Open vSwitch

Release 3.7.90

The Open vSwitch Development Community

Jun 09, 2026

CONTENTS

CHAPTER

ONE

PROJECT

1.1 Community

- *Mailing Lists*
- *Reporting Bugs*
- *Patchwork*
- *Release Process*
- *Security Process*
- *Authors*

1.2 Contributing

- *Submitting Patches*
- *Backporting patches*
- *Inclusive Language*
- *Coding Style*

1.3 Maintaining

- *The Linux Foundation Open vSwitch Project Charter*
- *Committers*
- *Expectations for Developers with Open vSwitch Repo Access*
- *OVS Committer Grant/Revocation Policy*
- *Emeritus Status for OVS Committers*

1.4 Documentation

- *Getting Started*
- *Tutorials*
- *How-to Guides*
- *Deep Dive*
- *Reference Guide*

- *Open vSwitch Internals*
- *Open vSwitch Documentation*
- *FAQ*
- Looking for specific information?
  - *Full Table of Contents*
  - Index

1.5 Getting Help

- Reach out to us *here*.

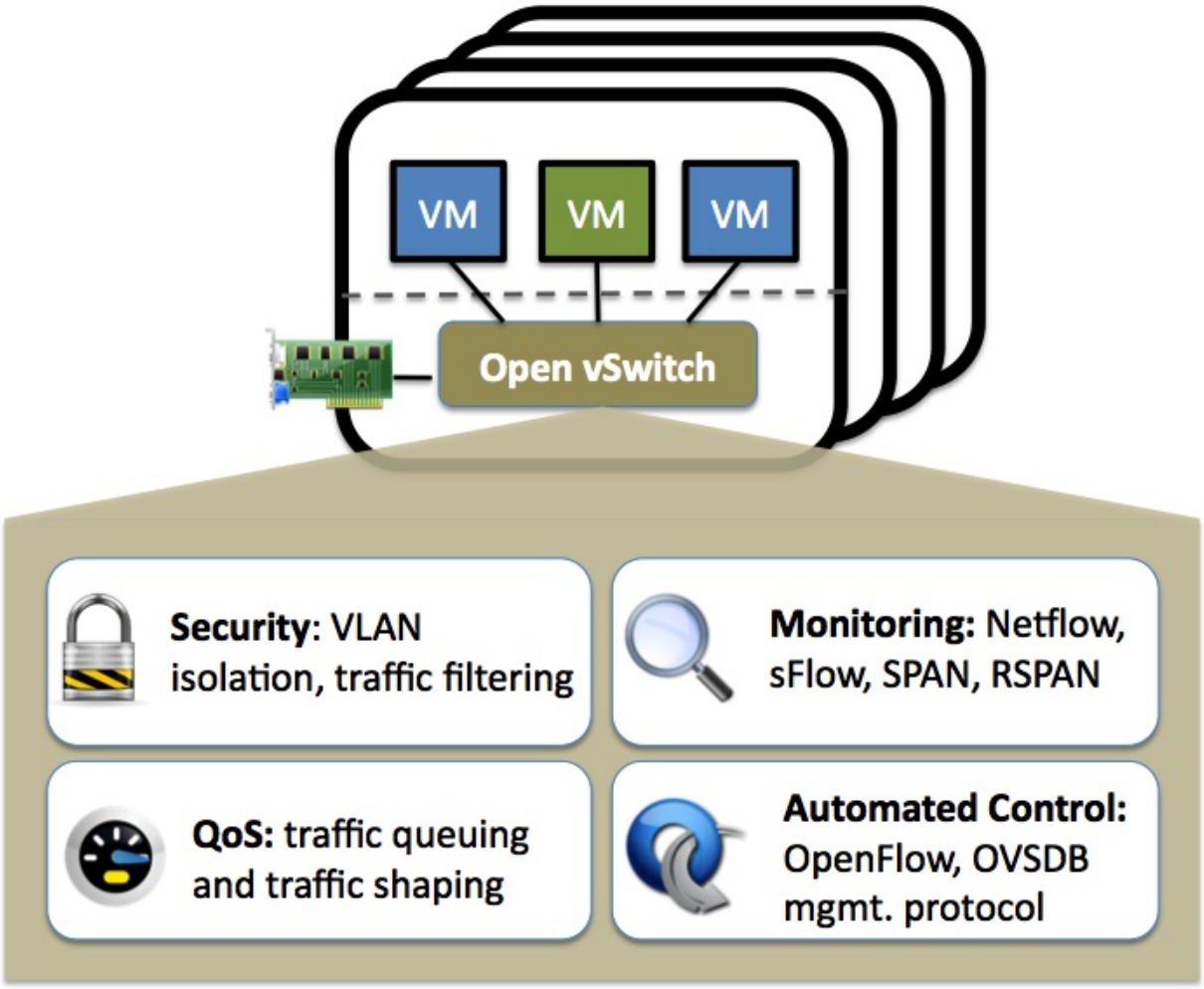
CHAPTER

TWO

GETTING STARTED

How to get started with Open vSwitch.

2.1 What Is Open vSwitch?



2.1.1 Overview

Open vSwitch is a multilayer software switch licensed under the open source Apache 2 license. Our goal is to implement a production quality switch platform that supports standard management interfaces and opens the forwarding functions to programmatic extension and control.

Open vSwitch is well suited to function as a virtual switch in VM environments. In addition to exposing standard control and visibility interfaces to the virtual networking layer, it was designed to support distribution across multiple physical servers. Open vSwitch supports multiple Linux-based virtualization technologies including KVM, and VirtualBox.

The bulk of the code is written in platform-independent C and is easily ported to other environments. The current release of Open vSwitch supports the following features:

- Standard 802.1Q VLAN model with trunk and access ports
- NIC bonding with or without LACP on upstream switch
- NetFlow, sFlow(R), and mirroring for increased visibility
- QoS (Quality of Service) configuration, plus policing
- Geneve, GRE, VXLAN, ERSPAN, GTP-U, SRv6, and Bareudp tunneling
- 802.1ag connectivity fault management
- OpenFlow 1.0 plus numerous extensions
- Transactional configuration database with C and Python bindings
- High-performance forwarding using a Linux kernel module

Open vSwitch can also operate entirely in userspace without assistance from a kernel module. This userspace implementation should be easier to port than the kernel-based switch. OVS in userspace can access Linux or DPDK devices. Note Open vSwitch with userspace datapath and non DPDK devices is considered experimental and comes with a cost in performance.

2.1.2 What's here?

The main components of this distribution are:

- ovs-vswitchd, a daemon that implements the switch, along with a companion Linux kernel module for flow-based switching.
- ovsdb-server, a lightweight database server that ovs-vswitchd queries to obtain its configuration.
- ovs-dpctl, a tool for configuring the switch kernel module.
- Scripts and specs for building RPMs for Red Hat Enterprise Linux and deb packages for Ubuntu/Debian.
- ovs-vsctl, a utility for querying and updating the configuration of ovs-vswitchd.
- ovs-appctl, a utility that sends commands to running Open vSwitch daemons.

Open vSwitch also provides some tools:

- ovs-ofctl, a utility for querying and controlling OpenFlow switches and controllers.
- ovs-pki, a utility for creating and managing the public-key infrastructure for OpenFlow switches.
- ovs-testcontroller, a simple OpenFlow controller that may be useful for testing (though not for production).
- A patch to tcpdump that enables it to parse OpenFlow messages.

2.2 Why Open vSwitch?

Hypervisors need the ability to bridge traffic between VMs and with the outside world. On Linux-based hypervisors, this used to mean using the built-in L2 switch (the Linux bridge), which is fast and reliable. So, it is reasonable to ask why Open vSwitch is used.

The answer is that Open vSwitch is targeted at multi-server virtualization deployments, a landscape for which the previous stack is not well suited. These environments are often characterized by highly dynamic end-points, the maintenance of logical abstractions, and (sometimes) integration with or offloading to special purpose switching hardware.

The following characteristics and design considerations help Open vSwitch cope with the above requirements.

2.2.1 The mobility of state

All network state associated with a network entity (say a virtual machine) should be easily identifiable and migratable between different hosts. This may include traditional "soft state" (such as an entry in an L2 learning table), L3 forwarding state, policy routing state, ACLs, QoS policy, monitoring configuration (e.g. NetFlow, IPFIX, sFlow), etc.

Open vSwitch has support for both configuring and migrating both slow (configuration) and fast network state between instances. For example, if a VM migrates between end-hosts, it is possible to not only migrate associated configuration (SPAN rules, ACLs, QoS) but any live network state (including, for example, existing state which may be difficult to reconstruct). Further, Open vSwitch state is typed and backed by a real data-model allowing for the development of structured automation systems.

2.2.2 Responding to network dynamics

Virtual environments are often characterized by high-rates of change. VMs coming and going, VMs moving backwards and forwards in time, changes to the logical network environments, and so forth.

Open vSwitch supports a number of features that allow a network control system to respond and adapt as the environment changes. This includes simple accounting and visibility support such as NetFlow, IPFIX, and sFlow. But perhaps more useful, Open vSwitch supports a network state database (OVSDB) that supports remote triggers. Therefore, a piece of orchestration software can "watch" various aspects of the network and respond if/when they change. This is used heavily today, for example, to respond to and track VM migrations.

Open vSwitch also supports OpenFlow as a method of exporting remote access to control traffic. There are a number of uses for this including global network discovery through inspection of discovery or link-state traffic (e.g. LLDP, CDP, OSPF, etc.).

2.2.3 Maintenance of logical tags

Distributed virtual switches (such as VMware vDS and Cisco's Nexus 1000V) often maintain logical context within the network through appending or manipulating tags in network packets. This can be used to uniquely identify a VM (in a manner resistant to hardware spoofing), or to hold some other context that is only relevant in the logical domain. Much of the problem of building a distributed virtual switch is to efficiently and correctly manage these tags.

Open vSwitch includes multiple methods for specifying and maintaining tagging rules, all of which are accessible to a remote process for orchestration. Further, in many cases these tagging rules are stored in an optimized form so they don't have to be coupled with a heavyweight network device. This allows, for example, thousands of tagging or address remapping rules to be configured, changed, and migrated.

In a similar vein, Open vSwitch supports a GRE implementation that can handle thousands of simultaneous GRE tunnels and supports remote configuration for tunnel creation, configuration, and tear-down. This, for example, can be used to connect private VM networks in different data centers.

2.2.4 Hardware integration

Open vSwitch's forwarding path (the in-kernel datapath) is designed to be amenable to "offloading" packet processing to hardware chipsets, whether housed in a classic hardware switch chassis or in an end-host NIC. This allows for the Open vSwitch control path to be able to both control a pure software implementation or a hardware switch.

There are many ongoing efforts to port Open vSwitch to hardware chipsets. These include multiple merchant silicon chipsets (Broadcom and Marvell), as well as a number of vendor-specific platforms. The "Porting" section in the documentation discusses how one would go about making such a port.

The advantage of hardware integration is not only performance within virtualized environments. If physical switches also expose the Open vSwitch control abstractions, both bare-metal and virtualized hosting environments can be managed using the same mechanism for automated network control.

2.2.5 Summary

In many ways, Open vSwitch targets a different point in the design space than previous hypervisor networking stacks, focusing on the need for automated and dynamic network control in large-scale Linux-based virtualization environments.

The goal with Open vSwitch is to keep the in-kernel code as small as possible (as is necessary for performance) and to reuse existing subsystems when applicable (for example Open vSwitch uses the existing QoS stack). As of Linux 3.3, Open vSwitch is included as a part of the kernel and packaging for the userspace utilities are available on most popular distributions.

2.3 Installing Open vSwitch

A collection of guides detailing how to install Open vSwitch in a variety of different environments and using different configurations.

2.3.1 Installation from Source

Open vSwitch on Linux, FreeBSD and NetBSD

This document describes how to build and install Open vSwitch on a generic Linux, FreeBSD, or NetBSD host. For specifics around installation on a specific platform, refer to one of the other installation guides listed in *Installing Open vSwitch*.

Obtaining Open vSwitch Sources

The canonical location for Open vSwitch source code is its Git repository, which you can clone into a directory named "ovs" with:

```
$ git clone https://github.com/openvswitch/ovs.git
```

Cloning the repository leaves the "main" branch initially checked out. This is the right branch for general development. If, on the other hand, if you want to build a particular released version, you can check it out by running a command such as the following from the "ovs" directory:

```
$ git checkout v2.7.0
```

The repository also has a branch for each release series. For example, to obtain the latest fixes in the Open vSwitch 2.7.x release series, which might include bug fixes that have not yet been in any released version, you can check it out from the "ovs" directory with:

```
$ git checkout origin/branch-2.7
```

If you do not want to use Git, you can also obtain tarballs for Open vSwitch release versions via http://openvswitch.org/download/, or download a ZIP file for any snapshot from the web interface at https://github.com/openvswitch/ovs.

Build Requirements

To compile the userspace programs in the Open vSwitch distribution, you will need the following software:

- GNU make
- A C compiler, such as:
  - GCC 4.6 or later.
  - Clang 3.4 or later.

  While OVS may be compatible with other compilers, optimal support for atomic operations may be missing, making OVS very slow (see `lib/ovs-atomic.h`).
- libssl, from OpenSSL, is optional but recommended if you plan to connect the Open vSwitch to an OpenFlow controller. libssl is required to establish confidentiality and authenticity in the connections from an Open vSwitch to an OpenFlow controller. If libssl is installed, then Open vSwitch will automatically build with support for it.
- libcap-ng, written by Steve Grubb, is optional but recommended. It is required to run OVS daemons as a non-root user with dropped root privileges. If libcap-ng is installed, then Open vSwitch will automatically build with support for it.
- Python 3.7 or later.
- Unbound library, from http://www.unbound.net, is optional but recommended if you want to enable ovs-vswitchd and other utilities to use DNS names when specifying OpenFlow and OVSDB remotes. If unbound library is already installed, then Open vSwitch will automatically build with support for it. The environment variable OVS_RESOLV_CONF can be used to specify DNS server configuration file (the default file on Linux is /etc/resolv.conf), and environment variable OVS_UNBOUND_CONF can be used to specify the configuration file for unbound.

On Linux, you may use the kernel module distributed with the upstream Linux kernel 3.3 or later. You may also use the userspace-only implementation, at some cost in features and performance. Refer to *Open vSwitch without Kernel Support* for details.

If you are working from a Git tree or snapshot (instead of from a distribution tarball), or if you modify the Open vSwitch build system or the database schema, you will also need the following software:

- Autoconf version 2.63 or later.
- Automake version 1.10 or later.
- libtool version 2.4 or later. (Older versions might work too.)

The datapath tests for userspace and Linux datapaths also rely upon:

- pyftpdlib. Version 1.2.0 is known to work. Earlier versions should also work.
- netcat. Several common implementations are known to work.
- curl. Version 7.47.0 is known to work. Earlier versions should also work.
- tftpy. Version 0.6.2 is known to work. Earlier versions should also work.
- netstat. Available from various distro specific packages

The ovs-vswitchd.conf.db(5) manpage will include an E-R diagram, in formats other than plain text, only if you have the following:

- dot from graphviz (http://www.graphviz.org/).

If you are going to extensively modify Open vSwitch, consider installing the following to obtain better warnings:

- "sparse" version 0.6.4 or later (https://git.kernel.org/pub/scm/devel/sparse/sparse.git/).
- GNU make.
- clang, version 3.4 or later
- flake8 (for Python code)
- the python packages listed in "python/test_requirements.txt" (compatible with pip). If they are installed, the pytest-based Python unit tests will be run.

You may find the ovs-dev script found in `utilities/ovs-dev.py` useful.

Installation Requirements

The machine you build Open vSwitch on may not be the one you run it on. To simply install and run Open vSwitch you require the following software:

- Shared libraries compatible with those used for the build.
- On Linux, if you want to use the kernel-based datapath (which is the most common use case), then a kernel with a compatible kernel module. The kernel module is distributed with the upstream Linux kernel 3.3 and later. Open vSwitch features and performance can vary based on the kernel version. Refer to *Releases* for more information.
- For optional support of ingress policing on Linux, the "tc" program from iproute2 (part of all major distributions and available at https://wiki.linuxfoundation.org/networking/iproute2).
- Python 3.7 or later.

On Linux you should ensure that `/dev/urandom` exists. To support TAP devices, you must also ensure that `/dev/net/tun` exists.

Bootstrapping

This step is not needed if you have downloaded a released tarball. If you pulled the sources directly from an Open vSwitch Git tree or got a Git tree snapshot, then run boot.sh in the top source directory to build the "configure" script:

```
$ ./boot.sh
```

Configuring

Configure the package by running the configure script. You can usually invoke configure without any arguments. For example:

```
$ ./configure
```

By default all files are installed under `/usr/local`. Open vSwitch also expects to find its database in `/usr/local/etc/openvswitch` by default. If you want to install all files into, e.g., `/usr` and `/var` instead of `/usr/local` and `/usr/local/var` and expect to use `/etc/openvswitch` as the default database directory, add options as shown here:

```
$ ./configure --prefix=/usr --localstatedir=/var --sysconfdir=/etc
```

> **Note**
>
> Open vSwitch installed with packages like .rpm (e.g. via `yum install` or `rpm -ivh`) and .deb (e.g. via `apt-get install` or `dpkg -i`) use the above configure options.

By default, static libraries are built and linked against. If you want to use shared libraries instead:

```
$ ./configure --enable-shared
```

To use a specific C compiler for compiling Open vSwitch user programs, also specify it on the configure command line, like so:

```
$ ./configure CC=gcc-4.2
```

To use 'clang' compiler:

```
$ ./configure CC=clang
```

To supply special flags to the C compiler, specify them as `CFLAGS` on the configure command line. If you want the default CFLAGS, which include `-g` to build debug symbols and `-O2` to enable optimizations, you must include them yourself. For example, to build with the default CFLAGS plus `-mssse3`, you might run configure as follows:

```
$ ./configure CFLAGS="-g -O2 -mssse3"
```

For efficient hash computation special flags can be passed to leverage built-in intrinsics. For example on X86_64 with SSE4.2 instruction set support, CRC32 intrinsics can be used by passing `-msse4.2`:

```
$ ./configure CFLAGS="-g -O2 -msse4.2"`
```

Also builtin popcnt instruction can be used to speedup the counting of the bits set in an integer. For example on X86_64 with POPCNT support, it can be enabled by passing `-mpopcnt`:

```
$ ./configure CFLAGS="-g -O2 -mpopcnt"`
```

If you are on a different processor and don't know what flags to choose, it is recommended to use `-march=native` settings:

```
$ ./configure CFLAGS="-g -O2 -march=native"
```

With this, GCC will detect the processor and automatically set appropriate flags for it. This should not be used if you are compiling OVS outside the target machine.

If you are a developer and want to enable Address Sanitizer for debugging purposes, at about a 2x runtime cost, you can add `-fsanitize=address -fno-omit-frame-pointer -fno-common` to CFLAGS. For example:

```
$ ./configure CFLAGS="-g -O2 -fsanitize=address -fno-omit-frame-pointer -fno-common"
```

If you plan to do much Open vSwitch development, you might want to add `--enable-Werror`, which adds the `-Werror` option to the compiler command line, turning warnings into errors. That makes it impossible to miss warnings generated by the build. For example:

```
$ ./configure --enable-Werror
```

If you're building with GCC, then, for improved warnings, install `sparse` (see "Prerequisites") and enable it for the build by adding `--enable-sparse`. Use this with `--enable-Werror` to avoid missing both compiler and `sparse` warnings, e.g.:

```
$ ./configure --enable-Werror --enable-sparse
```

To build with gcov code coverage support, add `--enable-coverage`:

```
$ ./configure --enable-coverage
```

The configure script accepts a number of other options and honors additional environment variables. For a full list, invoke configure with the `--help` option:

```
$ ./configure --help
```

You can also run configure from a separate build directory. This is helpful if you want to build Open vSwitch in more than one way from a single source directory, e.g. to try out both GCC and Clang builds. For example:

```
$ mkdir _gcc && (cd _gcc && ./configure CC=gcc)
$ mkdir _clang && (cd _clang && ./configure CC=clang)
```

Under certain loads the ovsdb-server and other components perform better when using the jemalloc memory allocator, instead of the glibc memory allocator. If you wish to link with jemalloc add it to LIBS:

```
$ ./configure LIBS=-ljemalloc
```

Note

Linking Open vSwitch with the jemalloc shared library may not work as expected in certain operating system development environments. You can override the automatic compiler decision to avoid possible linker issues by passing `-fno-lto` or `-fno-builtin` flag since the jemalloc override standard built-in memory allocation functions such as malloc, calloc, etc. Both options can solve possible jemalloc linker issues with pros and cons for each case, feel free to choose the path that appears best to you. Disabling LTO flag example:

```
$ ./configure LIBS=-ljemalloc CFLAGS="-g -O2 -fno-lto"
```

Disabling built-in flag example:

```
$ ./configure LIBS=-ljemalloc CFLAGS="-g -O2 -fno-builtin"
```

Building

1. Run GNU make in the build directory, e.g.:

   ```
   $ make
   ```

   or if GNU make is installed as "gmake":

   ```
   $ gmake
   ```

   If you used a separate build directory, run make or gmake from that directory, e.g.:

   ```
   $ make -C _gcc
   $ make -C _clang
   ```

> **Note**
>
> Some versions of Clang and ccache are not completely compatible. If you see unusual warnings when you use both together, consider disabling ccache.

2. Consider running the testsuite. Refer to *Testing* for instructions.
3. Run `make install` to install the executables and manpages into the running system, by default under `/usr/local`:

```
$ make install
```

Starting

On Unix-alike systems, such as BSDs and Linux, starting the Open vSwitch suite of daemons is a simple process. Open vSwitch includes a shell script, and helpers, called ovs-ctl which automates much of the tasks for starting and stopping ovsdb-server, and ovs-vswitchd. After installation, the daemons can be started by using the ovs-ctl utility. This will take care to setup initial conditions, and start the daemons in the correct order. The ovs-ctl utility is located in '$(pkgdatadir)/scripts', and defaults to '/usr/local/share/openvswitch/scripts'. An example after install might be:

```
$ export PATH=$PATH:/usr/local/share/openvswitch/scripts
$ ovs-ctl start
```

Additionally, the ovs-ctl script allows starting / stopping the daemons individually using specific options. To start just the ovsdb-server:

```
$ export PATH=$PATH:/usr/local/share/openvswitch/scripts
$ ovs-ctl --no-ovs-vswitchd start
```

Likewise, to start just the ovs-vswitchd:

```
$ export PATH=$PATH:/usr/local/share/openvswitch/scripts
$ ovs-ctl --no-ovsdb-server start
```

Refer to ovs-ctl(8) for more information on ovs-ctl.

In addition to using the automated script to start Open vSwitch, you may wish to manually start the various daemons. Before starting ovs-vswitchd itself, you need to start its configuration database, ovsdb-server. Each machine on which Open vSwitch is installed should run its own copy of ovsdb-server. Before ovsdb-server itself can be started, configure a database that it can use:

```
$ mkdir -p /usr/local/etc/openvswitch
$ ovsdb-tool create /usr/local/etc/openvswitch/conf.db \
    vswitchd/vswitch.ovsschema
```

Configure ovsdb-server to use database created above, to listen on a Unix domain socket, to connect to any managers specified in the database itself, and to use the SSL/TLS configuration in the database:

```
$ mkdir -p /usr/local/var/run/openvswitch
$ ovsdb-server --remote=punix:/usr/local/var/run/openvswitch/db.sock \
    --remote=db:Open_vSwitch,Open_vSwitch,manager_options \
    --private-key=db:Open_vSwitch,SSL,private_key \
    --certificate=db:Open_vSwitch,SSL,certificate \
```

(continues on next page)

(continued from previous page)

```
    --bootstrap-ca-cert=db:Open_vSwitch,SSL,ca_cert \
    --pidfile --detach --log-file
```

Note

If you built Open vSwitch without SSL/TLS support, then omit `--private-key`, `--certificate`, and `--bootstrap-ca-cert`.)

Initialize the database using ovs-vsctl. This is only necessary the first time after you create the database with ovsdb-tool, though running it at any time is harmless:

```
$ ovs-vsctl --no-wait init
```

Start the main Open vSwitch daemon, telling it to connect to the same Unix domain socket:

```
$ ovs-vswitchd --pidfile --detach --log-file
```

Starting OVS in container

For ovs vswitchd, we need to load ovs kernel modules on host.

Hence, OVS containers kernel version needs to be same as that of host kernel.

Export following variables in .env and place it under project root:

```
$ OVS_BRANCH=<BRANCH>
$ OVS_VERSION=<VERSION>
$ DISTRO=<LINUX_DISTRO>
$ KERNEL_VERSION=<LINUX_KERNEL_VERSION>
$ GITHUB_SRC=<GITHUB_URL>
$ DOCKER_REPO=<REPO_TO_PUSH_IMAGE>
```

To build ovs modules:

```
$ cd utilities/docker
$ make build
```

Compiled Modules will be tagged with docker image

To Push ovs modules:

```
$ make push
```

OVS docker image will be pushed to specified docker repo.

Start ovsdb-server using below command:

```
$ docker run -itd --net=host --name=ovsdb-server \
  <docker_repo>:<tag> ovsdb-server
```

Start ovs-vswitchd with privileged mode as it needs to load kernel module in host using below command:

```
$ docker run -itd --net=host --name=ovs-vswitchd \
  --volumes-from=ovsdb-server -v /lib:/lib --privileged \
  <docker_repo>:<tag> ovs-vswitchd
```

Note

The debian docker file uses ubuntu 16.04 as a base image for reference.

User can use any other base image for debian, e.g. u14.04, etc.

RHEL based docker build support needs to be added.

Validating

At this point you can use ovs-vsctl to set up bridges and other Open vSwitch features. For example, to create a bridge named `br0` and add ports `eth0` and `vif1.0` to it:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 eth0
$ ovs-vsctl add-port br0 vif1.0
```

Refer to ovs-vsctl(8) for more details. You may also wish to refer to *Testing* for information on more generic testing of OVS.

When using ovs in container, exec to container to run above commands:

```
$ docker exec -it <ovsdb-server/ovs-vswitchd> /bin/bash
```

Upgrading

When you upgrade Open vSwitch from one version to another you should also upgrade the database schema:

Note

The following manual steps may also be accomplished by using ovs-ctl to stop and start the daemons after upgrade. The ovs-ctl script will automatically upgrade the schema.

1. Stop the Open vSwitch daemons, e.g.:

   ```
   $ kill `cd /usr/local/var/run/openvswitch && cat ovsdb-server.pid ovs-vswitchd.pid`
   ```

2. Install the new Open vSwitch release by using the same configure options as was used for installing the previous version. If you do not use the same configure options, you can end up with two different versions of Open vSwitch executables installed in different locations.

3. Upgrade the database, in one of the following two ways:

   - If there is no important data in your database, then you may delete the database file and recreate it with ovsdb-tool, following the instructions under "Building and Installing Open vSwitch for Linux, FreeBSD or NetBSD".
   - If you want to preserve the contents of your database, back it up first, then use `ovsdb-tool convert` to upgrade it, e.g.:

```
$ ovsdb-tool convert /usr/local/etc/openvswitch/conf.db \
    vswitchd/vswitch.ovsschema
```

4. Start the Open vSwitch daemons as described under *Starting* above.

Hot Upgrading

Upgrading Open vSwitch from one version to the next version with minimum disruption of traffic going through the system that is using that Open vSwitch needs some considerations:

1. If the upgrade only involves upgrading the userspace utilities and daemons of Open vSwitch, make sure that the new userspace version is compatible with the previously loaded kernel module.
2. An upgrade of userspace daemons means that they have to be restarted. Restarting the daemons means that the OpenFlow flows in the ovs-vswitchd daemon will be lost. One way to restore the flows is to let the controller re-populate it. Another way is to save the previous flows using a utility like ovs-ofctl and then re-add them after the restart. Restoring the old flows is accurate only if the new Open vSwitch interfaces retain the old 'ofport' values.
3. When the new userspace daemons get restarted, they automatically flush the old flows setup in the kernel. This can be expensive if there are hundreds of new flows that are entering the kernel but userspace daemons are busy setting up new userspace flows from either the controller or an utility like ovs-ofctl. Open vSwitch database provides an option to solve this problem through the `other_config:flow-restore-wait` column of the `Open_vSwitch` table. Refer to the ovs-vswitchd.conf.db(5) manpage for details.
4. If the upgrade also involves upgrading the kernel module, the old kernel module needs to be unloaded and the new kernel module should be loaded. This means that the kernel network devices belonging to Open vSwitch is recreated and the kernel flows are lost. The downtime of the traffic can be reduced if the userspace daemons are restarted immediately and the userspace flows are restored as soon as possible.
5. When upgrading ovs running in container on host that is managed by ovn, simply stop the docker container, remove and re-run with new docker image that has newer ovs version.
6. When running ovs in container, if ovs is used in bridged mode where management interface is managed by ovs, docker restart will result in loss of network connectivity. Hence, make sure to delete the bridge mapping of physical interface from ovs, upgrade ovs via docker and then add back the interface to ovs bridge. This mapping need not be deleted in case of multi nics if management interface is not managed by ovs.

The ovs-ctl utility's `restart` function only restarts the userspace daemons, makes sure that the 'ofport' values remain consistent across restarts, restores userspace flows using the ovs-ofctl utility and also uses the `other_config:flow-restore-wait` column to keep the traffic downtime to the minimum. The ovs-ctl utility's `force-reload-kmod` function does all of the above, but also replaces the old kernel module with the new one. Open vSwitch startup scripts for Debian and RHEL use ovs-ctl's functions and it is recommended that these functions be used for other software platforms too.

Reporting Bugs

Report problems to bugs@openvswitch.org.

Open vSwitch on NetBSD

On NetBSD, you might want to install requirements from pkgsrc. In that case, you need at least the following packages.

- automake
- libtool-base
- gmake

- python37

Some components have additional requirements. Refer to *Open vSwitch on Linux, FreeBSD and NetBSD* for more information.

Assuming you are running NetBSD/amd64 7.0.2, you can download and install pre-built binary packages as the following:

```
$ PKG_PATH=http://ftp.netbsd.org/pub/pkgsrc/packages/NetBSD/amd64/7.0.2/All/
$ export PKG_PATH
$ pkg_add automake libtool-base gmake python37 pkg_alternatives
```

Note

You might get some warnings about minor version mismatch. These can be safely ignored.

NetBSD's `/usr/bin/make` is not GNU make. GNU make is installed as `/usr/pkg/bin/gmake` by the above mentioned `gmake` package.

As all executables installed with pkgsrc are placed in `/usr/pkg/bin/` directory, it might be a good idea to add it to your PATH. Or install OVS by `gmake` and `gmake install`.

Open vSwitch on NetBSD is currently "userspace switch" implementation in the sense described in *Open vSwitch without Kernel Support* and *Porting Open vSwitch to New Software or Hardware*.

Open vSwitch without Kernel Support

Open vSwitch can operate, at a cost in performance, entirely in userspace, without assistance from a kernel module. This file explains how to install Open vSwitch in such a mode.

This version of Open vSwitch should be built manually with `configure` and `make`. Debian packaging for Open vSwitch is also included, but it has not been recently tested, and so Debian packages are not a recommended way to use this version of Open vSwitch.

Warning

The userspace-only mode of Open vSwitch without DPDK is considered experimental. It has not been thoroughly tested.

Building and Installing

The requirements and procedure for building, installing, and configuring Open vSwitch are the same as those given in *Open vSwitch on Linux, FreeBSD and NetBSD*. You may omit configuring, building, and installing the kernel module, and the related requirements.

On Linux, the userspace switch additionally requires the kernel TUN/TAP driver to be available, either built into the kernel or loaded as a module. If you are not sure, check for a directory named `/sys/class/misc/tun`. If it does not exist, then attempt to load the module with `modprobe tun`.

The tun device must also exist as `/dev/net/tun`. If it does not exist, then create `/dev/net` (if necessary) with `mkdir /dev/net`, then create `/dev/net/tun` with `mknod /dev/net/tun c 10 200`.

On FreeBSD and NetBSD, the userspace switch additionally requires the kernel tap(4) driver to be available, either built into the kernel or loaded as a module.

Using the Userspace Datapath with ovs-vswitchd

To use ovs-vswitchd in userspace mode, create a bridge with `datapath_type=netdev` in the configuration database. For example:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl set bridge br0 datapath_type=netdev
$ ovs-vsctl add-port br0 eth0
$ ovs-vsctl add-port br0 eth1
$ ovs-vsctl add-port br0 eth2
```

ovs-vswitchd will create a TAP device as the bridge's local interface, named the same as the bridge, as well as for each configured internal interface.

Currently, on FreeBSD, the functionality required for in-band control support is not implemented. To avoid related errors, you can disable the in-band support with the following command:

```
$ ovs-vsctl set bridge br0 other_config:disable-in-band=true
```

Firewall Rules

On Linux, when a physical interface is in use by the userspace datapath, packets received on the interface still also pass into the kernel TCP/IP stack. This can cause surprising and incorrect behavior. You can use "iptables" to avoid this behavior, by using it to drop received packets. For example, to drop packets received on eth0:

```
$ iptables -A INPUT -i eth0 -j DROP
$ iptables -A FORWARD -i eth0 -j DROP
```

Other Settings

On NetBSD, depending on your network topology and applications, the following configuration might help. See sysctl(7).:

```
$ sysctl -w net.inet.ip.checkinterface=1
```

Reporting Bugs

Report problems to bugs@openvswitch.org.

Open vSwitch with DPDK

This document describes how to build and install Open vSwitch using a DPDK datapath. Open vSwitch can use the DPDK library to operate entirely in userspace.

Important

The *releases FAQ* lists support for the required versions of DPDK for each version of Open vSwitch. If building OVS and DPDK outside of the main build tree users should consult this list first.

Build requirements

In addition to the requirements described in *Open vSwitch on Linux, FreeBSD and NetBSD*, building Open vSwitch with DPDK will require the following:

- DPDK 25.11.2
- A DPDK supported NIC

  Only required when physical ports are in use

- A suitable kernel

  On Linux Distros running kernel version >= 3.0, only *IOMMU* needs to enabled via the grub cmdline, assuming you are using **VFIO**. For older kernels, ensure the kernel is built with `UIO`, `HUGETLBFS`, `PROC_PAGE_MONITOR`, `HPET`, `HPET_MMAP` support. If these are not present, it will be necessary to upgrade your kernel or build a custom kernel with these flags enabled.

Detailed system requirements can be found at DPDK requirements.

Installing

Install DPDK

1. Download the DPDK sources, extract the file and set `DPDK_DIR`:

   ```
   $ cd /usr/src/
   $ wget https://fast.dpdk.org/rel/dpdk-25.11.2.tar.xz
   $ tar xf dpdk-25.11.2.tar.xz
   $ export DPDK_DIR=/usr/src/dpdk-stable-25.11.2
   $ cd $DPDK_DIR
   ```

2. Configure and install DPDK using Meson

   Build and install the DPDK library:

   ```
   $ export DPDK_BUILD=$DPDK_DIR/build
   $ meson build
   $ ninja -C build
   $ sudo ninja -C build install
   $ sudo ldconfig
   ```

   Check if libdpdk can be found by pkg-config:

   ```
   $ pkg-config --modversion libdpdk
   ```

   The above command should return the DPDK version installed. If not found, export the path to the installed DPDK libraries:

   ```
   $ export PKG_CONFIG_PATH=/path/to/installed/".pc" file/for/DPDK
   ```

   For example, On Fedora 32:

```
$ export PKG_CONFIG_PATH=/usr/local/lib64/pkgconfig
```

Detailed information can be found at DPDK documentation.

3. (Optional) Configure and export the DPDK shared library location

   Since DPDK is built both as static and shared library by default, no extra configuration is required for the build.

   Exporting the path to library is not necessary if the DPDK libraries are system installed. For libraries installed using a prefix, export the path to this library:

```
$ export LD_LIBRARY_PATH=/path/to/installed/DPDK/libraries
```

> **Note**
>
> Minor performance loss is expected when using OVS with a shared DPDK library compared to a static DPDK library.

Install OVS

OVS can be installed using different methods. For OVS to use DPDK, it has to be configured to build against the DPDK library (`--with-dpdk`).

> **Note**
>
> This section focuses on generic recipe that suits most cases. For distribution specific instructions, refer to one of the more relevant guides.

1. Ensure the standard OVS requirements, described in *Build Requirements*, are installed
2. Bootstrap, if required, as described in *Bootstrapping*
3. Configure the package using the `--with-dpdk` flag:

   If OVS must consume DPDK static libraries (also equivalent to `--with-dpdk=yes`):

```
$ ./configure --with-dpdk=static
```

   If OVS must consume DPDK shared libraries:

```
$ ./configure --with-dpdk=shared
```

> **Note**
>
> While `--with-dpdk` is required, you can pass any other configuration option described in *Configuring*.

   It is strongly recommended to build OVS with at least `-msse4.2` and `-mpopcnt` optimization flags.

4. Build and install OVS, as described in *Building*

Additional information can be found in *Open vSwitch on Linux, FreeBSD and NetBSD*.

> **Note**
>
> If you are running using the Fedora or Red Hat package, the Open vSwitch daemon will run as a non-root user. This implies that you must have a working IOMMU. Visit the RHEL README for additional information.

Setup

Setup Hugepages

Allocate a number of 2M Huge pages:

- For persistent allocation of huge pages, write to hugepages.conf file in */etc/sysctl.d*:

```
$ echo 'vm.nr_hugepages=2048' > /etc/sysctl.d/hugepages.conf
```

- For run-time allocation of huge pages, use the `sysctl` utility:

```
$ sysctl -w vm.nr_hugepages=N  # where N = No. of 2M huge pages
```

To verify hugepage configuration:

```
$ grep HugePages_ /proc/meminfo
```

Mount the hugepages, if not already mounted by default:

```
$ mount -t hugetlbfs none /dev/hugepages
```

> **Note**
>
> The amount of hugepage memory required can be affected by various aspects of the datapath and device configuration. Refer to *DPDK Device Memory Models* for more details.

Setup DPDK devices using VFIO

VFIO is preferred to the UIO driver when using recent versions of DPDK. VFIO support required support from both the kernel and BIOS. For the former, kernel version > 3.6 must be used. For the latter, you must enable VT-d in the BIOS and ensure this is configured via grub. To ensure VT-d is enabled via the BIOS, run:

```
$ dmesg | grep -e DMAR -e IOMMU
```

If VT-d is not enabled in the BIOS, enable it now.

To ensure VT-d is enabled in the kernel, run:

```
$ cat /proc/cmdline | grep iommu=pt
$ cat /proc/cmdline | grep intel_iommu=on
```

If VT-d is not enabled in the kernel, enable it now.

Once VT-d is correctly configured, load the required modules and bind the NIC to the VFIO driver:

```
$ modprobe vfio-pci
$ /usr/bin/chmod a+x /dev/vfio
$ /usr/bin/chmod 0666 /dev/vfio/*
```

(continues on next page)

(continued from previous page)

```
$ $DPDK_DIR/usertools/dpdk-devbind.py --bind=vfio-pci eth1
$ $DPDK_DIR/usertools/dpdk-devbind.py --status
```

Setup OVS

Open vSwitch should be started as described in *Open vSwitch on Linux, FreeBSD and NetBSD* with the exception of ovs-vswitchd, which requires some special configuration to enable DPDK functionality. DPDK configuration arguments can be passed to ovs-vswitchd via the `other_config` column of the `Open_vSwitch` table. At a minimum, the `dpdk-init` option must be set to either `true` or `try`. For example:

```
$ export PATH=$PATH:/usr/local/share/openvswitch/scripts
$ export DB_SOCK=/usr/local/var/run/openvswitch/db.sock
$ ovs-vsctl --no-wait set Open_vSwitch . other_config:dpdk-init=true
$ ovs-ctl --no-ovsdb-server --db-sock="$DB_SOCK" start
```

There are many other configuration options, the most important of which are listed below. Defaults will be provided for all values not explicitly set.

dpdk-init
: Specifies whether OVS should initialize and support DPDK ports. This field can either be `true` or `try`. A value of `true` will cause the ovs-vswitchd process to abort on initialization failure. A value of `try` will imply that the ovs-vswitchd process should continue running even if the EAL initialization fails.

dpdk-lcore-mask
: Specifies the CPU cores on which dpdk lcore threads should be spawned and expects hex string (eg '0x123').

dpdk-socket-mem
: Comma separated list of memory to pre-allocate from hugepages on specific sockets. If not specified, this option will not be set by default. DPDK default will be used instead.

dpdk-probe-at-init
: Specifies whether DPDK should probe all devices available at initialization. This option is needed when using the `class=eth,mac=XX:XX:XX:XX:XX:XX` syntax for DPDK ports. Defaults to false. See `ovs-vswitchd.conf.db(5)` for more details.

dpdk-hugepage-dir
: Directory where hugetlbfs is mounted

vhost-sock-dir
: Option to set the path to the vhost-user unix socket files.

If allocating more than one GB hugepage, you can configure the amount of memory used from any given NUMA nodes. For example, to use 1GB from NUMA node 0 and 0GB for all other NUMA nodes, run:

```
$ ovs-vsctl --no-wait set Open_vSwitch . \
    other_config:dpdk-socket-mem="1024,0"
```

or:

```
$ ovs-vsctl --no-wait set Open_vSwitch . \
    other_config:dpdk-socket-mem="1024"
```

> **Note**
>
> Changing any of these options requires restarting the ovs-vswitchd application

See the section `Performance Tuning` for important DPDK customizations.

Validating

DPDK support can be confirmed by validating the `dpdk_initialized` boolean value from the ovsdb. A value of `true` means that the DPDK EAL initialization succeeded:

```
$ ovs-vsctl get Open_vSwitch . dpdk_initialized
true
```

Additionally, the library version linked to ovs-vswitchd can be confirmed with either the ovs-vswitchd logs, or by running either of the commands:

```
$ ovs-vswitchd --version
ovs-vswitchd (Open vSwitch) 2.9.0
DPDK 17.11.0
$ ovs-vsctl get Open_vSwitch . dpdk_version
"DPDK 17.11.0"
```

At this point you can use ovs-vsctl to set up bridges and other Open vSwitch features. Seeing as we've configured DPDK support, we will use DPDK-type ports. For example, to create a userspace bridge named `br0` and add two `dpdk` ports to it, run:

```
$ ovs-vsctl add-br br0 -- set bridge br0 datapath_type=netdev
$ ovs-vsctl add-port br0 myportnameone -- set Interface myportnameone \
    type=dpdk options:dpdk-devargs=0000:06:00.0
$ ovs-vsctl add-port br0 myportnametwo -- set Interface myportnametwo \
    type=dpdk options:dpdk-devargs=0000:06:00.1
```

DPDK devices will not be available for use until a valid dpdk-devargs is specified.

Refer to ovs-vsctl(8) and *Using Open vSwitch with DPDK* for more details.

Performance Tuning

To achieve optimal OVS performance, the system can be configured and that includes BIOS tweaks, Grub cmdline additions, better understanding of NUMA nodes and apt selection of PCIe slots for NIC placement.

> **Note**
>
> This section is optional. Once installed as described above, OVS with DPDK will work out of the box.

Recommended BIOS Settings

Table 1: Recommended BIOS Settings

Setting	Value
C3 Power State	Disabled
C6 Power State	Disabled
MLC Streamer	Enabled
MLC Spatial Prefetcher	Enabled
DCU Data Prefetcher	Enabled
DCA	Enabled
CPU Power and Performance	Performance
Memory RAS and Performance Config -> NUMA optimized	Enabled

PCIe Slot Selection

The fastpath performance can be affected by factors related to the placement of the NIC, such as channel speeds between PCIe slot and CPU or the proximity of PCIe slot to the CPU cores running the DPDK application. Listed below are the steps to identify right PCIe slot.

1. Retrieve host details using `dmidecode`. For example:

   ```
   $ dmidecode -t baseboard | grep "Product Name"
   ```

2. Download the technical specification for product listed, e.g: S2600WT2

3. Check the Product Architecture Overview on the Riser slot placement, CPU sharing info and also PCIe channel speeds

   For example: On S2600WT, CPU1 and CPU2 share Riser Slot 1 with Channel speed between CPU1 and Riser Slot1 at 32GB/s, CPU2 and Riser Slot1 at 16GB/s. Running DPDK app on CPU1 cores and NIC inserted in to Riser card Slots will optimize OVS performance in this case.

4. Check the Riser Card #1 - Root Port mapping information, on the available slots and individual bus speeds. In S2600WT slot 1, slot 2 has high bus speeds and are potential slots for NIC placement.

Advanced Hugepage Setup

Allocate and mount 1 GB hugepages.

- For persistent allocation of huge pages, add the following options to the kernel bootline:

  ```
  default_hugepagesz=1GB hugepagesz=1G hugepages=N
  ```

  For platforms supporting multiple huge page sizes, add multiple options:

  ```
  default_hugepagesz=<size> hugepagesz=<size> hugepages=N
  ```

  where:

  N
  : number of huge pages requested

size
: huge page size with an optional suffix `[kKmMgG]`

- For run-time allocation of huge pages:

```
$ echo N > /sys/devices/system/node/nodeX/hugepages/hugepages-1048576kB/nr_hugepages
```

where:

N
: number of huge pages requested

X
: NUMA Node

> **Note**
>
> For run-time allocation of 1G huge pages, Contiguous Memory Allocator (`CONFIG_CMA`) has to be supported by kernel, check your Linux distro.

Now mount the huge pages, if not already done so:

```
$ mount -t hugetlbfs -o pagesize=1G none /dev/hugepages
```

Isolate Cores

The `isolcpus` option can be used to isolate cores from the Linux scheduler. The isolated cores can then be used to dedicatedly run HPC applications or threads. This helps in better application performance due to zero context switching and minimal cache thrashing. To run platform logic on core 0 and isolate cores between 1 and 19 from scheduler, add `isolcpus=1-19` to GRUB cmdline.

> **Note**
>
> It has been verified that core isolation has minimal advantage due to mature Linux scheduler in some circumstances.

Compiler Optimizations

The default compiler optimization level is `-O2`. Changing this to more aggressive compiler optimization such as `-O3 -march=native` with gcc (verified on 5.3.1) can produce performance gains though not significant. `-march=native` will produce optimized code on local machine and should be used when software compilation is done on Testbed.

Multiple Poll-Mode Driver Threads

With pmd multi-threading support, OVS creates one pmd thread for each NUMA node by default, if there is at least one DPDK interface from that NUMA node added to OVS. However, in cases where there are multiple ports/rxq's producing traffic, performance can be improved by creating multiple pmd threads running on separate cores. These pmd threads can share the workload by each being responsible for different ports/rxq's. Assignment of ports/rxq's to pmd threads is done automatically.

A set bit in the mask means a pmd thread is created and pinned to the corresponding CPU core. For example, to run pmd threads on core 1 and 2:

```
$ ovs-vsctl set Open_vSwitch . other_config:pmd-cpu-mask=0x6
```

When using dpdk and dpdkvhostuser ports in a bi-directional VM loopback as shown below, spreading the workload over 2 or 4 pmd threads shows significant improvements as there will be more total CPU occupancy available:

```
NIC port0 <-> OVS <-> VM <-> OVS <-> NIC port 1
```

Refer to ovs-vswitchd.conf.db(5) for additional information on configuration options.

Affinity

For superior performance, DPDK pmd threads and Qemu vCPU threads need to have affinity set accordingly.

- PMD thread Affinity

  A poll mode driver (pmd) thread handles the I/O of all DPDK interfaces assigned to it. A pmd thread shall poll the ports for incoming packets, switch the packets and send to tx port. A pmd thread is CPU bound, and needs to be have affinity set to isolated cores for optimum performance. Even though a PMD thread may exist, the thread only starts consuming CPU cycles if there is at least one receive queue assigned to the pmd.

  > **Note**
  >
  > On NUMA systems, PCI devices are also local to a NUMA node. Unbound rx queues for a PCI device will be assigned to a pmd on it's local NUMA node if a non-isolated PMD exists on that NUMA node. If not, the queue will be assigned to a non-isolated pmd on a remote NUMA node. This will result in reduced maximum throughput on that device and possibly on other devices assigned to that pmd thread. If such a queue assignment is made a warning message will be logged: "There's no available (non-isolated) pmd thread on numa node N. Queue Q on port P will be assigned to the pmd on core C (numa node N'). Expect reduced performance."

  Binding PMD threads to cores is described in the above section `Multiple Poll-Mode Driver Threads`.

- QEMU vCPU thread Affinity

  A VM performing simple packet forwarding or running complex packet pipelines has to ensure that the vCPU threads performing the work has as much CPU occupancy as possible.

  For example, on a multicore VM, multiple QEMU vCPU threads shall be spawned. When the DPDK `testpmd` application that does packet forwarding is invoked, the `taskset` command should be used to affinitize the vCPU threads to the dedicated isolated cores on the host system.

Enable HyperThreading

With HyperThreading, or SMT, enabled, a physical core appears as two logical cores. SMT can be utilized to spawn worker threads on logical cores of the same physical core there by saving additional cores.

With DPDK, when pinning pmd threads to logical cores, care must be taken to set the correct bits of the `pmd-cpu-mask` to ensure that the pmd threads are pinned to SMT siblings.

Take a sample system configuration, with 2 sockets, 2 * 10 core processors, HT enabled. This gives us a total of 40 logical cores. To identify the physical core shared by two logical cores, run:

```
$ cat /sys/devices/system/cpu/cpuN/topology/thread_siblings_list
```

where `N` is the logical core number.

In this example, it would show that cores `1` and `21` share the same physical core. Logical cores can be specified in pmd-cpu-masks similarly to physical cores, as described in `Multiple Poll-Mode Driver Threads`.

NUMA/Cluster-on-Die

Ideally inter-NUMA datapaths should be avoided where possible as packets will go across QPI and there may be a slight performance penalty when compared with intra NUMA datapaths. On Intel Xeon Processor E5 v3, Cluster On Die is introduced on models that have 10 cores or more. This makes it possible to logically split a socket into two NUMA regions and again it is preferred where possible to keep critical datapaths within the one cluster.

It is good practice to ensure that threads that are in the datapath are pinned to cores in the same NUMA area. e.g. pmd threads and QEMU vCPUs responsible for forwarding. If DPDK is built with `CONFIG_RTE_LIBRTE_VHOST_NUMA=y`, vHost User ports automatically detect the NUMA socket of the QEMU vCPUs and will be serviced by a PMD from the same node provided a core on this node is enabled in the `pmd-cpu-mask`. `libnuma` packages are required for this feature.

Binding PMD threads is described in the above section `Multiple Poll-Mode Driver Threads`.

DPDK Physical Port Rx Queues

```
$ ovs-vsctl set Interface <DPDK interface> options:n_rxq=<integer>
```

The above command sets the number of rx queues for DPDK physical interface. The rx queues are assigned to pmd threads on the same NUMA node in a round-robin fashion.

DPDK Physical Port Queue Sizes

```
$ ovs-vsctl set Interface dpdk0 options:n_rxq_desc=<integer>
$ ovs-vsctl set Interface dpdk0 options:n_txq_desc=<integer>
```

The above command sets the number of rx/tx descriptors that the NIC associated with dpdk0 will be initialised with.

Different `n_rxq_desc` and `n_txq_desc` configurations yield different benefits in terms of throughput and latency for different scenarios. Generally, smaller queue sizes can have a positive impact for latency at the expense of throughput. The opposite is often true for larger queue sizes. Note: increasing the number of rx descriptors eg. to 4096 may have a negative impact on performance due to the fact that non-vectorised DPDK rx functions may be used. This is dependent on the driver in use, but is true for the commonly used i40e and ixgbe DPDK drivers.

Exact Match Cache

Each pmd thread contains one Exact Match Cache (EMC). After initial flow setup in the datapath, the EMC contains a single table and provides the lowest level (fastest) switching for DPDK ports. If there is a miss in the EMC then the next level where switching will occur is the datapath classifier. Missing in the EMC and looking up in the datapath classifier incurs a significant performance penalty. If lookup misses occur in the EMC because it is too small to handle the number of flows, its size can be increased. The EMC size can be modified by editing the define `EM_FLOW_HASH_SHIFT` in `lib/dpif-netdev.c`.

As mentioned above, an EMC is per pmd thread. An alternative way of increasing the aggregate amount of possible flow entries in EMC and avoiding datapath classifier lookups is to have multiple pmd threads running.

Rx Mergeable Buffers

Rx mergeable buffers is a virtio feature that allows chaining of multiple virtio descriptors to handle large packet sizes. Large packets are handled by reserving and chaining multiple free descriptors together. Mergeable buffer support is negotiated between the virtio driver and virtio device and is supported by the DPDK vhost library. This behavior is supported and enabled by default, however in the case where the user knows that rx mergeable buffers are not needed i.e. jumbo frames are not needed, it can be forced off by adding `mrg_rxbuf=off` to the QEMU command line options. By not reserving multiple chains of descriptors it will make more individual virtio descriptors available for rx to the guest using dpdkvhost ports and this can improve performance.

Output Packet Batching

To make advantage of batched transmit functions, OVS collects packets in intermediate queues before sending when processing a batch of received packets. Even if packets are matched by different flows, OVS uses a single send operation for all packets destined to the same output port.

Furthermore, OVS is able to buffer packets in these intermediate queues for a configurable amount of time to reduce the frequency of send bursts at medium load levels when the packet receive rate is high, but the receive batch size still very small. This is particularly beneficial for packets transmitted to VMs using an interrupt-driven virtio driver, where the interrupt overhead is significant for the OVS PMD, the host operating system and the guest driver.

The `tx-flush-interval` parameter can be used to specify the time in microseconds OVS should wait between two send bursts to a given port (default is `0`). When the intermediate queue fills up before that time is over, the buffered packet batch is sent immediately:

```
$ ovs-vsctl set Open_vSwitch . other_config:tx-flush-interval=50
```

This parameter influences both throughput and latency, depending on the traffic load on the port. In general lower values decrease latency while higher values may be useful to achieve higher throughput.

