/network/interfaces
```

Change the `vxrd-svcnode-ip` field to the Anycast address:

```
auto lo
iface lo inet loopback
    address 10.2.1.2
    vxrd-svcnode-ip 10.10.10.10
```

Run `ifreload -a`:

```
cumulus@leaf2$ sudo ifreload -a
```

Verify the new service node is configured:

```
cumulus@leaf2$ ip -d link show vni-10
35: vni-10: <BROADCAST, MULTICAST, UP, LOWER_UP> mtu 1500 qdisc noqueue master br-10 state UNKNOWN mode DEFAULT
    link/ether 4e:03:a7:47:a7:9d brd ff:ff:ff:ff:ff:ff
        vxlan id 10 remote 10.10.10.10 local 10.2.1.2 srcport 32768 dstport 4789 ageing 300 svcnode 10.10.10.10 bridge_slave
```

```
cumulus@leaf2$ ip -d link show vni-2000
39: vni-2000: <BROADCAST, MULTICAST, UP, LOWER_UP> mtu 1500 qdisc noqueue master br-20 state UNKNOWN mode DEFAULT
```

```
link/ether 4a:fd:88:c3:fa:  
df brd ff:ff:ff:ff:ff:ff  
  vxlan id 2000 remote  
10.10.10.10 local 10.2.1.1  
srcport 32768 61000 dstport  
4789 ageing 300 svcnode  
10.10.10.10  
  bridge_slave
```

```
cumulus@leaf1$ ip -d link show  
vni-30  
37: vni-30: <BROADCAST,  
MULTICAST,UP,LOWER_UP> mtu  
1500 qdisc noqueue master br-  
30 state UNKNOWN mode DEFAULT  
    link/ether 3e:b3:dc:f3:bd:  
2b brd ff:ff:ff:ff:ff:ff  
      vxlan id 30 remote  
10.10.10.10 local 10.2.1.1  
srcport 32768 61000 dstport  
4789 ageing 300 svcnode  
10.10.10.10  
  bridge_slave
```

**⚠️** The svcnode 10.10.10.10 means the interface has the correct service node configured.

Use the `vxrdctl vxlans` command to check the service node:

```
cumulus@leaf1$ vxrdctl vxlans  
VNI      Local Addr      Svc  
Node  
====  ======  
=====  
 10      10.2.1.1  
10.2.1.3  
 30      10.2.1.1  
10.2.1.3  
2000     10.2.1.1  
10.2.1.3
```

```
link/ether 72:3a:bd:06:00:  
b7 brd ff:ff:ff:ff:ff:ff  
  vxlan id 2000 remote  
10.10.10.10 local 10.2.1.2  
srcport 32768 61000 dstport  
4789 ageing 300 svcnode  
10.10.10.10  
  bridge_slave
```

```
cumulus@leaf2$ ip -d link show  
vni-30  
37: vni-30: <BROADCAST,  
MULTICAST,UP,LOWER_UP> mtu  
1500 qdisc noqueue master br-  
30 state UNKNOWN mode DEFAULT  
    link/ether 22:65:3f:63:08:  
bd brd ff:ff:ff:ff:ff:ff  
      vxlan id 30 remote  
10.10.10.10 local 10.2.1.2  
srcport 32768 61000 dstport  
4789 ageing 300 svcnode  
10.10.10.10  
  bridge_slave
```

**⚠️** The svcnode 10.10.10.10 means the interface has the correct service node configured.

Use the `vxrdctl vxlans` command to check the service node:

```
cumulus@leaf2$ vxrdctl vxlans  
VNI      Local Addr      Svc  
Node  
====  ======  
=====  
 10      10.2.1.2  
10.2.1.3  
 30      10.2.1.2  
10.2.1.3  
2000     10.2.1.2  
10.2.1.3
```

## Testing Connectivity

Repeat the ping tests from the previous section. Here is the table again for reference:

VNI	server1	server2
10	10.10.10.1	10.10.10.2
2000	10.10.20.1	10.10.20.2
30	10.10.30.1	10.10.30.2

```
cumulus@server1:~$ ping 10.10.10.2
PING 10.10.10.2 (10.10.10.2) 56(84) bytes of data.
64 bytes from 10.10.10.2: icmp_seq=1 ttl=64 time=5.32 ms
64 bytes from 10.10.10.2: icmp_seq=2 ttl=64 time=0.206 ms
^C
--- 10.10.10.2 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 1001ms
rtt min/avg/max/mdev = 0.206/2.767/5.329/2.562 ms

PING 10.10.20.2 (10.10.20.2) 56(84) bytes of data.
64 bytes from 10.10.20.2: icmp_seq=1 ttl=64 time=1.64 ms
64 bytes from 10.10.20.2: icmp_seq=2 ttl=64 time=0.187 ms
^C
--- 10.10.20.2 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 1001ms
rtt min/avg/max/mdev = 0.187/0.914/1.642/0.728 ms

cumulus@server1:~$ ping 10.10.30.2
PING 10.10.30.2 (10.10.30.2) 56(84) bytes of data.
64 bytes from 10.10.30.2: icmp_seq=1 ttl=64 time=1.63 ms
64 bytes from 10.10.30.2: icmp_seq=2 ttl=64 time=0.191 ms
^C
--- 10.10.30.2 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 1001ms
rtt min/avg/max/mdev = 0.191/0.913/1.635/0.722 ms
```

## Additional Resources

Both vxsnd and vxrd have man pages in Cumulus Linux.

For vxsnd:

```
cumulus@spine1$ man vxsnd
```

For vxrd:

```
cumulus@leaf1$ man vxrd
```

## See Also

- <https://tools.ietf.org/html/rfc7348>
- <http://en.wikipedia.org/wiki/Anycast>

## LNV Full Example

Lightweight Network Virtualization (LNV) is a technique for deploying [VXLANS](#) (see page 220) without a central controller on bare metal switches. This a full example complete with diagram. Please reference the [Lightweight Network Virtualization chapter](#) (see page 252) for more detailed information. This full example uses the **recommended way** of deploying LNV, which is to use Anycast to load balance the service nodes.



LNV is a lightweight controller option. Please contact Cumulus Networks with your scale requirements and we can make sure this is the right fit for you. There are also other controller options that can work on Cumulus Linux.

## Contents

(Click to expand)

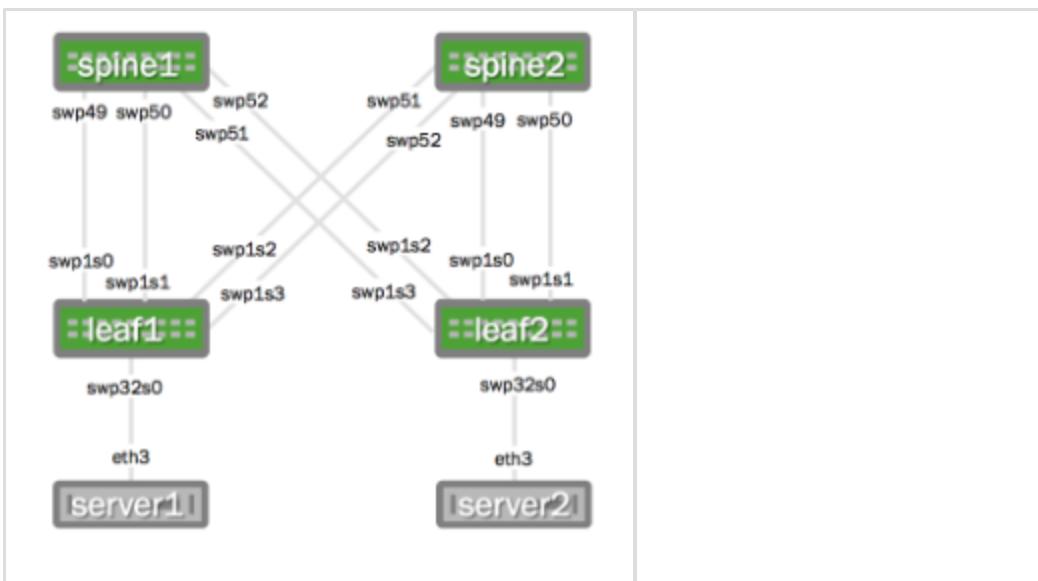
- [Contents](#) (see page 280)
- [Example LNV Configuration](#) (see page 280)
  - [Layer 3 IP Addressing](#) (see page 281)
  - [Quagga Configuration](#) (see page 283)
  - [Host Configuration](#) (see page 284)
  - [Service Node Configuration](#) (see page 286)
- [See Also](#) (see page 287)

## Example LNV Configuration

The following images illustrate the configuration:

Physical Cabling Diagram

Network Virtualization Diagram



 Want to try out configuring LNV and don't have a Cumulus Linux switch? Sign up to use the [Cumulus Workbench](#), which has this exact topology.



Feeling Overwhelmed? Come join a [Cumulus Boot Camp](#) and get instructor-led training!

## Layer 3 IP Addressing

Here is the configuration for the IP addressing information used in this example:

**spine1:** /etc/network/interfaces

```

auto lo
iface lo inet loopback
    address 10.2.1.3/32
    address 10.10.10.10/32

auto eth0
iface eth0 inet dhcp

auto swp49
iface swp49
    address 10.1.1.2/30

auto swp50
iface swp50
    address 10.1.1.6/30
  
```

**spine2:** /etc/network/interfaces

```

auto lo
iface lo inet loopback
    address 10.2.1.4/32
    address 10.10.10.10/32

auto eth0
iface eth0 inet dhcp

auto swp49
iface swp49
    address 10.1.1.18/30

auto swp50
iface swp50
    address 10.1.1.22/30
  
```

```
auto swp51
iface swp51
  address 10.1.1.50/30

auto swp52
iface swp52
  address 10.1.1.54/30
```

```
auto swp51
iface swp51
  address 10.1.1.34/30

auto swp52
iface swp52
  address 10.1.1.38/30
```

**leaf1:** /etc/network/interfaces

```
auto lo
iface lo inet loopback
  address 10.2.1.1/32
  vxrd-src-ip 10.2.1.1
  vxrd-svcnode-ip 10.10.10.10

auto eth0
iface eth0 inet dhcp

auto swp1s0
iface swp1s0
  address 10.1.1.1/30

auto swp1s1
iface swp1s1
  address 10.1.1.5/30

auto swp1s2
iface swp1s2
  address 10.1.1.33/30

auto swp1s3
iface swp1s3
  address 10.1.1.37/30

auto vni-10
iface vni-10
  vxlan-id 10
  vxlan-local-tunnelip 10.2.1.1

auto vni-2000
iface vni-2000
  vxlan-id 2000
  vxlan-local-tunnelip 10.2.1.1

auto vni-30
iface vni-30
```

**leaf2:** /etc/network/interfaces

```
auto lo
iface lo inet loopback
  address 10.2.1.2/32
  vxrd-src-ip 10.2.1.2
  vxrd-svcnode-ip 10.10.10.10

auto eth0
iface eth0 inet dhcp

auto swp1s0
iface swp1s0 inet static
  address 10.1.1.17/30

auto swp1s1
iface swp1s1 inet static
  address 10.1.1.21/30

auto swp1s2
iface swp1s2 inet static
  address 10.1.1.49/30

auto swp1s3
iface swp1s3 inet static
  address 10.1.1.53/30

auto vni-10
iface vni-10
  vxlan-id 10
  vxlan-local-tunnelip 10.2.1.2

auto vni-2000
iface vni-2000
  vxlan-id 2000
  vxlan-local-tunnelip 10.2.1.2

auto vni-30
iface vni-30
```

```

vxlan-id 30
vxlan-local-tunnelip 10.2.1.1

auto br-10
iface br-10
    bridge-ports swp32s0.10 vni-
10

auto br-20
iface br-20
    bridge-ports swp32s0.20 vni-
2000

auto br-30
iface br-30
    bridge-ports swp32s0.30 vni-
30

```

```

vxlan-id 30
vxlan-local-tunnelip 10.2.1.2

auto br-10
iface br-10
    bridge-ports swp32s0.10 vni-
10

auto br-20
iface br-20
    bridge-ports swp32s0.20 vni-
2000

auto br-30
iface br-30
    bridge-ports swp32s0.30 vni-
30

```

## Quagga Configuration

The service nodes and registration nodes must all be routable between each other. The L3 fabric on Cumulus Linux can either be [BGP \(see page 345\)](#) or [OSPF \(see page 332\)](#). In this example, OSPF is used to demonstrate full reachability.

Here is the Quagga configuration using OSPF:

**spine1:** /etc/quagga/Quagga.conf

```

interface lo
    ip ospf area 0.0.0.0
interface swp49
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0
!
interface swp50
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0
!
interface swp51
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0
!
interface swp52
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0

```

**spine2:** /etc/quagga/Quagga.conf

```

interface lo
    ip ospf area 0.0.0.0
interface swp49
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0
!
interface swp50
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0
!
interface swp51
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0
!
interface swp52
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0

```

```
!
!
!
!
!
router-id 10.2.1.3
router ospf
  ospf router-id 10.2.1.3
```

```
!
!
!
!
!
router-id 10.2.1.4
router ospf
  ospf router-id 10.2.1.4
```

**leaf1:** /etc/quagga/Quagga.conf

```
interface lo
  ip ospf area 0.0.0.0
interface swp1s0
  ip ospf network point-to-
  point
  ip ospf area 0.0.0.0
!
interface swp1s1
  ip ospf network point-to-
  point
  ip ospf area 0.0.0.0
!
interface swp1s2
  ip ospf network point-to-
  point
  ip ospf area 0.0.0.0
!
interface swp1s3
  ip ospf network point-to-
  point
  ip ospf area 0.0.0.0
!
!
!
!
!
!
router-id 10.2.1.1
router ospf
  ospf router-id 10.2.1.1
```

**leaf2:** /etc/quagga/Quagga.conf

```
interface lo
  ip ospf area 0.0.0.0
interface swp1s0
  ip ospf network point-to-
  point
  ip ospf area 0.0.0.0
!
interface swp1s1
  ip ospf network point-to-
  point
  ip ospf area 0.0.0.0
!
interface swp1s2
  ip ospf network point-to-
  point
  ip ospf area 0.0.0.0
!
interface swp1s3
  ip ospf network point-to-
  point
  ip ospf area 0.0.0.0
!
!
!
!
!
!
router-id 10.2.1.2
router ospf
  ospf router-id 10.2.1.2
```

## Host Configuration

In this example, the servers are running Ubuntu 14.04. A trunk must be mapped from server1 and server2 to the respective switch. In Ubuntu this is done with subinterfaces.

**server1**

```
auto eth3.10
iface eth3.10 inet
static
    address 10.10.10.1/24

auto eth3.20
iface eth3.20 inet
static
    address 10.10.20.1/24

auto eth3.30
iface eth3.30 inet
static
    address 10.10.30.1/24
```

**server2**

```
auto eth3.10
iface eth3.10 inet
static
    address 10.10.10.2/24

auto eth3.20
iface eth3.20 inet
static
    address 10.10.20.2/24

auto eth3.30
iface eth3.30 inet
static
    address 10.10.30.2/24
```

## Service Node Configuration

**spine1:**/etc/vxsnrd.conf

```
[common]
# Log level is one of DEBUG,
INFO, WARNING, ERROR, CRITICAL
#loglevel = INFO
# Destination for log
message. Can be a file name, 'stdout', or 'syslog'
#logdest = syslog
# log file size in bytes. Used
when logdest is a file
#logfilesize = 512000
# maximum number of log files
stored on disk. Used when
logdest is a file
#logbackupcount = 14
# The file to write the pid.
If using monit, this must
match the one
# in the vxsnrd.rc
#pidfile = /var/run/vxsnrd.pid
# The file name for the unix
domain socket used for mgmt.
#udsfile = /var/run/vxsnrd.sock
# UDP port for vxfld control
messages
#vxfld_port = 10001
# This is the address to which
registration daemons send
control messages for
# registration and/or BUM
packets for replication
svcnod_ip = 10.10.10.10
# Holdtime (in seconds) for
soft state. It is used when
sending a
# register msg to peers in
response to learning a <vni,
addr> from a
# VXLAN data pkt
#holdtime = 90
# Local IP address to bind to f
or receiving inter-vxsnrd
control traffic
src_ip = 10.2.1.3
```

**spine2:**/etc/vxsnrd.conf

```
[common]
# Log level is one of DEBUG,
INFO, WARNING, ERROR, CRITICAL
#loglevel = INFO
# Destination for log
message. Can be a file name, 'stdout', or 'syslog'
#logdest = syslog
# log file size in bytes. Used
when logdest is a file
#logfilesize = 512000
# maximum number of log files
stored on disk. Used when
logdest is a file
#logbackupcount = 14
# The file to write the pid.
If using monit, this must
match the one
# in the vxsnrd.rc
#pidfile = /var/run/vxsnrd.pid
# The file name for the unix
domain socket used for mgmt.
#udsfile = /var/run/vxsnrd.sock
# UDP port for vxfld control
messages
#vxfld_port = 10001
# This is the address to which
registration daemons send
control messages for
# registration and/or BUM
packets for replication
svcnod_ip = 10.10.10.10
# Holdtime (in seconds) for
soft state. It is used when
sending a
# register msg to peers in
response to learning a <vni,
addr> from a
# VXLAN data pkt
#holdtime = 90
# Local IP address to bind to f
or receiving inter-vxsnrd
control traffic
src_ip = 10.2.1.4
```

```
[vxsnd]
# Space separated list of IP
addresses of vxsnd to share
state with
svcnode_peers = 10.2.1.4
# When set to true, the
service node will listen for
vxlan data traffic
# Note: Use 1, yes, true, or
on, for True and 0, no, false,
or off,
# for False
#enable_vxlan_listen = true
# When set to true, the
svcnode_ip will be installed
on the loopback
# interface, and it will be
withdrawn when the vxsnd is no
longer in
# service. If set to true,
the svcnode_ip configuration
# variable must be defined.
# Note: Use 1, yes, true, or
on, for True and 0, no, false,
or off,
# for False
#install_svcnode_ip = false
# Seconds to wait before
checking the database to age
out stale entries
#age_check = 90
```

```
[vxsnd]
# Space separated list of IP
addresses of vxsnd to share
state with
svcnode_peers = 10.2.1.3
# When set to true, the
service node will listen for
vxlan data traffic
# Note: Use 1, yes, true, or
on, for True and 0, no, false,
or off,
# for False
#enable_vxlan_listen = true
# When set to true, the
svcnode_ip will be installed
on the loopback
# interface, and it will be
withdrawn when the vxsnd is no
longer in
# service. If set to true,
the svcnode_ip configuration
# variable must be defined.
# Note: Use 1, yes, true, or
on, for True and 0, no, false,
or off,
# for False
#install_svcnode_ip = false
# Seconds to wait before
checking the database to age
out stale entries
#age_check = 90
```

## See Also

- <https://tools.ietf.org/html/rfc7348>
- <http://en.wikipedia.org/wiki/Anycast>
- Detailed LNV Configuration Guide (see page 252)
- Cumulus Networks Training

## Static MAC Bindings with VXLAN

Cumulus Linux includes native Linux VXLAN kernel support.

## Contents

(Click to expand)

- [Contents \(see page 287\)](#)

- Requirements (see page 288)
- Example VXLAN Configuration (see page 288)
- Configuring the Static MAC Bindings VXLAN (see page 288)
- Troubleshooting VXLANs in Cumulus Linux (see page 292)

## Requirements

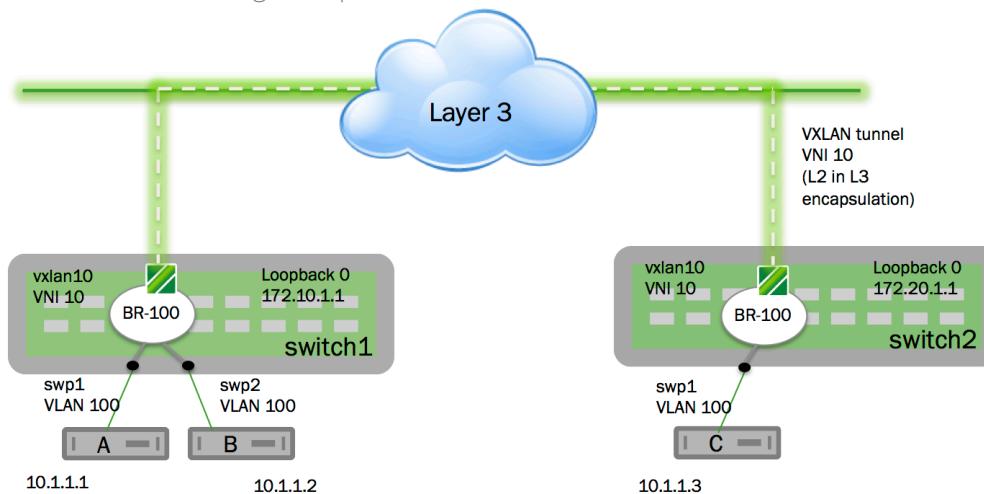
A VXLAN configuration requires a switch with a Trident II chipset running Cumulus Linux 2.0 or later.

For a basic VXLAN configuration, you should ensure that:

- The VXLAN has a network identifier (VNI); do not use 0 or 16777215 as the VNI ID, as they are reserved values under Cumulus Linux.
- The VXLAN link and local interfaces are added to bridge to create the association between port, VLAN and VXLAN instance.
- Each bridge on the switch has only one VXLAN interface. Cumulus Linux does not support more than one VXLAN link in a bridge; however a switch can have multiple bridges.

## Example VXLAN Configuration

Consider the following example:



Preconfiguring remote MAC addresses does not scale. A better solution is to use the Cumulus Networks [Lightweight Network Virtualization](#) feature, or a controller-based option like [Midokura MidoNet](#) and [OpenStack](#) or [VMware NSX](#).

## Configuring the Static MAC Bindings VXLAN

To configure the example illustrated above, edit `/etc/network/interfaces` with a text editor like `vi`, `nano` or `zile`.

Add the following configuration to the `/etc/network/interfaces` file on **switch1**:

```
auto vtep1000
```

```

iface vtep1000
    vxlan-id 1000
    vxlan-local-tunnelip 172.10.1.1

auto br-100
iface br-100
    bridge-ports swp1.100 swp2.100 vtep1000
    post-up bridge fdb add 0:00:10:00:00:0C dev vtep1000 dst 172.20.1.
1 vni 1000

```

Add the following configuration to the /etc/network/interfaces file on switch2:

```

auto vtep1000
iface vtep1000
    vxlan-id 1000
    vxlan-local-tunnelip 172.20.1.1

auto br-100
iface br-100
    bridge-ports swp1.100 swp2.100 vtep1000
    post-up bridge fdb add 00:00:10:00:00:0A dev vtep1000 dst 172.10.1
.1 vni 1000
    post-up bridge fdb add 00:00:10:00:00:0B dev vtep1000 dst 172.10.1
.1 vni 1000

```

#### Runtime Configuration (Advanced)



A runtime configuration is non-persistent, which means the configuration you create here does not persist after you reboot the switch.

In general, to configure a VXLAN in Cumulus Linux without a controller, run the following commands in a terminal connected to the switch:

1. Create a VXLAN link:

```

cumulus@switch1:~$ sudo ip link add <name> type vxlan id <vnid> local
<ip addr> [group <mcast group address>] [no] nolearning [ttl] [tos]
[dev] [port MIN MAX] [ageing <value>] [svcnode addr]

```



If you are specifying ageing, you **must** specify the service node (svcnode).

2. Add a VXLAN link to a bridge:

```
cumulus@switch1:~$ sudo brctl addif br-vxlan <name>
```

3. Install a static MAC binding to a remote tunnel IP:

```
cumulus@switch1:~$ sudo bridge fdb add <mac addr> dev <device> dst <ip
addr> vni <vni> port <port> via <device>
```

4. Show VXLAN link and FDB:

```
cumulus@switch1:~$ sudo ip -d link show
cumulus@switch1:~$ sudo bridge fdb show
```

To create a runtime configuration that matches the image above, do the following:

1. Configure hosts A and B as part of the same tenant as C (VNI 10) on switch1. Hosts A and B are part of VLAN 100. To configure the VTEP interface with VNI 10, run the following commands in a terminal connected to switch1 running Cumulus Linux:

```
cumulus@switch1:~$ sudo ip link add link swp1 name swp1.100 type vlan
id 100
cumulus@switch1:~$ sudo ip link add link swp2 name swp2.100 type vlan
id 100
cumulus@switch1:~$ sudo ip link add vtep1000 type vxlan id 10 local
172.10.1.1 nolearning
cumulus@switch1:~$ sudo ip link set swp1 up
cumulus@switch1:~$ sudo ip link set swp2 up
cumulus@switch1:~$ sudo ip link set vtep1000 up
```

2. Configure VLAN 100 and VTEP 1000 to be part of the same bridge br-100 on switch1:

```
cumulus@switch1:~$ sudo brctl addbr br-100
cumulus@switch1:~$ sudo ip link set br-100 up
cumulus@switch1:~$ sudo brctl addif br-100 swp1.100 swp2.100
cumulus@switch1:~$ sudo brctl addif br-100 vtep1000
```

3. Install a static MAC binding to a remote tunnel IP, assuming the MAC address for host C is 00:00:10:00:00:0C:

```
cumulus@switch1:~$ sudo bridge fdb add 00:00:10:00:00:0C dev vtep1000
dst 172.20.1.1
```

- Configure host C as part of the same tenant as hosts A and B on switch2:

```
cumulus@switch2:~$ sudo ip link add link swp1 name swp1.100 type vlan
id 100
cumulus@switch2:~$ sudo ip link add name vtep1000 type vxlan id 10
local 172.20.1.1 nolearning
cumulus@switch2:~$ sudo ip link set swp1 up
cumulus@switch2:~$ sudo ip link set vtep1000 up
```

- Configure VLAN 100 and VTEP 1000 to be part of the same bridge br-100 on switch2:

```
cumulus@switch2:~$ sudo brctl addbr br-100
cumulus@switch2:~$ sudo ip link set br-100 up
cumulus@switch2:~$ sudo brctl addif br-100 swp1.100
cumulus@switch2:~$ sudo brctl addif br-100 vtep1000
```

- Install a static MAC binding to a remote tunnel IP on switch2, assuming the MAC address for host A is 00:00:10:00:00:0A and the MAC address for host B is 00:00:10:00:00:0B:

```
cumulus@switch2:~$ sudo bridge fdb add 00:00:10:00:00:0A dev vtep1000
dst 172.10.1.1
cumulus@switch2:~$ sudo bridge fdb add 00:00:10:00:00:0B dev vtep1000
dst 172.10.1.1
```

- Verify the configuration on switch1, then on switch2:

```
cumulus@switch1:~$ sudo ip -d link show
cumulus@switch1:~$ sudo bridge fdb show

cumulus@switch2:~$ sudo ip -d link show
cumulus@switch2:~$ sudo bridge fdb show
```

- Set the static arp for hosts B and C on host A:

```
root@hostA:~# sudo arp -s 10.1.1.3 00:00:10:00:00:0C
```

9. Set the static arp for hosts A and C on host B:

```
root@hostB:~# sudo arp -s 10.1.1.3 00:00:10:00:00:0C
```

10. Set the static arp for hosts A and B on host C:

```
root@hostC:~# arp -s 10.1.1.1 00:00:10:00:00:0A
root@hostC:~# arp -s 10.1.1.2 00:00:10:00:00:0B
```

## Troubleshooting VXLANs in Cumulus Linux

Use the following commands to troubleshoot issues on the switch:

- `brctl show`: Verifies the VXLAN configuration in a bridge:

```
cumulus@switch:~$ sudo brctl show
bridge name      bridge id          STP enabled
interfaces
br-vxln100      8000.44383900480d    no
swp2s0.100
00
                                         swp2s1.1
                                         vxln100
```

- `bridge fdb show`: Displays the list of MAC addresses in an FDB:

```
cumulus@switch1:~$ sudo bridge fdb show
52:54:00:ae:2a:e0 dev vxln100 dst 172.16.21.150 self permanent
d2:ca:78:bb:7c:9b dev vxln100 permanent
90:e2:ba:3f:ce:34 dev swp2s1.100
90:e2:ba:3f:ce:35 dev swp2s0.100
44:38:39:00:48:0e dev swp2s1.100 permanent
44:38:39:00:48:0d dev swp2s0.100 permanent
```

- `ip -d link show`: Displays information about the VXLAN link:

```
cumulus@switch1:~$ sudo ip -d link show vxln100
71: vxln100: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue
  master br-vxln100 state UNKNOWN mode DEFAULT
    link/ether d2:ca:78:bb:7c:9b brd ff:ff:ff:ff:ff:ff
    vxlan id 100 local 172.16.20.103 port 32768 61000 nolearning
    ageing 300 svcnode 172.16.21.125
```

## VXLAN Active-Active Mode

VXLAN active-active mode allows a pair of MLAG (see page 191) switches to act as a single VTEP, providing active-active VXLAN termination for bare metal as well as virtualized workloads.

### Contents

- Contents (see page 293)
- Requirements (see page 293)
- Anycast IP Addresses (see page 293)
- Checking VXLAN Interface Configuration Consistency (see page 294)
- Configuring VXLAN Active-Active Mode (see page 294)
  - Configuring the Anycast IP Address (see page 294)
  - Configuring MLAG (see page 295)
  - Configuration LNV (see page 295)
  - Configuring STP (see page 295)
- Example VXLAN Active-Active Configuration (see page 295)
  - leaf1 Configuration (see page 295)
  - leaf2 Configuration (see page 297)
  - Quagga Configuration (see page 299)
  - LNV Configuration (see page 300)
    - leaf1 Configuration (see page 300)
    - leaf2 Configuration (see page 300)
- VXLAN PROTO\_DOWN State (see page 300)
- Caveats and Errata (see page 301)

### Requirements

- Each MLAG switch should be provisioned with a virtual IP address in the form of an anycast IP address for VXLAN datapath termination.
- All MLAG requirements (see page 193) apply for VXLAN Active-Active mode.
- LNV (see page 252) is the only supported control plane option for VXLAN active-active mode in this release. LNV can be configured for either service node replication or head-end replication.
- If STP (see page 124) is enabled on the bridge that is connected to VXLAN, then BPDU filter and BPDU guard (see page 135) should be enabled in the VXLAN interface.

### Anycast IP Addresses

The VXLAN termination address is an anycast IP address that you configure as a `c1agd` parameter (`c1agd-vxlan-anycast-ip`) under the loopback interface. `c1agd` dynamically adds and removes this address as the loopback interface address as follows:

- When the switches come up, `ifupdown2` places all VXLAN interfaces in a PROTO\_DOWN state (see page 300).

- Upon MLAG peering and a successful VXLAN interface consistency check between the switches, `c1agd` adds the anycast address as the interface address to the loopback interface. It then changes the local IP address of the VXLAN interface from a unique non-virtual IP address to an anycast virtual IP address and puts the interface in an UP state.
- If after establishing MLAG peering, the peer link goes down, then the primary switch continues to keep all VXLAN interfaces up with the anycast IP address while the secondary switch brings down all VXLAN interfaces and places them in a PROTO\_DOWN state. It also removes the anycast IP address from the loopback interface and changes the local IP address of the VXLAN interface to a unique non-virtual IP address.
- If after establishing MLAG peering, one of the switches goes down, then the other running switch continues to use the anycast IP address.
- If after establishing MLAG peering, `c1agd` is stopped, all VXLAN interfaces are put in a PROTO\_DOWN state. The anycast IP address is removed from the loopback interface and the local IP addresses of the VXLAN interfaces are changed from the anycast IP address to unique non-virtual IP addresses.
- If MLAG peering could not be established between the switches, `c1agd` brings up all the VXLAN interfaces after the reload timer expires with unique non-virtual IP addresses. This allows the VXLAN interface to be up and running on both switches even though peering is not established.

## ***Checking VXLAN Interface Configuration Consistency***

The VXLAN active-active configuration for a given VNI has to be consistent between the MLAG switches for correct traffic behavior. `c1agd` ensures that the configuration consistency is met before bringing the VXLAN interfaces operationally up. The consistency checks include:

- The anycast virtual IP address for VXLAN termination must be the same on both switches
- A VXLAN interface with the same VNI must be configured and administratively up on both switches

## ***Configuring VXLAN Active-Active Mode***

### ***Configuring the Anycast IP Address***

With MLAG peering, both switches use an anycast IP address for VXLAN encapsulation and decapsulation. This allows remote VTEPs to learn the host MAC addresses attached to the MLAG switches against one logical VTEP even though the switches independently encapsulate and decapsulate layer 2 traffic originating from the host. You configure this anycast address under the loopback interface as shown below.

```
auto lo
iface lo
  address 27.0.0.11/32
  c1agd-vxlan-anycast-ip 36.0.0.11
```



This is not a loopback interface address configuration. It's a `c1agd` parameter configuration under the loopback interface. Only `c1agd` can add or remove an anycast virtual IP address as an interface address to the loopback interface.

## Configuring MLAG

Refer to the MLAG chapter (see page 195) for configuration information.

## Configuration LNV

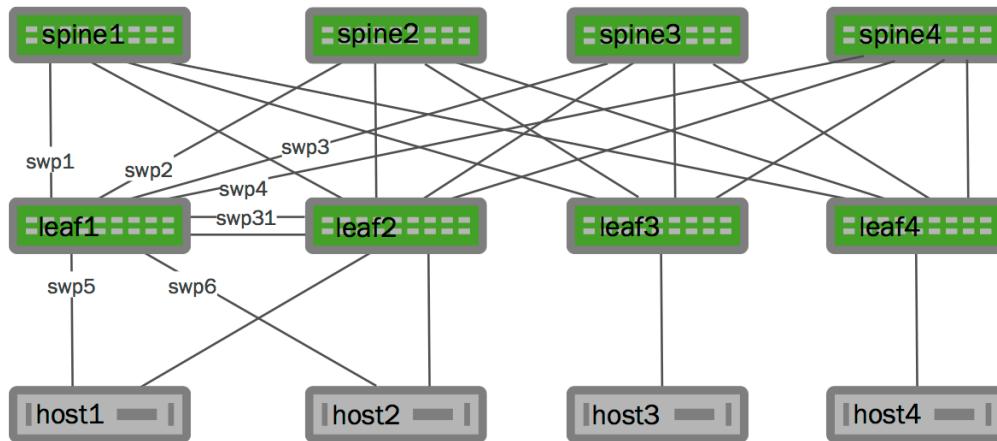
Refer to the LNV chapter (see page 252) for configuration information.

## Configuring STP

You should enable BPDU filter and BPDU guard (see page 135) in the VXLAN interfaces if STP (see page 124) is enabled in the bridge that is connected to the VXLAN.

## Example VXLAN Active-Active Configuration

The following example configures two bonds for MLAG, each with a single port, a peer link that is a bond with two member ports, and two traditional Linux bridges. It is a Clos network with spine nodes (spine1-4), 2 MLAG switches (leaf1, leaf2), 2 hosts connected to those switches and 2 standalone switches (leaf3 and leaf4) with hosts connected to them. The configuration is stored in /etc/network/interfaces on each peer switch.



Note the configuration of the local IP address in the VXLAN interfaces below. They are configured with individual IP addresses, which clagd changes to anycast upon MLAG peering.

### leaf1 Configuration

leaf1 configuration; click here to expand...

```

auto eth0
    address 10.0.0.1
    netmask 255.255.255.0
auto lo
iface lo
    address 27.0.0.11/32
    clagd-vxlan-anycast-ip 36.0.0.11
auto swp1
iface swp1

```

```
        address 10.1.1.1/30
        mtu 9050
auto swp2
iface swp2
        address 10.1.1.5/30
        mtu 9050
auto swp3
iface swp3
        address 10.1.1.33/30
        mtu 9050
auto swp4
iface swp4
        address 10.1.1.37/30
        mtu 9050
auto peerlink
iface peerlink
        bond-slaves swp31 swp32
        bond-mode 802.3ad
        bond-miimon 100
        bond-min-links 1
        bond-xmit_hash_policy layer3+4
        bond-lacp-rate 1
        mtu 9050
auto peerlink.4094
iface peerlink.4094
        address 27.0.0.11/32
        address 169.254.0.1/30
        mtu 9050
        clagd-priority 4096
        clagd-sys-mac 44:38:39:ff:ff:01
        clagd-peer-ip 169.254.0.2
        clagd-backup-ip 10.0.0.2
auto host1
iface host1
        bond-slaves swp5
        bond-mode 802.3ad
        bond-miimon 100
        bond-min-links 1
        bond-xmit_hash_policy layer3+4
        bond-lacp-rate 1
        mtu 9050
        clag-id 1
auto host2
iface host2
        bond-slaves swp6
        bond-mode 802.3ad
        bond-miimon 100
        bond-min-links 1
        bond-xmit_hash_policy layer3+4
        bond-lacp-rate 1
        mtu 9050
        clag-id 2
```

```

auto vxlan-1000
iface vxlan-1000
    vxlan-id 1000
    vxlan-local-tunnelip 27.0.0.11
    mtu 9000
auto vxlan-2000
iface vxlan-2000
    vxlan-id 2000
    vxlan-local-tunnelip 27.0.0.11
    mtu 9000
auto br1000
iface br1000
    bridge-ports host1 host2.1000 peerlink.1000 vxlan-1000
    bridge-stp on
    mstpcctl-portbpdufilter vxlan-1000=yes
    mstpcctl-bpduguard vxlan-1000=yes
    mstpcctl-portautoedge host1=yes host2.1000=yes peerlink.1000=yes
auto br2000
iface br2000
    bridge-ports host1.2000 host2 peerlink.2000 vxlan-2000
    bridge-stp on
    mstpcctl-portbpdufilter vxlan-2000=yes
    mstpcctl-bpduguard vxlan-2000=yes
    mstpcctl-portautoedge host1.2000=yes host2=yes peerlink.2000=yes

```

## ***leaf2 Configuration***

leaf2 configuration; click here to expand...

```

auto eth0
    address 10.0.0.2
    netmask 255.255.255.0
auto lo
iface lo
    address 27.0.0.12/32
    clagd-vxlan-anycast-ip 36.0.0.11
auto swp1
iface swp1
    address 10.1.1.17/30
    mtu 9050
auto swp2
iface swp2
    address 10.1.1.21/30
    mtu 9050
auto swp3
iface swp3
    address 10.1.1.49/30
    mtu 9050
auto swp4
iface swp4
    address 10.1.1.53/30
    mtu 9050

```

```
address 10.1.1.53/30
mtu 9050
auto peerlink
iface peerlink
    bond-slaves swp31 swp32
    bond-mode 802.3ad
    bond-miimon 100
    bond-min-links 1
    bond-xmit_hash_policy layer3+4
    bond-lacp-rate 1
    mtu 9050
auto peerlink.4094
iface peerlink.4094
    address 27.0.0.12/32
    address 169.254.0.2/30
    mtu 9050
    clagd-priority 4096
    clagd-sys-mac 44:38:39:ff:ff:01
    clagd-peer-ip 169.254.0.1
    clagd-backup-ip 10.0.0.1
auto host1
iface host1
    bond-slaves swp5
    bond-mode 802.3ad
    bond-miimon 100
    bond-min-links 1
    bond-xmit_hash_policy layer3+4
    bond-lacp-rate 1
    mtu 9050
    clag-id 1
auto host2
iface host2
    bond-slaves swp6
    bond-mode 802.3ad
    bond-miimon 100
    bond-min-links 1
    bond-xmit_hash_policy layer3+4
    bond-lacp-rate 1
    mtu 9050
    clag-id 2
auto vxlan-1000
iface vxlan-1000
    vxlan-id 1000
    vxlan-local-tunnelip 27.0.0.12
    mtu 9000
auto vxlan-2000
iface vxlan-2000
    vxlan-id 2000
    vxlan-local-tunnelip 27.0.0.12
    mtu 9000
auto br1000
iface br1000
```

```

bridge-ports host1 host2.1000 peerlink.1000 vxlan-1000
bridge-stp on
mstpcctl-portbpdufilter vxlan-1000=yes
mstpcctl-bpduguard vxlan-1000=yes
mstpcctl-portautoedge host1=yes host2.1000=yes peerlink.1000=yes
auto br2000
iface br2000
    bridge-ports host1.2000 host2 peerlink.2000 vxlan-2000
    bridge-stp on
    mstpcctl-portbpdufilter vxlan-2000=yes
    mstpcctl-bpduguard vxlan-2000=yes
    mstpcctl-portautoedge host1.2000=yes host2=yes peerlink.2000=yes

```

## Quagga Configuration

The layer 3 fabric can be configured using [BGP](#) (see page 345) or [OSPF](#) (see page 332). The following example uses OSPF; the configuration needed in the MLAG switches in the above specified topology is as follows:

**leaf1:** /etc/quagga/Quagga.conf

```

interface lo
    ip ospf area 0.0.0.0
interface swp1
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0
!
interface swp2
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0
!
interface swp3
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0
!
interface swp4
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0
!
!
!
!
router-id 10.2.1.1

```

**leaf2:** /etc/quagga/Quagga.conf

```

interface lo
    ip ospf area 0.0.0.0
interface swp1
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0
!
interface swp2
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0
!
interface swp3
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0
!
interface swp4
    ip ospf network point-to-
    point
    ip ospf area 0.0.0.0
!
!
!
!
router-id 10.2.1.2

```

```
router ospf
  ospf router-id 10.2.1.1
```

```
router ospf
  ospf router-id 10.2.1.2
```

## LNV Configuration

The following configuration variables should be set in leaf1 and leaf2 in `/etc/vxrd.conf`. This configuration assumes head-end replication is used to replicate BUM traffic. If service node based replication is used, then `svcnod_ip` variable has to be set with service node address. Please refer to [Configuring the Registration Node \(see page 265\)](#) for setting that variable.

### *leaf1 Configuration*

```
# Local IP address to bind to for receiving control traffic from the snd
src_ip = 27.0.0.11

# Enable self replication
# Note: Use true, or on, for True and 0, no, false, or off,
# for False
head_rep = true
```

### *leaf2 Configuration*

```
# Local IP address to bind to for receiving control traffic from the snd
src_ip = 27.0.0.12

# Enable self replication
# Note: Use true, or on, for True and 0, no, false, or off,
# for False
head_rep = true
```

## VXLAN PROTO\_DOWN State

Similar to a bond interface, if MLAG detects a problem that could result in connectivity issues such as traffic black-holing or a network meltdown if the link carrier was left in an UP state, it can put VXLAN interface into a [PROTO\\_DOWN state \(see page \)](#). Such connectivity issues include:

- When the peer link goes down but the peer switch is up (that is, the backup link is active).
- When an MLAG-enabled node is booted or rebooted, VXLAN interfaces are placed in a PROTO\_DOWN state until the node establishes a connection to its peer switch, detects existence of corresponding VXLAN interfaces in the peer switch, or five minutes have elapsed.

- If the anycast address is not configured or if it is not the same in both MLAG switches, the VXLAN interfaces are placed into a PROTO\_DOWN state.
- A configuration mismatch between the MLAG switches, such as the VXLAN interface is configured on just one of the switches or if the interface is shut down on one of the switches, then the VXLAN interface is placed into a PROTO\_DOWN state on the secondary switch.

You can use the `clagctl` command to check if any VXLAN devices are in a PROTO\_DOWN state. As shown below, VXLAN devices are kept in a PROTO\_DOWN state due to the missing anycast configuration.

```

cumulus@switch$ clagctl
The peer is alive
    Our Priority, ID, and Role: 4096 c4:54:44:bd:01:71 primary
    Peer Priority, ID, and Role: 8192 00:02:00:00:00:36 secondary
        Peer Interface and IP: peerlink.4094 169.254.0.2
                                Backup IP: 10.0.0.2 (active)
                                System MAC: 44:38:39:ff:ff:01

CLAG Interfaces
Our Interface      Peer Interface      CLAG Id      Conflicts
Proto-Down Reason
-----  -----  -----  -----
-----  -----
host1          host2          1          -
-
host1          host2          2          -
-
vxlan-1000      -          -          -
vxlan-single,no-anycast-ip
vxlan-2000      -          -          -
vxlan-single,no-anycast-ip

```

## Caveats and Errata

- VLAN-aware bridge mode (see page 182) is not supported for VXLAN active-active mode in this release.
- The VLAN used for the peer link layer 3 subinterface should not be reused for any other interface in the system. It is recommended to use a high VLAN ID value. Read more about the [range of VLAN IDs you can use \(see page 191\)](#).
- Active-active mode works only with LNV in this release. Integration with controller-based VXLANs such as VMware NSX and Midokura MidoNet will be supported in the future.

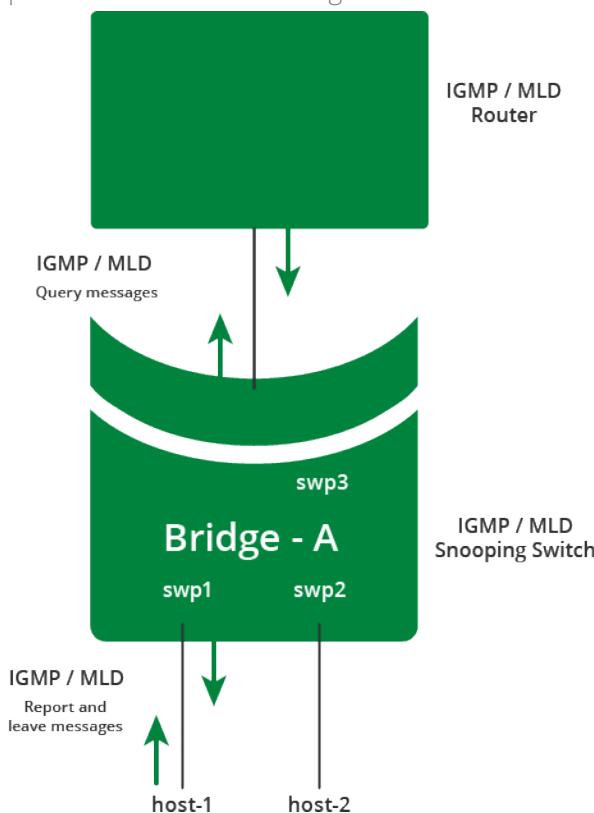
## IGMP and MLD Snooping

IGMP (Internet Group Management Protocol) and MLD (Multicast Listener Discovery) snooping functionality is implemented in the bridge driver in the kernel. IGMP snooping processes IGMP v1/v2/v3 reports received on a bridge port in a bridge to identify the hosts which would like to receive multicast traffic destined to that group.

When an IGMPv2 leave message is received, a group specific query is sent to identify if there are any other hosts interested in that group, before the group is deleted.

An IGMP query message received on a port is used to identify the port that is connected to a router and is interested in receiving multicast traffic.

MLD snooping processes MLD v1/v2 reports, queries and v1 done messages for IPv6 groups. If IGMP or MLD snooping is disabled, multicast traffic will be flooded to all the bridge ports in the bridge. The multicast group IP address is mapped to a multicast MAC address and a forwarding entry is created with a list of ports interested in receiving multicast traffic destined to that group.



## Contents

(Click to expand)

- [Contents \(see page 302\)](#)
- [Commands \(see page 303\)](#)
- [Creating a Bridge and Configuring IGMP/MLD Snooping \(see page 303\)](#)
- [Configuration Files \(see page 307\)](#)
- [Man Pages \(see page 307\)](#)

- Useful Links (see page 308)

## **Commands**

- brctl
- bridge

## ***Creating a Bridge and Configuring IGMP/MLD Snooping***

You need to set a number of parameters for IGMP and MLD snooping, but the setting to enable it is `bridge-mcsnoop 1`. The following configuration in `/etc/network/interfaces` is for the example bridge above. For an explanation of the relevant parameters, see the `bridge-utils-interfaces` man page:

```

auto br0
iface br0
    bridge-vlan-aware yes
    bridge-ports swp1 swp2 swp3
    bridge-vids 100 200
    bridge-pvid 1
    bridge-stp on
    bridge-mclmc 2
    bridge-mcroouter 1
    bridge-mcsnoop 1
    bridge-mcsqc 2
    bridge-mcqifaddr 0
    bridge-mcquerier 0
    bridge-hashel 4096
    bridge-hashmax 4096
    bridge-mclmi 1
    bridge-mcmi 260
    bridge-mcqpi 255
    bridge-mcqci 125
    bridge-mcqri 10
    bridge-mcsqi 31

auto swp1
iface swp1
    bridge-vids 100
    bridge-portmcroouter 1
    bridge-portmcfl 0

auto swp2
iface swp2
    bridge-vids 200

```

```
bridge-portmcrouter 1  
bridge-portmcfl 0  
  
auto swp3  
iface swp3  
    bridge-access 100
```

## Runtime Configuration (Advanced)

- !** A runtime configuration is non-persistent, which means the configuration you create here does not persist after you reboot the switch.

To enable snooping at runtime, use the `brctl` command. Create a bridge and add bridge ports to the bridge. IGMP and MLD snooping are enabled by default on the bridge:

```
cumulus@switch:~$ sudo brctl addbr br0  
cumulus@switch:~$ sudo brctl addif br0 swp1 swp2 swp3  
cumulus@switch:~$ sudo ifconfig br0 up
```

To get the IGMP/MLD snooping bridge state, use:

```
cumulus@switch:~# sudo brctl showstp br0  
br0  
bridge id          8000.7072cf8c272c  
designated root    8000.7072cf8c272c  
root port          0                      path cost      0  
max age           20.00                  bridge max age  
20.00  
hello time         2.00                  bridge hello time  
2.00  
forward delay      15.00                 bridge forward delay  
15.00  
ageing time        300.00                tcn timer  
hello timer        0.00                  gc timer  
0.00  
topology change timer 0.00  
263.70  
hash elasticity    4096                 hash max       4096  
mc last member count 2                   mc init query count 2  
mc router          1                   mc snooping     1  
mc last member timer 1.00  
260.00  
mc membership timer
```

mc querier timer	255.00	mc query interval	
125.00			
mc response interval	10.00	mc init query interval	
31.25			
mc querier flags	0	mc query ifaddr	0
 swp1 (1)			
port id	8001	state	
forwarding			
designated root	8000.7072cf8c272c	path cost	2
designated bridge	8000.7072cf8c272c	message age timer	
0.00			
designated port	8001	forward delay timer	
0.00			
designated cost	0	hold timer	
0.00			
mc router flags	1	mc fast leave	0
 swp2 (2)			
port id	8002	state	
forwarding			
designated root	8000.7072cf8c272c	path cost	2
designated bridge	8000.7072cf8c272c	message age timer	
0.00			
designated port	8002	forward delay timer	
0.00			
designated cost	0	hold timer	
0.00			
mc router flags	1	mc fast leave	0
 swp3 (3)			
port id	8003	state	
forwarding			
designated root	8000.7072cf8c272c	path cost	2
designated bridge	8000.7072cf8c272c	message age timer	
0.00			
designated port	8003	forward delay timer	
8.98			
designated cost	0	hold timer	
0.00			
mc router flags	1	mc fast leave	0

To get the groups and bridge port state, use `bridge mdb show` command. To display router ports and group information use `bridge -d mdb show` command:

```
cumulus@switch:~# sudo bridge -d mdb show
dev br0 port swp2 grp 234.10.10.10 temp
dev br0 port swp1 grp 238.39.20.86 permanent
dev br0 port swp1 grp 234.1.1.1 temp
dev br0 port swp2 grp ff1a::9 permanent
router ports on br0: swp3

cumulus@switch:~# sudo bridge mdb show
dev br0 port swp2 grp 234.10.10.10 temp
dev br0 port swp1 grp 238.39.20.86 permanent
dev br0 port swp1 grp 234.1.1.1 temp
dev br0 port swp2 grp ff1a::9 permanent
```

To disable IGMP and MLD snooping, use:

```
cumulus@switch:~$ sudo brctl setmcsnoop br0 0
```

## ***Configuring IGMP/MLD Snooping Parameters***

For an explanation of these parameters, see the `brctl` and `bridge-utils-interfaces` man pages:

```
cumulus@switch:~$ sudo brctl setmclmc br0 2
cumulus@switch:~$ sudo brctl setmcrouter br0 1
cumulus@switch:~$ sudo brctl setmcsrc br0 2
cumulus@switch:~$ sudo brctl sethashel br0 4096
cumulus@switch:~$ sudo brctl sethashmax br0 4096
cumulus@switch:~$ sudo brctl setmclmi br0 1
cumulus@switch:~$ sudo brctl setmcmi br0 260
cumulus@switch:~$ sudo brctl setmcqpi br0 255
cumulus@switch:~$ sudo brctl setmcqi br0 125
cumulus@switch:~$ sudo brctl setmcqri br0 10
cumulus@switch:~$ sudo brctl setmsqi br0 31
```

## ***Querier and Fast Leave Configuration***

If there is no multicast router in the VLAN, the IGMP/MLD snooping querier can be configured to generate query messages.

To send queries with a non-zero IP address, configure an IP address on the bridge device, then set `setmcqifaddr` to 1:

```
cumulus@switch:~# sudo brctl setmcquerier br0 1  
cumulus@switch:~$ sudo brctl setmcqifaddr br0 1
```

If only one host is attached to each host port, fast leave can be configured on that port. When a leave message is received on that port, no query messages will be sent and the group will be deleted immediately:

```
cumulus@switch:~# sudo brctl setportmcfl br0 swp1 1
```

## Static Group and Router Port Configuration

To configure static permanent multicast group on a port, use:

```
cumulus@switch:~# sudo bridge mdb add dev br0 port swp2 grp ff1a::9  
permanent  
cumulus@switch:~# sudo bridge mdb add dev br0 port swp1 grp 238.39.20.86  
permanent
```

A static temporary multicast group can also be configured on a port, which would be deleted after the membership timer expires, if no report is received on that port:

```
cumulus@switch:~# sudo bridge mdb add dev br0 port swp1 grp 238.39.20.86  
temp
```

To configure a static router port, use:

```
cumulus@switch:~# sudo brctl setportmcrouter br0 swp3 2
```

## Configuration Files

- /etc/network/interfaces

## Man Pages

- brctl(8)
- bridge(8)
- bridge-utils-interfaces(5)

## ***Useful Links***

- <http://www.linuxfoundation.org/collaborate/workgroups/networking/bridge#Snooping>
- <https://tools.ietf.org/html/rfc4541>
- [http://en.wikipedia.org/wiki/IGMP\\_snooping](http://en.wikipedia.org/wiki/IGMP_snooping)
- <http://tools.ietf.org/rfc/rfc2236.txt>
- <http://tools.ietf.org/html/rfc3376>
- <http://tools.ietf.org/search/rfc2710>
- <http://tools.ietf.org/html/rfc3810>

# Layer 3 Features

## Routing

This chapter discusses routing on switches running Cumulus Linux.

### Contents

(Click to expand)

- [Contents \(see page 309\)](#)
- [Commands \(see page 309\)](#)
- [Static Routing via ip route \(see page 309\)](#)
  - [Persistently Adding a Static Route \(see page 311\)](#)
- [Static Routing via quagga \(see page 311\)](#)
  - [Persistent Configuration \(see page 313\)](#)
- [Supported Route Table Entries \(see page 313\)](#)
- [Configuration Files \(see page 314\)](#)
- [Useful Links \(see page 314\)](#)
- [Caveats and Errata \(see page 314\)](#)

### Commands

- [ip route](#)

### Static Routing via ip route

The `ip route` command allows manipulating the kernel routing table directly from the Linux shell. See `man ip(8)` for details. `quagga` monitors the kernel routing table changes and updates its own routing table accordingly.

To display the routing table:

```
cumulus@switch:~$ ip route show
default via 10.0.1.2 dev eth0
10.0.1.0/24 dev eth0  proto kernel  scope link  src 10.0.1.52
192.0.2.0/24 dev swp1  proto kernel  scope link  src 192.0.2.12
192.0.2.10/24 via 192.0.2.1 dev swp1  proto zebra  metric 20
192.0.2.20/24  proto zebra  metric 20
    nexthop via 192.0.2.1  dev swp1 weight 1
    nexthop via 192.0.2.2  dev swp2 weight 1
192.0.2.30/24 via 192.0.2.1 dev swp1  proto zebra  metric 20
192.0.2.40/24 dev swp2  proto kernel  scope link  src 192.0.2.42
192.0.2.50/24 via 192.0.2.2 dev swp2  proto zebra  metric 20
```

```
192.0.2.60/24 via 192.0.2.2 dev swp2 proto zebra metric 20
192.0.2.70/24 proto zebra metric 30
    nexthop via 192.0.2.1 dev swp1 weight 1
    nexthop via 192.0.2.2 dev swp2 weight 1
198.51.100.0/24 dev swp3 proto kernel scope link src 198.51.100.1
198.51.100.10/24 dev swp4 proto kernel scope link src 198.51.100.11
198.51.100.20/24 dev br0 proto kernel scope link src 198.51.100.21
```

To add a static route (does not persist across reboots):

```
cumulus@switch:~$ sudo ip route add 203.0.113.0/24 via 198.51.100.2
cumulus@switch:~$ ip route
default via 10.0.1.2 dev eth0
10.0.1.0/24 dev eth0 proto kernel scope link src 10.0.1.52
192.0.2.0/24 dev swp1 proto kernel scope link src 192.0.2.12
192.0.2.10/24 via 192.0.2.1 dev swp1 proto zebra metric 20
192.0.2.20/24 proto zebra metric 20
    nexthop via 192.0.2.1 dev swp1 weight 1
    nexthop via 192.0.2.2 dev swp2 weight 1
192.0.2.30/24 via 192.0.2.1 dev swp1 proto zebra metric 20
192.0.2.40/24 dev swp2 proto kernel scope link src 192.0.2.42
192.0.2.50/24 via 192.0.2.2 dev swp2 proto zebra metric 20
192.0.2.60/24 via 192.0.2.2 dev swp2 proto zebra metric 20
192.0.2.70/24 proto zebra metric 30
    nexthop via 192.0.2.1 dev swp1 weight 1
    nexthop via 192.0.2.2 dev swp2 weight 1
198.51.100.0/24 dev swp3 proto kernel scope link src 198.51.100.1
198.51.100.10/24 dev swp4 proto kernel scope link src 198.51.100.11
198.51.100.20/24 dev br0 proto kernel scope link src 198.51.100.21
203.0.113.0/24 via 198.51.100.2 dev swp3
```

To delete a static route (does not persist across reboots):

```
cumulus@switch:~$ sudo ip route del 203.0.113.0/24
cumulus@switch:~$ ip route
default via 10.0.1.2 dev eth0
10.0.1.0/24 dev eth0 proto kernel scope link src 10.0.1.52
192.0.2.0/24 dev swp1 proto kernel scope link src 192.0.2.12
192.0.2.10/24 via 192.0.2.1 dev swp1 proto zebra metric 20
192.0.2.20/24 proto zebra metric 20
    nexthop via 192.0.2.1 dev swp1 weight 1
    nexthop via 192.0.2.2 dev swp2 weight 1
```

```

192.0.2.30/24 via 192.0.2.1 dev swp1 proto zebra metric 20
192.0.2.40/24 dev swp2 proto kernel scope link src 192.0.2.42
192.0.2.50/24 via 192.0.2.2 dev swp2 proto zebra metric 20
192.0.2.60/24 via 192.0.2.2 dev swp2 proto zebra metric 20
192.0.2.70/24 proto zebra metric 30
    nexthop via 192.0.2.1 dev swp1 weight 1
    nexthop via 192.0.2.2 dev swp2 weight 1
198.51.100.0/24 dev swp3 proto kernel scope link src 198.51.100.1
198.51.100.10/24 dev swp4 proto kernel scope link src 198.51.100.11
198.51.100.20/24 dev br0 proto kernel scope link src 198.51.100.21

```

## Persistently Adding a Static Route

A static route can be persistently added by adding `up ip route add ..` into `/etc/network/interfaces`. For example:

```

cumulus@switch:~$ cat /etc/network/interfaces
# This file describes the network interfaces available on your system
# and how to activate them. For more information, see interfaces(5).

# The loopback network interface
auto lo
iface lo inet loopback

auto swp3
iface swp3
    address 198.51.100.1/24
    up ip route add 203.0.113.0/24 via 198.51.100.2

```



Notice the simpler configuration of `swp3` due to `ifupdown2`. For more information, see [Configuring Network Interfaces with ifupdown \(see page 94\)](#).

## Static Routing via quagga

Static routes can also be managed via the `quagga` CLI. The routes are added to the `quagga` routing table, and then will be updated into the kernel routing table as well.

To add a static route (does not persist across reboot):

```

cumulus@switch:~$ sudo vtysh

```

```
Hello, this is Quagga (version 0.99.21).
Copyright 1996-2005 Kunihiro Ishiguro, et al.

switch# conf t
switch(config)# ip route 203.0.113.0/24 198.51.100.2
switch(config)# end
switch# show ip route
Codes: K - kernel route, C - connected, S - static, R - RIP,
       O - OSPF, I - IS-IS, B - BGP, A - Babel,
       > - selected route, * - FIB route

K>* 0.0.0.0/0 via 10.0.1.2, eth0
C>* 10.0.1.0/24 is directly connected, eth0
O 192.0.2.0/24 [110/10] is directly connected, swp1, 00:13:25
C>* 192.0.2.0/24 is directly connected, swp1
O>* 192.0.2.10/24 [110/20] via 192.0.2.1, swp1, 00:13:09
O>* 192.0.2.20/24 [110/20] via 192.0.2.1, swp1, 00:13:09
*           via 192.0.2.41, swp2, 00:13:09
O>* 192.0.2.30/24 [110/20] via 192.0.2.1, swp1, 00:13:09
O 192.0.2.40/24 [110/10] is directly connected, swp2, 00:13:25
C>* 192.0.2.40/24 is directly connected, swp2
O>* 192.0.2.50/24 [110/20] via 192.0.2.41, swp2, 00:13:09
O>* 192.0.2.60/24 [110/20] via 192.0.2.41, swp2, 00:13:09
O>* 192.0.2.70/24 [110/30] via 192.0.2.1, swp1, 00:13:09
*           via 192.0.2.41, swp2, 00:13:09
O 198.51.100.0/24 [110/10] is directly connected, swp3, 00:13:22
C>* 198.51.100.0/24 is directly connected, swp3
O 198.51.100.10/24 [110/10] is directly connected, swp4, 00:13:22
C>* 198.51.100.10/24 is directly connected, swp4
O 198.51.100.20/24 [110/10] is directly connected, br0, 00:13:22
C>* 198.51.100.20/24 is directly connected, br0
S>* 203.0.113.0/24 [1/0] via 198.51.100.2, swp3
C>* 127.0.0.0/8 is directly connected, lo
```

To delete a static route (does not persist across reboot):

```
cumulus@switch:~$ sudo vtysh

Hello, this is Quagga (version 0.99.21).
Copyright 1996-2005 Kunihiro Ishiguro, et al.

switch# conf t
switch(config)# no ip route 203.0.113.0/24 198.51.100.2
```

```

switch(config)# end
switch# show ip route
Codes: K - kernel route, C - connected, S - static, R - RIP,
       O - OSPF, I - IS-IS, B - BGP, A - Babel,
       > - selected route, * - FIB route

K>* 0.0.0.0/0 via 10.0.1.2, eth0
C>* 10.0.1.0/24 is directly connected, eth0
O 192.0.2.0/24 [110/10] is directly connected, swp1, 00:13:55
C>* 192.0.2.0/24 is directly connected, swp1
O>* 192.0.2.10/24 [110/20] via 11.0.0.1, swp1, 00:13:39
O>* 192.0.2.20/24 [110/20] via 11.0.0.1, swp1, 00:13:39
*           via 11.0.4.1, swp2, 00:13:39
O>* 192.0.2.30/24 [110/20] via 11.0.0.1, swp1, 00:13:39
O 192.0.2.40/24 [110/10] is directly connected, swp2, 00:13:55
C>* 192.0.2.40/24 is directly connected, swp2
O>* 192.0.2.50/24 [110/20] via 11.0.4.1, swp2, 00:13:39
O>* 192.0.2.60/24 [110/20] via 11.0.4.1, swp2, 00:13:39
O>* 192.0.2.70/24 [110/30] via 11.0.0.1, swp1, 00:13:39
*           via 11.0.4.1, swp2, 00:13:39
O 198.51.100.0/24 [110/10] is directly connected, swp3, 00:13:52
C>* 198.51.100.0/24 is directly connected, swp3
O 198.51.100.10/24 [110/10] is directly connected, swp4, 00:13:52
C>* 198.51.100.10/24 is directly connected, swp4
O 198.51.100.20/24 [110/10] is directly connected, br0, 00:13:52
C>* 198.51.100.20/24 is directly connected, br0
C>* 127.0.0.0/8 is directly connected, lo
switch#

```

## Persistent Configuration

From the quagga CLI, the running configuration can be saved so it persists between reboots:

```

switch# write mem
Configuration saved to /etc/quagga/zebra.conf
switch# end

```

## Supported Route Table Entries

Cumulus Linux supports different numbers of route entries, depending upon your switch platform (Trident, Trident+, or Trident II; see the [HCL](#)) and whether the routes are IPv4 or IPv6.

In addition, switches on the Trident II platform are configured to manage route table entries using Algorithm Longest Prefix Match (ALPM). In ALPM mode, the hardware can store significantly more route entries.

Following are the number of route supported on Trident II switches with ALPM:

- 32K IPv4 routes
- 16K IPv6 routes
- 32K total routes (both IPv4 and IPv6)

Following are the number of route supported on Trident and Trident+ switches:

- 16K IPv4 routes
- 8K IPv6 routes
- 16K total routes (both IPv4 and IPv6)

## Configuration Files

- /etc/network/interfaces
- /etc/quagga/zebra.conf

## Useful Links

- <http://linux-ip.net/html/tools-ip-route.html>
- <http://www.nongnu.org/quagga/docs/docs-info.html#Static-Route-Commands>

## Caveats and Errata

- Static routes added via `quagga` can be deleted via Linux shell. This operation, while possible, should be avoided. Routes added by `quagga` should only be deleted by `quagga`, otherwise `quagga` might not be able to clean up all its internal state completely and incorrect routing can occur as a result.

# Introduction to Routing Protocols

This chapter discusses the various routing protocols, and how to configure them.

## Contents

(Click to expand)

- Contents (see page 314)
- Defining Routing Protocols (see page 314)
- Configuring Routing Protocols (see page 315)
- Protocol Tuning (see page 315)
- Configuration Files (see page 316)

## Defining Routing Protocols

A *routing protocol* dynamically computes reachability between various end points. This enables communication to work around link and node failures, and additions and withdrawals of various addresses.

*IP routing protocols* are typically distributed; that is, an instance of the routing protocol runs on each of the routers in a network.



Cumulus Linux does **not** support running multiple instances of the same protocol on a router.

*Distributed routing protocols* compute reachability between end points by disseminating relevant information and running a routing algorithm on this information to determine the routes to each end station. To scale the amount of information that needs to be exchanged, routes are computed on address prefixes rather than on every end point address.

## Configuring Routing Protocols

A routing protocol needs to know three pieces of information, at a minimum:

- Who am I (my identity)
- To whom to disseminate information
- What to disseminate

Most routing protocols use the concept of a router ID to identify a node. Different routing protocols answer the last two questions differently.

The way they answer these questions affects the network design and thereby configuration. For example, in a link-state protocol such as OSPF (see [Open Shortest Path First \(OSPF\) Protocol \(see page 332\)](#)) or IS-IS, complete local information (links and attached address prefixes) about a node is disseminated to every other node in the network. Since the state that a node has to keep grows rapidly in such a case, link-state protocols typically limit the number of nodes that communicate this way. They allow for bigger networks to be built by breaking up a network into a set of smaller subnetworks (which are called areas or levels), and by advertising summarized information about an area to other areas.

Besides the two critical pieces of information mentioned above, protocols have other parameters that can be configured. These are usually specific to each protocol.

## Protocol Tuning

Most protocols provide certain tunable parameters that are specific to convergence during changes.

Wikipedia defines [convergence](#) as the “state of a set of routers that have the same topological information about the network in which they operate”. It is imperative that the routers in a network have the same topological state for the proper functioning of a network. Without this, traffic can be blackholed, and thus not reach its destination. It is normal for different routers to have differing topological states during changes, but this difference should vanish as the routers exchange information about the change and recompute the forwarding paths. Different protocols converge at different speeds in the presence of changes.

A key factor that governs how quickly a routing protocol converges is the time it takes to detect the change. For example, how quickly can a routing protocol be expected to act when there is a link failure. Routing protocols classify changes into two kinds: hard changes such as link failures, and soft changes such as a peer dying silently. They’re classified differently because protocols provide different mechanisms for dealing with these failures.

It is important to configure the protocols to be notified immediately on link changes. This is also true when a node goes down, causing all of its links to go down.

Even if a link doesn’t fail, a routing peer can crash. This causes that router to usually delete the routes it has computed or worse, it makes that router impervious to changes in the network, causing it to go out of sync with the other routers in the network because it no longer shares the same topological information as its peers.

The most common way to detect a protocol peer dying is to detect the absence of a heartbeat. All routing protocols send a heartbeat (or “hello”) packet periodically. When a node does not see a consecutive set of these hello packets from a peer, it declares its peer dead and informs other routers in the network about this. The period of each heartbeat and the number of heartbeats that need to be missed before a peer is declared dead are two popular configurable parameters.

If you configure these timers very low, the network can quickly descend into instability under stressful conditions when a router is not able to keep sending the heartbeats quickly as it is busy computing routing state; or the traffic is so much that the hellos get lost. Alternately, configuring this timer to very high values also causes blackholing of communication because it takes much longer to detect peer failures. Usually, the default values initialized within each protocol are good enough for most networks. Cumulus Networks recommends you do not adjust these settings.

## Configuration Files

- /etc/quagga/daemons

## Network Topology

In computer networks, *topology* refers to the structure of interconnecting various nodes. Some commonly used topologies in networks are star, hub and spoke, leaf and spine, and broadcast.

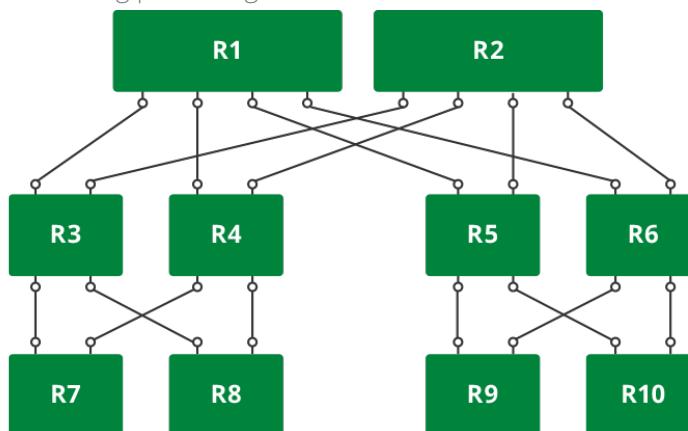
## Contents

(Click to expand)

- Contents (see page 316)
- Clos Topologies (see page 316)
- Over-Subscribed and Non-Blocking Configurations (see page 317)
- Containing the Failure Domain (see page 317)
- Load Balancing (see page 318)

## Clos Topologies

In the vast majority of modern data centers, **Clos or fat tree topology** is very popular. This topology is shown in the figure below. It is also commonly referred to as leaf-spine topology. We shall use this topology throughout the routing protocol guide.



This topology allows the building of networks of varying size using nodes of different port counts and/or by increasing the tiers. The picture above is a three-tiered Clos network. We number the tiers from the bottom to the top. Thus, in the picture, the lowermost layer is called tier 1 and the topmost tier is called tier 3.

The number of end stations (such as servers) that can be attached to such a network is determined by a very simple mathematical formula.

In a 2-tier network, if each node is made up of  $m$  ports, then the total number of end stations that can be connected is  $m^2/2$ . In more general terms, if tier-1 nodes are  $m$ -port nodes and tier-2 nodes are  $n$ -port nodes, then the total number of end stations that can be connected are  $(m \cdot n)/2$ . In a three tier network, where tier-3 nodes are  $o$ -port nodes, the total number of end stations that can be connected are  $(m \cdot n \cdot o)/2^{(\text{number of tiers}-1)}$ .

Let's consider some practical examples. In many data centers, it is typical to connect 40 servers to a top-of-rack (ToR) switch. The ToRs are all connected via a set of spine switches. If a ToR switch has 64 ports, then after hooking up 40 ports to the servers, the remaining 24 ports can be hooked up to 24 spine switches of the same link speed or to a smaller number of higher link speed switches. For example, if the servers are all hooked up as 10GE links, then the ToRs can connect to the spine switches via 40G links. So, instead of connecting to 24 spine switches with 10G links, the ToRs can connect to 6 spine switches with each link being 40G. If the spine switches are also 64-port switches, then the total number of end stations that can be connected is 2560 ( $40 \cdot 64$ ) stations.

In a three tier network of 64-port switches, the total number of servers that can be connected are  $(40 \cdot 64 \cdot 64)/2 = 81920$ . As you can see, this kind of topology can serve quite a large network with three tiers.

## ***Over-Subscribed and Non-Blocking Configurations***

In the above example, the network is *over-subscribed*; that is, 400G of bandwidth from end stations (40 servers \* 10GE links) is serviced by only 240G of inter-rack bandwidth. The over-subscription ratio is 0.6 ( $240/400$ ).

This can lead to congestion in the network and hot spots. Instead, if network operators connected 32 servers per rack, then 32 ports are left to be connected to spine switches. Now, the network is said to be *rerrangably non-blocking*. Now any server in a rack can talk to any other server in any other rack without necessarily blocking traffic between other servers.

In such a network, the total number of servers that can be connected are  $(64 \cdot 64)/2 = 2048$ . Similarly, a three-tier version of the same can serve up to  $(64 \cdot 64 \cdot 64)/4 = 65536$  servers.

## ***Containing the Failure Domain***

Traditional data centers were built using just two spine switches. This means that if one of those switches fails, the network bandwidth is cut in half, thereby greatly increasing network congestion and adversely affecting many applications. To avoid this, vendors typically try and make the spine switches resilient to failures by providing such features as dual control line cards and attempting to make the software highly available. However, as Douglas Adams famously noted, “>>>”. In many cases, HA is among the top two or three causes of software failure (and thereby switch failure).

To support a fairly large network with just two spine switches also means that these switches have a large port count. This can make the switches quite expensive.

If the number of spine switches were to be merely doubled, the effect of a single switch failure is halved. With 8 spine switches, the effect of a single switch failure only causes a 12% reduction in available bandwidth.

So, in modern data centers, people build networks with anywhere from 4 to 32 spine switches.

## Load Balancing

In a Clos network, traffic is load balanced across the multiple links using equal cost multi-pathing (ECMP).

Routing algorithms compute shortest paths between two end stations where shortest is typically the lowest path cost. Each link is assigned a metric or cost. By default, a link's cost is a function of the link speed. The higher the link speed, the lower its cost. A 10G link has a higher cost than a 40G or 100G link, but a lower cost than a 1G link. Thus, the link cost is a measure of its traffic carrying capacity.

In the modern data center, the links between tiers of the network are homogeneous; that is, they have the same characteristics (same speed and therefore link cost) as the other links. As a result, the first hop router can pick any of the spine switches to forward a packet to its destination (assuming that there is no link failure between the spine and the destination switch). Most routing protocols recognize that there are multiple equal-cost paths to a destination and enable any of them to be selected for a given traffic flow.

## Quagga Overview

Cumulus Linux uses **quagga**, an open source routing software suite, to provide the routing protocols for dynamic routing. Cumulus Linux supports the latest Quagga version, 0.99.23.1. Quagga is a fork of the **GNU Zebra** project.

Quagga provides many routing protocols, of which Cumulus Linux supports the following:

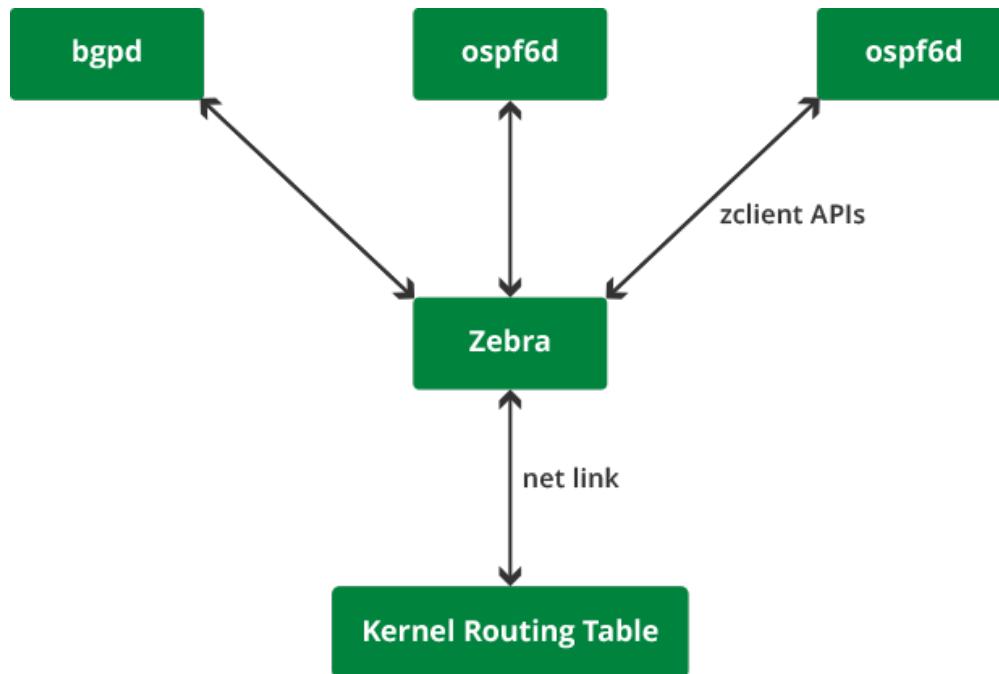
- Open Shortest Path First ([v2 \(see page 332\)](#) and [v3 \(see page 343\)](#))
- Border Gateway Protocol ([see page 345](#))

## Contents

(Click to expand)

- [Contents \(see page 318\)](#)
- [Architecture \(see page 319\)](#)
- [Zebra \(see page 319\)](#)
- [Configuration Files \(see page 319\)](#)
- [Useful Links \(see page 320\)](#)

## Architecture



As shown in the figure above, the Quagga routing suite consists of various protocol-specific daemons and a protocol-independent daemon called **zebra**. Each of the protocol-specific daemons are responsible for running the relevant protocol and building the routing table based on the information exchanged.

It is not uncommon to have more than one protocol daemon running at the same time. For example, at the edge of an enterprise, protocols internal to an enterprise (called IGP for Interior Gateway Protocol) such as [OSPF \(see page 332\)](#) or RIP run alongside the protocols that connect an enterprise to the rest of the world (called EGP or Exterior Gateway Protocol) such as [BGP \(see page 345\)](#).

**zebra** is the daemon that resolves the routes provided by multiple protocols (including static routes specified by the user) and programs these routes in the Linux kernel via **netlink** (in Linux). **zebra** does more than this, of course.

## Zebra

The [quagga documentation](#) defines **zebra** as the IP routing manager for **quagga** that “provides kernel routing table updates, interface lookups, and redistribution of routes between different routing protocols.”

## Configuration Files

- /etc/quagga/bgpd.conf
- /etc/quagga/daemons
- /etc/quagga/debian.conf
- /etc/quagga/ospf6d.conf
- /etc/quagga/ospfd.conf
- /etc/quagga/vtysh.conf
- /etc/quagga/zebra.conf

## Useful Links

- <http://www.quagga.net/>
- <http://packages.debian.org/quagga>

## Configuring Quagga

This section provides an overview of configuring quagga.

Before you run quagga, make sure all relevant daemons, such as zebra, are running. Make your changes in /etc/quagga/daemons then restart quagga with service quagga restart.

## Contents

(Click to expand)

- Contents (see page 320)
- Configuration Files (see page 321)
  - Starting Quagga (see page 321)
  - Understanding Integrated Configurations (see page 321)
  - Restoring the Default Quagga Configuration (see page 323)
- Interface IP Addresses (see page 323)
- Using the vtysh Modal CLI (see page 323)
- Using the Cumulus Linux Non-Modal CLI (see page 328)
- Comparing vtysh and Cumulus Linux Commands (see page 328)
  - Displaying the Routing Table (see page 328)
  - Creating a New Neighbor (see page 329)
  - Redistributing Routing Information (see page 329)
  - Defining a Static Route (see page 329)
  - Configuring an IPv6 Interface (see page 329)
  - Enabling PTM (see page 330)
  - Configuring MTU in IPv6 Network Discovery (see page 330)
  - Logging OSPF Adjacency Changes (see page 330)
  - Setting OSPF Interface Priority (see page 331)
  - Configuring Timing for OSPF SPF Calculations (see page 331)
  - Configuring Hello Packet Intervals (see page 331)
  - Displaying OSPF Debugging Status (see page 332)
  - Displaying BGP Information (see page 332)
- Useful Links (see page 332)

## Configuration Files

At startup, `quagga` reads a set of files to determine the startup configuration. The files and what they contain are specified below:

File	Description
Quagga.conf	The default, integrated, single configuration file for all <code>quagga</code> daemons.
daemons	Contains the list of <code>quagga</code> daemons that must be started.
zebra.conf	Configuration file for the <code>zebra</code> daemon.
ospfd.conf	Configuration file for the OSPFv2 daemon.
ospf6d.conf	Configuration file for the OSPFv3 daemon.
bgpd.conf	Configuration file for the BGP daemon.

## Starting Quagga

Quagga does not start by default in Cumulus Linux 2.0 and later versions.

Before you start `quagga`, modify `/etc/quagga/daemons` to enable the corresponding daemons:

```
zebra=yes (* this one is mandatory to bring the others up)
bgpd=yes
ospfd=yes
ospf6d=yes
ripd=no
ripngd=no
isisd=no
babeld=no
```

Then, start `quagga`:

```
cumulus@switch1:~$ sudo service quagga start
```

## Understanding Integrated Configurations

By default in Cumulus Linux, `quagga` saves the configuration of all daemons in a single integrated configuration file, `Quagga.conf`.

You can disable this mode by running:

```
quagga(config)# no service integrated-vtysh-config  
quagga(config)#
```

To enable the integrated configuration file mode again, run:

```
quagga(config)# service integrated-vtysh-config  
quagga(config)#
```

If you disable the integrated configuration mode, `quagga` saves each daemon-specific configuration file in a separate file. At a minimum for a daemon to start, that daemon must be specified in the `daemons` file and the daemon-specific configuration file must be present, even if that file is empty.

For example, to start `bgpd`, the `daemons` file needs to be formatted as follows, at minimum:

```
cumulus@switch:~$ sudo cat /etc/quagga/daemons  
zebra=yes  
bgpd=yes
```

The current configuration can be saved by running:

```
quagga# write mem  
Building Configuration...  
Integrated configuration saved to /etc/quagga/Quagga.conf  
[OK]
```



You can use `write file` instead of `write mem`.

When the integrated configuration mode disabled, the output looks like this:

```
quagga# write mem  
Building Configuration...  
Configuration saved to /etc/quagga/zebra.conf  
Configuration saved to /etc/quagga/bgpd.conf  
[OK]
```



The `daemons` file is not written using the `write mem` command.

## Restoring the Default Quagga Configuration

If you need to restore the Quagga configuration to the default running configuration, you need to delete the Quagga.conf file and restart the quagga service.

1. Confirm service integrated-vtysh-config is enabled:

```
cumulus@switch$ sudo cl-rctl running-config |grep integrated  
service integrated-vtysh-config
```

2. Remove /etc/quagga/Quagga.conf:

```
cumulus@switch$ sudo rm /etc/quagga/Quagga.conf
```

3. Restart the quagga service:

```
cumulus@switch$ sudo service quagga restart
```



If for some reason service integrated-vtysh-config is not configured, then you should remove zebra.conf instead of Quagga.conf in step 2 above.

## Interface IP Addresses

Quagga inherits the IP addresses for the network interfaces from the /etc/network/interfaces file. This is the recommended way to define the addresses. For more information, see [Configuring IP Addresses](#) (see page 97).

## Using the vtysh Modal CLI

Quagga provides a CLI – vtysh – for configuring and displaying the state of the protocols. It is invoked by running:

```
cumulus@switch:~$ sudo vtysh  
  
Hello, this is Quagga (version 0.99.21).  
Copyright 1996-2005 Kunihiro Ishiguro, et al.  
  
quagga#
```

Launching `vtysh` brings you into `zebra` initially. From here, you can log into other protocol daemons, such as `bgpd`, `ospf` or `babeld`.

`vtysh` provides a Cisco-like modal CLI, and many of the commands are similar to Cisco IOS commands. By modal CLI, we mean that there are different modes to the CLI, and certain commands are only available within a specific mode. Configuration is available with the `configure terminal` command, which is invoked thus:

```
quagga# configure terminal  
quagga(config)#
```

The prompt displays the mode the CLI is in. For example, when the interface-specific commands are invoked, the prompt changes to:

```
quagga(config)# interface swp1  
quagga(config-if)#
```

When the routing protocol specific commands are invoked, the prompt changes to:

```
quagga(config)# router ospf  
quagga(config-router)#
```

At any level, "?" displays the list of available top-level commands at that level:

```
quagga(config-if)# ?  
babel      Babel interface commands  
bandwidth   Set bandwidth informational parameter  
description Interface specific description  
end        End current mode and change to enable mode  
exit        Exit current mode and down to previous mode  
ip          Interface Internet Protocol config commands  
ipv6       Interface IPv6 config commands  
isis        IS-IS commands  
link-detect Enable link detection on interface  
list        Print command list  
mpls-te    MPLS-TE specific commands  
multicast   Set multicast flag to interface  
no          Negate a command or set its defaults  
ospf        OSPF interface commands  
quit        Exit current mode and down to previous mode  
shutdown   Shutdown the selected interface
```

?-based completion is also available to see the parameters that a command takes:

```
quagga(config-if)# bandwidth ?
<1-10000000> Bandwidth in kilobits
quagga(config-if)# ip ?
address Set the IP address of an interface
irdp Alter ICMP Router discovery preference this interface
ospf OSPF interface commands
rip Routing Information Protocol
router IP router interface commands
```

Displaying state can be done at any level, including the top level. For example, to see the routing table as seen by zebra, you use:

```
quagga# show ip route
Codes: K - kernel route, C - connected, S - static, R - RIP,
       O - OSPF, I - IS-IS, B - BGP, A - Babel,
       > - selected route, * - FIB route

K>* 0.0.0.0/0 via 192.168.0.2, eth0
C>* 192.0.2.11/24 is directly connected, swp1
C>* 192.0.2.12/24 is directly connected, swp2
B>* 203.0.113.30/24 [200/0] via 192.0.2.2, swp1, 10:43:05
B>* 203.0.113.31/24 [200/0] via 192.0.2.2, swp1, 10:43:05
B>* 203.0.113.32/24 [200/0] via 192.0.2.2, swp1, 10:43:05
C>* 127.0.0.0/8 is directly connected, lo
C>* 192.168.0.0/24 is directly connected, eth0
```

To run the same command at a config level, you prepend do to it:

```
quagga(config-router)# do show ip route
Codes: K - kernel route, C - connected, S - static, R - RIP,
       O - OSPF, I - IS-IS, B - BGP, A - Babel,
       > - selected route, * - FIB route

K>* 0.0.0.0/0 via 192.168.0.2, eth0
C>* 192.0.2.11/24 is directly connected, swp1
C>* 192.0.2.12/24 is directly connected, swp2
B>* 203.0.113.30/24 [200/0] via 192.0.2.2, swp1, 10:43:05
B>* 203.0.113.31/24 [200/0] via 192.0.2.2, swp1, 10:43:05
```

```
B>* 203.0.113.32/24 [200/0] via 192.0.2.2, swp1, 10:43:05
C>* 127.0.0.0/8 is directly connected, lo
C>* 192.168.0.0/24 is directly connected, eth0
```

Running single commands with `vtysh` is possible using the `-c` option of `vtysh`:

```
cumulus@switch:~$ sudo vtysh -c 'sh ip route'
Codes: K - kernel route, C - connected, S - static, R - RIP,
       O - OSPF, I - IS-IS, B - BGP, A - Babel,
       > - selected route, * - FIB route

K>* 0.0.0.0/0 via 192.168.0.2, eth0
C>* 192.0.2.11/24 is directly connected, swp1
C>* 192.0.2.12/24 is directly connected, swp2
B>* 203.0.113.30/24 [200/0] via 192.0.2.2, swp1, 11:05:10
B>* 203.0.113.31/24 [200/0] via 192.0.2.2, swp1, 11:05:10
B>* 203.0.113.32/24 [200/0] via 192.0.2.2, swp1, 11:05:10
C>* 127.0.0.0/8 is directly connected, lo
C>* 192.168.0.0/24 is directly connected, eth0
```

Running a command multiple levels down is done thus:

```
cumulus@switch:~$ sudo vtysh -c 'configure terminal' -c 'router ospf' -c
'area 0.0.0.1 range 10.10.10.0/24'
```

Notice that the commands also take a partial command name (for example, `sh ip route` above) as long as the partial command name is not aliased:

```
cumulus@switch:~$ sudo vtysh -c 'sh ip r'
% Ambiguous command.
```

A command or feature can be disabled by prepending the command with `no`. For example:

```
quagga(config-router)# no area 0.0.0.1 range 10.10.10.0/24
```

The current state of the configuration can be viewed via:

```
quagga# show running-config
```

```
Building configuration...
```

```
Current configuration:
```

```
!
hostname quagga
log file /media/node/zebra.log
log file /media/node/bgpd.log
log timestamp precision 6
!
service integrated-vtysh-config
!
password xxxxxx
enable password xxxxxx
!
interface eth0
ipv6 nd suppress-ra
link-detect
!
interface lo
link-detect
!
interface swp1
ipv6 nd suppress-ra
link-detect
!
interface swp2
ipv6 nd suppress-ra
link-detect
!
router bgp 65000
bgp router-id 0.0.0.9
bgp log-neighbor-changes
bgp scan-time 20
network 29.0.1.0/24
timers bgp 30 90
neighbor tier-2 peer-group
neighbor 192.0.2.2 remote-as 65000
neighbor 192.0.2.2 ttl-security hops 1
neighbor 192.0.2.2 advertisement-interval 30
neighbor 192.0.2.2 timers 30 90
neighbor 192.0.2.2 timers connect 30
neighbor 192.0.2.2 next-hop-self
neighbor 192.0.2.12 remote-as 65000
neighbor 192.0.2.12 next-hop-self
neighbor 203.0.113.1 remote-as 65000
```

```
!
ip forwarding
ipv6 forwarding
!
line vty
exec-timeout 0 0
!
end
```

## **Using the Cumulus Linux Non-Modal CLI**

The `vtysh` modal CLI can be difficult to work with and even more difficult to script. As an alternative to this, Cumulus Linux contains a non-modal version of these commands, structured similar to the Linux `ip` command. The available commands are:

Command	Description
<code>cl-bgp</code>	BGP (see page 345) commands. See <code>man cl-bgp</code> for details.
<code>cl-ospf</code>	OSPFv2 (see page 332) commands. For example: <code>cumulus@switch:~\$ sudo cl-ospf area 0.0.0.1 range 10.10.10.0/24</code>
<code>cl-ospf6</code>	OSPFv3 (see page 343) commands.
<code>cl-ra</code>	Route advertisement commands. See <code>man cl-ra</code> for details.
<code>cl-rctl</code>	Zebra and non-routing protocol-specific commands. See <code>man cl-rctl</code> for details.

## **Comparing `vtysh` and Cumulus Linux Commands**

This section describes how you can use the various Cumulus Linux CLI commands to configure Quagga, without using `vtysh`.

### **Displaying the Routing Table**

To display the routing table under Quagga, you would run:

```
quagga# show ip route
```

To display the routing table with the Cumulus Linux CLI, run:

```
cumulus@switch:~$ sudo cl-rctl route
```

## ***Creating a New Neighbor***

To create a new neighbor under Quagga, you would run:

```
quagga(config)# router bgp 65002
quagga(config-router)# neighbor 14.0.0.22 remote-as 65007
```

To create a new neighbor with the Cumulus Linux CLI, run:

```
cumulus@switch:~$ sudo cl-bgp as 65002 neighbor add 14.0.0.22 remote-as
65007
```

## ***Redistributing Routing Information***

To redistribute routing information from static route entries into RIP tables under Quagga, you would run:

```
quagga(config)# router bgp 65002
quagga(config-router)# redistribute static
```

To redistribute routing information from static route entries into RIP tables with the Cumulus Linux CLI, run:

```
cumulus@switch:~$ sudo cl-bgp as 65002 redistribute add static
```

## ***Defining a Static Route***

To define a static route under Quagga, you would run:

```
quagga(config)# ip route 155.1.2.20/24 br2 45
```

To define a static route with the Cumulus Linux CLI, run:

```
cumulus@switch:~$ sudo cl-rctl ip route add 175.0.0.0/28 interface br1
distance 25
```

## ***Configuring an IPv6 Interface***

To configure an IPv6 address under Quagga, you would run:

```
quagga(config)# int br3
quagga(config-if)# ipv6 address 3002:2123:1234:1abc::21/64
```

To configure an IPv6 address with the Cumulus Linux CLI, run:

```
cumulus@switch:~$ sudo cl-rctl interface add swp3 ipv6 address 3002:2123:
abcd:2120::41/64
```

## **Enabling PTM**

To enable topology checking (PTM) under Quagga, you would run:

```
quagga(config)# ptm-enable
```

To enable topology checking (PTM) with the Cumulus Linux CLI, run:

```
cumulus@switch:~$ sudo cl-rctl ptm-enable set
```

## **Configuring MTU in IPv6 Network Discovery**

To configure MTU (see page 111) in IPv6 network discovery for an interface under Quagga, you would run:

```
quagga(config)# int swp3
quagga(config-if)# ipv6 nd mtu 9000
```

To configure MTU in IPv6 network discovery for an interface with the Cumulus Linux CLI, run:

```
cumulus@switch:~$ sudo cl-ra interface swp3 set mtu 9000
```

## **Logging OSPF Adjacency Changes**

To log adjacency of OSPF changes under Quagga, you would run:

```
quagga(config)# router ospf
quagga(config-router)# router-id 2.0.0.21
quagga(config-router)# log-adjacency-changes
```

To log adjacency changes of OSPF with the Cumulus Linux CLI, run:

```
cumulus@switch:~$ sudo cl-ospf log-adjacency-changes set  
cumulus@switch:~$ sudo cl-ospf router-id set 3.0.0.21
```

## ***Setting OSPF Interface Priority***

To set the OSPF interface priority under Quagga, you would run:

```
quagga(config)# int swp3  
quagga(config-if)# ip ospf priority 120
```

To set the OSPF interface priority with the Cumulus Linux CLI, run:

```
cumulus@switch:~$ sudo cl-ospf interface set swp3 priority 120
```

## ***Configuring Timing for OSPF SPF Calculations***

To configure timing for OSPF SPF calculations under Quagga, you would run:

```
quagga(config)# router ospf6  
quagga(config-ospf6)# timer throttle spf 40 50 60
```

To configure timing for OSPF SPF calculations with the Cumulus Linux CLI, run:

```
cumulus@switch:~$ sudo cl-ospf6 timer add throttle spf 40 50 60
```

## ***Configuring Hello Packet Intervals***

To configure the OSPF Hello packet interval in number of seconds for an interface under Quagga, you would run:

```
quagga(config)# int swp4  
quagga(config-if)# ipv6 ospf6 hello-interval 60
```

To configure the OSPF Hello packet interval in number of seconds for an interface with the Cumulus Linux CLI, run:

```
cumulus@switch:~$ sudo cl-ospf6 interface set swp4 hello-interval 60
```

## Displaying OSPF Debugging Status

To display OSPF debugging status under Quagga, you would run:

```
quagga# show debugging ospf
```

To display OSPF debugging status with the Cumulus Linux CLI, run:

```
cumulus@switch:~$ sudo cl-ospf debug show
```

## Displaying BGP Information

To display BGP information under Quagga, you would run:

```
quagga# show ip bgp summary
```

To display BGP information with the Cumulus Linux CLI, run:

```
cumulus@switch:~$ sudo cl-bgp summary
```

## Useful Links

- <http://www.nongnu.org/quagga/docs/docs-info.html#BGP>
- <http://www.nongnu.org/quagga/docs/docs-info.html#IPv6-Support>
- <http://www.nongnu.org/quagga/docs/docs-info.html#Zebra>

## Open Shortest Path First - OSPF - Protocol

OSPFv2 is a [link-state routing protocol](#) for IPv4. OSPF maintains the view of the network topology conceptually as a directed graph. Each router represents a vertex in the graph. Each link between neighboring routers represents a unidirectional edge. Each link has an associated weight (called cost) that is either automatically derived from its bandwidth or administratively assigned. Using the weighted topology graph, each router computes a shortest path tree (SPT) with itself as the root, and applies the results to build its forwarding table. The computation is generally referred to as *SPF computation* and the resultant tree as the *SPF tree*.

An LSA (*link-state advertisement*) is the fundamental quantum of information that OSPF routers exchange with each other. It seeds the graph building process on the node and triggers SPF computation. LSAs originated by a node are distributed to all the other nodes in the network through a mechanism called *flooding*. Flooding is done hop-by-hop. OSPF ensures reliability by using link state acknowledgement packets. The set of LSAs in a router's memory is termed *link-state database* (LSDB), a representation of the network graph. Thus, OSPF ensures a consistent view of LSDB on each node in the network in a distributed fashion (eventual consistency model); this is key to the protocol's correctness.

## Contents

(Click to expand)

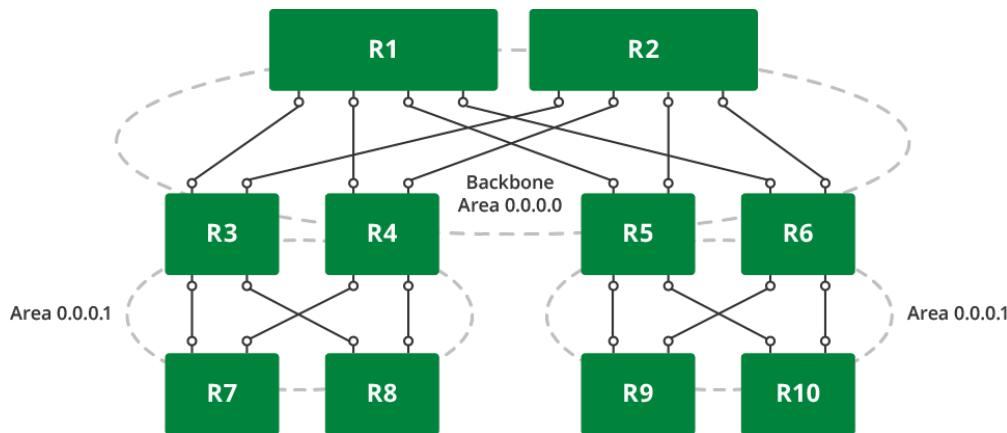
- [Contents \(see page 333\)](#)
- [Scalability and Areas \(see page 333\)](#)
- [Configuring OSPFv2 \(see page 334\)
  - \[Activating the OSPF and Zebra Daemons \\(see page 334\\)\]\(#\)
  - \[Enabling OSPF \\(see page 334\\)\]\(#\)
  - \[Defining \\(Custom\\) OSPF Parameters on the Interfaces \\(see page 336\\)\]\(#\)](#)
- [Scaling Tip: Summarization \(see page 337\)](#)
- [Scaling Tip: Stub Areas \(see page 338\)](#)
- [Configuration Tip: Unnumbered Interfaces \(see page 339\)](#)
- [ECMP \(see page 340\)](#)
- [Topology Changes and OSPF Reconvergence \(see page 340\)
  - \[Example Configurations \\(see page 340\\)\]\(#\)](#)
- [Debugging OSPF \(see page 341\)](#)
- [Configuration Files \(see page 342\)](#)
- [Supported RFCs \(see page 342\)](#)
- [Useful Links \(see page 342\)](#)

## Scalability and Areas

An increase in the number of nodes affects OSPF scalability in the following ways:

- Memory footprint to hold the entire network topology,
- Flooding performance,
- SPF computation efficiency.

The OSPF protocol advocates hierarchy as a *divide and conquer* approach to achieve high scale. The topology may be divided into areas, resulting in a two-level hierarchy. Area 0 (or 0.0.0.0), called the backbone area, is the top level of the hierarchy. Packets traveling from one non-zero area to another must go via the backbone area. As an example, the leaf-spine topology we have been referring to in the routing section can be divided into areas as follows:



Here are some points to note about areas and OSPF behavior:

- Routers that have links to multiple areas are called *area border routers* (ABR). For example, routers R3, R4, R5, R6 are ABRs in the diagram. An ABR performs a set of specialized tasks, such as SPF computation per area and summarization of routes across areas.
- Most of the LSAs have an area-level flooding scope. These include router LSA, network LSA, and summary LSA.



In the diagram, we reused the same non-zero area address. This is fine since the area address is only a scoping parameter provided to all routers within that area. It has no meaning outside the area. Thus, in the cases where ABRs do not connect to multiple non-zero areas, the same area address can be used, thus reducing the operational headache of coming up with area addresses.

## Configuring OSPFv2

Configuring OSPF involves the following tasks:

- Activating the OSPF daemon
- Enabling OSPF
- Defining (Custom) OSPF parameters on the interfaces

## Activating the OSPF and Zebra Daemons

1. Add the following to `/etc/quagga/daemons`:

```
zebra=yes
ospfd=yes
```

2. Restart the quagga service to start the new daemons:

```
cumulus@switch:~$ sudo service quagga restart
```

## Enabling OSPF

As we discussed in [Introduction to Routing Protocols](#) (see page 314), there are three steps to the configuration:

1. Identifying the router with the router ID.

2. With whom should the router communicate?
3. What information (most notably the prefix reachability) to advertise?

There are two ways to achieve (2) and (3) in the Quagga OSPF:

1. The `network` statement under `router ospf` does both. The statement is specified with an IP subnet prefix and an area address. All the interfaces on the router whose IP address matches the `network` subnet are put into the specified area. OSPF process starts bringing up peering adjacency on those interfaces. It also advertises the interface IP addresses formatted into LSAs (of various types) to the neighbors for proper reachability.

From the Cumulus Linux shell:

```
cumulus@switch:~$ sudo vtysh

Hello, this is Quagga (version 0.99.21).
Copyright 1996-2005 Kunihiro Ishiguro, et al.

R3# configure terminal
R3(config)# router ospf
R3(config-router)# router-id 0.0.0.1
R3(config-router)# log-adjacency-changes detail
R3(config-router)# network 10.0.0.0/16 area 0.0.0.0
R3(config-router)# network 192.0.2.0/16 area 0.0.0.1
R3(config-router)#

```

Or through `cl-ospf`, from the Cumulus Linux shell:

```
cumulus@switch:~$ sudo cl-ospf router set id 0.0.0.1
cumulus@switch:~$ sudo cl-ospf router set log-adjacency-changes detail
cumulus@switch:~$ sudo cl-ospf router set network 10.0.0.0/16 area
0.0.0.0
cumulus@switch:~$ sudo cl-ospf router set network 192.0.2.0/16 area
0.0.0.1

```

The subnets in the `network` subnet can be as coarse as possible to cover the most number of interfaces on the router that should run OSPF.

There may be interfaces where it's undesirable to bring up OSPF adjacency. For example, in a data center topology, the host-facing interfaces need not run OSPF; however the corresponding IP addresses should still be advertised to neighbors. This can be achieved using the `passive-interface` construct.

From the vtysh/quagga CLI:

```
R3# configure terminal  
R3(config)# router ospf  
R3(config-router)# passive-interface swp10  
R3(config-router)# passive-interface swp11
```

Or use the `passive-interface default` command to put all interfaces as passive and selectively remove certain interfaces to bring up protocol adjacency:

```
R3# configure terminal  
R3(config)# router ospf  
R3(config-router)# passive-interface default  
R3(config-router)# no passive-interface swp1
```

2. Explicitly enable OSPF for each interface by configuring it under the interface configuration mode:

```
R3# configure terminal  
R3(config)# interface swp1  
R3(config-if)# ip ospf area 0.0.0.0
```

If OSPF adjacency bringup is not desired, you should configure the corresponding interfaces as passive as explained above.

This model of configuration is required for unnumbered interfaces as discussed later in this guide.

For achieving step (3) alone, the `quagga` configuration provides another method: *redistribution*. For example:

```
R3# configure terminal  
R3(config)# router ospf  
R3(config-router)# redistribute connected
```

Redistribution, however, unnecessarily loads the database with type-5 LSAs and should be limited to generating real external prefixes (for example, prefixes learned from BGP). In general, it is a good practice to generate local prefixes using `network` and/or `passive-interface` statements.

## **Defining (Custom) OSPF Parameters on the Interfaces**

1. Network type, such as point-to-point, broadcast.
2. Timer tuning, like hello interval.
3. For unnumbered interfaces (see below), enable OSPF.

Using Quagga's vtysh:

```
R3(config)# interface swp1
R3(config-if)# ospf network point-to-point
R3(config-if)# ospf hello-interval 5
```

Or through cl-ospf, from the Cumulus Linux shell:

```
cumulus@switch:~$ sudo cl-ospf interface swp1 set network point-to-point
cumulus@switch:~$ sudo cl-ospf interface swp1 set hello-interval 5
```

The OSPF configuration is saved in `/etc/quagga/ospfd.conf`.

### ***Scaling Tip: Summarization***

By default, an ABR creates a summary (type-3) LSA for each route in an area and advertises it in adjacent areas. Prefix range configuration optimizes this behavior by creating and advertising one summary LSA for multiple routes.

To configure a range:

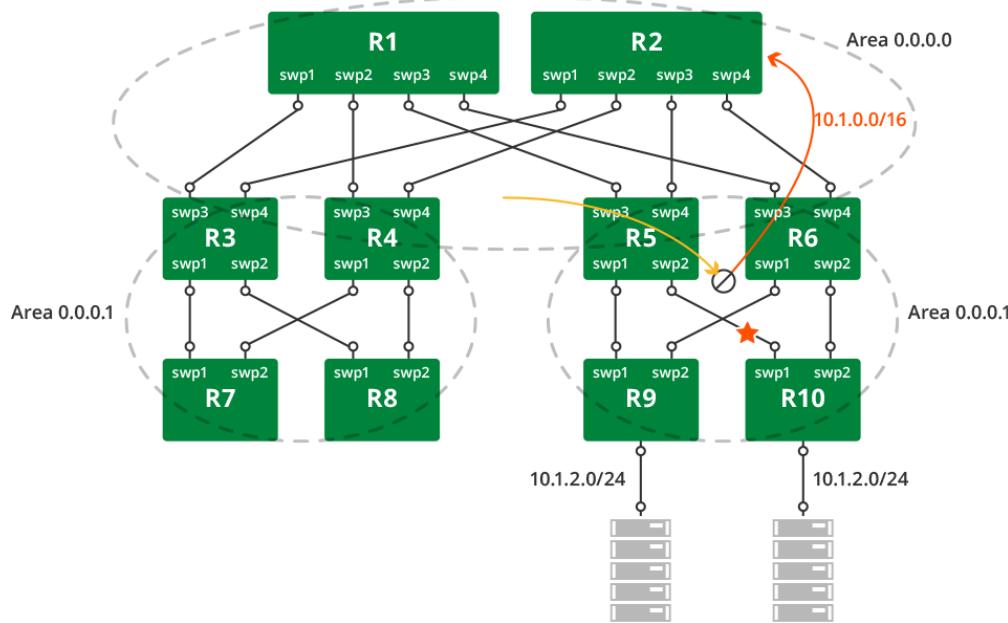
```
R3(config)# router ospf
R3(config-router)# area 0.0.0.1 range 30.0.0.0/8
```



Summarize in the direction to the backbone. The backbone receives summarized routes and injects them to other areas already summarized.



Summarization can cause non-optimal forwarding of packets during failures. Here is an example scenario:



As shown in the diagram, the ABRs in the right non-zero area summarize the host prefixes as 10.1.0.0/16. When the link between R5 and R10 fails, R5 will send a worse metric for the summary route (metric for the summary route is the maximum of the metrics of intra-area routes that are covered by the summary route. Upon failure of the R5-R10 link, the metric for 10.1.2.0/24 goes higher at R5 as the path is R5-R9-R6-R10). As a result, other backbone routers shift traffic destined to 10.1.0.0/16 towards R6. This breaks ECMP and is an under-utilization of network capacity for traffic destined to 10.1.1.0/24.

### Scaling Tip: Stub Areas

Nodes in an area receive and store intra-area routing information and summarized information about other areas from the ABRs. In particular, a good summarization practice about inter-area routes through prefix range configuration helps scale the routers and keeps the network stable.

Then there are external routes. External routes are the routes redistributed into OSPF from another protocol. They have an AS-wide flooding scope. In many cases, external link states make up a large percentage of the LSDB.

*Stub areas* alleviate this scaling problem. A stub area is an area that does not receive external route advertisements.

To configure a stub area:

```
R3(config)# router ospf
R3(config-router)# area 0.0.0.1 stub
```

Stub areas still receive information about networks that belong to other areas of the same OSPF domain. Especially, if summarization is not configured (or is not comprehensive), the information can be overwhelming for the nodes. *Totally stubby areas* address this issue. Routers in totally stubby areas keep in their LSDB information about routing within their area, plus the default route.

To configure a totally stubby area:

```
R3(config)# router ospf
R3(config-router)# area 0.0.0.1 stub no-summary
```

Here is a brief tabular summary of the area type differences:

Type	Behavior
Normal non- zero area	LSA types 1, 2, 3, 4 area-scoped, type 5 externals, inter-area routes summarized
Stub area	LSA types 1, 2, 3, 4 area-scoped, No type 5 externals, inter-area routes summarized
Totally stubby area	LSA types 1, 2 area-scoped, default summary, No type 3, 4, 5 LSA types allowed

### **Configuration Tip: Unnumbered Interfaces**

Unnumbered interfaces are interfaces without unique IP addresses. In OSPFv2, configuring unnumbered interfaces reduces the links between routers into pure topological elements, which dramatically simplifies network configuration and reconfiguration. In addition, the routing database contains only the real networks, so the memory footprint is reduced and SPF is faster.



Unnumbered is usable for point-to-point interfaces only.



If there is a `network <network number>/<mask> area <area ID>` command present in the Quagga configuration, the `ip ospf area <area ID>` command is rejected with the error "Please remove network command first." This prevents you from configuring other areas on some of the unnumbered interfaces. You can use either the `network area` command or the `ospf area` command in the configuration, but not both.



Unless the Ethernet media is intended to be used as a LAN with multiple connected routers, we recommend configuring the interface as point-to-point. It has the additional advantage of a simplified adjacency state machine; there is no need for DR/BDR election and *LSA reflection*. See [RFC5309](#) for a more detailed discussion.

To configure an unnumbered interface, take the IP address of another interface (called the *anchor*) and use that as the IP address of the unnumbered interface:

```
cumulus@switch:~$ sudo ifconfig lo 192.0.2.1/24
cumulus@switch:~$ sudo ifconfig swp1 192.0.2.1/24
cumulus@switch:~$ sudo ifconfig swp2 192.0.2.1/24
```

To enable OSPF on an unnumbered interface from within Quagga's vtysh:

```
R3(config)# interface swp1
R3(config-if)# ip ospf area 0.0.0.1
```

## ECMP

During SPF computation for an area, if OSPF finds multiple paths with equal cost (metric), all those paths are used for forwarding. For example, in the reference topology diagram, R8 uses both R3 and R4 as next hops to reach a destination attached to R9.

## Topology Changes and OSPF Reconvergence

Topology changes usually occur due to one of four events:

1. Maintenance of a router node
2. Maintenance of a link
3. Failure of a router node
4. Failure of a link

For the maintenance events, operators typically raise the OSPF administrative weight of the link(s) to ensure that all traffic is diverted from the link or the node (referred to as *costing out*). The speed of reconvergence does not matter. Indeed, changing the OSPF cost causes LSAs to be reissued, but the links remain in service during the SPF computation process of all routers in the network.

For the failure events, traffic may be lost during reconvergence; that is, until SPF on all nodes computes an alternative path around the failed link or node to each of the destinations. The reconvergence depends on layer 1 failure detection capabilities and at the worst case *DeadInterval* OSPF timer.

## Example Configurations

Example configuration for event 1, using vtysh:

```
R3(config)# router ospf
R3(config-router)# max-metric router-lsa administrative
```

Or, with the non-modal shell command approach:

```
cumulus@switch:~$ sudo cl-ospf router set max-metric router-lsa
administrative
```

Example configuration for event 2, using vtysh:

```
R3(config)# interface swp1
R3(config-if)# ospf cost 65535
```

Or, with the non-modal shell command approach:

```
cumulus@switch:~$ sudo cl-ospf interface swp1 set cost 65535
```

## Debugging OSPF

`OperState` lists all the commands to view the operational state of OSPF.

The three most important states while troubleshooting the protocol are:

1. Neighbors, with `show ip ospf neighbor`. This is the starting point to debug neighbor states (also see `tcpdump` below).
2. Database, with `show ip ospf database`. This is the starting point to verify that the LSDB is, in fact, synchronized across all routers in the network. For example, sweeping through the output of `show ip ospf database router` taken from all routers in an area will ensure if the topology graph building process is complete; that is, every node has seen all the other nodes in the area.
3. Routes, with `show ip ospf route`. This is the outcome of SPF computation that gets downloaded to the forwarding table, and is the starting point to debug, for example, why an OSPF route is not being forwarded correctly.



Compare the route output with kernel by using `show ip route | grep zebra` and with the hardware entries using `cl-route-check -V`.

Using `cl-ospf`:

```
cumulus@switch:~$ sudo cl-ospf neighbor show [all | detail]

cumulus@switch:~$ sudo cl-ospf database show [asbr-summary | network |
opaque-area |
                    opaque-link | summary | external |
nssa-external | opaque-as | router]

cumulus@switch:~$ sudo cl-ospf route show
```

`Debugging-OSPF` lists all of the OSPF debug options.

Using `cl-ospf`:

```
Usage: cl-ospf debug { COMMAND | help }

COMMANDS
  { set | clear } (all | event | ism | ism [OBJECT] | lsa | lsa
[OBJECT] |
  nsm | nsm [OBJECT] | nssa | packet | packet [OBJECT] |
  zebra [OBJECT] | zebra all)
```

Using zebra under vtysh:

```
cumulus@switch:~$ sudo vtysh
R3# show [zebra]

IOBJECT := { events | status | timers }
OOBJECT := { interface | redistribute }
POBJECT := { all | dd | hello | ls-ack | ls-request | ls-update }
ZOBJECT := { all | events | kernel | packet | rib |
```

Using tcpdump to capture OSPF packets:

```
cumulus@switch:~$ sudo tcpdump -v -i swp1 ip proto ospf
```

## ***Configuration Files***

- /etc/quagga/daemons
- /etc/quagga/ospfd.conf

## ***Supported RFCs***

- RFC2328
- RFC3137
- RFC5309

## ***Useful Links***

- Bidirectional forwarding detection (see page 367) (BFD) and OSPF
- [http://en.wikipedia.org/wiki/Open\\_Shortest\\_Path\\_First](http://en.wikipedia.org/wiki/Open_Shortest_Path_First)
- <http://www.nongnu.org/quagga/docs/docs-info.html#OSPFv2>
- Perlman, Radia (1999). Interconnections: Bridges, Routers, Switches, and Internetworking Protocols (2 ed.). Addison-Wesley.
- Moy, John T. OSPF: Anatomy of an Internet Routing Protocol. Addison-Wesley.

## Open Shortest Path First v3 - OSPFv3 - Protocol

OSPFv3 is a revised version of OSPFv2 to support the IPv6 address family. Refer to [Open Shortest Path First \(OSPF\) Protocol \(see page 332\)](#) for a discussion on the basic concepts, which remain the same between the two versions.

OSPFv3 has changed the formatting in some of the packets and LSAs either as a necessity to support IPv6 or to improve the protocol behavior based on OSPFv2 experience. Most notably, v3 defines a new LSA, called intra-area prefix LSA to separate out the advertisement of stub networks attached to a router from the router LSA. It is a clear separation of node topology from prefix reachability and lends itself well to an optimized SPF computation.



IETF has defined extensions to OSPFv3 to support multiple address families (that is, both IPv6 and IPv4). Quagga (see page 318) does not support it yet.

### Contents

(Click to expand)

- [Contents \(see page 343\)](#)
- [Configuring OSPFv3 \(see page 343\)](#)
- [Unnumbered Interfaces \(see page 345\)](#)
- [Debugging OSPF \(see page 345\)](#)
- [Configuration Files \(see page 345\)](#)
- [Supported RFCs \(see page 345\)](#)
- [Useful Links \(see page 345\)](#)

### Configuring OSPFv3

Configuring OSPFv3 involves the following tasks:

1. Activating the OSPF6 and Zebra daemons:
  - a. Add the following to `/etc/quagga/daemons`:

```
zebra=yes
ospf6d=yes
```
  - b. Restart the `quagga` service to start the new daemons:

```
cumulus@switch:~$ sudo service quagga restart
```

2. Enabling OSPF6 and map interfaces to areas. From Quagga's `vtysh` shell:

```
cumulus@switch:~$ sudo vtysh
```

```
Hello, this is Quagga (version 0.99.21).
Copyright 1996-2005 Kunihiro Ishiguro, et al.

R3# conf t
R3# configure terminal
R3(config)# router ospf6
R3(config-router)# router-id 0.0.1
R3(config-router)# log-adjacency-changes detail
R3(config-router)# interface swp1 area 0.0.0.0
R3(config-router)# interface swp2 area 0.0.0.1
R3(config-router)#
R3#
```

Or through `cl-ospf6`, from the Cumulus Linux shell:

```
cumulus@switch:~$ sudo cl-ospf6 router set id 0.0.0.1
cumulus@switch:~$ sudo cl-ospf6 router set log-adjacency-changes detail
cumulus@switch:~$ sudo cl-ospf6 interface swp1 set area 0.0.0.0
cumulus@switch:~$ sudo cl-ospf6 interface swp2 set area 0.0.0.1
```

3. Defining (custom) OSPF6 parameters on the interfaces:
  - a. Network type (such as point-to-point, broadcast)
  - b. Timer tuning (for example, hello interval)

Using Quagga's vtysh:

```
R3(config)# interface swp1
R3(config-if)# ipv6 ospf6 network point-to-point
R3(config-if)# ipv6 ospf6 hello-interval 5
```

Or through `cl-ospf6`, from the Cumulus Linux shell:

```
cumulus@switch:~$ sudo cl-ospf6 interface swp1 set network point-to-
point
cumulus@switch:~$ sudo cl-ospf6 interface swp1 set hello-interval 5
```

The OSPFv3 configuration is saved in `/etc/quagga/ospf6d.conf`.

## Unnumbered Interfaces

Unlike OSPFv2, OSPFv3 intrinsically supports unnumbered interfaces. Forwarding to the next hop router is done entirely using IPv6 link local addresses. Therefore, you are not required to configure any global IPv6 address to interfaces between routers.

## Debugging OSPF

See [Debugging OSPF \(see page 341\)](#) for OSPFv2 for the troubleshooting discussion. The equivalent commands are:

```
cumulus@switch:~$ sudo vtysh
R3# show ipv6 ospf6 neighbor
R3# show ipv6 ospf6 database [detail | dump | internal |
                                as-external | group-membership |
                                inter-prefix | inter-router |
                                intra-prefix | link | network |
                                router | type-7 | * | adv-router |
                                linkstate-id | self-originated]
R3# show ip ospf route
```

Another helpful command is `show ipv6 ospf6 [area <id>] spf tree`. It dumps the node topology as computed by SPF to help visualize the network view.

## Configuration Files

- /etc/quagga/daemons
- /etc/quagga/ospf6d.conf

## Supported RFCs

- RFC5340
- RFC3137

## Useful Links

- Bidirectional forwarding detection ([see page 367](#)) (BFD) and OSPF
- [http://en.wikipedia.org/wiki/Open\\_Shortest\\_Path\\_First](http://en.wikipedia.org/wiki/Open_Shortest_Path_First)
- <http://www.nongnu.org/quagga/docs/docs-info.html#OSPFv3>

## Configuring Border Gateway Protocol - BGP

BGP is the routing protocol that runs the Internet. It is an increasingly popular protocol for use in the data center as it lends itself well to the rich interconnections in a Clos topology. Specifically:

- It does not require routing state to be periodically refreshed unlike OSPF.
- It is less chatty than its link-state siblings. For example, a link or node transition can result in a bestpath change, causing BGP to send updates.
- It is multi-protocol and extensible.
- There are many robust vendor implementations.
- The protocol is very mature and comes with many years of operational experience.

This IETF draft provides further details of the use of BGP within the data center.

## Contents

(Click to expand)

- [Contents \(see page 346\)](#)
- [Commands \(see page 347\)](#)
- [Autonomous System Number \(ASN\) \(see page 347\)](#)
- [eBGP and iBGP \(see page 347\)](#)
- [Route Reflectors \(see page 348\)](#)
- [ECMP with BGP \(see page 348\)](#)
- [BGP for both IPv4 and IPv6 \(see page 348\)](#)
- [Configuring BGP \(see page 348\)](#)
- [Using BGP Unnumbered Interfaces \(see page 350\)
  - \[BGP and Extended Next-hop Encoding \\(see page 351\\)\]\(#\)
  - \[Configuring BGP Unnumbered Interfaces \\(see page 351\\)\]\(#\)
  - \[Managing Unnumbered Interfaces \\(see page 351\\)\]\(#\)
  - \[How traceroute Interacts with BGP Unnumbered Interfaces \\(see page 353\\)\]\(#\)
  - \[Advanced: Understanding How Next-hop Fields Are Set \\(see page 353\\)\]\(#\)
  - \[Limitations \\(see page 355\\)\]\(#\)](#)
- [Fast Convergence Design Considerations \(see page 355\)
  - \[Specifying the Interface Name in the neighbor Command \\(see page 355\\)\]\(#\)](#)
- [Configuring BGP Peering Relationships across Switches \(see page 356\)](#)
- [Configuration Tips \(see page 357\)
  - \[Using peer-group to Simplify Configuration \\(see page 357\\)\]\(#\)
  - \[Preserving the AS\\\_PATH Setting \\(see page 358\\)\]\(#\)
  - \[Utilizing Multiple Routing Tables and Forwarding \\(see page 358\\)\]\(#\)](#)
- [Troubleshooting \(see page 358\)
  - \[Debugging Tip: Logging Neighbor State Changes \\(see page 361\\)\]\(#\)
  - \[Troubleshooting Link-local Addresses \\(see page 361\\)\]\(#\)](#)
- [Enabling Read-only Mode \(see page 362\)](#)
- [Applying a Route Map for Route Updates \(see page 363\)](#)
- [Protocol Tuning \(see page 363\)
  - \[Converging Quickly On Link Failures \\(see page 363\\)\]\(#\)](#)

- Converging Quickly On Soft Failures (see page 364)
- Reconnecting Quickly (see page 365)
- Advertisement Interval (see page 365)
- Configuration Files (see page 366)
- Useful Links (see page 366)
- Caveats and Errata (see page 366)
  - ttl-security Issue (see page 366)

## Commands

Cumulus Linux:

- bgp
- vtysh

Quagga:

- bgp
- neighbor
- router
- show

## Autonomous System Number (ASN)

One of the key concepts in BGP is an *autonomous system number* or ASN. An **autonomous system** is defined as a set of routers under a common administration. Since BGP was originally designed to peer between independently managed enterprises and/or service providers, each such enterprise is treated as an autonomous system, responsible for a set of network addresses. Each such autonomous system is given a unique number called its ASN. ASNs are handed out by a central authority (ICANN). However, ASNs between 64512 and 65535 are reserved for private use. Using BGP within the data center relies on either using this number space or else using the single ASN you own.

The ASN is central to how BGP builds a forwarding topology. A BGP route advertisement carries with it not only the originator's ASN, but also the list of ASNs that this route advertisement has passed through. When forwarding a route advertisement, a BGP speaker adds itself to this list. This list of ASNs is called the *AS path*. BGP uses the AS path to detect and avoid loops.

ASNs were originally 16-bit numbers, but were later modified to be 32-bit. Quagga supports both 16-bit and 32-bit ASNs, but most implementations still run with 16-bit ASNs.

## eBGP and iBGP

When BGP is used to peer between autonomous systems, the peering is referred to as *external BGP* or eBGP. When BGP is used within an autonomous system, the peering used is referred to as *internal BGP* or iBGP. eBGP peers have different ASNs while iBGP peers have the same ASN.

While the heart of the protocol is the same when used as eBGP or iBGP, there is a key difference in the protocol behavior between use as eBGP and iBGP: an iBGP node does not forward routing information learned from one iBGP peer to another iBGP peer. It expects the originating iBGP peer to send this information to all iBGP peers.

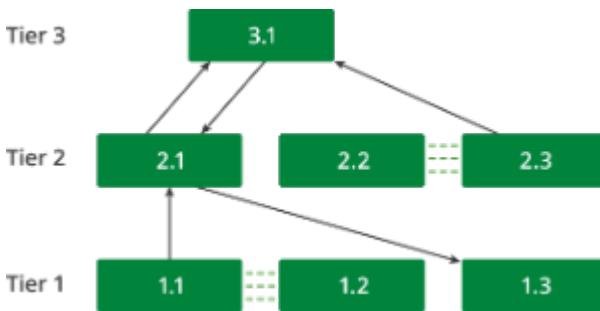
This implies that iBGP peers are all connected to each other. In a large network, this requirement can quickly become unscalable. The most popular method to avoid this problem is to introduce a *route reflector*.

## Route Reflectors

Route reflectors are quite easy to understand in a Clos topology. In a two-tier Clos network, the leaf (or tier 1) switches are the only ones connected to end stations. Subsequently, this means that the spines themselves do not have any routes to announce. They're merely **reflecting** the routes announced by one leaf to the other leaves. Thus, the spine switches function as route reflectors while the leaf switches serve as route reflector clients.

In a three-tier network, the tier 2 nodes (or mid-tier spines) act as both route reflector servers and route reflector clients. They act as route reflectors because they announce the routes learned from the tier 1 nodes to other tier 1 nodes and to tier 3 nodes. They also act as route reflector clients to the tier 3 nodes, receiving routes learned from other tier 2 nodes. Tier 3 nodes act only as route reflectors.

In the following illustration, tier 2 node 2.1 is acting as a route reflector server, announcing the routes between tier 1 nodes 1.1 and 1.2 to tier 1 node 1.3. It is also a route reflector client, learning the routes between tier 2 nodes 2.2 and 2.3 from the tier 3 node, 3.1.



## ECMP with BGP

If a BGP node hears a prefix **p** from multiple peers, it has all the information necessary to program the routing table to forward traffic for that prefix **p** through all of these peers. Thus, BGP supports equal-cost multipathing.

## BGP for both IPv4 and IPv6

Unlike OSPF, which has separate versions of the protocol to announce IPv4 and IPv6 routes, BGP is a multi-protocol routing engine, capable of announcing both IPv4 and IPv6 prefixes. It supports announcing IPv4 prefixes over an IPv4 session and IPv6 prefixes over an IPv6 session. It also supports announcing prefixes of both these address families over a single IPv4 session or over a single IPv6 session.

## Configuring BGP

1. Activate the BGP and Zebra daemons:
  - Add the following line to `/etc/quagga/daemons`:

```

zebra=yes
bgpd = yes
  
```

- Touch an empty `bgpd` configuration file:

```
cumulus@switch:~$ sudo touch /etc/quagga/bgpd.conf
```

A slightly more useful configuration file would contain the following lines:

```
hostname R7
password *****
enable password *****
log timestamp precision 6
log file /var/log/quagga/bgpd.log
!
line vty
exec-timeout 0 0
!
```

The most important information here is the specification of the location of the log file, where the BGP process can log debugging and other useful information. A common convention is to store the log files under `/var/log/quagga`.

You must restart `quagga` when a new daemon is enabled:

```
cumulus@switch:~$ sudo service quagga restart
```

- Identify the BGP node by assigning an ASN and `router-id`:

```
cumulus@switch:~$ sudo vtysh

Hello, this is Quagga (version 0.99.21).
Copyright 1996-2005 Kunihiro Ishiguro, et al.

R7# configure terminal
R7(config)# router bgp 65000
R7(config-router)# bgp router-id 0.0.0.1
```

- Specify to whom it must disseminate routing information:

```
R7(config-router)# neighbor 10.0.0.2 remote-as 65001
```

If it is an iBGP session, the `remote-as` is the same as the local AS:

```
R7(config-router)# neighbor 10.0.0.2 remote-as 65000
```

Specifying the peer's IP address allows BGP to set up a TCP socket with this peer, but it doesn't distribute any prefixes to it, unless it is explicitly told that it must via the `activate` command:

```
R7(config-router)# address-family ipv4 unicast
R7(config-router-af)# neighbor 10.0.0.2 activate
R7(config-router-af)# exit
R7(config-router)# address-family ipv6
R7(config-router-af)# neighbor 2002:0a00:0002::0a00:0002 activate
R7(config-router-af)# exit
```

As you can see, `activate` has to be specified for each address family that is being announced by the BGP session.

4. Specify some properties of the BGP session:

```
R7(config-router)# neighbor 10.0.0.2 next-hop-self
R7(config-router)# address-family ipv4 unicast
R7(config-router-af)# maximum-paths 64
```

For iBGP, the `maximum-paths` is selected by typing:

```
R7(config-router-af)# maximum-paths ibgp 64
```

If this is a route-reflector client, it can be specified as follows:

```
R3(config-router-af)# neighbor 10.0.0.1 route-reflector-client
```



It is node R3, the route reflector, on which the peer is specified as a client.

5. Specify what prefixes to originate:

```
R7(config-router)# address-family ipv4 unicast
R7(config-router-af)# network 192.0.2.0/24
R7(config-router-af)# network 203.0.113.1/24
```

## Using BGP Unnumbered Interfaces

Unnumbered interfaces are interfaces without unique IP addresses. In BGP, you configure unnumbered interfaces using *extended next-hop encoding* (ENHE), which is defined by [RFC 5549](#). BGP unnumbered interfaces provide a means of advertising an IPv4 route with an IPv6 next-hop. Prior to RFC 5549, an IPv4 route could be advertised only with an IPv4 next-hop.

BGP unnumbered interfaces are particularly useful in deployments where IPv4 prefixes are advertised through BGP over a section without any IPv4 address configuration on links. As a result, the routing entries are also IPv4 for destination lookup and have IPv6 next-hops for forwarding purposes.

### BGP and Extended Next-hop Encoding

Once enabled and active, BGP makes use of the available IPv6 next-hops for advertising any IPv4 prefixes. BGP learns the prefixes, calculates the routes and installs them in IPv4 AFI to IPv6 AFI format. However, ENHE in Cumulus Linux does not install routes into the kernel in IPv4 prefix to IPv6 next-hop format. For link-local peerings enabled by dynamically learning the other end's link-local address using IPv6 neighbor discovery router advertisements, an IPv6 next-hop is converted into an IPv4 link-local address and a static neighbor entry is installed for this IPv4 link-local address with the MAC address derived from the link-local address of the other end.



It is assumed that the IPv6 implementation on the peering device will use the MAC address as the interface ID when assigning the IPv6 link-local address, as suggested by RFC 4291.

## Configuring BGP Unnumbered Interfaces

Configuring a BGP unnumbered interface requires enabling IPv6 neighbor discovery router advertisements. The `interval` you specify is measured in seconds, and defaults to 600 seconds. Extended next-hop encoding is sent only for the link-local address peerings:

```
interface swp1
    no ipv6 nd suppress-ra
    ipv6 nd ra-interval 5
!
router bgp 10
    neighbor swp1 interface
    neighbor swp1 remote-as 20
    neighbor swp1 capability extended-nexthop
!
```

## Managing Unnumbered Interfaces

All the relevant BGP commands are now capable of showing IPv6 next-hops and/or the interface name for any IPv4 prefix:

```
# show ip bgp
BGP table version is 66, local router ID is 6.0.0.5
```

```

Status codes: s suppressed, d damped, h history, * valid, > best, =
multipath,
                  i internal, r RIB-failure, S Stale, R Removed
Origin codes: i - IGP, e - EGP, ? - incomplete
      Network          Next Hop            Metric LocPrf Weight Path
*-> 6.0.0.5/32        0.0.0.0                      0       32768  ?
*= 6.0.0.6/32        swp2                         0   65534 64503 ?
*=                     swp6                         0   65002 64503 ?
*=                     swp5                         0   65001 64503 ?
*=                     swp1                         0   65534 64503 ?
*=                     swp4                         0   65534 64503 ?
*>                     swp3                         0   65534 64503 ?