Low traffic (`packet rate < 1 / tx-flush-interval`) should not experience any significant latency or throughput increase as packets are forwarded immediately.

At intermediate load levels (`1 / tx-flush-interval < packet rate < 32 / tx-flush-interval`) traffic should experience an average latency increase of up to `1 / 2 * tx-flush-interval` and a possible throughput improvement.

Very high traffic (`packet rate >> 32 / tx-flush-interval`) should experience the average latency increase equal to `32 / (2 * packet rate)`. Most send batches in this case will contain the maximum number of packets (`32`).

A `tx-burst-interval` value of `50` microseconds has shown to provide a good performance increase in a `PHY-VM-PHY` scenario on `x86` system for interrupt-driven guests while keeping the latency increase at a reasonable level:

https://mail.openvswitch.org/pipermail/ovs-dev/2017-December/341628.html

> **Note**
>
> Throughput impact of this option significantly depends on the scenario and the traffic patterns. For example: `tx-burst-interval` value of `50` microseconds shows performance degradation in `PHY-VM-PHY` with bonded PHY scenario while testing with `256 - 1024` packet flows:
>
> https://mail.openvswitch.org/pipermail/ovs-dev/2017-December/341700.html

The average number of packets per output batch can be checked in PMD stats:

```
$ ovs-appctl dpif-netdev/pmd-perf-show
```

Limitations

- Network Interface Firmware requirements: Each release of DPDK is validated against a specific firmware version for a supported Network Interface. New firmware versions introduce bug fixes, performance improvements and new functionality that DPDK leverages. The validated firmware versions are available as part of the release notes for DPDK. It is recommended that users update Network Interface firmware to match what has been validated for the DPDK release.

  The latest list of validated firmware versions can be found in the DPDK release notes.

- Upper bound MTU: DPDK device drivers differ in how the L2 frame for a given MTU value is calculated e.g. i40e driver includes 2 x vlan headers in MTU overhead, em driver includes 1 x vlan header, ixgbe driver does not include a vlan header in overhead. Currently it is not possible for OVS DPDK to know what upper bound MTU value is supported for a given device. As such OVS DPDK must provision for the case where the L2 frame for a given MTU includes 2 x vlan headers. This reduces the upper bound MTU value for devices that do not include vlan headers in their L2 frames by 8 bytes e.g. ixgbe devices upper bound MTU is reduced from 9710 to 9702. This work around is temporary and is expected to be removed once a method is provided by DPDK to query the upper bound MTU value for a given device.
- Flow Control: When using i40e devices (Intel(R) 700 Series) it is recommended to set Link State Change detection to interrupt mode. Otherwise it has been observed that using the default polling mode, flow control changes may not be applied, and flow control states will not be reflected correctly. The issue is under investigation, this is a temporary work around.

  For information about setting Link State Change detection, refer to *Link State Change (LSC) detection configuration*.

Reporting Bugs

Report problems to bugs@openvswitch.org.

Open vSwitch with AF_XDP

This document describes how to build and install Open vSwitch using AF_XDP netdev.

> **Warning**
>
> The AF_XDP support of Open vSwitch is considered 'experimental'.

Introduction

AF_XDP, Address Family of the eXpress Data Path, is a new Linux socket type built upon the eBPF and XDP technology. It is aims to have comparable performance to DPDK but cooperate better with existing kernel's networking stack. An AF_XDP socket receives and sends packets from an eBPF/XDP program attached to the netdev, by-passing a couple of Linux kernel's subsystems. As a result, AF_XDP socket shows much better performance than AF_PACKET. For more details about AF_XDP, please see linux kernel's Documentation/networking/af_xdp.rst

AF_XDP Netdev

OVS has a couple of netdev types, i.e., system, tap, or dpdk. The AF_XDP feature adds a new netdev types called "afxdp", and implement its configuration, packet reception, and transmit functions. Since the AF_XDP socket, called xsk, operates in userspace, once ovs-vswitchd receives packets from xsk, the afxdp netdev re-uses the existing userspace dpif-netdev datapath. As a result, most of the packet processing happens at the userspace instead of linux kernel.

```
           |   +-------------------+
           |   |    ovs-vswitchd   |<-->ovsdb-server
           |   +-------------------+
           |   |      ofproto      |<-->OpenFlow controllers
           |   +--------+-+--------+
           |   | netdev | |ofproto-|
userspace  |   +--------+ |  dpif  |
           |   | afxdp  | +--------+
           |   | netdev | |  dpif  |
           |   +---||---+ +--------+
```

(continues on next page)

(continued from previous page)

```
         |       ||     |  dpif- |
         |       ||     | netdev |
         |_      ||     +--------+
                 ||
          _  +---||-----+--------+
         |   | AF_XDP prog +     |
  kernel |   |   xsk_map         |
         |_  +--------||---------+
                     ||
                  physical
                     NIC
```

Build requirements

In addition to the requirements described in *Open vSwitch on Linux, FreeBSD and NetBSD*, building Open vSwitch with AF_XDP will require the following:

- `libbpf` and `libxdp` (if version of `libbpf` if higher than `0.6`).
- Linux kernel v5.4+ or a nominally older distribution kernel that has required features backported.
- Linux kernel XDP support, with the following options (required)
  - CONFIG_BPF=y
  - CONFIG_BPF_SYSCALL=y
  - CONFIG_XDP_SOCKETS=y
- The following optional Kconfig options are also recommended, but not required:
  - CONFIG_BPF_JIT=y (Performance)
  - CONFIG_HAVE_EBPF_JIT=y (Performance)
  - CONFIG_XDP_SOCKETS_DIAG=y (Debugging)
- If you're building your own kernel, be sure that you're installing kernel headers too. For example, with the following command:

```
make headers_install INSTALL_HDR_PATH=/usr
```

- If you're using kernel from the distribution, be sure that corresponding kernel headers package installed.
- Once your AF_XDP-enabled kernel is ready, if possible, run **./xdpsock -r -N -z -i <your device>** under linux/samples/bpf. This is an OVS independent benchmark tools for AF_XDP. It makes sure your basic kernel requirements are met for AF_XDP.

Installing

For OVS to use AF_XDP netdev, it has to be configured with LIBBPF support.

First, install `libbpf` and `libxdp`. For example, on Fedora these libraries along with development headers can be obtained by installing `libbpf-devel` and `libxdp-devel` packages. For Ubuntu that will be `libbpf-dev` package with additional `libxdp-dev` on Ubuntu 22.10 or later.

Next, ensure the standard OVS requirements are installed and bootstrap/configure the package:

```
./boot.sh && ./configure --enable-afxdp
```

`--enable-afxdp` here is optional, but it will ensure that all dependencies are available at the build time.

Finally, build and install OVS:

```
make && make install
```

To kick start end-to-end autotesting:

```
uname -a # make sure having 5.4+ kernel
ethtool --version # make sure ethtool is installed
make check-afxdp TESTSUITEFLAGS='1'
```

> **Note**
>
> Not all test cases pass at this time. Currently all cvlan tests are skipped due to kernel issues.

If a test case fails, check the log at:

```
cat \
tests/system-afxdp-testsuite.dir/<test num>/system-afxdp-testsuite.log
```

Setup AF_XDP netdev

Before running OVS with AF_XDP, make sure the libbpf and libnuma are set-up right:

```
ldd vswitchd/ovs-vswitchd
```

Open vSwitch should be started using userspace datapath as described in *Open vSwitch on Linux, FreeBSD and NetBSD*:

```
ovs-vswitchd ...
ovs-vsctl -- add-br br0 -- set Bridge br0 datapath_type=netdev
```

Make sure your device driver support AF_XDP, netdev-afxdp supports the following additional options (see `man ovs-vswitchd.conf.db` for more details):

- `xdp-mode`: `best-effort`, `native-with-zerocopy`, `native` or `generic`. Defaults to `best-effort`, i.e. best of supported modes, so in most cases you don't need to change it.

For example, to use 1 PMD (on core 4) on 1 queue (queue 0) device, configure these options: `pmd-cpu-mask`, `pmd-rxq-affinity`, and `n_rxq`:

```
ethtool -L enp2s0 combined 1
ovs-vsctl set Open_vSwitch . other_config:pmd-cpu-mask=0x10
ovs-vsctl add-port br0 enp2s0 -- set interface enp2s0 type="afxdp" \
                                 other_config:pmd-rxq-affinity="0:4"
```

Or, use 4 pmds/cores and 4 queues by doing:

```
ethtool -L enp2s0 combined 4
ovs-vsctl set Open_vSwitch . other_config:pmd-cpu-mask=0x36
ovs-vsctl add-port br0 enp2s0 -- set interface enp2s0 type="afxdp" \
  options:n_rxq=4 other_config:pmd-rxq-affinity="0:1,1:2,2:3,3:4"
```

> **Note**
>
> `pmd-rxq-affinity` is optional. If not specified, system will auto-assign. `n_rxq` equals 1 by default.

To validate that the bridge has successfully instantiated, you can use the:

```
ovs-vsctl show
```

Should show something like:

```
Port "ens802f0"
 Interface "ens802f0"
    type: afxdp
    options: {n_rxq="1"}
```

Otherwise, enable debugging by:

```
ovs-appctl vlog/set netdev_afxdp::dbg
```

To check which XDP mode was chosen by `best-effort`, you can look for `xdp-mode` in the output of `ovs-vsctl get interface INT status:xdp-mode`:

```
# ovs-vsctl get interface ens802f0 status:xdp-mode
"native-with-zerocopy"
```

References

Most of the design details are described in the paper presented at Linux Plumber 2018, "Bringing the Power of eBPF to Open vSwitch"[1], section 4, and slides[2][4]. "The Path to DPDK Speeds for AF XDP"[3] gives a very good introduction about AF_XDP current and future work.

[1] http://vger.kernel.org/lpc_net2018_talks/ovs-ebpf-afxdp.pdf

[2] http://vger.kernel.org/lpc_net2018_talks/ovs-ebpf-lpc18-presentation.pdf

[3] http://vger.kernel.org/lpc_net2018_talks/lpc18_paper_af_xdp_perf-v2.pdf

[4] https://ovsfall2018.sched.com/event/IO7p/fast-userspace-ovs-with-afxdp

Performance Tuning

The name of the game is to keep your CPU running in userspace, allowing PMD to keep polling the AF_XDP queues without any interferences from kernel.

1. Make sure everything is in the same NUMA node (memory used by AF_XDP, pmd running cores, device plug-in slot)
2. Isolate your CPU by doing isolcpu at grub configure.
3. IRQ should not set to pmd running core.
4. The Spectre and Meltdown fixes increase the overhead of system calls.

Debugging performance issue

While running the traffic, use linux perf tool to see where your cpu spends its cycle:

```
cd bpf-next/tools/perf
make
./perf record -p `pidof ovs-vswitchd` sleep 10
./perf report
```

Measure your system call rate by doing:

```
pstree -p `pidof ovs-vswitchd`
strace -c -p <your pmd's PID>
```

Or, use OVS pmd tool:

```
ovs-appctl dpif-netdev/pmd-perf-show
```

Example Script

Below is a script using namespaces and veth peer:

```
#!/bin/bash
ovs-vswitchd --no-chdir --pidfile -vvconn -vofproto_dpif -vunixctl \
  --disable-system --detach \
ovs-vsctl -- add-br br0 -- set Bridge br0 \
  protocols=OpenFlow10,OpenFlow11,OpenFlow12,OpenFlow13,OpenFlow14 \
  fail-mode=secure datapath_type=netdev

ip netns add at_ns0
ovs-appctl vlog/set netdev_afxdp::dbg

ip link add p0 type veth peer name afxdp-p0
ip link set p0 netns at_ns0
ip link set dev afxdp-p0 up
ovs-vsctl add-port br0 afxdp-p0 -- \
  set interface afxdp-p0 external-ids:iface-id="p0" type="afxdp"

ip netns exec at_ns0 sh << NS_EXEC_HEREDOC
ip addr add "10.1.1.1/24" dev p0
ip link set dev p0 up
NS_EXEC_HEREDOC

ip netns add at_ns1
ip link add p1 type veth peer name afxdp-p1
ip link set p1 netns at_ns1
ip link set dev afxdp-p1 up

ovs-vsctl add-port br0 afxdp-p1 -- \
  set interface afxdp-p1 external-ids:iface-id="p1" type="afxdp"
ip netns exec at_ns1 sh << NS_EXEC_HEREDOC
ip addr add "10.1.1.2/24" dev p1
ip link set dev p1 up
NS_EXEC_HEREDOC
```

(continues on next page)

(continued from previous page)

```
ip netns exec at_ns0 ping -i .2 10.1.1.2
```

Limitations/Known Issues

1. No QoS support because AF_XDP netdev by-pass the Linux TC layer. A possible work-around is to use OpenFlow meter action.
2. Most of the tests are done using i40e single port. Multiple ports and also ixgbe driver also needs to be tested.
3. No latency test result (TODO items)
4. Due to limitations of current upstream kernel, various offloading (vlan, cvlan) is not working over virtual interfaces (i.e. veth pair). Also, TCP is not working over virtual interfaces (veth) in generic XDP mode. Some more information and possible workaround available here . For TAP interfaces generic mode seems to work fine (TCP works) and even could provide better performance than native mode in some cases.

PVP using tap device

Assume you have enp2s0 as physical nic, and a tap device connected to VM. First, start OVS, then add physical port:

```
ethtool -L enp2s0 combined 1
ovs-vsctl set Open_vSwitch . other_config:pmd-cpu-mask=0x10
ovs-vsctl add-port br0 enp2s0 -- set interface enp2s0 type="afxdp" \
  options:n_rxq=1 other_config:pmd-rxq-affinity="0:4"
```

Start a VM with virtio and tap device:

```
qemu-system-x86_64 -hda ubuntu1810.qcow \
  -m 4096 \
  -cpu host,+x2apic -enable-kvm \
  -device virtio-net-pci,mac=00:02:00:00:00:01,netdev=net0,mq=on,vectors=10,mrg_rxbuf=on,
↪rx_queue_size=1024 \
  -netdev type=tap,id=net0,vhost=on,queues=8 \
  -object memory-backend-file,id=mem,size=4096M,mem-path=/dev/hugepages,share=on \
  -numa node,memdev=mem -mem-prealloc -smp 2
```

Create OpenFlow rules:

```
ovs-vsctl add-port br0 tap0 -- set interface tap0
ovs-ofctl del-flows br0
ovs-ofctl add-flow br0 "in_port=enp2s0, actions=output:tap0"
ovs-ofctl add-flow br0 "in_port=tap0, actions=output:enp2s0"
```

Inside the VM, use xdp_rxq_info to bounce back the traffic:

```
./xdp_rxq_info --dev ens3 --action XDP_TX
```

PVP using vhostuser device

First, build OVS with DPDK and AFXDP:

```
./configure  --enable-afxdp --with-dpdk=shared|static
make -j4 && make install
```

Create a vhost-user port from OVS:

```
ovs-vsctl --no-wait set Open_vSwitch . other_config:dpdk-init=true
ovs-vsctl -- add-br br0 -- set Bridge br0 datapath_type=netdev \
  other_config:pmd-cpu-mask=0xfff
ovs-vsctl add-port br0 vhost-user-1 \
  -- set Interface vhost-user-1 type=dpdkvhostuserclient \
      options:vhost-server-path=/tmp/vhost-user-1
```

Start VM using vhost-user mode:

```
qemu-system-x86_64 -hda ubuntu1810.qcow \
 -m 4096 \
 -cpu host,+x2apic -enable-kvm \
 -chardev socket,id=char1,path=/tmp/vhost-user-1,server \
 -netdev type=vhost-user,id=mynet1,chardev=char1,vhostforce,queues=4 \
 -device virtio-net-pci,mac=00:00:00:00:00:01,netdev=mynet1,mq=on,vectors=10 \
 -object memory-backend-file,id=mem,size=4096M,mem-path=/dev/hugepages,share=on \
 -numa node,memdev=mem -mem-prealloc -smp 2
```

Setup the OpenFlow ruls:

```
ovs-ofctl del-flows br0
ovs-ofctl add-flow br0 "in_port=enp2s0, actions=output:vhost-user-1"
ovs-ofctl add-flow br0 "in_port=vhost-user-1, actions=output:enp2s0"
```

Inside the VM, use xdp_rxq_info to drop or bounce back the traffic:

```
./xdp_rxq_info --dev ens3 --action XDP_DROP
./xdp_rxq_info --dev ens3 --action XDP_TX
```

PCP container using veth

Create namespace and veth peer devices:

```
ip netns add at_ns0
ip link add p0 type veth peer name afxdp-p0
ip link set p0 netns at_ns0
ip link set dev afxdp-p0 up
ip netns exec at_ns0 ip link set dev p0 up
```

Attach the veth port to br0 (linux kernel mode):

```
ovs-vsctl add-port br0 afxdp-p0 -- set interface afxdp-p0
```

Or, use AF_XDP:

```
ovs-vsctl add-port br0 afxdp-p0 -- set interface afxdp-p0 type="afxdp"
```

Setup the OpenFlow rules:

```
ovs-ofctl del-flows br0
ovs-ofctl add-flow br0 "in_port=enp2s0, actions=output:afxdp-p0"
ovs-ofctl add-flow br0 "in_port=afxdp-p0, actions=output:enp2s0"
```

In the namespace, run drop or bounce back the packet:

```
ip netns exec at_ns0 ./xdp_rxq_info --dev p0 --action XDP_DROP
ip netns exec at_ns0 ./xdp_rxq_info --dev p0 --action XDP_TX
```

Bug Reporting

Please report problems to dev@openvswitch.org.

2.3.2 Installation from Packages

Open vSwitch is packaged on a variety of distributions. The tooling required to build these packages is included in the Open vSwitch tree. The instructions are provided below.

Distributions packaging Open vSwitch

This document lists various popular distributions packaging Open vSwitch. Open vSwitch is packaged by various distributions for multiple platforms and architectures.

> **Note**
>
> The packaged version available with distributions may not be latest Open vSwitch release.

Debian / Ubuntu

You can use `apt-get` or `aptitude` to install the .deb packages and must be superuser.

1. Debian and Ubuntu has `openvswitch-switch` and `openvswitch-common` packages that includes the core userspace components of the switch. Extra packages for documentation, ipsec, pki, VTEP and Python support are also available. The Open vSwitch kernel datapath is maintained as part of the upstream kernel available in the distribution.

2. For fast userspace switching, Open vSwitch with DPDK support is bundled in the package `openvswitch-switch-dpdk`.

Fedora

Fedora provides `openvswitch`, `openvswitch-devel`, `openvswitch-test` and `openvswitch-debuginfo` rpm packages. You can install `openvswitch` package in minimum installation. Use `yum` or `dnf` to install the rpm packages and must be superuser.

Red Hat

RHEL distributes `openvswitch` rpm package that supports kernel datapath. DPDK accelerated Open vSwitch can be installed using `openvswitch-dpdk` package.

OpenSuSE

OpenSUSE provides `openvswitch`, `openvswitch-switch` rpm packages. Also `openvswitch-dpdk` and `openvswitch-dpdk-switch` can be installed for Open vSwitch using DPDK accelerated datapath.

Debian Packaging for Open vSwitch

This document describes how to build Debian packages for Open vSwitch. To install Open vSwitch on Debian without building Debian packages, refer to *Open vSwitch on Linux, FreeBSD and NetBSD* instead.

> **Note**
>
> These instructions should also work on Ubuntu and other Debian derivative distributions.

Before You Begin

Before you begin, consider whether you really need to build packages yourself. Debian "wheezy" and "sid", as well as recent versions of Ubuntu, contain pre-built Debian packages for Open vSwitch. It is easier to install these than to build your own. To use packages from your distribution, skip ahead to "Installing .deb Packages", below.

Building Open vSwitch Debian packages

You may build from an Open vSwitch distribution tarball or from an Open vSwitch Git tree with these instructions.

You do not need to be the superuser to build the Debian packages.

1. Install the "build-essential" and "fakeroot" packages. For example:

   ```
   $ apt-get install build-essential fakeroot
   ```

2. Obtain and unpack an Open vSwitch source distribution and `cd` into its top level directory.

3. Install the build dependencies listed under "Build-Depends:" near the top of `debian/control.in`. You can install these any way you like, e.g. with `apt-get install`.

4. Prepare the package source.

   If you want to build the package with DPDK support execute the following command:

   ```
   $ ./boot.sh && ./configure --with-dpdk=shared && make debian
   ```

   If not:

   ```
   $ ./boot.sh && ./configure && make debian
   ```

Check your work by running `dpkg-checkbuilddeps` in the top level of your OVS directory. If you've installed all the dependencies properly, `dpkg-checkbuilddeps` will exit without printing anything. If you forgot to install some dependencies, it will tell you which ones.

5. Build the package:

   ```
   $ make debian-deb
   ```

5. The generated .deb files will be in the parent directory of the Open vSwitch source distribution.

Installing .deb Packages

These instructions apply to installing from Debian packages that you built yourself, as described in the previous section. In this case, use a command such as `dpkg -i` to install the .deb files that you build. You will have to manually install any missing dependencies.

You can also use these instruction to install from packages provided by Debian or a Debian derivative distribution such as Ubuntu. In this case, use a program such as `apt-get` or `aptitude` to download and install the provided packages. These programs will also automatically download and install any missing dependencies.

Important

You must be superuser to install Debian packages.

Install the `openvswitch-switch` and `openvswitch-common` packages. These packages include the core userspace components of the switch.

Open vSwitch `.deb` packages not mentioned above are rarely useful. Refer to their individual package descriptions to find out whether any of them are useful to you.

Reporting Bugs

Report problems to bugs@openvswitch.org.

Fedora, RHEL 7.x Packaging for Open vSwitch

This document provides instructions for building and installing Open vSwitch RPM packages on a Fedora Linux host. Instructions for the installation of Open vSwitch on a Fedora Linux host without using RPM packages can be found in the *Open vSwitch on Linux, FreeBSD and NetBSD*.

These instructions have been tested with Fedora 23, and are also applicable for RHEL 7.x and its derivatives, including CentOS 7.x and Scientific Linux 7.x.

Build Requirements

You will need to install all required packages to build the RPMs. Fedora 41 and newer version use `dnf5` by default. Other distributions typically use `dnf`. If neither `dnf5` nor `dnf` is available, then use `yum` instructions.

The command below will install RPM tools and generic build dependencies. And (optionally) include these packages: libcap-ng libcap-ng-devel dpdk-devel.

DNF5:

```
$ dnf5 install @development-tools rpm-build dnf-plugins-core
```

DNF:

```
$ dnf install @'Development Tools' rpm-build dnf-plugins-core
```

YUM:

```
$ yum install @'Development Tools' rpm-build yum-utils
```

Then it is necessary to install Open vSwitch specific build dependencies. The dependencies are listed in the SPEC file, but first it is necessary to replace the VERSION tag to be a valid SPEC.

The command below will create a temporary SPEC file:

```
$ sed -e 's/@VERSION@/0.0.1/' rhel/openvswitch-fedora.spec.in \
  > /tmp/ovs.spec
```

And to install specific dependencies, use the corresponding tool below. For some of the dependencies on RHEL you may need to add two additional repositories to help yum-builddep, e.g.:

```
$ subscription-manager repos --enable=rhel-7-server-extras-rpms
$ subscription-manager repos --enable=rhel-7-server-optional-rpms
```

or for RHEL 8:

```
$ subscription-manager repos \
  --enable=codeready-builder-for-rhel-8-x86_64-rpms
```

DNF:

```
$ dnf builddep /tmp/ovs.spec
```

YUM:

```
$ yum-builddep /tmp/ovs.spec
```

Once that is completed, remove the file `/tmp/ovs.spec`.

Bootstrapping

Refer to *Bootstrapping*.

Configuring

Refer to *Configuring*.

Building

User Space RPMs

To build Open vSwitch user-space RPMs, execute the following from the directory in which *./configure* was executed:

```
$ make rpm-fedora
```

This will create the RPMs *openvswitch*, *python3-openvswitch*, *openvswitch-test*, *openvswitch-devel* and *openvswitch-debuginfo*.

To enable DPDK support in the openvswitch package, the `--with dpdk` option can be added:

```
$ make rpm-fedora RPMBUILD_OPT="--with dpdk --without check"
```

To enable AF_XDP support in the openvswitch package, the `--with afxdp` option can be added:

```
$ make rpm-fedora RPMBUILD_OPT="--with afxdp --without check"
```

You can also have the above commands automatically run the Open vSwitch unit tests. This can take several minutes.

```
$ make rpm-fedora RPMBUILD_OPT="--with check"
```

Installing

RPM packages can be installed by using the command `rpm -i`. Package installation requires superuser privileges.

In most cases only the *openvswitch* RPM will need to be installed. The *python3-openvswitch*, *openvswitch-test*, *openvswitch-devel*, and *openvswitch-debuginfo* RPMs are optional unless required for a specific purpose.

Refer to the RHEL README for additional usage and configuration information.

Reporting Bugs

Report problems to bugs@openvswitch.org.

RHEL 5.6, 6.x Packaging for Open vSwitch

This document describes how to build and install Open vSwitch on a Red Hat Enterprise Linux (RHEL) host. If you want to install Open vSwitch on a generic Linux host, refer to *Open vSwitch on Linux, FreeBSD and NetBSD* instead.

We have tested these instructions with RHEL 5.6 and RHEL 6.0.

For RHEL 7.x (or derivatives, such as CentOS 7.x), you should follow the instructions in the *Fedora, RHEL 7.x Packaging for Open vSwitch*. The Fedora spec files are used for RHEL 7.x.

Prerequisites

You may build from an Open vSwitch distribution tarball or from an Open vSwitch Git tree.

The default RPM build directory, `_topdir`, has five directories in the top-level.

BUILD/
: where the software is unpacked and built

RPMS/
: where the newly created binary package files are written

SOURCES/
: contains the original sources, patches, and icon files

SPECS/
: contains the spec files for each package to be built

SRPMS/
: where the newly created source package files are written

Before you begin, note the RPM sources directory on your version of RHEL. The command `rpmbuild --showrc` will show the configuration for each of those directories. Alternatively, the command `rpm --eval '%{_topdir}'` shows the current configuration for the top level directory and the command `rpm --eval '%{_sourcedir}'` does the same for the sources directory. On RHEL 5, the default RPM `_topdir` is `/usr/src/redhat` and the default RPM sources directory is `/usr/src/redhat/SOURCES`. On RHEL 6, the default `_topdir` is `$HOME/rpmbuild` and the default RPM sources directory is `$HOME/rpmbuild/SOURCES`.

Build Requirements

You will need to install all required packages to build the RPMs. The command below will install RPM tools and generic build dependencies:

```
$ yum install @'Development Tools' rpm-build yum-utils
```

Then it is necessary to install Open vSwitch specific build dependencies. The dependencies are listed in the SPEC file, but first it is necessary to replace the VERSION tag to be a valid SPEC.

The command below will create a temporary SPEC file:

```
$ sed -e 's/@VERSION@/0.0.1/' rhel/openvswitch.spec.in > /tmp/ovs.spec
```

And to install specific dependencies, use yum-builddep tool:

```
$ yum-builddep /tmp/ovs.spec
```

Once that is completed, remove the file `/tmp/ovs.spec`.

If python3-sphinx package is not available in your version of RHEL, you can install it via pip with 'pip install sphinx'.

Open vSwitch requires python 3.7 or newer which is not available in older distributions. For those, one option is to build and install required version from source.

Bootstrapping and Configuring

If you are building from a distribution tarball, skip to *Building*. If not, you must be building from an Open vSwitch Git tree. Determine what version of Autoconf is installed (e.g. run `autoconf --version`). If it is not at least version 2.63, then you must upgrade or use another machine to build the packages.

Assuming all requirements have been met, build the tarball by running:

```
$ ./boot.sh
$ ./configure
$ make dist
```

You must run this on a machine that has the tools listed in *Build Requirements* as prerequisites for building from a Git tree. Afterward, proceed with the rest of the instructions using the distribution tarball.

Now you have a distribution tarball, named something like `openvswitch-x.y.z.tar.gz`. Copy this file into the RPM sources directory, e.g.:

```
$ cp openvswitch-x.y.z.tar.gz $HOME/rpmbuild/SOURCES
```

Broken `build` symlink

Some versions of the RHEL 6 kernel-devel package contain a broken `build` symlink. If you are using such a version, you must fix the problem before continuing.

To find out whether you are affected, run:

```
$ cd /lib/modules/<version>
$ ls -l build/
```

where <version> is the version number of the RHEL 6 kernel.

> **Note**
>
> The trailing slash in the final command is important. Be sure to include it.

If the `ls` command produces a directory listing, your kernel-devel package is OK. If it produces a `No such file or directory` error, your kernel-devel package is buggy.

If your kernel-devel package is buggy, then you can fix it with:

```
$ cd /lib/modules/<version>
$ rm build
$ ln -s /usr/src/kernels/<target> build
```

where <target> is the name of an existing directory under `/usr/src/kernels`, whose name should be similar to <version> but may contain some extra parts. Once you have done this, verify the fix with the same procedure you used above to check for the problem.

Building

You should have a distribution tarball named something like openvswitch-x.y.z.tar.gz. Copy this file into the RPM sources directory:

```
$ cp openvswitch-x.y.z.tar.gz $HOME/rpmbuild/SOURCES
```

Make another copy of the distribution tarball in a temporary directory. Then unpack the tarball and `cd` into its root:

```
$ tar xzf openvswitch-x.y.z.tar.gz
$ cd openvswitch-x.y.z
```

Userspace

To build Open vSwitch userspace, run:

```
$ rpmbuild -bb rhel/openvswitch.spec
```

This produces two RPMs: "openvswitch" and "openvswitch-debuginfo".

The above command automatically runs the Open vSwitch unit tests. To disable the unit tests, run:

```
$ rpmbuild -bb --without check rhel/openvswitch.spec
```

> **Note**
>
> If the build fails with `configure: error: source dir /lib/modules/2.6.32-279.el6.x86_64/build doesn't exist` or similar, then the kernel-devel package is missing or buggy.

Red Hat Network Scripts Integration

A RHEL host has default firewall rules that prevent any Open vSwitch tunnel traffic from passing through. If a user configures Open vSwitch tunnels like Geneve, GRE, VXLAN, etc., they will either have to manually add iptables firewall rules to allow the tunnel traffic or add it through a startup script Refer to the "enable-protocol" command in the ovs-ctl(8) manpage for more information.

In addition, simple integration with Red Hat network scripts has been implemented. Refer to README.RHEL.rst in the source tree or /usr/share/doc/openvswitch/README.RHEL.rst in the installed openvswitch package for details.

Reporting Bugs

Report problems to bugs@openvswitch.org.

2.3.3 Others

Bash command-line completion scripts

There are two completion scripts available: `ovs-appctl-bashcomp.bash` and `ovs-vsctl-bashcomp.bash`.

ovs-appctl-bashcomp

`ovs-appctl-bashcomp.bash` adds bash command-line completion support for `ovs-appctl`, `ovs-dpctl`, `ovs-ofctl` and `ovsdb-tool` commands.

Features

- Display available completion or complete on unfinished user input (long option, subcommand, and argument).
- Subcommand hints
- Convert between keywords like `bridge`, `port`, `interface`, or `dp` and the available record in ovsdb.

Limitations

- Only supports a small set of important keywords (`dp`, `datapath`, `bridge`, `switch`, `port`, `interface`, `iface`).
- Does not support parsing of nested options. For example:

```
$ ovsdb-tool create [db [schema]]
```

- Does not support expansion on repeated argument. For example:

```
$ ovs-dpctl show [dp...]).
```

- Only supports matching on long options, and only in the format `--option [arg]`. Do not use `--option=[arg]`.

ovs-vsctl-bashcomp

`ovs-vsctl-bashcomp.bash` adds Bash command-line completion support for `ovs-vsctl` command.

Features

- Display available completion and complete on user input for global/local options, command, and argument.
- Query database and expand keywords like `table`, `record`, `column`, or `key`, to available completions.
- Deal with argument relations like 'one and more', 'zero or one'.
- Complete multiple `ovs-vsctl` commands cascaded via `--`.

Limitations

Completion of very long `ovs-vsctl` commands can take up to several seconds.

Usage

The bashcomp scripts should be placed at `/etc/bash_completion.d/` to be available for all bash sessions. Running `make install` will place the scripts to `$(sysconfdir)/bash_completion.d/`, thus, the user should specify `--sysconfdir=/etc` at configuration. If OVS is installed from packages, the scripts will automatically be placed inside `/etc/bash_completion.d/`.

If you just want to run the scripts in one bash, you can remove them from `/etc/bash_completion.d/` and run the scripts via `. ovs-appctl-bashcomp.bash` or `. ovs-vsctl-bashcomp.bash`.

Tests

Unit tests are added in `tests/completion.at` and integrated into autotest framework. To run the tests, just run `make check`.

Open vSwitch Documentation

This document describes how to build the OVS documentation for use offline. A continuously updated, online version can be found at docs.openvswitch.org.

> **Note**
>
> These instructions provide information on building the documentation locally. For information on writing documentation, refer to *Documentation Style*

Build Requirements

As described in the *Documentation Style*, the Open vSwitch documentation is written in reStructuredText and built with Sphinx. A detailed guide on installing Sphinx in many environments is available on the Sphinx website but, for most Linux distributions, you can install with your package manager. For example, on Debian/Ubuntu run:

```
$ sudo apt-get install python3-sphinx
```

Similarly, on RHEL/Fedora run:

```
$ sudo dnf install python3-sphinx
```

A `requirements.txt` is also provided in the `/Documentation`, should you wish to install using `pip`:

```
$ virtualenv .venv
$ source .venv/bin/activate
$ pip install -r Documentation/requirements.txt
```

Configuring

It's unlikely that you'll need to customize any aspect of the configuration. However, the `Documentation/conf.py` is the go-to place for all configuration. This file is well documented and further information is available on the Sphinx website.

Building

Once Sphinx is installed, the documentation can be built using the provided Makefile targets:

```
$ make docs-check
```

> **Important**
>
> The `docs-check` target will fail if there are any syntax errors. However, it won't catch more succinct issues such as style or grammar issues. As a result, you should always inspect changes visually to ensure the result is as intended.

Once built, documentation is available in the `/Documentation/_build` folder. Open the root `index.html` to browse the documentation.

CHAPTER

THREE

TUTORIALS

Getting started with Open vSwitch (OVS).

3.1 OVS Faucet Tutorial

This tutorial demonstrates how Open vSwitch works with a general-purpose OpenFlow controller, using the Faucet controller as a simple way to get started. It was tested with the "main" branch of Open vSwitch and version 1.6.15 of Faucet. It does not use advanced or recently added features in OVS or Faucet, so other versions of both pieces of software are likely to work equally well.

The goal of the tutorial is to demonstrate Open vSwitch and Faucet in an end-to-end way, that is, to show how it works from the Faucet controller configuration at the top, through the OpenFlow flow table, to the datapath processing. Along the way, in addition to helping to understand the architecture at each level, we discuss performance and troubleshooting issues. We hope that this demonstration makes it easier for users and potential users to understand how Open vSwitch works and how to debug and troubleshoot it.

We provide enough details in the tutorial that you should be able to fully follow along by following the instructions.

3.1.1 Setting Up OVS

This section explains how to set up Open vSwitch for the purpose of using it with Faucet for the tutorial.

You might already have Open vSwitch installed on one or more computers or VMs, perhaps set up to control a set of VMs or a physical network. This is admirable, but we will be using Open vSwitch in a different way to set up a simulation environment called the OVS "sandbox". The sandbox does not use virtual machines or containers, which makes it more limited, but on the other hand it is (in this writer's opinion) easier to set up.

There are two ways to start a sandbox: one that uses the Open vSwitch that is already installed on a system, and another that uses a copy of Open vSwitch that has been built but not yet installed. The latter is more often used and thus better tested, but both should work. The instructions below explain both approaches:

1. Get a copy of the Open vSwitch source repository using Git, then `cd` into the new directory:

   ```
   $ git clone https://github.com/openvswitch/ovs.git
   $ cd ovs
   ```

   The default checkout is the main branch. You will need to use the main branch for this tutorial as it includes some functionality required for this tutorial.

2. If you do not already have an installed copy of Open vSwitch on your system, or if you do not want to use it for the sandbox (the sandbox will not disturb the functionality of any existing switches), then proceed to step 3. If you do have an installed copy and you want to use it for the sandbox, try to start the sandbox by running:

   ```
   $ tutorial/ovs-sandbox
   ```

> **Note**
>
> The default behaviour for some of the commands used in this tutorial changed in Open vSwitch versions 2.9.x and 2.10.x which breaks the tutorial. We recommend following step 3 and building main from source or using a system Open vSwitch that is version 2.8.x or older.

If it is successful, you will find yourself in a subshell environment, which is the sandbox (you can exit with `exit` or Control+D). If so, you're finished and do not need to complete the rest of the steps. If it fails, you can proceed to step 3 to build Open vSwitch anyway.

3. Before you build, you might want to check that your system meets the build requirements. Read *Open vSwitch on Linux, FreeBSD and NetBSD* to find out. For this tutorial, there is no need to compile the Linux kernel module, or to use any of the optional libraries such as OpenSSL, DPDK, or libcap-ng.

   If you are using a Linux system that uses apt and have some `deb-src` repos listed in `/etc/apt/sources.list`, often an easy way to install the build dependencies for a package is to use `build-dep`:

```
$ sudo apt-get build-dep openvswitch
```

4. Configure and build Open vSwitch:

```
$ ./boot.sh
$ ./configure
$ make -j4
```

5. Try out the sandbox by running:

```
$ make sandbox
```

   You can exit the sandbox with `exit` or Control+D.

3.1.2 Setting up Faucet

This section explains how to get a copy of Faucet and set it up appropriately for the tutorial. There are many other ways to install Faucet, but this simple approach worked well for me. It has the advantage that it does not require modifying any system-level files or directories on your machine. It does, on the other hand, require Docker, so make sure you have it installed and working.

It will be a little easier to go through the rest of the tutorial if you run these instructions in a separate terminal from the one that you're using for Open vSwitch, because it's often necessary to switch between one and the other.

1. Get a copy of the Faucet source repository using Git, then `cd` into the new directory:

```
$ git clone https://github.com/faucetsdn/faucet.git
$ cd faucet
```

   At this point I checked out the latest tag:

```
$ latest_tag=$(git describe --tags $(git rev-list --tags --max-count=1))
$ git checkout $latest_tag
```

2. Build a docker container image:

```
$ sudo docker build -t faucet/faucet -f Dockerfile.faucet .
```

   This will take a few minutes.

3. Create an installation directory under the `faucet` directory for the docker image to use:

```
$ mkdir inst
```

   The Faucet configuration will go in `inst/faucet.yaml` and its main log will appear in `inst/faucet.log`. (The official Faucet installation instructions call to put these in `/etc/ryu/faucet` and `/var/log/ryu/faucet`, respectively, but we avoid modifying these system directories.)

4. Create a container and start Faucet:

```
$ sudo docker run -d --name faucet --restart=always -v $(pwd)/inst/:/etc/faucet/ -v
↪$(pwd)/inst/:/var/log/faucet/ -p 6653:6653 -p 9302:9302 faucet/faucet
```

5. Look in `inst/faucet.log` to verify that Faucet started. It will probably start with an exception and traceback because we have not yet created `inst/faucet.yaml`.

6. Later on, to make a new or updated Faucet configuration take effect quickly, you can run:

```
$ sudo docker exec faucet pkill -HUP -f faucet.faucet
```

   Another way is to stop and start the Faucet container:

```
$ sudo docker restart faucet
```

   You can also stop and delete the container; after this, to start it again, you need to rerun the `docker run` command:

```
$ sudo docker stop faucet
$ sudo docker rm faucet
```

3.1.3 Overview

Now that Open vSwitch and Faucet are ready, here's an overview of what we're going to do for the remainder of the tutorial:

1. Switching: Set up an L2 network with Faucet.
2. Routing: Route between multiple L3 networks with Faucet.
3. ACLs: Add and modify access control rules.

At each step, we will take a look at how the features in question work from Faucet at the top to the data plane layer at the bottom. From the highest to lowest level, these layers and the software components that connect them are:

Faucet.
: As the top level in the system, this is the authoritative source of the network configuration.

  Faucet connects to a variety of monitoring and performance tools, but we won't use them in this tutorial. Our main insights into the system will be through `faucet.yaml` for configuration and `faucet.log` to observe state, such as MAC learning and ARP resolution, and to tell when we've screwed up configuration syntax or semantics.

The OpenFlow subsystem in Open vSwitch.
: OpenFlow is the protocol, standardized by the Open Networking Foundation, that controllers like Faucet use to control how Open vSwitch and other switches treat packets in the network.

  We will use `ovs-ofctl`, a utility that comes with Open vSwitch, to observe and occasionally modify Open vSwitch's OpenFlow behavior. We will also use `ovs-appctl`, a utility for communicating with `ovs-vswitchd` and other Open vSwitch daemons, to ask "what-if?" type questions.

In addition, the OVS sandbox by default raises the Open vSwitch logging level for OpenFlow high enough that we can learn a great deal about OpenFlow behavior simply by reading its log file.

Open vSwitch datapath.
This is essentially a cache designed to accelerate packet processing. Open vSwitch includes a few different datapaths, such as one based on the Linux kernel and a userspace-only datapath (sometimes called the "DPDK" datapath). The OVS sandbox uses the latter, but the principles behind it apply equally well to other datapaths.

At each step, we discuss how the design of each layer influences performance. We demonstrate how Open vSwitch features can be used to debug, troubleshoot, and understand the system as a whole.

3.1.4 Switching

Layer-2 (L2) switching is the basis of modern networking. It's also very simple and a good place to start, so let's set up a switch with some VLANs in Faucet and see how it works at each layer. Begin by putting the following into `inst/faucet.yaml`:

```
dps:
    switch-1:
        dp_id: 0x1
        timeout: 3600
        arp_neighbor_timeout: 3600
        interfaces:
            1:
                native_vlan: 100
            2:
                native_vlan: 100
            3:
                native_vlan: 100
            4:
                native_vlan: 200
            5:
                native_vlan: 200
vlans:
    100:
    200:
```

This configuration file defines a single switch ("datapath" or "dp") named `switch-1`. The switch has five ports, numbered 1 through 5. Ports 1, 2, and 3 are in VLAN 100, and ports 4 and 5 are in VLAN 2. Faucet can identify the switch from its datapath ID, which is defined to be 0x1.

> **Note**
>
> This also sets high MAC learning and ARP timeouts. The defaults are 5 minutes and about 8 minutes, which are fine in production but sometimes too fast for manual experimentation.

Now restart Faucet so that the configuration takes effect, e.g.:

```
$ sudo docker restart faucet
```

Assuming that the configuration update is successful, you should now see a new line at the end of `inst/faucet.log`:

```
Sep 10 06:44:10 faucet INFO     Add new datapath DPID 1 (0x1)
```

Faucet is now waiting for a switch with datapath ID 0x1 to connect to it over OpenFlow, so our next step is to create a switch with OVS and make it connect to Faucet. To do that, switch to the terminal where you checked out OVS and start a sandbox with `make sandbox` or `tutorial/ovs-sandbox` (as explained earlier under *Setting Up OVS*). You should see something like this toward the end of the output:

```
----------------------------------------------------------------------
You are running in a dummy Open vSwitch environment.  You can use
ovs-vsctl, ovs-ofctl, ovs-appctl, and other tools to work with the
dummy switch.

Log files, pidfiles, and the configuration database are in the
"sandbox" subdirectory.

Exit the shell to kill the running daemons.
blp@sigabrt:~/nicira/ovs/tutorial(0)$
```

Inside the sandbox, create a switch ("bridge") named `br0`, set its datapath ID to 0x1, add simulated ports to it named `p1` through `p5`, and tell it to connect to the Faucet controller. To make it easier to understand, we request for port `p1` to be assigned OpenFlow port 1, `p2` port 2, and so on. As a final touch, configure the controller to be "out-of-band" (this is mainly to avoid some annoying messages in the `ovs-vswitchd` logs; for more information, run `man ovs-vswitchd.conf.db` and search for `connection_mode`):

```
$ ovs-vsctl add-br br0 \
        -- set bridge br0 other-config:datapath-id=0000000000000001 \
        -- add-port br0 p1 -- set interface p1 ofport_request=1 \
        -- add-port br0 p2 -- set interface p2 ofport_request=2 \
        -- add-port br0 p3 -- set interface p3 ofport_request=3 \
        -- add-port br0 p4 -- set interface p4 ofport_request=4 \
        -- add-port br0 p5 -- set interface p5 ofport_request=5 \
        -- set-controller br0 tcp:127.0.0.1:6653 \
        -- set controller br0 connection-mode=out-of-band
```

Note

You don't have to run all of these as a single `ovs-vsctl` invocation. It is a little more efficient, though, and since it updates the OVS configuration in a single database transaction it means that, for example, there is never a time when the controller is set but it has not yet been configured as out-of-band.

Faucet requires ports to be in the up state before it will configure them. In Open vSwitch versions earlier than 2.11.0 dummy ports started in the down state. You will need to force them to come up with the following `ovs-appctl` command (please skip this step if using a newer version of Open vSwitch):

```
$ ovs-appctl netdev-dummy/set-admin-state up
```

Now, if you look at `inst/faucet.log` again, you should see that Faucet recognized and configured the new switch and its ports:

```
Sep 10 06:45:03 faucet.valve INFO     DPID 1 (0x1) switch-1 Cold start configuring DP
Sep 10 06:45:03 faucet.valve INFO     DPID 1 (0x1) switch-1 Configuring VLAN 100 vid:100␣
↪ports:Port 1,Port 2,Port 3
Sep 10 06:45:03 faucet.valve INFO     DPID 1 (0x1) switch-1 Configuring VLAN 200 vid:200␣
↪ports:Port 4,Port 5
Sep 10 06:45:24 faucet.valve INFO     DPID 1 (0x1) switch-1 Port 1 (1) up
```

(continues on next page)

(continued from previous page)

```
Sep 10 06:45:24 faucet.valve INFO     DPID 1 (0x1) switch-1 Port 2 (2) up
Sep 10 06:45:24 faucet.valve INFO     DPID 1 (0x1) switch-1 Port 3 (3) up
Sep 10 06:45:24 faucet.valve INFO     DPID 1 (0x1) switch-1 Port 4 (4) up
Sep 10 06:45:24 faucet.valve INFO     DPID 1 (0x1) switch-1 Port 5 (5) up
```

Over on the Open vSwitch side, you can see a lot of related activity if you take a look in `sandbox/ovs-vswitchd.log`. For example, here is the basic OpenFlow session setup and Faucet's probe of the switch's ports and capabilities:

```
rconn|INFO|br0<->tcp:127.0.0.1:6653: connecting...
vconn|DBG|tcp:127.0.0.1:6653: sent (Success): OFPT_HELLO (OF1.4) (xid=0x1):
 version bitmap: 0x01, 0x02, 0x03, 0x04, 0x05
vconn|DBG|tcp:127.0.0.1:6653: received: OFPT_HELLO (OF1.3) (xid=0xdb9dab08):
 version bitmap: 0x01, 0x02, 0x03, 0x04
vconn|DBG|tcp:127.0.0.1:6653: negotiated OpenFlow version 0x04 (we support version 0x05␣
↪and earlier, peer supports version 0x04 and earlier)
rconn|INFO|br0<->tcp:127.0.0.1:6653: connected
vconn|DBG|tcp:127.0.0.1:6653: received: OFPT_FEATURES_REQUEST (OF1.3) (xid=0xdb9dab09):
00040|vconn|DBG|tcp:127.0.0.1:6653: sent (Success): OFPT_FEATURES_REPLY (OF1.3)␣
↪(xid=0xdb9dab09): dpid:0000000000000001
n_tables:254, n_buffers:0
capabilities: FLOW_STATS TABLE_STATS PORT_STATS GROUP_STATS QUEUE_STATS
vconn|DBG|tcp:127.0.0.1:6653: received: OFPST_PORT_DESC request (OF1.3)␣
↪(xid=0xdb9dab0a): port=ANY
vconn|DBG|tcp:127.0.0.1:6653: sent (Success): OFPST_PORT_DESC reply (OF1.3)␣
↪(xid=0xdb9dab0a):
 1(p1): addr:aa:55:aa:55:00:14
     config:     0
     state:      LIVE
     speed: 0 Mbps now, 0 Mbps max
 2(p2): addr:aa:55:aa:55:00:15
     config:     0
     state:      LIVE
     speed: 0 Mbps now, 0 Mbps max
 3(p3): addr:aa:55:aa:55:00:16
     config:     0
     state:      LIVE
     speed: 0 Mbps now, 0 Mbps max
 4(p4): addr:aa:55:aa:55:00:17
     config:     0
     state:      LIVE
     speed: 0 Mbps now, 0 Mbps max
 5(p5): addr:aa:55:aa:55:00:18
     config:     0
     state:      LIVE
     speed: 0 Mbps now, 0 Mbps max
 LOCAL(br0): addr:42:51:a1:c4:97:45
     config:     0
     state:      LIVE
     speed: 0 Mbps now, 0 Mbps max
```

After that, you can see Faucet delete all existing flows and then start adding new ones:

```
vconn|DBG|tcp:127.0.0.1:6653: received: OFPT_FLOW_MOD (OF1.3) (xid=0xdb9dab0f): DEL␣
↪table:255 priority=0 actions=drop
vconn|DBG|tcp:127.0.0.1:6653: received: OFPT_FLOW_MOD (OF1.3) (xid=0xdb9dab10): ADD␣
↪priority=0 cookie:0x5adc15c0 out_port:0 actions=drop
vconn|DBG|tcp:127.0.0.1:6653: received: OFPT_FLOW_MOD (OF1.3) (xid=0xdb9dab11): ADD␣
↪table:1 priority=0 cookie:0x5adc15c0 out_port:0 actions=goto_table:2
vconn|DBG|tcp:127.0.0.1:6653: received: OFPT_FLOW_MOD (OF1.3) (xid=0xdb9dab12): ADD␣
↪table:2 priority=0 cookie:0x5adc15c0 out_port:0 actions=goto_table:3
...
```

OpenFlow Layer

Let's take a look at the OpenFlow tables that Faucet set up. Before we do that, it's helpful to take a look at `docs/architecture.rst` in the Faucet documentation to learn how Faucet structures its flow tables. In summary, this document says that when all features are enabled our table layout will be:

Table 0
: Port-based ACLs

Table 1
: Ingress VLAN processing

Table 2
: VLAN-based ACLs

Table 3
: Ingress L2 processing, MAC learning

Table 4
: L3 forwarding for IPv4

Table 5
: L3 forwarding for IPv6

Table 6
: Virtual IP processing, e.g. for router IP addresses implemented by Faucet

Table 7
: Egress L2 processing

Table 8
: Flooding

With that in mind, let's dump the flow tables. The simplest way is to just run plain `ovs-ofctl dump-flows`:

```
$ ovs-ofctl dump-flows br0
```

If you run that bare command, it produces a lot of extra junk that makes the output harder to read, like statistics and "cookie" values that are all the same. In addition, for historical reasons `ovs-ofctl` always defaults to using OpenFlow 1.0 even though Faucet and most modern controllers use OpenFlow 1.3, so it's best to force it to use OpenFlow 1.3. We could throw in a lot of options to fix these, but we'll want to do this more than once, so let's start by defining a shell function for ourselves:

```
$ dump-flows () {
  ovs-ofctl -OOpenFlow13 --names --no-stat dump-flows "$@" \
    | sed 's/cookie=0x5adc15c0, //'
}
```

Let's also define `save-flows` and `diff-flows` functions for later use:

```
$ save-flows () {
  ovs-ofctl -OOpenFlow13 --no-names --sort dump-flows "$@"
}
$ diff-flows () {
  ovs-ofctl -OOpenFlow13 diff-flows "$@" | sed 's/cookie=0x5adc15c0 //'
}
```

Now let's take a look at the flows we've got and what they mean, like this:

```
$ dump-flows br0
```

To reduce resource utilisation on hardware switches, Faucet will try to install the minimal set of OpenFlow tables to match the features enabled in `faucet.yaml`. Since we have only enabled switching we will end up with 4 tables. If we inspect the contents of `inst/faucet.log` Faucet will tell us what each table does:

```
Sep 10 06:44:10 faucet.valve INFO     DPID 1 (0x1) switch-1 table ID 0 table config dec_
↪ttl: None exact_match: None match_types: (('eth_dst', True), ('eth_type', False), ('in_
↪port', False), ('vlan_vid', False)) meter: None miss_goto: None name: vlan next_
↪tables: ['eth_src'] output: True set_fields: ('vlan_vid',) size: 32 table_id: 0 vlan_
↪port_scale: 1.5
Sep 10 06:44:10 faucet.valve INFO     DPID 1 (0x1) switch-1 table ID 1 table config dec_
↪ttl: None exact_match: None match_types: (('eth_dst', True), ('eth_src', False), ('eth_
↪type', False), ('in_port', False), ('vlan_vid', False)) meter: None miss_goto: eth_dst␣
↪name: eth_src next_tables: ['eth_dst', 'flood'] output: True set_fields: ('vlan_vid',
↪'eth_dst') size: 32 table_id: 1 vlan_port_scale: 4.1
Sep 10 06:44:10 faucet.valve INFO     DPID 1 (0x1) switch-1 table ID 2 table config dec_
↪ttl: None exact_match: True match_types: (('eth_dst', False), ('vlan_vid', False))␣
↪meter: None miss_goto: flood name: eth_dst next_tables: [] output: True set_fields:␣
↪None size: 41 table_id: 2 vlan_port_scale: 4.1
Sep 10 06:44:10 faucet.valve INFO     DPID 1 (0x1) switch-1 table ID 3 table config dec_
↪ttl: None exact_match: None match_types: (('eth_dst', True), ('in_port', False), (
↪'vlan_vid', False)) meter: None miss_goto: None name: flood next_tables: [] output:␣
↪True set_fields: None size: 32 table_id: 3 vlan_port_scale: 2.1
```

Currently, we have:

Table 0 (vlan)
: Ingress VLAN processing

Table 1 (eth_src)
: Ingress L2 processing, MAC learning

Table 2 (eth_dst)
: Egress L2 processing

Table 3 (flood)
: Flooding

In Table 0 we see flows that recognize packets without a VLAN header on each of our ports (`vlan_tci=0x0000/0x1fff`), push on the VLAN configured for the port, and proceed to table 3. There is also a fallback flow to drop other packets, which in practice means that if any received packet already has a VLAN header then it will be dropped:

```
priority=9000,in_port=p1,vlan_tci=0x0000/0x1fff actions=push_vlan:0x8100,set_field:4196->
↪vlan_vid,goto_table:1
priority=9000,in_port=p2,vlan_tci=0x0000/0x1fff actions=push_vlan:0x8100,set_field:4196->
↪vlan_vid,goto_table:1
```

(continues on next page)

(continued from previous page)

```
priority=9000,in_port=p3,vlan_tci=0x0000/0x1fff actions=push_vlan:0x8100,set_field:4196->
↪vlan_vid,goto_table:1
priority=9000,in_port=p4,vlan_tci=0x0000/0x1fff actions=push_vlan:0x8100,set_field:4296->
↪vlan_vid,goto_table:1
priority=9000,in_port=p5,vlan_tci=0x0000/0x1fff actions=push_vlan:0x8100,set_field:4296->
↪vlan_vid,goto_table:1
priority=0 actions=drop
```

Note

The syntax `set_field:4196->vlan_vid` is curious and somewhat misleading. OpenFlow 1.3 defines the `vlan_vid` field as a 13-bit field where bit 12 is set to 1 if the VLAN header is present. Thus, since 4196 is 0x1064, this action sets VLAN value 0x64, which in decimal is 100.