# show ip bgp 6.0.0.14/32
BGP routing table entry for 6.0.0.14/32
Paths: (1 available, best #1, table Default-IP-Routing-Table)
  Advertised to non peer-group peers:
    swp1 swp2 swp3 swp4 swp5 swp6
    65534
      fe80::202:ff:fe00:3d from swp2 (6.0.0.14)
      (fe80::202:ff:fe00:3d) (used)
        Origin incomplete, metric 0, localpref 100, valid, external, best
        Last update: Tue May 12 17:18:41 2015
  
```

Quagga RIB commands are also modified:

```

# show ip route
Codes: K - kernel route, C - connected, S - static, R - RIP,
       O - OSPF, I - IS-IS, B - BGP, A - Babel, T - Table,
       > - selected route, * - FIB route
K>* 0.0.0.0/0 via 192.168.0.2, eth0
C>* 6.0.0.5/32 is directly connected, lo
B>* 6.0.0.6/32 [20/0] via fe80::202:ff:fe00:45, swp3, 00:46:12
  *                               via fe80::202:ff:fe00:35, swp1, 00:46:12
  *                               via fe80::202:ff:fe00:3d, swp2, 00:46:12
  *                               via fe80::202:ff:fe00:4d, swp4, 00:46:12
  *                               via fe80::202:ff:fe00:55, swp5, 00:46:12
  *                               via fe80::202:ff:fe00:5a, swp6, 00:46:12
  
```

The following commands show how the IPv4 link-local address 169.254.0.1 is used to install the route and static neighbor entry to facilitate proper forwarding without having to install an IPv4 prefix with IPv6 next-hop in the kernel:

```
# ip route show 6.0.0.6
6.0.0.6 proto zebra metric 20
    nexthop via 169.254.0.1 dev swp3 weight 1 onlink
    nexthop via 169.254.0.1 dev swp1 weight 1 onlink
    nexthop via 169.254.0.1 dev swp2 weight 1 onlink
    nexthop via 169.254.0.1 dev swp4 weight 1 onlink
    nexthop via 169.254.0.1 dev swp5 weight 1 onlink
    nexthop via 169.254.0.1 dev swp6 weight 1 onlink

# ip neigh
fe80::202:ff:fe00:35 dev swp1 lladdr 00:02:00:00:00:35 router REACHABLE
fe80::202:ff:fe00:5a dev swp6 lladdr 00:02:00:00:00:5a router REACHABLE
fe80::202:ff:fe00:3d dev swp2 lladdr 00:02:00:00:00:3d router REACHABLE
fe80::202:ff:fe00:55 dev swp5 lladdr 00:02:00:00:00:55 router REACHABLE
fe80::202:ff:fe00:45 dev swp3 lladdr 00:02:00:00:00:45 router REACHABLE
fe80::202:ff:fe00:4d dev swp4 lladdr 00:02:00:00:00:4d router REACHABLE
169.254.0.1 dev swp5 lladdr 00:02:00:00:00:55 PERMANENT
192.168.0.2 dev eth0 lladdr 52:55:c0:a8:00:02 REACHABLE
169.254.0.1 dev swp3 lladdr 00:02:00:00:00:45 PERMANENT
169.254.0.1 dev swp1 lladdr 00:02:00:00:00:35 PERMANENT
169.254.0.1 dev swp4 lladdr 00:02:00:00:00:4d PERMANENT
169.254.0.1 dev swp6 lladdr 00:02:00:00:00:5a PERMANENT
169.254.0.1 dev swp2 lladdr 00:02:00:00:00:3d PERMANENT
```

## How traceroute Interacts with BGP Unnumbered Interfaces

Every router or end host must have an IPv4 address in order to complete a `traceroute` of IPv4 addresses. In this case, the IPv4 address used is that of the loopback device.

Even if ENHE is not used in the data center, link addresses are not typically advertised. This is because:

- Link addresses take up valuable FIB resources. In a large Clos environment, the number of such addresses can be quite large.
- Link addresses expose an additional attack vector for intruders to use to either break in or engage in DDOS attacks.

Therefore, assigning an IP address to the loopback device is essential.

## Advanced: Understanding How Next-hop Fields Are Set

This section describes how the IPv6 next-hops are set in the MP\_REACH\_NLRI (multiprotocol reachable NLRI) initiated by the system, which applies whether IPv6 prefixes or IPv4 prefixes are exchanged with ENHE. There are two main aspects to determine — how many IPv6 next-hops are included in the MP\_REACH\_NLRI (since the RFC allows either one or two next-hops) and the values of the next-hop(s). This section also describes how a received MP\_REACH\_NLRI is handled as far as processing IPv6 next-hops.

- Whether peering to a global IPv6 address or link-local IPv6 address, the determination whether to send one or two next-hops is as follows:

1. If reflecting the route, two next-hops are sent only if the peer has `nexthop-local unchanged` configured and the attribute of the received route has an IPv6 link-local next-hop; otherwise, only one next-hop is sent.
  2. Otherwise (if it's not reflecting the route), two next-hops are sent if explicitly configured (`nexthop-local unchanged`) or the peer is directly connected (that is, either peering is on link-local address or the global IPv4 or IPv6 address is *directly connected*) and the route is either a local/self-originated route or the peer is an eBGP peer.
  3. In all other cases, only one next-hop gets sent, unless an outbound route-map adds another next-hop.
- `route-map` can impose two next-hops in scenarios where Cumulus Linux would only send one next-hop — by specifying `set ipv6 nexthop link-local`.
  - For all routes to eBGP peers and self-originated routes to iBGP peers, the global next-hop (first value) is the peering address of the local system. If the peering is on the link-local address, this is the global IPv6 address on the peering interface, if present; otherwise, it is the link-local IPv6 address on the peering interface.
  - For other routes to iBGP peers (eBGP to iBGP or reflected), the global next-hop will be the global next-hop in the received attribute.



If this address were a link-local IPv6 address, it would get reset so that the link-local IPv6 address of the eBGP peer is not passed along to an iBGP peer, which most likely may be on a different link.

- `route-map` and/or the peer configuration can change the above behavior. For example, `route-map` can set the global IPv6 next-hop or the peer configuration can set it to `self` — which is relevant for *iBGP* peers. The `route-map` or peer configuration can also set the next-hop to `unchanged`, which ensures the source IPv6 global next-hop is passed around — which is relevant for *eBGP* peers.
- Whenever two next-hops are being sent, the link-local next-hop (the second value of the two) is the link-local IPv6 address on the peering interface unless it is due to `nh-local-unchanged` or `route-map` has set the link-local next-hop.
- Network administrators cannot set **martian values** for IPv6 next-hops in `route-map`. Also, global and link-local next-hops are validated to ensure they match the respective address types.
- In a received update, a martian check is imposed for the IPv6 global next-hop. If the check fails, it gets treated as an implicit withdraw.
- If two next-hops are received in an update and the second next-hop is not a link-local address, it gets ignored and the update is treated as if only one next-hop was received.
- Whenever two next-hops are received in an update, the second next-hop is used to install the route into `zebra`. As per the previous point, it is already assured that this is a link-local IPv6 address. Currently, this is assumed to be reachable and is not registered with NHT.
- When `route-map` specifies the next-hop as `peer-address`, the global IPv6 next-hop as well as the link-local IPv6 next-hop (if it's being sent) is set to the `peering address`. If the peering is on a link-local address, the former could be the link-local address on the peering interface, unless there is a global IPv6 address present on this interface.

The above rules imply that there are scenarios where a generated update has two IPv6 next-hops, and both of them are the IPv6 link-local address of the peering interface on the local system. If you are peering with a switch or router that is not running Cumulus Linux and expects the first next-hop to be a global IPv6 address, a route-map can be used on the sender to specify a global IPv6 address. This conforms with the recommendations in the Internet draft [draft-kato-bgp-ipv6-link-local-00.txt](#), "BGP4+ Peering Using IPv6 Link-local Address".

## ***Limitations***

- Interface-based peering with separate IPv4 and IPv6 sessions is not supported.
- ENHE is sent for IPv6 link-local peerings only.
- If a IPv4 /30 or /31 IP address is assigned to the interface IPv4 peering will be used over IPv6 link-local peering.

## ***Fast Convergence Design Considerations***

Without getting into the why (see the IETF draft cited in Useful Links below that talks about BGP use within the data center), we strongly recommend the following use of addresses in the design of a BGP-based data center network:

- Use of interface addresses: Set up BGP sessions only using interface-scoped addresses. This allows BGP to react quickly to link failures.
- Use of next-hop-self: Every BGP node says that it knows how to forward traffic to the prefixes it is announcing. This reduces the requirement to announce interface-specific addresses and thereby reduces the size of the forwarding table.

## ***Specifying the Interface Name in the neighbor Command***

When you are configuring BGP for the neighbors of a given interface, you can specify the interface name instead of its IP address. All the other `neighbor` command options remain the same.

This is equivalent to BGP peering to the link-local IPv6 address of the neighbor on the given interface. The link-local address is learned via IPv6 neighbor discovery router advertisements.

Consider the following example configuration:

```
router bgp 65000
  bgp router-id 0.0.0.1
  neighbor swp1 interface
  neighbor swp1 remote-as 65000
  neighbor swp1 next-hop-self
!
  address-family ipv6
  neighbor swp1 activate
  exit-address-family
```



Make sure that IPv6 neighbor discovery router advertisements are supported and not suppressed. In Quagga, you do this by checking the running configuration. Under the interface configuration, use `no ipv6 nd suppress-ra` to remove router suppression.

Cumulus Networks recommends you adjust the router advertisement's interval to a shorter value (`ipv6 nd ra-interval <interval>`) to address scenarios when nodes come up and miss router advertisement processing to relay the neighbor's link-local address to BGP. The `interval` is measured in seconds and defaults to 600 seconds.

## Configuring BGP Peering Relationships across Switches

A BGP peering relationship is typically initiated with the `neighbor x.x.x.x remote-as <AS number>` command. In order to simplify configuration across multiple switches, you can specify the *internal* or *external* keyword to the configuration instead of the AS number.

Specifying *internal* signifies an iBGP peering; that is, the neighbor will only create or accept a connection with the specified neighbor if the remote peer AS number matches this BGP's AS number.

Specifying *external* signifies an eBGP peering; that is, the neighbor will only create a connection with the neighbor if the remote peer AS number does **not** match this BGP AS number.

You can make this distinction using the `neighbor` command or the `peer-group` command.

In general, use the following syntax with the `neighbor` command:

```
neighbor (ipv4 addr|ipv6 addr|WORD) remote-as (<1-  
4294967295>|internal|external)
```

Some example configurations follow.

To connect to **the same AS** using the `neighbor` command, modify your configuration similar to the following:

```
router bgp 500  
neighbor 192.168.1.2 remote-as internal
```

To connect to a **different AS** using the `neighbor` command, modify your configuration similar to the following:

```
router bgp 500  
neighbor 192.168.1.2 remote-as external
```

To connect to **the same AS** using the `peer-group` command, modify your configuration similar to the following:

```
router bgp 500
neighbor swp1 interface
neighbor IBGP peer-group
neighbor IBGP remote-as internal
neighbor swp1 peer-group IBGP
neighbor 6.0.0.3 peer-group IBGP
neighbor 6.0.0.4 peer-group IBGP
```

To connect to a **different AS** using the `peer-group` command, modify your configuration similar to the following:

```
router bgp 500
neighbor swp2 interface
neighbor EBGP peer-group
neighbor EBGP remote-as external
neighbor 6.0.0.2 peer-group EBGP
neighbor swp2 peer-group EBGP
neighbor 6.0.0.4 peer-group EBGP
```

## Configuration Tips

### Using `peer-group` to Simplify Configuration

When there are many peers to connect to, the amount of redundant configuration becomes overwhelming. For example, repeating the `activate` and `next-hop-self` commands for even 60 neighbors makes for a very long configuration file. Using `peer-group` addresses this problem.

Instead of specifying properties of each individual peer, Quagga allows for defining one or more peer-groups and associating all the attributes common to that peer session to a peer-group.

After doing this, the only task is to associate an IP address with a peer-group. Here is an example of defining and using peer-groups:

```
R7(config-router)# neighbor tier-2 peer-group
R7(config-router)# neighbor tier-2 remote-as 65000
R7(config-router)# address-family ipv4 unicast
R7(config-router-af)# neighbor tier-2 activate
R7(config-router-af)# neighbor tier-2 next-hop-self
R7(config-router-af)# maximum-paths ibgp 64
R7(config-router-af)# exit
R7(config-router)# neighbor 10.0.0.2 peer-group tier-2
R7(config-router)# neighbor 192.0.2.2 peer-group tier-2
```

If you're using eBGP, besides specifying the neighbor's IP address, you also have to specify the neighbor's ASN, since it is different for each neighbor. In such a case, you wouldn't specify the `remote-as` for the peer-group.

## ***Preserving the AS\_PATH Setting***

If you plan to use multipathing with the `multipath-relax` option, Quagga generates an `AS_SET` in place of the current `AS_PATH` for the bestpath. This helps to prevent loops but is unusual behavior. To preserve the `AS_PATH` setting, use the `no-as-set` option when configuring bestpath:

```
R7(config-router)# bgp bestpath as-path multipath-relax no-as-set
```

## ***Utilizing Multiple Routing Tables and Forwarding***

You can run multiple routing tables (one for in-band/data plane traffic and one for out-of-band /management plane traffic) on the same switch using [management VRF \(see page 381\)](#) (multiple routing tables and forwarding).

## ***Troubleshooting***

The most common starting point for troubleshooting BGP is to view the summary of neighbors connected to and some information about these connections. A sample output of this command is as follows:

```
R7# show ip bgp summary
BGP router identifier 0.0.0.9, local AS number 65000
RIB entries 7, using 672 bytes of memory
Peers 2, using 9120 bytes of memory

Neighbor          V   AS MsgRcvd MsgSent     TblVer  InQ OutQ Up/Down  State
/PfxRcd
10.0.0.2          4 65000      11       10        0     0    0 00:06:38      3
192.0.2.2          4 65000      11       10        0     0    0 00:06:38      3

Total number of neighbors 2
```

(Pop quiz: Are these iBGP or eBGP sessions? Hint: Look at the ASNs.)

It is also useful to view the routing table as defined by BGP:

```
R7# show ip bgp
BGP table version is 0, local router ID is 0.0.0.9
Status codes: s suppressed, d damped, h history, * valid, > best, i -
internal,
r RIB-failure, S Stale, R Removed
```

Origin codes: i - IGP, e - EGP, ? - incomplete

Network	Next Hop	Metric	LocPrf	Weight	Path
*> 192.0.2.29/24	0.0.0.0	0		32768	i
*>i192.0.2.30/24	10.0.0.2	0	100	0	i
* i	192.0.2.2	0	100	0	i
*>i192.0.2.31/24	10.0.0.2	0	100	0	i
* i	192.0.2.2	0	100	0	i
*>i192.0.2.32/24	10.0.0.2	0	100	0	i
* i	192.0.2.2	0	100	0	i

Total number of prefixes 4

A more detailed breakdown of a specific neighbor can be obtained using `show ip bgp neighbor <neighbor ip address>`:

```
R7# show ip bgp neighbor 10.0.0.2
BGP neighbor is 10.0.0.2, remote AS 65000, local AS 65000, internal link
BGP version 4, remote router ID 0.0.0.5
BGP state = Established, up for 00:14:03
Last read 14:52:31, hold time is 180, keepalive interval is 60 seconds
Neighbor capabilities:
  4 Byte AS: advertised and received
  Route refresh: advertised and received(old & new)
  Address family IPv4 Unicast: advertised and received
Message statistics:
  Inq depth is 0
  Outq depth is 0
          Sent      Rcvd
Opens:           1          1
Notifications:   0          0
Updates:         1          3
Keepalives:     16         15
Route Refresh:  0          0
Capability:    0          0
Total:          18         19
Minimum time between advertisement runs is 5 seconds

For address family: IPv4 Unicast
NEXT_HOP is always this router
Community attribute sent to this neighbor(both)
3 accepted prefixes
```

```

Connections established 1; dropped 0
Last reset never
Local host: 10.0.0.1, Local port: 35258
Foreign host: 10.0.0.2, Foreign port: 179
Nexthop: 10.0.0.1
Nexthop global: fe80::202:ff:fe00:19
Nexthop local: ::
BGP connection: non shared network
Read thread: on Write thread: off
  
```

To see the details of a specific route such as from whom it was received, to whom it was sent, and so forth, use the `show ip bgp <ip address/prefix>` command:

```

R7# show ip bgp 192.0.2.0
BGP routing table entry for 192.0.2.0/24
Paths: (2 available, best #1, table Default-IP-Routing-Table)
  Not advertised to any peer
  Local
    10.0.0.2 (metric 1) from 10.0.0.2 (0.0.0.10)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      Originator: 0.0.0.10, Cluster list: 0.0.0.5
      Last update: Mon Jul  8 10:12:17 2013
  Local
    192.0.2.2 (metric 1) from 192.0.2.2 (0.0.0.10)
      Origin IGP, metric 0, localpref 100, valid, internal
      Originator: 0.0.0.10, Cluster list: 0.0.0.6
      Last update: Mon Jul  8 10:12:17 2013
  
```

This shows that the routing table prefix seen by BGP is 192.0.2.0/24, that this route was not advertised to any neighbor, and that it was heard by two neighbors, 10.0.0.2 and 192.0.2.2.

Here is another output of the same command, on a different node in the network:

```

cumulus@switch:~$ sudo vtysh -c 'sh ip bgp 192.0.2.0'
BGP routing table entry for 192.0.2.0/24
Paths: (1 available, best #1, table Default-IP-Routing-Table)
  Advertised to non peer-group peers:
    10.0.0.1 192.0.2.21 192.0.2.22
  Local, (Received from a RR-client)
    203.0.113.1 (metric 1) from 203.0.113.1 (0.0.0.10)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      Last update: Mon Jul  8 09:07:41 2013
  
```

## Debugging Tip: Logging Neighbor State Changes

It is very useful to log the changes that a neighbor goes through to troubleshoot any issues associated with that neighbor. This is done using the `log-neighbor-changes` command:

```
R7(config-router)# bgp log-neighbor-changes
```

The output is sent to the specified log file, usually `/var/log/quagga/bgpd.log`, and looks like this:

```
2013/07/08 10:12:06.572827 BGP: %NOTIFICATION: sent to neighbor 10.0.0.2 6
/3 (Cease/Peer Unconfigured) 0 bytes
2013/07/08 10:12:06.572954 BGP: Notification sent to neighbor 10.0.0.2:
type 6/3
2013/07/08 10:12:16.682071 BGP: %ADJCHANGE: neighbor 192.0.2.2 Up
2013/07/08 10:12:16.682660 BGP: %ADJCHANGE: neighbor 10.0.0.2 Up
```

## Troubleshooting Link-local Addresses

To verify that quagga learned the neighboring link-local IPv6 address via the IPv6 neighbor discovery router advertisements on a given interface, use the `show interface <if-name>` command. If `ipv6 nd suppress-ra` isn't enabled on both ends of the interface, then `Neighbor address(s)` should have the other end's link-local address. That is the address that BGP would use when BGP is enabled on that interface.

Use `vtysh` to run quagga, then verify the configuration:

```
cumulus@switch:~$ sudo vtysh

Hello, this is Quagga (version 0.99.21).
Copyright 1996-2005 Kunihiro Ishiguro, et al.

R7# show interface swp1
Interface swp1 is up, line protocol is up
  PTM status: disabled
  Description: rut
  index 3 metric 1 mtu 1500
  flags: <UP,BROADCAST,RUNNING,MULTICAST>
  HWaddr: 00:02:00:00:00:09
  inet 11.0.0.1/24 broadcast 11.0.0.255
    inet6 fe80::202:ff:fe00:9/64
      ND advertised reachable time is 0 milliseconds
      ND advertised retransmit interval is 0 milliseconds
      ND router advertisements are sent every 600 seconds
```

```
ND router advertisements lifetime tracks ra-interval
ND router advertisement default router preference is medium
Hosts use stateless autoconfig for addresses.
Neighbor address(s):
inet6 fe80::4638:39ff:fe00:129b/128
```

Instead of the IPv6 address, the peering interface name is displayed in the `show ip bgp summary` command and wherever else applicable:

```
R7# show ip bgp summary
BGP router identifier 0.0.0.1, local AS number 65000
RIB entries 1, using 112 bytes of memory
Peers 1, using 8712 bytes of memory

Neighbor          V     AS MsgRcvd MsgSent      TblVer  InQ OutQ Up/Down  State
/PfxRcd
swp1            4  65000       161       170          0      0      0 00:02:28      0
```

Most of the show commands can take the interface name instead of the IP address, if that level of specificity is needed:

```
R7# show ip bgp neighbors
<cr>
A.B.C.D  Neighbor to display information about
WORD      Neighbor on bgp configured interface
X:X::X:X  Neighbor to display information about
R7# show ip bgp neighbors swp1
```

## **Enabling Read-only Mode**

You can enable read-only mode for when the BGP process restarts or when the BGP process is cleared using `clear ip bgp *`. When enabled, read-only mode begins as soon as the first peer reaches its *established* state and a timer for `<max-delay>` seconds is started.

While in read-only mode, BGP doesn't run best-path or generate any updates to its peers. This mode continues until:

- All the configured peers, except the shutdown peers, have sent an explicit EOR (End-Of-RIB) or an implicit EOR. The first keep-alive after BGP has reached the established state is considered an implicit EOR. If the `<establish-wait>` option is specified, then BGP will wait for peers to reach the established state from the start of the `update-delay` until the `<establish-wait>` period is over; that is, the minimum set of established peers for which EOR is expected would be peers established during the `establish-wait` window, not necessarily all the configured neighbors.
- The `max-delay` period is over.

Upon reaching either of these two conditions, BGP resumes the decision process and generates updates to its peers.

To enable read-only mode:

```
cumulus@switch:$ sudo bgp update-delay <max-delay in seconds> [<establish-wait in seconds>]
```

The default <max-delay> is 0 — the feature is off by default.

Use output from `show ip bgp summary` for information about the state of the update delay.

This feature can be useful in reducing CPU/network usage as BGP restarts/clears. It's particularly useful in topologies where BGP learns a prefix from many peers. Intermediate best paths are possible for the same prefix as peers get established and start receiving updates at different times. This feature is also valuable if the network has a high number of such prefixes.

## Applying a Route Map for Route Updates

You can apply a route map on route updates from BGP to Zebra. All the applicable match operations are allowed, such as match on prefix, next-hop, communities, and so forth. Set operations for this attach-point are limited to metric and next-hop only. Any operation of this feature does not affect BGPs internal RIB.

Both IPv4 and IPv6 address families are supported. Route maps work on multi-paths as well. However, the metric setting is based on the best path only.

To apply a route map for route updates:

```
cumulus@switch:$ sudo cl-bgp table-map <route-map-name>
```

## Protocol Tuning

### Converging Quickly On Link Failures

In the Clos topology, we recommend that you only use interface addresses to set up peering sessions. This means that when the link fails, the BGP session is torn down immediately, triggering route updates to propagate through the network quickly. This requires the following commands be enabled for all links: `link-detect` and `ttl-security hops <hops>`. `ttl-security hops` specifies how many hops away the neighbor is. For example, in a Clos topology, every peer is at most 1 hop away.



See Caveats and Errata below for information regarding `ttl-security hops`.

Here is an example:

```
cumulus@switch:~$ sudo vtysh  
  
Hello, this is Quagga (version 0.99.21).
```

```
Copyright 1996-2005 Kunihiro Ishiguro, et al.
```

```
R7# configure terminal
R7(config)# interface swp1
R7(config-if)# link-detect
R7(config-if)# exit
R7(config)# router bgp 65000
R7(config-router)# neighbor 10.0.0.2 ttl-security hops 1
```

## Converging Quickly On Soft Failures

It is possible that the link is up, but the neighboring BGP process is hung or has crashed. If a BGP process crashes, Quagga's `watchquagga` daemon, which monitors the various `quagga` daemons, will attempt to restart it. If the process is also hung, `watchquagga` will attempt to restart the process. BGP itself has a `keepalive` timer that is exchanged between neighbors. By default, this `keepalive` timer is set to 60 seconds. This time can be reduced to a lower number, but this has the disadvantage of increasing the CPU load, especially in the presence of a lot of neighbors. `keepalive-time` is the periodicity with which the `keepalive` message is sent. `hold-time` specifies how many `keepalive` messages can be lost before the connection is considered invalid. It is usually set to 3 times the `keepalive` time. Here is an example of reducing these timers:

```
R7(config-router)# neighbor 10.0.0.2 timers 30 90
```

We can make these the default for all BGP neighbors using a different command:

```
R7(config-router)# timers bgp 30 90
```

The following display snippet shows that the default values have been modified for this neighbor:

```
R7(config-router)# do show ip bgp neighbor 10.0.0.2
BGP neighbor is 10.0.0.2, remote AS 65000, local AS 65000, internal link
  BGP version 4, remote router ID 0.0.0.5
  BGP state = Established, up for 05:53:59
  Last read 14:53:25, hold time is 180, keepalive interval is 60 seconds
  Configured hold time is 90, keepalive interval is 30 seconds
  ....
```



- When you're in a configuration mode, such as when you're configuring BGP parameters, you can run any `show` command by adding `do` to the original command. For example, `do show ip bgp neighbor` was shown above. Under a non-configuration mode, you'd simply run:

```
show ip bgp neighbor 10.0.0.2
```

## Reconnecting Quickly

A BGP process attempts to connect to a peer after a failure (or on startup) every `connect-time` seconds. By default, this is 120 seconds. To modify this value, use:

```
R7(config-router)# neighbor 10.0.0.2 timers connect 30
```

This command has to be specified per each neighbor, peer-group doesn't support this option in quagga.

## Advertisement Interval

BGP by default chooses stability over fast convergence. This is very useful when routing for the Internet. For example, unlike link-state protocols, BGP typically waits for a duration of `advertisement-interval` seconds between sending consecutive updates to a neighbor. This ensures that an unstable neighbor flapping routes won't be propagated throughout the network. By default, this is set to 30 seconds for an eBGP session and 5 seconds for an iBGP session. For very fast convergence, set the timer to 0 seconds. You can modify this as follows:

```
R7(config-router)# neighbor 10.0.0.2 advertisement-interval 0
```

The following output shows the modified value:

```
R7(config-router)# do show ip bgp neighbor 10.0.0.2
BGP neighbor is 10.0.0.2, remote AS 65000, local AS 65000, internal link
  BGP version 4, remote router ID 0.0.0.5
  BGP state = Established, up for 06:01:49
  Last read 14:53:15, hold time is 180, keepalive interval is 60 seconds
  Configured hold time is 90, keepalive interval is 30 seconds
Neighbor capabilities:
  4 Byte AS: advertised and received
  Route refresh: advertised and received(old & new)
  Address family IPv4 Unicast: advertised and received
Message statistics:
  Inq depth is 0
  Outq depth is 0
          Sent          Rcvd
  Opens:           1           1
  Notifications:  0           0
```

```

Updates:          1      3
Keepalives:       363    362
Route Refresh:    0      0
Capability:       0      0
Total:           365    366
Minimum time between advertisement runs is 0 seconds
....
```

 This command is not supported with peer-groups.

See this [IETF draft](#) for more details on the use of this value.

## Configuration Files

- /etc/quagga/daemons
- /etc/quagga/bgpd.conf

## Useful Links

- Bidirectional forwarding detection (see page 367) (BFD) and BGP
- Wikipedia entry for BGP (includes list of useful RFCs)
- Quagga online documentation for BGP (may not be up to date)
- IETF draft discussing BGP use within data centers

## Caveats and Errata

### ttl-security Issue

Enabling `ttl-security` does not cause the hardware to be programmed with the relevant information. This means that frames will come up to the CPU and be dropped there. It is recommended that you use the `cl-acltool` command to explicitly add the relevant entry to hardware.

For example, you can configure a file, like `/etc/cumulus/acl/policy.d/01control_plane_bgp.rules`, with a rule like this for TTL:

```

INGRESS_INTF = swp1
INGRESS_CHAIN = INPUT, FORWARD

[iptables]
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -p tcp --dport bgp -m
ttl --ttl 255 POLICE --set-mode pkt --set-rate 2000 --set-burst 1000
-A $INGRESS_CHAIN --in-interface $INGRESS_INTF -p tcp --dport bgp DROP
```



For more information about ACLs and `c1-acltool`, see [Netfilter \(ACLs\) \(see page 76\)](#).

## Bidirectional Forwarding Detection - BFD

*Bidirectional Forwarding Detection (BFD)* provides low overhead and rapid detection of failures in the paths between two network devices. It provides a unified mechanism for link detection over all media and protocol layers. Use BFD to detect failures for IPv4 and IPv6 single or multihop paths between any two network devices, including unidirectional path failure detection.



Cumulus Linux does not support demand mode in BFD.

### Using BFD Multihop Routed Paths

BFD multihop sessions are built over arbitrary paths between two systems, which results in some complexity that does not exist for single hop sessions. Here are some best practices for using multihop paths:

- **Spoofing:** To avoid spoofing with multihop paths, configure `max_hop_cnt` (maximum hop count) for each peer, which limits the number of hops for a BFD session. All BFD packets exceeding the max hop count will be dropped.
- **Demultiplexing:** Since multihop BFD sessions can take arbitrary paths, demultiplex the initial BFD packet based on the source/destination IP address pair. Use Quagga, which monitors connectivity to the peer, to determine the source/destination IP address pairs.

Multihop BFD sessions are supported for both IPv4 and IPv6 peers. See below for more details.

### BFD Parameters

You can configure the following BFD parameters for both IPv4 and IPv6 sessions:

- The required minimum interval between the received BFD control packets.
- The minimum interval for transmitting BFD control packets.
- The detection time multiplier.

### Configuring BFD

You configure BFD one of two ways: by specifying the configuration in the [PTM topology.dot file \(see page 145\)](#), or using [Quagga \(see page 318\)](#).

The [Quagga CLI \(see page \)](#) can track IPv4 and IPv6 peer connectivity — both single hop and multihop, and both link-local IPv6 peers and global IPv6 peers — using BFD sessions without needing the `topology.dot` file. Use Quagga to register multihop peers with PTM and BFD as well as for monitoring the connectivity to the remote BGP multihop peer. Quagga can dynamically register and unregister both IPv4 and IPv6 peers with BFD when the BFD-enabled peer connectivity is established or de-established, respectively. Also, you can configure BFD parameters for each BGP or OSPF peer using Quagga.



The BFD parameter configured in the topology file is given higher precedence over the client-configured BFD parameters for a BFD session that has been created by both topology file and client (Quagga).

## BFD in BGP

For Quagga when using **BGP**, neighbors are registered and de-registered with [PTM](#) (see page 145) dynamically when you enable BFD in BGP:

```
quagga(config)# router bgp X
quagga(config-router)# neighbor <neighbor ip> bfd
```

You can configure BFD parameters for each BGP neighbor. For example:

### BFD in BGP

```
quagga(config-router)# neighbor <neighbor ip> bfd
  <2-255> Detect Multiplier
  <cr>
quagga(config-router)# neighbor <neighbor ip> bfd 4
  <50-60000> Required min receive interval
quagga(config-router)# neighbor <neighbor ip> bfd 4 400
  <50-60000> Desired min transmit interval
quagga(config-router)# neighbor <neighbor ip> bfd 4 400 400
  <cr>
quagga(config-router)# neighbor <neighbor ip> bfd 4 400 400
```

To see neighbor information in BGP, including BFD status, run `show bgp neighbors <IP address>`.

### Show BGP Neighbor

```
quagga# show bgp neighbors 12.12.12.1
BGP neighbor is 12.12.12.1, remote AS 65001, local AS 65000, external
link
Hostname: r1
  BGP version 4, remote router ID 0.0.0.1
  BGP state = Established, up for 00:01:39
  Last read 00:00:39, Last write 00:01:09
  Hold time is 180, keepalive interval is 60 seconds
  Neighbor capabilities:
    4 Byte AS: advertised and received
    AddPath:
      IPv4 Unicast: RX advertised and received
      Route refresh: advertised and received(old & new)
      Address family IPv4 Unicast: advertised and received
      Hostname Capability: advertised and received
```

Graceful Restart Capabilty: advertised and received

  Remote Restart timer is 120 seconds

  Address families by peer:

    none

  Graceful restart informations:

    End-of-RIB send: IPv4 Unicast

    End-of-RIB received: IPv4 Unicast

  Message statistics:

    Inq depth is 0

    Outq depth is 0

	Sent	Rcvd
Opens:	1	1
Notifications:	0	0
Updates:	2	2
Keepalives:	2	1
Route Refresh:	0	0
Capability:	0	0
Total:	5	4

  Minimum **time** between advertisement runs is 30 seconds

  Update **source** is 12.12.12.7

For address family: IPv4 Unicast

  Update group 1, subgroup 1

  Packet Queue length 0

  NEXT\_HOP is always this router

  Community attribute sent to this neighbor(both)

  1 accepted prefixes

  Connections established 1; dropped 0

  Last reset never

  External BGP neighbor may be up to 2 hops away.

  Local host: 12.12.12.7, Local port: 34274

  Foreign host: 12.12.12.1, Foreign port: 179

  Nexthop: 12.12.12.7

  Nexthop global: ::

  Nexthop **local**: ::

  BGP connection: non shared network

  Read thread: on Write thread: off

  BFD: Type: multi hop

    Detect Mul: 3, Min Rx interval: 300, Min Tx interval: 300

    Status: Down, Last update: 0:00:00:13

## BFD in OSPF

For Quagga using **OSFP**, neighbors are registered and de-registered dynamically with **PTM** (see page 145) when you enable or disable BFD in OSPF. A neighbor is registered with BFD when two-way adjacency is established and deregistered when adjacency goes down if the BFD is enabled on the interface. The BFD configuration is per interface and any IPv4 and IPv6 neighbors discovered on that interface inherit the configuration.

### BFD in OSPF

```

quagga(config)# interface X
quagga(config-if)# ipv6 ospf6 bfd
  <2-255> Detect Multiplier
  <cr>
quagga(config-if)# ipv6 ospf6 bfd 5
  <50-60000> Required min receive interval
quagga(config-if)# ipv6 ospf6 bfd 5 500
  <50-60000> Desired min transmit interval
quagga(config-if)# ipv6 ospf6 bfd 5 500 500
  <cr>
quagga(config-if)# ipv6 ospf6 bfd 5 500 500
  
```

## OSPF Show Commands

The BFD lines at the end of each code block shows the corresponding IPv6 or IPv4 OSPF interface or neighbor information.

### Show IPv6 OSPF Interface

```

quagga# show ipv6 ospf6 interface swp2s0
swp2s0 is up, type BROADCAST
  Interface ID: 4
  Internet Address:
    inet : 11.0.0.21/30
    inet6: fe80::4638:39ff:fe00:6c8e/64
  Instance ID 0, Interface MTU 1500 (autodetect: 1500)
  MTU mismatch detection: enabled
  Area ID 0.0.0.0, Cost 10
  State PointToPoint, Transmit Delay 1 sec, Priority 1
  Timer intervals configured:
    Hello 10, Dead 40, Retransmit 5
  DR: 0.0.0.0 BDR: 0.0.0.0
  Number of I/F scoped LSAs is 2
    0 Pending LSAs for LSUpdate in Time 00:00:00 [thread off]
    0 Pending LSAs for LSAck in Time 00:00:00 [thread off]
  BFD: Detect Mul: 3, Min Rx interval: 300, Min Tx interval: 300
  
```

### Show IPv6 OSPF Neighbor

```

quagga# show ipv6 ospf6 neighbor detail
Neighbor 0.0.0.4%swp2s0
  Area 0.0.0.0 via interface swp2s0 (ifindex 4)
  His IfIndex: 3 Link-local address: fe80::202:ff:fe00:a
  State Full for a duration of 02:32:33
  His choice of DR/BDR 0.0.0.0/0.0.0.0, Priority 1
  DbDesc status: Slave SeqNum: 0x76000000
  Summary-List: 0 LSAs
  
```

```

Request-List: 0 LSAs
Retrans-List: 0 LSAs
0 Pending LSAs for DbDesc in Time 00:00:00 [thread off]
0 Pending LSAs for LSReq in Time 00:00:00 [thread off]
0 Pending LSAs for LSUpdate in Time 00:00:00 [thread off]
0 Pending LSAs for LSAck in Time 00:00:00 [thread off]
BFD: Type: single hop
    Detect Mul: 3, Min Rx interval: 300, Min Tx interval: 300
    Status: Up, Last update: 0:00:00:20

```

#### Show IPv4 OSPF Interface

```

quagga# show ip ospf interface swp2s0
swp2s0 is up
    ifindex 4, MTU 1500 bytes, BW 0 Kbit <UP,BROADCAST,RUNNING,
MULTICAST>
    Internet Address 11.0.0.21/30, Area 0.0.0.0
    MTU mismatch detection:enabled
    Router ID 0.0.0.3, Network Type POINTOPOINT, Cost: 10
    Transmit Delay is 1 sec, State Point-To-Point, Priority 1
    No designated router on this network
    No backup designated router on this network
    Multicast group memberships: OSPFAllRouters
    Timer intervals configured, Hello 10s, Dead 40s, Wait 40s,
Retransmit 5
        Hello due in 7.056s
    Neighbor Count is 1, Adjacent neighbor count is 1
    BFD: Detect Mul: 5, Min Rx interval: 500, Min Tx interval: 500

```

#### Show IPv4 OSPF Neighbor

```

quagga# show ip ospf neighbor detail
Neighbor 0.0.0.4, interface address 11.0.0.22
    In the area 0.0.0.0 via interface swp2s0
    Neighbor priority is 1, State is Full, 5 state changes
    Most recent state change statistics:
        Progressive change 3h59m04s ago
    DR is 0.0.0.0, BDR is 0.0.0.0
    Options 2 *|-|-|---|E|*
    Dead timer due in 38.501s
    Database Summary List 0
    Link State Request List 0
    Link State Retransmission List 0
    Thread Inactivity Timer on
    Thread Database Description Retransmission off
    Thread Link State Request Retransmission on
    Thread Link State Update Retransmission on
    BFD: Type: single hop

```

```
Detect Mul: 5, Min Rx interval: 500, Min Tx interval: 500
Status: Down, Last update: 0:00:01:29
```

## Troubleshooting BFD

To troubleshoot BFD, use `ptmctl -b`. For more information, see [Prescriptive Topology Manager - PTM](#) (see page 145).

## Equal Cost Multipath Load Sharing - Hardware ECMP

Cumulus Linux supports hardware-based equal cost multipath (ECMP) load sharing. ECMP is enabled by default in Cumulus Linux. Load sharing occurs automatically for all routes with multiple next hops installed. ECMP load sharing supports both IPv4 and IPv6 routes.



ECMP is not supported in Cumulus RMP.

## Contents

(Click to expand)

- [Contents \(see page 372\)](#)
- [Understanding Equal Cost Routing \(see page 372\)](#)
- [Understanding ECMP Hashing \(see page 373\)
  - \[Using cl-ecmpcalc to Determine the Hash Result \\(see page 373\\)\]\(#\)
  - \[cl-ecmpcalc Limitations \\(see page 374\\)\]\(#\)
  - \[ECMP Hash Buckets \\(see page 374\\)\]\(#\)](#)
- [Resilient Hashing \(see page 376\)
  - \[Resilient Hash Buckets \\(see page 377\\)\]\(#\)
  - \[Removing Next Hops \\(see page 377\\)\]\(#\)
  - \[Adding Next Hops \\(see page 379\\)\]\(#\)
  - \[Configuring Resilient Hashing \\(see page 379\\)\]\(#\)](#)
- [Caveats \(see page 380\)](#)
- [Useful Links \(see page 380\)](#)

## Understanding Equal Cost Routing

ECMP operates only on equal cost routes in the Linux routing table.

In this example, the 10.1.1.0/24 route has two possible next hops that have been installed in the routing table:

```
$ ip route show 10.1.1.0/24
10.1.1.0/24 proto zebra metric 20
```

```
nexthop via 192.168.1.1 dev swp1 weight 1 onlink
nexthop via 192.168.2.1 dev swp2 weight 1 onlink
```

For routes to be considered equal they must:

- Originate from the same routing protocol. Routes from different sources are not considered equal. For example, a static route and an OSPF route are not considered for ECMP load sharing.
- Have equal cost. If two routes from the same protocol are unequal, only the best route is installed in the routing table.



BGP does not install multiple routes by default. To do so, use the `maximum-paths` command. See the [ECMP section \(see page 345\)](#) of the BGP chapter for more information.

## ***Understanding ECMP Hashing***

Once multiple routes are installed in the routing table, a hash is used to determine which path a packet follows.

Cumulus Linux hashes on the following fields:

- IP protocol
- Ingress interface
- Source IPv4 or IPv6 address
- Destination IPv4 or IPv6 address

For TCP/UDP frames, Cumulus Linux also hashes on:

- Source port
- Destination port

ECMP Hash Fields					
Source IP	Destination IP	Layer 4 Protocol	Source Port	Destination Port	Payload

To prevent out of order packets, ECMP hashing is done on a per-packet basis. However, all packets with the same source and destination IP addresses and the same source and destination ports always hash to the same next hop. ECMP hashing does not keep a record of flow states.

ECMP hashing does not keep a record of packets that have hashed to each next hop and does not guarantee that traffic sent to each next hop is equal.

## ***Using `cl-ecmpcalc` to Determine the Hash Result***

Since the hash is deterministic and always provides the same result for the same input, you can query the hardware and determine the hash result of a given input. This is useful when determining exactly which path a flow takes through a network.

On Cumulus Linux, use the `cl-ecmpcalc` command to determine a hardware hash result.

In order to use `cl-ecmpcalc`, all fields that are used in the hash must be provided. This includes ingress interface, layer 3 source IP, layer 3 destination IP, layer 4 source port and layer 4 destination port.

```
$ sudo cl-ecmpcalc -i swp1 -s 10.0.0.1 -d 10.0.0.1 -p tcp --sport 20000 --  
dport 80  
ecmpcalc: will query hardware  
swp3
```

If any field is omitted, `cl-ecmpcalc` fails.

```
$ sudo cl-ecmpcalc -i swp1 -s 10.0.0.1 -d 10.0.0.1 -p tcp  
ecmpcalc: will query hardware  
usage: cl-ecmpcalc [-h] [-v] [-p PROTOCOL] [-s SRC] [--sport SPORT] [-d  
DST]  
                  [--dport DPORT] [--vid VID] [-i IN_INTERFACE]  
                  [--sportid SPORTID] [--smodid SMODID] [-o OUT_INTERFACE]  
                  [--dportid DPORTID] [--dmodid DMODID] [--hardware]  
                  [--nohardware] [-hs HASHSEED]  
                  [-hf HASHFIELDS [HASHFIELDS ...]]  
                  [--hashfunction {crc16-ccitt,crc16-bisync}] [-e EGRESS]  
                  [-c MCOUNT]
```