Table 1 starts off with a flow that drops some inappropriate packets, in this case EtherType 0x9000 (Ethernet Configuration Testing Protocol), which should not be forwarded by a switch:

```
table=1, priority=9099,dl_type=0x9000 actions=drop
```

Table 1 is primarily used for MAC learning but the controller hasn't learned any MAC addresses yet. It also drops some more inappropriate packets such as those that claim to be from a broadcast source address (why not from all multicast source addresses, though?). We'll come back here later:

```
table=1, priority=9099,dl_src=ff:ff:ff:ff:ff:ff actions=drop
table=1, priority=9001,dl_src=0e:00:00:00:00:01 actions=drop
table=1, priority=9000,dl_vlan=100 actions=CONTROLLER:96,goto_table:2
table=1, priority=9000,dl_vlan=200 actions=CONTROLLER:96,goto_table:2
table=1, priority=0 actions=goto_table:2
```

Table 2 is used to direct packets to learned MACs but Faucet hasn't learned any MACs yet, so it just sends all the packets along to table 3:

```
table=2, priority=0 actions=goto_table:3
```

Table 3 does some more dropping of packets we don't want to forward, in this case STP:

```
table=3, priority=9099,dl_dst=01:00:0c:cc:cc:cd actions=drop
table=3, priority=9099,dl_dst=01:80:c2:00:00:00/ff:ff:ff:ff:ff:f0 actions=drop
```

Table 3 implements flooding, broadcast, and multicast. The flows for broadcast and flood are easy to understand: if the packet came in on a given port and needs to be flooded or broadcast, output it to all the other ports in the same VLAN:

```
table=3, priority=9004,dl_vlan=100,dl_dst=ff:ff:ff:ff:ff:ff actions=pop_vlan,output:p1,
↪output:p2,output:p3
table=3, priority=9004,dl_vlan=200,dl_dst=ff:ff:ff:ff:ff:ff actions=pop_vlan,output:p4,
↪output:p5
table=3, priority=9000,dl_vlan=100 actions=pop_vlan,output:p1,output:p2,output:p3
table=3, priority=9000,dl_vlan=200 actions=pop_vlan,output:p4,output:p5
```

There are also some flows for handling some standard forms of multicast, and a fallback drop flow:

```
table=3, priority=9003,dl_vlan=100,dl_dst=33:33:00:00:00:00/ff:ff:00:00:00:00␣
↪actions=pop_vlan,output:p1,output:p2,output:p3
table=3, priority=9003,dl_vlan=200,dl_dst=33:33:00:00:00:00/ff:ff:00:00:00:00␣
↪actions=pop_vlan,output:p4,output:p5
table=3, priority=9001,dl_vlan=100,dl_dst=01:80:c2:00:00:00/ff:ff:ff:00:00:00␣
↪actions=pop_vlan,output:p1,output:p2,output:p3
table=3, priority=9002,dl_vlan=100,dl_dst=01:00:5e:00:00:00/ff:ff:ff:00:00:00␣
↪actions=pop_vlan,output:p1,output:p2,output:p3
table=3, priority=9001,dl_vlan=200,dl_dst=01:80:c2:00:00:00/ff:ff:ff:00:00:00␣
↪actions=pop_vlan,output:p4,output:p5
table=3, priority=9002,dl_vlan=200,dl_dst=01:00:5e:00:00:00/ff:ff:ff:00:00:00␣
↪actions=pop_vlan,output:p4,output:p5
table=3, priority=0 actions=drop
```

Tracing

Let's go a level deeper. So far, everything we've done has been fairly general. We can also look at something more specific: the path that a particular packet would take through Open vSwitch. We can use the `ofproto/trace` command to play "what-if?" games. This command is one that we send directly to `ovs-vswitchd`, using the `ovs-appctl` utility.

> **Note**
>
> `ovs-appctl` is actually a very simple-minded JSON-RPC client, so you could also use some other utility that speaks JSON-RPC, or access it from a program as an API.

The `ovs-vswitchd`(8) manpage has a lot of detail on how to use `ofproto/trace`, but let's just start by building up from a simple example. You can start with a command that just specifies the datapath (e.g. `br0`), an input port, and nothing else; unspecified fields default to all-zeros. Let's look at the full output for this trivial example:

```
$ ovs-appctl ofproto/trace br0 in_port=p1
Flow: in_port=1,vlan_tci=0x0000,dl_src=00:00:00:00:00:00,dl_dst=00:00:00:00:00:00,dl_
↪type=0x0000

bridge("br0")
-------------
 0. in_port=1,vlan_tci=0x0000/0x1fff, priority 9000, cookie 0x5adc15c0
    push_vlan:0x8100
    set_field:4196->vlan_vid
    goto_table:1
 1. dl_vlan=100, priority 9000, cookie 0x5adc15c0
    CONTROLLER:96
    goto_table:2
 2. priority 0, cookie 0x5adc15c0
    goto_table:3
 3. dl_vlan=100, priority 9000, cookie 0x5adc15c0
    pop_vlan
    output:1
     >> skipping output to input port
    output:2
    output:3

Final flow: unchanged
```

(continues on next page)

(continued from previous page)

```
Megaflow: recirc_id=0,eth,in_port=1,vlan_tci=0x0000,dl_src=00:00:00:00:00:00,dl_
↪dst=00:00:00:00:00:00,dl_type=0x0000
Datapath actions: push_vlan(vid=100,pcp=0),userspace(pid=0,controller(reason=1,dont_
↪send=1,continuation=0,recirc_id=1,rule_cookie=0x5adc15c0,controller_id=0,max_len=96)),
↪pop_vlan,2,3
```

The first line of output, beginning with `Flow:`, just repeats our request in a more verbose form, including the L2 fields that were zeroed.

Each of the numbered items under `bridge("br0")` shows what would happen to our hypothetical packet in the table with the given number. For example, we see in table 0 that the packet matches a flow that push on a VLAN header, set the VLAN ID to 100, and goes on to further processing in table 1. In table 1, the packet gets sent to the controller to allow MAC learning to take place, and then table 3 floods the packet to the other ports in the same VLAN.

Summary information follows the numbered tables. The packet hasn't been changed (overall, even though a VLAN was pushed and then popped back off) since ingress, hence `Final flow: unchanged`. We'll look at the `Megaflow` information later. The `Datapath actions` summarize what would actually happen to such a packet.

Triggering MAC Learning

We just saw how a packet gets sent to the controller to trigger MAC learning. Let's actually send the packet and see what happens. But before we do that, let's save a copy of the current flow tables for later comparison:

```
$ save-flows br0 > flows1
```

Now use `ofproto/trace`, as before, with a few new twists: we specify the source and destination Ethernet addresses and append the `-generate` option so that side effects like sending a packet to the controller actually happen:

```
$ ovs-appctl ofproto/trace br0 in_port=p1,dl_src=00:11:11:00:00:00,dl_
↪dst=00:22:22:00:00:00 -generate
```

The output is almost identical to that before, so it is not repeated here. But, take a look at `inst/faucet.log` now. It should now include a line at the end that says that it learned about our MAC 00:11:11:00:00:00, like this:

```
Sep 10 08:16:28 faucet.valve INFO     DPID 1 (0x1) switch-1 L2 learned 00:11:11:00:00:00␣
↪(L2 type 0x0000, L3 src None, L3 dst None) Port 1 VLAN 100  (1 hosts total)
```

Now compare the flow tables that we saved to the current ones:

```
diff-flows flows1 br0
```

The result should look like this, showing new flows for the learned MACs:

```
+table=1 priority=9098,in_port=1,dl_vlan=100,dl_src=00:11:11:00:00:00 hard_timeout=3605␣
↪actions=goto_table:2
+table=2 priority=9099,dl_vlan=100,dl_dst=00:11:11:00:00:00 idle_timeout=3605␣
↪actions=pop_vlan,output:1
```

To demonstrate the usefulness of the learned MAC, try tracing (with side effects) a packet arriving on `p2` (or `p3`) and destined to the address learned on `p1`, like this:

```
$ ovs-appctl ofproto/trace br0 in_port=p2,dl_src=00:22:22:00:00:00,dl_
↪dst=00:11:11:00:00:00 -generate
```

The first time you run this command, you will notice that it sends the packet to the controller, to learn `p2`'s 00:22:22:00:00:00 source address:

```
bridge("br0")
-------------
 0. in_port=2,vlan_tci=0x0000/0x1fff, priority 9000, cookie 0x5adc15c0
    push_vlan:0x8100
    set_field:4196->vlan_vid
    goto_table:1
 1. dl_vlan=100, priority 9000, cookie 0x5adc15c0
    CONTROLLER:96
    goto_table:2
 2. dl_vlan=100,dl_dst=00:11:11:00:00:00, priority 9099, cookie 0x5adc15c0
    pop_vlan
    output:1
```

If you check `inst/faucet.log`, you can see that `p2`'s MAC has been learned too:

```
Sep 10 08:17:45 faucet.valve INFO     DPID 1 (0x1) switch-1 L2 learned 00:22:22:00:00:00␣
↪(L2 type 0x0000, L3 src None, L3 dst None) Port 2 VLAN 100  (2 hosts total)
```

Similarly for `diff-flows`:

```
$ diff-flows flows1 br0
+table=1 priority=9098,in_port=1,dl_vlan=100,dl_src=00:11:11:00:00:00 hard_timeout=3605␣
↪actions=goto_table:2
+table=1 priority=9098,in_port=2,dl_vlan=100,dl_src=00:22:22:00:00:00 hard_timeout=3599␣
↪actions=goto_table:2
+table=2 priority=9099,dl_vlan=100,dl_dst=00:11:11:00:00:00 idle_timeout=3605␣
↪actions=pop_vlan,output:1
+table=2 priority=9099,dl_vlan=100,dl_dst=00:22:22:00:00:00 idle_timeout=3599␣
↪actions=pop_vlan,output:2
```

Then, if you re-run either of the `ofproto/trace` commands (with or without `-generate`), you can see that the packets go back and forth without any further MAC learning, e.g.:

```
$ ovs-appctl ofproto/trace br0 in_port=p2,dl_src=00:22:22:00:00:00,dl_
↪dst=00:11:11:00:00:00 -generate
Flow: in_port=2,vlan_tci=0x0000,dl_src=00:22:22:00:00:00,dl_dst=00:11:11:00:00:00,dl_
↪type=0x0000

bridge("br0")
-------------
 0. in_port=2,vlan_tci=0x0000/0x1fff, priority 9000, cookie 0x5adc15c0
    push_vlan:0x8100
    set_field:4196->vlan_vid
    goto_table:1
 1. in_port=2,dl_vlan=100,dl_src=00:22:22:00:00:00, priority 9098, cookie 0x5adc15c0
    goto_table:2
 2. dl_vlan=100,dl_dst=00:11:11:00:00:00, priority 9099, cookie 0x5adc15c0
    pop_vlan
    output:1

Final flow: unchanged
Megaflow: recirc_id=0,eth,in_port=2,vlan_tci=0x0000/0x1fff,dl_src=00:22:22:00:00:00,dl_
↪dst=00:11:11:00:00:00,dl_type=0x0000
Datapath actions: 1
```

Performance

Open vSwitch has a concept of a "fast path" and a "slow path"; ideally all packets stay in the fast path. This distinction between slow path and fast path is the key to making sure that Open vSwitch performs as fast as possible.

Some factors can force a flow or a packet to take the slow path. As one example, all CFM, BFD, LACP, STP, and LLDP processing takes place in the slow path, in the cases where Open vSwitch processes these protocols itself instead of delegating to controller-written flows. As a second example, any flow that modifies ARP fields is processed in the slow path. These are corner cases that are unlikely to cause performance problems in practice because these protocols send packets at a relatively slow rate, and users and controller authors do not normally need to be concerned about them.

To understand what cases users and controller authors should consider, we need to talk about how Open vSwitch optimizes for performance. The Open vSwitch code is divided into two major components which, as already mentioned, are called the "slow path" and "fast path" (aka "datapath"). The slow path is embedded in the `ovs-vswitchd` userspace program. It is the part of the Open vSwitch packet processing logic that understands OpenFlow. Its job is to take a packet and run it through the OpenFlow tables to determine what should happen to it. It outputs a list of actions in a form similar to OpenFlow actions but simpler, called "ODP actions" or "datapath actions". It then passes the ODP actions to the datapath, which applies them to the packet.

> **Note**
>
> Open vSwitch contains a single slow path and multiple fast paths. The difference between using Open vSwitch with the Linux kernel versus with DPDK is the datapath.

If every packet passed through the slow path and the fast path in this way, performance would be terrible. The key to getting high performance from this architecture is caching. Open vSwitch includes a multi-level cache. It works like this:

1. A packet initially arrives at the datapath. Some datapaths, such as DPDK, have a first-level cache called the "microflow cache". The microflow cache is the key to performance for relatively long-lived, high packet rate flows. If the datapath has a microflow cache, then it consults it and, if there is a cache hit, the datapath executes the associated actions. Otherwise, it proceeds to step 2.

2. The datapath consults its second-level cache, called the "megaflow cache". The megaflow cache is the key to performance for shorter or low packet rate flows. If there is a megaflow cache hit, the datapath executes the associated actions. Otherwise, it proceeds to step 3.

3. The datapath passes the packet to the slow path, which runs it through the OpenFlow table to yield ODP actions, a process that is often called "flow translation". It then passes the packet back to the datapath to execute the actions and to, if possible, install a megaflow cache entry so that subsequent similar packets can be handled directly by the fast path. (We already described above most of the cases where a cache entry cannot be installed.)

The megaflow cache is the key cache to consider for performance tuning. Open vSwitch provides tools for understanding and optimizing its behavior. The `ofproto/trace` command that we have already been using is the most common tool for this use. Let's take another look at the most recent `ofproto/trace` output:

```
$ ovs-appctl ofproto/trace br0 in_port=p2,dl_src=00:22:22:00:00:00,dl_
↪dst=00:11:11:00:00:00 -generate
Flow: in_port=2,vlan_tci=0x0000,dl_src=00:22:22:00:00:00,dl_dst=00:11:11:00:00:00,dl_
↪type=0x0000

bridge("br0")
-------------
 0. in_port=2,vlan_tci=0x0000/0x1fff, priority 9000, cookie 0x5adc15c0
    push_vlan:0x8100
    set_field:4196->vlan_vid
```

(continues on next page)

(continued from previous page)

```
    goto_table:1
 1. in_port=2,dl_vlan=100,dl_src=00:22:22:00:00:00, priority 9098, cookie 0x5adc15c0
    goto_table:2
 2. dl_vlan=100,dl_dst=00:11:11:00:00:00, priority 9099, cookie 0x5adc15c0
    pop_vlan
    output:1

Final flow: unchanged
Megaflow: recirc_id=0,eth,in_port=2,vlan_tci=0x0000/0x1fff,dl_src=00:22:22:00:00:00,dl_
↪dst=00:11:11:00:00:00,dl_type=0x0000
Datapath actions: 1
```

This time, it's the last line that we're interested in. This line shows the entry that Open vSwitch would insert into the megaflow cache given the particular packet with the current flow tables. The megaflow entry includes:

- `recirc_id`. This is an implementation detail that users don't normally need to understand.
- `eth`. This just indicates that the cache entry matches only Ethernet packets; Open vSwitch also supports other types of packets, such as IP packets not encapsulated in Ethernet.
- All of the fields matched by any of the flows that the packet visited:

  `in_port`
  : In tables 0 and 1.

  `vlan_tci`
  : In tables 0, 1, and 2 (`vlan_tci` includes the VLAN ID and PCP fields and``dl_vlan`` is just the VLAN ID).

  `dl_src`
  : In table 1.

  `dl_dst`
  : In table 2.

- All of the fields matched by flows that had to be ruled out to ensure that the ones that actually matched were the highest priority matching rules.

The last one is important. Notice how the megaflow matches on `dl_type=0x0000`, even though none of the tables matched on `dl_type` (the Ethernet type). One reason is because of this flow in OpenFlow table 1 (which shows up in `dump-flows` output):

```
table=1, priority=9099,dl_type=0x9000 actions=drop
```

This flow has higher priority than the flow in table 1 that actually matched. This means that, to put it in the megaflow cache, `ovs-vswitchd` has to add a match on `dl_type` to ensure that the cache entry doesn't match ECTP packets (with Ethertype 0x9000).

> **Note**
>
> In fact, in some cases `ovs-vswitchd` matches on fields that aren't strictly required according to this description. `dl_type` is actually one of those, so deleting the LLDP flow probably would not have any effect on the megaflow. But the principle here is sound.

So why does any of this matter? It's because, the more specific a megaflow is, that is, the more fields or bits within fields that a megaflow matches, the less valuable it is from a caching viewpoint. A very specific megaflow might match

on L2 and L3 addresses and L4 port numbers. When that happens, only packets in one (half-)connection match the megaflow. If that connection has only a few packets, as many connections do, then the high cost of the slow path translation is amortized over only a few packets, so the average cost of forwarding those packets is high. On the other hand, if a megaflow only matches a relatively small number of L2 and L3 packets, then the cache entry can potentially be used by many individual connections, and the average cost is low.

For more information on how Open vSwitch constructs megaflows, including about ways that it can make megaflow entries less specific than one would infer from the discussion here, please refer to the 2015 NSDI paper, "The Design and Implementation of Open vSwitch", which focuses on this algorithm.

3.1.5 Routing

We've looked at how Faucet implements switching in OpenFlow, and how Open vSwitch implements OpenFlow through its datapath architecture. Now let's start over, adding L3 routing into the picture.

It's remarkably easy to enable routing. We just change our `vlans` section in `inst/faucet.yaml` to specify a router IP address for each VLAN and define a router between them. The `dps` section is unchanged:

```
dps:
    switch-1:
        dp_id: 0x1
        timeout: 3600
        arp_neighbor_timeout: 3600
        interfaces:
            1:
                native_vlan: 100
            2:
                native_vlan: 100
            3:
                native_vlan: 100
            4:
                native_vlan: 200
            5:
                native_vlan: 200
vlans:
    100:
        faucet_vips: ["10.100.0.254/24"]
    200:
        faucet_vips: ["10.200.0.254/24"]
routers:
    router-1:
        vlans: [100, 200]
```

Then we can tell Faucet to reload its configuration:

```
$ sudo docker exec faucet pkill -HUP -f faucet.faucet
```

OpenFlow Layer

Now that we have an additional feature enabled (routing) we will notice some additional OpenFlow tables if we check `inst/faucet.log`:

```
Sep 10 08:28:14 faucet.valve INFO     DPID 1 (0x1) switch-1 table ID 0 table config dec_
↪ttl: None exact_match: None match_types: (('eth_dst', True), ('eth_type', False), ('in_
↪port', False), ('vlan_vid', False)) meter: None miss_goto: None name: vlan next_
```

(continues on next page)

(continued from previous page)

```
↪tables: ['eth_src'] output: True set_fields: ('vlan_vid',) size: 32 table_id: 0 vlan_
↪port_scale: 1.5
Sep 10 08:28:14 faucet.valve INFO     DPID 1 (0x1) switch-1 table ID 1 table config dec_
↪ttl: None exact_match: None match_types: (('eth_dst', True), ('eth_src', False), ('eth_
↪type', False), ('in_port', False), ('vlan_vid', False)) meter: None miss_goto: eth_dst␣
↪name: eth_src next_tables: ['ipv4_fib', 'vip', 'eth_dst', 'flood'] output: True set_
↪fields: ('vlan_vid', 'eth_dst') size: 32 table_id: 1 vlan_port_scale: 4.1
Sep 10 08:28:14 faucet.valve INFO     DPID 1 (0x1) switch-1 table ID 2 table config dec_
↪ttl: True exact_match: None match_types: (('eth_type', False), ('ipv4_dst', True), (
↪'vlan_vid', False)) meter: None miss_goto: None name: ipv4_fib next_tables: ['vip',
↪'eth_dst', 'flood'] output: True set_fields: ('eth_dst', 'eth_src', 'vlan_vid') size:␣
↪32 table_id: 2 vlan_port_scale: 3.1
Sep 10 08:28:14 faucet.valve INFO     DPID 1 (0x1) switch-1 table ID 3 table config dec_
↪ttl: None exact_match: None match_types: (('arp_tpa', False), ('eth_dst', False), (
↪'eth_type', False), ('icmpv6_type', False), ('ip_proto', False)) meter: None miss_
↪goto: None name: vip next_tables: ['eth_dst', 'flood'] output: True set_fields: None␣
↪size: 32 table_id: 3 vlan_port_scale: None
Sep 10 08:28:14 faucet.valve INFO     DPID 1 (0x1) switch-1 table ID 4 table config dec_
↪ttl: None exact_match: True match_types: (('eth_dst', False), ('vlan_vid', False))␣
↪meter: None miss_goto: flood name: eth_dst next_tables: [] output: True set_fields:␣
↪None size: 41 table_id: 4 vlan_port_scale: 4.1
Sep 10 08:28:14 faucet.valve INFO     DPID 1 (0x1) switch-1 table ID 5 table config dec_
↪ttl: None exact_match: None match_types: (('eth_dst', True), ('in_port', False), (
↪'vlan_vid', False)) meter: None miss_goto: None name: flood next_tables: [] output:␣
↪True set_fields: None size: 32 table_id: 5 vlan_port_scale: 2.1
```

So now we have an additional FIB and VIP table:

Table 0 (vlan)
: Ingress VLAN processing

Table 1 (eth_src)
: Ingress L2 processing, MAC learning

Table 2 (ipv4_fib)
: L3 forwarding for IPv4

Table 3 (vip)
: Virtual IP processing, e.g. for router IP addresses implemented by Faucet

Table 4 (eth_dst)
: Egress L2 processing

Table 5 (flood)
: Flooding

Back in the OVS sandbox, let's see what new flow rules have been added, with:

```
$ diff-flows flows1 br0 | grep +
```

First, table 1 has new flows to direct ARP packets to table 3 (the virtual IP processing table), presumably to handle ARP for the router IPs. New flows also send IP packets destined to a particular Ethernet address to table 2 (the L3 forwarding table); we can make the educated guess that the Ethernet address is the one used by the Faucet router:

```
+table=1 priority=9131,arp,dl_vlan=100 actions=goto_table:3
+table=1 priority=9131,arp,dl_vlan=200 actions=goto_table:3
```

(continues on next page)

(continued from previous page)

```
+table=1 priority=9099,ip,dl_vlan=100,dl_dst=0e:00:00:00:00:01 actions=goto_table:2
+table=1 priority=9099,ip,dl_vlan=200,dl_dst=0e:00:00:00:00:01 actions=goto_table:2
```

In the new `ipv4_fib` table (table 2) there appear to be flows for verifying that the packets are indeed addressed to a network or IP address that Faucet knows how to route:

```
+table=2 priority=9131,ip,dl_vlan=100,nw_dst=10.100.0.254 actions=goto_table:3
+table=2 priority=9131,ip,dl_vlan=200,nw_dst=10.200.0.254 actions=goto_table:3
+table=2 priority=9123,ip,dl_vlan=200,nw_dst=10.100.0.0/24 actions=goto_table:3
+table=2 priority=9123,ip,dl_vlan=100,nw_dst=10.100.0.0/24 actions=goto_table:3
+table=2 priority=9123,ip,dl_vlan=200,nw_dst=10.200.0.0/24 actions=goto_table:3
+table=2 priority=9123,ip,dl_vlan=100,nw_dst=10.200.0.0/24 actions=goto_table:3
```

In our new `vip` table (table 3) there are a few different things going on. It sends ARP requests for the router IPs to the controller; presumably the controller will generate replies and send them back to the requester. It switches other ARP packets, either broadcasting them if they have a broadcast destination or attempting to unicast them otherwise. It sends all other IP packets to the controller:

```
+table=3 priority=9133,arp,arp_tpa=10.100.0.254 actions=CONTROLLER:128
+table=3 priority=9133,arp,arp_tpa=10.200.0.254 actions=CONTROLLER:128
+table=3 priority=9132,arp,dl_dst=ff:ff:ff:ff:ff:ff actions=goto_table:4
+table=3 priority=9131,arp actions=goto_table:4
+table=3 priority=9130,ip actions=CONTROLLER:128
```

Performance is clearly going to be poor if every packet that needs to be routed has to go to the controller, but it's unlikely that's the full story. In the next section, we'll take a closer look.

Tracing

As in our switching example, we can play some "what-if?" games to figure out how this works. Let's suppose that a machine with IP 10.100.0.1, on port `p1`, wants to send a IP packet to a machine with IP 10.200.0.1 on port `p4`. Assuming that these hosts have not been in communication recently, the steps to accomplish this are normally the following:

1. Host 10.100.0.1 sends an ARP request to router 10.100.0.254.
2. The router sends an ARP reply to the host.
3. Host 10.100.0.1 sends an IP packet to 10.200.0.1, via the router's Ethernet address.
4. The router broadcasts an ARP request to `p4` and `p5`, the ports that carry the 10.200.0.<x> network.
5. Host 10.200.0.1 sends an ARP reply to the router.
6. Either the router sends the IP packet (which it buffered) to 10.200.0.1, or eventually 10.100.0.1 times out and resends it.

Let's use `ofproto/trace` to see whether Faucet and OVS follow this procedure.

Before we start, save a new snapshot of the flow tables for later comparison:

```
$ save-flows br0 > flows2
```

Step 1: Host ARP for Router

Let's simulate the ARP from 10.100.0.1 to its gateway router 10.100.0.254. This requires more detail than any of the packets we've simulated previously:

```
$ ovs-appctl ofproto/trace br0 in_port=p1,dl_src=00:01:02:03:04:05,dl_
↪dst=ff:ff:ff:ff:ff:ff,dl_type=0x806,arp_spa=10.100.0.1,arp_tpa=10.100.0.254,arp_
↪sha=00:01:02:03:04:05,arp_tha=ff:ff:ff:ff:ff:ff,arp_op=1 -generate
```

The important part of the output is where it shows that the packet was recognized as an ARP request destined to the router gateway and therefore sent to the controller:

```
3. arp,arp_tpa=10.100.0.254, priority 9133, cookie 0x5adc15c0
  CONTROLLER:128
```

The Faucet log shows that Faucet learned the host's MAC address, its MAC-to-IP mapping, and responded to the ARP request:

```
Sep 10 08:52:46 faucet.valve INFO     DPID 1 (0x1) switch-1 Adding new route 10.100.0.1/
↪32 via 10.100.0.1 (00:01:02:03:04:05) on VLAN 100
Sep 10 08:52:46 faucet.valve INFO     DPID 1 (0x1) switch-1 Resolve response to 10.100.0.
↪254 from 00:01:02:03:04:05 (L2 type 0x0806, L3 src 10.100.0.1, L3 dst 10.100.0.254)␣
↪Port 1 VLAN 100
Sep 10 08:52:46 faucet.valve INFO     DPID 1 (0x1) switch-1 L2 learned 00:01:02:03:04:05␣
↪(L2 type 0x0806, L3 src 10.100.0.1, L3 dst 10.100.0.254) Port 1 VLAN 100  (1 hosts␣
↪total)
```

We can also look at the changes to the flow tables:

```
$ diff-flows flows2 br0
+table=1 priority=9098,in_port=1,dl_vlan=100,dl_src=00:01:02:03:04:05 hard_timeout=3605␣
↪actions=goto_table:4
+table=2 priority=9131,ip,dl_vlan=200,nw_dst=10.100.0.1 actions=set_field:4196->vlan_vid,
↪set_field:0e:00:00:00:00:01->eth_src,set_field:00:01:02:03:04:05->eth_dst,dec_ttl,goto_
↪table:4
+table=2 priority=9131,ip,dl_vlan=100,nw_dst=10.100.0.1 actions=set_field:4196->vlan_vid,
↪set_field:0e:00:00:00:00:01->eth_src,set_field:00:01:02:03:04:05->eth_dst,dec_ttl,goto_
↪table:4
+table=4 priority=9099,dl_vlan=100,dl_dst=00:01:02:03:04:05 idle_timeout=3605␣
↪actions=pop_vlan,output:1
```

The new flows include one in table 1 and one in table 4 for the learned MAC, which have the same forms we saw before. The new flows in table 2 are different. They matches packets directed to 10.100.0.1 (in two VLANs) and forward them to the host by updating the Ethernet source and destination addresses appropriately, decrementing the TTL, and skipping ahead to unicast output in table 7. This means that packets sent **to** 10.100.0.1 should now get to their destination.

Step 2: Router Sends ARP Reply

`inst/faucet.log` said that the router sent an ARP reply. How can we see it? Simulated packets just get dropped by default. One way is to configure the dummy ports to write the packets they receive to a file. Let's try that. First configure the port:

```
$ ovs-vsctl set interface p1 options:pcap=p1.pcap
```

Then re-run the "trace" command:

```
$ ovs-appctl ofproto/trace br0 in_port=p1,dl_src=00:01:02:03:04:05,dl_
↪dst=ff:ff:ff:ff:ff:ff,dl_type=0x806,arp_spa=10.100.0.1,arp_tpa=10.100.0.254,arp_
↪sha=00:01:02:03:04:05,arp_tha=ff:ff:ff:ff:ff:ff,arp_op=1 -generate
```

And dump the reply packet:

```
$ /usr/sbin/tcpdump -evvvr sandbox/p1.pcap
reading from file sandbox/p1.pcap, link-type EN10MB (Ethernet)
20:55:13.186932 0e:00:00:00:00:01 (oui Unknown) > 00:01:02:03:04:05 (oui Unknown),␣
↪ethertype ARP (0x0806), length 60: Ethernet (len 6), IPv4 (len 4), Reply 10.100.0.254␣
↪is-at 0e:00:00:00:00:01 (oui Unknown), length 46
```

We clearly see the ARP reply, which tells us that the Faucet router's Ethernet address is 0e:00:00:00:00:01 (as we guessed before from the flow table.

Let's configure the rest of our ports to log their packets, too:

```
$ for i in 2 3 4 5; do ovs-vsctl set interface p$i options:pcap=p$i.pcap; done
```

Step 3: Host Sends IP Packet

Now that host 10.100.0.1 has the MAC address for its router, it can send an IP packet to 10.200.0.1 via the router's MAC address, like this:

```
$ ovs-appctl ofproto/trace br0 in_port=p1,dl_src=00:01:02:03:04:05,dl_
↪dst=0e:00:00:00:00:01,udp,nw_src=10.100.0.1,nw_dst=10.200.0.1,nw_ttl=64 -generate
Flow: udp,in_port=1,vlan_tci=0x0000,dl_src=00:01:02:03:04:05,dl_dst=0e:00:00:00:00:01,nw_
↪src=10.100.0.1,nw_dst=10.200.0.1,nw_tos=0,nw_ecn=0,nw_ttl=64,tp_src=0,tp_dst=0

bridge("br0")
-------------
 0. in_port=1,vlan_tci=0x0000/0x1fff, priority 9000, cookie 0x5adc15c0
    push_vlan:0x8100
    set_field:4196->vlan_vid
    goto_table:1
 1. ip,dl_vlan=100,dl_dst=0e:00:00:00:00:01, priority 9099, cookie 0x5adc15c0
    goto_table:2
 2. ip,dl_vlan=100,nw_dst=10.200.0.0/24, priority 9123, cookie 0x5adc15c0
    goto_table:3
 3. ip, priority 9130, cookie 0x5adc15c0
    CONTROLLER:128

Final flow: udp,in_port=1,dl_vlan=100,dl_vlan_pcp=0,vlan_tci1=0x0000,dl_
↪src=00:01:02:03:04:05,dl_dst=0e:00:00:00:00:01,nw_src=10.100.0.1,nw_dst=10.200.0.1,nw_
↪tos=0,nw_ecn=0,nw_ttl=64,tp_src=0,tp_dst=0
Megaflow: recirc_id=0,eth,ip,in_port=1,vlan_tci=0x0000/0x1fff,dl_src=00:01:02:03:04:05,
↪dl_dst=0e:00:00:00:00:01,nw_dst=10.200.0.0/25,nw_frag=no
Datapath actions: push_vlan(vid=100,pcp=0),userspace(pid=0,controller(reason=1,dont_
↪send=0,continuation=0,recirc_id=6,rule_cookie=0x5adc15c0,controller_id=0,max_len=128))
```

Observe that the packet gets recognized as destined to the router, in table 1, and then as properly destined to the 10.200.0.0/24 network, in table 2. In table 3, however, it gets sent to the controller. Presumably, this is because Faucet has not yet resolved an Ethernet address for the destination host 10.200.0.1. It probably sent out an ARP request. Let's take a look in the next step.

Step 4: Router Broadcasts ARP Request

The router needs to know the Ethernet address of 10.200.0.1. It knows that, if this machine exists, it's on port p4 or p5, since we configured those ports as VLAN 200.

Let's make sure:

```
$ /usr/sbin/tcpdump -evvvr sandbox/p4.pcap
reading from file sandbox/p4.pcap, link-type EN10MB (Ethernet)
20:57:31.116097 0e:00:00:00:00:01 (oui Unknown) > Broadcast, ethertype ARP (0x0806),␣
↪length 60: Ethernet (len 6), IPv4 (len 4), Request who-has 10.200.0.1 tell 10.200.0.
↪254, length 46
```

and:

```
$ /usr/sbin/tcpdump -evvvr sandbox/p5.pcap
reading from file sandbox/p5.pcap, link-type EN10MB (Ethernet)
20:58:04.129735 0e:00:00:00:00:01 (oui Unknown) > Broadcast, ethertype ARP (0x0806),␣
↪length 60: Ethernet (len 6), IPv4 (len 4), Request who-has 10.200.0.1 tell 10.200.0.
↪254, length 46
```

For good measure, let's make sure that it wasn't sent to p3:

```
$ /usr/sbin/tcpdump -evvvr sandbox/p3.pcap
reading from file sandbox/p3.pcap, link-type EN10MB (Ethernet)
```

Step 5: Host 2 Sends ARP Reply

The Faucet controller sent an ARP request, so we can send an ARP reply:

```
$ ovs-appctl ofproto/trace br0 in_port=p4,dl_src=00:10:20:30:40:50,dl_
↪dst=0e:00:00:00:00:01,dl_type=0x806,arp_spa=10.200.0.1,arp_tpa=10.200.0.254,arp_
↪sha=00:10:20:30:40:50,arp_tha=0e:00:00:00:00:01,arp_op=2 -generate
Flow: arp,in_port=4,vlan_tci=0x0000,dl_src=00:10:20:30:40:50,dl_dst=0e:00:00:00:00:01,
↪arp_spa=10.200.0.1,arp_tpa=10.200.0.254,arp_op=2,arp_sha=00:10:20:30:40:50,arp_
↪tha=0e:00:00:00:00:01

bridge("br0")
-------------
 0. in_port=4,vlan_tci=0x0000/0x1fff, priority 9000, cookie 0x5adc15c0
    push_vlan:0x8100
    set_field:4296->vlan_vid
    goto_table:1
 1. arp,dl_vlan=200, priority 9131, cookie 0x5adc15c0
    goto_table:3
 3. arp,arp_tpa=10.200.0.254, priority 9133, cookie 0x5adc15c0
    CONTROLLER:128

Final flow: arp,in_port=4,dl_vlan=200,dl_vlan_pcp=0,vlan_tci1=0x0000,dl_
↪src=00:10:20:30:40:50,dl_dst=0e:00:00:00:00:01,arp_spa=10.200.0.1,arp_tpa=10.200.0.254,
↪arp_op=2,arp_sha=00:10:20:30:40:50,arp_tha=0e:00:00:00:00:01
Megaflow: recirc_id=0,eth,arp,in_port=4,vlan_tci=0x0000/0x1fff,arp_tpa=10.200.0.254
Datapath actions: push_vlan(vid=200,pcp=0),userspace(pid=0,controller(reason=1,dont_
↪send=0,continuation=0,recirc_id=7,rule_cookie=0x5adc15c0,controller_id=0,max_len=128))
```

It shows up in `inst/faucet.log`:

```
Sep 10 08:59:02 faucet.valve INFO     DPID 1 (0x1) switch-1 Adding new route 10.200.0.1/
↪32 via 10.200.0.1 (00:10:20:30:40:50) on VLAN 200
Sep 10 08:59:02 faucet.valve INFO     DPID 1 (0x1) switch-1 Received advert for 10.200.0.
↪1 from 00:10:20:30:40:50 (L2 type 0x0806, L3 src 10.200.0.1, L3 dst 10.200.0.254) Port␣
↪4 VLAN 200
Sep 10 08:59:02 faucet.valve INFO     DPID 1 (0x1) switch-1 L2 learned 00:10:20:30:40:50␣
↪(L2 type 0x0806, L3 src 10.200.0.1, L3 dst 10.200.0.254) Port 4 VLAN 200  (1 hosts␣
↪total)
```

and in the OVS flow tables:

```
$ diff-flows flows2 br0
+table=1 priority=9098,in_port=4,dl_vlan=200,dl_src=00:10:20:30:40:50 hard_timeout=3598␣
↪actions=goto_table:4
...
+table=2 priority=9131,ip,dl_vlan=200,nw_dst=10.200.0.1 actions=set_field:4296->vlan_vid,
↪set_field:0e:00:00:00:00:01->eth_src,set_field:00:10:20:30:40:50->eth_dst,dec_ttl,goto_
↪table:4
+table=2 priority=9131,ip,dl_vlan=100,nw_dst=10.200.0.1 actions=set_field:4296->vlan_vid,
↪set_field:0e:00:00:00:00:01->eth_src,set_field:00:10:20:30:40:50->eth_dst,dec_ttl,goto_
↪table:4
...
+table=4 priority=9099,dl_vlan=200,dl_dst=00:10:20:30:40:50 idle_timeout=3598␣
↪actions=pop_vlan,output:4
```

Step 6: IP Packet Delivery

Now both the host and the router have everything they need to deliver the packet. There are two ways it might happen. If Faucet's router is smart enough to buffer the packet that trigger ARP resolution, then it might have delivered it already. If so, then it should show up in `p4.pcap`. Let's take a look:

```
$ /usr/sbin/tcpdump -evvvr sandbox/p4.pcap ip
reading from file sandbox/p4.pcap, link-type EN10MB (Ethernet)
```

Nope. That leaves the other possibility, which is that Faucet waits for the original sending host to re-send the packet. We can do that by re-running the trace:

```
$ ovs-appctl ofproto/trace br0 in_port=p1,dl_src=00:01:02:03:04:05,dl_
↪dst=0e:00:00:00:00:01,udp,nw_src=10.100.0.1,nw_dst=10.200.0.1,nw_ttl=64 -generate

Flow: udp,in_port=1,vlan_tci=0x0000,dl_src=00:01:02:03:04:05,dl_dst=0e:00:00:00:00:01,nw_
↪src=10.100.0.1,nw_dst=10.200.0.1,nw_tos=0,nw_ecn=0,nw_ttl=64,tp_src=0,tp_dst=0
bridge("br0")
-------------
 0. in_port=1,vlan_tci=0x0000/0x1fff, priority 9000, cookie 0x5adc15c0
    push_vlan:0x8100
    set_field:4196->vlan_vid
    goto_table:1
 1. ip,dl_vlan=100,dl_dst=0e:00:00:00:00:01, priority 9099, cookie 0x5adc15c0
    goto_table:2
 2. ip,dl_vlan=100,nw_dst=10.200.0.1, priority 9131, cookie 0x5adc15c0
    set_field:4296->vlan_vid
```

(continues on next page)

(continued from previous page)

```
    set_field:0e:00:00:00:00:01->eth_src
    set_field:00:10:20:30:40:50->eth_dst
    dec_ttl
    goto_table:4
 4. dl_vlan=200,dl_dst=00:10:20:30:40:50, priority 9099, cookie 0x5adc15c0
    pop_vlan
    output:4

Final flow: udp,in_port=1,vlan_tci=0x0000,dl_src=0e:00:00:00:00:01,dl_
↪dst=00:10:20:30:40:50,nw_src=10.100.0.1,nw_dst=10.200.0.1,nw_tos=0,nw_ecn=0,nw_ttl=63,
↪tp_src=0,tp_dst=0
Megaflow: recirc_id=0,eth,ip,in_port=1,vlan_tci=0x0000/0x1fff,dl_src=00:01:02:03:04:05,
↪dl_dst=0e:00:00:00:00:01,nw_dst=10.200.0.1,nw_ttl=64,nw_frag=no
Datapath actions: set(eth(src=0e:00:00:00:00:01,dst=00:10:20:30:40:50)),set(ipv4(dst=10.
↪200.0.1,ttl=63)),4
```

Finally, we have working IP packet forwarding!

Performance

Take another look at the megaflow line above:

```
Megaflow: recirc_id=0,eth,ip,in_port=1,vlan_tci=0x0000/0x1fff,dl_src=00:01:02:03:04:05,
↪dl_dst=0e:00:00:00:00:01,nw_dst=10.200.0.1,nw_ttl=64,nw_frag=no
```

This means that (almost) any packet between these Ethernet source and destination hosts, destined to the given IP host, will be handled by this single megaflow cache entry. So regardless of the number of UDP packets or TCP connections that these hosts exchange, Open vSwitch packet processing won't need to fall back to the slow path. It is quite efficient.

> **Note**
>
> The exceptions are packets with a TTL other than 64, and fragmented packets. Most hosts use a constant TTL for outgoing packets, and fragments are rare. If either of those did change, then that would simply result in a new megaflow cache entry.

The datapath actions might also be worth a look:

```
Datapath actions: set(eth(src=0e:00:00:00:00:01,dst=00:10:20:30:40:50)),set(ipv4(dst=10.
↪200.0.1,ttl=63)),4
```

This just means that, to process these packets, the datapath changes the Ethernet source and destination addresses and the IP TTL, and then transmits the packet to port `p4` (also numbered 4). Notice in particular that, despite the OpenFlow actions that pushed, modified, and popped back off a VLAN, there is nothing in the datapath actions about VLANs. This is because the OVS flow translation code "optimizes out" redundant or unneeded actions, which saves time when the cache entry is executed later.

> **Note**
>
> It's not clear why the actions also re-set the IP destination address to its original value. Perhaps this is a minor performance bug.

3.1.6 ACLs

Let's try out some ACLs, since they do a good job illustrating some of the ways that OVS tries to optimize megaflows. Update `inst/faucet.yaml` to the following:

```
dps:
    switch-1:
        dp_id: 0x1
        timeout: 3600
        arp_neighbor_timeout: 3600
        interfaces:
            1:
                native_vlan: 100
                acl_in: 1
            2:
                native_vlan: 100
            3:
                native_vlan: 100
            4:
                native_vlan: 200
            5:
                native_vlan: 200
vlans:
    100:
        faucet_vips: ["10.100.0.254/24"]
    200:
        faucet_vips: ["10.200.0.254/24"]
routers:
    router-1:
        vlans: [100, 200]
acls:
    1:
        - rule:
            dl_type: 0x800
            nw_proto: 6
            tcp_dst: 8080
            actions:
                allow: 0
        - rule:
            actions:
                allow: 1
```

Then reload Faucet:

```
$ sudo docker exec faucet pkill -HUP -f faucet.faucet
```

We will now find Faucet has added a new table to the start of the pipeline for processing port ACLs. Let's take a look at our new table 0 with `dump-flows br0`:

```
priority=9099,tcp,in_port=p1,tp_dst=8080 actions=drop
priority=9098,in_port=p1 actions=goto_table:1
priority=9099,in_port=p2 actions=goto_table:1
priority=9099,in_port=p3 actions=goto_table:1
priority=9099,in_port=p4 actions=goto_table:1
```

(continues on next page)

(continued from previous page)

```
priority=9099,in_port=p5 actions=goto_table:1
priority=0 actions=drop
```

We now have a flow that just jumps to table 1 (vlan) for each configured port, and a low priority rule to drop other unrecognized packets. We also see a flow rule for dropping TCP port 8080 traffic on port 1. If we compare this rule to the ACL we configured, we can clearly see how Faucet has converted this ACL to fit into the OpenFlow pipeline.

The most interesting question here is performance. If you recall the earlier discussion, when a packet through the flow table encounters a match on a given field, the resulting megaflow has to match on that field, even if the flow didn't actually match. This is expensive.

In particular, here you can see that any TCP packet is going to encounter the ACL flow, even if it is directed to a port other than 8080. If that means that every megaflow for a TCP packet is going to have to match on the TCP destination, that's going to be bad for caching performance because there will be a need for a separate megaflow for every TCP destination port that actually appears in traffic, which means a lot more megaflows than otherwise. (Really, in practice, if such a simple ACL blew up performance, OVS wouldn't be a very good switch!)

Let's see what happens, by sending a packet to port 80 (instead of 8080):

```
$ ovs-appctl ofproto/trace br0 in_port=p1,dl_src=00:01:02:03:04:05,dl_
↪dst=0e:00:00:00:00:01,tcp,nw_src=10.100.0.1,nw_dst=10.200.0.1,nw_ttl=64,tp_dst=80 -
↪generate
src=10.100.0.1,nw_dst=10.200.0.1,nw_ttl=64,tp_dst=80 -generate
Flow: tcp,in_port=1,vlan_tci=0x0000,dl_src=00:01:02:03:04:05,dl_dst=0e:00:00:00:00:01,nw_
↪src=10.100.0.1,nw_dst=10.200.0.1,nw_tos=0,nw_ecn=0,nw_ttl=64,tp_src=0,tp_dst=80,tcp_
↪flags=0

bridge("br0")
-------------
 0. in_port=1, priority 9098, cookie 0x5adc15c0
    goto_table:1
 1. in_port=1,vlan_tci=0x0000/0x1fff, priority 9000, cookie 0x5adc15c0
    push_vlan:0x8100
    set_field:4196->vlan_vid
    goto_table:2
 2. ip,dl_vlan=100,dl_dst=0e:00:00:00:00:01, priority 9099, cookie 0x5adc15c0
    goto_table:3
 3. ip,dl_vlan=100,nw_dst=10.200.0.0/24, priority 9123, cookie 0x5adc15c0
    goto_table:4
 4. ip, priority 9130, cookie 0x5adc15c0
    CONTROLLER:128

Final flow: tcp,in_port=1,dl_vlan=100,dl_vlan_pcp=0,vlan_tci1=0x0000,dl_
↪src=00:01:02:03:04:05,dl_dst=0e:00:00:00:00:01,nw_src=10.100.0.1,nw_dst=10.200.0.1,nw_
↪tos=0,nw_ecn=0,nw_ttl=64,tp_src=0,tp_dst=80,tcp_flags=0
Megaflow: recirc_id=0,eth,tcp,in_port=1,vlan_tci=0x0000/0x1fff,dl_src=00:01:02:03:04:05,
↪dl_dst=0e:00:00:00:00:01,nw_dst=10.200.0.0/25,nw_frag=no,tp_dst=0x0/0xf000
Datapath actions: push_vlan(vid=100,pcp=0),userspace(pid=0,controller(reason=1,dont_
↪send=0,continuation=0,recirc_id=8,rule_cookie=0x5adc15c0,controller_id=0,max_len=128))
```

Take a look at the Megaflow line and in particular the match on `tp_dst`, which says `tp_dst=0x0/0xf000`. What this means is that the megaflow matches on only the top 4 bits of the TCP destination port. That works because:

```
  80 (base 10) == 0000,0000,0101,0000 (base 2)
8080 (base 10) == 0001,1111,1001,0000 (base 2)
```

and so by matching on only the top 4 bits, rather than all 16, the OVS fast path can distinguish port 80 from port 8080. This allows this megaflow to match one-sixteenth of the TCP destination port address space, rather than just 1/65536th of it.

> **Note**
>
> The algorithm OVS uses for this purpose isn't perfect. In this case, a single-bit match would work (e.g. tp_dst=0x0/0x1000), and would be superior since it would only match half the port address space instead of one-sixteenth.

For details of this algorithm, please refer to `lib/classifier.c` in the Open vSwitch source tree, or our 2015 NSDI paper "The Design and Implementation of Open vSwitch".

3.1.7 Finishing Up

When you're done, you probably want to exit the sandbox session, with Control+D or `exit`, and stop the Faucet controller with `sudo docker stop faucet; sudo docker rm faucet`.

3.1.8 Further Directions

We've looked a fair bit at how Faucet interacts with Open vSwitch. If you still have some interest, you might want to explore some of these directions:

- Adding more than one switch. Faucet can control multiple switches but we've only been simulating one of them. It's easy enough to make a single OVS instance act as multiple switches (just `ovs-vsctl add-br` another bridge), or you could use genuinely separate OVS instances.
- Additional features. Faucet has more features than we've demonstrated, such as IPv6 routing and port mirroring. These should also interact gracefully with Open vSwitch.
- Real performance testing. We've looked at how flows and traces **should** demonstrate good performance, but of course there's no proof until it actually works in practice. We've also only tested with trivial configurations. Open vSwitch can scale to millions of OpenFlow flows, but the scaling in practice depends on the particular flow tables and traffic patterns, so it's valuable to test with large configurations, either in the way we've done it or with real traffic.

3.2 OVS IPsec Tutorial

This document provides a step-by-step guide for running IPsec tunnel in Open vSwitch. A more detailed description on OVS IPsec tunnel and its configuration modes can be found in *Encrypt Open vSwitch Tunnels with IPsec*.

3.2.1 Requirements

OVS IPsec tunnel requires Linux kernel (>= v3.10.0) and OVS out-of-tree kernel module. The compatible IKE daemons are LibreSwan (>= v3.23) and StrongSwan (>= v5.3.5).

3.2.2 Installing OVS and IPsec Packages

OVS IPsec has .deb and .rpm packages. You should use the right package based on your Linux distribution. This tutorial uses Ubuntu 22.04 and Fedora 32 as examples.

Ubuntu

1. Install the related packages:

```
# apt-get install openvswitch-ipsec
```

   If the installation is successful, you should be able to see the ovs-monitor-ipsec daemon is running in your system.

Fedora

1. Install the related packages. Fedora 32 does not require installation of the out-of-tree kernel module:

```
# dnf install python3-openvswitch libreswan \
              openvswitch openvswitch-ipsec
```

2. Install firewall rules to allow ESP and IKE traffic:

```
# systemctl start firewalld
# firewall-cmd --add-service ipsec
```

   Or to make permanent:

```
# systemctl enable firewalld
# firewall-cmd --permanent --add-service ipsec
```

3. Run the openvswitch-ipsec service:

```
# systemctl start openvswitch-ipsec.service
```

> **Note**
>
> The SELinux policies might prevent openvswitch-ipsec.service to access certain resources. You can configure SELinux to remove such restrictions.

3.2.3 Configuring IPsec tunnel

Suppose you want to build an IPsec tunnel between two hosts. Assume *host_1*'s external IP is 1.1.1.1, and *host_2*'s external IP is 2.2.2.2. Make sure *host_1* and *host_2* can ping each other via these external IPs.

0. Set up some variables to make life easier. On both hosts, set `ip_1` and `ip_2` variables, e.g.:

```
# ip_1=1.1.1.1
# ip_2=2.2.2.2
```

1. Set up OVS bridges in both hosts.

   In *host_1*:

```
# ovs-vsctl add-br br-ipsec
# ip addr add 192.0.0.1/24 dev br-ipsec
# ip link set br-ipsec up
```

In *host_2*:

```
# ovs-vsctl add-br br-ipsec
# ip addr add 192.0.0.2/24 dev br-ipsec
# ip link set br-ipsec up
```

2. Set up IPsec tunnel.

   There are three authentication methods. Choose one method to set up your IPsec tunnel and follow the steps below.

   a) Using pre-shared key:

   In *host_1*:

```
# ovs-vsctl add-port br-ipsec tun -- \
           set interface tun type=gre \
                         options:remote_ip=$ip_2 \
                         options:psk=swordfish
```

   In *host_2*:

```
# ovs-vsctl add-port br-ipsec tun -- \
           set interface tun type=gre \
                         options:remote_ip=$ip_1 \
                         options:psk=swordfish
```

   Note

   Pre-shared key (PSK) based authentication is easy to set up but less secure compared with other authentication methods. You should use it cautiously in production systems.

   b) Using self-signed certificate:

   Generate self-signed certificate in both *host_1* and *host_2*. Then copy the certificate of *host_1* to *host_2* and the certificate of *host_2* to *host_1*.

   In *host_1*:

```
# ovs-pki req -u host_1
# ovs-pki self-sign host_1
# scp host_1-cert.pem $ip_2:/etc/keys/host_1-cert.pem
```

   In *host_2*:

```
# ovs-pki req -u host_2
# ovs-pki self-sign host_2
# scp host_2-cert.pem $ip_1:/etc/keys/host_2-cert.pem
```

   Note

   If you use StrongSwan as IKE daemon, please move the host certificates to /etc/ipsec.d/certs/ and private key to /etc/ipsec.d/private/ so that StrongSwan has permission to access those files.

   Configure IPsec tunnel to use self-signed certificates.

In *host_1*:

```
# ovs-vsctl set Open_vSwitch . \
          other_config:certificate=/etc/keys/host_1-cert.pem \
          other_config:private_key=/etc/keys/host_1-privkey.pem
# ovs-vsctl add-port br-ipsec tun -- \
           set interface tun type=gre \
                  options:remote_ip=$ip_2 \
                  options:remote_cert=/etc/keys/host_2-cert.pem
```

In *host_2*:

```
# ovs-vsctl set Open_vSwitch . \
          other_config:certificate=/etc/keys/host_2-cert.pem \
          other_config:private_key=/etc/keys/host_2-privkey.pem
# ovs-vsctl add-port br-ipsec tun -- \
           set interface tun type=gre \
                  options:remote_ip=$ip_1 \
                  options:remote_cert=/etc/keys/host_1-cert.pem
```

Note

The confidentiality of the private key is very critical. Don't copy it to places where it might be compromised. (The certificate need not be kept confidential.)

c) Using CA-signed certificate:

First you need to establish a public key infrastructure (PKI). Suppose you choose *host_1* to host PKI.

In *host_1*:

```
# ovs-pki init
```

Generate certificate requests and copy the certificate request of *host_2* to *host_1*.

In *host_1*:

```
# ovs-pki req -u host_1
```

In *host_2*:

```
# ovs-pki req -u host_2
# scp host_2-req.pem $ip_1:/etc/keys/host_2-req.pem
```

Sign the certificate requests with the CA key. Copy *host_2*'s signed certificate and the CA certificate to *host_2*.

In *host_1*:

```
# ovs-pki sign host_1 switch
# ovs-pki sign host_2 switch
# scp host_2-cert.pem $ip_2:/etc/keys/host_2-cert.pem
# scp /var/lib/openvswitch/pki/switchca/cacert.pem \
          $ip_2:/etc/keys/cacert.pem
```

> **Note**
>
> If you use StrongSwan as IKE daemon, please move the host certificates to /etc/ipsec.d/certs/, CA certificate to /etc/ipsec.d/cacerts/, and private key to /etc/ipsec.d/private/ so that StrongSwan has permission to access those files.

Configure IPsec tunnel to use CA-signed certificate.

In *host_1*:

```
# ovs-vsctl set Open_vSwitch . \
        other_config:certificate=/etc/keys/host_1-cert.pem \
        other_config:private_key=/etc/keys/host_1-privkey.pem \
        other_config:ca_cert=/etc/keys/cacert.pem
# ovs-vsctl add-port br-ipsec tun -- \
         set interface tun type=gre \
                       options:remote_ip=$ip_2 \
                       options:remote_name=host_2
```

In *host_2*:

```
# ovs-vsctl set Open_vSwitch . \
        other_config:certificate=/etc/keys/host_2-cert.pem \
        other_config:private_key=/etc/keys/host_2-privkey.pem \
        other_config:ca_cert=/etc/keys/cacert.pem
# ovs-vsctl add-port br-ipsec tun -- \
         set interface tun type=gre \
                       options:remote_ip=$ip_1 \
                       options:remote_name=host_1
```

> **Note**
>
> remote_name is the common name (CN) of the signed-certificate. It must match the name given as the argument to the `ovs-pki sign command`. It ensures that only certificate with the expected CN can be authenticated; otherwise, any certificate signed by the CA would be accepted.

3. Set the *local_ip* field in the Interface table (Optional)

   Make sure that the *local_ip* field in the Interface table is set to the NIC used for egress traffic.

   On *host 1*:

   ```
   # ovs-vsctl set Interface tun options:local_ip=$ip_1
   ```

   Similarly, on *host 2*:

   ```
   # ovs-vsctl set Interface tun options:local_ip=$ip_2
   ```

   > **Note**
   >
   > It is not strictly necessary to set the *local_ip* field if your system only has one NIC or the default gateway interface is set to the NIC used for egress traffic.

4. Test IPsec tunnel.

   Now you should have an IPsec GRE tunnel running between two hosts. To verify it, in *host_1*:

   ```
   # ping 192.0.0.2 &
   # tcpdump -ni any net $ip_2
   ```

   You should be able to see that ESP packets are being sent from *host_1* to *host_2*.

3.2.4 Custom options

Any parameter prefixed with *ipsec_* will be added to the connection profile. For example:

```
# ovs-vsctl set interface tun options:ipsec_encapsulation=yes
```

Will result in:

```
#  ovs-appctl -t ovs-monitor-ipsec tunnels/show
Interface name: tun v7 (CONFIGURED)
Tunnel Type:    vxlan
Local IP:       192.0.0.1
Remote IP:      192.0.0.2
Address Family: IPv4
SKB mark:       None
Local cert:     None
Local name:     None
Local key:      None
Remote cert:    None
Remote name:    None
CA cert:        None
PSK:            swordfish
Custom Options: {'encapsulation': 'yes'}
```

And in the following connection profiles:

```
conn tun-in-7
    left=192.0.0.1
    right=192.0.0.2
    authby=secret
    encapsulation=yes
    leftprotoport=udp/4789
    rightprotoport=udp

conn tun-out-7
    left=192.0.0.1
    right=192.0.0.2
    authby=secret
    encapsulation=yes
    leftprotoport=udp
    rightprotoport=udp/4789
```

3.2.5 Troubleshooting

The `ovs-monitor-ipsec` daemon manages and monitors the IPsec tunnel state. Use the following `ovs-appctl` command to view `ovs-monitor-ipsec` internal representation of tunnel configuration:

```
# ovs-appctl -t ovs-monitor-ipsec tunnels/show
```

If there is misconfiguration, then `ovs-appctl` should indicate why. For example:

```
Interface name: gre0 v5 (CONFIGURED) <--- Should be set to CONFIGURED.
                                          Otherwise, error message will
                                          be provided
Tunnel Type:    gre
Local IP:       %defaultroute
Remote IP:      2.2.2.2
SKB mark:       None
Local cert:     None
Local name:     None
Local key:      None
Remote cert:    None
Remote name:    None
CA cert:        None
PSK:            swordfish
Custom Options: {}
Ofport:         1          <--- Whether ovs-vswitchd has assigned Ofport
                                number to this Tunnel Port
CFM state:      Up         <--- Whether CFM declared this tunnel healthy
Kernel policies installed:
...                          <--- IPsec policies for this OVS tunnel in
                                  Linux Kernel installed by strongSwan
Kernel security associations installed:
...                          <--- IPsec security associations for this OVS
                                  tunnel in Linux Kernel installed by
                                  strongswan
IPsec connections that are active:
...                          <--- IPsec "connections" for this OVS
                                  tunnel
```

If you don't see any active connections, try to run the following command to refresh the `ovs-monitor-ipsec` daemon:

```
# ovs-appctl -t ovs-monitor-ipsec refresh
```

You can also check the logs of the `ovs-monitor-ipsec` daemon and the IKE daemon to locate issues. `ovs-monitor-ipsec` outputs log messages to /var/log/openvswitch/ovs-monitor-ipsec.log.

3.2.6 Bug Reporting

If you think you may have found a bug with security implications, like

1. IPsec protected tunnel accepted packets that came unencrypted; OR
2. IPsec protected tunnel allowed packets to leave unencrypted;

Then report such bugs according to *Security Process*.

If bug does not have security implications, then report it according to instructions in *Reporting Bugs*.

If you have suggestions to improve this tutorial, please send a email to ovs-discuss@openvswitch.org.

3.3 Open vSwitch Advanced Features

Many tutorials cover the basics of OpenFlow. This is not such a tutorial. Rather, a knowledge of the basics of OpenFlow is a prerequisite. If you do not already understand how an OpenFlow flow table works, please go read a basic tutorial and then continue reading here afterward.

It is also important to understand the basics of Open vSwitch before you begin. If you have never used ovs-vsctl or ovs-ofctl before, you should learn a little about them before proceeding.

Most of the features covered in this tutorial are Open vSwitch extensions to OpenFlow. Also, most of the features in this tutorial are specific to the software Open vSwitch implementation. If you are using an Open vSwitch port to an ASIC-based hardware switch, this tutorial will not help you.

This tutorial does not cover every aspect of the features that it mentions. You can find the details elsewhere in the Open vSwitch documentation, especially `ovs-ofctl(8)` and the comments in the `include/openflow/nicira-ext.h` and `include/openvswitch/meta-flow.h` header files.

3.3.1 Getting Started

This is a hands-on tutorial. To get the most out of it, you will need Open vSwitch binaries. You do not, on the other hand, need any physical networking hardware or even supervisor privilege on your system. Instead, we will use a script called `ovs-sandbox`, which accompanies the tutorial, that constructs a software simulated network environment based on Open vSwitch.

You can use `ovs-sandbox` three ways:

- If you have already installed Open vSwitch on your system, then you should be able to just run `ovs-sandbox` from this directory without any options.
- If you have not installed Open vSwitch (and you do not want to install it), then you can build Open vSwitch according to the instructions in *Open vSwitch on Linux, FreeBSD and NetBSD*, without installing it. Then run `./ovs-sandbox -b DIRECTORY` from this directory, substituting the Open vSwitch build directory for `DIRECTORY`.
- As a slight variant on the latter, you can run `make sandbox` from an Open vSwitch build directory.

When you run `ovs-sandbox`, it does the following:

1. **CAUTION:** Deletes any subdirectory of the current directory named "sandbox" and any files in that directory.
2. Creates a new directory "sandbox" in the current directory.
3. Sets up special environment variables that ensure that Open vSwitch programs will look inside the "sandbox" directory instead of in the Open vSwitch installation directory.
4. If you are using a built but not installed Open vSwitch, installs the Open vSwitch manpages in a subdirectory of "sandbox" and adjusts the `MANPATH` environment variable to point to this directory. This means that you can use, for example, `man ovs-vsctl` to see a manpage for the `ovs-vsctl` program that you built.
5. Creates an empty Open vSwitch configuration database under "sandbox".
6. Starts `ovsdb-server` running under "sandbox".
7. Starts `ovs-vswitchd` running under "sandbox", passing special options that enable a special "dummy" mode for testing.
8. Starts a nested interactive shell inside "sandbox".

At this point, you can run all the usual Open vSwitch utilities from the nested shell environment. You can, for example, use `ovs-vsctl` to create a bridge:

```
$ ovs-vsctl add-br br0
```

From Open vSwitch's perspective, the bridge that you create this way is as real as any other. You can, for example, connect it to an OpenFlow controller or use `ovs-ofctl` to examine and modify it and its OpenFlow flow table. On the other hand, the bridge is not visible to the operating system's network stack, so `ip` cannot see it or affect it, which means that utilities like `ping` and `tcpdump` will not work either. (That has its good side, too: you can't screw up your computer's network stack by manipulating a sandboxed OVS.)

When you're done using OVS from the sandbox, exit the nested shell (by entering the "exit" shell command or pressing Control+D). This will kill the daemons that `ovs-sandbox` started, but it leaves the "sandbox" directory and its contents in place.

The sandbox directory contains log files for the Open vSwitch dameons. You can examine them while you're running in the sandboxed environment or after you exit.

3.3.2 Using GDB

GDB support is not required to go through the tutorial. It is added in case user wants to explore the internals of OVS programs.

GDB can already be used to debug any running process, with the usual `gdb <program> <process-id>` command.

`ovs-sandbox` also has a `-g` option for launching ovs-vswitchd under GDB. This option can be handy for setting break points before ovs-vswitchd runs, or for catching early segfaults. Similarly, a `-d` option can be used to run ovsdb-server under GDB. Both options can be specified at the same time.

In addition, a `-e` option also launches ovs-vswitchd under GDB. However, instead of displaying a `gdb>` prompt and waiting for user input, ovs-vswitchd will start to execute immediately. `-r` option is the corresponding option for running ovsdb-server under gdb with immediate execution.

To avoid GDB mangling with the sandbox sub shell terminal, `ovs-sandbox` starts a new xterm to run each GDB session. For systems that do not support X windows, GDB support is effectively disabled.

When launching sandbox through the build tree's make file, the `-g` option can be passed via the `SANDBOXFLAGS` environment variable. `make sandbox SANDBOXFLAGS=-g` will start the sandbox with ovs-vswitchd running under GDB in its own xterm if X is available.

In addition, a set of GDB macros are available in `utilities/gdb/ovs_gdb.py`. Which are able to dump various internal data structures. See the header of the file itself for some more details and an example.

3.3.3 Motivation

The goal of this tutorial is to demonstrate the power of Open vSwitch flow tables. The tutorial works through the implementation of a MAC-learning switch with VLAN trunk and access ports. Outside of the Open vSwitch features that we will discuss, OpenFlow provides at least two ways to implement such a switch:

1. An OpenFlow controller to implement MAC learning in a "reactive" fashion. Whenever a new MAC appears on the switch, or a MAC moves from one switch port to another, the controller adjusts the OpenFlow flow table to match.

2. The "normal" action. OpenFlow defines this action to submit a packet to "the traditional non-OpenFlow pipeline of the switch". That is, if a flow uses this action, then the packets in the flow go through the switch in the same way that they would if OpenFlow was not configured on the switch.

Each of these approaches has unfortunate pitfalls. In the first approach, using an OpenFlow controller to implement MAC learning, has a significant cost in terms of network bandwidth and latency. It also makes the controller more difficult to scale to large numbers of switches, which is especially important in environments with thousands of hypervisors (each of which contains a virtual OpenFlow switch). MAC learning at an OpenFlow controller also behaves poorly if the OpenFlow controller fails, slows down, or becomes unavailable due to network problems.

The second approach, using the "normal" action, has different problems. First, little about the "normal" action is standardized, so it behaves differently on switches from different vendors, and the available features and how those

features are configured (usually not through OpenFlow) varies widely. Second, "normal" does not work well with other OpenFlow actions. It is "all-or-nothing", with little potential to adjust its behavior slightly or to compose it with other features.

3.3.4 Scenario

We will construct Open vSwitch flow tables for a VLAN-capable, MAC-learning switch that has four ports:

p1
: a trunk port that carries all VLANs, on OpenFlow port 1.

p2
: an access port for VLAN 20, on OpenFlow port 2.

p3, p4
: both access ports for VLAN 30, on OpenFlow ports 3 and 4, respectively.

> **Note**
>
> The ports' names are not significant. You could call them eth1 through eth4, or any other names you like.

> **Note**
>
> An OpenFlow switch always has a "local" port as well. This scenario won't use the local port.

Our switch design will consist of five main flow tables, each of which implements one stage in the switch pipeline:

Table 0
: Admission control.

Table 1
: VLAN input processing.

Table 2
: Learn source MAC and VLAN for ingress port.

Table 3
: Look up learned port for destination MAC and VLAN.

Table 4
: Output processing.

The section below describes how to set up the scenario, followed by a section for each OpenFlow table.

You can cut and paste the `ovs-vsctl` and `ovs-ofctl` commands in each of the sections below into your `ovs-sandbox` shell. They are also available as shell scripts in this directory, named `t-setup`, `t-stage0`, `t-stage1`, ..., `t-stage4`. The `ovs-appctl` test commands are intended for cutting and pasting and are not supplied separately.

3.3.5 Setup

To get started, start `ovs-sandbox`. Inside the interactive shell that it starts, run this command:

```
$ ovs-vsctl add-br br0 -- set Bridge br0 fail-mode=secure
```

This command creates a new bridge "br0" and puts "br0" into so-called "fail-secure" mode. For our purpose, this just means that the OpenFlow flow table starts out empty.

> **Note**
>
> If we did not do this, then the flow table would start out with a single flow that executes the "normal" action. We could use that feature to yield a switch that behaves the same as the switch we are currently building, but with the caveats described under "Motivation" above.)

The new bridge has only one port on it so far, the "local port" br0. We need to add `p1`, `p2`, `p3`, and `p4`. A shell `for` loop is one way to do it:

```
for i in 1 2 3 4; do
    ovs-vsctl add-port br0 p$i -- set Interface p$i ofport_request=$i
    ovs-ofctl mod-port br0 p$i up
done
```

In addition to adding a port, the `ovs-vsctl` command above sets its `ofport_request` column to ensure that port `p1` is assigned OpenFlow port 1, `p2` is assigned OpenFlow port 2, and so on.

> **Note**
>
> We could omit setting the ofport_request and let Open vSwitch choose port numbers for us, but it's convenient for the purposes of this tutorial because we can talk about OpenFlow port 1 and know that it corresponds to `p1`.

The `ovs-ofctl` command above brings up the simulated interfaces, which are down initially, using an OpenFlow request. The effect is similar to `ip link up`, but the sandbox's interfaces are not visible to the operating system and therefore `ip` would not affect them.

We have not configured anything related to VLANs or MAC learning. That's because we're going to implement those features in the flow table.

To see what we've done so far to set up the scenario, you can run a command like `ovs-vsctl show` or `ovs-ofctl show br0`.

3.3.6 Implementing Table 0: Admission control

Table 0 is where packets enter the switch. We use this stage to discard packets that for one reason or another are invalid. For example, packets with a multicast source address are not valid, so we can add a flow to drop them at ingress to the switch with:

```
$ ovs-ofctl add-flow br0 \
    "table=0, dl_src=01:00:00:00:00:00/01:00:00:00:00:00, actions=drop"
```

A switch should also not forward IEEE 802.1D Spanning Tree Protocol (STP) packets, so we can also add a flow to drop those and other packets with reserved multicast protocols:

```
$ ovs-ofctl add-flow br0 \
    "table=0, dl_dst=01:80:c2:00:00:00/ff:ff:ff:ff:ff:f0, actions=drop"
```

We could add flows to drop other protocols, but these demonstrate the pattern.

We need one more flow, with a priority lower than the default, so that flows that don't match either of the "drop" flows we added above go on to pipeline stage 1 in OpenFlow table 1:

```
$ ovs-ofctl add-flow br0 "table=0, priority=0, actions=resubmit(,1)"
```

> **Note**
>
> The "resubmit" action is an Open vSwitch extension to OpenFlow.

3.3.7 Testing Table 0

If we were using Open vSwitch to set up a physical or a virtual switch, then we would naturally test it by sending packets through it one way or another, perhaps with common network testing tools like `ping` and `tcpdump` or more specialized tools like Scapy. That's difficult with our simulated switch, since it's not visible to the operating system.

But our simulated switch has a few specialized testing tools. The most powerful of these tools is `ofproto/trace`. Given a switch and the specification of a flow, `ofproto/trace` shows, step-by-step, how such a flow would be treated as it goes through the switch.

Example 1

Try this command:

```
$ ovs-appctl ofproto/trace br0 in_port=1,dl_dst=01:80:c2:00:00:05
```

The output should look something like this:

```
Flow: in_port=1,vlan_tci=0x0000,dl_src=00:00:00:00:00:00,dl_dst=01:80:c2:00:00:05,dl_
↪type=0x0000

bridge("br0")
-------------
 0. dl_dst=01:80:c2:00:00:00/ff:ff:ff:ff:ff:f0, priority 32768
    drop

Final flow: unchanged
Megaflow: recirc_id=0,in_port=1,dl_src=00:00:00:00:00:00/01:00:00:00:00:00,dl_
↪dst=01:80:c2:00:00:00/ff:ff:ff:ff:ff:f0,dl_type=0x0000
Datapath actions: drop
```

The first line shows the flow being traced, in slightly greater detail than specified on the command line. It is mostly zeros because unspecified fields default to zeros.

The second group of lines shows the packet's trip through bridge br0. We see, in table 0, the OpenFlow flow that the fields matched, along with its priority, followed by its actions, one per line. In this case, we see that this packet that has a reserved multicast destination address matches the flow that drops those packets.

The final block of lines summarizes the results, which are not very interesting here.

Example 2

Try another command:

```
$ ovs-appctl ofproto/trace br0 in_port=1,dl_dst=01:80:c2:00:00:10
```

The output should be:

```
Flow: in_port=1,vlan_tci=0x0000,dl_src=00:00:00:00:00:00,dl_dst=01:80:c2:00:00:10,dl_
↪type=0x0000
```

(continues on next page)

(continued from previous page)

```
bridge("br0")
-------------
 0. priority 0
    resubmit(,1)
 1. No match.
    drop

Final flow: unchanged
Megaflow: recirc_id=0,in_port=1,dl_src=00:00:00:00:00:00/01:00:00:00:00:00,dl_
↪dst=01:80:c2:00:00:10/ff:ff:ff:ff:ff:f0,dl_type=0x0000
Datapath actions: drop
```

This time the flow we handed to `ofproto/trace` doesn't match any of our "drop" flows in table 0, so it falls through to the low-priority "resubmit" flow. The "resubmit" causes a second lookup in OpenFlow table 1, described by the block of text that starts with "1." We haven't yet added any flows to OpenFlow table 1, so no flow actually matches in the second lookup. Therefore, the packet is still actually dropped, which means that the externally observable results would be identical to our first example.

3.3.8 Implementing Table 1: VLAN Input Processing

A packet that enters table 1 has already passed basic validation in table 0. The purpose of table 1 is validate the packet's VLAN, based on the VLAN configuration of the switch port through which the packet entered the switch. We will also use it to attach a VLAN header to packets that arrive on an access port, which allows later processing stages to rely on the packet's VLAN always being part of the VLAN header, reducing special cases.

Let's start by adding a low-priority flow that drops all packets, before we add flows that pass through acceptable packets. You can think of this as a "default drop" flow:

```
$ ovs-ofctl add-flow br0 "table=1, priority=0, actions=drop"
```

Our trunk port `p1`, on OpenFlow port 1, is an easy case. `p1` accepts any packet regardless of whether it has a VLAN header or what the VLAN was, so we can add a flow that resubmits everything on input port 1 to the next table:

```
$ ovs-ofctl add-flow br0 \
    "table=1, priority=99, in_port=1, actions=resubmit(,2)"
```

On the access ports, we want to accept any packet that has no VLAN header, tag it with the access port's VLAN number, and then pass it along to the next stage:

```
$ ovs-ofctl add-flows br0 - <<'EOF'
table=1, priority=99, in_port=2, vlan_tci=0, actions=mod_vlan_vid:20, resubmit(,2)
table=1, priority=99, in_port=3, vlan_tci=0, actions=mod_vlan_vid:30, resubmit(,2)
table=1, priority=99, in_port=4, vlan_tci=0, actions=mod_vlan_vid:30, resubmit(,2)
EOF
```

We don't write any flows that match packets with 802.1Q that enter this stage on any of the access ports, so the "default drop" flow we added earlier causes them to be dropped, which is ordinarily what we want for access ports.

> **Note**
>
> Another variation of access ports allows ingress of packets tagged with VLAN 0 (aka 802.1p priority tagged packets). To allow such packets, replace `vlan_tci=0` by `vlan_tci=0/0xfff` above.

3.3.9 Testing Table 1

`ofproto/trace` allows us to test the ingress VLAN flows that we added above.

Example 1: Packet on Trunk Port

Here's a test of a packet coming in on the trunk port:

```
$ ovs-appctl ofproto/trace br0 in_port=1,vlan_tci=5
```

The output shows the lookup in table 0, the resubmit to table 1, and the resubmit to table 2 (which does nothing because we haven't put anything there yet):

```
Flow: in_port=1,vlan_tci=0x0005,dl_src=00:00:00:00:00:00,dl_dst=00:00:00:00:00:00,dl_
↪type=0x0000

bridge("br0")
-------------
 0. priority 0
    resubmit(,1)
 1. in_port=1, priority 99
    resubmit(,2)
 2. No match.
    drop

Final flow: unchanged
Megaflow: recirc_id=0,in_port=1,dl_src=00:00:00:00:00:00/01:00:00:00:00:00,dl_
↪dst=00:00:00:00:00:00/ff:ff:ff:ff:ff:f0,dl_type=0x0000
Datapath actions: drop
```

Example 2: Valid Packet on Access Port

Here's a test of a valid packet (a packet without an 802.1Q header) coming in on access port `p2`:

```
$ ovs-appctl ofproto/trace br0 in_port=2
```

The output is similar to that for the previous case, except that it additionally tags the packet with `p2`'s VLAN 20 before it passes it along to table 2:

```
Flow: in_port=2,vlan_tci=0x0000,dl_src=00:00:00:00:00:00,dl_dst=00:00:00:00:00:00,dl_
↪type=0x0000

bridge("br0")
-------------
 0. priority 0
    resubmit(,1)
 1. in_port=2,vlan_tci=0x0000, priority 99
    mod_vlan_vid:20
    resubmit(,2)
 2. No match.
    drop

Final flow: in_port=2,dl_vlan=20,dl_vlan_pcp=0,dl_src=00:00:00:00:00:00,dl_
↪dst=00:00:00:00:00:00,dl_type=0x0000
Megaflow: recirc_id=0,in_port=2,vlan_tci=0x0000,dl_src=00:00:00:00:00:00/
```

(continues on next page)

(continued from previous page)

```
↪01:00:00:00:00:00,dl_dst=00:00:00:00:00:00/ff:ff:ff:ff:ff:f0,dl_type=0x0000
Datapath actions: drop
```

Example 3: Invalid Packet on Access Port

This tests an invalid packet (one that includes an 802.1Q header) coming in on access port `p2`:

```
$ ovs-appctl ofproto/trace br0 in_port=2,vlan_tci=5
```

The output shows the packet matching the default drop flow:

```
Flow: in_port=2,vlan_tci=0x0005,dl_src=00:00:00:00:00:00,dl_dst=00:00:00:00:00:00,dl_
↪type=0x0000

bridge("br0")
-------------
 0. priority 0
    resubmit(,1)
 1. priority 0
    drop

Final flow: unchanged
Megaflow: recirc_id=0,in_port=2,vlan_tci=0x0005,dl_src=00:00:00:00:00:00/
↪01:00:00:00:00:00,dl_dst=00:00:00:00:00:00/ff:ff:ff:ff:ff:f0,dl_type=0x0000
Datapath actions: drop
```

3.3.10 Implementing Table 2: MAC+VLAN Learning for Ingress Port

This table allows the switch we're implementing to learn that the packet's source MAC is located on the packet's ingress port in the packet's VLAN.

> **Note**
>
> This table is a good example why table 1 added a VLAN tag to packets that entered the switch through an access port. We want to associate a MAC+VLAN with a port regardless of whether the VLAN in question was originally part of the packet or whether it was an assumed VLAN associated with an access port.

It only takes a single flow to do this. The following command adds it:

```
$ ovs-ofctl add-flow br0 \
    "table=2 actions=learn(table=10, NXM_OF_VLAN_TCI[0..11], \
                           NXM_OF_ETH_DST[]=NXM_OF_ETH_SRC[], \
                           load:NXM_OF_IN_PORT[]->NXM_NX_REG0[0..15]), \
                     resubmit(,3)"
```

The "learn" action (an Open vSwitch extension to OpenFlow) modifies a flow table based on the content of the flow currently being processed. Here's how you can interpret each part of the "learn" action above:

`table=10`
: Modify flow table 10. This will be the MAC learning table.

`NXM_OF_VLAN_TCI[0..11]`
: Make the flow that we add to flow table 10 match the same VLAN ID that the packet we're currently processing

contains. This effectively scopes the MAC learning entry to a single VLAN, which is the ordinary behavior for a VLAN-aware switch.

NXM_OF_ETH_DST[]=NXM_OF_ETH_SRC[]
Make the flow that we add to flow table 10 match, as Ethernet destination, the Ethernet source address of the packet we're currently processing.

load:NXM_OF_IN_PORT[]->NXM_NX_REG0[0..15]
Whereas the preceding parts specify fields for the new flow to match, this specifies an action for the flow to take when it matches. The action is for the flow to load the ingress port number of the current packet into register 0 (a special field that is an Open vSwitch extension to OpenFlow).

> **Note**
>
> A real use of "learn" for MAC learning would probably involve two additional elements. First, the "learn" action would specify a hard_timeout for the new flow, to enable a learned MAC to eventually expire if no new packets were seen from a given source within a reasonable interval. Second, one would usually want to limit resource consumption by using the Flow_Table table in the Open vSwitch configuration database to specify a maximum number of flows in table 10.

This definitely calls for examples.

3.3.11 Testing Table 2

Example 1

Try the following test command:

```
$ ovs-appctl ofproto/trace br0 \
    in_port=1,vlan_tci=20,dl_src=50:00:00:00:00:01 -generate
```

The output shows that "learn" was executed in table 2 and the particular flow that was added:

```
Flow: in_port=1,vlan_tci=0x0014,dl_src=50:00:00:00:00:01,dl_dst=00:00:00:00:00:00,dl_
↪type=0x0000

bridge("br0")
-------------
 0. priority 0
    resubmit(,1)
 1. in_port=1, priority 99
    resubmit(,2)
 2. priority 32768
    learn(table=10,NXM_OF_VLAN_TCI[0..11],NXM_OF_ETH_DST[]=NXM_OF_ETH_SRC[],load:NXM_OF_
↪IN_PORT[]->NXM_NX_REG0[0..15])
     -> table=10 vlan_tci=0x0014/0x0fff,dl_dst=50:00:00:00:00:01 priority=32768␣
↪actions=load:0x1->NXM_NX_REG0[0..15]
    resubmit(,3)
 3. No match.
    drop

Final flow: unchanged
Megaflow: recirc_id=0,in_port=1,vlan_tci=0x0014/0x1fff,dl_src=50:00:00:00:00:01,dl_
↪dst=00:00:00:00:00:00/ff:ff:ff:ff:ff:f0,dl_type=0x0000
Datapath actions: drop
```

The `-generate` keyword is new. Ordinarily, `ofproto/trace` has no side effects: "output" actions do not actually output packets, "learn" actions do not actually modify the flow table, and so on. With `-generate`, though, `ofproto/trace` does execute "learn" actions. That's important now, because we want to see the effect of the "learn" action on table 10. You can see that by running:

```
$ ovs-ofctl dump-flows br0 table=10
```

which (omitting the `duration` and `idle_age` fields, which will vary based on how soon you ran this command after the previous one, as well as some other uninteresting fields) prints something like:

```
NXST_FLOW reply (xid=0x4):
 table=10, vlan_tci=0x0014/0x0fff,dl_dst=50:00:00:00:00:01 actions=load:0x1->NXM_NX_
↪REG0[0..15]
```

You can see that the packet coming in on VLAN `20` with source MAC `50:00:00:00:00:01` became a flow that matches VLAN `20` (written in hexadecimal) and destination MAC `50:00:00:00:00:01`. The flow loads port number 1, the input port for the flow we tested, into register 0.

Example 2

Here's a second test command:

```
$ ovs-appctl ofproto/trace br0 \
    in_port=2,dl_src=50:00:00:00:00:01 -generate
```

The flow that this command tests has the same source MAC and VLAN as example 1, although the VLAN comes from an access port VLAN rather than an 802.1Q header. If we again dump the flows for table 10 with:

```
$ ovs-ofctl dump-flows br0 table=10
```

then we see that the flow we saw previously has changed to indicate that the learned port is port 2, as we would expect:

```
NXST_FLOW reply (xid=0x4):
 table=10, vlan_tci=0x0014/0x0fff,dl_dst=50:00:00:00:00:01 actions=load:0x2->NXM_NX_
↪REG0[0..15]
```

3.3.12 Implementing Table 3: Look Up Destination Port

This table figures out what port we should send the packet to based on the destination MAC and VLAN. That is, if we've learned the location of the destination (from table 2 processing some previous packet with that destination as its source), then we want to send the packet there.

We need only one flow to do the lookup:

```
$ ovs-ofctl add-flow br0 \
    "table=3 priority=50 actions=resubmit(,10), resubmit(,4)"
```

The flow's first action resubmits to table 10, the table that the "learn" action modifies. As you saw previously, the learned flows in this table write the learned port into register 0. If the destination for our packet hasn't been learned, then there will be no matching flow, and so the "resubmit" turns into a no-op. Because registers are initialized to 0, we can use a register 0 value of 0 in our next pipeline stage as a signal to flood the packet.

The second action resubmits to table 4, continuing to the next pipeline stage.

We can add another flow to skip the learning table lookup for multicast and broadcast packets, since those should always be flooded:

```
$ ovs-ofctl add-flow br0 \
    "table=3 priority=99 dl_dst=01:00:00:00:00:00/01:00:00:00:00:00 \
      actions=resubmit(,4)"
```

Note

We don't strictly need to add this flow, because multicast addresses will never show up in our learning table. (In turn, that's because we put a flow into table 0 to drop packets that have a multicast source address.)

3.3.13 Testing Table 3

Example

Here's a command that should cause OVS to learn that `f0:00:00:00:00:01` is on `p1` in VLAN `20`:

```
$ ovs-appctl ofproto/trace br0 \
    in_port=1,dl_vlan=20,dl_src=f0:00:00:00:00:01,dl_dst=90:00:00:00:00:01 \
    -generate
```

The output shows (from the "no match" looking up the resubmit to table 10) that the flow's destination was unknown:

```
Flow: in_port=1,dl_vlan=20,dl_vlan_pcp=0,dl_src=f0:00:00:00:00:01,dl_
↪dst=90:00:00:00:00:01,dl_type=0x0000

bridge("br0")
-------------
 0. priority 0
    resubmit(,1)
 1. in_port=1, priority 99
    resubmit(,2)
 2. priority 32768
    learn(table=10,NXM_OF_VLAN_TCI[0..11],NXM_OF_ETH_DST[]=NXM_OF_ETH_SRC[],load:NXM_OF_
↪IN_PORT[]->NXM_NX_REG0[0..15])
     -> table=10 vlan_tci=0x0014/0x0fff,dl_dst=f0:00:00:00:00:01 priority=32768␣
↪actions=load:0x1->NXM_NX_REG0[0..15]
    resubmit(,3)
 3. priority 50
    resubmit(,10)
    10. No match.
            drop
    resubmit(,4)
 4. No match.
    drop

Final flow: unchanged
Megaflow: recirc_id=0,in_port=1,dl_vlan=20,dl_src=f0:00:00:00:00:01,dl_
↪dst=90:00:00:00:00:01,dl_type=0x0000
Datapath actions: drop
```

There are two ways that you can verify that the packet's source was learned. The most direct way is to dump the learning table with:

```
$ ovs-ofctl dump-flows br0 table=10
```

which ought to show roughly the following, with extraneous details removed:

```
table=10, vlan_tci=0x0014/0x0fff,dl_dst=f0:00:00:00:00:01 actions=load:0x1->NXM_NX_
↪REG0[0..15]
```

Note

If you tried the examples for the previous step, or if you did some of your own experiments, then you might see additional flows there. These additional flows are harmless. If they bother you, then you can remove them with *ovs-ofctl del-flows br0 table=10*.

The other way is to inject a packet to take advantage of the learning entry. For example, we can inject a packet on p2 whose destination is the MAC address that we just learned on p1:

```
$ ovs-appctl ofproto/trace br0 \
    in_port=2,dl_src=90:00:00:00:00:01,dl_dst=f0:00:00:00:00:01 -generate
```

Here is this command's output. Take a look at the lines that trace the `resubmit(,10)`, showing that the packet matched the learned flow for the first MAC we used, loading the OpenFlow port number for the learned port `p1` into register `0`:

```
Flow: in_port=2,vlan_tci=0x0000,dl_src=90:00:00:00:00:01,dl_dst=f0:00:00:00:00:01,dl_
↪type=0x0000

bridge("br0")
-------------
 0. priority 0
    resubmit(,1)
 1. in_port=2,vlan_tci=0x0000, priority 99
    mod_vlan_vid:20
    resubmit(,2)
 2. priority 32768
    learn(table=10,NXM_OF_VLAN_TCI[0..11],NXM_OF_ETH_DST[]=NXM_OF_ETH_SRC[],load:NXM_OF_
↪IN_PORT[]->NXM_NX_REG0[0..15])
     -> table=10 vlan_tci=0x0014/0x0fff,dl_dst=90:00:00:00:00:01 priority=32768␣
↪actions=load:0x2->NXM_NX_REG0[0..15]
    resubmit(,3)
 3. priority 50
    resubmit(,10)
    10. vlan_tci=0x0014/0x0fff,dl_dst=f0:00:00:00:00:01, priority 32768
            load:0x1->NXM_NX_REG0[0..15]
    resubmit(,4)
 4. No match.
    drop

Final flow: reg0=0x1,in_port=2,dl_vlan=20,dl_vlan_pcp=0,dl_src=90:00:00:00:00:01,dl_
↪dst=f0:00:00:00:00:01,dl_type=0x0000
Megaflow: recirc_id=0,in_port=2,vlan_tci=0x0000,dl_src=90:00:00:00:00:01,dl_
↪dst=f0:00:00:00:00:01,dl_type=0x0000
Datapath actions: drop
```

If you read the commands above carefully, then you might have noticed that they simply have the Ethernet source and destination addresses exchanged. That means that if we now rerun the first `ovs-appctl` command above, e.g.:

```
$ ovs-appctl ofproto/trace br0 \
    in_port=1,dl_vlan=20,dl_src=f0:00:00:00:00:01,dl_dst=90:00:00:00:00:01 \
    -generate
```

then we see in the output, looking at the indented "load" action executed in table 10, that the destination has now been learned:

```
Flow: in_port=1,dl_vlan=20,dl_vlan_pcp=0,dl_src=f0:00:00:00:00:01,dl_
↪dst=90:00:00:00:00:01,dl_type=0x0000

bridge("br0")
-------------
 0. priority 0
    resubmit(,1)
 1. in_port=1, priority 99
    resubmit(,2)
 2. priority 32768
    learn(table=10,NXM_OF_VLAN_TCI[0..11],NXM_OF_ETH_DST[]=NXM_OF_ETH_SRC[],load:NXM_OF_
↪IN_PORT[]->NXM_NX_REG0[0..15])
     -> table=10 vlan_tci=0x0014/0x0fff,dl_dst=f0:00:00:00:00:01 priority=32768␣
↪actions=load:0x1->NXM_NX_REG0[0..15]
    resubmit(,3)
 3. priority 50
    resubmit(,10)
    10. vlan_tci=0x0014/0x0fff,dl_dst=90:00:00:00:00:01, priority 32768
            load:0x2->NXM_NX_REG0[0..15]
    resubmit(,4)
 4. No match.
    drop
```

3.3.14 Implementing Table 4: Output Processing

At entry to stage 4, we know that register 0 contains either the desired output port or is zero if the packet should be flooded. We also know that the packet's VLAN is in its 802.1Q header, even if the VLAN was implicit because the packet came in on an access port.

The job of the final pipeline stage is to actually output packets. The job is trivial for output to our trunk port `p1`:

```
$ ovs-ofctl add-flow br0 "table=4 reg0=1 actions=1"
```

For output to the access ports, we just have to strip the VLAN header before outputting the packet:

```
$ ovs-ofctl add-flows br0 - <<'EOF'
table=4 reg0=2 actions=strip_vlan,2
table=4 reg0=3 actions=strip_vlan,3
table=4 reg0=4 actions=strip_vlan,4
EOF
```

The only slightly tricky part is flooding multicast and broadcast packets and unicast packets with unlearned destinations. For those, we need to make sure that we only output the packets to the ports that carry our packet's VLAN, and that we include the 802.1Q header in the copy output to the trunk port but not in copies output to access ports:

```
$ ovs-ofctl add-flows br0 - <<'EOF'
table=4 reg0=0 priority=99 dl_vlan=20 actions=1,strip_vlan,2
table=4 reg0=0 priority=99 dl_vlan=30 actions=1,strip_vlan,3,4
table=4 reg0=0 priority=50            actions=1
EOF
```

Note

Our flows rely on the standard OpenFlow behavior that an output action will not forward a packet back out the port it came in on. That is, if a packet comes in on p1, and we've learned that the packet's destination MAC is also on p1, so that we end up with `actions=1` as our actions, the switch will not forward the packet back out its input port. The multicast/broadcast/unknown destination cases above also rely on this behavior.

3.3.15 Testing Table 4

Example 1: Broadcast, Multicast, and Unknown Destination

Try tracing a broadcast packet arriving on `p1` in VLAN `30`:

```
$ ovs-appctl ofproto/trace br0 \
    in_port=1,dl_dst=ff:ff:ff:ff:ff:ff,dl_vlan=30
```

The interesting part of the output is the final line, which shows that the switch would remove the 802.1Q header and then output the packet to `p3` and `p4`, which are access ports for VLAN `30`:

```
Datapath actions: pop_vlan,3,4
```

Similarly, if we trace a broadcast packet arriving on `p3`:

```
$ ovs-appctl ofproto/trace br0 in_port=3,dl_dst=ff:ff:ff:ff:ff:ff
```

then we see that it is output to `p1` with an 802.1Q tag and then to `p4` without one:

```
Datapath actions: push_vlan(vid=30,pcp=0),1,pop_vlan,4
```

Note

Open vSwitch could simplify the datapath actions here to just `4,push_vlan(vid=30,pcp=0),1` but it is not smart enough to do so.

The following are also broadcasts, but the result is to drop the packets because the VLAN only belongs to the input port:

```
$ ovs-appctl ofproto/trace br0 \
    in_port=1,dl_dst=ff:ff:ff:ff:ff:ff
$ ovs-appctl ofproto/trace br0 \
    in_port=1,dl_dst=ff:ff:ff:ff:ff:ff,dl_vlan=55
```

Try some other broadcast cases on your own:

```
$ ovs-appctl ofproto/trace br0 \
    in_port=1,dl_dst=ff:ff:ff:ff:ff:ff,dl_vlan=20
$ ovs-appctl ofproto/trace br0 \
    in_port=2,dl_dst=ff:ff:ff:ff:ff:ff
$ ovs-appctl ofproto/trace br0 \
    in_port=4,dl_dst=ff:ff:ff:ff:ff:ff
```

You can see the same behavior with multicast packets and with unicast packets whose destination has not been learned, e.g.:

```
$ ovs-appctl ofproto/trace br0 \
    in_port=4,dl_dst=01:00:00:00:00:00
$ ovs-appctl ofproto/trace br0 \
    in_port=1,dl_dst=90:12:34:56:78:90,dl_vlan=20
$ ovs-appctl ofproto/trace br0 \
    in_port=1,dl_dst=90:12:34:56:78:90,dl_vlan=30
```

Example 2: MAC Learning

Let's follow the same pattern as we did for table 3. First learn a MAC on port `p1` in VLAN `30`:

```
$ ovs-appctl ofproto/trace br0 \
    in_port=1,dl_vlan=30,dl_src=10:00:00:00:00:01,dl_dst=20:00:00:00:00:01 \
    -generate
```

You can see from the last line of output that the packet's destination is unknown, so it gets flooded to both `p3` and `p4`, the other ports in VLAN `30`:

```
Datapath actions: pop_vlan,3,4
```

Then reverse the MACs and learn the first flow's destination on port `p4`:

```
$ ovs-appctl ofproto/trace br0 \
    in_port=4,dl_src=20:00:00:00:00:01,dl_dst=10:00:00:00:00:01 -generate
```

The last line of output shows that the this packet's destination is known to be `p1`, as learned from our previous command:

```
Datapath actions: push_vlan(vid=30,pcp=0),1
```

Now, if we rerun our first command:

```
$ ovs-appctl ofproto/trace br0 \
    in_port=1,dl_vlan=30,dl_src=10:00:00:00:00:01,dl_dst=20:00:00:00:00:01 \
    -generate
```

. . . we can see that the result is no longer a flood but to the specified learned destination port `p4`:

```
Datapath actions: pop_vlan,4
```

Contact

bugs@openvswitch.org http://openvswitch.org/

3.4 OVS Conntrack Tutorial

OVS can be used with the Connection tracking system where OpenFlow flow can be used to match on the state of a TCP, UDP, ICMP, etc., connections. (Connection tracking system supports tracking of both statefull and stateless protocols)

This tutorial demonstrates how OVS can use the connection tracking system to match on the TCP segments from connection setup to connection tear down. It will use OVS with the Linux kernel module as the datapath for this tutorial. (The datapath that utilizes the openvswitch kernel module to do the packet processing in the Linux kernel)

3.4.1 Definitions

conntrack: is a connection tracking module for stateful packet inspection.

pipeline: is the packet processing pipeline which is the path taken by the packet when traversing through the tables where the packet matches the match fields of a flow in the table and performs the actions present in the matched flow.

network namespace: is a way to create virtual routing domains within a single instance of linux kernel. Each network namespace has it's own instance of network tables (arp, routing) and certain interfaces attached to it.

flow: used in this tutorial refers to the OpenFlow flow which can be programmed using an OpenFlow controller or OVS command line tools like ovs-ofctl which is used here. A flow will have match fields and actions.

3.4.2 Conntrack Related Fields

Match Fields

OVS supports following match fields related to conntrack:

1. **ct_state**: The state of a connection matching the packet. Possible values:

   - *new*
   - *est*
   - *rel*
   - *rpl*
   - *inv*
   - *trk*
   - *snat*
   - *dnat*

Each of these flags is preceded by either a "+" for a flag that must be set, or a "-" for a flag that must be unset. Multiple flags can also be specified e.g. ct_state=+trk+new. We will see the usage of some of these flags below. For a detailed description, please see the OVS fields documentation at: http://openvswitch.org/support/dist-docs/ovs-fields.7.txt

2. **ct_zone**: A zone is an independent connection tracking context which can be set by a ct action. A 16-bit ct_zone set by the most recent ct action (by an OpenFlow flow on a conntrack entry) can be used as a match field in another flow entry.

3. **ct_mark**: The 32-bit metadata committed, by an action within the exec parameter to the ct action, to the connection to which the current packet belongs.

4. **ct_label**: The 128-bit label committed by an action within the exec parameter to the ct action, to the connection to which the current packet belongs.

5. **ct_nw_src / ct_ipv6_src**: Matches IPv4/IPv6 conntrack original direction tuple source address.

6. **ct_nw_dst / ct_ipv6_dst**: Matches IPv4/IPv6 conntrack original direction tuple destination address.

7. **ct_nw_proto**: Matches conntrack original direction tuple IP protocol type.

8. **ct_tp_src**: Matches on the conntrack original direction tuple transport source port.

9. **ct_tp_dst**: Matches on the conntrack original direction tuple transport destination port.

Actions

OVS supports "ct" action related to conntrack.

ct([argument][,argument...])

The **ct** action sends the packet through the connection tracker.

The following arguments are supported:

1. **commit**: Commit the connection to the connection tracking module which will be stored beyond the lifetime of packet in the pipeline.

2. **force**: The force flag may be used in addition to commit flag to effectively terminate the existing connection and start a new one in the current direction.

3. **table=number**: Fork pipeline processing in two. The original instance of the packet will continue processing the current actions list as an untracked packet. An additional instance of the packet will be sent to the connection tracker, which will be re-injected into the OpenFlow pipeline to resume processing in table number, with the ct_state and other ct match fields set.

4. **zone=value OR zone=src[start..end]**: A 16-bit context id that can be used to isolate connections into separate domains, allowing over-lapping network addresses in different zones. If a zone is not provided, then the default is to use zone zero.

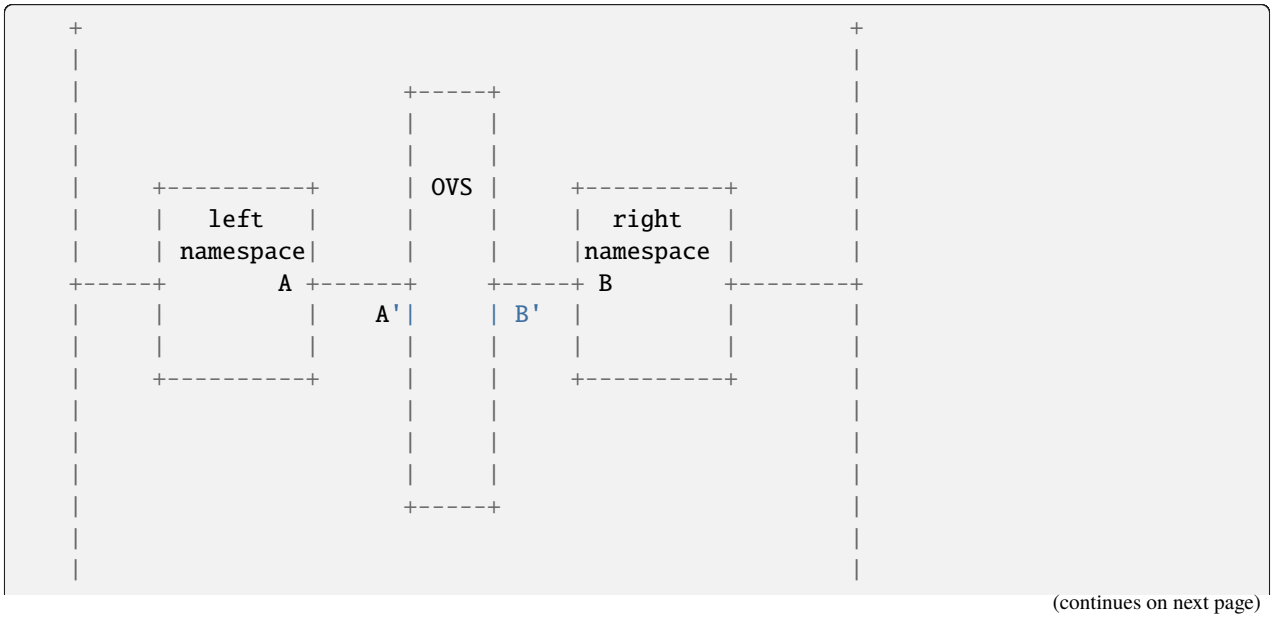
5. **exec([action][,action...])**: Perform restricted set of actions within the context of connection tracking. Only actions which modify the *ct_mark* or *ct_label* fields are accepted within the exec action.

6. **alg=<ftp/tftp>**: Specify alg (application layer gateway) to track specific connection types.

7. **nat**: Specifies the address and port translation for the connection being tracked.

3.4.3 Sample Topology

This tutorial uses the following topology to carry out the tests.



(continues on next page)

(continued from previous page)

```
    +                                                    +
192.168.0.X n/w                                      10.0.0.X n/w

A  = veth_l1
A' = veth_l0
B  = veth_r1
B' = veth_r0

Diagram: Sample Topology for conntrack testing
```

The steps for creation of the setup are mentioned below.

Create "left" network namespace:

```
$ ip netns add left
```

Create "right" network namespace:

```
$ ip netns add right
```

Create first pair of veth interfaces:

```
$ ip link add veth_l0 type veth peer name veth_l1
```

Add veth_l1 to "left" network namespace:

```
$ ip link set veth_l1 netns left
```

Create second pair of veth interfaces:

```
$ ip link add veth_r0 type veth peer name veth_r1
```

Add veth_r1 to "right" network namespace:

```
$ ip link set veth_r1 netns right
```

Create a bridge br0:

```
$ ovs-vsctl add-br br0
```

Add veth_l0 and veth_r0 to br0:

```
$ ovs-vsctl add-port br0 veth_l0
$ ovs-vsctl add-port br0 veth_r0
```

Packets generated with src/dst IP set to 192.168.0.X / 10.0.0.X in the "left" and the inverse in the "right" namespaces will appear to OVS as hosts in two networks (192.168.0.X and 10.0.0.X) communicating with each other. This is basically a simulation of two networks / subnets with hosts communicating with each other with OVS in middle.

3.4.4 Tool used to generate TCP segments

You can use scapy to generate the TCP segments. We used scapy on Ubuntu 16.04 for the steps carried out in this testing. (Installation of scapy is not discussed and is out of scope of this document.)

You can keep two scapy sessions active on each of the namespaces:

```
$ sudo ip netns exec left sudo `which scapy`

$ sudo ip netns exec right sudo `which scapy`
```

Note: In case you encounter this error:

```
ifreq = ioctl(s, SIOCGIFADDR,struct.pack("16s16x",LOOPBACK_NAME))

IOError: [Errno 99] Cannot assign requested address
```

run the command:

```
$ sudo ip netns exec <namespace> sudo ip link set lo up
```

3.4.5 Matching TCP packets

TCP Connection setup

Two simple flows can be added in OVS which will forward packets from "left" to "right" and from "right" to "left":

```
$ ovs-ofctl add-flow br0 \
        "table=0, priority=10, in_port=veth_l0, actions=veth_r0"

$ ovs-ofctl add-flow br0 \
        "table=0, priority=10, in_port=veth_r0, actions=veth_l0"
```

Instead of adding these two flows, we will add flows to match on the states of the TCP segments.

We will send the TCP connection setup segments namely: syn, syn-ack and ack between hosts 192.168.0.2 in the "left" namespace and 10.0.0.2 in the "right" namespace.

First, let's add a flow to start "tracking" a packet received at OVS.

How do we start tracking a packet?

To start tracking a packet, it first needs to match a flow, which has action as "ct". This action sends the packet through the connection tracker. To identify that a packet is an "untracked" packet, the ct_state in the flow match field must be set to "-trk", which means it is not a tracked packet. Once the packet is sent to the connection tracker, then only we will know about its conntrack state. (i.e. whether this packet represents start of a new connection or the packet belongs to an existing connection or it is a malformed packet and so on.)

Let's add that flow:

```
(flow #1)
$ ovs-ofctl add-flow br0 \
   "table=0, priority=50, ct_state=-trk, tcp, in_port=veth_l0, actions=ct(table=0)"
```

A TCP syn packet sent from "left" namespace will match flow #1 because the packet is coming to OVS from veth_l0 port and it is not being tracked. This is because the packet just entered OVS. When a packet enters a namespace for the first time, a new connection tracker context is entered, hence, the packet will be initially "untracked" in that namespace. When a packet (re)enters the same datapath that it already belongs to there is no need to discard the namespace and other information associated with the conntrack flow. In this case the packet will remain in the tracked state. If the namespace has changed then it is discarded and a new connection tracker is created since connection tracking information is logically separate for different namespaces. The flow will send the packet to the connection tracker due to the action "ct". Also "table=0" in the "ct" action forks the pipeline processing in two. The original instance of packet will continue processing the current action list as untracked packet. (Since there are no actions after this, the original packet gets dropped.) The forked instance of the packet will be sent to the connection tracker, which

will be re-injected into the OpenFlow pipeline to resume processing in table number, with the ct_state and other ct match fields set. In this case, the packet with the ct_state and other ct match fields comes back to table 0.