```
cl-ecmpcalc: error: --sport and --dport required for TCP and UDP frames
```

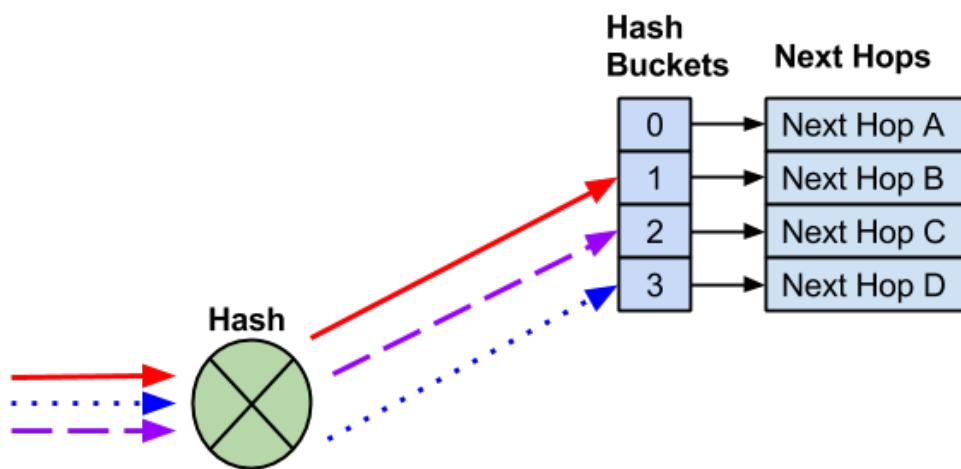
## ***cl-ecmpcalc Limitations***

`cl-ecmpcalc` can only take input interfaces that can be converted to a single physical port in the port tab file, like the physical switch ports (`swp`). Virtual interfaces like bridges, bonds, and subinterfaces are not supported.

## ***ECMP Hash Buckets***

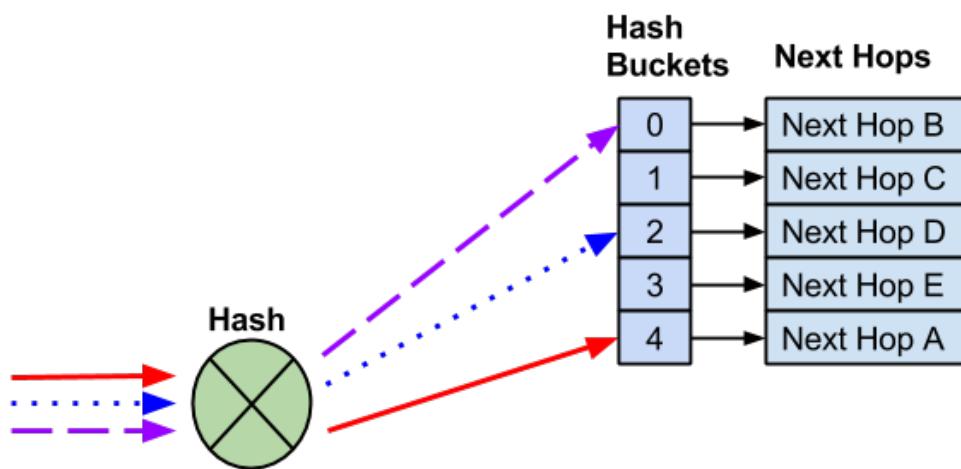
When multiple routes are installed in the routing table, each route is assigned to an ECMP *bucket*. When the ECMP hash is executed the result of the hash determines which bucket gets used.

In the following example, 4 next hops exist. Three different flows are hashed to different hash buckets. Each next hop is assigned to a unique hash bucket.



### ***Adding a Next Hop***

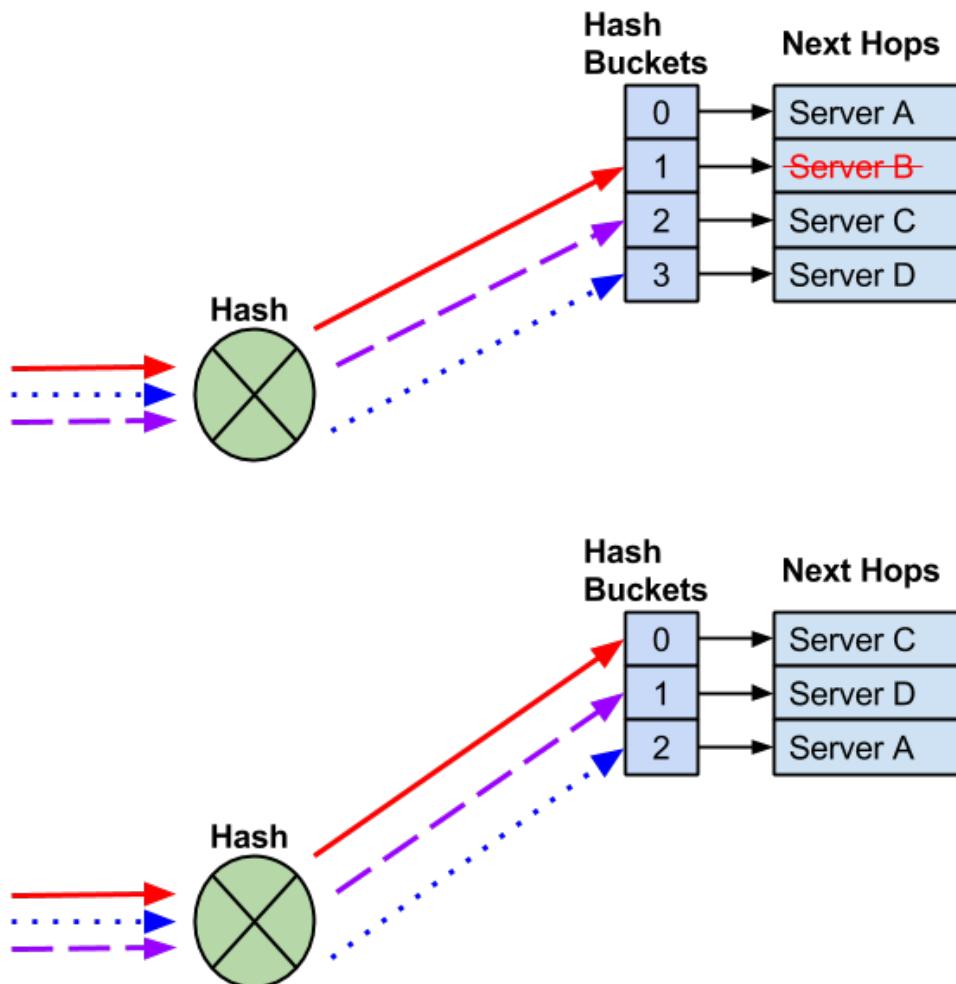
When a next hop is added, a new hash bucket is created. The assignment of next hops to hash buckets, as well as the hash result, may change when additional next hops are added.



A new next hop is added and a new hash bucket is created. As a result, the hash and hash bucket assignment changed, causing the existing flows to be sent to different next hops.

### ***Removing a Next Hop***

When a next hop is removed, the remaining hash bucket assignments may change, again, potentially changing the next hop selected for an existing flow.



A next hop fails and the next hop and hash bucket are removed. The remaining next hops may be reassigned.

In most cases, the modification of hash buckets has no impact on traffic flows as traffic is being forward to a single end host. In deployments where multiple end hosts are using the same IP address (anycast), *resilient hashing* must be used.

## Resilient Hashing

In Cumulus Linux when a next hop fails or is removed from an ECMP pool, the hashing or hash bucket assignment can change. For deployments where there is a need for flows to always use the same next hop, like TCP anycast deployments, this can create session failures.

The ECMP hash performed with resilient hashing is exactly the same as the default hashing mode. Only the method in which next hops are assigned to hash buckets differs.

Resilient hashing supports both IPv4 and IPv6 routes.

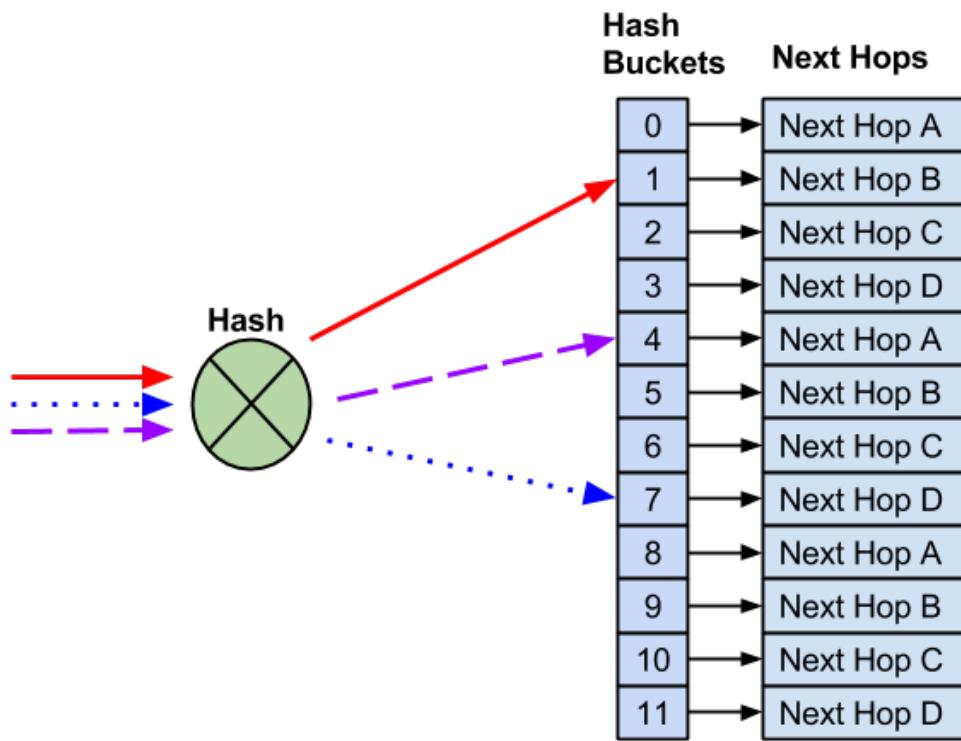
Resilient hashing is not enabled by default. See below for steps on configuring it.



Resilient hashing prevents disruptions when new next hops are removed. It does not prevent disruption when next hops are added.

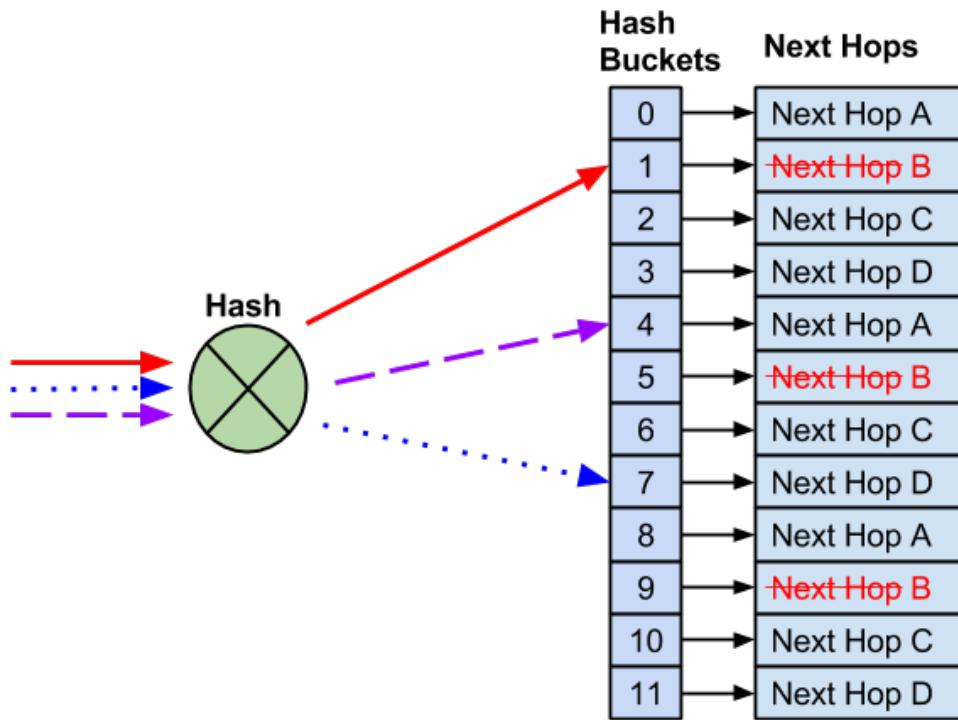
## Resilient Hash Buckets

When resilient hashing is configured, a fixed number of buckets are defined. Next hops are then assigned in round robin fashion to each of those buckets. In this example, 12 buckets are created and four next hops are assigned.

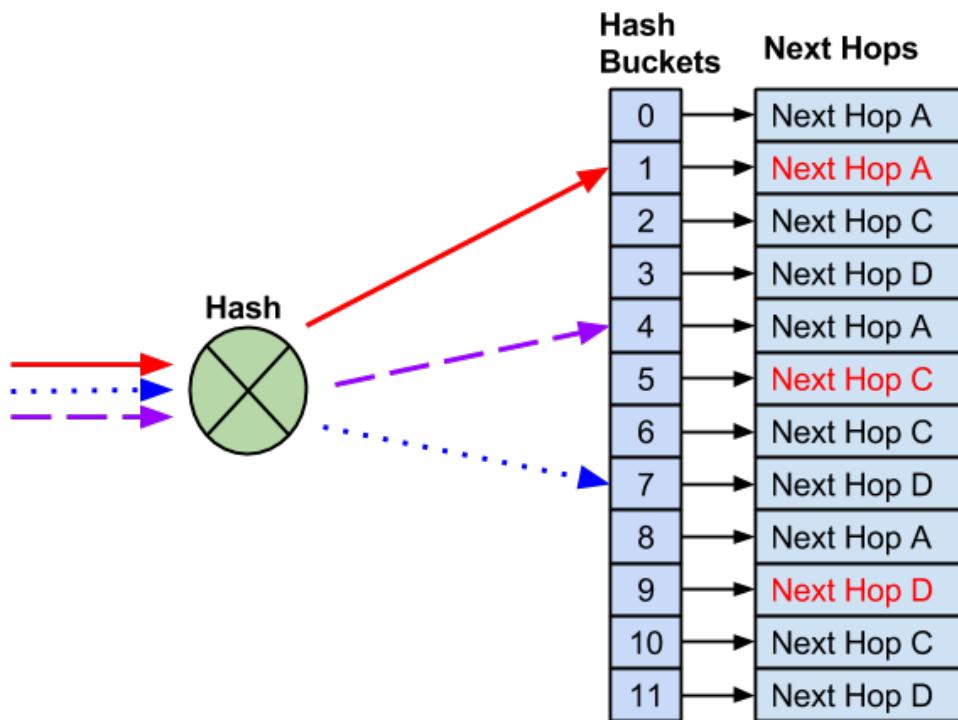


## Removing Next Hops

Unlike default ECMP hashing, when a next hop needs to be removed, the number of hash buckets does not change.



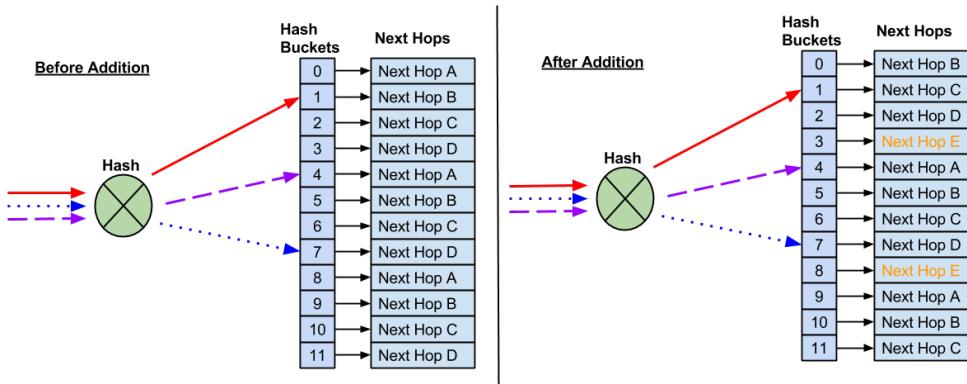
With 12 buckets assigned and four next hops, instead of reducing the number of buckets — which would impact flows to known good hosts — the remaining next hops replace the failed next hop.



After the failed next hop is removed, the remaining next hops are installed as replacements. This prevents impact to any flows that hash to working next hops.

## Adding Next Hops

Resilient hashing does not prevent possible impact to existing flows when new next hops are added. Due to the fact there are a fixed number of buckets, a new next hop requires reassigning next hops to buckets.



As a result, some flows may hash to new next hops, which can impact anycast deployments.

## Configuring Resilient Hashing

Resilient hashing is not enabled by default. When resilient hashing is enabled, 65,536 buckets are created to be shared among all ECMP routes.



An ECMP route counts as a single route with multiple next hops. The following example is considered to be a single ECMP route:

```
$ ip route show 10.1.1.0/24
10.1.1.0/24 proto zebra metric 20
nexthop via 192.168.1.1 dev swp1 weight 1 onlink
nexthop via 192.168.2.1 dev swp2 weight 1 onlink
```

All ECMP routes must use the same number of buckets (the number of buckets cannot be configured per ECMP route).

The number of buckets can be configured as 64, 128, 256, 512 or 1024; the default is 128:

Number of Hash Buckets	Number of Supported ECMP Routes
64	1024
<b>128</b>	<b>512</b>

Number of Hash Buckets	Number of Supported ECMP Routes
256	256
512	128
1024	64

A larger number of ECMP buckets reduces the impact on adding new next hops to an ECMP route. However, the system supports fewer ECMP routes. If the maximum number of ECMP routes have been installed, new ECMP routes log an error and are not installed.

To enable resilient hashing, edit `/etc/cumulus/datapath/traffic.conf`:

1. Enable resilient hashing:

```
# Enable resilient hashing
resilient_hash_enable = TRUE
```

2. **(Optional)** Edit the number of hash buckets:

```
# Resilient hashing flowset entries per ECMP group
# Valid values - 64, 128, 256, 512, 1024
resilient_hash_entries_ecmp = 256
```

3. Restart the `switchd` service:

```
cumulus@switch:~$ sudo service switchd restart
```

## Caveats

Resilient hashing is only supported on switches with the [Trident II chipsets](#). You can run `netshow system` to determine the chipset.

## Useful Links

- [http://en.wikipedia.org/wiki/Equal-cost\\_multi-path\\_routing](http://en.wikipedia.org/wiki/Equal-cost_multi-path_routing)

## Management VRF

*Management VRF* (multiple routing tables and forwarding) provides routing separation between the out-of-band management network and the in-band data plane network. When management VRF is enabled, applications running on control plane processor communicate out from the management network unless configured otherwise.

Management VRF creates two routing tables within the Linux kernel:

- *main*: This is the routing table for all the data plane switch ports.
- *mgmt*: This is the routing table for eth0.

Cumulus Linux only supports eth0 as the management interface. VLAN subinterfaces, bonds, bridges and the front panel switch ports are not supported as management interfaces.

Management VRF assumes all traffic *generated by the switch* (except via Quagga) will exit eth0 by default, so unless there is application-level intervention, any packet generated by an application on the switch will only reference the eth0 routing table (the *mgmt* table). Applications that need to communicate over the data plane network (the *main* table) **must** bind to the loopback IP address.

For example, if the switch is responding to an inbound SSH connection or inbound ping, management VRF does not force the traffic out through eth0. However, if you attempt to SSH from the switch outbound, then management VRF will force the traffic to exit eth0, unless you specify otherwise. For example, when initiating an SSH connection, you can use `-b <loopback IP address>` to SSH to a device via the data plane network.

### Enabling Management VRF

To enable management VRF, complete the following steps:

1. Update the `apt` source list:

```
$ sudo apt-get update
```

2. Install the management VRF package:

```
sudo apt-get install cl-mgmtvrf
```

3. Run the management VRF script:

```
sudo cl-mgmtvrf --enable
```



Management VRF has hooks in the eth0 DHCP client to force the correct mgmt table routes when the DHCP address is obtained. If you use static IP address assignment on eth0, you have to manually configure the routes before you execute this step. See the 'Using Static IP Addresses on eth0' section below for more information.

#### 4. Restart Quagga:

```
sudo service quagga restart
```



You can also bounce adjacency to the peer advertising the default route to get the default route from the data plane network into the main routing table.

## Verifying Management VRF

To check the status of management VRF, run:

```
cl-mgmtvrf --status
```

This will display `cl-mgmtvrf` is NOT enabled or `cl-mgmtvrf` is enabled, depending upon whether management VRF is disabled or enabled.

## Disabling Management VRF

To disable management VRF, run:

```
sudo cl-mgmtvrf --disable
```



If management VRF is disabled and the data plane adds a default route, the default route via the management interface will **not** be added to main routing table.

## Using ping or traceroute

By default, issuing a `ping` or `traceroute` assumes the packet should be sent to the dataplane network (the main routing table). If you wish to use `ping` or `traceroute` on the control plane network, use the `-I` flag for `ping` and `-i` for `traceroute`.

```
ping -I eth0
```

or

```
sudo traceroute -i eth0
```



DNS does not work with `traceroute` or `ping` unless you explicitly add support for the DNS server in the *main* routing table.

## **OSPF and BGP**

No changes are required for either BGP or OSPF. Quagga has been updated in Cumulus Linux 2.5.3 to be aware of the management VRF and automatically sends packets based on the switch port routing table. This includes BGP peering via loopback interfaces. BGP does routing lookups in the default table.

## **SNMP and sFlow**

Both SNMP and sFlow do not currently have a method to use a switch port to send data. For any netflow collectors or SNMP traps, this traffic gets sent out to eth0. Cumulus Networks will support switch ports in the future.

**Note:** For SNMP, this restriction only applies to traps. SNMP polling is not affected.

## **SSH**

If you SSH to the switch through a switch port, it works as expected. If you need to SSH from the device out a switch port, use `ssh -b <ip_address_of_swp_port>`. For example:

```
cumulus@leaf1$ ip addr show swp17
19: swp17: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
state UP qlen 500
    link/ether ec:f4:bb:fc:19:23 brd ff:ff:ff:ff:ff:ff
    inet 10.23.23.2/24 scope global swp17
        inet6 fe80::eef4:bbff:fe:1923/64 scope link
            valid_lft forever preferred_lft forever

cumulus@leaf1$ ssh -b 10.23.23.2 10.3.3.3
```

## **Viewing the Routing Tables**

When you look at the routing table with `ip route show`, you are looking at the switch port (*main*) table. You can also see the dataplane routing table with `ip route show table main`.

To look at information about eth0 (the management routing table), use `ip route show table mgmt`.

```
cumulus@leaf1$ ip route show table mgmt
default via 192.168.0.1 dev eth0
```

```
cumulus@leaf1$ ip route show
default via 10.23.23.3 dev swp17 proto zebra metric 20
10.3.3.3 via 10.23.23.3 dev swp17
10.23.23.0/24 dev swp17 proto kernel scope link src 10.23.23.2
192.168.0.0/24 dev eth0 proto kernel scope link src 192.168.0.11
```

## Viewing a Single Route

Note that if you use `ip route get` to return information about a single route, the command resolves over the `mgmt` table by default. To get information about the route in the switching silicon, use:

```
ip route get <addr> from <loopback IP>
```

Or:

```
sudo cl-rtcl ip route show <addr>
```

## Using Static IP Addresses on eth0

If you're using DHCP on your management network, the Management VRF feature has hooks in the `eth0` DHCP client to automatically add the correct default and interface routes into the `mgmt` table when the DHCP address is obtained.

If a static IP address is used in the `eth0` definition, you must manually control the connected and static routes attached to `eth0` *before running `cl-mgmtvrf --enable`*

To do this change the configuration in the `/etc/network/interfaces` as follows:

```
auto eth0
iface eth0 inet static
    address 192.1.1.254/24
    post-up ip route add 192.1.1.0/24 dev eth0 table mgmt
    post-up ip route add default via 192.1.1.1 dev eth0 table mgmt
    post-up ip route del 192.1.1.0/24 dev eth0 table main
    post-down ip route del 192.1.1.0/24 dev eth0 table mgmt
    post-down ip route del default via 192.1.1.1 dev eth0 table mgmt
```

Then bounce `eth0`:

```
sudo ifdown eth0; sudo ifup eth0
```

Enabling management VRF via `cl-mgmtvrf --enable` after this step should lead to the expected routing behavior.



The `post-down` commands are there to ensure that no routing race condition can occur on an interface experiencing route flapping. As a result, the following error messages during a link flap are harmless and can be ignored:

```
warning: eth0: post-down cmd 'ip route del 192.1.1.0/24 dev eth0  
table mgmt' failed (RTNETLINK answers: No such process)  
warning: eth0: post-down cmd 'ip route del default 192.1.1.1 via  
eth0 table mgmt' failed (Error: either "to" is duplicate, or  
"192.1.1.1" is a garbage.)
```

## Incompatibility with `cl-ns-mgmt`



If you are using the Cumulus Linux [management namespace](#) feature (via the `cl-ns-mgmt` utility), you cannot enable management VRF, as the two features are incompatible. Management VRF does not run if Cumulus Linux detects that you have management namespaces enabled, and vice versa.

## Log Files

`/var/log/cl-mgmtvrf.log`

## Caveats and Errata

- If you are using an [MLAG configuration \(see page 191\)](#) when the `eth0` management interface is enabled, you cannot specify a backup link (via `clagd-backup-ip`) over the switch ports.
- Duplicate IP addresses are not supported i.e. you cannot have the same IP address in both the management network and the data network.
- DHCP relay does not work with `cl-mgmtvrf` when the host and server facing ports are in the data plane. For more information, refer to the [knowledge base article](#).

# Monitoring and Troubleshooting

This chapter introduces monitoring and troubleshooting Cumulus Linux.

## Contents

(Click to expand)

- [Contents \(see page 386\)](#)
- [Commands \(see page 386\)](#)
- [Using the Serial Console \(see page 386\)
  - \[Configuring the Serial Console on PowerPC or ARM Switches \\(see page 386\\)\]\(#\)
  - \[Configuring the Serial Console on x86 Switches \\(see page 387\\)\]\(#\)](#)
- [Diagnostics Using cl-support \(see page 388\)](#)
- [Sending Log Files to a syslog Server \(see page 389\)](#)
- [Next Steps \(see page 391\)](#)

## Commands

- [cl-support](#)
- [fw\\_setenv](#)

## Using the Serial Console

The serial console can be a useful tool for debugging issues, especially when you find yourself rebooting the switch often or if you don't have a reliable network connection.

The default serial console baud rate is 115200, which is the baud rate ONIE uses.

### ***Configuring the Serial Console on PowerPC or ARM Switches***

On PowerPC switches, the U-Boot environment variable `baudrate` identifies the baud rate of the serial console. To change the `baudrate` variable, use the `fw_setenv` command:

```
cumulus@switch:~$ sudo fw_setenv baudrate 9600
Updating environment variable: `baudrate'
Proceed with update [N/y]? y
```

You must reboot the switch for the `baudrate` change to take effect.

The valid values for `baudrate` are:

- 300

- 600
- 1200
- 2400
- 4800
- 9600
- 19200
- 38400
- 115200

## Configuring the Serial Console on x86 Switches

On x86 switches, you configure serial console baud rate by editing `grub`.



Incorrect configuration settings in `grub` can cause the switch to be inaccessible via the console. Grub changes should be carefully reviewed before implementation.

The valid values for the baud rate are:

- 300
- 600
- 1200
- 2400
- 4800
- 9600
- 19200
- 38400
- 115200

To change the serial console baud rate:

1. Edit `/etc/default/grub`. The two relevant lines in `/etc/default/grub` are as follows; replace the 115200 value with a valid value specified above in the `--speed` variable in the first line and in the `console` variable in the second line:

```
GRUB_SERIAL_COMMAND="serial --port=0x2f8 --speed=115200 --word=8 --
parity=no --stop=1"
GRUB_CMDLINE_LINUX="console=ttyS1,115200n8
cl_platform=accton_as5712_54x"
```

2. After you save your changes to the grub configuration, type the following at the command prompt:

```
cumulus@switch:~$ update-grub
```

3. If you plan on accessing your switch's BIOS over the serial console, you need to update the baud rate in the switch BIOS. For more information, see [this knowledge base article](#).
4. Reboot the switch.

## Diagnostics Using cl-support

You can use `cl-support` to generate a single export file that contains various details and the configuration from a switch. This is useful for remote debugging and troubleshooting.

You should run `cl-support` before you submit a support request to Cumulus Networks as this file helps in the investigation of issues:

```
cumulus@switch:~$ sudo cl-support -h
Usage: cl-support [-h] [reason]...
Args:
[reason]: Optional reason to give for invoking cl-support.
          Saved into tarball's reason.txt file.
Options:
-h: Print this usage statement
```

Example output:

```
cumulus@switch:~$ ls /var/support
cl_support_20130806_032720.tar.xz
```

The directory structure is compressed using LZMA2 compression and can be extracted using the `unxz` command:

```
cumulus@switch:~$ cd /var/support
cumulus@switch:~$ sudo unxz cl_support_20130729_140040.tar.xz
cumulus@switch:~$ sudo tar xf cl_support_20130729_140040.tar
cumulus@switch:~$ ls -l cl_support_20130729_140040/
-rwxr-xr-x 1 root root 7724 Jul 29 14:00 cl-support
-rw-r--r-- 1 root root 52 Jul 29 14:00 cmdline.args
drwxr-xr-x 2 root root 4096 Jul 29 14:00 core
drwxr-xr-x 64 root root 4096 Jul 29 13:51 etc
drwxr-xr-x 4 root root 4096 Jul 29 14:00 proc
drwxr-xr-x 2 root root 4096 Jul 29 14:01 support
drwxr-xr-x 3 root root 4096 Jul 29 14:00 sys
drwxr-xr-x 3 root root 4096 Aug  8 15:22 var
```

The directory contains the following elements:

Directory	Description
core	Contains the core files generated from Cumulus Linux HAL process, <code>switchd</code> .
etc	Is a replica of the switch's <code>/etc</code> directory. <code>/etc</code> contains all the general Linux configuration files, as well as configurations for the system's network interfaces, <code>quagga</code> , <code>jdoe</code> , and other packages.
log	Is a replica of the switch's <code>/var/log</code> directory. Most Cumulus Linux log files are located in this directory. Notable log files include <code>switchd.log</code> , <code>daemon.log</code> , <code>quagga</code> log files, and <code>syslog</code> . For more information, read this <a href="#">knowledge base article</a> .
proc	Is a replica of the switch's <code>/proc</code> directory. In Linux, <code>/proc</code> contains runtime system information (like system memory, devices mounted, and hardware configuration). These files are not actual files but the current state of the system.
support	Is a set of files containing further system information, which is obtained by <code>c1-support</code> running commands such as <code>ps -aux</code> , <code>netstat -i</code> , and so forth — even the routing tables.

`c1-support`, when untarred, contains a `reason.txt` file. This file indicates what reason triggered it. When contacting Cumulus Networks technical support, please attach the `c1-support` file if possible. For more information about `c1-support`, read [Understanding and Decoding the c1-support Output File](#) (see page 417).

## Sending Log Files to a syslog Server

All logging on Cumulus Linux is done with `rsyslog`. `rsyslog` provides both local logging to the `syslog` file as well as the ability to export logs to an external `syslog` server.

**Local logging:** Most logs within Cumulus Linux are sent to files in the `/var/log` directory. Most relevant information is placed within the `/var/log/syslog` file. For more information on specific log files, see [Troubleshooting Log Files](#) (see page 419).

**Export logging:** To send `syslog` files to an external `syslog` server, add a rule specifying to copy all messages (\*) to the IP address and switch port of your `syslog` server in the `rsyslog` configuration files as described below.

In the following example, 192.168.1.2 is the remote `syslog` server and 514 is the port number. For UDP-based `syslog`, use a single @ before the IP address: @192.168.1.2:514. For TCP-based `syslog`, use two @@ before the IP address: @@192.168.1.2:514.

1. Create a file called something like `/etc/rsyslog.d/90-remotesyslog.conf`. Make sure it starts with a number lower than 99 so that it executes before `99-rsyslog.conf`. Add content like the following:

```
## Copy all messages to the remote syslog server at 192.168.1.2 port  
514  
*.* @192.168.1.2:514
```

## 2. Restart rsyslog.

```
service rsyslog restart
```



Starting with Cumulus Linux 2.5.4, all Cumulus Linux rules have been moved from `/etc/rsyslog.conf` into separate files in `/etc/rsyslog.d/`, which are called at the end of the GLOBAL DIRECTIVES section of `/etc/rsyslog.conf`. As a result, the RULES section at the end of `rsyslog.conf` is ignored because the messages have to be processed by the rules in `/etc/rsyslog.d` and then dropped by the last line in `/etc/rsyslog.d/99-syslog.conf`.



In the case of the `switchd` rules file, the file must be numbered lower than 25. For example, `13-switchd-remote.conf`.

If you need to send other log files (e.g. `switchd` logs) to a `syslog` server, configure a new file in `/etc/rsyslog.d`, as described above, and add lines similar to the following lines:

```
## Logging switchd messages to remote syslog server  
$ModLoad imfile  
$InputFileName /var/log/switchd.log  
$InputFileStateFile logfile-log  
$InputFileTag switchd:  
$InputFileSeverity info  
$InputFileFacility local7  
$InputFilePollInterval 5  
$InputRunFileMonitor  
  
if $programname == 'switchd' then @192.168.1.2:514
```

Then restart `syslog`:

```
service rsyslog restart
```

In the above configuration, each setting is defined as follows:

Setting	Description
\$ModLoad <i>imfile</i>	Enables the <code>rsyslog</code> module to watch file contents.
\$InputFileName	The file to be sent to the <code>syslog</code> server. In this example, you are going to send changes made to <code>/var/log/switchd.log</code> to the <code>syslog</code> server.
\$InputFileStateFile	This is used by <code>rsyslog</code> to track state of the file being monitored. This must be unique for each file being monitored.
\$InputFileTag	Defines the <code>syslog</code> tag that will precede the <code>syslog</code> messages. In this example, all logs are prefaced with <code>switchd</code> .
\$InputFileSeverity	Defines the logging severity level sent to the <code>syslog</code> server.
\$InputFileFacility	Defines the logging format. <code>local7</code> is common.
\$InputFilePollInterval	Defines how frequently in seconds <code>rsyslog</code> looks for new information in the file. Lower values provide faster updates but create slightly more load on the CPU.
\$InputRunFileMonitor	Enables the file monitor module with the configured settings.

In most cases, the settings to customize include:

Setting	Description
\$InputFileName	The file to stream to the <code>syslog</code> server.
\$InputFileStateFile	A unique name for each file being watched.
\$InputFileTag	A prefix to the log message on the server.

Finally, the `if $programname` line is what sends the log files to the `syslog` server. It follows the same syntax as the `/var/log/syslog` file, where @ indicates UDP, 192.168.1.2 is the IP address of the `syslog` server, and 514 is the UDP port. The value `switchd` must match the value in `$InputFileTag`.

## Next Steps

The links below discuss more specific monitoring topics.

## Single User Mode - Boot Recovery

Use single user mode to assist in troubleshooting system boot issues or for password recovery. Entering single user mode is [platform-specific](#), so follow the appropriate steps for your x86, ARM or PowerPC switch.

### Contents

(Click to expand)

- [Contents \(see page 392\)](#)
- [Entering Single User Mode on a PowerPC or ARM Switch \(see page 392\)](#)
- [Entering Single User Mode on an x86 Switch \(see page 392\)](#)

### Entering Single User Mode on a PowerPC or ARM Switch

1. From the console, boot the switch, interrupting the U-Boot countdown to enter the U-Boot prompt. Enter the following:

```
=> setenv lbootargs init=/bin/sh  
=> boot
```

2. After the system boots, the shell command prompt appears. In this mode, you can change the root password or test a boot service that is hanging the boot process.
3. Reboot the system.

```
cumulus@switch:~$ sudo reboot -f  
Restarting the system.
```

### Entering Single User Mode on an x86 Switch

From the console, boot the switch. At the GRUB menu, select the image slot you wish to boot into with a password:

```
GNU GRUB version 1.99-27+deb7u2  
+-----+  
| Cumulus Linux 2.5.0-be24dc3-201412021541-build - slot 1 |  
| Cumulus Linux 2.5.0-be24dc3-201412021541-build - slot 1 (recovery mode) |  
| Cumulus Linux 2.5.0-b1bb3b7-201412090640-build - slot 2 |  
| Cumulus Linux 2.5.0-b1bb3b7-201412090640-build - slot 2 (recovery mode) |  
| ONIE |  
+-----|
```

In this example, you are selecting the slot2 image. Under the `linux` option, add `init=/bin/bash`:

```
GNU GRUB  version 1.99-27+deb7u2
+-----
| insmod part_gpt
|^
| insmod ext2
| set root='(hd0,gpt3)'
| search --no-floppy --fs-uuid --set=root c42be287-5321-4e77-975f-54e237a\
| d72b0
| echo 'Loading Linux ...'
| linux /cl-vmlinuz-3.2.60-1+deb7u1+cl2.5-slot-2 root=UUID=f01a2d40-d2fe-\|
| 435b-b3d1-7edc1eb0c42f console=ttyS0,115200n8 cl_platform=dell_s6000_s1\|
| 220 quiet active=2 init=/bin/bash
| echo 'Loading initial ramdisk ...' A
| initrd /cl-initrd.img-3.2.60-1+deb7u1+cl2.5-slot-2
|
|
+-----+
```

Type `Ctrl+x` or `F10` to boot with this change.

When you are done making changes as a single user, run `reboot -f` to boot the switch back to a normal state:

```
Begin: Running /scripts/init-bottom ... done.
bash: cannot set terminal process group (-1): Inappropriate ioctl for device
bash: no job control in this shell
cumulus@switch:/# sudo reboot -f
```

## Using netshow to Troubleshoot Your Network Configuration

`netshow` is a tool in Cumulus Linux that quickly returns a lot of information about your network configuration. It's a tool designed by network operators for network troubleshooters since existing command line tools have too many options. `netshow` addresses this by leveraging the network troubleshooting experience from a wide group of troubleshooters and boiling it down to just a few important options. `netshow` quickly aggregates basic network information on Linux devices with numerous interfaces. `netshow` intelligently informs the administrator what network type an interface belongs to, and shows the most relevant information to a network administrator.

`netshow` can be used on any distribution of Linux, not just Cumulus Linux.

## Installing netshow

Starting with Cumulus Linux 2.5.5, `netshow` is included in the main repository for Cumulus Linux. However, it is not installed by default if you upgraded to this version using `apt-get dist-upgrade`. You install `netshow` in Cumulus Linux in one of two ways:

- By doing a [binary image install \(see page 16\)](#) of Cumulus Linux 2.5.5 using `cl-img-install`
- Install the `netshow` package using `apt-get install netshow`

## Installing netshow on a Linux Server or in OpenStack

To install `netshow` on a Linux server, run:

```
pip install netshow-linux-lib
```



Debian and Red Hat packages will be available in the near future.

## Using netshow

Running `netshow` with no arguments displays all available command line arguments usable by `netshow`. (Running `netshow --help` gives you the same information.) The output looks like this:

```
cumulus@leaf1$ netshow
Usage:
    netshow system [--json | -j ]
    netshow counters [errors] [all] [--json | -j | -l | --legend ]
    netshow lldp [--json | -j | -l | --legend ]
    netshow interface [<iface>] [all] [--mac | -m ] [--oneline | -1 | --
json | -j | -l | --legend ]
    netshow access [all] [--mac | -m ] [--oneline | -1 | --json | -j | -l
| --legend ]
    netshow bridges [all] [--mac | -m ] [--oneline | -1 | --json | -j | -l
| --legend ]
    netshow bonds [all] [--mac | -m ] [--oneline | -1 | --json | -j | -l |
--legend ]
    netshow bondmems [all] [--mac | -m ] [--oneline | -1 | --json | -j | -
l | --legend ]
    netshow mgmt [all] [--mac | -m ] [--oneline | -1 | --json | -j | -l |
--legend ]
    netshow l2 [all] [--mac | -m ] [--oneline | -1 | --json | -j | -l | --
legend ]
```

```

netshow 13 [all] [--mac | -m ] [--oneline | -1 | --json | -j | -l | --
legend ]
    netshow trunks [all] [--mac | -m ] [--oneline | -1 | --json | -j | -l
| --legend ]
        netshow (--version | -V)

```

**Help:**

* default is to show interfaces only in the UP state.	
counters	summary of physical port counters.
interface	summary info of all interfaces
access	summary of physical ports with 12 or 13 config
bonds	summary of bonds
bondmems	summary of bond members
bridges	summary of ports with bridge members
mgmt	summary of mgmt ports
13	summary of ports with an IP.
12	summary of access, trunk and bridge interfaces
phy	summary of physical ports
trunks	summary of trunk interfaces
lldp	physical device neighbor information
interface <iface>	list summary of a single interface
system	system information

**Options:**

all	show all ports include those are down or admin down
--mac	show interface MAC in output
--version	netshow software version
--oneline	output each entry on one line
-1	alias for --oneline
--json	print output in json
-l	alias for --legend
--legend	print legend key explaining abbreviations

cumulus@leaf1\$

A Linux administrator can quickly see the few options available with the tool. One core tenet of `netshow` is for it to have a small number of command options. `netshow` is not designed to solve your network problem, but to help answer this simple question: "What is the basic network setup of my Linux device?" By helping to answer that question, a Linux administrator can spend more time troubleshooting the specific network problem instead of spending most of their time understanding the basic network state.

Originally developed for Cumulus Linux, `netshow` works on Debian-based servers and switches and Red Hat-based Linux systems.

`netshow` is designed by network operators, which has rarely occurred in the networking industry, where most command troubleshooting tools are designed by developers and are most useful in the network application development process.

## Showing Interfaces

To show all available interfaces that are physically UP, run `netshow interface`:

```
cumulus@leaf1$ netshow interface
-----
To view the legend, rerun "netshow" cmd with the "--legend" option
-----
      Name    Speed     MTU   Mode    Summary
--  -----  -----  -----  -----
UP  eth0     1G       1500  Mgmt   IP: 192.168.0.12/24(DHCP)
UP  lo       N/A      16436 Mgmt   IP: 127.0.0.1/8, ::1/128
cumulus@leaf1$
```

Whereas `netshow interface all` displays every interface regardless of state:

```
cumulus@leaf1$ netshow interface all
      Name    Speed     Mtu   Mode    Summary
--  -----  -----  -----  -----
UP  lo       N/A      16436 Loopback  IP: 127.0.0.1/8, ::1/128
UP  eth0     1G       1500  Mgmt   IP: 192.168.0.11/24 (DHCP)
ADMDN swp1s0  10G(4x10) 1500  Unknwn
ADMDN swp1s1  10G(4x10) 1500  Unknwn
ADMDN swp1s2  10G(4x10) 1500  Unknwn
ADMDN swp1s3  10G(4x10) 1500  Unknwn
ADMDN swp2    40G(QSFP) 1500  Unknwn
ADMDN swp3    40G(QSFP) 1500  Unknwn
ADMDN swp4    40G(QSFP) 1500  Unknwn
ADMDN swp5    40G(QSFP) 1500  Unknwn
ADMDN swp6    40G(QSFP) 1500  Unknwn
ADMDN swp7    40G(QSFP) 1500  Unknwn
ADMDN swp8    40G(QSFP) 1500  Unknwn
ADMDN swp9    40G(QSFP) 1500  Unknwn
ADMDN swp10   40G(QSFP) 1500  Unknwn
ADMDN swp11   40G(QSFP) 1500  Unknwn
ADMDN swp12   40G(QSFP) 1500  Unknwn
ADMDN swp13   40G(QSFP) 1500  Unknwn
ADMDN swp14   40G(QSFP) 1500  Unknwn
ADMDN swp15   40G(QSFP) 1500  Unknwn
ADMDN swp16   40G(QSFP) 1500  Unknwn
ADMDN swp17   40G(QSFP) 1500  Unknwn
ADMDN swp18   40G(QSFP) 1500  Unknwn
```

ADMDN	swp19	40G(QSFP)	1500	Unknwn
ADMDN	swp20	40G(QSFP)	1500	Unknwn
ADMDN	swp21	40G(QSFP)	1500	Unknwn
ADMDN	swp22	40G(QSFP)	1500	Unknwn
ADMDN	swp23	40G(QSFP)	1500	Unknwn
ADMDN	swp24	40G(QSFP)	1500	Unknwn
ADMDN	swp25	40G(QSFP)	1500	Unknwn
ADMDN	swp26	40G(QSFP)	1500	Unknwn
ADMDN	swp27	40G(QSFP)	1500	Unknwn
ADMDN	swp28	40G(QSFP)	1500	Unknwn
ADMDN	swp29	40G(QSFP)	1500	Unknwn
ADMDN	swp30	40G(QSFP)	1500	Unknwn
ADMDN	swp31	40G(QSFP)	1500	Unknwn
ADMDN	swp32s0	10G(4x10)	1500	Unknwn
ADMDN	swp32s1	10G(4x10)	1500	Unknwn
ADMDN	swp32s2	10G(4x10)	1500	Unknwn
ADMDN	swp32s3	10G(4x10)	1500	Unknwn

You can get information about the switch itself by running `netshow system`:

```
cumulus@leaf1$ netshow system

Quanta QuantaMesh BMS T1048-LB9
Cumulus Version 2.5.4
Build: 2.5.4-ecb2027-201510091646-build

Chipset: Broadcom Firebolt3 BCM56538

Port Config: 48x1G-T and 4x10G-SFP+

CPU: (ppc) Freescale MPC8541 e500 825MHz

UpTime: 20:01:17

cumulus@leaf1$
```

## Troubleshooting Example: OpenStack

Looking at an OpenStack Environment, here is the physical diagram:



For server2, `netshow` can help us see the OpenStack network configuration. The `netshow` output below shows a summary of a Kilo-based OpenStack server running 3 tenants.

```
[root@server2 ~]# netshow int
-----
To view the legend, rerun "netshow" cmd with the "--legend" option
-----
      Name          Speed     MTU     Mode           Summary
--  -----
UP   brq0b6f10c7-42  N/A      1500   Bridge/L2      802.1q Tag: 141
                                         STP: Disabled
                                         Untagged Members:
                                         tap079cf993-c7
                                         Tagged Members: eth1.141
                                         802.1q Tag: 155
                                         STP: Disabled
                                         Untagged Members:
                                         tap5353b20a-68
                                         Tagged Members: eth1.155
                                         802.1q Tag: 168
                                         STP: Disabled
                                         Untagged Members:
                                         tapfc2203e4-5b
                                         Tagged Members: eth1.168
                                         IP: 192.168.0.105/24
                                         UP   eth0          N/A      1500   Interface/L3
                                         UP   eth1          N/A      1500   IntTypeUnknown
                                         UP   eth1          N/A      1500   Trunk/L2      Bridge Membership:
                                         Tagged: brq0b6f10c7-42
                                         (141), brq8cdc0589-9b(155), brq8ff99102-29(168)
                                         UP   lo            N/A      65536  Loopback      IP: 127.0.0.1/8, ::1/128
                                         UP   tap079cf993-c7 10M      1500   Access/L2    Untagged: brq0b6f10c7-42
                                         UP   tap5353b20a-68 10M      1500   Access/L2    Untagged: brq8cdc0589-9b
                                         UP   tapfc2203e4-5b 10M      1500   Access/L2    Untagged: brq8ff99102-29
```

OpenStack interface numbering is not the easiest read, but here `netshow` can quickly show you:

- A list of all the interfaces in admin UP state and carrier UP state
- 3 bridges
- That STP is disabled for all the bridges
- An uplink trunk interface with 3 VLANs configured on it
- Many tap interfaces, most likely the virtual machines

This output took about 5 seconds to get and another 1 minute to analyze. To get this same level of understanding using traditional tools such as:

- `ip link show`

- brctl show
- ip addr show

... could take about 10 minutes. This is a significant improvement in productivity!

`netshow` uses a plugin architecture and can be easily expanded. An OpenStack interface discovery module is currently in development. If `netshow` is run on a hypervisor with OpenStack Keystone login environment variables like `OS_TENANT_NAME`, `netshow` should show the above output with a better interface discovery state, where `netshow` collects from OpenStack information from `libvirt`, `nova` and `neutron` to overlay the virtual machine and tenant subnet information over the interface kernel state information.

Interface discovery is one of the most powerful features of `netshow`. The ability to expand its interface discovery capabilities further simplifies understanding basic network troubleshooting, making the Linux administrator more productive and improving time to resolution while investigating network problems.

## Other Useful `netshow` Features

`netshow` uses the `python network-docopt` package. This is inspired by `docopt` and provides the ability to specify partial commands, without tab completion and running the complete option. For example:

```
netshow int runs netshow interface  
netshow sys runs netshow system
```

`netshow` will eventually support interface name autocompletion. In the near future, if you run `netshow int tap123` and there is only one interface starting with `tap123`, `netshow` will autocomplete the command option with the full interface.

## Contributions Welcome!

`netshow` is an open source project licensed under GPLv2. To contribute please contact Cumulus Networks through the [Cumulus Community Forum](#) or the [Netshow Linux Provider Github Repository Home](#). You can find developer documentation at [netshow.readthedocs.org](#). The documentation is still under development.

## Monitoring Interfaces and Transceivers Using `ethtool`

The `ethtool` command enables you to query or control the network driver and hardware settings. It takes the device name (like `swp1`) as an argument. When the device name is the only argument to `ethtool`, it prints the current settings of the network device. See `man ethtool(8)` for details. Not all options are currently supported on switch port interfaces.

## Contents

(Click to expand)

- [Contents \(see page 399\)](#)
- [Commands \(see page 400\)](#)
- [Monitoring Interfaces Using ethtool \(see page 400\)
  - \[Viewing and Clearing Interface Counters \\(see page 401\\)\]\(#\)](#)
- [Monitoring Switch Port SFP/QSFP Using ethtool \(see page 402\)](#)

## Commands

- cl-netstat
- ethtool

## Monitoring Interfaces Using ethtool

To check the status of an interface using ethtool:

```
cumulus@switch:~$ ethtool swp1
Settings for swp1:
    Supported ports: [ FIBRE ]
    Supported link modes:  1000baseT/Full
                           10000baseT/Full
    Supported pause frame use: No
    Supports auto-negotiation: No
    Advertised link modes:  1000baseT/Full
    Advertised pause frame use: No
    Advertised auto-negotiation: No
    Speed: 10000Mb/s
    Duplex: Full
    Port: FIBRE
    PHYAD: 0
    Transceiver: external
    Auto-negotiation: off
    Current message level: 0x00000000 (0)

    Link detected: yes
```

To query interface statistics:

```
cumulus@switch:~$ sudo ethtool -S swp1
NIC statistics:
    HwIfInOctets: 1435339
    HwIfInUcastPkts: 11795
    HwIfInBcastPkts: 3
    HwIfInMcastPkts: 4578
    HwIfOutOctets: 14866246
    HwIfOutUcastPkts: 11791
    HwIfOutMcastPkts: 136493
    HwIfOutBcastPkts: 0
    HwIfInDiscards: 0
```

```

HwIfInL3Drops: 0
HwIfInBufferDrops: 0
HwIfInAclDrops: 28
HwIfInDot3LengthErrors: 0
HwIfInErrors: 0
SoftInErrors: 0
SoftInDrops: 0
SoftInFrameErrors: 0
HwIfOutDiscards: 0
HwIfOutErrors: 0
HwIfOutQDrops: 0
HwIfOutNonQDrops: 0
SoftOutErrors: 0
SoftOutDrops: 0
SoftOutTxFifoFull: 0
HwIfOutQLen: 0

```

## ***Viewing and Clearing Interface Counters***

Interface counters contain information about an interface. You can view this information when you run `cl-netstat`, `ifconfig`, or `cat /proc/net/dev`. You can also use `cl-netstat` to save or clear this information:

```

cumulus@switch:~# sudo cl-netstat
Kernel Interface table
Iface      MTU Met          RX_OK RX_ERR RX_DRP RX_OVR          TX_OK TX_ERR
TX_DRP TX_OVR Flg
-----
eth0      1500 0            611   0     0     0             487   0
0          0   BMRU
lo       16436 0            0     0     0     0             0     0
0          0   LRU
swp1      1500 0            0     0     0     0             0     0
0          0   BMU

cumulus@switch:~# sudo :~# cl-netstat -c
Cleared counters

```

Option	Description
-c	Copies and clears statistics. It does not clear counters in the kernel or hardware.

Option	Description
-d	Deletes saved statistics, either the uid or the specified tag.
-D	Deletes all saved statistics.
-l	Lists saved tags.
-r	Displays raw statistics (unmodified output of cl-netstat).
-t <tag name>	Saves statistics with <tag name>.
-v	Prints cl-netstat version and exits.

## Monitoring Switch Port SFP/QSFP Using ethtool

The `ethtool -m` command provides switch port SFP information. It shows connector information, vendor data, and more:

```
cumulus@switch:~$ sudo ethtool -m swp1
swp1: SFP detected
      Connector : CopperPigtail
      EncodingCodes : Unspecified
      ExtIdentOfTypeOfTransceiver : GBIC/SFP defined by twowire interface ID
      LengthCable(UnitsOfM) : 1
      NominalSignallingRate(UnitsOf100Mbd) : 103
      RateIdentifier : Unspecified
      ReceivedPowerMeasurementType : OMA
      TransceiverCodes :
          SFP+CableTechnology : Passive Cable
      TypeOfTransceiver : SFP or SFP Plus
          VendorDataCode(yyymmdd) : 110830
      VendorName : Amphenol
          VendorOUI : Amp
      VendorPN : 571540001
          VendorRev : M
      VendorSN : APF11350017C4V
```

## Resource Diagnostics Using cl-resource-query

You can use `cl-resource-query` to retrieve information about host entries, MAC entries, L2 and L3 routes, and ECMPs (equal-cost multi-path routes, see [Load Balancing \(see page 317\)](#)) that are in use. This is especially useful because Cumulus Linux syncs routes between the kernel and the switching silicon. If the required resource pools in hardware fill up, new kernel routes can cause existing routes to move from being fully allocated to being partially allocated.

In order to avoid this, routes in the hardware should be monitored and kept below the ASIC limits. For example, on systems with a Trident II chipset, the limits are as follows:

```
routes: 8092 <<< if all routes are IPv6, or 16384 if all routes are IPv4
long mask routes 2048 <<< these are routes with a mask longer than the
route mask limit
route mask limit 64
host_routes: 8192
ecmp_nhs: 16346
ecmp_nhs_per_route: 52
```

This translates to about 314 routes with ECMP next hops, if every route has the maximum ECMP NHs.

For systems with a Trident+ chipset, the limits are as follows:

```
routes: 16384 <<< if all routes are IPv4
long mask routes 256 <<< these are routes with a mask longer than the
route mask limit
route mask limit 64
host_routes: 8192
ecmp_nhs: 4044
ecmp_nhs_per_route: 52
```

This translates to about 77 routes with ECMP next hops, if every route has the maximum ECMP NHs.

You can monitor this in Cumulus Linux with the `cl-resource-query` command. Results vary between switches running on Trident+ and Trident II chipsets.

`cl-resource-query` results for a Trident II switch:

```
cumulus@switch:~$ sudo cl-resource-query
Host entries:           1,   0% of maximum value    8192 <<< this is
the default software-imposed limit, 50% of the hardware limit
IPv4 neighbors:         1           <<< these are counts of the number
of valid entries in the table
IPv6 neighbors:         0
```

IPv4 entries:	13,	0% of maximum value	32668
IPv6 entries:	18,	0% of maximum value	16384
IPv4 Routes:	13		
IPv6 Routes:	18		
Total Routes:	31,	0% of maximum value	32768
ECMP nexthops:	0,	0% of maximum value	16346
MAC entries:	12,	0% of maximum value	32768

cl-resource-query results for a Trident+ switch:

cumulus@switch:~\$ sudo cl-resource-query			
Host entries:	6,	0% of maximum value	4096 <<< same as
above			
IPv4 neighbors:	6		
IPv6 neighbors:	0		
IPv4/IPv6 entries:	33,	0% of maximum value	16284
Long IPv6 entries:	0,	0% of maximum value	256
IPv4 Routes:	29		
IPv6 Routes:	2		
Total Routes:	31,	0% of maximum value	32768
ECMP nexthops:	0,	0% of maximum value	4041
MAC entries:	0,	0% of maximum value	131072

## Monitoring System Hardware

You monitor system hardware in these ways, using:

- decode-syseeprom
- sensors
- smond
- Net-SNMP (see page 454)
- watchdog

### Contents

(Click to expand)

- Contents (see page 404)
- Commands (see page 405)
- Monitoring Hardware Using decode-syseeprom (see page 405)
  - Command Options (see page 405)
  - Related Commands (see page 406)
- Monitoring Hardware Using sensors (see page 406)

- Command Options (see page 407)
- Monitoring Switch Hardware Using SNMP (see page 408)
- Monitoring System Units Using smond (see page 408)
  - Command Options (see page 408)
- Keeping the Switch Alive Using the Hardware Watchdog (see page 409)
- Configuration Files (see page 409)
- Useful Links (see page 409)

## Commands

- decode-syseeprom
- dmidecode
- lshw
- sensors
- smond

## Monitoring Hardware Using *decode-syseeprom*

The `decode-syseeprom` command enables you to retrieve information about the switch's EEPROM. If the EEPROM is writable, you can set values on the EEPROM.

For example:

```
cumulus@switch:~$ decode-syseeprom
TlvInfo Header:
  Id String:      TlvInfo
  Version:       1
  Total Length: 114

  TLV Name          Code Len Value
  -----  -----
Product Name        0x21   4  4804
Part Number         0x22   14 R0596-F0009-00
Device Version     0x26    1  2
Serial Number       0x23   19 D1012023918PE000012
Manufacture Date   0x25   19 10/09/2013 20:39:02
Base MAC Address   0x24    6  00:E0:EC:25:7B:D0
MAC Addresses       0x2A    2  53
Vendor Name         0x2D   17 Penguin Computing
Label Revision     0x27    4  4804
Manufacture Country 0x2C    2  CN
CRC-32              0xFE    4  0x96543BC5
  (checksum valid)
```

## Command Options

Usage: /usr/cumulus/bin/decode-syseeprom [-a][-r][-s [args]][-t]

Option	Description
-h, --help	Displays the help message and exits.
-a	Prints the base MAC address for switch interfaces.
-r	Prints the number of MACs allocated for switch interfaces.
-s	Sets the EEPROM content if the EEPROM is writable. args can be supplied in command line in a comma separated list of the form '<field>=<value>, ...'. ', '=' and '!' are illegal characters in field names and values. Fields that are not specified will default to their current values. If args are supplied in the command line, they will be written without confirmation. If args is empty, the values will be prompted interactively.
-t TARGET	Selects the target EEPROM (board, psu2, psu1) for the read or write operation; default is board.
-e, --serial	Prints the device serial number.

## Related Commands

You can also use the `dmidecode` command to retrieve hardware configuration information that's been populated in the BIOS.

You can use `apt-get` to install the `lshw` program on the switch, which also retrieves hardware configuration information.

## Monitoring Hardware Using sensors

The `sensors` command provides a method for monitoring the health of your switch hardware, such as power, temperature and fan speeds. This command executes `lm-sensors`.

For example:

```
cumulus@switch:~$ sensors
tmp75-i2c-6-48
Adapter: i2c-1-mux (chan_id 0)
temp1:      +39.0°C  (high = +75.0°C, hyst = +25.0°C)

tmp75-i2c-6-49
Adapter: i2c-1-mux (chan_id 0)
temp1:      +35.5°C  (high = +75.0°C, hyst = +25.0°C)
```

```

ltc4215-i2c-7-40
Adapter: i2c-1-mux (chan_id 1)
in1:          +11.87 V
in2:          +11.98 V
power1:       12.98 W
curr1:        +1.09 A

max6651-i2c-8-48
Adapter: i2c-1-mux (chan_id 2)
fan1:         13320 RPM (div = 1)
fan2:         13560 RPM

```



Output from the `sensors` command varies depending upon the switch hardware you use, as each platform ships with a different type and number of sensors.

## **Command Options**

Usage: `sensors [OPTION]... [CHIP]...`

Option	Description
<code>-c, --config-file</code>	Specify a config file; use <code>-</code> after <code>-c</code> to read the config file from <code>stdin</code> ; by default, <code>sensors</code> references the configuration file in <code>/etc/sensors.d/</code> .
<code>-s, --set</code>	Executes set statements in the config file (root only); <code>sensors -s</code> is run once at boot time and applies all the settings to the boot drivers.
<code>-f, --fahrenheits</code>	Show temperatures in degrees Fahrenheit.
<code>-A, --no-adapter</code>	Do not show the adapter for each chip.
<code>--bus-list</code>	Generate bus statements for <code>sensors.conf</code> .

If `[CHIP]` is not specified in the command, all chip info will be printed. Example chip names include:

- `lm78-i2c-0-2d *-i2c-0-2d`
- `lm78-i2c-0-* *-i2c-0-*`
- `lm78-i2c-*-2d *-i2c-*-2d`
- `lm78-i2c-*-* *-i2c-*-*`
- `lm78-isa-0290 *-isa-0290`

- lm78-isa-\* \*-isa-\*
- lm78-\*

## Monitoring Switch Hardware Using SNMP

The Net-SNMP documentation has been moved to a new chapter, [available here \(see page 454\)](#).

## Monitoring System Units Using smond

The `smond` daemon monitors system units like power supply and fan, updates their corresponding LEDs, and logs the change in the state. Changes in system unit state are detected via the `cpld` registers. `smond` utilizes these registers to read all sources, which impacts the health of the system unit, determines the unit's health, and updates the system LEDs.

Use `smonctl` to display sensor information for the various system units:

```
cumulus@switch:~$ smonctl
Board : OK
Fan   : OK
PSU1  : OK
PSU2  : BAD
Temp1  (Networking ASIC Die Temp Sensor) : OK
Temp10 (Right side of the board)        : OK
Temp2  (Near the CPU (Right))          : OK
Temp3  (Top right corner)             : OK
Temp4  (Right side of Networking ASIC) : OK
Temp5  (Middle of the board)           : OK
Temp6  (P2020 CPU die sensor)         : OK
Temp7  (Left side of the board)        : OK
Temp8  (Left side of the board)        : OK
Temp9  (Right side of the board)       : OK
```

## Command Options

Usage: `smonctl [OPTION]... [CHIP]...`

Option	Description
<code>-s SENSOR, --sensor SENSOR</code>	Displays data for the specified sensor.
<code>-v, --verbose</code>	Displays detailed hardware sensors data.

For more information, read `man smond` and `man smonctl`.

## Keeping the Switch Alive Using the Hardware Watchdog

Cumulus Linux includes a simplified version of the `wd_keepalive(8)` daemon from the standard Debian package `watchdog`. `wd_keepalive` writes to a file called `/dev/watchdog` periodically to keep the switch from resetting, at least once per minute. Each write delays the reboot time by another minute. After one minute of inactivity where `wd_keepalive` doesn't write to `/dev/watchdog`, the switch resets itself.

The watchdog is enabled by default on QuantaMesh BMS T1048-LB9 switches only; you must enable the watchdog on all other switch platforms. When enabled, it starts when you boot the switch, before `switchd` starts.

To enable the hardware watchdog, edit the `/etc/watchdog.d/<your_platform>` file and set `run_watchdog` to 1:

```
run_watchdog=1
```

To disable the watchdog, edit the `/etc/watchdog.d/<your_platform>` file and set `run_watchdog` to 0 :

```
run_watchdog=0
```

Then stop the daemon:

```
cumulus@switch:~$ sudo service wd_keepalive stop
```

You can modify the settings for the watchdog — like the timeout setting and scheduler priority — in its configuration file, `/etc/watchdog.conf`.

## Configuration Files

- `/etc/cumulus/switchd.conf`
- `/etc/cumulus/sysledcontrol.conf`
- `/etc/sensors.d/<switch>.conf` - sensor configuration file (do **not** edit it!)
- `/etc/watchdog.conf`

## Useful Links

- <http://packages.debian.org/search?keywords=lshw>
- <http://lm-sensors.org>
- Net-SNMP tutorials

# Monitoring System Statistics and Network Traffic with sFlow

sFlow is a monitoring protocol that samples network packets, application operations, and system counters. sFlow enables you to monitor your network traffic as well as your switch state and performance metrics. An outside server, known as an *sFlow collector*, is required to collect and analyze this data.

`hsflowd` is the daemon that samples and sends sFlow data to configured collectors. `hsflowd` is not included in the base Cumulus Linux installation. After installation, `hsflowd` will automatically start when the switch boots up.

## Contents

(Click to expand)

- [Contents \(see page 410\)](#)
- [Installing hsflowd \(see page 410\)](#)
- [Configuring sFlow \(see page 410\)
  - \[Configuring sFlow via DNS-SD \\(see page 410\\)\]\(#\)
  - \[Manually Configuring /etc/hsflowd.conf \\(see page 411\\)\]\(#\)](#)
- [Configuring sFlow Visualization Tools \(see page 412\)](#)
- [Configuration Files \(see page 412\)](#)
- [Useful Links \(see page 412\)](#)

## Installing hsflowd

To download and install the `hsflowd` package, use `apt-get`:

```
cumulus@switch:~$ sudo apt-get update
cumulus@switch:~$ sudo apt-get install -y hsflowd
```

## Configuring sFlow

You can configure `hsflowd` to send to the designated collectors via two methods:

- DNS service discovery (DNS-SD)
- Manually configuring `/etc/hsflowd.conf`

## Configuring sFlow via DNS-SD

With this method, you need to configure your DNS zone to advertise the collectors and polling information to all interested clients. Add the following content to the zone file on your DNS server:

```
_sflow._udp SRV 0 0 6343 collector1
_sflow._udp SRV 0 0 6344 collector2
```

```
_sflow._udp TXT (
  "txtvers=1"
  "sampling.1G=2048"
  "sampling.10G=4096"
  "sampling.40G=8192"
  "polling=20"
)
```

The above snippet instructs `hsflowd` to send sFlow data to collector1 on port 6343 and to collector2 on port 6344. `hsflowd` will poll counters every 20 seconds and sample 1 out of every 2048 packets.

After the initial configuration is ready, bring up the sFlow daemon by running:

```
cumulus@switch:~$ sudo service hsflowd start
```

No additional configuration is required in `/etc/hsflowd.conf`.

## **Manually Configuring /etc/hsflowd.conf**

With this method you will set up the collectors and variables on each switch.

Edit `/etc/hsflowd.conf` and change `DNSSD = on` to `DNSSD = off`:

```
DNSSD = off
```

Then set up your collectors and sampling rates in `/etc/hsflowd.conf`:

```
# Manual Configuration (requires DNSSD=off above)
#####
# Typical configuration is to send every 30 seconds
polling = 20

sampling.1G=2048
sampling.10G=4096
sampling.40G=8192

collector {
  ip = 192.0.2.100
  udpport = 6343
}

collector {
```

```
ip = 192.0.2.200
udpport = 6344
}
```

This configuration polls the counters every 20 seconds, samples 1 of every 2048 packets and sends this information to a collector at 192.0.2.100 on port 6343 and to another collector at 192.0.2.200 on port 6344.



Some collectors require each source to transmit on a different port, others may listen on only one port. Please refer to the documentation for your collector for more information.

## ***Configuring sFlow Visualization Tools***

For information on configuring various sFlow visualization tools, read this [Help Center article](#).

## ***Configuration Files***

- /etc/hsflowd.conf

## ***Useful Links***

- [sFlow Collectors](#)
- [sFlow Wikipedia page](#)

## **Monitoring Virtual Device Counters**

Cumulus Linux gathers statistics for VXLANS and VLANs using virtual device counters. These counters are supported on Trident II-based platforms only; see the [Cumulus Networks HCL](#) for a list of supported Trident II platforms.

You can retrieve the data from these counters using tools like `ip -s link show`, `ifconfig`, `/proc/net/dev`, or `netstat -i`.

## ***Contents***

(Click to expand)

- [Contents \(see page 412\)](#)
- [Sample VXLAN Statistics \(see page 413\)](#)
- [Sample VLAN Statistics \(see page 414\)
  - \[For VLANs Using the non-VLAN-aware Bridge Driver \\(see page 414\\)\]\(#\)
  - \[For VLANs Using the VLAN-aware Bridge Driver \\(see page 414\\)\]\(#\)](#)
- [Configuring the Counters in switchd \(see page 415\)
  - \[Configuring the Poll Interval \\(see page 415\\)\]\(#\)
  - \[Configuring Internal VLAN Statistics \\(see page 416\\)\]\(#\)](#)

- [Clearing Statistics \(see page 416\)](#)
- [Caveats and Errata \(see page 416\)](#)

## Sample VXLAN Statistics

VXLAN statistics are available as follows:

- Aggregate statistics are available per VNI; this includes access and network statistics.
- Network statistics are available for each VNI and displayed against the VXLAN device. This is independent of the VTEP used, so this is a summary of the VNI statistics across all tunnels.
- Access statistics are available per VLAN subinterface.

First, get interface information regarding the VXLAN bridge:

```
cumulus@switch:~$ brctl show br-vxln16757104
bridge name          bridge id      STP enabled    interfaces
-br-vxln16757104    8000.443839006988    no           swp2s0.6
                                         swp2s1.6
                                         swp2s2.6
                                         swp2s3.6
                                         vxln16757104
```

To get VNI statistics, run:

```
cumulus@switch:~$ ip -s link show br-vxln16757104
62: br-vxln16757104: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc
noqueue state UP mode DEFAULT
    link/ether 44:38:39:00:69:88 brd ff:ff:ff:ff:ff:ff
    RX: bytes   packets   errors   dropped overrun mcast
        10848       158       0        0        0        0
    TX: bytes   packets   errors   dropped carrier collsns
        27816       541       0        0        0        0
```

To get access statistics, run:

```
cumulus@switch:~$ ip -s link show swp2s0.6
63: swp2s0.6@swp2s0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc
noqueue master br-vxln16757104 state UP mode DEFAULT
    link/ether 44:38:39:00:69:88 brd ff:ff:ff:ff:ff:ff
    RX: bytes   packets   errors   dropped overrun mcast
        2680        39       0        0        0        0
    TX: bytes   packets   errors   dropped carrier collsns
        7558       140       0        0        0        0
```

To get network statistics, run:

```
cumulus@switch:~$ ip -s link show vxln16757104
61: vxln16757104: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue
  master br-vxln16757104 state UNKNOWN mode DEFAULT
    link/ether e2:37:47:db:f1:94 brd ff:ff:ff:ff:ff:ff
      RX: bytes  packets  errors  dropped overrun mcast
        0         0       0       0       0       0
      TX: bytes  packets  errors  dropped carrier collsns
        0         0       0       9       0       0
```

## **Sample VLAN Statistics**

### **For VLANs Using the non-VLAN-aware Bridge Driver**

In this case, each bridge is a single L2 broadcast domain and is associated with an internal VLAN. This internal VLAN's counters are displayed as bridge netdev stats.

```
cumulus@switch:~$ brctl show br0
bridge name     bridge id          STP enabled     interfaces
br0            8000.443839006989    yes           bond0.100
                                         swp2s2.100

cumulus@switch:~$ ip -s link show br0
42: br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue state UP
  mode DEFAULT
    link/ether 44:38:39:00:69:89 brd ff:ff:ff:ff:ff:ff
      RX: bytes  packets  errors  dropped overrun mcast
        23201498   227514   0       0       0       0
      TX: bytes  packets  errors  dropped carrier collsns
        18198262   178443   0       0       0       0
```

### **For VLANs Using the VLAN-aware Bridge Driver**

For a bridge using the [VLAN-aware driver](#) (see page 182), the bridge is a just a container and each VLAN (VID /PVID) in the bridge is an independent L2 broadcast domain. As there is no netdev available to display these VLAN statistics, the switchd nodes are used instead:

```
cumulus@switch:~$ ifquery bridge
auto bridge
iface bridge inet static
  bridge-vlan-aware yes
```

```

bridge-ports swp2s0 swp2s1
bridge-stp on
bridge-vids 2000-2002 4094
cumulus@switch:~$ ls /cumulus/switchd/run/stats/vlan/
2 2000 2001 2002 all
cumulus@switch:~$ cat /cumulus/switchd/run/stats/vlan/2000/aggregate
Vlan id                      : 2000
L3 Routed In Octets           : -
L3 Routed In Packets          : -
L3 Routed Out Octets          : -
L3 Routed Out Packets         : -
Total In Octets               : 375
Total In Packets              : 3
Total Out Octets              : 387
Total Out Packets             : 3

```

## Configuring the Counters in switchd

These counters are enabled by default. To configure them, use `c1-cfg` and configure them as you would any other `switchd` parameter (see page 87). The `switchd` parameters are as follows:

- `stats.vlan.aggregate`, which controls the statistics available for each VLAN. Its value defaults to *BRIEF*.
- `stats.vxlan.aggregate`, which controls the statistics available for each VNI (access and network). Its value defaults to *DETAIL*.
- `stats.vxlan.member`, which controls the statistics available for each local/access port in a VXLAN bridge. Its value defaults to *BRIEF*.

The values for each parameter can be one of the following:

- NONE: This disables the counter.
- BRIEF: This provides tx/rx packet/byte counters for the associated parameter.
- DETAIL: This provides additional feature-specific counters. In the case of `stats.vxlan.aggregate`, DETAIL provides access vs. network statistics. For the other types, DETAIL has the same effect as BRIEF.



If you change one of these settings on the fly, the new configuration applies only to those VNIs or VLANs set up after the configuration changed; previously allocated counters remain as is.

## Configuring the Poll Interval

The virtual device counters are polled periodically. This can be CPU intensive, so the interval is configurable in `switchd`, with a default of 2 seconds.

```
# Virtual devices hw-stat poll interval (in seconds)
#stats.vdev_hw_poll_interval = 2
```

## Configuring Internal VLAN Statistics

For debugging purposes, you may need to access packet statistics associated with internal VLAN IDs. These statistics are hidden by default, but can be configured in `switchd`:

```
#stats.vlan.show_internal_vlans = FALSE
```

## Clearing Statistics

Since `ethtool` is not supported for virtual devices, you cannot clear the statistics cache maintained by the kernel. You can clear the hardware statistics via `switchd`:

```
cumulus@switch:~$ sudo echo 1 > /cumulus/switchd/clear/stats/vlan
cumulus@switch:~$ sudo echo 1 > /cumulus/switchd/clear/stats/vxlan
cumulus@switch:~$
```

## Caveats and Errata

- Currently the CPU port is internally added as a member of all VLANs. Because of this, packets sent to the CPU are counted against the corresponding VLAN's tx packets/bytes. There is no workaround.
- When checking the virtual counters for the bridge, the TX count is the number of packets destined to the CPU before any hardware policers take effect. For example, if 500 broadcast packets are sent into the bridge, the CPU is also sent 500 packets. These 500 packets are policed by the default ACLs in Cumulus Linux, so the CPU might receive fewer than the 500 packets if the incoming packet rate is too high. The TX counter for the bridge should be equal to  $500 * (\text{number of ports in the bridge} - \text{incoming port} + \text{CPU port})$  or just  $500 * \text{number of ports in the bridge}$ .
- You cannot use `ethtool -s` for virtual devices. This is because the counters available via `netdev` are sufficient to display the vlan/vxlan counters currently supported in the hardware (only rx/tx packets/bytes are supported currently).

## Understanding and Decoding the cl-support Output File

***The cl-support command generates a tar archive of useful information for troubleshooting that can be auto-generated or manually created. To manually create it, run the cl-support command. The cl-support file is automatically generated when:***

- There is a [core file dump](#) of any application (not specific to Cumulus Linux, but something all Linux distributions support)
- Memory usage surpasses 90% of the total system memory (memory usage > 90% for 1 cycle)
- The [loadavg](#) over 15 minutes has on average greater than 2 (loadavg (15min) > 2)

All of these conditions are triggered by [jdoe](#), located at `/etc/jdoe/jdoorc`.

The Cumulus Networks support team may request you submit the output from `cl-support` to help with the investigation of issues you might experience with Cumulus Linux.

```
cumulus@switch:~$ sudo cl-support -h
Usage: cl-support [-h] [reason]...
Args:
[reason]: Optional reason to give for invoking cl-support.
          Saved into tarball's reason.txt file.
Options:
-h: Print this usage statement
```

Example output:

```
cumulus@switch:~$ ls /var/support
cl_support__switch_20141204_203833
```

(Click to expand)

- The `cl-support` command generates a tar archive of useful information for troubleshooting that can be auto-generated or manually created. To manually create it, run the `cl-support` command. The `cl-support` file is automatically generated when: (see page 417)
- Understanding the File Naming Scheme (see page 417)
- Decoding the Output (see page 418)

## Understanding the File Naming Scheme

The `cl-support` command generates a file under `/var/support` with the following naming scheme. The following example describes the file called `cl_support__switch_20141204_203833.tar.xz`.

<b>cl_support</b>	<b>switch</b>	<b>20141204</b>	<b>203833</b>
This is always prepended to the <code>tar.gz</code> output.	This is the hostname of the switch where <code>cl-support</code> was executed.	The date in year, month, day; so 20141204 is December, 4th, 2014.	The time in hours, minutes, seconds; so 203833 is 20, 38, 33 (20:38:33) or the equivalent to 8:38:33 PM.

## Decoding the Output

Decoding a `cl_support` file is a simple process performed using the `tar` command. The following example illustrates extracting the `cl_support` file:

```
tar -xf cl_support__switch_20141204_203834.tar.xz
```

The `-xf` options are defined here:

<b>Option</b>	<b>Description</b>
<code>-x</code>	Extracts to disk from the archive.
<code>-f</code>	Reads the archive from the specified file.

```
cumulus@switch:~$ ls -l cl_support__switch_20141204_203834/
-rwxr-xr-x 1 root root 7724 Jul 29 14:00 cl-support
-rw-r--r-- 1 root root    52 Jul 29 14:00 cmdline.args
drwxr-xr-x 2 root root 4096 Jul 29 14:00 core
drwxr-xr-x 64 root root 4096 Jul 29 13:51 etc
drwxr-xr-x  4 root root 4096 Jul 29 14:00 proc
drwxr-xr-x  2 root root 4096 Jul 29 14:01 support
drwxr-xr-x  3 root root 4096 Jul 29 14:00 sys
drwxr-xr-x  3 root root 4096 Aug  8 15:22 var
```

The `cl_support` file, when untarred, contains a `reason.txt` file. This file indicates what reason triggered the event. When contacting Cumulus Networks technical support, please attach the `cl-support` file if possible.

The directory contains the following elements:

<b>Directory</b>	<b>Description</b>
<code>cl-support</code>	

Directory	Description
	This is a copy of the <code>c1-support</code> script that generated the <code>c1_support</code> file. It is copied so Cumulus Networks knows exactly which files were included and which weren't. This helps to fix future <code>c1-support</code> requests in the future.
core	Contains the core files generated from the Cumulus Linux HAL (hardware abstraction layer) process, <code>switchd</code> .
etc	<code>etc</code> is the core system configuration directory. <code>c1-support</code> replicates the switch's <code>/etc</code> directory. <code>/etc</code> contains all the general Linux configuration files, as well as configurations for the system's network interfaces, <code>quagga</code> , <code>jdoe</code> , and other packages.
var/log	<code>/var</code> is the "variable" subdirectory, where programs record runtime information. System logging, user tracking, caches and other files that system programs create and monitor go into <code>/var</code> . <code>c1-support</code> includes only the <code>log</code> subdirectory of the <code>var</code> system-level directory and replicates the switch's <code>/var/log</code> directory. Most Cumulus Linux log files are located in this directory. Notable log files include <code>switchd.log</code> , <code>daemon.log</code> , <code>quagga</code> log files, and <code>syslog</code> . For more information, read this <a href="#">knowledge base article</a> .
proc	<code>proc</code> (short for processes) provides system statistics through a directory-and-file interface. In Linux, <code>/proc</code> contains runtime system information (like system memory, devices mounted, and hardware configuration). <code>c1-support</code> simply replicates the switch's <code>/proc</code> directory to determine the current state of the system.
support	<code>support</code> is <b>not</b> a replica of the Linux file system like the other folders listed above. Instead, it is a set of files containing the output of commands from the command line. Examples include the output of <code>ps -aux</code> , <code>netstat -i</code> , and so forth — even the routing tables are included.

Here is more information on the file structure:

- [Troubleshooting the etc Directory \(see page 422\)](#) — In terms of sheer numbers of files, `/etc` contains the largest number of files to send to Cumulus Networks by far. However, log files could be significantly larger in file size.
- [Troubleshooting Log Files \(see page 419\)](#) — This guide highlights the most important log files to look at. Keep in mind, `c1-support` includes all of the log files.
- [Troubleshooting the support Directory \(see page 433\)](#) — This is an explanation of the `support` directory included in the `c1-support` output.

## Troubleshooting Log Files

The only real unique entity for logging on Cumulus Linux compared to any other Linux distribution is `switchd.log`, which logs the HAL (hardware abstraction layer) from hardware like the Broadcom ASIC.

[This guide on NixCraft](#) is amazing for understanding how `/var/log` works. The green highlighted rows below are the most important logs and usually looked at first when debugging.

Log	Description	Why is this important?
/var/log/alternatives.log	Information from the update-alternatives are logged into this log file.	
/var/log/apt	Information the <code>apt</code> utility can send logs here; for example, from <code>apt-get install</code> and <code>apt-get remove</code> .	
/var/log/audit/	Contains log information stored by the Linux audit daemon, <code>audited</code> .	
/var/log/auth.log	Authentication logs.	
/var/log/boot.log	Contains information that is logged when the system boots.	
/var/log/btmp	<p>This file contains information about failed login attempts. Use the <code>last</code> command to view the <code>btmp</code> file. For example:</p> <pre data-bbox="383 1100 833 1132"><code>last -f /var/log/btmp   more</code></pre>	
/var/log/daemon.log	Contains information logged by the various background daemons that run on the system.	
/var/log/dmesg	Contains kernel ring buffer information. When the system boots up, it prints number of messages on the screen that display information about the hardware devices that the kernel detects during boot process. These messages are available in the kernel ring buffer and whenever a new message arrives, the old message gets overwritten. You can also view the content of this file using the <code>dmesg</code> command.	dmesg is one of the few places to determine hardware errors.
/var/log/dpkg.log	Contains information that is logged when a package is installed or removed using the <code>dpkg</code> command.	
/var/log/faillog	Contains failed user login attempts. Use the <code>faillog</code> command to display the contents of this file.	
/var/log/fsck/*	The <code>fsck</code> utility is used to check and optionally repair one or more Linux filesystems.	

Log	Description	Why is this important?
/var/log/jdoo.log	jdoo is a utility for managing and monitoring processes, files, directories and filesystems on a Unix system.	
/var/log/mail.log	Mail server logs.	
/var/log/messages	General messages and system related information.	
/var/log/news/*	The news command keeps you informed of news concerning the system.	
/var/log/ntpstats	Logs for network configuration protocol.	
/var/log/kern.log	Kernel logs.	
/var/log/quagga/*	Where Quagga logs to once enabled.	This is how Cumulus Networks troubleshoots routing. For example an md5 or mtu mismatch with OSPF.
/var/log/switchd.log/	The HAL log for Cumulus Linux.	This is specific to Cumulus Linux. Any switchd crashes are logged here.
/var/log/syslog	The main system log, which logs everything except auth-related messages.	The primary log; it's easiest to grep this file to see what occurred during a problem.
/var/log/wtmp	Login records file.	
	apt command log file.	

Log	Description	Why is this important?
/var/log/yum.log		

## Troubleshooting the etc Directory

The [c1-support](#) (see page 417) script replicates the /etc directory.

Files that c1-support deliberately excludes are:

File	Description
/etc/nologin	nologin prevents unprivileged users from logging into the system.
/etc/alternatives	update-alternatives creates, removes, maintains and displays information about the symbolic links comprising the Debian alternatives system.

This is the alphabetical of the output from running `ls -1` on the /etc directory structure created by c1-support. The green highlighted rows are the ones Cumulus Networks finds most important when troubleshooting problems.

File	Description	Why is this important?
adduser.conf	The file /etc/adduser.conf contains defaults for the programs adduser, addgroup, deluser, and delgroup.	
adjtime	Corrects the time to synchronize the <a href="#">system clock</a> .	
apt	apt (Advanced Package Tool) is the command-line <a href="#">tool for handling packages</a> . This folder contains all the configurations.	apt interactions or unsupported apps can affect machine performance.
audisp	The directory that contains audisp-remote.conf, which is the file that controls the <a href="#">configuration of the audit remote logging subsystem</a> .	
audit	The directory that contains the /etc/audit/auditd.conf, which contains configuration information specific to the <a href="#">audit daemon</a> .	
bash.bashrc	Bash is an <a href="#">sh-compatible command language interpreter</a> that executes commands read from standard input or from a file.	

File	Description	Why is this important?
bash_completion	This points to <code>/usr/share/bash-completion/bash_completion</code> .	
bash_completion.d	This folder contains app-specific code for Bash completion on Cumulus Linux, such as <code>mstpcctl</code> .	
bcm.d	Broadcom-specific ASIC file structure (hardware interaction). If there are questions contact the Cumulus Networks Support team. This is unique to Cumulus Linux.	
bindresvport.blacklist	This file contains a list of port numbers between 600 and 1024, which should not be used by <code>bindresvport</code> .	
ca-certificates	The folder for <code>ca-certificates</code> . It is empty by default on Cumulus Linux; see below for more information.	
ca-certificates.conf	Each lines list the pathname of activated CA certificates under <code>/usr/share/ca-certificates</code> .	
calendar	The system-wide <a href="#">default calendar file</a> .	
chef	This is an example of something that is not included by default. In this instance, <code>c1-support</code> included the <a href="#">chef folder</a> for some reason.	This is not installed by default, but this tool could have been installed or configured incorrectly, which is why it's included in the <code>c1-support</code> output.
cron.d	<code>cron</code> is a daemon that <a href="#">executes scheduled commands</a> .	
cron.daily	See above.	
cron.hourly	See above.	
cron.monthly	See above.	
cron.weekly	See above.	
crontab	See above.	
cumulus	This directory contains the following:	

File	Description	Why is this important?
	<ul style="list-style-type: none"> <li>• ACL information, stored in the <code>acl</code> directory.</li> <li>• <code>switchd</code> configuration file, <code>switchd.conf</code>.</li> <li>• <code>qos</code>, which is under the <code>datapath</code> directory.</li> <li>• The routing protocol process priority, <code>nice.conf</code>.</li> <li>• The breakout cable configuration, under <code>ports.conf</code>.</li> </ul>	This folder is specific to Cumulus Linux and does not exist on other Linux platforms. For example, while you can configure <code>iptables</code> , to hardware accelerate rules into the hardware you need to use <code>cl-acltool</code> and have the rules under the <code>/etc/cumulus/acl/policy.d/&lt;filename.rules</code> )
<code>debconf.conf</code>	Debconf is a <a href="#">configuration system for Debian packages</a> .	
<code>debian_version</code>	The complete <a href="#">Debian version string</a> .	
<code>debsums-ignore</code>	<code>debsums</code> <a href="#">verifies installed package files</a> against their MD5 checksums. This file identifies the packages to ignore.	
<code>default</code>	This folder contains files with configurable flags for many different applications (most installed by default or added manually). For example, <code>/etc/default/networking</code> has a flag for <code>EXCLUDE_INTERFACES=</code> , which is set to nothing by default, but a user could change it to something like <code>sdp3</code> .	
<code>deluser.conf</code>	The file <code>/etc/deluser.conf</code> contains defaults for the programs <code>deluser</code> and <code>delgroup</code> .	
<code>dhcp</code>	This directory contains <a href="#">DHCP-specific information</a> .	
<code>dpkg</code>	The <a href="#">package manager</a> for Debian.	
<code>e2fsck.conf</code>	The <a href="#">configuration file for e2fsck</a> . It controls the default behavior of <code>e2fsck</code> while it checks ext2, ext3 or ext4 filesystems.	
<code>environment</code>	Utilized by <code>pam_env</code> for setting and unsetting environment variables.	
<code>ethertypes</code>	This file can be used to <a href="#">show readable characters</a> instead of hexadecimal numbers for the protocols. For example, <code>0x0800</code> will be represented by IPv4.	

File	Description	Why is this important?
fstab	Static information about the <a href="#">filesystems</a> .	
fstab.d	The directory that can contain additional <code>fstab</code> information; it is empty by default.	
fw_env.config	Configuration file utilized by <a href="#">U-Boot</a> .	
gai.conf	Configuration file for sorting the return information from <a href="#">getaddrinfo</a> .	
groff	The directory containing information for <code>groffer</code> , an application used for displaying <a href="#">Unix man pages</a> .	
group	The <code>/etc/group</code> file is a text file that <a href="#">defines the groups</a> on the system.	
group-	Backup for the <code>/etc/group</code> file.	
gshadow	<code>/etc/gshadow</code> contains the <a href="#">shadowed information for group accounts</a> .	
gshadow-	Backup for the <code>/etc/gshadow</code> file.	
host.conf	<a href="#">Resolver configuration file</a> , which contains options like <code>multi</code> that determines whether <code>/etc/hosts</code> will respond with multiple entries for DNS names.	
hostname	The <a href="#">system host name</a> , such as leaf1, spine1, sw1.	
hosts	The <a href="#">static table lookup</a> for hostnames.	
hosts.allow	The part of the <a href="#">host_access</a> program for controlling a simple access control language. <code>hosts.allow=Access</code> is granted when a daemon/client pair matches an entry.	
hosts.deny	See hosts.allow above, except that access is denied when a daemon/client pair matches an entry.	
init	Default location of the <a href="#">system job configuration files</a> .	
init.d		

File	Description	Why is this important?
	In order for a service to start when the switch boots, you should add the <a href="#">necessary script</a> to the director here. The differences between <code>init</code> and <code>init.d</code> are explained well <a href="#">here</a> .	
inittab	The format of the <code>inittab</code> file used by the sysv-compatible <code>init</code> process.	
inputrc	The initialization file utilized by <code>readline</code> .	
insserv	This application <a href="#">enables installed system init scripts</a> ; this directory is empty by default.	
insserv.conf	Configuration file for <code>insserv</code> .	
insserv.conf.d	Additional directory for <code>insserv</code> configurations.	
iproute2	Directory containing values for the Linux command line tool <code>ip</code> .	
issue	<code>/etc/issue</code> is a text file that contains a <a href="#">message or system identification</a> to be printed before the login prompt.	
issue.net	Identification file for <code>telnet</code> sessions.	
jdoo	<code>jdoo</code> is a utility for <a href="#">monitoring services</a> (see page 404) on a Cumulus Linux system; this directory has configuration files beneath it.	
ld.so.cache	Contains a <a href="#">compiled list of candidate libraries</a> previously found in the augmented library path.	
ld.so.conf	Used by the <code>ldconfig</code> tool, which <a href="#">configures dynamic linker run-time bindings</a> .	
ld.so.conf.d	The directory that contains additional <code>ld.so.conf</code> configuration (see above).	
ldap	The directory containing the <code>ldap.conf</code> configuration file used to set the system-wide default to be applied when running LDAP clients.	
libaudit.conf	Configuration file utilized by <code>get_auditfail_action</code> .	

File	Description	Why is this important?
libnl-3	Directory for the configuration relating to the <a href="#">libnl library</a> , which is the core library for implementing the fundamentals required to use the netlink protocol such as socket handling, message construction and parsing, and sending and receiving of data.	
lldpd.d	Directory containing configuration files whose commands are executed by <code>lldpccli</code> at startup.	
localtime	Copy of the original data file for <code>/etc/timezone</code> .	
logcheck	Directory containing <code>logcheck.conf</code> and logfiles utilized by the <code>log check</code> program, which scans system logs for interesting lines.	
login.defs	<a href="#">Shadow password suite configuration</a> .	
logrotate.conf	Rotates, compresses and mails <a href="#">system logs</a> .	
logrotate.d	Directory containing additional log rotate configurations.	
lsb-release	Shows the current version of <a href="#">Linux</a> on the system. Run <code>cat /etc/lsb-release</code> for output.	This shows you the version of the operating system you are running; also compare this to the output of <code>c1-img-select</code> .
magic	Used by the <code>file</code> command to determine file type. <code>magic</code> tests check for files with data in particular fixed formats.	
magic.mime	The <code>magic MIME type</code> causes the <code>file</code> command to output MIME type strings rather than the more traditional human readable ones.	
mailcap	The <code>mailcap</code> file is read by the metamail program to determine how to <a href="#">display non-text at the local site</a> .	
mailcap.order	The <code>order of entries</code> in the <code>/etc/mailcap</code> file can be altered by editing the <code>/etc/mailcap.order</code> file.	
manpath.config		

File	Description	Why is this important?
	The <b>manpath configuration file</b> is used by the manual page utilities to assess users' manpaths at run time, to indicate which manual page hierarchies (manpaths) are to be treated as system hierarchies and to assign them directories to be used for storing cat files.	
mime.types	MIME type description file for <b>cups</b> .	
mke2fs.conf	Configuration file for <b>mke2fs</b> , which is a program that <b>creates an ext, ext3 or ext4 filesystem</b> .	
modprobe.d	Configuration directory for <b>modprobe</b> , which is a utility that can <b>add and remove modules from the Linux kernel</b> .	
modules	The kernel modules to load at boot time.	
motd	The contents of <b>/etc/motd</b> ("message of the day") are displayed by <b>pam_motd</b> after a successful login but just before it executes the login shell.	
mtab	The programs <b>mount</b> and <b>umount</b> maintain a list of <b>currently mounted filesystems</b> in the <b>/etc/mtab</b> file. If no arguments are given to <b>mount</b> , this list is printed.	
nanorc	The GNU <b>nano</b> <b>rcfile</b> .	
network	Contains the <b>network interface configuration</b> for <b>ifup</b> and <b>ifdown</b> .	The main configuration file is under <b>/etc/network/interfaces</b> . This is where you configure L2 and L3 information for all of your front panel ports (swp interfaces). Settings like MTU, link speed, IP address information, VLANs are all done here.
networks	<b>Network name information</b> .	
nsswitch.conf	<b>System databases and name service switch configuration file</b> .	
ntp.conf	<b>NTP (network time protocol) server configuration file</b> .	
openvswitch		

File	Description	Why is this important?
	The directory containing the <a href="#">conf.db file</a> , which is used by <code>ovsdb-server</code> .	
openvswitch-vtep	Configuration files used for the VTEP daemon and <code>ovsdb-server</code> .	
opt	Host-specific configuration files for <a href="#">add-on applications</a> installed in <code>/opt</code> .	
os-release	<a href="#">Operating system identification</a> .	
pam.conf	The <a href="#">PAM (pluggable authentication module)</a> configuration file. When a PAM-aware privilege granting application is started, it activates its attachment to the PAM-API. This activation performs a number of tasks, the most important being the reading of the configuration file(s).	
pam.d	Alternate directory to configure PAM (see above).	
passwd	<a href="#">User account information</a> .	
passwd-	Backup file for <code>/etc/passwd</code> .	
perl	<a href="#">Perl</a> is an available scripting language. <code>/etc/perl</code> contains configuration files specific to Perl.	
profile	<code>/etc/profile</code> is utilized by <code>sysprofile</code> , a modular centralized shell configuration.	
profile.d	The directory version of the above, which contains configuration files.	
protocols	The <a href="#">protocols definition file</a> , a plain ASCII file that describes the various DARPA net protocols that are available from the TCP/IP subsystem.	
ptm.d	The directory containing scripts that are run if <a href="#">PTM</a> (see page 145) passes or fails.	Cumulus Linux-specific folder for PTM (prescriptive topology manager).
python	<code>python</code> is an available scripting language.	
python2.6	The 2.6 version of <code>python</code> .	

File	Description	Why is this important?
python2.7	The 2.7 version of <code>python</code> .	
quagga	Contains the configuration files for the <a href="#">Quagga routing suite</a> (see page 320), the preferred Cumulus Linux routing engine.	
rc.local	The <code>/etc/rc.local</code> script is used by the system administrator to <a href="#">execute after all the normal system services are started</a> , at the end of the process of switching to a multiuser runlevel. You can use it to start a custom service, for example, a server that's installed in <code>/usr/local</code> . Most installations don't need <code>/etc/rc.local</code> ; it's provided for the minority of cases <a href="#">where it's needed</a> .	
rc0.d	Like <code>rc.local</code> , these scripts are booted by default, but the number of the folder represents the <a href="#">Linux runlevel</a> . This folder 0 represents runlevel 0 (halt the system).	
rc1.d	This is run level 1, which is single-user/minimal mode.	
rc2.d	Runlevels 2 through 5 are multiuser modes. Debian systems (such as Cumulus Linux) come with <code>id=2</code> , which indicates that the <a href="#">default runlevel will be 2 when the multi-user state is entered</a> , and the scripts in <code>/etc/rc2.d/</code> will be run.	
rc3.d	See above.	
rc4.d	See above.	
rc5.d	See above.	
rc6.d	Runlevel 6 is reboot the system.	
rcS.d	S stands for <i>single</i> and is equivalent to rc1.	
resolv.conf	Resolver configuration file, which is where DNS is set (domain, nameserver and search).	You need DNS to reach the Cumulus Linux repository.
rmt	This is not a mistake. The <a href="#">shell script</a> <code>/etc/rmt</code> is provided for compatibility with other Unix-like systems, some of which have utilities that expect to find (and execute) <code>rmt</code> in the <code>/etc</code> directory on remote systems.	

File	Description	Why is this important?
rpc	The <code>rpc</code> file contains <b>human-readable names</b> that can be used in place of RPC program numbers.	
rsyslog.conf	The <code>rsyslog.conf</code> file is the main configuration file for <code>rsyslogd</code> , which logs system messages on *nix systems.	
rsyslog.d	The directory containing additional configuration for <code>rsyslog.conf</code> (see above).	
securetty	This file lists terminals into which the <b>root user can log in</b> .	
security	The <code>/etc/security</code> directory contains <b>security-related configurations files</b> . Whereas PAM concerns itself with the methods used to authenticate any given user, the files under <code>/etc/security</code> are concerned with just what a user can or cannot do. For example, the <code>/etc/security/access.conf</code> file contains a list of which users are allowed to log in and from what host (for example, using telnet). The <code>/etc/security/limits.conf</code> file contains various system limits, such as maximum number of processes.	
selinux	NSA <b>Security-Enhanced Linux</b> .	
sensors.d	The directory from which the <code>sensors</code> program loads its configuration; this is unique for each hardware platform. See also <b>Monitoring System Hardware</b> (see page 404).	
sensors3.conf	The <code>sensors.conf</code> file describes how <code>libsensors</code> , and thus all programs using it, should translate the raw readings from the kernel modules to real-world values.	
services	<code>services</code> is a plain ASCII file providing a mapping between human-readable textual names for internet services and their underlying assigned port numbers and protocol types.	
shadow	shadow is a file that <b>contains the password information</b> for the system's accounts and optional aging information.	
shadow-	The backup for the <code>/etc/shadow</code> file.	

File	Description	Why is this important?
shells	The pathnames of <b>valid login shells</b> .	
skel	The skeleton directory (usually <code>/etc/skel</code> ) is used to copy default files and also sets a umask for the creation used by <code>pam_mkhomedir</code> .	
snmp	Interface functions to the <b>SNMP</b> (simple network management protocol) toolkit.	
ssh	The <b>ssh configuration</b> .	
ssl	The <b>OpenSSL ssl library</b> implements the Secure Sockets Layer (SSL v2/v3) and Transport Layer Security (TLS v1) protocols. This directory holds certificates and configuration.	
staff-group-forusr-local	Use <code>cat</code> or <code>more</code> on this file to learn more information, see <a href="http://bugs.debian.org/299007">http://bugs.debian.org/299007</a> .	
sudoers	The <code>sudoers</code> policy plugin determines a user's <b>sudo privileges</b> .	
sudoers.d	The directory file containing additional <code>sudoers</code> configuration (see above).	
sysctl.conf	Configures <b>kernel parameters at boot</b> .	
sysctl.d	The directory file containing additional configuration (see above).	
systemd	<code>systemd</code> system and service manager.	
terminfo	Terminal capability database.	
timezone	If this file exists, it is read and its contents are used as the <b>time zone name</b> .	
ucf.conf	The update configuration file <b>preserves user changes</b> in configuration files.	
udev	Dynamic device management.	
ufw	Provides both a command line interface and a framework for managing a <b>netfilter firewall</b> .	

File	Description	Why is this important?
vim	Configuration file for command line tool vim.	
wgetrc	Configuration file for command line tool wget.	

## Troubleshooting the support Directory

The `support` directory is unique in the fact that it is not a copy of the switch's filesystem. Actually, it is the output from various commands. For example:

File	Equivalent Command	Description
<code>support /ip addr</code>	<code>cumulus@switch:~\$ ip addr show</code>	This shows you all the interfaces (including swp front panel ports), IP address information, admin state and physical state.

## Managing Application Daemons

You manage application daemons in Cumulus Linux in the following ways:

- Identifying active listener ports
- Identifying daemons currently active or stopped
- Identifying boot time state of a specific daemon
- Disabling or enabling a specific daemon

### Contents

(Click to expand)

- [Contents \(see page 433\)](#)
- [Identifying Active Listener Ports for IPv4 and IPv6 \(see page 433\)](#)
- [Identifying Daemons Currently Active or Stopped \(see page 434\)](#)
- [Identifying Boot Time State of a Specific Daemon \(see page 434\)](#)
- [Disabling or Enabling a Specific Daemon \(see page 435\)](#)

### Identifying Active Listener Ports for IPv4 and IPv6

You can identify the active listener ports under both IPv4 and IPv6 using the `lsof` command:

```
cumulus@switch:~$ sudo lsof -Pnl +M -i4
COMMAND PID USER FD TYPE DEVICE SIZE/OFF NODE NAME
ntpd 1882 104 16u IPv4 3954 0t0 UDP *:123
ntpd 1882 104 18u IPv4 3963 0t0 UDP 127.0.0.1:123
```

```

  ntpd 1882 104 19u IPv4 3964 0t0 UDP 192.168.8.37:123
  snmpd 1987 105 8u IPv4 5423 0t0 UDP *:161
  zebra 1993 103 10u IPv4 5151 0t0 TCP 127.0.0.1:2601 (LISTEN)
  sshd 2496 0 3u IPv4 5809 0t0 TCP *:22 (LISTEN)
  jdoo 2622 0 6u IPv4 6132 0t0 TCP 127.0.0.1:2812 (LISTEN)
  sshd 31700 0 3r IPv4 187630 0t0 TCP 192.168.8.37:22->192.168.8.3:50386
  (ESTABLISHED)

cumulus@switch:~$ sudo lsof -Pnl +M -i6
COMMAND PID USER FD TYPE DEVICE SIZE/OFF NODE NAME
  ntpd 1882 104 17u IPv6 3955 0t0 UDP *:123
  ntpd 1882 104 20u IPv6 3965 0t0 UDP [::1]:123
  ntpd 1882 104 21u IPv6 3966 0t0 UDP [fe80::7272:cfff:fe96:6639]:123
  sshd 2496 0 4u IPv6 5811 0t0 TCP *:22 (LISTEN)

```

## ***Identifying Daemons Currently Active or Stopped***

To determine which daemons are currently active or stopped, use the `service --status-all` command, then pipe the results to `grep`, using the `-` or `+` operators:

```

cumulus@switch:~$ sudo service --status-all | grep +
[ ? ] acinit
[ + ] arp_refresh
[ + ] auditd
...
cumulus@switch:~$ sudo service --status-all | grep -
[ - ] isc-dhcp-server
[ - ] openvswitch-vtep
[ - ] ptmd
...

```

## ***Identifying Boot Time State of a Specific Daemon***

The `ls` command can provide the boot time state of a daemon. A file link with a name starting with **S** identifies a boot-time-enabled daemon. A file link with a name starting with **K** identifies a disabled daemon.