Next, we add a flow to match on the packet coming back from conntrack:

```
(flow #2)
$ ovs-ofctl add-flow br0 \
    "table=0, priority=50, ct_state=+trk+new, tcp, in_port=veth_l0, actions=ct(commit),
↪veth_r0"
```

Now that the packet is coming back from conntrack, the ct_state would have the "trk" set. Also, if this is the first packet of the TCP connection, the ct_state "new" would be set. (Which is the condition here as there does not exist any TCP connection between hosts 192.168.0.2 and 10.0.0.2) The ct argument "commit" will commit the connection to the connection tracking module. The significance of this action is that the information about the connection will now be stored beyond the lifetime of the packet in the pipeline.

Let's send the TCP syn segment using scapy (at the "left" scapy session) (flags=0x02 is syn):

```
$ >>> sendp(Ether()/IP(src="192.168.0.2", dst="10.0.0.2")/TCP(sport=1024, dport=2048,␣
↪flags=0x02, seq=100), iface="veth_l1")
```

This packet will match flow #1 and flow #2.

The conntrack module will now have an entry for this connection:

```
$ ovs-appctl dpctl/dump-conntrack | grep "192.168.0.2"
tcp,orig=(src=192.168.0.2,dst=10.0.0.2,sport=1024,dport=2048),reply=(src=10.0.0.2,
↪dst=192.168.0.2,sport=2048,dport=1024),protoinfo=(state=SYN_SENT)
```

Note: At this stage, if the TCP syn packet is re-transmitted, it will again match flow #1 (since a new packet is untracked) and it will match flow #2. The reason it will match flow #2 is that although conntrack has information about the connection, but it is not in "ESTABLISHED" state, therefore it matches the "new" state again.

Next for the TCP syn-ack from the opposite/server direction, we need following flows at OVS:

```
(flow #3)
$ ovs-ofctl add-flow br0 \
    "table=0, priority=50, ct_state=-trk, tcp, in_port=veth_r0, actions=ct(table=0)"
(flow #4)
$ ovs-ofctl add-flow br0 \
    "table=0, priority=50, ct_state=+trk+est, tcp, in_port=veth_r0, actions=veth_l0"
```

flow #3 matches untracked packets coming back from server (10.0.0.2) and sends this to conntrack. (Alternatively, we could have also combined flow #1 and flow #3 into one flow by not having the "in_port" match)

The syn-ack packet which has now gone through the conntrack has the ct_state of "est".

Note: Conntrack puts the ct_state of the connection to "est" state when it sees bidirectional traffic, but till it does not get the third ack from client, it puts a short cleanup timer on the conntrack entry.

Sending TCP syn-ack segment using scapy (at the "right" scapy session) (flags=0x12 is ack and syn):

```
$ >>> sendp(Ether()/IP(src="10.0.0.2", dst="192.168.0.2")/TCP(sport=2048, dport=1024,␣
↪flags=0x12, seq=200, ack=101), iface="veth_r1")
```

This packet will match flow #3 and flow #4.

conntrack entry:

```
$ ovs-appctl dpctl/dump-conntrack | grep "192.168.0.2"

tcp,orig=(src=192.168.0.2,dst=10.0.0.2,sport=1024,dport=2048),reply=(src=10.0.0.2,
↪dst=192.168.0.2,sport=2048,dport=1024),protoinfo=(state=ESTABLISHED)
```

The conntrack state is "ESTABLISHED" on receiving just syn and syn-ack packets, but at this point if it does not receive the third ack (from client), the connection gets cleared up from conntrack quickly.

Next, for a TCP ack from client direction, we can add following flows to match on the packet:

```
(flow #5)
$ ovs-ofctl add-flow br0 \
    "table=0, priority=50, ct_state=+trk+est, tcp, in_port=veth_l0, actions=veth_r0"
```

Send the third TCP ack segment using scapy (at the "left" scapy session) (flags=0x10 is ack):

```
$ >>> sendp(Ether()/IP(src="192.168.0.2", dst="10.0.0.2")/TCP(sport=1024, dport=2048,␣
↪flags=0x10, seq=101, ack=201), iface="veth_l1")
```

This packet will match on flow #1 and flow #5.

conntrack entry:

```
$ ovs-appctl dpctl/dump-conntrack | grep "192.168.0.2"

 tcp,orig=(src=192.168.0.2,dst=10.0.0.2,sport=1024,dport=2048), \
     reply=(src=10.0.0.2,dst=192.168.0.2,sport=2048,dport=1024), \
                                       protoinfo=(state=ESTABLISHED)
```

The conntrack state stays in "ESTABLISHED" state, but now since it has received the ack from client, it will stay in this state for a longer time even without receiving any data on this connection.

TCP Data

When a data segment, carrying one byte of TCP payload, is sent from 192.168.0.2 to 10.0.0.2, the packet carrying the segment would hit flow #1 and then flow #5.

Send a TCP segment with one byte data using scapy (at the "left" scapy session) (flags=0x10 is ack):

```
$ >>> sendp(Ether()/IP(src="192.168.0.2", dst="10.0.0.2")/TCP(sport=1024, dport=2048,␣
↪flags=0x10, seq=101, ack=201)/"X", iface="veth_l1")
```

Send the TCP ack for the above segment using scapy (at the "right" scapy session) (flags=0x10 is ack):

```
$ >>> sendp(Ether()/IP(src="10.0.0.2", dst="192.168.0.2")/TCP(sport=2048, dport=1024,␣
↪flags=0X10, seq=201, ack=102), iface="veth_r1")
```

The acknowledgement for the data would hit flow #3 and flow #4.

TCP Connection Teardown

There are different ways to tear down TCP connection. We will tear down the connection by sending "fin" from client, "fin-ack" from server followed by the last "ack" by client.

All the packets from client to server would hit flow #1 and flow #5. All the packets from server to client would hit flow #3 and flow #4. Interesting point to note is that even when the TCP connection is going down, all the packets (which

are actually tearing down the connection) still hits "+est" state. A packet, for which the conntrack entry *is* or *was* in "ESTABLISHED" state, would continue to match "+est" ct_state in OVS.

Note: In fact, when the conntrack connection state is in "TIME_WAIT" state (after all the TCP fins and their acks are exchanged), a re-transmitted data packet (from 192.168.0.2 -> 10.0.0.2), still hits flows #1 and #5.

Sending TCP fin segment using scapy (at the "left" scapy session) (flags=0x11 is ack and fin):

```
$ >>> sendp(Ether()/IP(src="192.168.0.2", dst="10.0.0.2")/TCP(sport=1024, dport=2048,␣
↪flags=0x11, seq=102, ack=201), iface="veth_l1")
```

This packet hits flow #1 and flow #5.

conntrack entry:

```
$ sudo ovs-appctl dpctl/dump-conntrack | grep "192.168.0.2"

  tcp,orig=(src=192.168.0.2,dst=10.0.0.2,sport=1024,dport=2048),reply=(src=10.0.0.2,
↪dst=192.168.0.2,sport=2048,dport=1024),protoinfo=(state=FIN_WAIT_1)
```

Sending TCP fin-ack segment using scapy (at the "right" scapy session) (flags=0x11 is ack and fin):

```
$ >>> sendp(Ether()/IP(src="10.0.0.2", dst="192.168.0.2")/TCP(sport=2048, dport=1024,␣
↪flags=0X11, seq=201, ack=103), iface="veth_r1")
```

This packet hits flow #3 and flow #4.

conntrack entry:

```
$ sudo ovs-appctl dpctl/dump-conntrack | grep "192.168.0.2"

  tcp,orig=(src=192.168.0.2,dst=10.0.0.2,sport=1024,dport=2048),reply=(src=10.0.0.2,
↪dst=192.168.0.2,sport=2048,dport=1024),protoinfo=(state=LAST_ACK)
```

Sending TCP ack segment using scapy (at the "left" scapy session) (flags=0x10 is ack):

```
$ >>> sendp(Ether()/IP(src="192.168.0.2", dst="10.0.0.2")/TCP(sport=1024, dport=2048,␣
↪flags=0x10, seq=103, ack=202), iface="veth_l1")
```

This packet hits flow #1 and flow #5.

conntrack entry:

```
$ sudo ovs-appctl dpctl/dump-conntrack | grep "192.168.0.2"

  tcp,orig=(src=192.168.0.2,dst=10.0.0.2,sport=1024,dport=2048),reply=(src=10.0.0.2,
↪dst=192.168.0.2,sport=2048,dport=1024),protoinfo=(state=TIME_WAIT)
```

3.4.6 Summary

Following table summarizes the TCP segments exchanged against the flow match fields

TCP Segment	ct_state(flow#)
Connection Setup	
192.168.0.2 → 10.0.0.2 [SYN] Seq=0	-trk(#1) then +trk+new(#2)
10.0.0.2 → 192.168.0.2 [SYN, ACK] Seq=0 Ack=1	-trk(#3) then +trk+est(#4)
192.168.0.2 → 10.0.0.2 [ACK] Seq=1 Ack=1	-trk(#1) then +trk+est(#5)
Data Transfer	
192.168.0.2 → 10.0.0.2 [ACK] Seq=1 Ack=1	-trk(#1) then +trk+est(#5)
10.0.0.2 → 192.168.0.2 [ACK] Seq=1 Ack=2	-trk(#3) then +trk+est(#4)
Connection Teardown	
192.168.0.2 → 10.0.0.2 [FIN, ACK] Seq=2 Ack=1	-trk(#1) then +trk+est(#5)
10.0.0.2 → 192.168.0.2 [FIN, ACK] Seq=1 Ack=3	-trk(#3) then +trk+est(#4)
192.168.0.2 → 10.0.0.2 [ACK] Seq=3 Ack=2	-trk(#1) then +trk+est(#5)

Note: Relative sequence number and acknowledgement numbers are shown as captured from tshark.

Flows

```
 (flow #1)
 $ ovs-ofctl add-flow br0 \
    "table=0, priority=50, ct_state=-trk, tcp, in_port=veth_l0, actions=ct(table=0)"

(flow #2)
$ ovs-ofctl add-flow br0 \
    "table=0, priority=50, ct_state=+trk+new, tcp, in_port=veth_l0, actions=ct(commit),
↪veth_r0"

(flow #3)
$ ovs-ofctl add-flow br0 \
    "table=0, priority=50, ct_state=-trk, tcp, in_port=veth_r0, actions=ct(table=0)"

(flow #4)
$ ovs-ofctl add-flow br0 \
    "table=0, priority=50, ct_state=+trk+est, tcp, in_port=veth_r0, actions=veth_l0"

(flow #5)
$ ovs-ofctl add-flow br0 \
    "table=0, priority=50, ct_state=+trk+est, tcp, in_port=veth_l0, actions=veth_r0"
```

CHAPTER

FOUR

HOW-TO GUIDES

Answers to common "How do I?"-style questions. For more information on the topics covered herein, refer to *Deep Dive*.

4.1 OVS

4.1.1 Open vSwitch with KVM

This document describes how to use Open vSwitch with the Kernel-based Virtual Machine (KVM).

> **Note**
>
> This document assumes that you have Open vSwitch set up on a Linux system.

Setup

KVM uses tunctl to handle various bridging modes, which you can install with the Debian/Ubuntu package `uml-utilities`:

```
$ apt-get install uml-utilities
```

Next, you will need to modify or create custom versions of the `qemu-ifup` and `qemu-ifdown` scripts. In this guide, we'll create custom versions that make use of example Open vSwitch bridges that we'll describe in this guide.

Create the following two files and store them in known locations. For example:

```
$ cat << 'EOF' > /etc/ovs-ifup
#!/bin/sh

switch='br0'
ip link set $1 up
ovs-vsctl add-port ${switch} $1
EOF
```

```
$ cat << 'EOF' > /etc/ovs-ifdown
#!/bin/sh

switch='br0'
ip addr flush dev $1
ip link set $1 down
```

(continues on next page)

(continued from previous page)

```
ovs-vsctl del-port ${switch} $1
EOF
```

The basic usage of Open vSwitch is described at the end of *Open vSwitch on Linux, FreeBSD and NetBSD*. If you haven't already, create a bridge named `br0` with the following command:

```
$ ovs-vsctl add-br br0
```

Then, add a port to the bridge for the NIC that you want your guests to communicate over (e.g. `eth0`):

```
$ ovs-vsctl add-port br0 eth0
```

Refer to ovs-vsctl(8) for more details.

Next, we'll start a guest that will use our ifup and ifdown scripts:

```
$ kvm -m 512 -net nic,macaddr=00:11:22:EE:EE:EE -net \
    tap,script=/etc/ovs-ifup,downscript=/etc/ovs-ifdown -drive \
    file=/path/to/disk-image,boot=on
```

This will start the guest and associate a tap device with it. The `ovs-ifup` script will add a port on the br0 bridge so that the guest will be able to communicate over that bridge.

To get some more information and for debugging you can use Open vSwitch utilities such as ovs-dpctl and ovs-ofctl, For example:

```
$ ovs-dpctl show
$ ovs-ofctl show br0
```

You should see tap devices for each KVM guest added as ports to the bridge (e.g. tap0)

Refer to ovs-dpctl(8) and ovs-ofctl(8) for more details.

Bug Reporting

Please report problems to bugs@openvswitch.org.

4.1.2 Encrypt Open vSwitch Tunnels with IPsec

This document gives detailed description on the OVS IPsec tunnel and its configuration modes. If you want to follow a step-by-step guide to run and test IPsec tunnel, please refer to *OVS IPsec Tutorial*.

Overview

Why do encryption?

OVS tunnel packets are transported from one machine to another. Along the path, the packets are processed by physical routers and physical switches. There are risks that these physical devices might read or write the contents of the tunnel packets. IPsec encrypts IP payload and prevents the malicious party sniffing or manipulating the tunnel traffic.

OVS IPsec

OVS IPsec aims to provide a simple interface for user to add encryption on OVS tunnels. It supports GRE, GENEVE, and VXLAN tunnels. The IPsec configuration is done by setting options of the tunnel interface and other_config of Open_vSwitch. You can choose different authentication methods and plaintext tunnel policies based on your requirements.

OVS does not currently provide any support for IPsec encryption for traffic not encapsulated in a tunnel.

Configuration

Authentication Methods

Hosts of the IPsec tunnel need to authenticate each other to build a secure channel. There are three authentication methods:

1) You can use a pre-shared key (PSK) to do authentication. In both hosts, set the same PSK value. This PSK is like your password. You should never reveal it to untrusted parties. This method is easier to use but less secure than the certificate-based methods:

```
$ ovs-vsctl add-port br0 ipsec_gre0 -- \
            set interface ipsec_gre0 type=gre \
                                options:remote_ip=2.2.2.2 \
                                options:psk=swordfish
```

2) You can use a self-signed certificate to do authentication. In each host, generate a certificate and the paired private key. Copy the certificate of the remote host to the local host and configure the OVS as following:

```
$ ovs-vsctl set Open_vSwitch . \
            other_config:certificate=/path/to/local_cert.pem \
            other_config:private_key=/path/to/priv_key.pem
$ ovs-vsctl add-port br0 ipsec_gre0 -- \
            set interface ipsec_gre0 type=gre \
                            options:remote_ip=2.2.2.2 \
                            options:remote_cert=/path/to/remote_cert.pem
```

local_cert.pem is the certificate of the local host. *priv_key.pem* is the private key of the local host. *priv_key.pem* needs to be stored in a secure location. *remote_cert.pem* is the certificate of the remote host.

> **Note**
>
> OVS IPsec requires x.509 version 3 certificate with the subjectAltName DNS field setting the same string as the common name (CN) field. You can follow the tutorial in *OVS IPsec Tutorial* and use ovs-pki(8) to generate compatible certificate and key.
>
> (Before OVS version 2.10.90, ovs-pki(8) did not generate x.509 v3 certificates, so if your existing PKI was generated by an older version, it is not suitable for this purpose.)

3) You can also use CA-signed certificate to do authentication. First, you need to create a CA certificate and sign each host certificate with the CA key (please see *OVS IPsec Tutorial*). Copy the CA certificate to each host and configure the OVS as following:

```
$ ovs-vsctl set Open_vSwitch . \
            other_config:certificate=/path/to/local_cert.pem \
            other_config:private_key=/path/to/priv_key.pem \
            other_config:ca_cert=/path/to/ca_cert.pem
$ ovs-vsctl add-port br0 ipsec_gre0 -- \
            set interface ipsec_gre0 type=gre \
                                options:remote_ip=2.2.2.2 \
                                options:remote_name=remote_cn
```

ca_cert.pem is the CA certificate. You need to set *remote_cn* as the common name (CN) of the remote host's certificate so that only the certificate with the expected CN can be trusted in this connection. It is preferable to use this method than 2) if there are many remote hosts since you don't have to copy every remote certificate to the local host.

> **Note**
>
> When using certificate-based authentication, you should not set psk in the interface options. When using psk-based authentication, you should not set certificate, private_key, ca_cert, remote_cert, and remote_name.

Plaintext Policies

When an IPsec tunnel is configured in this database, multiple independent components take responsibility for implementing it. `ovs-vswitchd` and its datapath handle packet forwarding to the tunnel and a separate daemon pushes the tunnel's IPsec policy configuration to the kernel or other entity that implements it. There is a race: if the former configuration completes before the latter, then packets sent by the local host over the tunnel can be transmitted in plaintext. Using this setting, OVS users can avoid this undesirable situation.

1) The default setting allows unencrypted packets to be sent before IPsec completes negotiation:

```
$ ovs-vsctl add-port br0 ipsec_gre0 -- \
            set interface ipsec_gre0 type=gre \
                          options:remote_ip=2.2.2.2 \
                          options:psk=swordfish
```

This setting should be used only and only if tunnel configuration is static and/or if there is firewall that can drop the plain packets that occasionally leak the tunnel unencrypted on OVSDB (re)configuration events.

2) Setiing ipsec_skb_mark drops unencrypted packets by using skb_mark of tunnel packets:

```
$ ovs-vsctl set Open_vSwitch . other_config:ipsec_skb_mark=0/1
$ ovs-vsctl add-port br0 ipsec_gre0 -- \
            set interface ipsec_gre0 type=gre \
                          options:remote_ip=2.2.2.2 \
                          options:psk=swordfish
```

OVS IPsec drops unencrypted packets which carry the same skb_mark as *ipsec_skb_mark*. By setting the ipsec_skb_mark as 0/1, OVS IPsec prevents all unencrypted tunnel packets leaving the host since the default skb_mark value for tunnel packets are 0. This affects all OVS tunnels including those without IPsec being set up. You can install OpenFlow rules to enable those non-IPsec tunnels by setting the skb_mark of the tunnel traffic as non-zero value.

3) Setting *ipsec_skb_mark* as 1/1 only drops tunnel packets with skb_mark value being 1:

```
$ ovs-vsctl set Open_vSwitch . other_config:ipsec_skb_mark=1/1
$ ovs-vsctl add-port br0 ipsec_gre0 -- \
            set interface ipsec_gre0 type=gre \
                          options:remote_ip=2.2.2.2 \
                          options:psk=swordfish
```

Opposite to 2), this setting passes through unencrypted tunnel packets by default. To drop unencrypted IPsec tunnel traffic, you need to explicitly set skb_mark to a non-zero value for those tunnel traffic by installing OpenFlow rules.

Bug Reporting

If you think you may have found a bug with security implications, like

1) IPsec protected tunnel accepted packets that came unencrypted; OR
2) IPsec protected tunnel allowed packets to leave unencrypted

then please report such bugs according to *Security Process*.

If the bug does not have security implications, then report it according to instructions in *Reporting Bugs*.

4.1.3 Open vSwitch with SELinux

Security-Enhanced Linux (SELinux) is a Linux kernel security module that limits "the malicious things" that certain processes, including OVS, can do to the system in case they get compromised. In our case SELinux basically serves as the "second line of defense" that limits the things that OVS processes are allowed to do. The "first line of defense" is proper input validation that eliminates code paths that could be used by attacker to do any sort of "escape attacks", such as file name escape, shell escape, command line argument escape, buffer escape. Since developers don't always implement proper input validation, then SELinux Access Control's goal is to confine damage of such attacks, if they turned out to be possible.

Besides Type Enforcement there are other SELinux features, but they are out of scope for this document.

Currently there are two SELinux policies for Open vSwitch:

- the one that ships with your Linux distribution (i.e. selinux-policy-targeted package)
- the one that ships with OVS (i.e. openvswitch-selinux-policy package)

Limitations

If Open vSwitch is directly started from command line, then it will run under `unconfined_t` SELinux domain that basically lets daemon to do whatever it likes. This is very important for developers to understand, because they might introduced code in OVS that invokes new system calls that SELinux policy did not anticipate. This means that their feature may have worked out just fine for them. However, if someone else would try to run the same code when Open vSwitch is started through systemctl, then Open vSwitch would get Permission Denied errors.

Currently the only distributions that enforce SELinux on OVS by default are RHEL, CentOS and Fedora. While Ubuntu and Debian also have some SELinux support, they run Open vSwitch under the unrestricted `unconfined` domain. Also, it seems that Ubuntu is leaning towards Apparmor that works slightly differently than SELinux.

SELinux and Open vSwitch are moving targets. What this means is that, if you solely rely on your Linux distribution's SELinux policy, then this policy might not have correctly anticipated that a newer Open vSwitch version needs extra rules to allow behavior. However, if you solely rely on SELinux policy that ships with Open vSwitch, then Open vSwitch developers might not have correctly anticipated the feature set that your SELinux implementation supports.

Installation

Refer to *Fedora, RHEL 7.x Packaging for Open vSwitch* for instructions on how to build all Open vSwitch rpm packages.

Once the package is built, install it on your Linux distribution:

```
$ dnf install openvswitch-selinux-policy-2.4.1-1.el7.centos.noarch.rpm
```

Restart Open vSwitch:

```
$ systemctl restart openvswitch
```

Troubleshooting

When SELinux was implemented some of the standard system utilities acquired `-Z` flag (e.g. `ps -Z`, `ls -Z`). For example, to find out under which SELinux security domain process runs, use:

```
$ ps -AZ | grep ovs-vswitchd
system_u:system_r:openvswitch_t:s0 854 ?    ovs-vswitchd
```

To find out the SELinux label of file or directory, use:

```
$ ls -Z /etc/openvswitch/conf.db
system_u:object_r:openvswitch_rw_t:s0 /etc/openvswitch/conf.db
```

If, for example, SELinux policy for Open vSwitch is too strict, then you might see in Open vSwitch log files "Permission Denied" errors:

```
$ cat /var/log/openvswitch/ovs-vswitchd.log
vlog|INFO|opened log file /var/log/openvswitch/ovs-vswitchd.log
ovs_numa|INFO|Discovered 2 CPU cores on NUMA node 0
ovs_numa|INFO|Discovered 1 NUMA nodes and 2 CPU cores
reconnect|INFO|unix:/var/run/openvswitch/db.sock: connecting...
reconnect|INFO|unix:/var/run/openvswitch/db.sock: connected
netlink_socket|ERR|fcntl: Permission denied
dpif_netlink|ERR|Generic Netlink family 'ovs_datapath' does not exist.
                The Open vSwitch kernel module is probably not loaded.
dpif|WARN|failed to enumerate system datapaths: Permission denied
dpif|WARN|failed to create datapath ovs-system: Permission denied
```

However, not all "Permission denied" errors are caused by SELinux. So, before blaming too strict SELinux policy, make sure that indeed SELinux was the one that denied OVS access to certain resources, for example, run:

```
$ grep "openvswitch_t" /var/log/audit/audit.log | tail
type=AVC msg=audit(1453235431.640:114671): avc:  denied  { getopt } for  pid=4583 comm=
↪"ovs-vswitchd" scontext=system_u:system_r:openvswitch_t:s0 tcontext=system_u:system_
↪r:openvswitch_t:s0 tclass=netlink_generic_socket permissive=0
```

If SELinux denied OVS access to certain resources, then make sure that you have installed our SELinux policy package that "loosens" up distribution's SELinux policy:

```
$ rpm -qa | grep openvswitch-selinux
openvswitch-selinux-policy-2.4.1-1.el7.centos.noarch
```

Then verify that this module was indeed loaded:

```
# semodule -l | grep openvswitch
openvswitch-custom    1.0
openvswitch          1.1.1
```

If you still see Permission denied errors, then take a look into `selinux/openvswitch.te.in` file in the OVS source tree and try to add allow rules. This is really simple, just run SELinux audit2allow tool:

```
$ grep "openvswitch_t" /var/log/audit/audit.log | audit2allow -M ovslocal
```

Contributing SELinux policy patches

Here are few things to consider before proposing SELinux policy patches to Open vSwitch developer mailing list:

1. The SELinux policy that resides in Open vSwitch source tree amends SELinux policy that ships with your distributions.

   Implications of this are that it is assumed that the distribution's Open vSwitch SELinux module must be already loaded to satisfy dependencies.

2. The SELinux policy that resides in Open vSwitch source tree must work on all currently relevant Linux distributions.

   Implications of this are that you should use only those SELinux policy features that are supported by the lowest SELinux version out there. Typically this means that you should test your SELinux policy changes on the oldest RHEL or CentOS version that this OVS version supports. Refer to *Fedora, RHEL 7.x Packaging for Open vSwitch* to find out this.

3. The SELinux policy is enforced only when state transition to `openvswitch_t` domain happens.

   Implications of this are that perhaps instead of loosening SELinux policy you can do certain things at the time rpm package is installed.

Reporting Bugs

Report problems to bugs@openvswitch.org.

4.1.4 Open vSwitch with Libvirt

This document describes how to use Open vSwitch with Libvirt 0.9.11 or later. This document assumes that you followed *Open vSwitch on Linux, FreeBSD and NetBSD* or installed Open vSwitch from distribution packaging such as a .deb or .rpm. The Open vSwitch support is included by default in Libvirt 0.9.11. Consult www.libvirt.org for instructions on how to build the latest Libvirt, if your Linux distribution by default comes with an older Libvirt release.

Limitations

Currently there is no Open vSwitch support for networks that are managed by libvirt (e.g. NAT). As of now, only bridged networks are supported (those where the user has to manually create the bridge).

Setup

First, create the Open vSwitch bridge by using the ovs-vsctl utility (this must be done with administrative privileges):

```
$ ovs-vsctl add-br ovsbr
```

Once that is done, create a VM, if necessary, and edit its Domain XML file:

```
$ virsh edit <vm>
```

Lookup in the Domain XML file the `<interface>` section. There should be one such XML section for each interface the VM has:

```
<interface type='network'>
 <mac address='52:54:00:71:b1:b6'/>
 <source network='default'/>
 <address type='pci' domain='0x0000' bus='0x00' slot='0x03' function='0x0'/>
</interface>
```

And change it to something like this:

```
<interface type='bridge'>
 <mac address='52:54:00:71:b1:b6'/>
 <source bridge='ovsbr'/>
 <virtualport type='openvswitch'/>
 <address type='pci' domain='0x0000' bus='0x00' slot='0x03' function='0x0'/>
</interface>
```

The interface type must be set to `bridge`. The `<source>` XML element specifies to which bridge this interface will be attached to. The `<virtualport>` element indicates that the bridge in `<source>` element is an Open vSwitch bridge.

Then (re)start the VM and verify if the guest's vnet interface is attached to the ovsbr bridge:

```
$ ovs-vsctl show
```

Troubleshooting

If the VM does not want to start, then try to run the libvirtd process either from the terminal, so that all errors are printed in console, or inspect Libvirt/Open vSwitch log files for possible root cause.

Bug Reporting

Report problems to bugs@openvswitch.org.

4.1.5 Open vSwitch with SSL/TLS

If you plan to configure Open vSwitch to connect across the network to an OpenFlow controller, then we recommend that you build Open vSwitch with OpenSSL. SSL/TLS support ensures integrity and confidentiality of the OpenFlow connections, increasing network security.

This document describes how to configure an Open vSwitch to connect to an OpenFlow controller over SSL/TLS. Refer to *Open vSwitch on Linux, FreeBSD and NetBSD*. for instructions on building Open vSwitch with SSL/TLS support.

Open vSwitch uses TLS version 1.2 or later (TLSv1.2), as specified by RFC 5246. TLSv1.2 was released in August 2008, so all current software and hardware should implement it.

This document assumes basic familiarity with public-key cryptography and public-key infrastructure.

SSL/TLS Concepts for OpenFlow

This section is an introduction to the public-key infrastructure architectures that Open vSwitch supports for SSL/TLS authentication.

To connect over SSL/TLS, every Open vSwitch must have a unique private/public key pair and a certificate that signs that public key. Typically, the Open vSwitch generates its own public/private key pair. There are two common ways to obtain a certificate for a switch:

- Self-signed certificates: The Open vSwitch signs its certificate with its own private key. In this case, each switch must be individually approved by the OpenFlow controller(s), since there is no central authority.

  This is the only switch PKI model currently supported by NOX (http://noxrepo.org).

- Switch certificate authority: A certificate authority (the "switch CA") signs each Open vSwitch's public key. The OpenFlow controllers then check that any connecting switches' certificates are signed by that certificate authority.

  This is the only switch PKI model supported by the simple OpenFlow controller included with Open vSwitch.

Each Open vSwitch must also have a copy of the CA certificate for the certificate authority that signs OpenFlow controllers' keys (the "controller CA" certificate). Typically, the same controller CA certificate is installed on all of the

switches within a given administrative unit. There are two common ways for a switch to obtain the controller CA certificate:

- Manually copy the certificate to the switch through some secure means, e.g. using a USB flash drive, or over the network with "scp", or even FTP or HTTP followed by manual verification.
- Open vSwitch "bootstrap" mode, in which Open vSwitch accepts and saves the controller CA certificate that it obtains from the OpenFlow controller on its first connection. Thereafter the switch will only connect to controllers signed by the same CA certificate.

Establishing a Public Key Infrastructure

Open vSwitch can make use of your existing public key infrastructure. If you already have a PKI, you may skip forward to the next section. Otherwise, if you do not have a PKI, the ovs-pki script included with Open vSwitch can help. To create an initial PKI structure, invoke it as:

```
$ ovs-pki init
```

This will create and populate a new PKI directory. The default location for the PKI directory depends on how the Open vSwitch tree was configured (to see the configured default, look for the `--dir` option description in the output of `ovs-pki --help`).

The pki directory contains two important subdirectories. The *controllerca* subdirectory contains controller CA files, including the following:

cacert.pem
: Root certificate for the controller certificate authority. Each Open vSwitch must have a copy of this file to allow it to authenticate valid controllers.

private/cakey.pem
: Private signing key for the controller certificate authority. This file must be kept secret. There is no need for switches or controllers to have a copy of it.

The *switchca* subdirectory contains switch CA files, analogous to those in the *controllerca* subdirectory:

cacert.pem
: Root certificate for the switch certificate authority. The OpenFlow controller must have this file to enable it to authenticate valid switches.

private/cakey.pem
: Private signing key for the switch certificate authority. This file must be kept secret. There is no need for switches or controllers to have a copy of it.

After you create the initial structure, you can create keys and certificates for switches and controllers with ovs-pki. Refer to the ovs-pki(8) manage for complete details. A few examples of its use follow:

Controller Key Generation

To create a controller private key and certificate in files named ctl-privkey.pem and ctl-cert.pem, run the following on the machine that contains the PKI structure:

```
$ ovs-pki req+sign ctl controller
```

ctl-privkey.pem and ctl-cert.pem would need to be copied to the controller for its use at runtime. If, for testing purposes, you were to use ovs-testcontroller, the simple OpenFlow controller included with Open vSwitch, then the –private-key and –certificate options, respectively, would point to these files.

It is very important to make sure that no stray copies of ctl-privkey.pem are created, because they could be used to impersonate the controller.

Switch Key Generation with Self-Signed Certificates

If you are using self-signed certificates (see *SSL/TLS Concepts for OpenFlow*), this is one way to create an acceptable certificate for your controller to approve.

1. Run the following command on the Open vSwitch itself:

```
$ ovs-pki self-sign sc
```

> **Note**
>
> This command does not require a copy of any of the PKI files generated by `ovs-pki init`, and you should not copy them to the switch because some of them have contents that must remain secret for security.)

The `ovs-pki self-sign` command has the following output:

sc-privkey.pem
: the switch private key file. For security, the contents of this file must remain secret. There is ordinarily no need to copy this file off the Open vSwitch.

sc-cert.pem
: the switch certificate, signed by the switch's own private key. Its contents are not a secret.

2. Optionally, copy *controllerca/cacert.pem* from the machine that has the OpenFlow PKI structure and verify that it is correct. (Otherwise, you will have to use CA certificate bootstrapping when you configure Open vSwitch in the next step.)
3. Configure Open vSwitch to use the keys and certificates (see *Configuring SSL/TLS Support*, below).

Switch Key Generation with a Switch PKI (Easy Method)

If you are using a switch PKI (see *SSL/TLS Concepts for OpenFlow*, above), this method of switch key generation is a little easier than the alternate method described below, but it is also a little less secure because it requires copying a sensitive private key from file from the machine hosting the PKI to the switch.

1. Run the following on the machine that contains the PKI structure:

```
$ ovs-pki req+sign sc switch
```

This command has the following output:

sc-privkey.pem
: the switch private key file. For security, the contents of this file must remain secret.

sc-cert.pem
: the switch certificate. Its contents are not a secret.

2. Copy sc-privkey.pem and sc-cert.pem, plus controllerca/cacert.pem, to the Open vSwitch.
3. Delete the copies of sc-privkey.pem and sc-cert.pem on the PKI machine and any other copies that may have been made in transit. It is very important to make sure that there are no stray copies of sc-privkey.pem, because they could be used to impersonate the switch.

> **Warning**
>
> Don't delete controllerca/cacert.pem! It is not security-sensitive and you will need it to configure additional switches.

4. Configure Open vSwitch to use the keys and certificates (see *Configuring SSL/TLS Support*, below).

Switch Key Generation with a Switch PKI (More Secure)

If you are using a switch PKI (see *SSL/TLS Concepts for OpenFlow*, above), then, compared to the previous method, the method described here takes a little more work, but it does not involve copying the private key from one machine to another, so it may also be a little more secure.

1. Run the following command on the Open vSwitch itself:

```
$ ovs-pki req sc
```

> **Note**
>
> This command does not require a copy of any of the PKI files generated by "ovs-pki init", and you should not copy them to the switch because some of them have contents that must remain secret for security.

The "ovs-pki req" command has the following output:

sc-privkey.pem
: the switch private key file. For security, the contents of this file must remain secret. There is ordinarily no need to copy this file off the Open vSwitch.

sc-req.pem
: the switch "certificate request", which is essentially the switch's public key. Its contents are not a secret.

a fingerprint
: this is output on stdout.

2. Write the fingerprint down on a slip of paper and copy *sc-req.pem* to the machine that contains the PKI structure.
3. On the machine that contains the PKI structure, run:

```
$ ovs-pki sign sc switch
```

This command will output a fingerprint to stdout and request that you verify it. Check that it is the same as the fingerprint that you wrote down on the slip of paper before you answer "yes".

`ovs-pki sign` creates a file named *sc-cert.pem*, which is the switch certificate. Its contents are not a secret.

4. Copy the generated *sc-cert.pem*, plus *controllerca/cacert.pem* from the PKI structure, to the Open vSwitch, and verify that they were copied correctly.

   You may delete *sc-cert.pem* from the machine that hosts the PKI structure now, although it is not important that you do so.

> **Warning**
>
> Don't delete *controllerca/cacert.pem*! It is not security-sensitive and you will need it to configure additional switches.

5. Configure Open vSwitch to use the keys and certificates (see *Configuring SSL/TLS Support*, below).

Configuring SSL/TLS Support

SSL/TLS configuration requires three additional configuration files. The first two of these are unique to each Open vSwitch. If you used the instructions above to build your PKI, then these files will be named *sc-privkey.pem* and *sc-cert.pem*, respectively:

- A private key file, which contains the private half of an RSA or DSA key.

  This file can be generated on the Open vSwitch itself, for the greatest security, or it can be generated elsewhere and copied to the Open vSwitch.

  The contents of the private key file are secret and must not be exposed.

- A certificate file, which certifies that the private key is that of a trustworthy Open vSwitch.

  This file has to be generated on a machine that has the private key for the switch certification authority, which should not be an Open vSwitch; ideally, it should be a machine that is not networked at all.

  The certificate file itself is not a secret.

The third configuration file is typically the same across all the switches in a given administrative unit. If you used the instructions above to build your PKI, then this file will be named *cacert.pem*:

- The root certificate for the controller certificate authority. The Open vSwitch verifies it that is authorized to connect to an OpenFlow controller by verifying a signature against this CA certificate.

Once you have these files, configure ovs-vswitchd to use them using the `ovs-vsctl set-ssl` command, e.g.:

```
$ ovs-vsctl set-ssl /etc/openvswitch/sc-privkey.pem \
    /etc/openvswitch/sc-cert.pem /etc/openvswitch/cacert.pem
```

Substitute the correct file names, of course, if they differ from the ones used above. You should use absolute file names (ones that begin with `/`), because ovs-vswitchd's current directory is unrelated to the one from which you run ovs-vsctl.

If you are using self-signed certificates (see *SSL/TLS Concepts for OpenFlow*) and you did not copy controllerca/cacert.pem from the PKI machine to the Open vSwitch, then add the `--bootstrap` option, e.g.:

```
$ ovs-vsctl -- --bootstrap set-ssl /etc/openvswitch/sc-privkey.pem \
    /etc/openvswitch/sc-cert.pem /etc/openvswitch/cacert.pem
```

After you have added all of these configuration keys, you may specify `ssl:` connection methods elsewhere in the configuration database. `tcp:` connection methods are still allowed even after SSL/TLS has been configured, so for security you should use only `ssl:` connections.

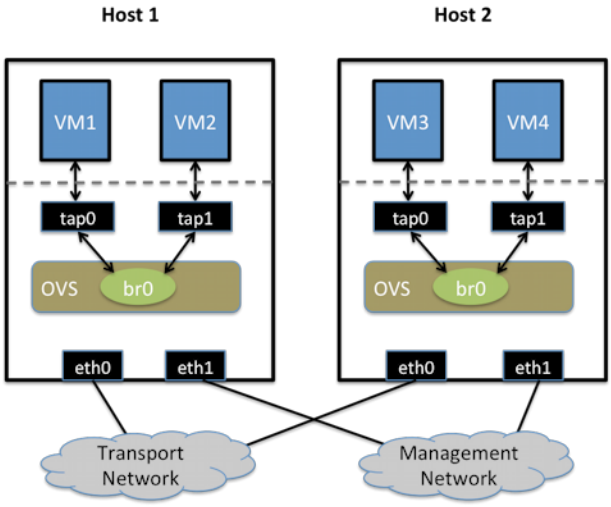
Reporting Bugs

Report problems to bugs@openvswitch.org.

4.1.6 Connecting VMs Using Tunnels

This document describes how to use Open vSwitch to allow VMs on two different hosts to communicate over port-based GRE tunnels.

> **Note**
>
> This guide covers the steps required to configure GRE tunneling. The same approach can be used for any of the other tunneling protocols supported by Open vSwitch.



Setup

This guide assumes the environment is configured as described below.

Two Physical Networks

- Transport Network

  Ethernet network for tunnel traffic between hosts running OVS. Depending on the tunneling protocol being used (this cookbook uses GRE), some configuration of the physical switches may be required (for example, it may be necessary to adjust the MTU). Configuration of the physical switching hardware is outside the scope of this cookbook entry.

- Management Network

  Strictly speaking this network is not required, but it is a simple way to give the physical host an IP address for remote access since an IP address cannot be assigned directly to a physical interface that is part of an OVS bridge.

Two Physical Hosts

The environment assumes the use of two hosts, named *host1* and *host2*. Both hosts are hypervisors running Open vSwitch. Each host has two NICs, *eth0* and *eth1*, which are configured as follows:

- *eth0* is connected to the Transport Network. *eth0* has an IP address that is used to communicate with Host2 over the Transport Network.
- *eth1* is connected to the Management Network. *eth1* has an IP address that is used to reach the physical host for management.

Four Virtual Machines

Each host will run two virtual machines (VMs). *vm1* and *vm2* are running on *host1*, while *vm3* and *vm4* are running on *host2*.

Each VM has a single interface that appears as a Linux device (e.g., `tap0`) on the physical host.

> **Note**
>
> VM interfaces may appear as Linux devices with names like `vnet0`, `vnet1`, etc.

Configuration Steps

Before you begin, you'll want to ensure that you know the IP addresses assigned to *eth0* on both *host1* and *host2*, as they will be needed during the configuration.

Perform the following configuration on *host1*.

1. Create an OVS bridge:

```
$ ovs-vsctl add-br br0
```

> **Note**
>
> You will *not* add *eth0* to the OVS bridge.

2. Boot *vm1* and *vm2* on *host1*. If the VMs are not automatically attached to OVS, add them to the OVS bridge you just created (the commands below assume `tap0` is for *vm1* and `tap1` is for *vm2*):

```
$ ovs-vsctl add-port br0 tap0
$ ovs-vsctl add-port br0 tap1
```

3. Add a port for the GRE tunnel:

```
$ ovs-vsctl add-port br0 gre0 \
    -- set interface gre0 type=gre options:remote_ip=<IP of eth0 on host2>
```

Create a mirrored configuration on *host2* using the same basic steps:

1. Create an OVS bridge, but do not add any physical interfaces to the bridge:

```
$ ovs-vsctl add-br br0
```

2. Launch *vm3* and *vm4* on *host2*, adding them to the OVS bridge if needed (again, `tap0` is assumed to be for *vm3* and `tap1` is assumed to be for *vm4*):

```
$ ovs-vsctl add-port br0 tap0
$ ovs-vsctl add-port br0 tap1
```

3. Create the GRE tunnel on *host2*, this time using the IP address for `eth0` on *host1* when specifying the `remote_ip` option:

```
$ ovs-vsctl add-port br0 gre0 \
  -- set interface gre0 type=gre options:remote_ip=<IP of eth0 on host1>
```

Testing

Pings between any of the VMs should work, regardless of whether the VMs are running on the same host or different hosts.

Using `ip route show` (or equivalent command), the routing table of the operating system running inside the VM should show no knowledge of the IP subnets used by the hosts, only the IP subnet(s) configured within the VM's

operating system. To help illustrate this point, it may be preferable to use very different IP subnet assignments within the guest VMs than what is used on the hosts.

Troubleshooting

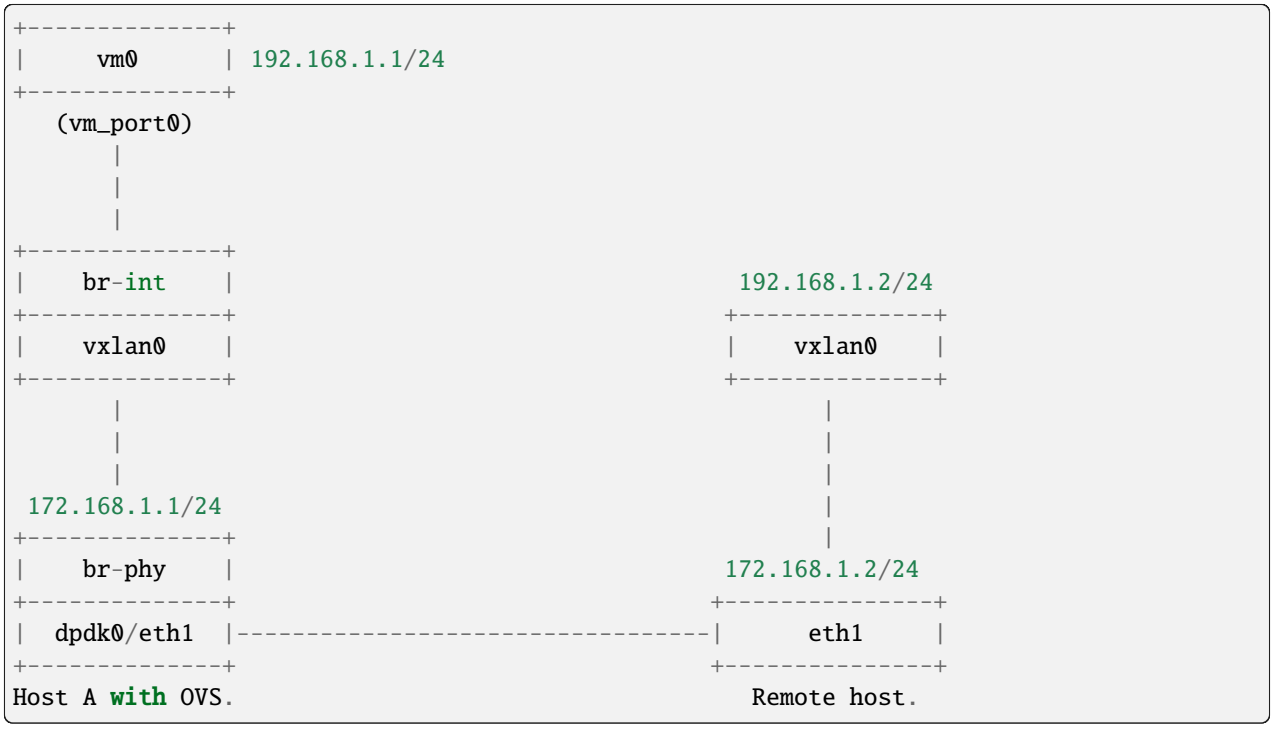
If connectivity between VMs on different hosts isn't working, check the following items:

- Make sure that *host1* and *host2* have full network connectivity over `eth0` (the NIC attached to the Transport Network). This may necessitate the use of additional IP routes or IP routing rules.
- Make sure that `gre0` on *host1* points to `eth0` on *host2*, and that `gre0` on *host2* points to `eth0` on *host1*.
- Ensure that all the VMs are assigned IP addresses on the same subnet; there is no IP routing functionality in this configuration.

4.1.7 Connecting VMs Using Tunnels (Userspace)

This document describes how to use Open vSwitch to allow VMs on two different hosts to communicate over VXLAN tunnels. Unlike *Connecting VMs Using Tunnels*, this configuration works entirely in userspace.

> **Note**
>
> This guide covers the steps required to configure VXLAN tunneling. The same approach can be used for any of the other tunneling protocols supported by Open vSwitch.



Setup

This guide assumes the environment is configured as described below.

Two Physical Hosts

The environment assumes the use of two hosts, named *host1* and *host2*. We only detail the configuration of *host1* but a similar configuration can be used for *host2*. Both hosts should be configured with Open vSwitch (with or without DPDK), QEMU/KVM and suitable VM images. Open vSwitch should be running before proceeding.

Configuration Steps

Perform the following configuration on *host1*:

1. Create a `br-int` bridge:

```
$ ovs-vsctl --may-exist add-br br-int \
  -- set Bridge br-int datapath_type=netdev \
  -- br-set-external-id br-int bridge-id br-int \
  -- set bridge br-int fail-mode=standalone
```

2. Add a port to this bridge. If using tap ports, first boot a VM and then add the port to the bridge:

```
$ ovs-vsctl add-port br-int tap0
```

   If using DPDK vhost-user ports, add the port and then boot the VM accordingly, using `vm_port0` as the interface name:

```
$ ovs-vsctl add-port br-int vm_port0 \
  -- set Interface vm_port0 type=dpdkvhostuserclient \
     options:vhost-server-path=/tmp/vm_port0
```

3. Configure the IP address of the VM interface *in the VM itself*:

```
$ ip addr add 192.168.1.1/24 dev eth0
$ ip link set eth0 up
```

4. On *host1*, add a port for the VXLAN tunnel:

```
$ ovs-vsctl add-port br-int vxlan0 \
  -- set interface vxlan0 type=vxlan options:remote_ip=172.168.1.2
```

   Note

   `172.168.1.2` is the remote tunnel end point address. On the remote host this will be `172.168.1.1`

5. Create a `br-phy` bridge:

```
$ ovs-vsctl --may-exist add-br br-phy \
    -- set Bridge br-phy datapath_type=netdev \
    -- br-set-external-id br-phy bridge-id br-phy \
    -- set bridge br-phy fail-mode=standalone \
         other_config:hwaddr=<mac address of eth1 interface>
```

   Note

   This additional bridge is required when running Open vSwitch in userspace rather than kernel-based Open vSwitch. The purpose of this bridge is to allow use of the kernel network stack for routing and ARP resolution.

The datapath needs to look-up the routing table and ARP table to prepare the tunnel header and transmit data to the output port.

Note

`eth1` is used rather than `eth0`. This is to ensure network connectivity is retained.

6. Attach `eth1/dpdk0` to the `br-phy` bridge.

   If the physical port `eth1` is operating as a kernel network interface, run:

```
$ ovs-vsctl --timeout 10 add-port br-phy eth1
$ ip addr add 172.168.1.1/24 dev br-phy
$ ip link set br-phy up
$ ip addr flush dev eth1 2>/dev/null
$ ip link set eth1 up
$ iptables -F
```

   If instead the interface is a DPDK interface and bound to the `igb_uio` or `vfio` driver, run:

```
$ ovs-vsctl --timeout 10 add-port br-phy dpdk0 \
  -- set Interface dpdk0 type=dpdk options:dpdk-devargs=0000:06:00.0
$ ip addr add 172.168.1.1/24 dev br-phy
$ ip link set br-phy up
$ iptables -F
```

   The commands are different as DPDK interfaces are not managed by the kernel, thus, the port details are not visible to any `ip` commands.

Important

Attempting to use the kernel network commands for a DPDK interface will result in a loss of connectivity through `eth1`. Refer to *Basic Configuration* for more details.

Once complete, check the cached routes using ovs-appctl command:

```
$ ovs-appctl ovs/route/show
```

If the tunnel route is missing, adding it now:

```
$ ovs-appctl ovs/route/add 172.168.1.1/24 br-phy
```

Repeat these steps if necessary for *host2*, but using the below commands for the VM interface IP address:

```
$ ip addr add 192.168.1.2/24 dev eth0
$ ip link set eth0 up
```

And the below command for the the *host2* VXLAN tunnel:

```
$ ovs-vsctl add-port br-int vxlan0 \
  -- set interface vxlan0 type=vxlan options:remote_ip=172.168.1.1
```

Testing

With this setup, ping to VXLAN target device (`192.168.1.2`) should work. Traffic will be VXLAN encapsulated and sent over the `eth1`/`dpdk0` interface.

Tunneling-related Commands

Tunnel routing table

To add route:

```
$ ovs-appctl ovs/route/add <IP address>/<prefix length>
                           <output-bridge-name> <gw> [table=ID]
```

To see all routes configured:

```
$ ovs-appctl ovs/route/show [table=ID|all]
```

To add/delete router rule:

```
$ ovs-appctl ovs/route/rule/add [-6] [not] from=all|IP/PLEN [prio=NUM]
                                table=local|main|default|ID
$ ovs-appctl ovs/route/rule/del [-6] [not] from=all|IP/PLEN [prio=NUM]
                                table=local|main|default|ID
```

To see all router rules configured:

```
$ ovs-appctl ovs/route/rule/show [-6]
```

To delete route:

```
$ ovs-appctl ovs/route/del <IP address>/<prefix length> [table=ID]
```

To look up and display the route for a destination:

```
$ ovs-appctl ovs/route/lookup <IP address> [src=IP]
```

ARP

To see arp cache content:

```
$ ovs-appctl tnl/arp/show
```

To flush arp cache:

```
$ ovs-appctl tnl/arp/flush
```

To set a specific arp entry:

```
$ ovs-appctl tnl/arp/set <bridge> <IP address> <MAC address>
```

Ports

To check tunnel ports listening in ovs-vswitchd:

```
$ ovs-appctl tnl/ports/show
```

To set range for VxLan UDP source port:

```
$ ovs-appctl tnl/egress_port_range <num1> <num2>
```

To show current range:

```
$ ovs-appctl tnl/egress_port_range
```

Datapath

To check datapath ports:

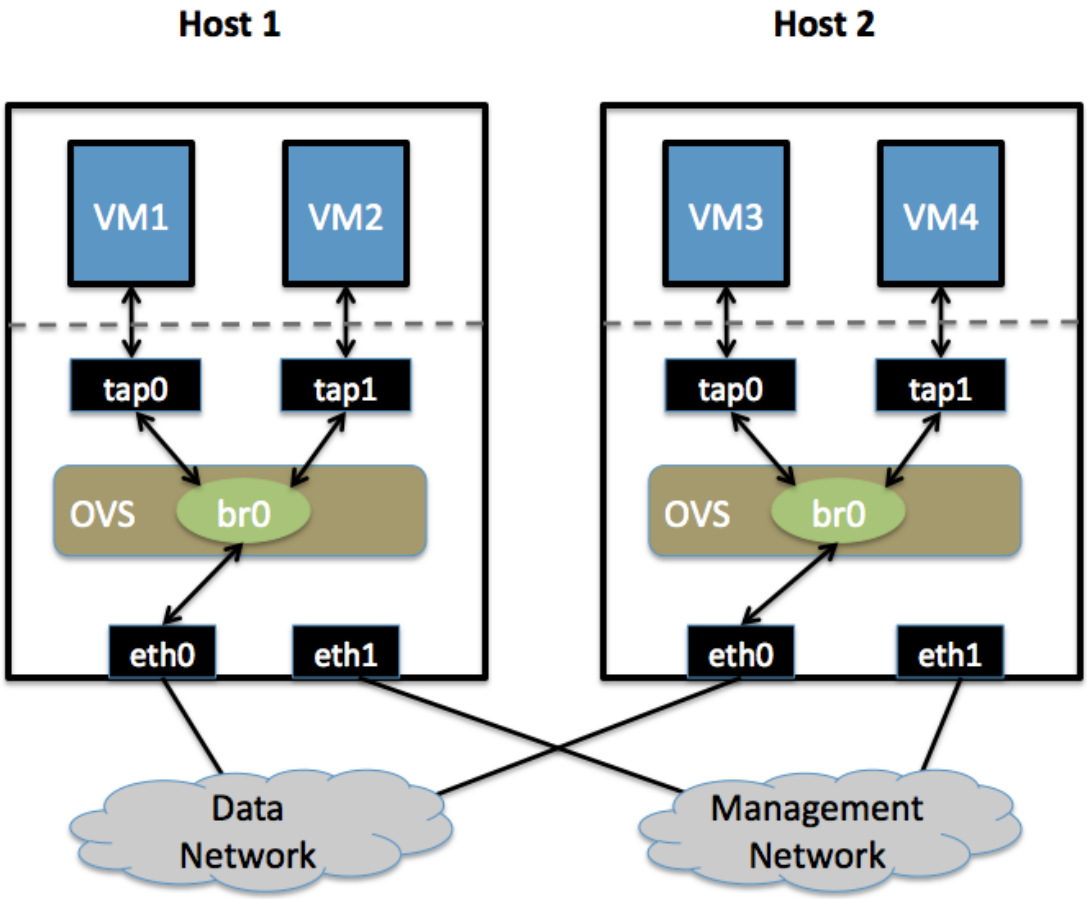
```
$ ovs-appctl dpif/show
```

To check datapath flows:

```
$ ovs-appctl dpif/dump-flows
```

4.1.8 Isolating VM Traffic Using VLANs

This document describes how to use Open vSwitch is to isolate VM traffic using VLANs.



Setup

This guide assumes the environment is configured as described below.

Two Physical Networks

- Data Network

  Ethernet network for VM data traffic, which will carry VLAN-tagged traffic between VMs. Your physical switch(es) must be capable of forwarding VLAN-tagged traffic and the physical switch ports should operate as VLAN trunks. (Usually this is the default behavior. Configuring your physical switching hardware is beyond the scope of this document.)

- Management Network

  This network is not strictly required, but it is a simple way to give the physical host an IP address for remote access, since an IP address cannot be assigned directly to eth0 (more on that in a moment).

Two Physical Hosts

The environment assumes the use of two hosts: *host1* and *host2*. Both hosts are running Open vSwitch. Each host has two NICs, eth0 and eth1, which are configured as follows:

- eth0 is connected to the Data Network. No IP address is assigned to eth0.
- eth1 is connected to the Management Network (if necessary). eth1 has an IP address that is used to reach the physical host for management.

Four Virtual Machines

Each host will run two virtual machines (VMs). *vm1* and *vm2* are running on *host1*, while *vm3* and *vm4* are running on *host2*.

Each VM has a single interface that appears as a Linux device (e.g., `tap0`) on the physical host.

> **Note**
>
> VM interfaces may appear as Linux devices with names like `vnet0`, `vnet1`, etc.

Configuration Steps

Perform the following configuration on *host1*:

1. Create an OVS bridge:

```
$ ovs-vsctl add-br br0
```

2. Add `eth0` to the bridge:

```
$ ovs-vsctl add-port br0 eth0
```

> **Note**
>
> By default, all OVS ports are VLAN trunks, so eth0 will pass all VLANs

> **Note**
>
> When you add eth0 to the OVS bridge, any IP addresses that might have been assigned to eth0 stop working. IP address assigned to eth0 should be migrated to a different interface before adding eth0 to the OVS bridge. This is the reason for the separate management connection via eth1.

3. Add *vm1* as an "access port" on VLAN 100. This means that traffic coming into OVS from VM1 will be untagged and considered part of VLAN 100:

```
$ ovs-vsctl add-port br0 tap0 tag=100
```

   Add VM2 on VLAN 200:

```
$ ovs-vsctl add-port br0 tap1 tag=200
```

Repeat these steps on *host2*:

1. Setup a bridge with eth0 as a VLAN trunk:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 eth0
```

2. Add VM3 to VLAN 100:

```
$ ovs-vsctl add-port br0 tap0 tag=100
```

3. Add VM4 to VLAN 200:

```
$ ovs-vsctl add-port br0 tap1 tag=200
```

Validation

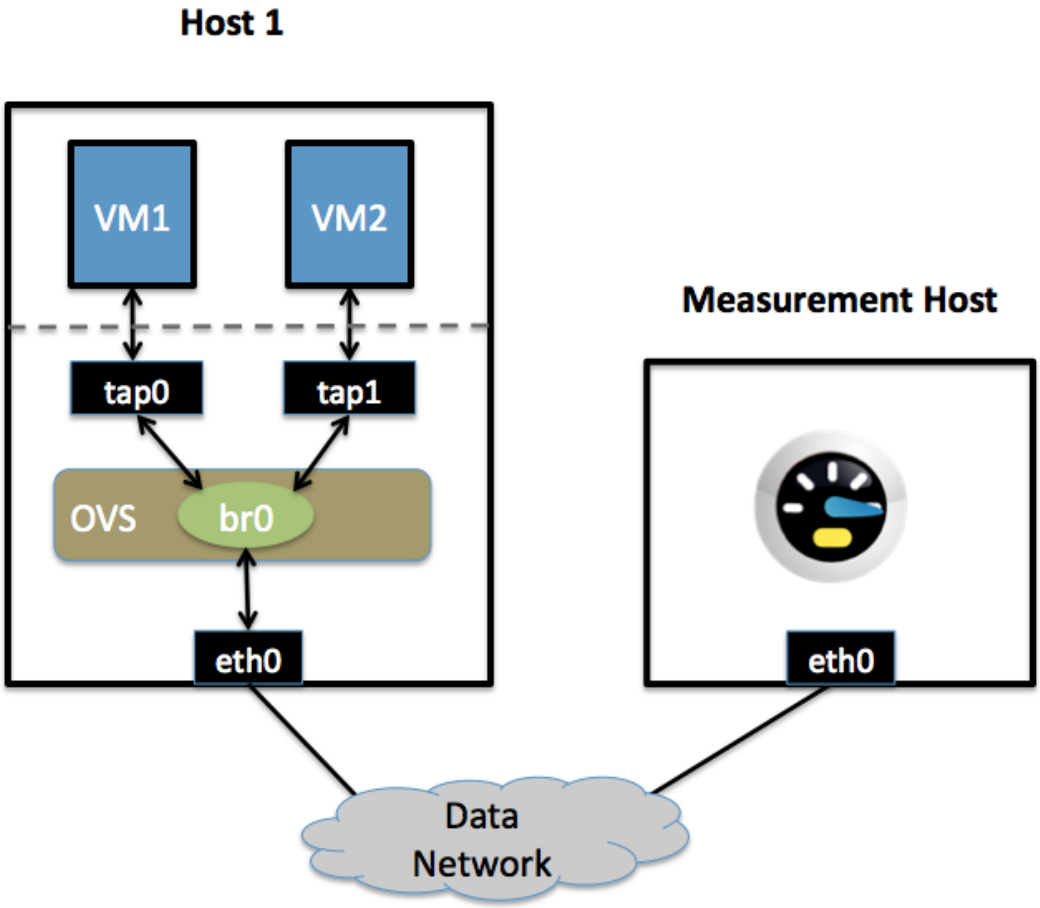
Pings from *vm1* to *vm3* should succeed, as these two VMs are on the same VLAN.

Pings from *vm2* to *vm4* should also succeed, since these VMs are also on the same VLAN as each other.

Pings from *vm1*/*vm3* to *vm2*/*vm4* should not succeed, as these VMs are on different VLANs. If you have a router configured to forward between the VLANs, then pings will work, but packets arriving at *vm3* should have the source MAC address of the router, not of *vm1*.

4.1.9 Quality of Service (QoS) Rate Limiting

This document explains how to use Open vSwitch to rate-limit traffic by a VM to either 1 Mbps or 10 Mbps.



Setup

This guide assumes the environment is configured as described below.

One Physical Network

- Data Network

  Ethernet network for VM data traffic. This network is used to send traffic to and from an external host used for measuring the rate at which a VM is sending. For experimentation, this physical network is optional; you can instead connect all VMs to a bridge that is not connected to a physical interface and use a VM as the measurement host.

There may be other networks (for example, a network for management traffic), but this guide is only concerned with the Data Network.

Two Physical Hosts

The first host, named *host1*, is a hypervisor that runs Open vSwitch and has one NIC. This single NIC, *eth0*, is connected to the Data Network. Because it is participating in an OVS bridge, no IP address can be assigned on *eth0*.

The second host, named Measurement Host, can be any host capable of measuring throughput from a VM. For this guide, we use netperf, a free tool for testing the rate at which one host can send to another. The Measurement Host has

only a single NIC, *eth0*, which is connected to the Data Network. *eth0* has an IP address that can reach any VM on *host1*.

Two VMs

Both VMs (*vm1* and *vm2*) run on *host1*.

Each VM has a single interface that appears as a Linux device (e.g., `tap0`) on the physical host.

> **Note**
>
> VM interfaces may appear as Linux devices with names like `vnet0`, `vnet1`, etc.

Configuration Steps

For both VMs, we modify the Interface table to configure an ingress policing rule. There are two values to set:

`ingress_policing_rate`
: the maximum rate (in Kbps) that this VM should be allowed to send

`ingress_policing_burst`
: a parameter to the policing algorithm to indicate the maximum amount of data (in Kb) that this interface can send beyond the policing rate.

To rate limit VM1 to 1 Mbps, use these commands:

```
$ ovs-vsctl set interface tap0 ingress_policing_rate=1000
$ ovs-vsctl set interface tap0 ingress_policing_burst=100
```

Similarly, to limit *vm2* to 10 Mbps, enter these commands on *host1*:

```
$ ovs-vsctl set interface tap1 ingress_policing_rate=10000
$ ovs-vsctl set interface tap1 ingress_policing_burst=1000
```

To see the current limits applied to VM1, run this command:

```
$ ovs-vsctl list interface tap0
```

Testing

To test the configuration, make sure *netperf* is installed and running on both VMs and on the Measurement Host. *netperf* consists of a client (`netperf`) and a server (`netserver`). In this example, we run `netserver` on the Measurement Host (installing Netperf usually starts `netserver` as a daemon, meaning this is running by default).

For this example, we assume that the Measurement Host has an IP of 10.0.0.100 and is reachable from both VMs.

From *vm1*, run this command:

```
$ netperf -H 10.0.0.100
```

This will cause VM1 to send TCP traffic as quickly as it can to the Measurement Host. After 10 seconds, this will output a series of values. We are interested in the "Throughput" value, which is measured in Mbps (10^6 bits/sec). For VM1 this value should be near 1. Running the same command on VM2 should give a result near 10.

Troubleshooting

Open vSwitch uses the Linux traffic-control capability for rate-limiting. If you are not seeing the configured rate-limit have any effect, make sure that your kernel is built with "ingress qdisc" enabled, and that the user-space utilities (e.g., `/sbin/tc`) are installed.

Additional Information

Open vSwitch's rate-limiting uses policing, which does not queue packets. It drops any packets beyond the specified rate. Specifying a larger burst size lets the algorithm be more forgiving, which is important for protocols like TCP that react severely to dropped packets. Setting a burst size of less than the MTU (e.g., 10 kb) should be avoided.

For TCP traffic, setting a burst size to be a sizeable fraction (e.g., > 10%) of the overall policy rate helps a flow come closer to achieving the full rate. If a burst size is set to be a large fraction of the overall rate, the client will actually experience an average rate slightly higher than the specific policing rate.

For UDP traffic, set the burst size to be slightly greater than the MTU and make sure that your performance tool does not send packets that are larger than your MTU (otherwise these packets will be fragmented, causing poor performance). For example, you can force netperf to send UDP traffic as 1000 byte packets by running:

```
$ netperf -H 10.0.0.100 -t UDP_STREAM -- -m 1000
```

4.1.10 How to Use the VTEP Emulator

This document explains how to use ovs-vtep, a VXLAN Tunnel Endpoint (VTEP) emulator that uses Open vSwitch for forwarding. VTEPs are the entities that handle VXLAN frame encapsulation and decapsulation in a network.

Requirements

The VTEP emulator is a Python script that invokes calls to tools like vtep-ctl and ovs-vsctl. It is only useful when Open vSwitch daemons like ovsdb-server and ovs-vswitchd are running and installed. To do this, either:

- Follow the instructions in *Open vSwitch on Linux, FreeBSD and NetBSD* (don't start any daemons yet).
- Follow the instructions in *Debian Packaging for Open vSwitch* and then install the `openvswitch-vtep` package (if operating on a debian based machine). This will automatically start the daemons.

Design

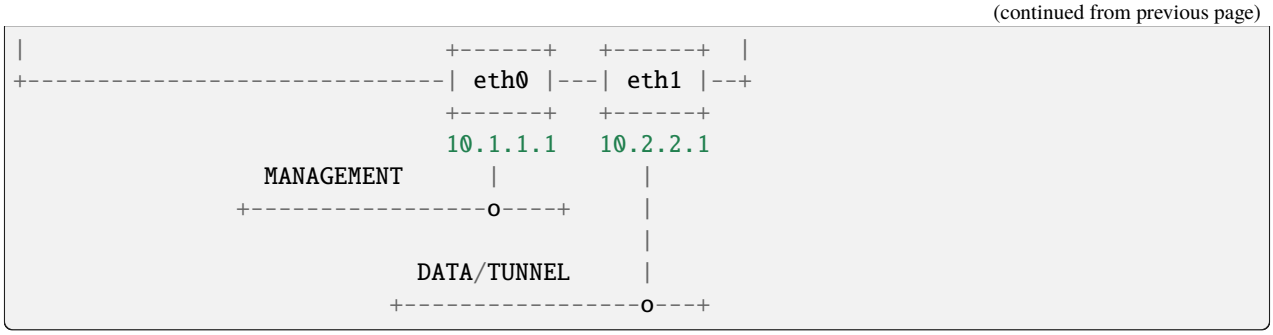
At the end of this process, you should have the following setup:

```
Architecture

+-------------------------------------------------+
| Host Machine                                    |
|                                                 |
|                                                 |
|       +---------+ +---------+                   |
|       |         | |         |                   |
|       |   VM1   | |   VM2   |                   |
|       |         | |         |                   |
|       +----o----+ +----o----+                   |
|            |           |                        |
| br0 +------o-----------o--------------------o--+ |
|           p0          p1                   br0  |
|                                                 |
|                                                 |
```

(continues on next page)

(continued from previous page)



Some important points.

- We will use Open vSwitch to create our "physical" switch labeled `br0`
- Our "physical" switch `br0` will have one internal port also named `br0` and two "physical" ports, namely `p0` and `p1`.
- The host machine may have two external interfaces. We will use `eth0` for management traffic and `eth1` for tunnel traffic (One can use a single interface to achieve both). Please take note of their IP addresses in the diagram. You do not have to use exactly the same IP addresses. Just know that the above will be used in the steps below.
- You can optionally connect physical machines instead of virtual machines to switch `br0`. In that case:
  - Make sure you have two extra physical interfaces in your host machine, `eth2` and `eth3`.
  - In the rest of this doc, replace `p0` with `eth2` and `p1` with `eth3`.

5. In addition to implementing `p0` and `p1` as physical interfaces, you can also optionally implement them as standalone TAP devices, or VM interfaces for simulation.
6. Creating and attaching the VMs is outside the scope of this document and is included in the diagram for reference purposes only.

Startup

These instructions describe how to run with a single ovsdb-server instance that handles both the OVS and VTEP schema. You can skip steps 1-3 if you installed using the debian packages as mentioned in step 2 of the "Requirements" section.

1. Create the initial OVS and VTEP schemas:

```
$ ovsdb-tool create /etc/openvswitch/ovs.db vswitchd/vswitch.ovsschema
$ ovsdb-tool create /etc/openvswitch/vtep.db vtep/vtep.ovsschema
```

2. Start ovsdb-server and have it handle both databases:

```
$ ovsdb-server --pidfile --detach --log-file \
    --remote punix:/var/run/openvswitch/db.sock \
    --remote=db:hardware_vtep,Global,managers \
    /etc/openvswitch/ovs.db /etc/openvswitch/vtep.db
```

3. Start ovs-vswitchd as normal:

```
$ ovs-vswitchd --log-file --detach --pidfile \
    unix:/var/run/openvswitch/db.sock
```

4. Create a "physical" switch and its ports in OVS:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 p0
$ ovs-vsctl add-port br0 p1
```

5. Configure the physical switch in the VTEP database:

```
$ vtep-ctl add-ps br0
$ vtep-ctl set Physical_Switch br0 tunnel_ips=10.2.2.1
```

6. Start the VTEP emulator. If you installed the components following *Open vSwitch on Linux, FreeBSD and NetBSD*, run the following from the `vtep` directory:

```
$ ./ovs-vtep --log-file=/var/log/openvswitch/ovs-vtep.log \
    --pidfile=/var/run/openvswitch/ovs-vtep.pid \
    --detach br0
```

   If the installation was done by installing the openvswitch-vtep package, you can find ovs-vtep at `/usr/share/openvswitch/scripts`.

7. Configure the VTEP database's manager to point at an NVC:

```
$ vtep-ctl set-manager tcp:<CONTROLLER IP>:6640
```

   Where <CONTROLLER IP> is your controller's IP address that is accessible via the Host Machine's eth0 interface.

Simulating an NVC

A VTEP implementation expects to be driven by a Network Virtualization Controller (NVC), such as NSX. If one does not exist, it's possible to use vtep-ctl to simulate one:

1. Create a logical switch:

```
$ vtep-ctl add-ls ls0
```

2. Bind the logical switch to a port:

```
$ vtep-ctl bind-ls br0 p0 0 ls0
$ vtep-ctl set Logical_Switch ls0 tunnel_key=33
```

3. Direct unknown destinations out a tunnel.

   For handling L2 broadcast, multicast and unknown unicast traffic, packets can be sent to all members of a logical switch referenced by a physical switch. The "unknown-dst" address below is used to represent these packets. There are different modes to replicate the packets. The default mode of replication is to send the traffic to a service node, which can be a hypervisor, server or appliance, and let the service node handle replication to other transport nodes (hypervisors or other VTEP physical switches). This mode is called *service node* replication. An alternate mode of replication, called *source node* replication, involves the source node sending to all other transport nodes. Hypervisors are always responsible for doing their own replication for locally attached VMs in both modes. Service node mode is the default. Service node replication mode is considered a basic requirement because it only requires sending the packet to a single transport node. The following configuration is for service node replication mode as only a single transport node destination is specified for the unknown-dst address:

```
$ vtep-ctl add-mcast-remote ls0 unknown-dst 10.2.2.2
```

4. Optionally, change the replication mode from a default of `service_node` to `source_node`, which can be done at the logical switch level:

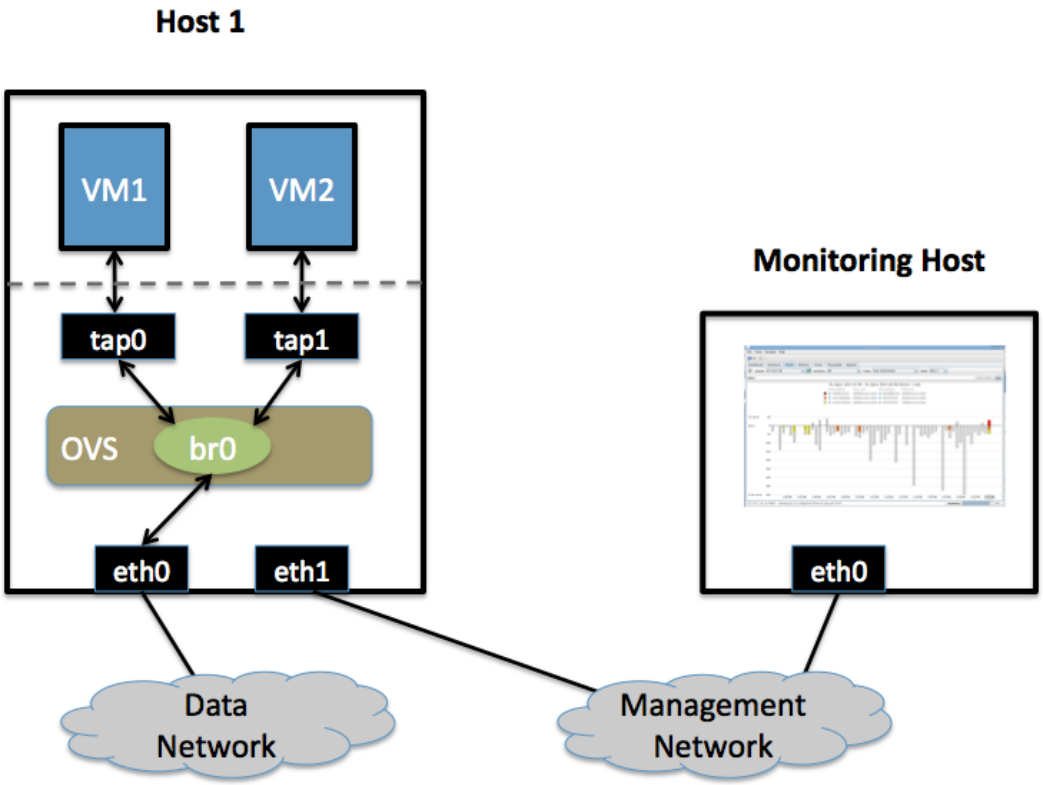
```
$ vtep-ctl set-replication-mode ls0 source_node
```

5. Direct unicast destinations out a different tunnel:

```
$ vtep-ctl add-ucast-remote ls0 00:11:22:33:44:55 10.2.2.3
```

4.1.11 Monitoring VM Traffic Using sFlow

This document describes how to use Open vSwitch is to monitor traffic sent between two VMs on the same host using an sFlow collector. VLANs.



Setup

This guide assumes the environment is configured as described below.

Two Physical Networks

- Data Network

  Ethernet network for VM data traffic. For experimentation, this physical network is optional. You can instead connect all VMs to a bridge that is not connected to a physical interface.

- Management Network

  This network must exist, as it is used to send sFlow data from the agent to the remote collector.

Two Physical Hosts

The environment assumes the use of two hosts: *host1* and *hostMon*. *host* is a hypervisor that run Open vSwitch and has two NICs:

- eth0 is connected to the Data Network. No IP address can be assigned on eth0 because it is part of an OVS bridge.
- eth1 is connected to the Management Network. eth1 has an IP address for management traffic, including sFlow.

hostMon can be any computer that can run the sFlow collector. For this cookbook entry, we use sFlowTrend, a free sFlow collector that is a simple cross-platform Java download. Other sFlow collectors should work equally well. *hostMon* has a single NIC, *eth0*, that is connected to the Management Network. *eth0* has an IP address that can reach *eth1* on *host1*.

Two Virtual Machines

This guide uses two virtual machines - *vm1* and *vm2*- running on *host1*.

> **Note**
>
> VM interfaces may appear as Linux devices with names like `vnet0`, `vnet1`, etc.

Configuration Steps

On *host1*, define the following configuration values in your shell environment:

```
COLLECTOR_IP=10.0.0.1
COLLECTOR_PORT=6343
AGENT_IP=eth1
HEADER_BYTES=128
SAMPLING_N=64
POLLING_SECS=10
```

Port 6343 (`COLLECTOR_PORT`) is the default port number for sFlowTrend. If you are using an sFlow collector other than sFlowTrend, set this value to the appropriate port for your particular collector. Set your own IP address for the collector in the place of 10.0.0.1 (`COLLECTOR_IP`). Setting the `AGENT_IP` value to eth1 indicates that the sFlow agent should send traffic from *eth1*'s IP address. The other values indicate settings regarding the frequency and type of packet sampling that sFlow should perform.

Still on *host1*, run the following command to create an sFlow configuration and attach it to bridge br0:

```
$ ovs-vsctl -- --id=@sflow create sflow agent=${AGENT_IP} \
    target="\"${COLLECTOR_IP}:${COLLECTOR_PORT}\"" header=${HEADER_BYTES} \
    sampling=${SAMPLING_N} polling=${POLLING_SECS} \
      -- set bridge br0 sflow=@sflow
```

Make note of the UUID that is returned by this command; this value is necessary to remove the sFlow configuration.

On *hostMon*, go to the sFlowTrend and click "Install" in the upper right-hand corner. If you have Java installed, this will download and start the sFlowTrend application. Once sFlowTrend is running, the light in the lower right-hand corner of the sFlowTrend application should blink green to indicate that the collector is receiving traffic.

The sFlow configuration is now complete, and sFlowTrend on *hostMon* should be receiving sFlow data from OVS on *host1*.

To configure sFlow on additional bridges, just replace `br0` in the above command with a different bridge name.

To remove sFlow configuration from a bridge (in this case, `br0`), run this command, where "sFlow UUID" is the UUID returned by the command used to set the sFlow configuration initially:

```
$ ovs-vsctl remove bridge br0 sflow <sFlow UUID>
```

To see all current sets of sFlow configuration parameters, run:

```
$ ovs-vsctl list sflow
```

Troubleshooting

If sFlow data isn't being collected and displayed by sFlowTrend, check the following items:

- Make sure the VMs are sending/receiving network traffic over bridge br0, preferably to multiple other hosts and using a variety of protocols.
- To confirm that the agent is sending traffic, check that running the following command shows that the agent on the physical server is sending traffic to the collector IP address (change the port below to match the port your collector is using):

  ```
  $ tcpdump -ni eth1 udp port 6343
  ```

If no traffic is being sent, there is a problem with the configuration of OVS. If traffic is being sent but nothing is visible in the sFlowTrend user interface, this may indicate a configuration problem with the collector.

Check to make sure the host running the collector (*hostMon*) does not have a firewall that would prevent UDP port 6343 from reaching the collector.

Credit

This document is heavily based on content from Neil McKee at InMon:

- https://mail.openvswitch.org/pipermail/ovs-dev/2010-July/165245.html
- https://blog.sflow.com/2010/01/open-vswitch.html

Note

The configuration syntax is out of date, but the high-level descriptions are correct.

4.1.12 Using Open vSwitch with DPDK

This document describes how to use Open vSwitch with DPDK.

Important

Using DPDK with OVS requires configuring OVS at build time to use the DPDK library. The version of DPDK that OVS supports varies from one OVS release to another, as described in the *releases FAQ*. For build instructions refer to *Open vSwitch with DPDK*.

Ports and Bridges

ovs-vsctl can be used to set up bridges and other Open vSwitch features. Bridges should be created with a `datapath_type=netdev`:

```
$ ovs-vsctl add-br br0 -- set bridge br0 datapath_type=netdev
```

ovs-vsctl can also be used to add DPDK devices. ovs-vswitchd should print the number of dpdk devices found in the log file:

```
$ ovs-vsctl add-port br0 dpdk-p0 -- set Interface dpdk-p0 type=dpdk \
    options:dpdk-devargs=0000:01:00.0
$ ovs-vsctl add-port br0 dpdk-p1 -- set Interface dpdk-p1 type=dpdk \
    options:dpdk-devargs=0000:01:00.1
```

Some NICs (i.e. Mellanox ConnectX-3) have only one PCI address associated with multiple ports. Using a PCI device like above won't work. Instead, below usage is suggested:

```
$ ovs-vsctl add-port br0 dpdk-p0 -- set Interface dpdk-p0 type=dpdk \
    options:dpdk-devargs="class=eth,mac=00:11:22:33:44:55"
$ ovs-vsctl add-port br0 dpdk-p1 -- set Interface dpdk-p1 type=dpdk \
    options:dpdk-devargs="class=eth,mac=00:11:22:33:44:56"
```

Important

Using this syntax requires that DPDK probes the device that owns those multiple ports. This can be achieved by either setting an allowlist of PCI devices in the `dpdk-extra` configuration, or by requesting that all available devices (including PCI devices) be probed at initialization (setting `dpdk-probe-at-init` to true).

After the DPDK ports get added to switch, a polling thread continuously polls DPDK devices and consumes 100% of the core, as can be checked from `top` and `ps` commands:

```
$ top -H
$ ps -eLo pid,psr,comm | grep pmd
```

Creating bonds of DPDK interfaces is slightly different to creating bonds of system interfaces. For DPDK, the interface type and devargs must be explicitly set. For example:

```
$ ovs-vsctl add-bond br0 dpdkbond p0 p1 \
    -- set Interface p0 type=dpdk options:dpdk-devargs=0000:01:00.0 \
    -- set Interface p1 type=dpdk options:dpdk-devargs=0000:01:00.1
```

To stop ovs-vswitchd & delete bridge, run:

```
$ ovs-appctl -t ovs-vswitchd exit
$ ovs-appctl -t ovsdb-server exit
$ ovs-vsctl del-br br0
```

OVS with DPDK Inside VMs

Additional configuration is required if you want to run ovs-vswitchd with DPDK backend inside a QEMU virtual machine. ovs-vswitchd creates separate DPDK TX queues for each CPU core available. This operation fails inside QEMU virtual machine because, by default, VirtIO NIC provided to the guest is configured to support only single TX queue and single RX queue. To change this behavior, you need to turn on `mq` (multiqueue) property of all `virtio-net-pci` devices emulated by QEMU and used by DPDK. You may do it manually (by changing QEMU command line) or, if you use Libvirt, by adding the following string to `<interface>` sections of all network devices used by DPDK:

```
<driver name='vhost' queues='N'/>
```

where:

N
: determines how many queues can be used by the guest.

This requires QEMU >= 2.2.

PHY-PHY

Add a userspace bridge and two dpdk (PHY) ports:

```
# Add userspace bridge
$ ovs-vsctl add-br br0 -- set bridge br0 datapath_type=netdev

# Add two dpdk ports
$ ovs-vsctl add-port br0 phy0 -- set Interface phy0 type=dpdk \
      options:dpdk-devargs=0000:01:00.0 ofport_request=1

$ ovs-vsctl add-port br0 phy1 -- set Interface phy1 type=dpdk
      options:dpdk-devargs=0000:01:00.1 ofport_request=2
```

Add test flows to forward packets between DPDK port 0 and port 1:

```
# Clear current flows
$ ovs-ofctl del-flows br0

# Add flows between port 1 (phy0) to port 2 (phy1)
$ ovs-ofctl add-flow br0 in_port=1,action=output:2
$ ovs-ofctl add-flow br0 in_port=2,action=output:1
```

Transmit traffic into either port. You should see it returned via the other.

PHY-VM-PHY (vHost Loopback)

Add a userspace bridge, two dpdk (PHY) ports, and two dpdkvhostuserclient ports:

```
# Add userspace bridge
$ ovs-vsctl add-br br0 -- set bridge br0 datapath_type=netdev

# Add two dpdk ports
$ ovs-vsctl add-port br0 phy0 -- set Interface phy0 type=dpdk \
      options:dpdk-devargs=0000:01:00.0 ofport_request=1

$ ovs-vsctl add-port br0 phy1 -- set Interface phy1 type=dpdk
      options:dpdk-devargs=0000:01:00.1 ofport_request=2

# Add two dpdkvhostuserclient ports
$ ovs-vsctl add-port br0 dpdkvhostclient0 \
    -- set Interface dpdkvhostclient0 type=dpdkvhostuserclient \
       options:vhost-server-path=/tmp/dpdkvhostclient0 ofport_request=3
$ ovs-vsctl add-port br0 dpdkvhostclient1 \
    -- set Interface dpdkvhostclient1 type=dpdkvhostuserclient \
       options:vhost-server-path=/tmp/dpdkvhostclient1 ofport_request=4
```

Add test flows to forward packets between DPDK devices and VM ports:

```
# Clear current flows
$ ovs-ofctl del-flows br0

# Add flows
$ ovs-ofctl add-flow br0 in_port=1,action=output:3
$ ovs-ofctl add-flow br0 in_port=3,action=output:1
$ ovs-ofctl add-flow br0 in_port=4,action=output:2
$ ovs-ofctl add-flow br0 in_port=2,action=output:4

# Dump flows
$ ovs-ofctl dump-flows br0
```

Create a VM using the following configuration:

Configuration	Values	Comments
QEMU version	2.2.0	n/a
QEMU thread affinity	core 5	taskset 0x20
Memory	4GB	n/a
Cores	2	n/a
Qcow2 image	CentOS7	n/a
mrg_rxbuf	off	n/a

You can do this directly with QEMU via the `qemu-system-x86_64` application:

```
$ export VM_NAME=vhost-vm
$ export GUEST_MEM=3072M
$ export QCOW2_IMAGE=/root/CentOS7_x86_64.qcow2
$ export VHOST_SOCK_DIR=/tmp

$ taskset 0x20 qemu-system-x86_64 -name $VM_NAME -cpu host -enable-kvm \
  -m $GUEST_MEM -drive file=$QCOW2_IMAGE --nographic -snapshot \
  -numa node,memdev=mem -mem-prealloc -smp sockets=1,cores=2 \
  -object memory-backend-file,id=mem,size=$GUEST_MEM,mem-path=/dev/hugepages,share=on \
  -chardev socket,id=char0,path=$VHOST_SOCK_DIR/dpdkvhostclient0,server \
  -netdev type=vhost-user,id=mynet1,chardev=char0,vhostforce \
  -device virtio-net-pci,mac=00:00:00:00:00:01,netdev=mynet1,mrg_rxbuf=off \
  -chardev socket,id=char1,path=$VHOST_SOCK_DIR/dpdkvhostclient1,server \
  -netdev type=vhost-user,id=mynet2,chardev=char1,vhostforce \
  -device virtio-net-pci,mac=00:00:00:00:00:02,netdev=mynet2,mrg_rxbuf=off
```

For a explanation of this command, along with alternative approaches such as booting the VM via libvirt, refer to *DPDK vHost User Ports*.

Once the guest is configured and booted, configure DPDK packet forwarding within the guest. To accomplish this, build the `testpmd` application as described in *DPDK in the Guest*. Once compiled, run the application:

```
$ cd $DPDK_DIR/app/test-pmd;
$ ./testpmd -c 0x3 -n 4 --socket-mem 1024 -- \
    --burst=64 -i --txqflags=0xf00 --disable-hw-vlan
$ set fwd mac retry
$ start
```

When you finish testing, bind the vNICs back to kernel:

```
$ $DPDK_DIR/usertools/dpdk-devbind.py --bind=virtio-pci 0000:00:03.0
$ $DPDK_DIR/usertools/dpdk-devbind.py --bind=virtio-pci 0000:00:04.0
```

Note

Valid PCI IDs must be passed in above example. The PCI IDs can be retrieved like so:

```
$ $DPDK_DIR/usertools/dpdk-devbind.py --status
```

More information on the dpdkvhostuserclient ports can be found in *DPDK vHost User Ports*.

PHY-VM-PHY (vHost Loopback) (Kernel Forwarding)

PHY-VM-PHY (vHost Loopback) details steps for PHY-VM-PHY loopback testcase and packet forwarding using DPDK testpmd application in the Guest VM. For users wishing to do packet forwarding using kernel stack below, you need to run the below commands on the guest:

```
$ ip addr add 1.1.1.2/24 dev eth1
$ ip addr add 1.1.2.2/24 dev eth2
$ ip link set eth1 up
$ ip link set eth2 up
$ systemctl stop firewalld.service
$ systemctl stop iptables.service
$ sysctl -w net.ipv4.ip_forward=1
$ sysctl -w net.ipv4.conf.all.rp_filter=0
$ sysctl -w net.ipv4.conf.eth1.rp_filter=0
$ sysctl -w net.ipv4.conf.eth2.rp_filter=0
$ route add -net 1.1.2.0/24 eth2
$ route add -net 1.1.1.0/24 eth1
$ arp -s 1.1.2.99 DE:AD:BE:EF:CA:FE
$ arp -s 1.1.1.99 DE:AD:BE:EF:CA:EE
```

PHY-VM-PHY (vHost Multiqueue)

vHost Multiqueue functionality can also be validated using the PHY-VM-PHY configuration. To begin, follow the steps described in *PHY-PHY* to create and initialize the database, start ovs-vswitchd and add `dpdk`-type devices to bridge `br0`. Once complete, follow the below steps:

1. Configure PMD and RXQs.

   For example, set the number of dpdk port rx queues to at least 2 The number of rx queues at vhost-user interface gets automatically configured after virtio device connection and doesn't need manual configuration:

   ```
   $ ovs-vsctl set Open_vSwitch . other_config:pmd-cpu-mask=0xc
   $ ovs-vsctl set Interface phy0 options:n_rxq=2
   $ ovs-vsctl set Interface phy1 options:n_rxq=2
   ```

2. Instantiate Guest VM using QEMU cmdline

   We must configure with appropriate software versions to ensure this feature is supported.

Table 1: VM Configuration

Setting	Value
QEMU version	2.5.0
QEMU thread affinity	2 cores (taskset 0x30)
Memory	4 GB
Cores	2
Distro	Fedora 22
Multiqueue	Enabled

To do this, instantiate the guest as follows:

```
$ export VM_NAME=vhost-vm
$ export GUEST_MEM=4096M
$ export QCOW2_IMAGE=/root/Fedora22_x86_64.qcow2
$ export VHOST_SOCK_DIR=/tmp
$ taskset 0x30 qemu-system-x86_64 -cpu host -smp 2,cores=2 -m 4096M \
    -drive file=$QCOW2_IMAGE --enable-kvm -name $VM_NAME \
    -nographic -numa node,memdev=mem -mem-prealloc \
    -object memory-backend-file,id=mem,size=$GUEST_MEM,mem-path=/dev/hugepages,
↪share=on \
    -chardev socket,id=char1,path=$VHOST_SOCK_DIR/dpdkvhostclient0,server \
    -netdev type=vhost-user,id=mynet1,chardev=char1,vhostforce,queues=2 \
    -device virtio-net-pci,mac=00:00:00:00:00:01,netdev=mynet1,mq=on,vectors=6 \
    -chardev socket,id=char2,path=$VHOST_SOCK_DIR/dpdkvhostclient1,server \
    -netdev type=vhost-user,id=mynet2,chardev=char2,vhostforce,queues=2 \
    -device virtio-net-pci,mac=00:00:00:00:00:02,netdev=mynet2,mq=on,vectors=6
```

Note

Queue value above should match the queues configured in OVS, The vector value should be set to "number of queues x 2 + 2"

3. Configure the guest interface

   Assuming there are 2 interfaces in the guest named eth0, eth1 check the channel configuration and set the number of combined channels to 2 for virtio devices:

```
$ ethtool -l eth0
$ ethtool -L eth0 combined 2
$ ethtool -L eth1 combined 2
```

   More information can be found in vHost walkthrough section.

4. Configure kernel packet forwarding

   Configure IP and enable interfaces:

```
$ ip addr add 5.5.5.1/24 dev eth0
$ ip addr add 90.90.90.1/24 dev eth1
$ ip link set eth0 up
$ ip link set eth1 up
```

   Configure IP forwarding and add route entries:

```
$ sysctl -w net.ipv4.ip_forward=1
$ sysctl -w net.ipv4.conf.all.rp_filter=0
$ sysctl -w net.ipv4.conf.eth0.rp_filter=0
$ sysctl -w net.ipv4.conf.eth1.rp_filter=0
$ ip route add 2.1.1.0/24 dev eth1
$ route add default gw 2.1.1.2 eth1
$ route add default gw 90.90.90.90 eth1
$ arp -s 90.90.90.90 DE:AD:BE:EF:CA:FE
$ arp -s 2.1.1.2 DE:AD:BE:EF:CA:FA
```

Check traffic on multiple queues:

```
$ cat /proc/interrupts | grep virtio
```

Flow Hardware Offload (Experimental)

The flow hardware offload is disabled by default and can be enabled by:

```
$ ovs-vsctl set Open_vSwitch . other_config:hw-offload=true
```

Matches and actions are programmed into HW to achieve full offload of the flow. If not all actions are supported, fallback to partial flow offload (offloading matches only). Moreover, it only works with PMD drivers that support the configured rte_flow actions. Partial flow offload requires support of "MARK + RSS" actions. Full hardware offload requires support of the actions listed below.

The validated NICs are:

- Mellanox (ConnectX-4, ConnectX-4 Lx, ConnectX-5)
- Napatech (NT200B01)

Supported protocols for hardware offload matches are:

- L2: Ethernet, VLAN
- L3: IPv4, IPv6
- L4: TCP, UDP, SCTP, ICMP

Supported actions for hardware offload are:

- Output.
- Drop.
- Modification of Ethernet (mod_dl_src/mod_dl_dst).
- Modification of IPv4 (mod_nw_src/mod_nw_dst/mod_nw_ttl).
- Modification of TCP/UDP (mod_tp_src/mod_tp_dst).
- VLAN Push/Pop (push_vlan/pop_vlan).
- Modification of IPv6 (set_field:<ADDR>->ipv6_src/ipv6_dst/mod_nw_ttl).
- Clone/output (tnl_push/push_vlan/output) for encapsulating over a tunnel.
- Tunnel pop, for packets received on physical ports.

> **Note**
>
> Tunnel offloads are experimental APIs in DPDK. In order to enable it, compile with -DALLOW_EXPERIMENTAL_API.

Multiprocess

This DPDK feature is not supported and disabled during OVS initialization.

Further Reading

More detailed information can be found in the *DPDK topics section* of the documentation. These guides are listed below.

DPDK Bridges

Bridge must be specially configured to utilize DPDK-backed *physical* and *virtual* ports.

Quick Example

This example demonstrates how to add a bridge that will take advantage of DPDK:

```
$ ovs-vsctl add-br br0 -- set bridge br0 datapath_type=netdev
```

This assumes Open vSwitch has been built with DPDK support. Refer to *Open vSwitch with DPDK* for more information.

Extended & Custom Statistics

The DPDK Extended Statistics API allows PMDs to expose a unique set of statistics. The Extended Statistics are implemented and supported only for DPDK physical and vHost ports. Custom statistics are a dynamic set of counters which can vary depending on the driver. Those statistics are implemented for DPDK physical ports and contain all "dropped", "error" and "management" counters from `XSTATS`. A list of all `XSTATS` counters can be found here.

> **Note**
>
> vHost ports only support RX packet size-based counters. TX packet size counters are not available.

To enable statistics, you have to enable OpenFlow 1.4 support for OVS. To configure a bridge, `br0`, to support OpenFlow version 1.4, run:

```
$ ovs-vsctl set bridge br0 datapath_type=netdev \
  protocols=OpenFlow10,OpenFlow11,OpenFlow12,OpenFlow13,OpenFlow14
```

Once configured, check the OVSDB protocols column in the bridge table to ensure OpenFlow 1.4 support is enabled:

```
$ ovsdb-client dump Bridge protocols
```

You can also query the port statistics by explicitly specifying the `-O OpenFlow14` option:

```
$ ovs-ofctl -O OpenFlow14 dump-ports br0
```

There are custom statistics that OVS accumulates itself and these stats has `ovs_` as prefix. These custom stats are shown along with other stats using the following command:

```
$ ovs-vsctl get Interface <iface> statistics
```

Simple Match Lookup

There are cases where users might want simple forwarding or drop rules for all packets received from a specific port, e.g

```
in_port=1,actions=2
in_port=2,actions=IN_PORT
in_port=3,vlan_tci=0x1234/0x1fff,actions=drop
in_port=4,actions=push_vlan:0x8100,set_field:4196->vlan_vid,output:3
```

There are also cases where complex OpenFlow rules can be simplified down to datapath flows with very simple match criteria.

In theory, for very simple forwarding, OVS doesn't need to parse packets at all in order to follow these rules. In practice, due to various implementation constraints, userspace datapath has to match at least on a small set of packet fields. Some matching criteria (for example, ingress port) are not related to the packet itself and others (for example, VLAN tag or Ethernet type) can be extracted without fully parsing the packet. This allows OVS to significantly speed up packet forwarding for these flows with simple match criteria. Statistics on the number of packets matched in this way can be found in a *Simple Match hits* counter of *ovs-appctl dpif-netdev/pmd-perf-show* command.

EMC Insertion Probability

By default 1 in every 100 flows is inserted into the Exact Match Cache (EMC). It is possible to change this insertion probability by setting the `emc-insert-inv-prob` option:

```
$ ovs-vsctl --no-wait set Open_vSwitch . other_config:emc-insert-inv-prob=N
```

where:

N
: A positive integer representing the inverse probability of insertion, i.e. on average 1 in every `N` packets with a unique flow will generate an EMC insertion.

If `N` is set to 1, an insertion will be performed for every flow. If set to 0, no insertions will be performed and the EMC will effectively be disabled.

With default `N` set to 100, higher megaflow hits will occur initially as observed with pmd stats:

```
$ ovs-appctl dpif-netdev/pmd-perf-show
```

For certain traffic profiles with many parallel flows, it's recommended to set `N` to '0' to achieve higher forwarding performance.

It is also possible to enable/disable EMC on per-port basis using:

```
$ ovs-vsctl set interface <iface> other_config:emc-enable={true,false}
```

> **Note**
>
> This could be useful for cases where different number of flows expected on different ports. For example, if one of the VMs encapsulates traffic using additional headers, it will receive large number of flows but only few flows will come out of this VM. In this scenario it's much faster to use EMC instead of classifier for traffic from the VM, but it's better to disable EMC for the traffic which flows to the VM.

For more information on the EMC refer to *Open vSwitch with DPDK* .

SMC cache

SMC cache or signature match cache is a new cache level after EMC cache. The difference between SMC and EMC is SMC only stores a signature of a flow thus it is much more memory efficient. With same memory space, EMC can store 8k flows while SMC can store 1M flows. When traffic flow count is much larger than EMC size, it is generally beneficial to turn off EMC and turn on SMC. It is currently turned off by default.

To turn on SMC:

```
$ ovs-vsctl --no-wait set Open_vSwitch . other_config:smc-enable=true
```

DPDK Physical Ports

The netdev datapath allows attaching of DPDK-backed physical interfaces in order to provide high-performance ingress/egress from the host.

Important

To use any DPDK-backed interface, you must ensure your bridge is configured correctly. For more information, refer to *DPDK Bridges*.

Changed in version 2.7.0: Before Open vSwitch 2.7.0, it was necessary to prefix port names with a `dpdk` prefix. Starting with 2.7.0, this is no longer necessary.

Quick Example

This example demonstrates how to bind two `dpdk` ports, bound to physical interfaces identified by hardware IDs `0000:01:00.0` and `0000:01:00.1`, to an existing bridge called `br0`:

```
$ ovs-vsctl add-port br0 dpdk-p0 \
   -- set Interface dpdk-p0 type=dpdk options:dpdk-devargs=0000:01:00.0
$ ovs-vsctl add-port br0 dpdk-p1 \
   -- set Interface dpdk-p1 type=dpdk options:dpdk-devargs=0000:01:00.1
```

For the above example to work, the two physical interfaces must be bound to the DPDK poll-mode drivers in userspace rather than the traditional kernel drivers. See the *binding NIC drivers* <*dpdk-binding-nics*> section for details.

Binding NIC Drivers

DPDK operates entirely in userspace and, as a result, requires use of its own poll-mode drivers in user space for physical interfaces and a passthrough-style driver for the devices in kernel space.

There are two different tools for binding drivers: **`driverctl`** which is a generic tool for persistently configuring alternative device drivers, and **`dpdk-devbind`** which is a DPDK-specific tool and whose changes do not persist across reboots. In addition, there are two options available for this kernel space driver - VFIO (Virtual Function I/O) and UIO (Userspace I/O) - along with a number of drivers for each option. We will demonstrate examples of both tools and will use the `vfio-pci` driver, which is the more secure, robust driver of those available. More information can be found in the DPDK drivers documentation.

To list devices using **`driverctl`**, run:

```
$ driverctl -v list-devices | grep -i net
0000:07:00.0 igb (I350 Gigabit Network Connection (Ethernet Server Adapter I350-T2))
0000:07:00.1 igb (I350 Gigabit Network Connection (Ethernet Server Adapter I350-T2))
```

You can then bind one or more of these devices using the same tool:

```
$ driverctl set-override 0000:07:00.0 vfio-pci
```

Alternatively, to list devices using **dpdk-devbind**, run:

```
$ dpdk-devbind --status
Network devices using DPDK-compatible driver
============================================
<none>

Network devices using kernel driver
===================================
0000:07:00.0 'I350 Gigabit Network Connection 1521' if=enp7s0f0 drv=igb unused=igb_uio
0000:07:00.1 'I350 Gigabit Network Connection 1521' if=enp7s0f1 drv=igb unused=igb_uio

Other Network devices
=====================
...
```

Once again, you can then bind one or more of these devices using the same tool:

```
$ dpdk-devbind --bind=vfio-pci 0000:07:00.0
```

Changed in version 2.6.0: Open vSwitch 2.6.0 added support for DPDK 16.07, which in turn renamed the former `dpdk_nic_bind` tool to `dpdk-devbind`.

For more information, refer to the DPDK drivers documentation.

Multiqueue

Poll Mode Driver (PMD) threads are the threads that do the heavy lifting for userspace switching. Correct configuration of PMD threads and the Rx queues they utilize is a requirement in order to deliver the high-performance possible with DPDK acceleration. It is possible to configure multiple Rx queues for `dpdk` ports, thus ensuring this is not a bottleneck for performance. For information on configuring PMD threads, refer to *PMD Threads*.

Traffic Rx Steering

> **Warning**
>
> This feature is experimental.

Some control protocols are used to maintain link status between forwarding engines. In SDN environments, these packets share the same physical network with the user data traffic.

When the system is not sized properly, the PMD threads may not be able to process all incoming traffic from the configured Rx queues. When a signaling packet of such protocols is dropped, it can cause link flapping, worsening the situation.

Some physical NICs can be programmed to put these protocols in a dedicated hardware Rx queue using the rte_flow API.

Warning

This feature is not compatible with all NICs. Refer to the DPDK compatibility matrix and vendor documentation for more details.

Rx steering must be enabled for specific protocols per port. The `rx-steering` option takes one of the following values:

`rss`
: Do regular RSS on all configured Rx queues. This is the default behaviour.

`rss+lacp`
: Do regular RSS on all configured Rx queues. An extra Rx queue is configured for LACP packets (ether type `0x8809`).

Example:

```
$ ovs-vsctl add-port br0 dpdk-p0 -- set Interface dpdk-p0 type=dpdk \
    options:dpdk-devargs=0000:01:00.0 options:n_rxq=2 \
    options:rx-steering=rss+lacp
```

Note

If multiple Rx queues are already configured, regular hash-based RSS (Receive Side Scaling) queue balancing is done on all but the extra Rx queue.

Tip

You can check if Rx steering is supported on a port with the following command:

```
$ ovs-vsctl get interface dpdk-p0 status
{..., rss_queues="0-1", rx_steering_queue="2"}
```

This will also show in `ovs-vswitchd.log`:

```
INFO|dpdk-p0: rx-steering: redirecting lacp traffic to queue 2
INFO|dpdk-p0: rx-steering: applying rss on queues 0-1
```

If the hardware does not support redirecting the specified protocols to a dedicated queue, it will be explicit:

```
$ ovs-vsctl get interface dpdk-p0 status
{..., rx-steering=unsupported}
```

More details can often be found in `ovs-vswitchd.log`:

```
WARN|dpdk-p0: rx-steering: failed to add lacp flow: Unsupported pattern
```

To disable Rx steering on a port, use the following command:

```
$ ovs-vsctl remove Interface dpdk-p0 options rx-steering
```

You can see that it has been disabled in `ovs-vswitchd.log`:

```
INFO|dpdk-p0: rx-steering: default rss
```

> **Warning**
>
> This feature is mutually exclusive with `other-config:hw-offload` as it may conflict with the offloaded flows. If both are enabled, `rx-steering` will fall back to default `rss` mode.

Flow Control

Flow control can be enabled only on DPDK physical ports. To enable flow control support at Tx side while adding a port, run:

```
$ ovs-vsctl add-port br0 dpdk-p0 -- set Interface dpdk-p0 type=dpdk \
    options:dpdk-devargs=0000:01:00.0 options:tx-flow-ctrl=true
```

Similarly, to enable Rx flow control, run:

```
$ ovs-vsctl add-port br0 dpdk-p0 -- set Interface dpdk-p0 type=dpdk \
    options:dpdk-devargs=0000:01:00.0 options:rx-flow-ctrl=true
```

To enable flow control auto-negotiation, run:

```
$ ovs-vsctl add-port br0 dpdk-p0 -- set Interface dpdk-p0 type=dpdk \
    options:dpdk-devargs=0000:01:00.0 options:flow-ctrl-autoneg=true
```

To turn on the Tx flow control at run time for an existing port, run:

```
$ ovs-vsctl set Interface dpdk-p0 options:tx-flow-ctrl=true
```

The flow control parameters can be turned off by setting `false` to the respective parameter. To disable the flow control at Tx side, run:

```
$ ovs-vsctl set Interface dpdk-p0 options:tx-flow-ctrl=false
```

Rx Checksum Offload

By default, DPDK physical ports are enabled with Rx checksum offload.

Rx checksum offload can offer performance improvement only for tunneling traffic in OVS-DPDK because the checksum validation of tunnel packets is offloaded to the NIC. Also enabling Rx checksum may slightly reduce the performance of non-tunnel traffic, specifically for smaller size packet.

Hotplugging

OVS supports port hotplugging, allowing the use of physical ports that were not bound to DPDK when ovs-vswitchd was started.

> **Warning**
>
> This feature is not compatible with all NICs. Refer to vendor documentation for more information.

Important

Ports must be bound to DPDK. Refer to *Binding NIC Drivers* for more information.

To *hotplug* a port, simply add it like any other port:

```
$ ovs-vsctl add-port br0 dpdkx -- set Interface dpdkx type=dpdk \
    options:dpdk-devargs=0000:01:00.0
```

Ports can be detached using the `del-port` command:

```
$ ovs-vsctl del-port dpdkx
```

This should both delete the port and detach the device. If successful, you should see an `INFO` log. For example:

```
INFO|Device '0000:04:00.1' has been detached
```

If the log is not seen then the port can be detached like so:

```
$ ovs-appctl netdev-dpdk/detach 0000:01:00.0
```

Warning

Detaching should not be done if a device is known to be non-detachable, as this may cause the device to behave improperly when added back with add-port. The Chelsio Terminator adapters which use the cxgbe driver seem to be an example of this behavior; check the driver documentation if this is suspected.

Hotplugging with IGB_UIO

Important

As of DPDK v20.11 IGB_UIO has been deprecated and is no longer built as part of the default DPDK library. Below is intended for those who wish to use IGB_UIO outside of the standard DPDK build from v20.11 onwards.