```
cumulus@switch:~/etc$ sudo ls -l rc*.d | grep <daemon name>
```

For example:

```
cumulus@switch:~/etc$ sudo ls -l rc*.d | grep snmpd
lrwxrwxrwx 1 root root 15 Apr 4 2014 K02snmpd -> ../init.d/snmpd
lrwxrwxrwx 1 root root 15 Apr 4 2014 K02snmpd -> ../init.d/snmpd
lrwxrwxrwx 1 root root 15 Apr 4 2014 S01snmpd -> ../init.d/snmpd
lrwxrwxrwx 1 root root 15 Apr 4 2014 S01snmpd -> ../init.d/snmpd
lrwxrwxrwx 1 root root 15 Apr 4 2014 S01snmpd -> ../init.d/snmpd
lrwxrwxrwx 1 root root 15 Apr 4 2014 S01snmpd -> ../init.d/snmpd
lrwxrwxrwx 1 root root 15 Apr 4 2014 K02snmpd -> ../init.d/snmpd
```

## ***Disabling or Enabling a Specific Daemon***

To enable or disable a specific daemon, run:

```
cumulus@switch:~$ update-rc.d <daemon> disable | enable
```

For example:

```
cumulus@switch:~/etc$ sudo update-rc.d snmpd disable
update-rc.d: using dependency based boot sequencing
insserv: warning: current start runlevel(s) (empty) of script `snmpd'
overrides LSB defaults (2 3 4 5).
insserv: warning: current stop runlevel(s) (0 1 2 3 4 5 6) of script
`snmpd' overrides LSB defaults (0 1 6).
insserv: warning: current start runlevel(s) (empty) of script `snmpd'
overrides LSB defaults (2 3 4 5).
insserv: warning: current stop runlevel(s) (0 1 2 3 4 5 6) of script
`snmpd' overrides LSB defaults (0 1 6).
```

```
cumulus@switch:~/etc$ sudo ls -l rc*.d | grep snmpd
lrwxrwxrwx 1 root root 15 Apr 4 2014 K02snmpd -> ../init.d/snmpd
lrwxrwxrwx 1 root root 15 Apr 4 2014 K02snmpd -> ../init.d/snmpd
lrwxrwxrwx 1 root root 15 Feb 13 17:35 K02snmpd -> ../init.d/snmpd
lrwxrwxrwx 1 root root 15 Feb 13 17:35 K02snmpd -> ../init.d/snmpd
lrwxrwxrwx 1 root root 15 Feb 13 17:35 K02snmpd -> ../init.d/snmpd
lrwxrwxrwx 1 root root 15 Feb 13 17:35 K02snmpd -> ../init.d/snmpd
lrwxrwxrwx 1 root root 15 Apr 4 2014 K02snmpd -> ../init.d/snmpd
```

```
cumulus@switch:~/etc$ sudo update-rc.d snmpd enable
update-rc.d: using dependency based boot sequencing
```

```
cumulus@switch:~/etc$ sudo ls -l rc*.d | grep snmpd
lrwxrwxrwx 1 root root 15 Apr 4 2014 K02snmpd -> ../../init.d/snmpd
lrwxrwxrwx 1 root root 15 Apr 4 2014 K02snmpd -> ../../init.d/snmpd
lrwxrwxrwx 1 root root 15 Feb 13 17:35 S01snmpd -> ../../init.d/snmpd
lrwxrwxrwx 1 root root 15 Feb 13 17:35 S01snmpd -> ../../init.d/snmpd
lrwxrwxrwx 1 root root 15 Feb 13 17:35 S01snmpd -> ../../init.d/snmpd
lrwxrwxrwx 1 root root 15 Feb 13 17:35 S01snmpd -> ../../init.d/snmpd
lrwxrwxrwx 1 root root 15 Apr 4 2014 K02snmpd -> ../../init.d/snmpd
```

## Troubleshooting Network Interfaces

The following sections describe various ways you can troubleshoot `ifupdown2`.

### Contents

(Click to expand)

- [Contents \(see page 436\)](#)
- [Enabling Logging for Networking \(see page 436\)](#)
- [Using ifquery to Validate and Debug Interface Configurations \(see page 437\)](#)
- [Debugging Mako Template Errors \(see page 438\)](#)
- [ifdown Cannot Find an Interface that Exists \(see page 439\)](#)
- [Removing All References to a Child Interface \(see page 439\)](#)
- [MTU Set on a Logical Interface Fails with Error: "Numerical result out of range" \(see page 441\)](#)
- [Interpreting iproute2 batch Command Failures \(see page 441\)](#)
- [Understanding the "RTNETLINK answers: Invalid argument" Error when Adding a Port to a Bridge \(see page 441\)](#)

### Enabling Logging for Networking

The `/etc/default/networking` file contains two settings for logging:

- To get `ifupdown2` logs when the switch boots (stored in `syslog`)
- To enable logging when you run `service networking [start|stop|reload]`

This file also contains an option for excluding interfaces when you boot the switch or run `service networking start|stop|reload`. You can exclude any interface specified in `/etc/network/interfaces`. These interfaces do not come up when you boot the switch or start/stop/reload the networking service.

```
$cat /etc/default/networking
#
#
# Parameters for the /etc/init.d/networking script
```

```

#
#
# Change the below to yes if you want verbose logging to be enabled
VERBOSE="no"

# Change the below to yes if you want debug logging to be enabled
DEBUG="no"

# Change the below to yes if you want logging to go to syslog
SYSLOG="no"

# Exclude interfaces
EXCLUDE_INTERFACES=

```

## ***Using ifquery to Validate and Debug Interface Configurations***

You use `ifquery` to print parsed `interfaces` file entries.

To use `ifquery` to pretty print `iface` entries from the `interfaces` file, run:

```

cumulus@switch:~$ sudo ifquery bond0
auto bond0
iface bond0
    address 14.0.0.9/30
    address 2001:ded:beef:2::1/64
    bond-slaves swp25 swp26
    bond-mode 802.3ad
    bond-miimon 100
    bond-use-carrier 1
    bond-lacp-rate 1
    bond-min-links 1
    bond-xmit-hash-policy layer3+4

```

Use `ifquery --check` to check the current running state of an interface within the `interfaces` file. It will return exit code 0 or 1 if the configuration does not match. The line `bond-xmit-hash-policy layer3+7` below fails because it should read `bond-xmit-hash-policy layer3+4`.

```

cumulus@switch:~$ sudo ifquery --check bond0
iface bond0
    bond-mode 802.3ad          [pass]
    bond-miimon 100           [pass]
    bond-use-carrier 1         [pass]

```

```
bond-lacp-rate 1          [pass]
bond-min-links 1          [pass]
bond-xmit-hash-policy layer3+7 [fail]
bond-slaves swp25 swp26    [pass]
address 14.0.0.9/30       [pass]
address 2001:ded:beef:2::1/64 [pass]
```



`ifquery --check` is an experimental feature.

Use `ifquery --running` to print the running state of interfaces in the `interfaces` file format:

```
cumulus@switch:~$ sudo ifquery --running bond0
auto bond0
iface bond0
    bond-xmit-hash-policy layer3+4
    bond-mimon 100
    bond-lacp-rate 1
    bond-min-links 1
    bond-slaves swp25 swp26
    bond-mode 802.3ad
    address 14.0.0.9/30
    address 2001:ded:beef:2::1/64
```

`ifquery --syntax-help` provides help on all possible attributes supported in the `interfaces` file. For complete syntax on the `interfaces` file, see `man interfaces` and `man ifupdown-addons-interfaces`.

You can use `ifquery --print-savedstate` to check the `ifupdown2` state database. `ifdown` works only on interfaces present in this state database.

```
cumulus@leaf1$ sudo ifquery --print-savedstate eth0
auto eth0
iface eth0 inet dhcp
```

## Debugging Mako Template Errors

An easy way to debug and get details about template errors is to use the `mako-render` command on your `interfaces` template file or on `/etc/network/interfaces` itself.

```
cumulus@switch:~$ sudo mako-render /etc/network/interfaces
# This file describes the network interfaces available on your system
```

```
# and how to activate them. For more information, see interfaces(5).

# The loopback network interface
auto lo
iface lo inet loopback

# The primary network interface
auto eth0
iface eth0 inet dhcp
#auto eth1
#iface eth1 inet dhcp

# Include any platform-specific interface configuration
source /etc/network/interfaces.d/*.*if

# ssim2 added

auto swp45
iface swp45

auto swp46
iface swp46

cumulus@switch:~$ sudo mako-render /etc/network/interfaces.d
/<interfaces_stub_file>
```

## ***ifdown Cannot Find an Interface that Exists***

If you are trying to bring down an interface that you know exists, use `ifdown` with the `--use-current-config` option to force `ifdown` to check the current `/etc/network/interfaces` file to find the interface. This can solve issues where the `ifup` command issues for that interface was interrupted before it updated the state database. For example:

```
cumulus@switch:~$ sudo ifdown br0
error: cannot find interfaces: br0 (interface was probably never up ?)

cumulus@switch:~$ sudo brctl show
bridge name      bridge id           STP enabled      interfaces
br0              8000.44383900279f    yes            downlink
                                         peerlink

cumulus@switch:~$ sudo ifdown br0 --use-current-config
```

## Removing All References to a Child Interface

If you have a configuration with a child interface, whether it's a VLAN, bond or another physical interface, and you remove that interface from a running configuration, you must remove every reference to it in the configuration. Otherwise, the interface continues to be used by the parent interface.

For example, consider the following configuration:

```
auto lo
iface lo inet loopback

auto eth0
iface eth0 inet dhcp

auto bond1
iface bond1
    bond-miimon 100
    bond-slaves swp2 swp1
    bond-mode 802.3ad
    bond-lACP-rate 1
    bond-min-links 1
    bond-xmit-hash-policy layer3+4

auto bond3
iface bond3
    bond-miimon 100
    bond-slaves swp8 swp6 swp7
    bond-mode 802.3ad
    bond-lACP-rate 1
    bond-min-links 1
    bond-xmit-hash-policy layer3+4

auto br0
iface br0
    bridge-ports swp3 swp5 bond1 swp4 bond3
    bridge-pathcosts swp3=4 swp5=4 swp4=4
    address 11.0.0.10/24
    address 2001::10/64
```

Notice that bond1 is a member of br0. If you comment out or simply delete bond1 from `/etc/network/interfaces`, you must remove the reference to it from the br0 configuration. Otherwise, if you reload the configuration with `ifreload -a`, bond1 is still part of br0.

## ***MTU Set on a Logical Interface Fails with Error: "Numerical result out of range"***

This error occurs when the MTU (see page 111) you are trying to set on an interface is higher than the MTU of the lower interface or dependent interface. Linux expects the upper interface to have an MTU less than or equal to the MTU on the lower interface.

In the example below, the swp1.100 VLAN interface is an upper interface to physical interface swp1. If you want to change the MTU to 9000 on the VLAN interface, you must include the new MTU on the lower interface swp1 as well.

```
auto swp1.100
iface swp1.100
    mtu 9000

auto swp1
iface swp1
    mtu 9000
```

## ***Interpreting iproute2 batch Command Failures***

ifupdown2 batches iproute2 commands for performance reasons. A batch command contains ip -force -batch - in the error message. The command number that failed is at the end of this line: Command failed -:1.

Below is a sample error for the command 1: link set dev host2 master bridge. There was an error adding the bond *host2* to the bridge named *bridge* because *host2* did not have a valid address.

```
error: failed to execute cmd 'ip -force -batch - [link set dev host2 master
bridge
addr flush dev host2
link set dev host1 master bridge
addr flush dev host1
]' (RTNETLINK answers: Invalid argument
Command failed -:1)
warning: bridge configuration failed (missing ports)
```

## ***Understanding the "RTNETLINK answers: Invalid argument" Error when Adding a Port to a Bridge***

This error can occur when the bridge port does not have a valid hardware address.

This can typically occur when the interface being added to the bridge is an incomplete bond; a bond without slaves is incomplete and does not have a valid hardware address.

# Network Troubleshooting

Cumulus Linux contains a number of command line and analytical tools to help you troubleshoot issues with your network.

## Contents

(Click to expand)

- [Contents \(see page 442\)](#)
- [Commands \(see page 442\)](#)
- [Checking Reachability Using ping \(see page 442\)](#)
- [Printing Route Trace Using traceroute \(see page 443\)](#)
- [Manipulating the System ARP Cache \(see page 443\)](#)
- [Generating Traffic Using mz \(see page 444\)](#)
- [Creating Counter ACL Rules \(see page 445\)](#)
- [Configuring SPAN and ERSPAN \(see page 446\)
  - \[Configuring SPAN for Switch Ports \\(see page 447\\)\]\(#\)
  - \[Configuring SPAN for Bonds \\(see page 450\\)\]\(#\)
  - \[Configuring ERSPAN \\(see page 451\\)\]\(#\)
  - \[Removing SPAN Rules \\(see page 452\\)\]\(#\)](#)
- [Monitoring Control Plane Traffic with tcpdump \(see page 452\)](#)
- [Configuration Files \(see page 453\)](#)
- [Useful Links \(see page 453\)](#)
- [Caveats and Errata \(see page 453\)](#)

## Commands

- [arp](#)
- [cl-acltool](#)
- [ip](#)
- [mz](#)
- [ping](#)
- [tcpdump](#)
- [traceroute](#)

## Checking Reachability Using ping

`ping` is used to check reachability of a host. `ping` also calculates the time it takes for packets to travel the round trip. See `man ping` for details.

To test the connection to an IPv4 host:

```
cumulus@switch:~$ ping 206.190.36.45
PING 206.190.36.45 (206.190.36.45) 56(84) bytes of data.
64 bytes from 206.190.36.45: icmp_req=1 ttl=53 time=40.4 ms
64 bytes from 206.190.36.45: icmp_req=2 ttl=53 time=39.6 ms
...
...
```

To test the connection to an IPv6 host:

```
cumulus@switch:~$ ping6 -I swp1 fe80::202:ff:fe00:2
PING fe80::202:ff:fe00:2(fe80::202:ff:fe00:2) from fe80::202:ff:fe00:1
swp1: 56 data bytes
64 bytes from fe80::202:ff:fe00:2: icmp_seq=1 ttl=64 time=1.43 ms
64 bytes from fe80::202:ff:fe00:2: icmp_seq=2 ttl=64 time=0.927 ms
```

## ***Printing Route Trace Using traceroute***

`traceroute` tracks the route that packets take from an IP network on their way to a given host. See `man traceroute` for details.

To track the route to an IPv4 host:

```
cumulus@switch:~$ traceroute www.google.com
traceroute to www.google.com (74.125.239.49), 30 hops max, 60 byte packets
1 fw.cumulusnetworks.com (192.168.1.1) 0.614 ms 0.863 ms 0.932 ms
2 router.hackerdojo.com (157.22.42.1) 15.459 ms 16.447 ms 16.818 ms
3 gw-cpe-hackerdojo.via.net (157.22.10.97) 18.470 ms 18.473 ms 18.897 ms
4 ge-1-5-v223.core1.uspao.via.net (157.22.10.81) 20.419 ms 20.422 ms
21.026 ms
5 core2-1-1-0.pao.net.google.com (198.32.176.31) 22.347 ms 22.584 ms
24.328 ms
6 216.239.49.250 (216.239.49.250) 24.371 ms 25.757 ms 25.987 ms
7 72.14.232.35 (72.14.232.35) 27.505 ms 22.925 ms 22.323 ms
8 nug04s19-in-f17.1e100.net (74.125.239.49) 23.544 ms 21.851 ms 22.604
ms
```

## ***Manipulating the System ARP Cache***

`arp` manipulates or displays the kernel's IPv4 network neighbor cache. See `man arp` for details.

To display the ARP cache:

```
cumulus@switch:~$ arp -a
? (11.0.2.2) at 00:02:00:00:00:10 [ether] on swp3
? (11.0.3.2) at 00:02:00:00:00:01 [ether] on swp4
? (11.0.0.2) at 44:38:39:00:01:c1 [ether] on swp1
```

To delete an ARP cache entry:

```
cumulus@switch:~$ arp -d 11.0.2.2
cumulus@switch:~$ arp -a
? (11.0.2.2) at <incomplete> on swp3
? (11.0.3.2) at 00:02:00:00:00:01 [ether] on swp4
? (11.0.0.2) at 44:38:39:00:01:c1 [ether] on swp1
```

To add a static ARP cache entry:

```
cumulus@switch:~$ arp -s 11.0.2.2 00:02:00:00:00:10
cumulus@switch:~$ arp -a
? (11.0.2.2) at 00:02:00:00:00:10 [ether] PERM on swp3
? (11.0.3.2) at 00:02:00:00:00:01 [ether] on swp4
? (11.0.0.2) at 44:38:39:00:01:c1 [ether] on swp1
```

## ***Generating Traffic Using mz***

**mz** is a fast traffic generator. It can generate a large variety of packet types at high speed. See `man mz` for details.

For example, to send two sets of packets to TCP port 23 and 24, with source IP 11.0.0.1 and destination 11.0.0.2, do the following:

```
cumulus@switch:~$ sudo mz swp1 -A 11.0.0.1 -B 11.0.0.2 -c 2 -v -t tcp
"dp=23-24"

Mausezahn 0.40 - (C) 2007-2010 by Herbert Haas - http://www.perihel.at/sec
/mz/
Use at your own risk and responsibility!
-- Verbose mode --

This system supports a high resolution clock.
The clock resolution is 4000250 nanoseconds.
Mausezahn will send 4 frames...
```

```

IP: ver=4, len=40, tos=0, id=0, frag=0, ttl=255, proto=6, sum=0, SA=11.
0.0.1, DA=11.0.0.2,
    payload=[see next layer]
TCP: sp=0, dp=23, S=42, A=42, flags=0, win=10000, len=20, sum=0,
    payload=

IP: ver=4, len=40, tos=0, id=0, frag=0, ttl=255, proto=6, sum=0, SA=11.
0.0.1, DA=11.0.0.2,
    payload=[see next layer]
TCP: sp=0, dp=24, S=42, A=42, flags=0, win=10000, len=20, sum=0,
    payload=

IP: ver=4, len=40, tos=0, id=0, frag=0, ttl=255, proto=6, sum=0, SA=11.
0.0.1, DA=11.0.0.2,
    payload=[see next layer]
TCP: sp=0, dp=23, S=42, A=42, flags=0, win=10000, len=20, sum=0,
    payload=

IP: ver=4, len=40, tos=0, id=0, frag=0, ttl=255, proto=6, sum=0, SA=11.
0.0.1, DA=11.0.0.2,
    payload=[see next layer]
TCP: sp=0, dp=24, S=42, A=42, flags=0, win=10000, len=20, sum=0,
    payload=

```

## Creating Counter ACL Rules

In Linux, all ACL rules are always counted. To create an ACL rule for counting purposes only, set the rule action to ACCEPT. See the [Netfilter \(see page 76\)](#) chapter for details on how to use cl-acltool to set up iptables-/ip6tables-/ebtables-based ACLs.



Always place your rules files under /etc/cumulus/acl/policy.d/.

To count all packets going to a Web server:

```

cumulus@switch$ cat sample_count.rules

[iptables]
-A FORWARD -p tcp --dport 80 -j ACCEPT

cumulus@switch:$ sudo cl-acltool -i -p sample_count.rules
Using user provided rule file sample_count.rules
Reading rule file sample_count.rules ...

```

```

Processing rules in file sample_count.rules ...
Installing acl policy... done.

cumulus@switch$ sudo iptables -L -v
Chain INPUT (policy ACCEPT 16 packets, 2224 bytes)
pkts bytes target     prot opt in      out      source
destination

Chain FORWARD (policy ACCEPT 0 packets, 0 bytes)
pkts bytes target     prot opt in      out      source
destination
      2   156 ACCEPT     tcp  --  any    any     anywhere
anywhere           tcp  dpt:http

Chain OUTPUT (policy ACCEPT 44 packets, 8624 bytes)
pkts bytes target     prot opt in      out      source
destination

```



The **-p** option clears out all other rules, and the **-i** option is used to reinstall all the rules.

## Configuring SPAN and ERSPAN

SPAN (Switched Port Analyzer) provides for the mirroring of all packets coming in from or going out of an interface to a local port for monitoring. This port is referred to as a mirror-to-port (MTP). The original packet is still switched, while a mirrored copy of the packet is sent to the MTP port.

ERSPAN (Encapsulated Remote SPAN) enables the mirrored packets to be sent to a monitoring node located anywhere across the routed network. The switch finds the outgoing port of the mirrored packets by doing a lookup of the destination IP address in its routing table. The original L2 packet is encapsulated with GRE for IP delivery. The encapsulated packets have the following format:

```

-----
| MAC_HEADER | IP_HEADER | GRE_HEADER | L2_Mirrored_Packet |
-----
```

SPAN and ERSPAN are configured via `c1-acltool`, the [same utility for security ACL configuration \(see page 76\)](#). The match criteria for SPAN and ERSPAN can only be an interface; more granular match terms are not supported. The interface can be a port, a subinterface or a bond interface. Both ingress and egress interfaces can be matched.

Cumulus Linux supports a maximum of 2 SPAN destinations. Multiple rules can point to the same SPAN destination. The MTP interface can be a physical port, a subinterface, or a bond interface. The SPAN /ERSPAN action is independent of security ACL actions. If packets match both a security ACL rule and a SPAN rule, both actions will be carried out.



Always place your rules files under `/etc/cumulus/acl/policy.d/`.

## Configuring SPAN for Switch Ports

This section describes how to set up, install, verify and uninstall SPAN rules. In the examples that follow, you will span (mirror) switch port `swp4` input traffic and `swp4` output traffic to destination switch port `swp19`.

First, create a rules file in `/etc/cumulus/acl/policy.d/`:

```
cumulus@switch:~$ sudo bash -c 'cat <<EOF > /etc/cumulus/acl/policy.d/span.rules
[iptables]
-A FORWARD --in-interface swp4 -j SPAN --dport swp19
-A FORWARD --out-interface swp4 -j SPAN --dport swp19
EOF'
```



Using `cl-acltool` with the `--out-interface` rule applies to transit traffic only; it does not apply to traffic sourced from the switch.

Install the rules:

```
cumulus@switch:~$ sudo cl-acltool -i
[sudo] password for cumulus:
Reading rule file /etc/cumulus/acl/policy.d/00control_plane.rules ...
Processing rules in file /etc/cumulus/acl/policy.d/00control_plane.rules ...
Reading rule file /etc/cumulus/acl/policy.d/99control_plane_catch_all.rules ...
...
Processing rules in file /etc/cumulus/acl/policy.d/99control_plane_catch_all.rules ...
Reading rule file /etc/cumulus/acl/policy.d/span.rules ...
Processing rules in file /etc/cumulus/acl/policy.d/span.rules ...
Installing acl policy
done.
```



Running the following command is incorrect and will remove **all** existing control-plane rules or other installed rules and only install the rules defined in `span.rules`:

```
cumulus@switch:~$ sudo cl-acltool -i -P /etc/cumulus/acl/policy.d/span.rules
```

Verify that the SPAN rules were installed:

```
cumulus@switch:~$ sudo iptables -L all | grep SPAN
38025 7034K SPAN      all -- swp4    any    anywhere
anywhere          dport:swp19
50832  55M SPAN      all -- any     swp4    anywhere
anywhere          dport:swp19
```

Or to verify all the rules are currently installed, run:

```
cumulus@switch:~$ sudo iptables -L -v
Chain INPUT (policy ACCEPT 0 packets, 0 bytes)
pkts bytes target  prot opt in     out     source
destination
      0     0 DROP    all   --  swp+   any    240.0.0.0/5
anywhere
      0     0 DROP    all   --  swp+   any    loopback/8
anywhere
      0     0 DROP    all   --  swp+   any    base-address.mcast.net/8
anywhere
      0     0 DROP    all   --  swp+   any    255.255.255.255
anywhere
      0     0 SETCLASS  ospf  --  swp+   any    anywhere
anywhere          SETCLASS  class:7
      0     0 POLICE   ospf  --  any    any    anywhere
anywhere          POLICE   mode:pkt rate:2000 burst:2000
      0     0 SETCLASS  tcp   --  swp+   any    anywhere
anywhere          tcp dpt:bgp SETCLASS  class:7
      0     0 POLICE   tcp   --  any    any    anywhere
anywhere          tcp dpt:bgp POLICE   mode:pkt rate:2000 burst:2000
      0     0 SETCLASS  tcp   --  swp+   any    anywhere
anywhere          tcp spt:bgp SETCLASS  class:7
      0     0 POLICE   tcp   --  any    any    anywhere
anywhere          tcp spt:bgp POLICE   mode:pkt rate:2000 burst:2000
      0     0 SETCLASS  tcp   --  swp+   any    anywhere
anywhere          tcp dpt:5342 SETCLASS  class:7
      0     0 POLICE   tcp   --  any    any    anywhere
anywhere          tcp dpt:5342 POLICE   mode:pkt rate:2000 burst:2000
      0     0 SETCLASS  tcp   --  swp+   any    anywhere
anywhere          tcp spt:5342 SETCLASS  class:7
      0     0 POLICE   tcp   --  any    any    anywhere
```

```

anywhere          tcp spt:5342 POLICE mode:pkt rate:2000 burst:2000
      0    0 SETCLASS  icmp -- swp+ any     anywhere
anywhere          SETCLASS class:2
      0    0 POLICE   icmp -- any   any     anywhere
anywhere          POLICE  mode:pkt rate:100 burst:40
      15   5205 SETCLASS  udp  -- swp+ any     anywhere
anywhere          udp dpts:bootps:bootpc SETCLASS class:2
      11   3865 POLICE   udp  -- any   any     anywhere
anywhere          udp dpt:bootps POLICE mode:pkt rate:100 burst:100
      0    0 POLICE   udp  -- any   any     anywhere
anywhere          udp dpt:bootpc POLICE mode:pkt rate:100 burst:100
      0    0 SETCLASS  tcp   -- swp+ any     anywhere
anywhere          tcp dpts:bootps:bootpc SETCLASS class:2
      0    0 POLICE   tcp   -- any   any     anywhere
anywhere          tcp dpt:bootps POLICE mode:pkt rate:100 burst:100
      0    0 POLICE   tcp   -- any   any     anywhere
anywhere          tcp dpt:bootpc POLICE mode:pkt rate:100 burst:100
      17   1088 SETCLASS  igmp -- swp+ any     anywhere
anywhere          SETCLASS class:6
      17   1156 POLICE   igmp -- any   any     anywhere
anywhere          POLICE  mode:pkt rate:300 burst:100
      394  41060 POLICE   all  -- swp+ any     anywhere
anywhere          ADDRTYPE match dst-type LOCAL POLICE mode:pkt rate:
1000 burst:1000 class:0
      0    0 POLICE   all  -- swp+ any     anywhere
anywhere          ADDRTYPE match dst-type IPROUTER POLICE mode:pkt rate:
400 burst:100 class:0
      988  279K SETCLASS  all  -- swp+ any     anywhere
anywhere          SETCLASS class:0

Chain FORWARD (policy ACCEPT 0 packets, 0 bytes)
pkts bytes target     prot opt in     out     source
destination
      0    0 DROP       all  -- swp+ any     240.0.0.0/5
anywhere
      0    0 DROP       all  -- swp+ any     loopback/8
anywhere
      0    0 DROP       all  -- swp+ any     base-address.mcast.net/8
anywhere
      0    0 DROP       all  -- swp+ any     255.255.255.255
anywhere
26864 4672K SPAN      all  -- swp4  any     anywhere
anywhere          dport:swp19 <---- input packets on swp4

40722  47M SPAN      all  -- any     swp4     anywhere

```

```
anywhere           dport:swp19  <---- output packets on swp4

Chain OUTPUT (policy ACCEPT 67398 packets, 5757K bytes)
 pkts bytes target     prot opt in     out     source
destination
```

## Configuring SPAN for Bonds

This section describes how to configure SPAN for all packets going out of bond0 locally to bond1.

First, create a rules file in /etc/cumulus/acl/policy.d/:

```
cumulus@switch:~$ sudo bash -c 'cat <<EOF > /etc/cumulus/acl/policy.d
/span_bond.rules
[iptables]
-A FORWARD --out-interface bond0 -j SPAN --dport bond1
EOF'
```



Using cl-acltool with the --out-interface rule applies to transit traffic only; it does not apply to traffic sourced from the switch.

Install the rules:

```
cumulus@switch:~$ sudo cl-acltool -i
[sudo] password for cumulus:
Reading rule file /etc/cumulus/acl/policy.d/00control_plane.rules ...
Processing rules in file /etc/cumulus/acl/policy.d/00control_plane.rules ...
Reading rule file /etc/cumulus/acl/policy.d/99control_plane_catch_all.rules
...
Processing rules in file /etc/cumulus/acl/policy.d/99control_plane_catch_all.rules ...
Reading rule file /etc/cumulus/acl/policy.d/span_bond.rules ...
Processing rules in file /etc/cumulus/acl/policy.d/span_bond.rules ...
Installing acl policy
done.
```

Verify that the SPAN rules were installed:

```
cumulus@switch:~$ sudo iptables -L -v | grep SPAN
 19  1938 SPAN      all  --  any    bond0    anywhere
 anywhere          dport:bond1
```

## Configuring ERSPAN

This section describes how to configure ERSPAN for all packets coming in from swp1 to 12.0.0.2:

First, create a rules file in /etc/cumulus/acl/policy.d/:

```
cumulus@switch:~$ sudo bash -c 'cat <<EOF > /etc/cumulus/acl/policy.d
/erspan.rules
[iptables]
-A FORWARD --in-interface swp1 -j ERSPAN --src-ip 12.0.0.1 --dst-ip
12.0.0.2 --ttl 64
EOF'
```

Install the rules:

```
cumulus@switch:~$ sudo cl-acltool -i
Reading rule file /etc/cumulus/acl/policy.d/00control_plane.rules ...
Processing rules in file /etc/cumulus/acl/policy.d/00control_plane.rules ...
Reading rule file /etc/cumulus/acl/policy.d/99control_plane_catch_all.rules
...
Processing rules in file /etc/cumulus/acl/policy.d
/99control_plane_catch_all.rules ...
Reading rule file /etc/cumulus/acl/policy.d/erspan.rules ...
Processing rules in file /etc/cumulus/acl/policy.d/erspan.rules ...
Installing acl policy
done.
```

Verify that the ERSPAN rules were installed:

```
cumulus@switch:~$ sudo iptables -L -v | grep SPAN
 69  6804 ERSPAN      all  --  swp1   any    anywhere
 anywhere          ERSPAN src-ip:12.0.0.1 dst-ip:12.0.0.2
```

The `src-ip` option can be any IP address, whether it exists in the routing table or not. The `dst-ip` option must be an IP address reachable via the routing table. The destination IP address must be reachable from a front-panel port, and not the management port. Use `ping` or `ip route get <ip>` to verify that the destination IP address is reachable. Setting the `--ttl1` option is recommended.



When using [Wireshark](#) to review the ERSPAN output, Wireshark may report the message "Unknown version, please report or test to use fake ERSPAN preference", and the trace is unreadable. To resolve this, go into the General preferences for Wireshark, then go to **Protocols** > **ERSPAN** and check the **Force to decode fake ERSPAN frame** option.

## Removing SPAN Rules

To remove your SPAN rules, run:

```
#Remove rules file:  
cumulus@switch:~$ sudo rm /etc/cumulus/acl/policy.d/span.rules  
#Reload the default rules  
cumulus@switch:~$ sudo cl-acltool -i  
cumulus@switch:~$
```

To verify that the SPAN rules were removed:

```
cumulus@switch:~$ sudo cl-acltool -L all | grep SPAN  
cumulus@switch:~$
```

## Monitoring Control Plane Traffic with `tcpdump`

You can use `tcpdump` to monitor control plane traffic — traffic sent to and coming from the switch CPUs. `tcpdump` does **not** monitor data plane traffic; use `cl-acltool` instead (see above).

For more information on `tcpdump`, read [the `tcpdump` documentation](#) and the [`tcpdump` man page](#).

The following example incorporates a few `tcpdump` options:

- `-i bond0`, which captures packets from bond0 to the CPU and from the CPU to bond0
- `host 169.254.0.2`, which filters for this IP address
- `-c 10`, which captures 10 packets then stops

```
cumulus@switch:~$ sudo tcpdump -i bond0 host 169.254.0.2 -c 10  
tcpdump: WARNING: bond0: no IPv4 address assigned  
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode  
listening on bond0, link-type EN10MB (Ethernet), capture size 65535 bytes  
16:24:42.532473 IP 169.254.0.2 > 169.254.0.1: ICMP echo request, id 27785,
```

```

seq 6, length 64
16:24:42.532534 IP 169.254.0.1 > 169.254.0.2: ICMP echo reply, id 27785,
seq 6, length 64
16:24:42.804155 IP 169.254.0.2.40210 > 169.254.0.1.5342: Flags [.], seq
266275591:266277039, ack 3813627681, win 58, options [nop,nop,TS val
590400681 ecr 530346691], length 1448
16:24:42.804228 IP 169.254.0.1.5342 > 169.254.0.2.40210: Flags [.], ack
1448, win 166, options [nop,nop,TS val 530348721 ecr 590400681], length 0
16:24:42.804267 IP 169.254.0.2.40210 > 169.254.0.1.5342: Flags [P.], seq
1448:1836, ack 1, win 58, options [nop,nop,TS val 590400681 ecr 530346691],
length 388
16:24:42.804293 IP 169.254.0.1.5342 > 169.254.0.2.40210: Flags [.], ack
1836, win 165, options [nop,nop,TS val 530348721 ecr 590400681], length 0
16:24:43.532389 IP 169.254.0.2 > 169.254.0.1: ICMP echo request, id 27785,
seq 7, length 64
16:24:43.532447 IP 169.254.0.1 > 169.254.0.2: ICMP echo reply, id 27785,
seq 7, length 64
16:24:43.838652 IP 169.254.0.1.59951 > 169.254.0.2.5342: Flags [.], seq
2555144343:2555145791, ack 2067274882, win 58, options [nop,nop,TS val
530349755 ecr 590399688], length 1448
16:24:43.838692 IP 169.254.0.1.59951 > 169.254.0.2.5342: Flags [P.], seq
1448:1838, ack 1, win 58, options [nop,nop,TS val 530349755 ecr 590399688],
length 390
10 packets captured
12 packets received by filter
0 packets dropped by kernel

```

## Configuration Files

- /etc/cumulus/acl/policy.conf

## Useful Links

- [www.perihel.at/sec/mz/mzguide.html](http://www.perihel.at/sec/mz/mzguide.html)
- [en.wikipedia.org/wiki/Ping](http://en.wikipedia.org/wiki/Ping)
- [www.tcpdump.org](http://www.tcpdump.org)
- [en.wikipedia.org/wiki/Traceroute](http://en.wikipedia.org/wiki/Traceroute)

## Caveats and Errata

- SPAN rules cannot match outgoing subinterfaces.
- ERSPAN rules must include `tt1` for versions 1.5.1 and earlier.

# SNMP Monitoring

Cumulus Linux 2.5.x utilizes the open source Net-SNMP agent `snmpd`, v5.4.3, which provides support for most of the common industry-wide MIBs, including interface counters and TCP/UDP IP stack data.



Cumulus Linux does not prevent customers from extending SNMP features. However, Cumulus Networks encourages the use of higher performance monitoring environments, rather than SNMP.

## Contents

(Click to expand)

- [Contents \(see page 454\)](#)
- [Starting the SNMP Daemon \(see page 454\)](#)
- [Configuring SNMP \(see page 456\)
  - \[Set up the Custom Cumulus MIBs \\(see page 456\\)\]\(#\)
  - \[Enable the .1.3.6.1.2.1 Range \\(see page 457\\)\]\(#\)
  - \[Enable Public Community \\(see page 457\\)\]\(#\)](#)
- [Nutanix Prism \(see page 458\)
  - \[Cumulus Configuration \\(see page 458\\)\]\(#\)
  - \[Nutanix Configuration \\(see page 459\\)\]\(#\)](#)
- [Switch Information Displayed on Nutanix Prism \(see page 462\)](#)
- [Troubleshooting \(see page 463\)](#)
- [Enabling LLDP / CDP on VMware ESXi \(Hypervisor on Nutanix\) \(see page 463\)](#)
- [Enabling LLDP / CDP on Nutanix Acropolis \(Hypervisor on Nutanix Acropolis\) \(see page 465\)](#)
- [snmpwalk the Switch from Another Linux Device \(see page 465\)](#)
- [Troubleshooting Connections without LLDP or CDP \(see page 467\)](#)
- [Generate Event Notification Traps \(see page 469\)
  - \[Enable MIB to OID Translation \\(see page 469\\)\]\(#\)
  - \[Configure Trap Events \\(see page 470\\)\]\(#\)](#)
- [Supported MIBs \(see page 473\)](#)

## Starting the SNMP Daemon

`snmpd` is disabled by default in Cumulus Linux 2.5.x. The following procedure is the recommended process to start `snmpd`, and monitor it using `jdoe`.



`jdoe` is the fork of `monit` version 5.2.5, and is included in Cumulus Linux 2.5.2 and later. For more information about upgrading from `monit` to `jdoe`, see the [jdoe upgrade knowledge base article](#).



jdoe and monit are mutually exclusive. If you would prefer to use monit, the installation process will uninstall jdoe. Cumulus Networks will not provide support for issues with monit.

To start the SNMP daemon:

1. Open `/etc/default/snmpd` to verify that `SNMPDRUN=yes`. If it does not, update the file to the correct value.
2. Create an `*.rc` configuration file in the `/etc/jdoe/jdoorc.d/` directory.



Cumulus Networks recommends using a name related to SNMP, for ease of troubleshooting. The rest of this process will use the filename `snmpd.rc`.

3. Add the following content to the `snmpd.rc` file created in step 2, under the Services banner, and save the file:

```
#####
## Services
#####
## Services
check process snmpd with pidfile /var/run/snmpd.pid
    every 6 cycles
    group networking
    start program = "/etc/init.d/snmpd start"
    stop program = "/etc/init.d/snmpd stop"
```

4. Configure `snmpd` to start automatically on boot:

```
# update-rc.d snmpd enable
```

5. Reload jdoe:

```
# sudo jdoe reload
```

6. Start the SNMP daemon, either with jdoe monitoring, or natively.

- With jdoe monitoring:

```
# sudo jdoo start snmpd
```

- Natively:

```
# sudo service snmpd start
```

Once the service is started, SNMP can be used to manage various components on the Cumulus Linux switch.

## Configuring SNMP

Cumulus Linux ships with a production usable default `snmpd.conf` file included. This section covers a few basic configuration options in `snmpd.conf`. For more information regarding further configuring this file, refer to the `snmpd.conf` man page.



The default `snmpd.conf` file does not include all supported MIBs or OIDs that can be exposed.



Customers are encouraged to at least change the default community string for v1 or v2c environments.

## Set up the Custom Cumulus MIBs



No changes are required in the `/etc/snmp/snmpd.conf` file on the switch, in order to support the custom Cumulus MIBs. The following lines are already included by default:

```
view systemonly included .1.3.6.1.4.1.40310.1
view systemonly included .1.3.6.1.4.1.40310.2
sysObjectID 1.3.6.1.4.1.40310
pass_persist .1.3.6.1.4.1.40310.1 /usr/share/snmp/resq_pp.py
pass_persist .1.3.6.1.4.1.40310.2 /usr/share/snmp/cl_drop_cntrs_pp.py
```

However, several files need to be copied to the server, in order for the custom Cumulus MIB to be recognized on the destination NMS server.

- `/usr/share/snmp/Cumulus-Snmp-MIB.txt`
- `/usr/share/snmp/Cumulus-Counters-MIB.txt`
- `/usr/share/snmp/Cumulus-Resource-Query-MIB.txt`

## ***Enable the .1.3.6.1.2.1 Range***

Some MIBs, including storage information, are not included by default in `snmpd.conf` in Cumulus Linux. This results in some default views on common network tools (like `librenms`) to return less than optimal data.

More MIBs can be included, by enabling all the .1.3.6.1.2.1 range. This simplifies the configuration file, removing concern that any required MIBs will be missed by the monitoring system.



This configuration grants access to a large number of MIBs, including all MIB2 MIBs, which could reveal more data than expected, and consume more CPU resources.

To enable the .1.3.6.1.2.1 range:

1. Open `/etc/snmp/snmpd.conf` in a text editor.
2. Replace lines 39 - 71 with the following code sample, and save the file.