As of DPDK v19.11, default igb_uio hotplugging behavior changed from previous DPDK versions.

From DPDK v19.11 onwards, if no device is bound to igb_uio when OVS is launched then the IOVA mode may be set to virtual addressing for DPDK. This is incompatible for hotplugging with igb_uio.

To hotplug a port with igb_uio in this case, DPDK must be configured to use physical addressing for IOVA mode. For more information regarding IOVA modes in DPDK please refer to the DPDK IOVA Mode Detection.

To configure OVS DPDK to use physical addressing for IOVA:

```
$ ovs-vsctl --no-wait set Open_vSwitch . \
    other_config:dpdk-extra="--iova-mode=pa"
```

Note

Changing IOVA mode requires restarting the ovs-vswitchd application.

Representors

DPDK representors enable configuring a phy port to a guest (VM) machine.

OVS resides in the hypervisor which has one or more physical interfaces also known as the physical functions (PFs). If a PF supports SR-IOV it can be used to enable communication with the VMs via Virtual Functions (VFs). The VFs are virtual PCIe devices created from the physical Ethernet controller.

DPDK models a physical interface as a rte device on top of which an eth device is created. DPDK (version 18.xx) introduced the representors eth devices. A representor device represents the VF eth device (VM side) on the hypervisor side and operates on top of a PF. Representors are multi devices created on top of one PF.

For more information, refer to the DPDK documentation.

Prior to port representors there was a one-to-one relationship between the PF and the eth device. With port representors the relationship becomes one PF to many eth devices. In case of two representors ports, when one of the ports is closed - the PCI bus cannot be detached until the second representor port is closed as well.

When configuring a PF-based port, OVS traditionally assigns the device PCI address in devargs. For an existing bridge called `br0` and PCI address `0000:08:00.0` an `add-port` command is written as:

```
$ ovs-vsctl add-port br0 dpdk-pf -- set Interface dpdk-pf type=dpdk \
   options:dpdk-devargs=0000:08:00.0
```

When configuring a VF-based port, DPDK uses an extended devargs syntax which has the following format:

```
BDBF,representor=<representor identifier>
```

This syntax shows that a representor is an enumerated eth device (with a representor identifier) which uses the PF PCI address. The following commands add representors of VF 3 and 5 using PCI device address `0000:08:00.0`:

```
$ ovs-vsctl add-port br0 dpdk-rep3 -- set Interface dpdk-rep3 type=dpdk \
   options:dpdk-devargs=0000:08:00.0,representor=vf3

$ ovs-vsctl add-port br0 dpdk-rep5 -- set Interface dpdk-rep5 type=dpdk \
   options:dpdk-devargs=0000:08:00.0,representor=vf5
```

Important

Representors ports are configured prior to OVS invocation and independently of it, or by other means as well. Please consult a NIC vendor instructions on how to establish representors.

Intel NICs ixgbe and i40e

In the following example we create one representor on PF address `0000:05:00.0`. Once the NIC is bounded to a DPDK compatible PMD the representor is created:

```
# echo 1 > /sys/bus/pci/devices/0000\:05\:00.0/max_vfs
```

Mellanox NICs ConnectX-4, ConnectX-5 and ConnectX-6

In the following example we create two representors on PF address `0000:05:00.0` and net device name `enp3s0f0`.

- Ensure SR-IOV is enabled on the system.

Enable IOMMU in Linux by adding `intel_iommu=on` to kernel parameters, for example, using GRUB (see /etc/grub/grub.conf).

- Verify the PF PCI address prior to representors creation:

```
# lspci | grep Mellanox
05:00.0 Ethernet controller: Mellanox Technologies MT27700 Family [ConnectX-4]
05:00.1 Ethernet controller: Mellanox Technologies MT27700 Family [ConnectX-4]
```

- Create the two VFs on the compute node:

```
# echo 2 > /sys/class/net/enp3s0f0/device/sriov_numvfs
```

Verify the VFs creation:

```
# lspci | grep Mellanox
05:00.0 Ethernet controller: Mellanox Technologies MT27700 Family [ConnectX-4]
05:00.1 Ethernet controller: Mellanox Technologies MT27700 Family [ConnectX-4]
05:00.2 Ethernet controller: Mellanox Technologies MT27700 Family [ConnectX-4␣
↪Virtual Function]
05:00.3 Ethernet controller: Mellanox Technologies MT27700 Family [ConnectX-4␣
↪Virtual Function]
```

- Unbind the relevant VFs 0000:05:00.2..0000:05:00.3:

```
# echo 0000:05:00.2 > /sys/bus/pci/drivers/mlx5_core/unbind
# echo 0000:05:00.3 > /sys/bus/pci/drivers/mlx5_core/unbind
```

- Change e-switch mode.

The Mellanox NIC has an e-switch on it. Change the e-switch mode from legacy to switchdev using the PF PCI address:

```
# sudo devlink dev eswitch set pci/0000:05:00.0 mode switchdev
```

This will create the VF representors network devices in the host OS.

- After setting the PF to switchdev mode bind back the relevant VFs:

```
# echo 0000:05:00.2 > /sys/bus/pci/drivers/mlx5_core/bind
# echo 0000:05:00.3 > /sys/bus/pci/drivers/mlx5_core/bind
```

- Restart Open vSwitch

To verify representors correct configuration, execute:

```
$ ovs-vsctl show
```

and make sure no errors are indicated.

Port representors are an example of multi devices. There are NICs which support multi devices by other methods than representors for which a generic devargs syntax is used. The generic syntax is based on the device mac address:

```
class=eth,mac=<MAC address>
```

For example, the following command adds a port to a bridge called `br0` using an eth device whose mac address is `00:11:22:33:44:55`:

```
$ ovs-vsctl add-port br0 dpdk-mac -- set Interface dpdk-mac type=dpdk \
   options:dpdk-devargs="class=eth,mac=00:11:22:33:44:55"
```

Representor specific configuration

In some topologies, a VF must be configured before being assigned to a guest (VM) machine. This configuration is done through VF-specific fields in the `options` column of the `Interface` table.

Important

Some DPDK port use bifurcated drivers, which means that a kernel netdevice remains when Open vSwitch is stopped.

In such case, any configuration applied to a VF would remain set on the kernel netdevice, and be inherited from it when Open vSwitch is restarted, even if the options described in this section are unset from Open vSwitch.

- Configure the VF MAC address:

```
$ ovs-vsctl set Interface dpdk-rep0 options:dpdk-vf-mac=00:11:22:33:44:55
```

The requested MAC address is assigned to the port and is listed as part of its options:

```
$ ovs-appctl dpctl/show
[...]
  port 3: dpdk-rep0 (dpdk: ..., dpdk-vf-mac=00:11:22:33:44:55, ...)

$ ovs-vsctl show
[...]
        Port dpdk-rep0
            Interface dpdk-rep0
                type: dpdk
                options: {dpdk-devargs="<representor devargs>", dpdk-vf-mac=
↪"00:11:22:33:44:55"}

$ ovs-vsctl get Interface dpdk-rep0 status
{dpdk-vf-mac="00:11:22:33:44:55", ...}

$ ovs-vsctl list Interface dpdk-rep0 | grep 'mac_in_use\|options'
mac_in_use          : "00:11:22:33:44:55"
options             : {dpdk-devargs="<representor devargs>", dpdk-vf-mac=
↪"00:11:22:33:44:55"}
```

The value listed as `dpdk-vf-mac` is only a request from the user and is possibly not yet applied.

When the requested configuration is successfully applied to the port, this MAC address is then also shown in the column `mac_in_use` of the `Interface` table. On failure however, `mac_in_use` will keep its previous value, which will thus differ from `dpdk-vf-mac`.

Jumbo Frames

DPDK physical ports can be configured to use Jumbo Frames. For more information, refer to *Jumbo Frames*.

Link State Change (LSC) detection configuration

There are two methods to get the information when Link State Change (LSC) happens on a network interface: by polling or interrupt.

Configuring the lsc detection mode has no direct effect on OVS itself, instead it configures the NIC how it should handle link state changes. Processing the link state update request triggered by OVS takes less time using interrupt mode, since

the NIC updates its link state in the background, while in polling mode the link state has to be fetched from the firmware every time to fulfil this request.

Note that not all PMD drivers support LSC interrupts.

The default configuration is interrupt mode. To set polling mode, option `dpdk-lsc-interrupt` has to be set to `false`.

Command to set interrupt mode for a specific interface::
$ ovs-vsctl set interface <iface_name> options:dpdk-lsc-interrupt=true

Command to set polling mode for a specific interface::
$ ovs-vsctl set interface <iface_name> options:dpdk-lsc-interrupt=false

DPDK vHost User Ports

OVS userspace switching supports vHost user ports as a primary way to interact with guests. For more information on vHost User, refer to the QEMU documentation on same.

> **Important**
>
> To use any DPDK-backed interface, you must ensure your bridge is configured correctly. For more information, refer to *DPDK Bridges*.

Quick Example

This example demonstrates how to add two `dpdkvhostuserclient` ports to an existing bridge called `br0`:

```
$ ovs-vsctl add-port br0 dpdkvhostclient0 \
    -- set Interface dpdkvhostclient0 type=dpdkvhostuserclient \
       options:vhost-server-path=/tmp/dpdkvhostclient0
$ ovs-vsctl add-port br0 dpdkvhostclient1 \
    -- set Interface dpdkvhostclient1 type=dpdkvhostuserclient \
       options:vhost-server-path=/tmp/dpdkvhostclient1
```

For the above examples to work, an appropriate server socket must be created at the paths specified (`/tmp/dpdkvhostclient0` and `/tmp/dpdkvhostclient1`). These sockets can be created with QEMU; see the *vhost-user client* section for details.

vhost-user vs. vhost-user-client

Open vSwitch provides two types of vHost User ports:

- vhost-user (`dpdkvhostuser`)
- vhost-user-client (`dpdkvhostuserclient`)

vHost User uses a client-server model. The server creates/manages/destroys the vHost User sockets, and the client connects to the server. Depending on which port type you use, `dpdkvhostuser` or `dpdkvhostuserclient`, a different configuration of the client-server model is used.

For vhost-user ports, Open vSwitch acts as the server and QEMU the client. This means if OVS dies, all VMs **must** be restarted. On the other hand, for vhost-user-client ports, OVS acts as the client and QEMU the server. This means OVS can die and be restarted without issue, and it is also possible to restart an instance itself. For this reason, vhost-user-client ports are the preferred type for all known use cases; the only limitation is that vhost-user client mode ports require QEMU version 2.7. Ports of type vhost-user are currently deprecated and will be removed in a future release.

vhost-user

Important

Use of vhost-user ports requires QEMU >= 2.2; vhost-user ports are *deprecated*.

To use vhost-user ports, you must first add said ports to the switch. DPDK vhost-user ports can have arbitrary names with the exception of forward and backward slashes, which are prohibited. For vhost-user, the port type is `dpdkvhostuser`:

```
$ ovs-vsctl add-port br0 vhost-user-1 -- set Interface vhost-user-1 \
    type=dpdkvhostuser
```

This action creates a socket located at `/usr/local/var/run/openvswitch/vhost-user-1`, which you must provide to your VM on the QEMU command line.

Note

If you wish for the vhost-user sockets to be created in a sub-directory of `/usr/local/var/run/openvswitch`, you may specify this directory in the ovsdb like so:

```
$ ovs-vsctl --no-wait \
    set Open_vSwitch . other_config:vhost-sock-dir=subdir
```

Once the vhost-user ports have been added to the switch, they must be added to the guest. There are two ways to do this: using QEMU directly, or using libvirt.

Note

IOMMU and Post-copy Live Migration are not supported with vhost-user ports.

Adding vhost-user ports to the guest (QEMU)

To begin, you must attach the vhost-user device sockets to the guest. To do this, you must pass the following parameters to QEMU:

```
-chardev socket,id=char1,path=/usr/local/var/run/openvswitch/vhost-user-1
-netdev type=vhost-user,id=mynet1,chardev=char1,vhostforce
-device virtio-net-pci,mac=00:00:00:00:00:01,netdev=mynet1
```

where `vhost-user-1` is the name of the vhost-user port added to the switch.

Repeat the above parameters for multiple devices, changing the chardev `path` and `id` as necessary. Note that a separate and different chardev `path` needs to be specified for each vhost-user device. For example you have a second vhost-user port named `vhost-user-2`, you append your QEMU command line with an additional set of parameters:

```
-chardev socket,id=char2,path=/usr/local/var/run/openvswitch/vhost-user-2
-netdev type=vhost-user,id=mynet2,chardev=char2,vhostforce
-device virtio-net-pci,mac=00:00:00:00:00:02,netdev=mynet2
```

In addition, QEMU must allocate the VM's memory on hugetlbfs. vhost-user ports access a virtio-net device's virtual rings and packet buffers mapping the VM's physical memory on hugetlbfs. To enable vhost-user ports to map the VM's

memory into their process address space, pass the following parameters to QEMU:

```
-object memory-backend-file,id=mem,size=4096M,mem-path=/dev/hugepages,share=on
-numa node,memdev=mem -mem-prealloc
```

Finally, you may wish to enable multiqueue support. This is optional but, should you wish to enable it, run:

```
-chardev socket,id=char2,path=/usr/local/var/run/openvswitch/vhost-user-2
-netdev type=vhost-user,id=mynet2,chardev=char2,vhostforce,queues=$q
-device virtio-net-pci,mac=00:00:00:00:00:02,netdev=mynet2,mq=on,vectors=$v
```

where:

$q
: The number of queues

$v
: The number of vectors, which is $q * 2 + 2

The vhost-user interface will be automatically reconfigured with required number of Rx and Tx queues after connection of virtio device. Manual configuration of `n_rxq` is not supported because OVS will work properly only if `n_rxq` will match number of queues configured in QEMU.

A least two PMDs should be configured for the vswitch when using multiqueue. Using a single PMD will cause traffic to be enqueued to the same vhost queue rather than being distributed among different vhost queues for a vhost-user interface.

If traffic destined for a VM configured with multiqueue arrives to the vswitch via a physical DPDK port, then the number of Rx queues should also be set to at least two for that physical DPDK port. This is required to increase the probability that a different PMD will handle the multiqueue transmission to the guest using a different vhost queue.

If one wishes to use multiple queues for an interface in the guest, the driver in the guest operating system must be configured to do so. It is recommended that the number of queues configured be equal to $q.

For example, this can be done for the Linux kernel virtio-net driver with:

```
$ ethtool -L <DEV> combined <$q>
```

where:

-L
: Changes the numbers of channels of the specified network device

combined
: Changes the number of multi-purpose channels.

Adding vhost-user ports to the guest (libvirt)

To begin, you must change the user and group that qemu runs under, and restart libvirtd.

- In `/etc/libvirt/qemu.conf` add/edit the following lines:

  ```
  user = "root"
  group = "root"
  ```

- Finally, restart the libvirtd process, For example, on Fedora:

  ```
  $ systemctl restart libvirtd.service
  ```

Once complete, instantiate the VM. A sample XML configuration file is provided at the *end of this file*. Save this file, then create a VM using this file:

```
$ virsh create demovm.xml
```

Once created, you can connect to the guest console:

```
$ virsh console demovm
```

The demovm xml configuration is aimed at achieving out of box performance on VM. These enhancements include:

- The vcpus are pinned to the cores of the CPU socket 0 using `vcpupin`.
- Configure NUMA cell and memory shared using `memAccess='shared'`.
- Disable `mrg_rxbuf='off'`

Refer to the libvirt documentation for more information.

vhost-user-client

Important

Use of vhost-user-client ports requires QEMU >= 2.7

To use vhost-user-client ports, you must first add said ports to the switch. Like DPDK vhost-user ports, DPDK vhost-user-client ports can have mostly arbitrary names. However, the name given to the port does not govern the name of the socket device. Instead, this must be configured by the user by way of a `vhost-server-path` option. For vhost-user-client, the port type is `dpdkvhostuserclient`:

```
$ VHOST_USER_SOCKET_PATH=/path/to/socket
$ ovs-vsctl add-port br0 vhost-client-1 \
    -- set Interface vhost-client-1 type=dpdkvhostuserclient \
         options:vhost-server-path=$VHOST_USER_SOCKET_PATH
```

Once the vhost-user-client ports have been added to the switch, they must be added to the guest. Like vhost-user ports, there are two ways to do this: using QEMU directly, or using libvirt. Only the QEMU case is covered here.

Adding vhost-user-client ports to the guest (QEMU)

Attach the vhost-user device sockets to the guest. To do this, you must pass the following parameters to QEMU:

```
-chardev socket,id=char1,path=$VHOST_USER_SOCKET_PATH,server
-netdev type=vhost-user,id=mynet1,chardev=char1,vhostforce
-device virtio-net-pci,mac=00:00:00:00:00:01,netdev=mynet1
```

where `vhost-user-1` is the name of the vhost-user port added to the switch.

If the corresponding `dpdkvhostuserclient` port has not yet been configured in OVS with `vhost-server-path=/path/to/socket`, QEMU will print a log similar to the following:

```
QEMU waiting for connection on: disconnected:unix:/path/to/socket,server
```

QEMU will wait until the port is created successfully in OVS to boot the VM. One benefit of using this mode is the ability for vHost ports to 'reconnect' in event of the switch crashing or being brought down. Once it is brought back up, the vHost ports will reconnect automatically and normal service will resume.

vhost-user-client IOMMU Support

vhost IOMMU is a feature which restricts the vhost memory that a virtio device can access, and as such is useful in deployments in which security is a concern.

IOMMU support may be enabled via a global config value, `vhost-iommu-support`. Setting this to true enables vhost IOMMU support for all vhost ports when/where available:

```
$ ovs-vsctl set Open_vSwitch . other_config:vhost-iommu-support=true
```

The default value is false.

Important

Changing this value requires restarting the daemon.

Important

Enabling the IOMMU feature also enables the vhost user reply-ack protocol; this is known to work on QEMU v2.10.0, but is buggy on older versions (2.7.0 - 2.9.0, inclusive). Consequently, the IOMMU feature is disabled by default (and should remain so if using the aforementioned versions of QEMU). Starting with QEMU v2.9.1, vhost-iommu-support can safely be enabled, even without having an IOMMU device, with no performance penalty.

vhost-user-client Post-copy Live Migration Support (experimental)

`Post-copy` migration is the migration mode where the destination CPUs are started before all the memory has been transferred. The main advantage is the predictable migration time. Mostly used as a second phase after the normal 'pre-copy' migration in case it takes too long to converge.

More information can be found in QEMU docs.

Post-copy support may be enabled via a global config value `vhost-postcopy-support`. Setting this to `true` enables Post-copy support for all vhost-user-client ports:

```
$ ovs-vsctl set Open_vSwitch . other_config:vhost-postcopy-support=true
```

The default value is `false`.

Important

Changing this value requires restarting the daemon.

Important

DPDK Post-copy migration mode uses userfaultfd syscall to communicate with the kernel about page fault handling and uses shared memory based on huge pages. So destination host linux kernel should support userfaultfd over shared hugetlbfs. This feature only introduced in kernel upstream version 4.11.

Post-copy feature supported in DPDK since 18.11.0 version and in QEMU since 2.12.0 version. But it's suggested to use QEMU >= 3.0.1 because migration recovery was fixed for post-copy in 3.0 and few additional bug fixes (like userfaulfd leak) was released in 3.0.1.

DPDK Post-copy feature requires avoiding to populate the guest memory (application must not call mlock* syscall without MCL_ONFAULT). So enabling mlockall is incompatible with post-copy feature in OVS 3.3 and older. Newer versions of OVS only lock memory pages that are faulted in, so both features can be used at the same time.

Note that during migration of vhost-user device, PMD threads hang for the time of faulted pages download from source host. Transferring 1GB hugepage across a 10Gbps link possibly unacceptably slow. So recommended hugepage size is 2MB.

vhost-user-client tx retries config

For vhost-user-client interfaces, the max amount of retries can be changed from the default 8 by setting `tx-retries-max`.

The minimum is 0 which means there will be no retries and if any packets in each batch cannot be sent immediately they will be dropped. The maximum is 32, which would mean that after the first packet(s) in the batch was sent there could be a maximum of 32 more retries.

Retries can help with avoiding packet loss when temporarily unable to send to a vhost interface because the virtqueue is full. However, spending more time retrying to send to one interface, will reduce the time available for rx/tx and processing packets on other interfaces, so some tuning may be required for best performance.

Tx retries max can be set for vhost-user-client ports:

```
$ ovs-vsctl set Interface vhost-client-1 options:tx-retries-max=0
```

Note

Configurable vhost tx retries are not supported with vhost-user ports.

vhost-user-client max queue pairs config

For vhost-user-client interfaces using the VDUSE backend, the maximum number of queue pairs the Virtio device will support can be set at port creation time. If not set, the default value is 1 queue pair. This value is ignored for Vhost-user backends.

Maximum number of queue pairs can be set for vhost-user-client-ports:

```
$ ovs-vsctl add-port br0 vduse0 \
    -- set Interface vduse0 type=dpdkvhostuserclient \
        options:vhost-server-path=/dev/vduse/vduse0 \
        options:vhost-max-queue-pairs=4
```

DPDK in the Guest

The DPDK `testpmd` application can be run in guest VMs for high speed packet forwarding between vhostuser ports. DPDK and testpmd application has to be compiled on the guest VM. Below are the steps for setting up the testpmd application in the VM.

Note

Support for DPDK in the guest requires QEMU >= 2.2

To begin, instantiate a guest as described in *vhost-user* or *vhost-user-client*. Once started, connect to the VM, download the DPDK sources to VM and build DPDK as described in *Installing*.

Setup huge pages and DPDK devices using UIO:

```
$ sysctl vm.nr_hugepages=1024
$ mkdir -p /dev/hugepages
$ mount -t hugetlbfs hugetlbfs /dev/hugepages  # only if not already mounted
$ modprobe uio
$ insmod $DPDK_BUILD/kmod/igb_uio.ko
$ $DPDK_DIR/usertools/dpdk-devbind.py --status
$ $DPDK_DIR/usertools/dpdk-devbind.py -b igb_uio 00:03.0 00:04.0
```

Note

vhost ports pci ids can be retrieved using:

```
lspci | grep Ethernet
```

Finally, start the application:

```
# TODO
```

Sample XML

```
<domain type='kvm'>
  <name>demovm</name>
  <uuid>4a9b3f53-fa2a-47f3-a757-dd87720d9d1d</uuid>
  <memory unit='KiB'>4194304</memory>
  <currentMemory unit='KiB'>4194304</currentMemory>
  <memoryBacking>
    <hugepages>
      <page size='2' unit='M' nodeset='0'/>
    </hugepages>
  </memoryBacking>
  <vcpu placement='static'>2</vcpu>
  <cputune>
    <shares>4096</shares>
    <vcpupin vcpu='0' cpuset='4'/>
    <vcpupin vcpu='1' cpuset='5'/>
    <emulatorpin cpuset='4,5'/>
  </cputune>
  <os>
    <type arch='x86_64' machine='pc'>hvm</type>
    <boot dev='hd'/>
  </os>
  <features>
```

(continues on next page)

(continued from previous page)

```
    <acpi/>
    <apic/>
  </features>
  <cpu mode='host-model'>
    <model fallback='allow'/>
    <topology sockets='2' cores='1' threads='1'/>
    <numa>
      <cell id='0' cpus='0-1' memory='4194304' unit='KiB' memAccess='shared'/>
    </numa>
  </cpu>
  <on_poweroff>destroy</on_poweroff>
  <on_reboot>restart</on_reboot>
  <on_crash>destroy</on_crash>
  <devices>
    <emulator>/usr/bin/qemu-system-x86_64</emulator>
    <disk type='file' device='disk'>
      <driver name='qemu' type='qcow2' cache='none'/>
      <source file='/root/CentOS7_x86_64.qcow2'/>
      <target dev='vda' bus='virtio'/>
    </disk>
    <interface type='vhostuser'>
      <mac address='00:00:00:00:00:01'/>
      <source type='unix' path='/usr/local/var/run/openvswitch/dpdkvhostuser0' mode=
↪'client'/>
       <model type='virtio'/>
      <driver queues='2'>
        <host mrg_rxbuf='on'/>
      </driver>
    </interface>
    <interface type='vhostuser'>
      <mac address='00:00:00:00:00:02'/>
      <source type='unix' path='/usr/local/var/run/openvswitch/dpdkvhostuser1' mode=
↪'client'/>
      <model type='virtio'/>
      <driver queues='2'>
        <host mrg_rxbuf='on'/>
      </driver>
    </interface>
    <serial type='pty'>
      <target port='0'/>
    </serial>
    <console type='pty'>
      <target type='serial' port='0'/>
    </console>
  </devices>
</domain>
```

Jumbo Frames

DPDK vHost User ports can be configured to use Jumbo Frames. For more information, refer to *Jumbo Frames*.

vhost tx retries

When sending a batch of packets to a vhost-user or vhost-user-client interface, it may happen that some but not all of the packets in the batch are able to be sent to the guest. This is often because there is not enough free descriptors in the virtqueue for all the packets in the batch to be sent. In this case there will be a retry, with a default maximum of 8 occurring. If at any time no packets can be sent, it may mean the guest is not accepting packets, so there are no (more) retries.

For information about configuring the maximum amount of tx retries for vhost-user-client interfaces see *vhost-user-client tx retries config*.

Note

Maximum vhost tx batch size is defined by NETDEV_MAX_BURST, and is currently as 32.

Tx Retries may be reduced or even avoided by some external configuration, such as increasing the virtqueue size through the `rx_queue_size` parameter introduced in QEMU 2.7.0 / libvirt 2.3.0:

```
<interface type='vhostuser'>
    <mac address='56:48:4f:53:54:01'/>
    <source type='unix' path='/tmp/dpdkvhostclient0' mode='server'/>
    <model type='virtio'/>
    <driver name='vhost' rx_queue_size='1024' tx_queue_size='1024'/>
    <address type='pci' domain='0x0000' bus='0x00' slot='0x10' function='0x0'/>
</interface>
```

The guest application will also need to provide enough descriptors. For example with `testpmd` the command line argument can be used:

```
--rxd=1024 --txd=1024
```

The guest should also have sufficient cores dedicated for consuming and processing packets at the required rate.

The amount of Tx retries on a vhost-user or vhost-user-client interface can be shown with:

```
$ ovs-vsctl get Interface dpdkvhostclient0 statistics:ovs_tx_retries
```

Further information can be found in the DPDK documentation

DPDK Virtual Devices

DPDK provides drivers for both physical and virtual devices. Physical DPDK devices are added to OVS by specifying a valid PCI address in `dpdk-devargs`. Virtual DPDK devices which do not have PCI addresses can be added using a different format for `dpdk-devargs`.

Important

To use any DPDK-backed interface, you must ensure your bridge is configured correctly. For more information, refer to *DPDK Bridges*.

> **Note**
>
> Not all DPDK virtual PMD drivers have been tested and verified to work.

Added in version 2.7.0.

Quick Example

To add a virtual `dpdk` devices, the `dpdk-devargs` argument should be of the format `eth_<driver_name><x>`, where `x`' is a unique identifier of your choice for the given port. For example to add a `dpdk` port that uses the `null` DPDK PMD driver, run:

```
$ ovs-vsctl add-port br0 null0 -- set Interface null0 type=dpdk \
    options:dpdk-devargs=eth_null0
```

Similarly, to add a `dpdk` port that uses the `af_packet` DPDK PMD driver, run:

```
$ ovs-vsctl add-port br0 myeth0 -- set Interface myeth0 type=dpdk \
    options:dpdk-devargs=eth_af_packet0,iface=eth0
```

More information on the different types of virtual DPDK PMDs can be found in the DPDK documentation.

PMD Threads

Poll Mode Driver (PMD) threads are the threads that do the heavy lifting for userspace switching. They perform tasks such as continuous polling of input ports for packets, classifying packets once received, and executing actions on the packets once they are classified.

PMD threads utilize Receive (Rx) and Transmit (Tx) queues, commonly known as *rxq*s and *txq*s to receive and send packets from/to an interface.

- For physical interfaces, the number of Tx Queues is automatically configured based on the number of PMD thread cores. The number of Rx queues can be configured with:

  ```
  $ ovs-vsctl set Interface <interface_name> options:n_rxq=N
  ```

- For virtual interfaces, the number of Tx and Rx queues are configured by libvirt/QEMU and enabled/disabled in the guest. Refer to :doc:'vhost-user' for more information.

The **`ovs-appctl`** utility provides a number of commands for querying PMD threads and their respective queues. This, and all of the above, is discussed here.

PMD Thread Statistics

To show current stats:

```
$ ovs-appctl dpif-netdev/pmd-perf-show
```

Detailed performance metrics for `pmd-perf-show` can also be enabled:

```
$ ovs-vsctl set Open_vSwitch . other_config:pmd-perf-metrics=true
```

See the ovs-vswitchd(8) manpage for more information.

To clear previous stats:

```
$ ovs-appctl dpif-netdev/pmd-stats-clear
```

Note

PMD stats are cumulative so they should be cleared in order to see how the PMDs are being used with current traffic.

Port/Rx Queue Assignment to PMD Threads

Correct configuration of PMD threads and the Rx queues they utilize is a requirement in order to achieve maximum performance. This is particularly true for enabling things like multiqueue for *physical* and *vhost-user* interfaces.

Rx queues will be assigned to PMD threads by OVS, or they can be manually pinned to PMD threads by the user.

To see the port/Rx queue assignment and current measured usage history of PMD core cycles for each Rx queue:

```
$ ovs-appctl dpif-netdev/pmd-rxq-show
```

Note

By default a history of one minute is recorded and shown for each Rx queue to allow for traffic pattern spikes. Any changes in the Rx queue's PMD core cycles usage, due to traffic pattern or reconfig changes, will take one minute to be fully reflected in the stats by default.

PMD thread usage of an Rx queue can be displayed for a shorter period of time, from the last 5 seconds up to the default 60 seconds in 5 second steps.

To see the port/Rx queue assignment and the last 5 secs of measured usage history of PMD core cycles for each Rx queue:

```
$ ovs-appctl dpif-netdev/pmd-rxq-show -secs 5
```

Changed in version 2.6.0: The `pmd-rxq-show` command was added in OVS 2.6.0.

Changed in version 2.16.0: A `overhead` statistics is shown per PMD: it represents the number of cycles inherently consumed by the OVS PMD processing loop.

Changed in version 3.1.0: The `-secs` parameter was added to the dpif-netdev/pmd-rxq-show command.

Rx queue to PMD assignment takes place whenever there are configuration changes or can be triggered by using:

```
$ ovs-appctl dpif-netdev/pmd-rxq-rebalance
```

Changed in version 2.9.0: Utilization-based allocation of Rx queues to PMDs and the `pmd-rxq-rebalance` command were added in OVS 2.9.0. Prior to this, allocation was round-robin and processing cycles were not taken into consideration.

In addition, the output of `pmd-rxq-show` was modified to include Rx queue utilization of the PMD as a percentage.

Port/Rx Queue assignment to PMD threads by manual pinning

Rx queues may be manually pinned to cores. This will change the default Rx queue assignment to PMD threads:

```
$ ovs-vsctl set Interface <iface> \
    other_config:pmd-rxq-affinity=<rxq-affinity-list>
```

where:

- <rxq-affinity-list> is a CSV list of <queue-id>:<core-id> values

For example:

```
$ ovs-vsctl set interface dpdk-p0 options:n_rxq=4 \
    other_config:pmd-rxq-affinity="0:3,1:7,3:8"
```

This will ensure there are *4* Rx queues for dpdk-p0 and that these queues are configured like so:

- Queue #0 pinned to core 3
- Queue #1 pinned to core 7
- Queue #2 not pinned
- Queue #3 pinned to core 8

PMD threads on cores where Rx queues are *pinned* will become *isolated* by default. This means that these threads will only poll the *pinned* Rx queues.

If using pmd-rxq-assign=group PMD threads with *pinned* Rxqs can be *non-isolated* by setting:

```
$ ovs-vsctl set Open_vSwitch . other_config:pmd-rxq-isolate=false
```

Warning

If there are no *non-isolated* PMD threads, *non-pinned* RX queues will not be polled. If the provided <core-id> is not available (e.g. the <core-id> is not in pmd-cpu-mask), the RX queue will be assigned to a *non-isolated* PMD, that will remain *non-isolated*.

Automatic Port/Rx Queue assignment to PMD threads

If pmd-rxq-affinity is not set for Rx queues, they will be assigned to PMDs (cores) automatically.

The algorithm used to automatically assign Rxqs to PMDs can be set by:

```
$ ovs-vsctl set Open_vSwitch . other_config:pmd-rxq-assign=<assignment>
```

By default, cycles assignment is used where the Rxqs will be ordered by their measured processing cycles, and then be evenly assigned in descending order to PMDs. The PMD that will be selected for a given Rxq will be the next one in alternating ascending/descending order based on core id. For example, where there are five Rx queues and three cores - 3, 7, and 8 - available and the measured usage of core cycles per Rx queue over the last interval is seen to be:

- Queue #0: 30%
- Queue #1: 80%
- Queue #3: 60%
- Queue #4: 70%
- Queue #5: 10%

The Rx queues will be assigned to the cores in the following order:

```
Core 3: Q1 (80%) |
Core 7: Q4 (70%) | Q5 (10%)
Core 8: Q3 (60%) | Q0 (30%)
```

`group` assignment is similar to `cycles` in that the Rxqs will be ordered by their measured processing cycles before being assigned to PMDs. It differs from `cycles` in that it uses a running estimate of the cycles that will be on each PMD to select the PMD with the lowest load for each Rxq.

This means that there can be a group of low traffic Rxqs on one PMD, while a high traffic Rxq may have a PMD to itself. Where `cycles` kept as close to the same number of Rxqs per PMD as possible, with `group` this restriction is removed for a better balance of the workload across PMDs.

For example, where there are five Rx queues and three cores - 3, 7, and 8 - available and the measured usage of core cycles per Rx queue over the last interval is seen to be:

- Queue #0: 10%
- Queue #1: 80%
- Queue #3: 50%
- Queue #4: 70%
- Queue #5: 10%

The Rx queues will be assigned to the cores in the following order:

```
Core 3: Q1 (80%) |
Core 7: Q4 (70%) |
Core 8: Q3 (50%) | Q0 (10%) | Q5 (10%)
```

Alternatively, `roundrobin` assignment can be used, where the Rxqs are assigned to PMDs in a round-robin fashion. This algorithm was used by default prior to OVS 2.9. For example, given the following ports and queues:

- Port #0 Queue #0 (P0Q0)
- Port #0 Queue #1 (P0Q1)
- Port #1 Queue #0 (P1Q0)
- Port #1 Queue #1 (P1Q1)
- Port #1 Queue #2 (P1Q2)

The Rx queues may be assigned to the cores in the following order:

```
Core 3: P0Q0 | P1Q1
Core 7: P0Q1 | P1Q2
Core 8: P1Q0 |
```

PMD Automatic Load Balance

Cycle or utilization based allocation of Rx queues to PMDs is done to give an efficient load distribution based at the time of assignment. However, over time it may become less efficient due to changes in traffic. This may cause an uneven load among the PMDs, which in the worst case may result in packet drops and lower throughput.

To address this, automatic load balancing of PMDs can be enabled by:

```
$ ovs-vsctl set open_vswitch . other_config:pmd-auto-lb="true"
```

The following are minimum configuration pre-requisites needed for PMD Auto Load Balancing to operate:

1. `pmd-auto-lb` is enabled.
2. `cycle` (default) or `group` based Rx queue assignment is selected.
3. There are two or more non-isolated PMDs present.

4. At least one non-isolated PMD is polling more than one Rx queue.

When PMD Auto Load Balance is enabled, a PMD core's CPU utilization percentage is measured. The PMD is considered above the threshold if that percentage utilization is greater than the load threshold every 10 secs for 1 minute.

The load threshold can be set by the user. For example, to set the load threshold to 70% utilization of a PMD core:

```
$ ovs-vsctl set open_vswitch .\
    other_config:pmd-auto-lb-load-threshold="70"
```

If not set, the default load threshold is 95%.

If a PMD core is detected to be above the load threshold and the minimum pre-requisites are met, a dry-run using the current PMD assignment algorithm is performed.

For each numa node, the current variance of load between the PMD cores and estimated variance from the dry-run are both calculated. If any numa's estimated dry-run variance is improved from the current one by the variance threshold, a new Rx queue to PMD assignment will be performed.

For example, to set the variance improvement threshold to 40%:

```
$ ovs-vsctl set open_vswitch .\
    other_config:pmd-auto-lb-improvement-threshold="40"
```

If not set, the default variance improvement threshold is 25%.

Note

PMD Auto Load Balancing will not operate if Rx queues are assigned to PMD cores on a different NUMA. This is because the processing load could change after a new assignment due to differing cross-NUMA datapaths, making it difficult to estimate the loads during a dry-run. The only exception is when all PMD threads are running on cores from a single NUMA node. In this case cross-NUMA datapaths will not change after reassignment.

The minimum time between 2 consecutive PMD auto load balancing iterations can also be configured by:

```
$ ovs-vsctl set open_vswitch .\
    other_config:pmd-auto-lb-rebal-interval="<interval>"
```

where `<interval>` is a value in minutes. The default interval is 1 minute.

A user can use this option to set a minimum frequency of Rx queue to PMD reassignment due to PMD Auto Load Balance. For example, this could be set (in min) such that a reassignment is triggered at most every few hours.

PMD load based sleeping

PMD threads constantly poll Rx queues which are assigned to them. In order to reduce the CPU cycles they use, they can sleep for small periods of time when there is no load or very-low load on all the Rx queues they poll.

This can be enabled by setting the max requested sleep time (in microseconds) for a PMD thread:

```
$ ovs-vsctl set open_vswitch . other_config:pmd-sleep-max=50
```

> **Note**
>
> Previous config name 'pmd-maxsleep' is deprecated and will be removed in a future release.

With a non-zero max value a PMD may request to sleep by an incrementing amount of time up to the maximum time. If at any point the threshold of at least half a batch of packets (i.e. 16) is received from an Rx queue that the PMD is polling is met, the requested sleep time will be reset to 0. At that point no sleeps will occur until the no/low load conditions return.

Sleeping in a PMD thread will mean there is a period of time when the PMD thread will not process packets. Sleep times requested are not guaranteed and can differ significantly depending on system configuration. The actual time not processing packets will be determined by the sleep and processor wake-up times and should be tested with each system configuration.

Sleep time statistics for 10 secs can be seen with:

```
$ ovs-appctl dpif-netdev/pmd-stats-clear \
    && sleep 10 && ovs-appctl dpif-netdev/pmd-perf-show
```

Example output, showing that during the last 10 seconds, 74.5% of iterations had a sleep of some length. The total amount of sleep time was 9.06 seconds and the average sleep time where a sleep was requested was 9 microseconds:

```
- sleep iterations:       977037  ( 74.5 % of iterations)
Sleep time (us):         9068841  (  9 us/iteration avg.)
```

Any potential power saving from PMD load based sleeping is dependent on the system configuration (e.g. enabling processor C-states) and workloads.

> **Note**
>
> If there is a sudden spike of packets while the PMD thread is sleeping and the processor is in a low-power state it may result in some lost packets or extra latency before the PMD thread returns to processing packets at full rate.

Maximum sleep values can also be set for individual PMD threads using key:value pairs in the form of core:max_sleep. Any PMD thread that has been assigned a specified value will use that. Any PMD thread that does not have a specified value will use the current global value.

Specified values for individual PMD threads can be added or removed at any time.

For example, to set PMD threads on cores 8 and 9 to never request a load based sleep and all others PMD threads to be able to request a max sleep of 50 microseconds (us):

```
$ ovs-vsctl set open_vswitch . other_config:pmd-sleep-max=50,8:0,9:0
```

The max sleep value for each PMD thread can be checked in the logs or with:

```
$ ovs-appctl dpif-netdev/pmd-sleep-show
pmd thread numa_id 0 core_id 8:
  max sleep:    0 us
pmd thread numa_id 1 core_id 9:
  max sleep:    0 us
pmd thread numa_id 0 core_id 10:
  max sleep:   50 us
pmd thread numa_id 1 core_id 11:
```

(continues on next page)

(continued from previous page)

```
  max sleep:   50 us
pmd thread numa_id 0 core_id 12:
  max sleep:   50 us
pmd thread numa_id 1 core_id 13:
  max sleep:   50 us
```

Quality of Service (QoS)

It is possible to apply both ingress and egress limiting when using the DPDK datapath. These are referred to as *QoS* and *Rate Limiting*, respectively.

Added in version 2.7.0.

QoS (Egress Policing)

Single Queue Policer

Assuming you have a *vhost-user port* transmitting traffic consisting of packets of size 64 bytes, the following command would limit the egress transmission rate of the port to ~1,000,000 packets per second:

```
$ ovs-vsctl set port vhost-user0 qos=@newqos -- \
    --id=@newqos create qos type=egress-policer other-config:cir=46000000 \
    other-config:cbs=2048`
```

To examine the QoS configuration of the port, run:

```
$ ovs-appctl -t ovs-vswitchd qos/show vhost-user0
```

To clear the QoS configuration from the port and ovsdb, run:

```
$ ovs-vsctl destroy QoS vhost-user0 -- clear Port vhost-user0 qos
```

Multi Queue Policer

In addition to the egress-policer OVS-DPDK also has support for a RFC 4115's Two-Rate, Three-Color marker meter. It's a two-level hierarchical policer which first does a color-blind marking of the traffic at the queue level, followed by a color-aware marking at the port level. At the end traffic marked as Green or Yellow is forwarded, Red is dropped. For details on how traffic is marked, see RFC 4115.

This egress policer can be used to limit traffic at different rated based on the queues the traffic is in. In addition, it can also be used to prioritize certain traffic over others at a port level.

For example, the following configuration will limit the traffic rate at a port level to a maximum of 2000 packets a second (64 bytes IPv4 packets). 1000pps as CIR (Committed Information Rate) and 1000pps as EIR (Excess Information Rate). CIR and EIR are measured in bytes without Ethernet header. As a result, 1000pps means (64-byte - 14-byte) * 1000 = 50,000 in the configuration below. High priority traffic is routed to queue 10, which marks all traffic as CIR, i.e. Green. All low priority traffic, queue 20, is marked as EIR, i.e. Yellow:

```
$ ovs-vsctl --timeout=5 set port dpdk1 qos=@myqos -- \
    --id=@myqos create qos type=trtcm-policer \
    other-config:cir=50000 other-config:cbs=2048 \
    other-config:eir=50000 other-config:ebs=2048  \
    queues:10=@dpdk1Q10 queues:20=@dpdk1Q20 -- \
```

(continues on next page)

(continued from previous page)

```
    --id=@dpdk1Q10 create queue \
     other-config:cir=100000 other-config:cbs=2048 \
     other-config:eir=0 other-config:ebs=0 -- \
    --id=@dpdk1Q20 create queue \
      other-config:cir=0 other-config:cbs=0 \
      other-config:eir=50000 other-config:ebs=2048
```

This configuration accomplishes that the high priority traffic has a guaranteed bandwidth egressing the ports at CIR (1000pps), but it can also use the EIR, so a total of 2000pps at max. These additional 1000pps is shared with the low priority traffic. The low priority traffic can use at maximum 1000pps.

Refer to `vswitch.xml` for more details on egress policer.

Rate Limiting (Ingress Policing)

Assuming you have a *vhost-user port* receiving traffic consisting of packets of size 64 bytes, the following command would limit the reception rate of the port to ~1,000,000 packets per second:

```
$ ovs-vsctl set interface vhost-user0 ingress_policing_rate=368000 \
    ingress_policing_burst=1000`
```

To examine the ingress policer configuration of the port:

```
$ ovs-vsctl list interface vhost-user0
```

To clear the ingress policer configuration from the port:

```
$ ovs-vsctl set interface vhost-user0 ingress_policing_rate=0
```

Refer to `vswitch.xml` for more details on ingress policer.

Flow Control

Flow control is available for *DPDK physical ports*. For more information, refer to *Flow Control*.

Jumbo Frames

Added in version 2.6.0.

By default, DPDK ports are configured with standard Ethernet MTU (1500B). To enable Jumbo Frames support for a DPDK port, change the Interface's `mtu_request` attribute to a sufficiently large value. For example, to add a *DPDK physical port* with an MTU of 9000, run:

```
$ ovs-vsctl add-port br0 dpdk-p0 -- set Interface dpdk-p0 type=dpdk \
     options:dpdk-devargs=0000:01:00.0 mtu_request=9000
```

Similarly, to change the MTU of an existing port to 6200, run:

```
$ ovs-vsctl set Interface dpdk-p0 mtu_request=6200
```

Some additional configuration is needed to take advantage of jumbo frames with *vHost User ports*:

- *Mergeable buffers* must be enabled for vHost User ports, as demonstrated in the QEMU command line snippet below:

```
-netdev type=vhost-user,id=mynet1,chardev=char0,vhostforce \
-device virtio-net-pci,mac=00:00:00:00:00:01,netdev=mynet1,mrg_rxbuf=on
```

- Where virtio devices are bound to the Linux kernel driver in a guest environment (i.e. interfaces are not bound to an in-guest DPDK driver), the MTU of those logical network interfaces must also be increased to a sufficiently large value. This avoids segmentation of Jumbo Frames received in the guest. Note that 'MTU' refers to the length of the IP packet only, and not that of the entire frame.

  To calculate the exact MTU of a standard IPv4 frame, subtract the L2 header and CRC lengths (i.e. 18B) from the max supported frame size. So, to set the MTU for a 9018B Jumbo Frame:

```
$ ip link set eth1 mtu 9000
```

When Jumbo Frames are enabled, the size of a DPDK port's mbuf segments are increased, such that a full Jumbo Frame of a specific size may be accommodated within a single mbuf segment.

Jumbo frame support has been validated against 9728B frames, which is the largest frame size supported by Fortville NIC using the DPDK i40e driver, but larger frames and other DPDK NIC drivers may be supported. These cases are common for use cases involving East-West traffic only.

DPDK Device Memory Models

DPDK device memory can be allocated in one of two ways in OVS DPDK, **shared memory** or **per port memory**. The specifics of both are detailed below.

Shared Memory

By default OVS DPDK uses a shared memory model. This means that multiple ports can share the same mempool. For example when a port is added it will have a given MTU and socket ID associated with it. If a mempool has been created previously for an existing port that has the same MTU and socket ID, that mempool is used for both ports. If there is no existing mempool supporting these parameters then a new mempool is created.

Per Port Memory

In the per port memory model, mempools are created per device and are not shared. The benefit of this is a more transparent memory model where mempools will not be exhausted by other DPDK devices. However this comes at a potential increase in cost for memory dimensioning for a given deployment. Users should be aware of the memory requirements for their deployment before using this model and allocate the required hugepage memory.

Per port mempool support may be enabled via a global config value, `per-port-memory`. Setting this to true enables the per port memory model for all DPDK devices in OVS:

```
$ ovs-vsctl set Open_vSwitch . other_config:per-port-memory=true
```

Important

This value should be set before setting dpdk-init=true. If set after dpdk-init=true then the daemon must be restarted to use per-port-memory.

Calculating Memory Requirements

The amount of memory required for a given mempool can be calculated by the **number mbufs in the mempool * mbuf size**.

Users should be aware of the following:

- The **number of mbufs** per mempool will differ between memory models.
- The **size of each mbuf** will be affected by the requested **MTU** size.

Important

An mbuf size in bytes is always larger than the requested MTU size due to alignment and rounding needed in OVS DPDK.

Below are a number of examples of memory requirement calculations for both shared and per port memory models.

Shared Memory Calculations

In the shared memory model the number of mbufs requested is directly affected by the requested MTU size as described in the table below.

MTU Size	Num MBUFS
1500 or greater	262144
Less than 1500	16384

Important

If a deployment does not have enough memory to provide 262144 mbufs then the requested amount is halved up until 16384.

Example 1

```
MTU = 1500 Bytes
Number of mbufs = 262144
Mbuf size = 3008 Bytes
Memory required = 262144 * 3008 = 788 MB
```

Example 2

```
MTU = 1800 Bytes
Number of mbufs = 262144
```

(continues on next page)

(continued from previous page)

```
Mbuf size = 3008 Bytes
Memory required = 262144 * 3008 = 788 MB
```

> **Note**
>
> Assuming the same socket is in use for example 1 and 2 the same mempool would be shared.

Example 3

```
MTU = 6000 Bytes
Number of mbufs = 262144
Mbuf size = 7104 Bytes
Memory required = 262144 * 7104 = 1862 MB
```

Example 4

```
MTU = 9000 Bytes
Number of mbufs = 262144
Mbuf size = 10176 Bytes
Memory required = 262144 * 10176 = 2667 MB
```

Per Port Memory Calculations

The number of mbufs requested in the per port model is more complicated and accounts for multiple dynamic factors in the datapath and device configuration.

A rough estimation of the number of mbufs required for a port is:

```
packets required to fill the device rxqs +
packets that could be stuck on other ports txqs +
packets on the pmd threads +
additional corner case memory.
```

The algorithm in OVS used to calculate this is as follows:

```
requested number of rxqs * requested rxq size +
requested number of txqs * requested txq size +
min(RTE_MAX_LCORE, requested number of rxqs) * netdev_max_burst +
MIN_NB_MBUF.
```

where:

- **requested number of rxqs**: Number of requested receive queues for a device.
- **requested rxq size**: The number of descriptors requested for a rx queue.
- **requested number of txqs**: Number of requested transmit queues for a device. Calculated as the number of PMDs configured +1.
- **requested txq size**: the number of descriptors requested for a tx queue.
- **min(RTE_MAX_LCORE, requested number of rxqs)**: Compare the maximum number of lcores supported by DPDK to the number of requested receive queues for the device and use the variable of lesser value.

- **NETDEV_MAX_BURST**: Maximum number of packets in a burst, defined as 32.
- **MIN_NB_MBUF**: Additional memory for corner case, defined as 16384.

For all examples below assume the following values:

- requested_rxq_size = 2048
- requested_txq_size = 2048
- RTE_MAX_LCORE = 128
- netdev_max_burst = 32
- MIN_NB_MBUF = 16384

Example 1: (1 rxq, 1 PMD, 1500 MTU)

```
MTU = 1500
Number of mbufs = (1 * 2048) + (2 * 2048) + (1 * 32) + (16384) = 22560
Mbuf size = 3008 Bytes
Memory required = 22560 * 3008 = 67 MB
```

Example 2: (1 rxq, 2 PMD, 6000 MTU)

```
MTU = 6000
Number of mbufs = (1 * 2048) + (3 * 2048) + (1 * 32) + (16384) = 24608
Mbuf size = 7104 Bytes
Memory required = 24608 * 7104 = 175 MB
```

Example 3: (2 rxq, 2 PMD, 9000 MTU)

```
MTU = 9000
Number of mbufs = (2 * 2048) + (3 * 2048) + (1 * 32) + (16384) = 26656
Mbuf size = 10176 Bytes
Memory required = 26656 * 10176 = 271 MB
```

Shared Mempool Configuration

In order to increase sharing of mempools, a user can configure the MTUs which mempools are based on by using `shared-mempool-config`.

An MTU configured by the user is adjusted to an mbuf size used for mempool creation and stored. If a port is subsequently added that has an MTU which can be accommodated by this mbuf size, it will be used for mempool creation/reuse.

This can increase sharing by consolidating mempools for ports with different MTUs which would otherwise use separate mempools. It can also help to remove the need for mempools being created after a port is added but before it's MTU is changed to a different value.

For example, on a 2 NUMA system:

```
$ ovs-vsctl ovs-vsctl --no-wait set Open_vSwitch . \
other_config:shared-mempool-config=9000,1500:1,6000:1
```

In this case, OVS stores the mbuf sizes based on the following MTUs.

- NUMA 0: 9000
- NUMA 1: 1500, 6000, 9000

Ports added will use mempools with the mbuf sizes based on the above MTUs where possible. If there is more than one suitable, the one closest to the MTU will be selected.

Port added on NUMA 0:

- MTU 1500, use mempool based on 9000 MTU
- MTU 6000, use mempool based on 9000 MTU
- MTU 9000, use mempool based on 9000 MTU
- MTU 9300, use mempool based on 9300 MTU (existing behaviour)

Port added on NUMA 1:

- MTU 1500, use mempool based on 1500 MTU
- MTU 6000, use mempool based on 6000 MTU
- MTU 9000, use mempool based on 9000 MTU
- MTU 9300, use mempool based on 9300 MTU (existing behaviour)

4.1.13 Flow Hardware offload with Linux TC flower

This document describes how to offload flows with TC flower.

Flow Hardware Offload

The flow hardware offload is disabled by default and can be enabled by:

```
$ ovs-vsctl set Open_vSwitch . other_config:hw-offload=true
```

TC flower has one additional configuration option caled `tc-policy`. For more details see `man ovs-vswitchd.conf.db`.

TC Meter Offload

Offloading meters to TC does not require any additional configuration and is enabled automatically when possible. Offloading with meters does require the tc-police action to be available in the Linux kernel. For more details on the tc-police action, see `man tc-police`.

Configuration

There is no parameter change in ovs-ofctl command, to configure a meter and use it for a flow in the offload way. Usually the commands are like:

```
$ ovs-ofctl -O OpenFlow13 add-meter br0 "meter=1 pktps bands=type=drop rate=1"
$ ovs-ofctl -O OpenFlow13 add-flow br0 "priority=10,in_port=ovs-p0,udp actions=meter:1,
↪normal"
```

For more details, see `man ovs-ofctl`.

> **Note**
>
> Each meter is mapped to one TC police action. To avoid conflicts, the police action indexes 0x10000000-0x1fffffff are reserved for this mapping. You can check the police actions using the command `tc action ls action police` on Linux systems.

Known TC flow offload limitations

General

These sections describe limitations to the general TC flow offload implementation.