```
#####
#####
#
# ACCESS CONTROL
#
#
# system
view systemonly included .1.3.6.1.2.1
# quagga ospf6
view systemonly included .1.3.6.1.3.102
# lldpd
view systemonly included .1.0.8802.1.1.2
#lmsensors
view systemonly included .1.3.6.1.4.1.2021.13.16
# Cumulus specific
view systemonly included .1.3.6.1.4.1.40310.1
view systemonly included .1.3.6.1.4.1.40310.2
```

3. Restart snmpd:

```
# sudo service snmpd start
```

## ***Enable Public Community***

Public community is disabled by default in Cumulus Linux. To enable querying by agent:

1. Open `/etc/snmp/snmpd.conf` in a text editor.

2. Add the following line to the end of the file, then save it:

```
rocommunity public default -V systemonly
```

3. Restart snmpd:

```
cumulus@switch:~$ sudo service snmpd restart
```

## Nutanix Prism

The Nutanix Prism is a graphical user interface (GUI) for managing infrastructures and virtual environments.

## Cumulus Configuration

1. SSH to the Cumulus Switch that needs to be configured, replacing [switch] below as appropriate:

```
cumulus@workbench:~$ ssh cumulus@[switch]
```

2. Confirm the switch is running Cumulus Linux 2.5.5 or newer:

```
cumulus@switch$ cat /etc/lsb-release
DISTRO_ID="Cumulus Linux"
DISTRO_RELEASE=2.5.5
DISTRO_DESCRIPTION=2.5.5-4cd66d9-201512071809-build
```

3. Open the /etc/snmp/snmpd.conf file in an editor:

4. Uncomment the following 3 lines in the /etc/snmp/snmpd.conf file, and save the file:

- bridge\_pp.py

```
pass_persist .1.3.6.1.2.1.17 /usr/share/snmp/bridge_pp.py
```

- Community

```
rocommunity public default -V systemonly
```

- Line directly below the Q-BRIDGE-MIB (.1.3.6.1.2.1.17)

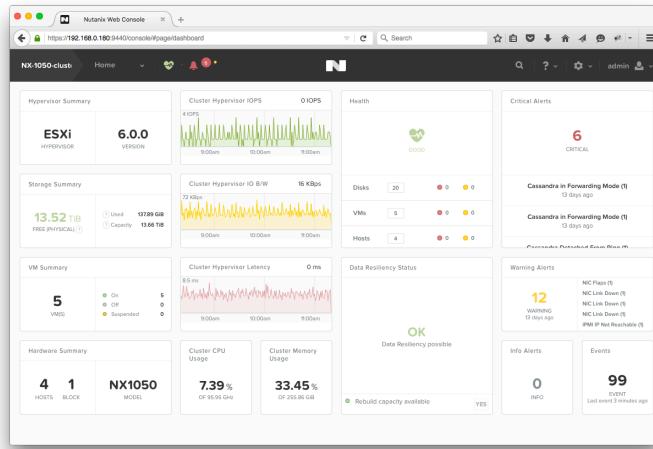
```
# BRIDGE-MIB and Q-BRIDGE-MIB tables
view      systemonly     included     .1.3.6.1.2.1.17
```

## 5. Restart snmpd:

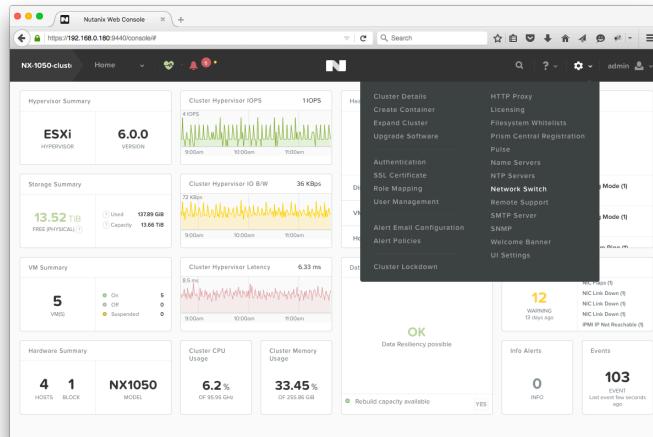
```
cumulus@switch$ sudo service snmpd restart
Restarting network management services: snmpd.
cumulus@switch$
```

## Nutanix Configuration

### 1. Log into the Nutanix Prism. Nutanix will default to the Home menu, referred to as the Dashboard:

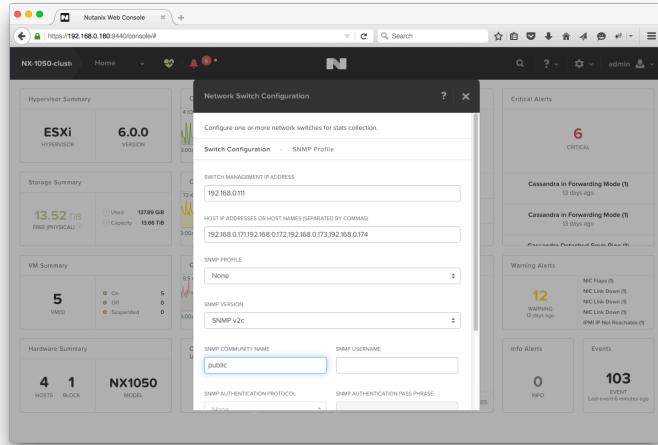


### 2. Click on the gear icon in the top right corner of the dashboard, and select NetworkSwitch:



### 3. Click the +Add Switch Configuration button in the Network Switch Configuration pop up window.

4. Fill out the **Network Switch Configuration** for the Top of Rack (ToR) switch configured for snmpd in the previous section:



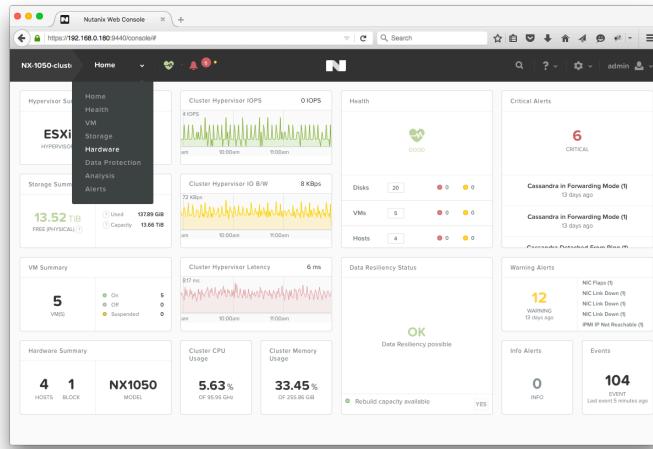
Configuration Parameter	Description	Value Used in Example
Switch Management IP Address	This can be any IP address on the box. In the screenshot above, the eth0 management IP is used.	192.168.0.111
Host IP Addresses or Host Names	IP addresses of Nutanix hosts connected to that particular ToR switch.	192.168.0.171, 192.168.0.172, 192.168.0.173, 192.168.0.174
SNMP Profile	Saved profiles, for easy configuration when hooking up to multiple switches.	None
SNMP Version	SNMP v2c or SNMP v3. Cumulus Linux has only been tested with SNMP v2c for Nutanix integration.	SNMP v2c
		public

Configuration Parameter	Description	Value Used in Example
SNMP Community Name	SNMP v2c uses communities to share MIBs. The default community for snmpd is 'public'.	



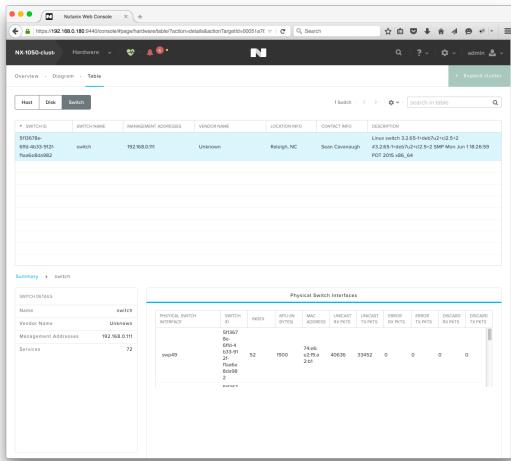
The rest of the values were not touched for this demonstration. They are usually used with SNMP v3.

- Save the configuration. The switch will now be present in the **Network Switch Configuration** menu now.
- Close the pop up window to return to the dashboard.
- Open the **Hardware** option from the **Home** dropdown menu:



- Select the **Table** button.

9. Select the **Switch** button. Configured switches are shown in the table, as indicated in the screenshot below, and can be selected in order to view interface statistics:



The screenshot shows the Nutanix Web Console interface. At the top, there's a navigation bar with tabs for Overview, Diagram, and Table. Below that, a search bar and a 'Switch' button are visible. The main area displays a table of configured switches. One row is highlighted in blue, showing details: SWITCH-ID is 5f13678e-6ffd-4b33-912f-f1aa6e8da982, SWITCH-NAME is 'switch', MANAGEMENT ADDRESS is 192.168.0.111, VENDOR NAME is Unknown, LOCATION INFO is Raleigh, NC, and DESCRIPTION is Line switch 3.2.65-fresh2/v3.5-2. The table also includes columns for INDEX, NETWORK MAC ADDRESS, UNICAST RX PKTS, UNICAST TX PKTS, ERROR RX PKTS, ERROR TX PKTS, DISCARD RX PKTS, and DISCARD TX PKTS. On the left, a sidebar titled 'Inventory' has a 'Switch' section with details: Name is 'switch', Vendor Name is 'Unknown', Management Address is '192.168.0.111', and Services is '72'. Below the table, a section titled 'Physical Switch Interfaces' lists interfaces: swp47, swp48, swp49, and swp50. Each interface has its own row with columns for SWITC... ID, INDEX, NETWORK MAC ADDRESS, UNICAST RX PKTS, UNICAST TX PKTS, ERROR RX PKTS, ERROR TX PKTS, DISCARD RX PKTS, and DISCARD TX PKTS.



The switch has been added correctly, when interfaces hooked up to the Nutanix hosts are visible.

## Switch Information Displayed on Nutanix Prism

- Physical Interface (e.g. swp1, swp2). This will only display swp interfaces connected to Nutanix hosts by default.
- Switch ID - Unique identifier that Nutanix keeps track of each port ID (see below)
- Index - interface index, in the above demonstration swp49 maps to Index 52 because there is a loopback and two ethernet interface before the swp starts.
- MTU of interface
- MAC Address of Interface
- Unicast RX Packets (Received)
- Unicast TX Packets (Transmitted)
- Error RX Packets (Received)
- Error TX Packets (Transmitted)
- Discard RX Packets (Received)
- Discard TX Packets (Transmitted)

The Nutanix appliance will use Switch IDs that can also be viewed on the Prism CLI (by SSHing to the box). To view information from the Nutanix CLI, login using the default username **nutanix**, and the password **nutanix/4u**.

```
nutanix@NTNX-14SM15270093-D-CVM:192.168.0.184:~$ ncli network list-switch
      Switch ID          : 00051a76-f711-89b6-0000-000000003bac:::
5f13678e-6ffd-4b33-912f-f1aa6e8da982
      Name          : switch
      Switch Management Address : 192.168.0.111
```

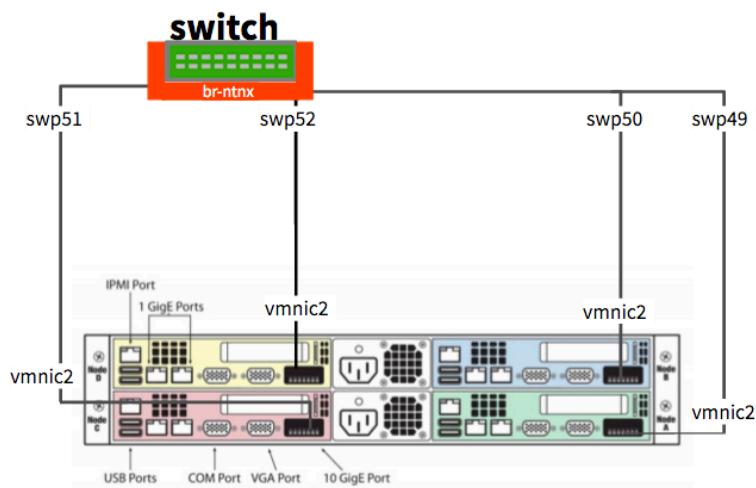
```

Description : Linux switch 3.2.65-1+deb7u2+c12.5+2 #3.
2.65-1+deb7u2+c12.5+2 SMP Mon Jun 1 18:26:59 PDT 2015 x86_64
Object ID : enterprises.40310
Contact Information : Admin <admin@company.com>
Location Information : Raleigh, NC
Services : 72
Switch Vendor Name : Unknown
Port IDs : 00051a76-f711-89b6-0000-00000003bac::5f13678e-6ffd-4b33-912f-f1aa6e8da982:52, 00051a76-f711-89b6-0000-00000003bac::5f13678e-6ffd-4b33-912f-f1aa6e8da982:53, 00051a76-f711-89b6-0000-00000003bac::5f13678e-6ffd-4b33-912f-f1aa6e8da982:54, 00051a76-f711-89b6-0000-00000003bac::5f13678e-6ffd-4b33-912f-f1aa6e8da982:55

```

## Troubleshooting

To help visualize the following diagram is provided:



Nutanix Node	Physical Port	Cumulus Linux Port
Node A (Green)	<code>vmnic2</code>	<code>swp49</code>
Node B (Blue)	<code>vmnic2</code>	<code>swp50</code>
Node C (Red)	<code>vmnic2</code>	<code>swp51</code>
Node D (Yellow)	<code>vmnic2</code>	<code>swp52</code>

## Enabling LLDP / CDP on VMware ESXi (Hypervisor on Nutanix)

1. Follow the directions on one of the following websites to enable CDP:

- a. [http://kb.vmware.com/selfservice/microsites/search.do?language=en\\_US&cmd=displayKC&externalId=1003885](http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=1003885)
- b. <http://wahlnetwork.com/2012/07/17/utilizing-cdp-and-lldp-with-vsphere-networking/>  
e.g. Switch CDP on:

```
root@NX-1050-A:~] esxcli network vswitch standard set -c both -v  
vSwitch0
```

Then confirm it is running:

```
root@NX-1050-A:~] esxcli network vswitch standard list -v vSwitch0  
vSwitch0  
  Name: vSwitch0  
  Class: etherswitch  
  Num Ports: 4082  
  Used Ports: 12  
  Configured Ports: 128  
  MTU: 1500  
  CDP Status: both  
  Beacon Enabled: false  
  Beacon Interval: 1  
  Beacon Threshold: 3  
  Beacon Required By:  
  Uplinks: vmnic3, vmnic2, vmnic1, vmnic0  
  Portgroups: VM Network, Management Network
```

The **both** means CDP is now running, and the lldp dameon on Cumulus Linux is capable of 'seeing' CDP devices.

2. After the next CDP interval, the Cumulus Linux box will pick up the interface via the lldp daemon:

```
cumulus@switch$ lldpctl show neighbor swp49  
-----  
-----  
LLDP neighbors:  
-----  
-----  
Interface:      swp49, via: CDPv2, RID: 6, Time: 0 day, 00:34:58  
  Chassis:  
    ChassisID:      local NX-1050-A  
    SysName:        NX-1050-A  
    SysDescr:       Releasebuild-2494585 running on VMware ESX
```

```

MgmtIP:      0.0.0.0
Capability:   Bridge, on
Port:
  PortID:     ifname vmnic2
  PortDescr:  vmnic2
-----
```

3. Use netshow to look at lldp information:

```

cumulus@switch$ netshow lldp
-----
To view the legend, rerun "netshow" cmd with the "--legend" option
-----
Local Port      Speed      Mode          Remote Port      Remote
Host           Summary
-----  -----  -----  -----  -----
-----  -----
eth0           1G        Mgmt        ===  swp32           swoob.vsokt.
local IP: 192.168.0.111/24(DHCP)
swp49          10G(SFP+)  Access/L2  ===  vmnic2         NX-1050-
A              Untagged: br-ntnx
swp50          10G(SFP+)  Access/L2  ===  vmnic2         NX-1050-
B              Untagged: br-ntnx
swp51          10G(SFP+)  Access/L2  ===  vmnic2         NX-1050-
C              Untagged: br-ntnx
swp52          10G(SFP+)  Access/L2  ===  vmnic2         NX-1050-
D              Untagged: br-ntnx
```

## ***Enabling LLDP / CDP on Nutanix Acropolis (Hypervisor on Nutanix Acropolis)***

Nutanix Acropolis is an alternate hypervisor that Nutanix supports. Acropolis Hypervisor uses the yum packaging system and is capable of installing normal Linux lldp daemons to operating just like Cumulus Linux. LLDP should be enabled for each interface on the host. Refer to <https://community.mellanox.com/docs/DOC-1522> for setup instructions.

## ***snmpwalk the Switch from Another Linux Device***

One of the most important ways to troubleshoot is to snmpwalk the switch from another Linux device that can reach the switch running Cumulus Linux. For this demonstration, another switch running Cumulus Linux within the network is used.

1. Open /etc/apt/sources.list in an editor.
2. Add the following line, and save the file:

```
deb http://ftp.us.debian.org/debian/ wheezy main non-free
```

3. Update the switch:

```
cumulus@switch2$ sudo apt-get update
```

4. Install the snmp and snmp-mibs-downloader packages:

```
cumulus@switch2$ sudo apt-get install snmp snmp-mibs-downloader
```

5. Verify that the "mibs :" line is commented out in /etc/snmp/snmp.conf:

```
#  
# As the snmp packages come without MIB files due to license reasons,  
loading  
# of MIBs is disabled by default. If you added the MIBs you can  
reenable  
# loading them by commenting out the following line.  
#mibs :
```

6. Perform an snmpwalk on the switch. The switch running snmpd in the demonstration is using IP address 192.168.0.111. It is possible to snmpwalk the switch from itself, following these instructions, ruling out an snmp problem vs networking problem.

```
cumulus@switch2$ snmpwalk -c public -v2c 192.168.0.111
```

## Output Examples

```
IF-MIB::ifPhysAddress.2 = STRING: 74:e6:e2:f5:a2:80  
IF-MIB::ifPhysAddress.3 = STRING: 0:e0:ec:25:b8:54  
IF-MIB::ifPhysAddress.4 = STRING: 74:e6:e2:f5:a2:81  
IF-MIB::ifPhysAddress.5 = STRING: 74:e6:e2:f5:a2:82  
IF-MIB::ifPhysAddress.6 = STRING: 74:e6:e2:f5:a2:83  
IF-MIB::ifPhysAddress.7 = STRING: 74:e6:e2:f5:a2:84  
IF-MIB::ifPhysAddress.8 = STRING: 74:e6:e2:f5:a2:85  
IF-MIB::ifPhysAddress.9 = STRING: 74:e6:e2:f5:a2:86  
IF-MIB::ifPhysAddress.10 = STRING: 74:e6:e2:f5:a2:87
```

```

IF-MIB::ifPhysAddress.11 = STRING: 74:e6:e2:f5:a2:88
IF-MIB::ifPhysAddress.12 = STRING: 74:e6:e2:f5:a2:89
IF-MIB::ifPhysAddress.13 = STRING: 74:e6:e2:f5:a2:8a
IF-MIB::ifPhysAddress.14 = STRING: 74:e6:e2:f5:a2:8b
IF-MIB::ifPhysAddress.15 = STRING: 74:e6:e2:f5:a2:8c
IF-MIB::ifPhysAddress.16 = STRING: 74:e6:e2:f5:a2:8d
IF-MIB::ifPhysAddress.17 = STRING: 74:e6:e2:f5:a2:8e
IF-MIB::ifPhysAddress.18 = STRING: 74:e6:e2:f5:a2:8f
IF-MIB::ifPhysAddress.19 = STRING: 74:e6:e2:f5:a2:90

```

Any information gathered here should verify that snmpd is running correctly on the Cumulus Linux side, reducing locations where a problem may reside.

## Troubleshooting Tips Table for snmp walks

Run snmpwalk from	If it works	If it does not work
<b>switch</b> (switch to be monitored)	snmpd is serving information correctly Problem resides somewhere else (e.g. network connectivity, Prism misconfiguration)	Is snmpd misconfigured or installed incorrectly?
<b>switch2</b> (another Cumulus Linux switch in the network)	snmpd is serving information correctly and network reachability works between <b>switch</b> and <b>switch2</b> Problems resides somewhere else (e.g. can Prism reach <b>switch</b> , Prism misconfiguration)	Network connectivity is not able to grab information? Is there an iptables rule blocking? Is the snmp walk being run correctly?
<b>Nutanix Prism CLI</b> (ssh to the cluster IP address)	snmpd is serving information correctly and network reachability works between <b>switch</b> and the <b>Nutanix Appliance</b> Problems resides somewhere else (e.g. The GUI might be misconfigured)	Is the right community name being used in the GUI? Is snmp v2c being used?

## Troubleshooting Connections without LLDP or CDP

- Find the MAC address information in the Prism GUI, located in: **Hardware -> Table -> Host -> Host NICs**
- Select a MAC address to troubleshoot (e.g. 0c:c4:7a:09:a2:43 represents vmnic0 which is tied to NX-1050-A).
- List out all the MAC addresses associated to the bridge:

```

cumulus@switch$ brctl showmacs br-ntnx
port name mac addr          vlan      is local?    ageing

```

timer					
swp9	00:02:00:00:00:06	0	no	66.94	
swp52	00:0c:29:3e:32:12	0	no	2.73	
swp49	00:0c:29:5a:f4:7f	0	no	2.73	
swp51	00:0c:29:6f:e1:e4	0	no	2.73	
swp49	00:0c:29:74:0c:ee	0	no	2.73	
swp50	00:0c:29:a9:36:91	0	no	2.73	
swp9	08:9e:01:f8:8f:0c	0	no	13.56	
swp9	08:9e:01:f8:8f:35	0	no	2.73	
swp4	0c:c4:7a:09:9e:d4	0	no	24.05	
swp1	0c:c4:7a:09:9f:8e	0	no	13.56	
swp3	0c:c4:7a:09:9f:93	0	no	13.56	
swp2	0c:c4:7a:09:9f:95	0	no	24.05	
swp52	0c:c4:7a:09:a0:c1	0	no	2.73	
swp51	0c:c4:7a:09:a2:35	0	no	2.73	
swp49	0c:c4:7a:09:a2:43	0	no	2.73	
swp9	44:38:39:00:82:04	0	no	2.73	
swp9	74:e6:e2:f5:a2:80	0	no	2.73	
swp1	74:e6:e2:f5:a2:81	0	yes	0.00	
swp2	74:e6:e2:f5:a2:82	0	yes	0.00	
swp3	74:e6:e2:f5:a2:83	0	yes	0.00	
swp4	74:e6:e2:f5:a2:84	0	yes	0.00	
swp5	74:e6:e2:f5:a2:85	0	yes	0.00	
swp6	74:e6:e2:f5:a2:86	0	yes	0.00	
swp7	74:e6:e2:f5:a2:87	0	yes	0.00	
swp8	74:e6:e2:f5:a2:88	0	yes	0.00	
swp9	74:e6:e2:f5:a2:89	0	yes	0.00	
swp10	74:e6:e2:f5:a2:8a	0	yes	0.00	
swp49	74:e6:e2:f5:a2:b1	0	yes	0.00	
swp50	74:e6:e2:f5:a2:b2	0	yes	0.00	
swp51	74:e6:e2:f5:a2:b3	0	yes	0.00	
swp52	74:e6:e2:f5:a2:b4	0	yes	0.00	
swp9	8e:0f:73:1b:f8:24	0	no	2.73	
swp9	c8:1f:66:ba:60:cf	0	no	66.94	

Alternatively, you can use grep:p

```
cumulus@switch$ brctl showmacs br-ntnx | grep 0c:c4:7a:09:a2:43
swp49      0c:c4:7a:09:a2:43          0        no           4.58
cumulus@switch$
```

vmnic1 is now hooked up to swp49. This matches what is seen in lldp:

```
cumulus@switch$ lldpctl show neighbor swp49
-----
-----
LLDP neighbors:
-----
-----
Interface:      swp49, via: CDPv2, RID: 6, Time: 0 day, 01:11:12
Chassis:
    ChassisID:      local NX-1050-A
    SysName:        NX-1050-A
    SysDescr:       Releasebuild-2494585 running on VMware ESX
    MgmtIP:         0.0.0.0
    Capability:    Bridge, on
Port:
    PortID:         ifname vmnic2
    PortDescr:      vmnic2
-----
-----
cumulus@switch$
```

## **Generate Event Notification Traps**

The Net-SNMP agent provides a method to generate SNMP trap events, via the Distributed Management (DisMan) Event MIB, for various system events, including linkup/down, exceeding the temperature sensor threshold, CPU load, or memory threshold, or other SNMP MIBs.

## **Enable MIB to OID Translation**

MIB names can be used instead of OIDs, by installing the `snmp-mibs-downloader`, to download SNMP MIBs to the switch prior to enabling traps. This greatly improves the readability of the `snmpd.conf` file.

1. Open `/etc/apt/sources.list` in a text editor.
2. Add the `non-free` repository, and save the file:

```
cumulus@switch:~$ deb http://ftp.us.debian.org/debian/ wheezy main non-free
```

3. Update the switch:

```
cumulus@switch:~$ apt-get update
```

4. Install the `snmp-mibs-downloader`:

```
apt-get snmp-mibs-downloader
```

5. Open the /etc/snmp/snmp.conf file to verify that the mibs : line is commented out:

```
#  
# As the snmp packages come without MIB files due to license reasons,  
loading  
# of MIBs is disabled by default. If you added the MIBs you can  
reenable  
# loading them by commenting out the following line.  
#mibs :
```

6. Open the /etc/default/snmpd file to verify that the export MIBS= line is commented out:

```
# This file controls the activity of snmpd and snmptrapd  
  
# Don't load any MIBs by default.  
# You might comment this lines once you have the MIBs Downloaded.  
#export MIBS=
```

7. Once the configuration has been confirmed, remove or comment out the non-free repository in /etc/apt/sources.list.

```
#deb http://ftp.us.debian.org/debian/ wheezy main non-free
```

## Configure Trap Events

The following configurations should be made in /etc/snmp/snmp.conf, in order to enable specific types of traps. Once configured, restart the snmpd service to apply the changes.

## Define Access Credentials

An SNMPv3 username is required to authorize the DisMan service. The example code below uses cumulusUser as the username.

```
createUser cumulusUser  
iquerySecName cumulusUser  
rouser cumulusUser
```

## Defining Trap Receivers

The example code below creates a trap receiver that is capable of receiving SNMPv2 traps.

```
trap2sink 192.168.1.1 public
```



Although the traps are sent to an SNMPV2 receiver, the SNMPV3 user is still required.



It is possible to define multiple trap receivers, and to use the domain name instead of IP address in the `trap2sink` directive.

## Configure LinkUp/Down Notifications

The `linkUpDownNotifications` directive is used to configure linkup/down notifications when the operational status of the link changes.

```
linkUpDownNotifications yes
```



The default frequency for checking link up/down is 60 seconds. The default frequency can be changed using the `monitor` directive directly instead of the `linkUpDownNotifications` directive. See `man snmpd.conf` for details.

## Configure Temperature Notifications

Temperature sensor information for each available sensor is maintained in the the `ImSensors` MIB. Each platform may contain a different number of temperature sensors. The example below generates a trap event when any temperature sensors exceeds a threshold of 68 degrees (centigrade). It monitors each `1mTempSensorsValue`. When the threshold value is checked and exceeds the `1mTempSensorsValue`, a trap is generated. The `-o lmTempSensorsDevice` option is used to instruct SNMP to also include the `ImTempSensorsDevice` MIB in the generated trap. The default frequency for the `monitor` directive is 600 seconds. The default frequency may be changed using the `-r` option.:.

```
monitor lmTemSensor -o lmTempSensorsDevice lmTempSensorsValue > 68000
```

Alternatively, temperature sensors may be monitored individually. To monitor the sensors individually, first use the `sensors` command to determine which sensors are available to be monitored on the platform.

```
#sensors

CY8C3245-i2c-4-2e
Adapter: i2c-0-mux (chan_id 2)
fan5: 7006 RPM (min = 2500 RPM, max = 23000 RPM)
fan6: 6955 RPM (min = 2500 RPM, max = 23000 RPM)
fan7: 6799 RPM (min = 2500 RPM, max = 23000 RPM)
fan8: 6750 RPM (min = 2500 RPM, max = 23000 RPM)
temp1: +34.0 C (high = +68.0 C)
temp2: +28.0 C (high = +68.0 C)
temp3: +33.0 C (high = +68.0 C)
temp4: +31.0 C (high = +68.0 C)
temp5: +23.0 C (high = +68.0 C)
```

Configure a `monitor` command for the specific sensor using the `-I` option. The `-I` option indicates that the monitored expression is applied to a single instance. In this example, there are five temperature sensors available. The following monitor directive can be used to monitor only temperature sensor three at five minute intervals.

```
monitor -I -r 300 lmTemSensor3 -o lmTempSensorsDevice.3 lmTempSensorsValue.
3 > 68000
```

## Configure Free Memory Notifications

You can monitor free memory using the following directives. The example below generates a trap when free memory drops below 1,000,000KB. The free memory trap also includes the amount of total real memory:

```
monitor MemFreeTotal -o memTotalReal memTotalFree < 1000000
```

## Configure Processor Load Notifications

To monitor CPU load for 1, 5 or 15 minute intervals, use the `load` directive in conjunction with the `monitor` directive. The following example will generate a trap when the 1 minute interval reaches 12%, the 5 minute interval reaches 10% or the 15 minute interval reaches 5%.

```
load 12 10 5
monitor -r 60 -o laNames -o laErrMessage "laTable" laErrorFlag !=0
```

## Configure Disk Utilization Notifications

To monitor disk utilization for all disks, use the `includeAllDisks` directive in conjunction with the `monitor` directive. The example code below generates a trap when a disk is 99% full:

```
includeAllDisks 1%
monitor -r 60 -o dskPath -o DiskErrMsg "dskTable" diskErrorFlag !=0
```

## Configure Authentication Notifications

To generate authentication failure traps, use the `authtrapenable` directive:

```
authtrapenable 1
```

## Supported MIBs

Below are the MIBs supported by Cumulus Linux 2.5.4, as well as suggested uses for them. The overall Cumulus Linux MIB is defined in `/usr/share/snmp/Cumulus-Snmp-MIB.txt`.

MIB Name	Suggested Uses
CUMULUS-COUNTERS-MIB	Discard counters: Cumulus Linux also includes its own counters MIB, defined in <code>/usr/share/snmp/Cumulus-Counters-MIB.txt</code> . It has the OID <code>1.3.6.1.4.1.40310.2</code>
CUMULUS-RESOURCE-QUERY-MIB	Cumulus Linux includes its own resource utilization MIB, which is similar to using <a href="#">cl-resource-query (see page 403)</a> . It monitors L3 entries by host, route, nexthops, ECMP groups and L2 MAC/BPDUs. The MIB is defined in <code>/usr/share/snmp/Cumulus-Resource-Query-MIB.txt</code> , and has the OID <code>.1.3.6.1.4.1.40310.1</code> .
DISMAN-EVENT	Trap monitoring
HOST-RESOURCES	Users, storage, interfaces, process info, run parameters
IF-MIB	Interface description, type, MTU, speed, MAC, admin, operation status, counters
IP (includes ICMP)	IPv4, IPv4 addresses, counters, netmasks
IPv6	IPv6 counters

MIB Name	Suggested Uses
IP-FORWARD	IP routing table
LLDP	L2 neighbor info from llldpd (note, you need to <a href="#">enable the SNMP subagent (see page 144)</a> in LLDP)
LM-SENSORS MIB	Fan speed, temperature sensor values, voltages
NET-SNMP-AGENT	Agent timers, user, group config
NET-SNMP-EXTEND	Agent timers, user, group config
NET-SNMP-EXTEND-MIB	(See also <a href="#">this knowledge base article</a> on extending NET-SNMP in Cumulus Linux to include data from power supplies, fans and temperature sensors.)
NET-SNMP-VACM	Agent timers, user, group config
NOTIFICATION-LOG	Local logging
SNMP-FRAMEWORK	Users, access
SNMP-MPD	Users, access
SNMP-TARGET	
SNMP-USER-BASED-SM	Users, access
SNMP-VIEW-BASED-ACM	Users, access
SNMPv2	SNMP counters (For information on exposing CPU and memory information via SNMP, see <a href="#">this knowledge base article</a> .)
TCP	TCP related information
UCD-SNMP	System memory, load, CPU, disk IO
UDP	UDP related information



The Quagga and Zebra routes MIB is disabled in Cumulus Linux.

# Index

4

40G ports 115  
logical limitations 115

8

802.1p 118  
    class of service 118  
802.3ad link aggregation 209

A

ABRs 334  
    area border routers 334  
access control lists 76  
access ports 176  
ACL policy files 80  
ACL rules 121  
ACLs 76  
active-active mode 217, 293  
    VRR 217  
    VXLAN 293  
active image slot 17  
active listener ports 433  
active-standby mode 217  
    VRR 217  
Algorithm Longest Prefix Match 313  
    routing 313  
ALPM mode 313  
    routing 313  
alternate image slot 17, 21  
    accessing 21  
AOC cables 11  
apt-get 46  
area border routers 334  
    ABRs 334  
arp cache 443  
ARP requests 218  
    VRR 218

AS\_PATH setting 358  
    BGP 358  
ASN 347  
    autonomous system number 347  
auto-negotiation 111  
autonomous system number 347  
    BGP 347  
autoprovision command 60  
autoprovisioning 52

## B

bestpath 358  
    BGP 358  
BFD 150, 152  
    Bidirectional Forwarding Detection 150  
    echo function 152  
BGP 345, 348  
    Border Gateway Protocol 345  
    ECMP 348  
BGP peering relationships 356, 356  
    external 356  
    internal 356  
bonds 158, 209  
    LACP Bypass 209  
boot recovery 392  
bpdufilter 138  
    and STP 138  
BPDU guard 135  
    and STP 135  
brctl 13, 125, 163, 164, 304, 304  
    and STP 125  
    IGMP snooping 304  
    MLD snooping 304  
bridge assurance 135  
    and STP 135  
bridges 162, 162, 163, 163, 164, 165, 165, 169, 171, 176, 176, 176, 182, 191  
    access ports 176  
    adding interfaces 163, 164  
    adding IP addresses 169  
    IGMP snooping 191  
    MAC addresses 165  
    MTU 165  
    physical interfaces 163

trunk ports 176  
untagged frames 171  
VLAN-aware 162, 182

## C

cable connectivity 11  
cabling 145  
    Prescriptive Topology Manager 145  
cl-acltool 76, 121, 445  
CLAG 217  
    and VRR 217  
clagctl 202  
class of service 118  
cl-bgp 328  
cl-cfg 89, 415  
cl-ecmpcalc 373  
cl-img-clear-overlay 21, 22  
cl-img-pkg 24  
cl-img-select 22, 22, 23  
cl-license 11  
cl-netstat 401  
cl-ospf 328, 335  
cl-ospf6 328, 343  
Clos topology 316  
cl-ra 328  
cl-rctl 328  
cl-resource-query 90, 403  
cl-route-check 341  
cl-support 388  
convergence 315  
    routing 315  
Cumulus Linux 7, 8, 21, 22, 22, 25, 191, 287  
    installing 7, 25  
    reprovisioning 22  
    reserved VLAN ranges 191  
    reverting 21  
    uninstalling 22  
    upgrading 8  
    VXLAN 287  
cumulus user 66

## D

DAC cables 11  
daemons 433  
datapath 118  
datapath.conf 118  
date 63  
    setting 63  
deb 51  
debugging 386  
decode-syseeprom 405  
differentiated services code point 118  
dmidecode 406  
dpkg 49  
dpkg-reconfigure 62  
DSCP 118  
    differentiated services code point 118  
DSCP marking 121  
dual-connected hosts 194  
duplex interfaces 110  
dynamic routing 154, 318  
    and PTM 154  
    quagga 318

## E

eBGP 347  
    external BGP 347  
ebtables 76, 79  
    memory spaces 79  
echo function 152, 152  
    BFD 152  
    PTM 152  
ECMP 318, 340, 348, 1  
    BGP 348  
    equal cost multi-pathing 318  
    monitoring 1  
    OSPF 340  
ECMP hashing 373, 376  
    resilient hashing 376  
EGP 319  
    Exterior Gateway Protocol 319  
equal cost multipath 373  
    ECMP hashing 373

equal cost multi-pathing 318  
  ECMP 318  
ERSPAN 446  
  network troubleshooting 446  
Ethernet management port 9  
ethtool 117, 399  
  switch ports 117  
external BGP 347  
  eBGP 347

## F

fast convergence 355  
  BGP 355  
fast leave 307  
  IGMP/MLD snooping 307  
First Hop Redundancy Protocol 217  
  VRR 217

## G

glob 105  
Graphviz 145

## H

hardware 404  
  monitoring 404  
hardware compatibility list 7  
hash distribution 161  
HCL 7  
head end replication 255  
  LNV 255  
high availability 191, 317  
host entries 403  
  monitoring 403  
Host HA 191  
hostname 9  
hsflowd 410  
hwclock 64

|

iBGP [347](#)  
    internal BGP [347](#)

ifdown [96](#)

ifplugged [218](#)  
    VRR [218](#)

ifquery [100, 437](#)

ifup [96](#)

ifupdown [95](#)

ifupdown2 [104, 174, 436, 436, 436](#)  
    excluding interfaces [436](#)  
    logging [436](#)  
    purging IP addresses [104](#)  
    troubleshooting [436](#)  
    VLAN tagging [174](#)

IGMP snooping [191, 205, 302](#)  
    MLAG [205](#)  
    VLAN-aware bridges [191](#)

IGP [319](#)  
    Interior Gateway Protocol [319](#)

image contents [24](#)

image slots [17, 18, 19, 19](#)  
    PowerPC [18](#)  
    resizing [19](#)  
    x86 [19](#)

installing [7](#)  
    Cumulus Linux [7](#)

interface counters [401](#)

interface dependencies [99](#)

interfaces [108, 116](#)  
    statistics [116](#)

internal BGP [347](#)  
    iBGP [347](#)

ip6tables [76](#)

IP addresses [104](#)  
    purging [104](#)

iproute2 [441](#)  
    failures [441](#)

iptables [76](#)

IPv4 routes [348](#)  
    BGP [348](#)

IPv6 routes [348](#)  
    BGP [348](#)

J

jdoo 206

L

LACP 158, 192  
    MLAG 192  
LACP Bypass 209  
layer 3 access ports 13  
    configuring 13  
LDAP 74  
leaf-spine topology 316  
license 10  
    installing 10  
lightweight network virtualization 253, 255, 255, 280  
    head end replication 255  
    service node replication 255  
link aggregation 158  
Link Layer Discovery Protocol 139  
link-local IPv6 addresses 361  
    BGP 361  
link pause 122  
    datapath 122  
link-state advertisement 333  
link state monitoring 218  
    VRR 218  
LLDP 139, 144  
    SNMP 144  
lldpcli 140  
lldpd 139, 147  
LNV 253, 253, 255, 255, 280, 280  
    head end replication 255  
    service node replication 255  
    VXLAN 253, 280  
load balancing 318  
logging 389, 436, 436  
    ifupdown2 436  
    networking service 436  
logging neighbor state changes 361  
    BGP 361  
logical switch 191  
longest prefix match 1  
    routing 1

loopback interface [14](#)  
    configuring [14](#)  
LSA [333](#)  
    link-state advertisement [333](#)  
LSDB [333](#)  
    link-state database [333](#)  
lshw [406](#)

## M

MAC entries [403](#)  
    monitoring [403](#)  
Mako templates [106, 438](#)  
    debugging [438](#)  
mangle table [122](#)  
    ACL rules [122](#)  
memory spaces [79](#)  
    ebtables [79](#)  
MLAG [191, 202, 202, 203, 205, 207, 207, 300, 300](#)  
    backup link [203](#)  
    IGMP snooping [205](#)  
    MTU [207](#)  
    peer link states [202, 300](#)  
    PROTO\_DOWN state [300](#)  
    protodown state [202](#)  
    STP [207](#)  
MLD snooping [302](#)  
monitoring [62, 386, 399, 403, 409, 410, 412, 454](#)  
    hardware watchdog [409](#)  
    Net-SNMP [454](#)  
    network traffic [410](#)  
mount points [19](#)  
mstpcctl [125, 178](#)  
MTU [111, 165, 207, 441](#)  
    bridges [165](#)  
    failures [441](#)  
    MLAG [207](#)  
multi-Chassis Link Aggregation [191](#)  
    MLAG [191](#)  
multiple bridges [166](#)  
mz [444](#)  
    traffic generator [444](#)

## N

name switch service 73  
Netfilter 76  
Net-SNMP 454  
networking service 436  
  logging 436  
network interfaces 95, 108  
  ifupdown 95  
network traffic 410  
  monitoring 410  
network troubleshooting 452  
  tcpdump 452  
network virtualization 220, 221, 287  
  VMware NSX 221  
no-as-set 358  
  BGP 358  
nonatomic updates 79  
  switchd 79  
non-blocking networks 317  
NSS 73  
  name switch service 73  
NTP 64  
  time 64  
ntpd 64

## O

ONIE 7, 23  
  rescue mode 23  
Open Network Install Environment 7  
Open Shortest Path First Protocol 332, 343  
  OSPFv2 332  
  OSPFv3 343  
open source contributions 6  
OSPF 337, 339, 340, 340, 342  
  ECMP 340  
  reconvergence 340  
  summary LSA 337  
  supported RFCs 342  
  unnumbered interfaces 339  
ospf6d.conf 344  
OSPFv2 332  
OSPFv3 343, 345, 345

supported RFCs 345  
unnumbered interfaces 345  
overlayfs file system 18  
over-subscribed networks 317

## P

packages 46  
managing 46  
packet buffering 118  
  datapath 118  
packet filtering 77  
packet queueing 118  
  datapath 118  
packet scheduling 118  
  datapath 118  
PAM 73  
  pluggable authentication modules 73  
parent interfaces 102  
password 66  
  default 66  
passwordless access 66  
passwords 9  
peer-groups 357  
  BGP 357  
Per VLAN Spanning Tree 125  
  PVST 125  
ping 442  
pluggable authentication modules 73  
policy.conf 82  
port lists 105  
port speeds 110  
Prescriptive Topology Manager 145  
primary image slot 17  
priority groups 118  
  datapath 118  
privileged commands 68  
PROTO\_DOWN state 300  
  MLAG 300  
protocol tuning 315, 363  
  BGP 363  
  routing 315  
protodown state 202  
  MLAG 202

PTM 145, 152  
echo function 152  
Prescriptive Topology Manager 145  
ptmctl 155  
ptmd 145  
PTM scripts 148  
PVRST 125  
    Rapid PVST 125  
PVST 125  
    Per VLAN Spanning Tree 125

## Q

QSFP 402  
Quagga 154, 154, 311, 318, 320  
    and PTM 154, 154  
    configuring 320  
    dynamic routing 318  
    static routing 311  
quality of service 123  
querier 306  
    IGMP/MLD snooping 306

## R

Rapid PVST 125  
    PVRST 125  
read-only mode 362  
    BGP 362  
recommended configuration 38  
reconvergence 340  
    OSPF 340  
remote access 65  
repositories 50  
    other packages 50  
rescue mode 23  
reserved VLAN ranges 191  
resilient hashing 376  
restart 90  
    switchd 90  
root user 9, 66  
route advertisements 347  
    BGP 347

route entries 314, 314

ALPM 314

limitations 314

route maps 363

BGP 363

route reflectors 348

BGP 348

routes 403

monitoring 403

routing protocols 314

RSTP 125

## S

sensors command 406

serial console management 9

service node replication 255

  LNV 255

services 433

sFlow 410

sFlow visualization tools 412

SFP 117, 402

  switch ports 117

single user mode 392

smonctl 408

smond 408

snmpd 454

sources.list 50

SPAN 446

  network troubleshooting 446

spanning tree parameters 127

Spanning Tree Protocol 124, 183

  STP 124

  VLAN-aware bridges 183

SSH 65

SSH keys 65

static routing 309, 311

  with ip route 309

  with Quagga 311

STP 124, 135, 207

  and bridge assurance 135

  MLAG 207

  Spanning Tree Protocol 124

stub areas 338

- OSPF 338
- sudo 67, 67
- sudoers 67, 68
  - examples 68
- summary LSA 337
  - OSPF 337
- SVI 196
  - switched virtual interface 196
- switchd 79, 87, 88, 90, 415
  - configuring 87
  - counters 415
  - file system 88
  - nonatomic updates 79
  - restarting 90
- switched virtual interface 196
  - SVI 196
- switch ports 12, 115
  - configuring 12
  - logical limitations 115
- syslog 389
- system management 386

## T

- tcpdump 452
  - network troubleshooting 452
- templates 106
- time 63
  - setting 63
- time zone 10, 62
- topology 145, 316
  - data center 145
- traceroute 443
- traffic.conf 118
- traffic distribution 161
- traffic generator 444
  - mz 444
- traffic marking 121
  - datapath 121
- troubleshooting 386, 392, 452
  - single user mode 392
  - tcpdump 452
- trunk ports 171, 176
- tzdata 62

## U

U-Boot [7](#), [386](#)  
unnumbered interfaces [339](#), [345](#)  
    OSPF [339](#)  
    OSPFv3 [345](#)  
untagged frames [171](#)  
    bridges [171](#)  
upgrading [8](#)  
    Cumulus Linux [8](#)  
user accounts [67](#), [67](#)  
    cumulus [67](#)  
    root [67](#)  
user authentication [73](#)  
user commands [104](#)  
    interfaces [104](#)

## V

virtual device counters [412](#), [415](#), [416](#)  
    monitoring [412](#)  
    poll interval [415](#)  
    VLAN statistics [416](#)  
virtual router redundancy [215](#)  
visudo [67](#)  
VLAN [196](#), [412](#)  
    statistics [412](#)  
    switched virtual interface [196](#)  
VLAN-aware bridges [162](#), [182](#), [183](#), [183](#), [191](#)  
    configuring [183](#)  
    IGMP snooping [191](#)  
    Spanning Tree Protocol [183](#)  
VLAN tagging [174](#), [174](#), [176](#)  
    advanced example [176](#)  
    basic example [174](#)  
VLAN translation [181](#)  
VRR [215](#)  
    virtual router redundancy [215](#)  
VTEP [220](#), [222](#)  
vtys [323](#)  
    quagga CLI [323](#)  
VXLAN [220](#), [222](#), [253](#), [280](#), [287](#), [293](#), [412](#)  
    active-active mode [293](#)

LNV 253, 280  
no controller 287  
statistics 412  
VMware NSX 222

## W

watchdog 409  
monitoring 409

## Z

zebra 319  
routing 319  
zero touch provisioning 52, 55  
USB 55  
ZTP 52