Flow bytes count

Flows that are offloaded with TC do not include the L2 bytes in the packet byte count. Take the datapath flow dump below as an example. The first one is from the none-offloaded case the second one is from a TC offloaded flow:

```
in_port(2),eth(macs),eth_type(0x0800),ipv4(proto=17,frag=no), packets:10, bytes:470,␣
↪used:0.001s, actions:outputmeter(0),3

in_port(2),eth(macs),eth_type(0x0800),ipv4(proto=17,frag=no), packets:10, bytes:330,␣
↪used:0.001s, actions:outputmeter(0),3
```

As you can see above the none-offload case reports 140 bytes more, which is 14 bytes per packet. This represents the L2 header, in this case, 2 * *Ethernet address* + *Ethertype*.

TC Meter Offload

These sections describe limitations related to the TC meter offload implementation.

Missing byte count drop statistics

The kernel's TC infrastructure is only counting the number of dropped packet, not their byte size. This results in the meter statistics always showing 0 for byte_count. Here is an example:

```
$ ovs-ofctl -O OpenFlow13 meter-stats br0
OFPST_METER reply (OF1.3) (xid=0x2):
meter:1 flow_count:1 packet_in_count:11 byte_in_count:377 duration:3.199s bands:
0: packet_count:9 byte_count:0
```

First flow packet not processed by meter

Packets that are received by ovs-vswitchd through an upcall before the actual meter flow is installed, are not passing TC police action and therefore are not considered for policing.

Conntrack Application Layer Gateways (ALG)

TC does not support conntrack helpers, i.e., ALGs. TC will not offload flows if the ALG keyword is present within the ct() action. However, this will not allow ALGs to work within the datapath, as the return traffic without the ALG keyword might run through a TC rule, which internally will not call the conntrack helper required.

So if ALG support is required, tc offload must be disabled.

CHAPTER

FIVE

DEEP DIVE

How Open vSwitch is implemented and, where necessary, why it was implemented that way.

5.1 OVS

5.1.1 Design Decisions In Open vSwitch

This document describes design decisions that went into implementing Open vSwitch. While we believe these to be reasonable decisions, it is impossible to predict how Open vSwitch will be used in all environments. Understanding assumptions made by Open vSwitch is critical to a successful deployment. The end of this document contains contact information that can be used to let us know how we can make Open vSwitch more generally useful.

Asynchronous Messages

Over time, Open vSwitch has added many knobs that control whether a given controller receives OpenFlow asynchronous messages. This section describes how all of these features interact.

First, a service controller never receives any asynchronous messages unless it changes its miss_send_len from the service controller default of zero in one of the following ways:

- Sending an `OFPT_SET_CONFIG` message with nonzero `miss_send_len`.
- Sending any `NXT_SET_ASYNC_CONFIG` message: as a side effect, this message changes the `miss_send_len` to `OFP_DEFAULT_MISS_SEND_LEN` (128) for service controllers.

Second, `OFPT_FLOW_REMOVED` and `NXT_FLOW_REMOVED` messages are generated only if the flow that was removed had the `OFPFF_SEND_FLOW_REM` flag set.

Third, `OFPT_PACKET_IN` and `NXT_PACKET_IN` messages are sent only to OpenFlow controller connections that have the correct connection ID (see `struct nx_controller_id` and `struct nx_action_controller`):

- For packet-in messages generated by a `NXAST_CONTROLLER` action, the controller ID specified in the action.
- For other packet-in messages, controller ID zero. (This is the default ID when an OpenFlow controller does not configure one.)

Finally, Open vSwitch consults a per-connection table indexed by the message type, reason code, and current role. The following table shows how this table is initialized by default when an OpenFlow connection is made. An entry labeled `yes` means that the message is sent, an entry labeled `---` means that the message is suppressed.

Table 1: OFPT_PACKET_IN / NXT_PACKET_IN

message and reason code	other	secondary
OFPR_NO_MATCH	yes	—
OFPR_ACTION	yes	—
OFPR_INVALID_TTL	—	—
OFPR_ACTION_SET (OF1.4+)	yes	—
OFPR_GROUP (OF1.4+)	yes	—
OFPR_PACKET_OUT (OF1.4+)	yes	—

Table 2: OFPT_FLOW_REMOVED / NXT_FLOW_REMOVED

message and reason code	other	secondary
OFPRR_IDLE_TIMEOUT	yes	—
OFPRR_HARD_TIMEOUT	yes	—
OFPRR_DELETE	yes	—
OFPRR_GROUP_DELETE (OF1.3+)	yes	—
OFPRR_METER_DELETE (OF1.4+)	yes	—
OFPRR_EVICTION (OF1.4+)	yes	—

Table 3: OFPT_PORT_STATUS

message and reason code	other	secondary
OFPPR_ADD	yes	—
OFPPR_DELETE	yes	—
OFPPR_MODIFY	yes	—

Table 4: OFPT_ROLE_REQUEST / OFPT_ROLE_REPLY (OF1.4+)

message and reason code	other	secondary
OFPCRR_PROMOTE_REQUEST	—	—
OFPCRR_CONFIG	—	—
OFPCRR_EXPERIMENTER	—	—

Table 5: OFPT_TABLE_STATUS (OF1.4+)

message and reason code	other	secondary
OFPTR_VACANCY_DOWN	—	—
OFPTR_VACANCY_UP	—	—

Table 6: OFPT_REQUESTFORWARD (OF1.4+)

message and reason code	other	secondary
OFPRFR_GROUP_MOD	—	—
OFPRFR_METER_MOD	—	—

The NXT_SET_ASYNC_CONFIG message directly sets all of the values in this table for the current connection. The OFPC_INVALID_TTL_TO_CONTROLLER bit in the OFPT_SET_CONFIG message controls the setting for

OFPR_INVALID_TTL for the "primary" role.

OFPAT_ENQUEUE

The OpenFlow 1.0 specification requires the output port of the OFPAT_ENQUEUE action to "refer to a valid physical port (i.e. < OFPP_MAX) or OFPP_IN_PORT". Although OFPP_LOCAL is not less than OFPP_MAX, it is an 'internal' port which can have QoS applied to it in Linux. Since we allow the OFPAT_ENQUEUE to apply to 'internal' ports whose port numbers are less than OFPP_MAX, we interpret OFPP_LOCAL as a physical port and support OFPAT_ENQUEUE on it as well.

OFPT_FLOW_MOD

The OpenFlow specification for the behavior of OFPT_FLOW_MOD is confusing. The following tables summarize the Open vSwitch implementation of its behavior in the following categories:

"match on priority"
: Whether the flow_mod acts only on flows whose priority matches that included in the flow_mod message.

"match on out_port"
: Whether the flow_mod acts only on flows that output to the out_port included in the flow_mod message (if out_port is not OFPP_NONE). OpenFlow 1.1 and later have a similar feature (not listed separately here) for out_group.

"match on flow_cookie":
: Whether the flow_mod acts only on flows whose flow_cookie matches an optional controller-specified value and mask.

"updates flow_cookie":
: Whether the flow_mod changes the flow_cookie of the flow or flows that it matches to the flow_cookie included in the flow_mod message.

"updates OFPFF_ flags":
: Whether the flow_mod changes the OFPFF_SEND_FLOW_REM flag of the flow or flows that it matches to the setting included in the flags of the flow_mod message.

"honors OFPFF_CHECK_OVERLAP":
: Whether the OFPFF_CHECK_OVERLAP flag in the flow_mod is significant.

"updates idle_timeout" and "updates hard_timeout":
: Whether the idle_timeout and hard_timeout in the flow_mod, respectively, have an effect on the flow or flows matched by the flow_mod.

"updates idle timer":
: Whether the flow_mod resets the per-flow timer that measures how long a flow has been idle.

"updates hard timer":
: Whether the flow_mod resets the per-flow timer that measures how long it has been since a flow was modified.

"zeros counters":
: Whether the flow_mod resets per-flow packet and byte counters to zero.

"may add a new flow":
: Whether the flow_mod may add a new flow to the flow table. (Obviously this is always true for "add" commands but in some OpenFlow versions "modify" and "modify-strict" can also add new flows.)

"sends flow_removed message":
: Whether the flow_mod generates a flow_removed message for the flow or flows that it affects.

An entry labeled yes means that the flow mod type does have the indicated behavior, --- means that it does not, an empty cell means that the property is not applicable, and other values are explained below the table.

OpenFlow 1.0

RULE	ADD	MODIFY	STRICT	DELETE	STRICT
match on `priority`	yes	—	yes	—	yes
match on `out_port`	—	—	—	yes	yes
match on `flow_cookie`	—	—	—	—	—
match on `table_id`	—	—	—	—	—
controller chooses `table_id`	—	—	—		
updates `flow_cookie`	yes	yes	yes		
updates `OFPFF_SEND_FLOW_REM`	yes	•	•		
honors `OFPFF_CHECK_OVERLAP`	yes	•	•		
updates `idle_timeout`	yes	•	•		
updates `hard_timeout`	yes	•	•		
resets idle timer	yes	•	•		
resets hard timer	yes	yes	yes		
zeros counters	yes	•	•		
may add a new flow	yes	yes	yes		
sends `flow_removed` message	—	—	—	%	%

where:

+

"modify" and "modify-strict" only take these actions when they create a new flow, not when they update an existing flow.

%

"delete" and "delete_strict" generates a flow_removed message if the deleted flow or flows have the `OFPFF_SEND_FLOW_REM` flag set. (Each controller can separately control whether it wants to receive the generated messages.)

OpenFlow 1.1

OpenFlow 1.1 makes these changes:

- The controller now must specify the `table_id` of the flow match searched and into which a flow may be inserted. Behavior for a `table_id` of 255 is undefined.
- A `flow_mod`, except an "add", can now match on the `flow_cookie`.
- When a `flow_mod` matches on the `flow_cookie`, "modify" and "modify-strict" never insert a new flow.

RULE	ADD	MODIFY	STRICT	DELETE	STRICT
match on `priority`	yes	—	yes	—	yes
match on `out_port`	—	—	—	yes	yes
match on `flow_cookie`	—	yes	yes	yes	yes
match on `table_id`	yes	yes	yes	yes	yes
controller chooses `table_id`	yes	yes	yes		
updates `flow_cookie`	yes	—	—		
updates `OFPFF_SEND_FLOW_REM`	yes	•	•		
honors `OFPFF_CHECK_OVERLAP`	yes	•	•		
updates `idle_timeout`	yes	•	•		
updates `hard_timeout`	yes	•	•		
resets idle timer	yes	•	•		
resets hard timer	yes	yes	yes		
zeros counters	yes	•	•		
may add a new flow	yes	#	#		
sends `flow_removed` message	—	—	—	%	%

where:

+
: “modify” and “modify-strict” only take these actions when they create a new flow, not when they update an existing flow.

%
: “delete” and “delete_strict” generates a flow_removed message if the deleted flow or flows have the `OFPFF_SEND_FLOW_REM` flag set. (Each controller can separately control whether it wants to receive the generated messages.)

#
: “modify” and “modify-strict” only add a new flow if the flow_mod does not match on any bits of the flow cookie

OpenFlow 1.2

OpenFlow 1.2 makes these changes:

- Only “add” commands ever add flows, “modify” and “modify-strict” never do.
- A new flag `OFPFF_RESET_COUNTS` now controls whether “modify” and “modify-strict” reset counters, whereas previously they never reset counters (except when they inserted a new flow).

RULE	ADD	MODIFY	STRICT	DELETE	STRICT
match on `priority`	yes	—	yes	—	yes
match on `out_port`	—	—	—	yes	yes
match on `flow_cookie`	—	yes	yes	yes	yes
match on `table_id`	yes	yes	yes	yes	yes
controller chooses `table_id`	yes	yes	yes		
updates `flow_cookie`	yes	—	—		
updates `OFPFF_SEND_FLOW_REM`	yes	—	—		
honors `OFPFF_CHECK_OVERLAP`	yes	—	—		
updates `idle_timeout`	yes	—	—		
updates `hard_timeout`	yes	—	—		
resets idle timer	yes	—	—		
resets hard timer	yes	yes	yes		
zeros counters	yes	&	&		
may add a new flow	yes	—	—		
sends `flow_removed` message	—	—	—	%	%

%

"delete" and "delete_strict" generates a flow_removed message if the deleted flow or flows have the `OFPFF_SEND_FLOW_REM` flag set. (Each controller can separately control whether it wants to receive the generated messages.)

&

"modify" and "modify-strict" reset counters if the `OFPFF_RESET_COUNTS` flag is specified.

OpenFlow 1.3

OpenFlow 1.3 makes these changes:

- Behavior for a table_id of 255 is now defined, for "delete" and "delete-strict" commands, as meaning to delete from all tables. A table_id of 255 is now explicitly invalid for other commands.
- New flags `OFPFF_NO_PKT_COUNTS` and `OFPFF_NO_BYT_COUNTS` for "add" operations.

The table for 1.3 is the same as the one shown above for 1.2.

OpenFlow 1.4

OpenFlow 1.4 makes these changes:

- Adds the "importance" field to `flow_mods`, but it does not explicitly specify which kinds of `flow_mods` set the importance. For consistency, Open vSwitch uses the same rule for importance as for `idle_timeout` and `hard_timeout`, that is, only an "ADD" flow_mod sets the importance. (This issue has been filed with the ONF as EXT-496.)
- Eviction Mechanism to automatically delete entries of lower importance to make space for newer entries.

OpenFlow 1.4 Bundles

Open vSwitch makes all flow table modifications atomically, i.e., any datapath packet only sees flow table configurations either before or after any change made by any `flow_mod`. For example, if a controller removes all flows with a single OpenFlow `flow_mod`, no packet sees an intermediate version of the OpenFlow pipeline where only some of the flows have been deleted.

It should be noted that Open vSwitch caches datapath flows, and that the cached flows are *NOT* flushed immediately when a flow table changes. Instead, the datapath flows are revalidated against the new flow table as soon as possible,

and usually within one second of the modification. This design amortizes the cost of datapath cache flushing across multiple flow table changes, and has a significant performance effect during simultaneous heavy flow table churn and high traffic load. This means that different cached datapath flows may have been computed based on a different flow table configurations, but each of the datapath flows is guaranteed to have been computed over a coherent view of the flow tables, as described above.

With OpenFlow 1.4 bundles this atomicity can be extended across an arbitrary set of `flow_mod`. Bundles are supported for `flow_mod` and port_mod messages only. For `flow_mod`, both `atomic` and `ordered` bundle flags are trivially supported, as all bundled messages are executed in the order they were added and all flow table modifications are now atomic to the datapath. Port mods may not appear in atomic bundles, as port status modifications are not atomic.

To support bundles, ovs-ofctl has a `--bundle` option that makes the flow mod commands (`add-flow`, `add-flows`, `mod-flows`, `del-flows`, and `replace-flows`) use an OpenFlow 1.4 bundle to operate the modifications as a single atomic transaction. If any of the flow mods in a transaction fail, none of them are executed. All flow mods in a bundle appear to datapath lookups simultaneously.

Furthermore, ovs-ofctl `add-flow` and `add-flows` commands now accept arbitrary flow mods as an input by allowing the flow specification to start with an explicit `add`, `modify`, `modify_strict`, `delete`, or `delete_strict` keyword. A missing keyword is treated as `add`, so this is fully backwards compatible. With the new `--bundle` option all the flow mods are executed as a single atomic transaction using an OpenFlow 1.4 bundle. Without the `--bundle` option the flow mods are executed in order up to the first failing `flow_mod`, and in case of an error the earlier successful `flow_mod` calls are not rolled back.

OFPT_PACKET_IN

The OpenFlow 1.1 specification for `OFPT_PACKET_IN` is confusing. The definition in OF1.1 `openflow.h` is[*]:

```
/* Packet received on port (datapath -> controller). */
struct ofp_packet_in {
    struct ofp_header header;
    uint32_t buffer_id;     /* ID assigned by datapath. */
    uint32_t in_port;       /* Port on which frame was received. */
    uint32_t in_phy_port;   /* Physical Port on which frame was received. */
    uint16_t total_len;     /* Full length of frame. */
    uint8_t reason;         /* Reason packet is being sent (one of OFPR_*) */
    uint8_t table_id;       /* ID of the table that was looked up */
    uint8_t data[0];        /* Ethernet frame, halfway through 32-bit word,
                               so the IP header is 32-bit aligned.  The
                               amount of data is inferred from the length
                               field in the header.  Because of padding,
                               offsetof(struct ofp_packet_in, data) ==
                               sizeof(struct ofp_packet_in) - 2. */
};
OFP_ASSERT(sizeof(struct ofp_packet_in) == 24);
```

The confusing part is the comment on the `data[]` member. This comment is a leftover from OF1.0 `openflow.h`, in which the comment was correct: `sizeof(struct ofp_packet_in)` is 20 in OF1.0 and `ffsetof(struct ofp_packet_in, data)` is 18. When OF1.1 was written, the structure members were changed but the comment was carelessly not updated, and the comment became wrong: `sizeof(struct ofp_packet_in)` and offsetof(struct ofp_packet_in, data) are both 24 in OF1.1.

That leaves the question of how to implement `ofp_packet_in` in OF1.1. The OpenFlow reference implementation for OF1.1 does not include any padding, that is, the first byte of the encapsulated frame immediately follows the `table_id` member without a gap. Open vSwitch therefore implements it the same way for compatibility.

For an earlier discussion, please see the thread archived at: https://www.mail-archive.com/openflow-discuss@lists.stanford.edu/msg00663.html

[*] The quoted definition is directly from OF1.1. Definitions used inside OVS omit the 8-byte `ofp_header` members, so the sizes in this discussion are 8 bytes larger than those declared in OVS header files.

VLAN Matching

The 802.1Q VLAN header causes more trouble than any other 4 bytes in networking. More specifically, three versions of OpenFlow and Open vSwitch have among them four different ways to match the contents and presence of the VLAN header. The following table describes how each version works.

Match	NXM	OF1.0	OF1.1	OF1.2
[1]	0000/0000	????/1,??/?	????/1,??/?	0000/0000,--
[2]	0000/ffff	ffff/0,??/?	ffff/0,??/?	0000/ffff,--
[3]	1xxx/1fff	0xxx/0,??/1	0xxx/0,??/1	1xxx/ffff,--
[4]	z000/f000	????/1,0y/0	fffe/0,0y/0	1000/1000,0y
[5]	zxxx/ffff	0xxx/0,0y/0	0xxx/0,0y/0	1xxx/ffff,0y
[6]	0000/0fff	<none>	<none>	<none>
[7]	0000/f000	<none>	<none>	<none>
[8]	0000/efff	<none>	<none>	<none>
[9]	1001/1001	<none>	<none>	1001/1001,--
[10]	3000/3000	<none>	<none>	<none>
[11]	1000/1000	<none>	fffe/0,??/1	1000/1000,--

where:

Match:
See the list below.

NXM:
`xxxx/yyyy` means `NXM_OF_VLAN_TCI_W` with value `xxxx` and mask `yyyy`. A mask of `0000` is equivalent to omitting `NXM_OF_VLAN_TCI(_W)`, a mask of `ffff` is equivalent to `NXM_OF_VLAN_TCI`.

OF1.0, OF1.1:
`wwww/x,yy/z` means `dl_vlan` `wwww`, `OFPFW_DL_VLAN` `x`, `dl_vlan_pcp` `yy`, and `OFPFW_DL_VLAN_PCP` `z`. If `OFPFW_DL_VLAN` or `OFPFW_DL_VLAN_PCP` is 1, the corresponding field value is wildcarded, otherwise it is matched. `?` means that the given bits are ignored (their conventional values are `0000/x,00/0` in OF1.0, `0000/x,00/1` in OF1.1; `x` is never ignored). <none> means that the given match is not supported.

OF1.2:
`xxxx/yyyy,zz` means `OXM_OF_VLAN_VID_W` with value `xxxx` and mask `yyyy`, and `OXM_OF_VLAN_PCP` (which is not maskable) with value `zz`. A mask of `0000` is equivalent to omitting `OXM_OF_VLAN_VID(_W)`, a mask of `ffff` is equivalent to `OXM_OF_VLAN_VID`. `--` means that `OXM_OF_VLAN_PCP` is omitted. <none> means that the given match is not supported.

The matches are:

[1]:
Matches any packet, that is, one without an 802.1Q header or with an 802.1Q header with any TCI value.

[2]
Matches only packets without an 802.1Q header.

NXM:
Any match with `vlan_tci == 0` and `(vlan_tci_mask & 0x1000) != 0` is equivalent to the one listed in the table.

OF1.0:
The spec doesn't define behavior if `dl_vlan` is set to `0xffff` and `OFPFW_DL_VLAN_PCP` is not set.

OF1.1:
The spec says explicitly to ignore `dl_vlan_pcp` when `dl_vlan` is set to `0xffff`.

OF1.2:
The spec doesn't say what should happen if `vlan_vid == 0` and `(vlan_vid_mask & 0x1000) != 0` but `vlan_vid_mask != 0x1000`, but it would be straightforward to also interpret as `[2]`.

[3]
Matches only packets that have an 802.1Q header with VID `xxx` (and any PCP).

[4]
Matches only packets that have an 802.1Q header with PCP `y` (and any VID).

NXM:
`z` is `(y << 1) | 1`.

OF1.0:
The spec isn't very clear, but OVS implements it this way.

OF1.2:
Presumably other masks such that `(vlan_vid_mask & 0x1fff) == 0x1000` would also work, but the spec doesn't define their behavior.

[5]
Matches only packets that have an 802.1Q header with VID `xxx` and PCP `y`.

NXM:
`z` is `((y << 1) | 1)`.

OF1.2:
Presumably other masks such that `(vlan_vid_mask & 0x1fff) == 0x1fff` would also work.

[6]
Matches packets with no 802.1Q header or with an 802.1Q header with a VID of 0. Only possible with NXM.

[7]
Matches packets with no 802.1Q header or with an 802.1Q header with a PCP of 0. Only possible with NXM.

[8]
Matches packets with no 802.1Q header or with an 802.1Q header with both VID and PCP of 0. Only possible with NXM.

[9]
Matches only packets that have an 802.1Q header with an odd-numbered VID (and any PCP). Only possible with NXM and OF1.2. (This is just an example; one can match on any desired VID bit pattern.)

[10]
Matches only packets that have an 802.1Q header with an odd-numbered PCP (and any VID). Only possible with NXM. (This is just an example; one can match on any desired VID bit pattern.)

[11]
Matches any packet with an 802.1Q header, regardless of VID or PCP.

Additional notes:

OF1.2:
The top three bits of `OXM_OF_VLAN_VID` are fixed to zero, so bits 13, 14, and 15 in the masks listed in the table may be set to arbitrary values, as long as the corresponding value bits are also zero. The suggested `ffff` mask for [2], [3], and [5] allows a shorter OXM representation (the mask is omitted) than the minimal `1fff` mask.

Flow Cookies

OpenFlow 1.0 and later versions have the concept of a "flow cookie", which is a 64-bit integer value attached to each flow. The treatment of the flow cookie has varied greatly across OpenFlow versions, however.

In OpenFlow 1.0:

- `OFPFC_ADD` set the cookie in the flow that it added.
- `OFPFC_MODIFY` and `OFPFC_MODIFY_STRICT` updated the cookie for the flow or flows that it modified.
- `OFPST_FLOW` messages included the flow cookie.
- `OFPT_FLOW_REMOVED` messages reported the cookie of the flow that was removed.

OpenFlow 1.1 made the following changes:

- Flow mod operations `OFPFC_MODIFY`, `OFPFC_MODIFY_STRICT`, `OFPFC_DELETE`, and `OFPFC_DELETE_STRICT`, plus flow stats requests and aggregate stats requests, gained the ability to match on flow cookies with an arbitrary mask.
- `OFPFC_MODIFY` and `OFPFC_MODIFY_STRICT` were changed to add a new flow, in the case of no match, only if the flow table modification operation did not match on the cookie field. (In OpenFlow 1.0, modify operations always added a new flow when there was no match.)
- `OFPFC_MODIFY` and `OFPFC_MODIFY_STRICT` no longer updated flow cookies.

OpenFlow 1.2 made the following changes:

- `OFPC_MODIFY` and `OFPFC_MODIFY_STRICT` were changed to never add a new flow, regardless of whether the flow cookie was used for matching.

Open vSwitch support for OpenFlow 1.0 implements the OpenFlow 1.0 behavior with the following extensions:

- An NXM extension field `NXM_NX_COOKIE(_W)` allows the NXM versions of `OFPFC_MODIFY`, `OFPFC_MODIFY_STRICT`, `OFPFC_DELETE`, and `OFPFC_DELETE_STRICT` `flow_mod` calls, plus flow stats requests and aggregate stats requests, to match on flow cookies with arbitrary masks. This is much like the equivalent OpenFlow 1.1 feature.
- Like OpenFlow 1.1, `OFPC_MODIFY` and `OFPFC_MODIFY_STRICT` add a new flow if there is no match and the mask is zero (or not given).
- The `cookie` field in `OFPT_FLOW_MOD` and `NXT_FLOW_MOD` messages is used as the cookie value for `OFPFC_ADD` commands, as described in OpenFlow 1.0. For `OFPFC_MODIFY` and `OFPFC_MODIFY_STRICT` commands, the `cookie` field is used as a new cookie for flows that match unless it is `UINT64_MAX`, in which case the flow's cookie is not updated.
- `NXT_PACKET_IN` (the Nicira extended version of `OFPT_PACKET_IN`) reports the cookie of the rule that generated the packet, or all-1-bits if no rule generated the packet. (Older versions of OVS used all-0-bits instead of all-1-bits.)

The following table shows the handling of different protocols when receiving `OFPFC_MODIFY` and `OFPFC_MODIFY_STRICT` messages. A mask of 0 indicates either an explicit mask of zero or an implicit one by not specifying the `NXM_NX_COOKIE(_W)` field.

OpenFlow 1.0	no	yes	(add on miss)	(add on miss)
OpenFlow 1.1	yes	no	no	yes
OpenFlow 1.2	yes	no	no	no
NXM	yes	yes*	no	yes

* Updates the flow's cookie unless the `cookie` field is `UINT64_MAX`.

Multiple Table Support

OpenFlow 1.0 has only rudimentary support for multiple flow tables. Notably, OpenFlow 1.0 does not allow the controller to specify the flow table to which a flow is to be added. Open vSwitch adds an extension for this purpose, which is enabled on a per-OpenFlow connection basis using the NXT_FLOW_MOD_TABLE_ID message. When the extension is enabled, the upper 8 bits of the command member in an OFPT_FLOW_MOD or NXT_FLOW_MOD message designates the table to which a flow is to be added.

The Open vSwitch software switch implementation offers 255 flow tables. On packet ingress, only the first flow table (table 0) is searched, and the contents of the remaining tables are not considered in any way. Tables other than table 0 only come into play when an NXAST_RESUBMIT_TABLE action specifies another table to search.

Tables 128 and above are reserved for use by the switch itself. Controllers should use only tables 0 through 127.

OFPTC_* Table Configuration

This section covers the history of the OFPTC_* table configuration bits across OpenFlow versions.

OpenFlow 1.0 flow tables had fixed configurations.

OpenFlow 1.1 enabled controllers to configure behavior upon flow table miss and added the OFPTC_MISS_* constants for that purpose. OFPTC_* did not control anything else but it was nevertheless conceptualized as a set of bit-fields instead of an enum. OF1.1 added the OFPT_TABLE_MOD message to set OFPTC_MISS_* for a flow table and added the config field to the OFPST_TABLE reply to report the current setting.

OpenFlow 1.2 did not change anything in this regard.

OpenFlow 1.3 switched to another means to changing flow table miss behavior and deprecated OFPTC_MISS_* without adding any more OFPTC_* constants. This meant that OFPT_TABLE_MOD now had no purpose at all, but OF1.3 kept it around "for backward compatibility with older and newer versions of the specification." At the same time, OF1.3 introduced a new message OFPMP_TABLE_FEATURES that included a field config documented as reporting the OFPTC_* values set with OFPT_TABLE_MOD; of course this served no real purpose because no OFPTC_* values are defined. OF1.3 did remove the OFPTC_* field from OFPMP_TABLE (previously named OFPST_TABLE).

OpenFlow 1.4 defined two new OFPTC_* constants, OFPTC_EVICTION and OFPTC_VACANCY_EVENTS, using bits that did not overlap with OFPTC_MISS_* even though those bits had not been defined since OF1.2. OFPT_TABLE_MOD still controlled these settings. The field for OFPTC_* values in OFPMP_TABLE_FEATURES was renamed from config to capabilities and documented as reporting the flags that are supported in a OFPT_TABLE_MOD message. The OFPMP_TABLE_DESC message newly added in OF1.4 reported the OFPTC_* setting.

OpenFlow 1.5 did not change anything in this regard.

Table 7: Revisions

OpenFlow	OFPTC_* flags	TABLE_MOD	Statistics	TABLE_FEATURES	TABLE_DESC
OF1.0	none	no (*)(+)	no (*)	nothing (*)(+)	no (*)(+)
OF1.1/1.2	MISS_*	yes	yes	nothing (+)	no (+)
OF1.3	none	yes (*)	no (*)	config (*)	no (*)(+)
OF1.4/1.5	EVICTION/VACANCY_EVENTS	yes	no	capabilities	yes

where:

OpenFlow:
The OpenFlow version(s).

OFPTC_* flags:
The OFPTC_* flags defined in those versions.

TABLE_MOD:
Whether OFPT_TABLE_MOD can modify OFPTC_* flags.

Statistics:
Whether `OFPST_TABLE/OFPMP_TABLE` reports the `OFPTC_*` flags.

`TABLE_FEATURES`:
What `OFPMP_TABLE_FEATURES` reports (if it exists): either the current configuration or the switch's capabilities.

`TABLE_DESC`:
Whether `OFPMP_TABLE_DESC` reports the current configuration.

(*): Nothing to report/change anyway.

(+): No such message.

IPv6

Open vSwitch supports stateless handling of IPv6 packets. Flows can be written to support matching TCP, UDP, and ICMPv6 headers within an IPv6 packet. Deeper matching of some Neighbor Discovery messages is also supported.

IPv6 was not designed to interact well with middle-boxes. This, combined with Open vSwitch's stateless nature, have affected the processing of IPv6 traffic, which is detailed below.

Extension Headers

The base IPv6 header is incredibly simple with the intention of only containing information relevant for routing packets between two endpoints. IPv6 relies heavily on the use of extension headers to provide any other functionality. Unfortunately, the extension headers were designed in such a way that it is impossible to move to the next header (including the layer-4 payload) unless the current header is understood.

Open vSwitch will process the following extension headers and continue to the next header:

- Fragment (see the next section)
- AH (Authentication Header)
- Hop-by-Hop Options
- Routing
- Destination Options

When a header is encountered that is not in that list, it is considered "terminal". A terminal header's IPv6 protocol value is stored in `nw_proto` for matching purposes. If a terminal header is TCP, UDP, or ICMPv6, the packet will be further processed in an attempt to extract layer-4 information.

Fragments

IPv6 requires that every link in the internet have an MTU of 1280 octets or greater (RFC 2460). As such, a terminal header (as described above in "Extension Headers") in the first fragment should generally be reachable. In this case, the terminal header's IPv6 protocol type is stored in the `nw_proto` field for matching purposes. If a terminal header cannot be found in the first fragment (one with a fragment offset of zero), the `nw_proto` field is set to 0. Subsequent fragments (those with a non-zero fragment offset) have the `nw_proto` field set to the IPv6 protocol type for fragments (44).

Jumbograms

An IPv6 jumbogram (RFC 2675) is a packet containing a payload longer than 65,535 octets. A jumbogram is only relevant in subnets with a link MTU greater than 65,575 octets, and are not required to be supported on nodes that do not connect to link with such large MTUs. Currently, Open vSwitch doesn't process jumbograms.

In-Band Control

Motivation

An OpenFlow switch must establish and maintain a TCP network connection to its controller. There are two basic ways to categorize the network that this connection traverses: either it is completely separate from the one that the switch is otherwise controlling, or its path may overlap the network that the switch controls. We call the former case "out-of-band control", the latter case "in-band control".

Out-of-band control has the following benefits:

- Simplicity: Out-of-band control slightly simplifies the switch implementation.
- Reliability: Excessive switch traffic volume cannot interfere with control traffic.
- Integrity: Machines not on the control network cannot impersonate a switch or a controller.
- Confidentiality: Machines not on the control network cannot snoop on control traffic.

In-band control, on the other hand, has the following advantages:

- No dedicated port: There is no need to dedicate a physical switch port to control, which is important on switches that have few ports (e.g. wireless routers, low-end embedded platforms).
- No dedicated network: There is no need to build and maintain a separate control network. This is important in many environments because it reduces proliferation of switches and wiring.

Open vSwitch supports both out-of-band and in-band control. This section describes the principles behind in-band control. See the description of the Controller table in ovs-vswitchd.conf.db(5) to configure OVS for in-band control.

Principles

The fundamental principle of in-band control is that an OpenFlow switch must recognize and switch control traffic without involving the OpenFlow controller. All the details of implementing in-band control are special cases of this principle.

The rationale for this principle is simple. If the switch does not handle in-band control traffic itself, then it will be caught in a contradiction: it must contact the controller, but it cannot, because only the controller can set up the flows that are needed to contact the controller.

The following points describe important special cases of this principle.

- In-band control must be implemented regardless of whether the switch is connected.

  It is tempting to implement the in-band control rules only when the switch is not connected to the controller, using the reasoning that the controller should have complete control once it has established a connection with the switch.

  This does not work in practice. Consider the case where the switch is connected to the controller. Occasionally it can happen that the controller forgets or otherwise needs to obtain the MAC address of the switch. To do so, the controller sends a broadcast ARP request. A switch that implements the in-band control rules only when it is disconnected will then send an `OFPT_PACKET_IN` message up to the controller. The controller will be unable to respond, because it does not know the MAC address of the switch. This is a deadlock situation that can only be resolved by the switch noticing that its connection to the controller has hung and reconnecting.

- In-band control must override flows set up by the controller.

  It is reasonable to assume that flows set up by the OpenFlow controller should take precedence over in-band control, on the basis that the controller should be in charge of the switch.

  Again, this does not work in practice. Reasonable controller implementations may set up a "last resort" fallback rule that wildcards every field and, e.g., sends it up to the controller or discards it. If a controller does that, then it will isolate itself from the switch.

- The switch must recognize all control traffic.

  The fundamental principle of in-band control states, in part, that a switch must recognize control traffic without involving the OpenFlow controller. More specifically, the switch must recognize *all* control traffic. "False negatives", that is, packets that constitute control traffic but that the switch does not recognize as control traffic, lead to control traffic storms.

  Consider an OpenFlow switch that only recognizes control packets sent to or from that switch. Now suppose that two switches of this type, named A and B, are connected to ports on an Ethernet hub (not a switch) and that an OpenFlow controller is connected to a third hub port. In this setup, control traffic sent by switch A will be seen by switch B, which will send it to the controller as part of an OFPT_PACKET_IN message. Switch A will then see the OFPT_PACKET_IN message's packet, re-encapsulate it in another OFPT_PACKET_IN, and send it to the controller. Switch B will then see that OFPT_PACKET_IN, and so on in an infinite loop.

  Incidentally, the consequences of "false positives", where packets that are not control traffic are nevertheless recognized as control traffic, are much less severe. The controller will not be able to control their behavior, but the network will remain in working order. False positives do constitute a security problem.

- The switch should use echo-requests to detect disconnection.

  TCP will notice that a connection has hung, but this can take a considerable amount of time. For example, with default settings the Linux kernel TCP implementation will retransmit for between 13 and 30 minutes, depending on the connection's retransmission timeout, according to kernel documentation. This is far too long for a switch to be disconnected, so an OpenFlow switch should implement its own connection timeout. OpenFlow `OFPT_ECHO_REQUEST` messages are the best way to do this, since they test the OpenFlow connection itself.

Implementation

This section describes how Open vSwitch implements in-band control. Correctly implementing in-band control has proven difficult due to its many subtleties, and has thus gone through many iterations. Please read through and understand the reasoning behind the chosen rules before making modifications.

Open vSwitch implements in-band control as "hidden" flows, that is, flows that are not visible through OpenFlow, and at a higher priority than wildcarded flows can be set up through OpenFlow. This is done so that the OpenFlow controller cannot interfere with them and possibly break connectivity with its switches. It is possible to see all flows, including in-band ones, with the ovs-appctl "bridge/dump-flows" command.

The Open vSwitch implementation of in-band control can hide traffic to arbitrary "remotes", where each remote is one TCP port on one IP address. Currently the remotes are automatically configured as the in-band OpenFlow controllers plus the OVSDB managers, if any. (The latter is a requirement because OVSDB managers are responsible for configuring OpenFlow controllers, so if the manager cannot be reached then OpenFlow cannot be reconfigured.)

The following rules (with the OFPP_NORMAL action) are set up on any bridge that has any remotes:

(a) DHCP requests sent from the local port.

(b) ARP replies to the local port's MAC address.

(c) ARP requests from the local port's MAC address.

In-band also sets up the following rules for each unique next-hop MAC address for the remotes' IPs (the "next hop" is either the remote itself, if it is on a local subnet, or the gateway to reach the remote):

(d) ARP replies to the next hop's MAC address.

(e) ARP requests from the next hop's MAC address.

In-band also sets up the following rules for each unique remote IP address:

(f) ARP replies containing the remote's IP address as a target.

(g) ARP requests containing the remote's IP address as a source.

In-band also sets up the following rules for each unique remote (IP,port) pair:

(h) TCP traffic to the remote's IP and port.

(i) TCP traffic from the remote's IP and port.

The goal of these rules is to be as narrow as possible to allow a switch to join a network and be able to communicate with the remotes. As mentioned earlier, these rules have higher priority than the controller's rules, so if they are too broad, they may prevent the controller from implementing its policy. As such, in-band actively monitors some aspects of flow and packet processing so that the rules can be made more precise.

In-band control monitors attempts to add flows into the datapath that could interfere with its duties. The datapath only allows exact match entries, so in-band control is able to be very precise about the flows it prevents. Flows that miss in the datapath are sent to userspace to be processed, so preventing these flows from being cached in the "fast path" does not affect correctness. The only type of flow that is currently prevented is one that would prevent DHCP replies from being seen by the local port. For example, a rule that forwarded all DHCP traffic to the controller would not be allowed, but one that forwarded to all ports (including the local port) would.

As mentioned earlier, packets that miss in the datapath are sent to the userspace for processing. The userspace has its own flow table, the "classifier", so in-band checks whether any special processing is needed before the classifier is consulted. If a packet is a DHCP response to a request from the local port, the packet is forwarded to the local port, regardless of the flow table. Note that this requires L7 processing of DHCP replies to determine whether the 'chaddr' field matches the MAC address of the local port.

It is interesting to note that for an L3-based in-band control mechanism, the majority of rules are devoted to ARP traffic. At first glance, some of these rules appear redundant. However, each serves an important role. First, in order to determine the MAC address of the remote side (controller or gateway) for other ARP rules, we must allow ARP traffic for our local port with rules (b) and (c). If we are between a switch and its connection to the remote, we have to allow the other switch's ARP traffic to through. This is done with rules (d) and (e), since we do not know the addresses of the other switches a priori, but do know the remote's or gateway's. Finally, if the remote is running in a local guest VM that is not reached through the local port, the switch that is connected to the VM must allow ARP traffic based on the remote's IP address, since it will not know the MAC address of the local port that is sending the traffic or the MAC address of the remote in the guest VM.

With a few notable exceptions below, in-band should work in most network setups. The following are considered "supported" in the current implementation:

- Locally Connected. The switch and remote are on the same subnet. This uses rules (a), (b), (c), (h), and (i).
- Reached through Gateway. The switch and remote are on different subnets and must go through a gateway. This uses rules (a), (b), (c), (h), and (i).
- Between Switch and Remote. This switch is between another switch and the remote, and we want to allow the other switch's traffic through. This uses rules (d), (e), (h), and (i). It uses (b) and (c) indirectly in order to know the MAC address for rules (d) and (e). Note that DHCP for the other switch will not work unless an OpenFlow controller explicitly lets this switch pass the traffic.
- Between Switch and Gateway. This switch is between another switch and the gateway, and we want to allow the other switch's traffic through. This uses the same rules and logic as the "Between Switch and Remote" configuration described earlier.
- Remote on Local VM. The remote is a guest VM on the system running in-band control. This uses rules (a), (b), (c), (h), and (i).
- Remote on Local VM with Different Networks. The remote is a guest VM on the system running in-band control, but the local port is not used to connect to the remote. For example, an IP address is configured on eth0 of the switch. The remote's VM is connected through eth1 of the switch, but an IP address has not been configured for that port on the switch. As such, the switch will use eth0 to connect to the remote, and eth1's rules about the local port will not work. In the example, the switch attached to eth0 would use rules (a), (b), (c), (h), and (i) on eth0. The switch attached to eth1 would use rules (f), (g), (h), and (i).

The following are explicitly *not* supported by in-band control:

- Specify Remote by Name. Currently, the remote must be identified by IP address. A naive approach would be to permit all DNS traffic. Unfortunately, this would prevent the controller from defining any policy over DNS. Since switches that are located behind us need to connect to the remote, in-band cannot simply add a rule that allows DNS traffic from the local port. The "correct" way to support this is to parse DNS requests to allow all traffic related to a request for the remote's name through. Due to the potential security problems and amount of processing, we decided to hold off for the time-being.
- Differing Remotes for Switches. All switches must know the L3 addresses for all the remotes that other switches may use, since rules need to be set up to allow traffic related to those remotes through. See rules (f), (g), (h), and (i).
- Differing Routes for Switches. In order for the switch to allow other switches to connect to a remote through a gateway, it allows the gateway's traffic through with rules (d) and (e). If the routes to the remote differ for the two switches, we will not know the MAC address of the alternate gateway.

Action Reproduction

It seems likely that many controllers, at least at startup, use the OpenFlow "flow statistics" request to obtain existing flows, then compare the flows' actions against the actions that they expect to find. Before version 1.8.0, Open vSwitch always returned exact, byte-for-byte copies of the actions that had been added to the flow table. The current version of Open vSwitch does not always do this in some exceptional cases. This section lists the exceptions that controller authors must keep in mind if they compare actual actions against desired actions in a bytewise fashion:

- Open vSwitch zeros padding bytes in action structures, regardless of their values when the flows were added.
- Open vSwitch "normalizes" the instructions in OpenFlow 1.1 (and later) in the following way:
  - OVS sorts the instructions into the following order: Apply-Actions, Clear-Actions, Write-Actions, Write-Metadata, Goto-Table.
  - OVS drops Apply-Actions instructions that have empty action lists.
  - OVS drops Write-Actions instructions that have empty action sets.

Please report other discrepancies, if you notice any, so that we can fix or document them.

Suggestions

Suggestions to improve Open vSwitch are welcome at discuss@openvswitch.org.

5.1.2 Open vSwitch Datapath Development Guide

The Open vSwitch kernel module allows flexible userspace control over flow-level packet processing on selected network devices. It can be used to implement a plain Ethernet switch, network device bonding, VLAN processing, network access control, flow-based network control, and so on.

The kernel module implements multiple "datapaths" (analogous to bridges), each of which can have multiple "vports" (analogous to ports within a bridge). Each datapath also has associated with it a "flow table" that userspace populates with "flows" that map from keys based on packet headers and metadata to sets of actions. The most common action forwards the packet to another vport; other actions are also implemented.

When a packet arrives on a vport, the kernel module processes it by extracting its flow key and looking it up in the flow table. If there is a matching flow, it executes the associated actions. If there is no match, it queues the packet to userspace for processing (as part of its processing, userspace will likely set up a flow to handle further packets of the same type entirely in-kernel).

Flow Key Compatibility

Network protocols evolve over time. New protocols become important and existing protocols lose their prominence. For the Open vSwitch kernel module to remain relevant, it must be possible for newer versions to parse additional protocols as part of the flow key. It might even be desirable, someday, to drop support for parsing protocols that have become obsolete. Therefore, the Netlink interface to Open vSwitch is designed to allow carefully written userspace applications to work with any version of the flow key, past or future.

To support this forward and backward compatibility, whenever the kernel module passes a packet to userspace, it also passes along the flow key that it parsed from the packet. Userspace then extracts its own notion of a flow key from the packet and compares it against the kernel-provided version:

- If userspace's notion of the flow key for the packet matches the kernel's, then nothing special is necessary.
- If the kernel's flow key includes more fields than the userspace version of the flow key, for example if the kernel decoded IPv6 headers but userspace stopped at the Ethernet type (because it does not understand IPv6), then again nothing special is necessary. Userspace can still set up a flow in the usual way, as long as it uses the kernel-provided flow key to do it.
- If the userspace flow key includes more fields than the kernel's, for example if userspace decoded an IPv6 header but the kernel stopped at the Ethernet type, then userspace can forward the packet manually, without setting up a flow in the kernel. This case is bad for performance because every packet that the kernel considers part of the flow must go to userspace, but the forwarding behavior is correct. (If userspace can determine that the values of the extra fields would not affect forwarding behavior, then it could set up a flow anyway.)

How flow keys evolve over time is important to making this work, so the following sections go into detail.

Flow Key Format

A flow key is passed over a Netlink socket as a sequence of Netlink attributes. Some attributes represent packet metadata, defined as any information about a packet that cannot be extracted from the packet itself, e.g. the vport on which the packet was received. Most attributes, however, are extracted from headers within the packet, e.g. source and destination addresses from Ethernet, IP, or TCP headers.

The <linux/openvswitch.h> header file defines the exact format of the flow key attributes. For informal explanatory purposes here, we write them as comma-separated strings, with parentheses indicating arguments and nesting. For example, the following could represent a flow key corresponding to a TCP packet that arrived on vport 1:

```
in_port(1), eth(src=e0:91:f5:21:d0:b2, dst=00:02:e3:0f:80:a4),
eth_type(0x0800), ipv4(src=172.16.0.20, dst=172.18.0.52, proto=6, tos=0,
frag=no), tcp(src=49163, dst=80)
```

Often we ellipsize arguments not important to the discussion, e.g.:

```
in_port(1), eth(...), eth_type(0x0800), ipv4(...), tcp(...)
```

Wildcarded Flow Key Format

A wildcarded flow is described with two sequences of Netlink attributes passed over the Netlink socket. A flow key, exactly as described above, and an optional corresponding flow mask.

A wildcarded flow can represent a group of exact match flows. Each `1` bit in the mask specifies an exact match with the corresponding bit in the flow key. A `0` bit specifies a don't care bit, which will match either a `1` or `0` bit of an incoming packet. Using a wildcarded flow can improve the flow set up rate by reducing the number of new flows that need to be processed by the user space program.

Support for the mask Netlink attribute is optional for both the kernel and user space program. The kernel can ignore the mask attribute, installing an exact match flow, or reduce the number of don't care bits in the kernel to less than what was specified by the user space program. In this case, variations in bits that the kernel does not implement will simply

result in additional flow setups. The kernel module will also work with user space programs that neither support nor supply flow mask attributes.

Since the kernel may ignore or modify wildcard bits, it can be difficult for the userspace program to know exactly what matches are installed. There are two possible approaches: reactively install flows as they miss the kernel flow table (and therefore not attempt to determine wildcard changes at all) or use the kernel's response messages to determine the installed wildcards.

When interacting with userspace, the kernel should maintain the match portion of the key exactly as originally installed. This will provides a handle to identify the flow for all future operations. However, when reporting the mask of an installed flow, the mask should include any restrictions imposed by the kernel.

The behavior when using overlapping wildcarded flows is undefined. It is the responsibility of the user space program to ensure that any incoming packet can match at most one flow, wildcarded or not. The current implementation performs best-effort detection of overlapping wildcarded flows and may reject some but not all of them. However, this behavior may change in future versions.

Unique Flow Identifiers

An alternative to using the original match portion of a key as the handle for flow identification is a unique flow identifier, or "UFID". UFIDs are optional for both the kernel and user space program.

User space programs that support UFID are expected to provide it during flow setup in addition to the flow, then refer to the flow using the UFID for all future operations. The kernel is not required to index flows by the original flow key if a UFID is specified.

Basic Rule for Evolving Flow Keys

Some care is needed to really maintain forward and backward compatibility for applications that follow the rules listed under "Flow key compatibility" above.

The basic rule is obvious:

> New network protocol support must only supplement existing flow key attributes. It must not change the meaning of already defined flow key attributes.

This rule does have less-obvious consequences so it is worth working through a few examples. Suppose, for example, that the kernel module did not already implement VLAN parsing. Instead, it just interpreted the 802.1Q TPID (`0x8100`) as the Ethertype then stopped parsing the packet. The flow key for any packet with an 802.1Q header would look essentially like this, ignoring metadata:

```
eth(...), eth_type(0x8100)
```

Naively, to add VLAN support, it makes sense to add a new "vlan" flow key attribute to contain the VLAN tag, then continue to decode the encapsulated headers beyond the VLAN tag using the existing field definitions. With this change, a TCP packet in VLAN 10 would have a flow key much like this:

```
eth(...), vlan(vid=10, pcp=0), eth_type(0x0800), ip(proto=6, ...), tcp(...)
```

But this change would negatively affect a userspace application that has not been updated to understand the new "vlan" flow key attribute. The application could, following the flow compatibility rules above, ignore the "vlan" attribute that it does not understand and therefore assume that the flow contained IP packets. This is a bad assumption (the flow only contains IP packets if one parses and skips over the 802.1Q header) and it could cause the application's behavior to change across kernel versions even though it follows the compatibility rules.

The solution is to use a set of nested attributes. This is, for example, why 802.1Q support uses nested attributes. A TCP packet in VLAN 10 is actually expressed as:

```
eth(...), eth_type(0x8100), vlan(vid=10, pcp=0), encap(eth_type(0x0800),
ip(proto=6, ...), tcp(...)))
```

Notice how the `eth_type`, `ip`, and `tcp` flow key attributes are nested inside the `encap` attribute. Thus, an application that does not understand the `vlan` key will not see either of those attributes and therefore will not misinterpret them. (Also, the outer `eth_type` is still `0x8100`, not changed to `0x0800`)

Handling Malformed Packets

Don't drop packets in the kernel for malformed protocol headers, bad checksums, etc. This would prevent userspace from implementing a simple Ethernet switch that forwards every packet.

Instead, in such a case, include an attribute with "empty" content. It doesn't matter if the empty content could be valid protocol values, as long as those values are rarely seen in practice, because userspace can always forward all packets with those values to userspace and handle them individually.

For example, consider a packet that contains an IP header that indicates protocol 6 for TCP, but which is truncated just after the IP header, so that the TCP header is missing. The flow key for this packet would include a tcp attribute with all-zero `src` and `dst`, like this:

```
eth(...), eth_type(0x0800), ip(proto=6, ...), tcp(src=0, dst=0)
```

As another example, consider a packet with an Ethernet type of 0x8100, indicating that a VLAN TCI should follow, but which is truncated just after the Ethernet type. The flow key for this packet would include an all-zero-bits vlan and an empty encap attribute, like this:

```
eth(...), eth_type(0x8100), vlan(0), encap()
```

Unlike a TCP packet with source and destination ports 0, an all-zero-bits VLAN TCI is not that rare, so the CFI bit (aka VLAN_TAG_PRESENT inside the kernel) is ordinarily set in a vlan attribute expressly to allow this situation to be distinguished. Thus, the flow key in this second example unambiguously indicates a missing or malformed VLAN TCI.

Other Rules

The other rules for flow keys are much less subtle:

- Duplicate attributes are not allowed at a given nesting level.
- Ordering of attributes is not significant.
- When the kernel sends a given flow key to userspace, it always composes it the same way. This allows userspace to hash and compare entire flow keys that it may not be able to fully interpret.

Coding Rules

Implement the headers and codes for compatibility with older kernel in `linux/compat/` directory. All public functions should be exported using `EXPORT_SYMBOL` macro. Public function replacing the same-named kernel function should be prefixed with `rpl_`. Otherwise, the function should be prefixed with `ovs_`. For special case when it is not possible to follow this rule (e.g., the `pskb_expand_head()` function), the function name must be added to `linux/compat/build-aux/export-check-allowlist`, otherwise, the compilation check `check-export-symbol` will fail.

5.1.3 Fuzzing

All about fuzzing in Open vSwitch.

What is Fuzzing?

Usually, software teams do functional testing (which is great) but not security testing of their code. For example:

```
func_add(int x, int y) { return x+y; }
```

may have a unit test like so:

```
ASSERT((func_add(4,5)==9))
```

However, corner cases are usually not tested so that *x=INT_MAX; y=1* demonstrates a problem in the implementation.

Fuzz testing is routinely used to probabilistically generate such corner cases and feed them to program APIs to test their behavior.

OVS Fuzzing Infrastructure

Open vSwitch is enrolled in the oss-fuzz program. oss-fuzz is a free continuous fuzzing service and infrastructure offered by Google. An enrolled project is continually fuzzed and bug reports are sent to maintainers as and when they are generated.

The technical requirements to enrol OvS in oss-fuzz program are:

- Dockerfile (hosted in Google's ossfuzz GitHub repo)
- Bash build script (hosted in Google's ossfuzz GitHub repo)

Each of these requirements is explained in the following paragraphs.

Dockerfile

Dockerfile defines the box in which OvS will be built by oss-fuzz builder bots. This file must be self sufficient in that it must define a build environment in which OvS can be built from source code.

The build environment comprises

- Linux box provided by Google (Ubuntu)
- Packages required to build OvS (e.g., libssl-dev python etc.)
- Source code of OvS (fetched by oss-fuzz builder bots from OvS' GitHub mirror on a daily basis)
- Build script written in Bash

The Dockerfile definition for OvS is located in the *projects/openvswitch/Dockerfile* sub-directory of Google's oss-fuzz repo.

Build script

The build script defines steps required to compile OvS from source code. The (Linux) box used by oss-fuzz builder bots (defined by Dockerfile) is different from the box in which fuzzing actually happens. It follows that the build script must ensure that fuzzing binaries are linked statically so that no assumption is made about packages available in the fuzzing box.

OvS contains a make target called *oss-fuzz-targets* for compiling and linking OvS fuzzer binaries. The line of bash script responsible for building statically linked OvS fuzzing binaries is the following:

```
./boot.sh && ./configure && make -j$(nproc) && make oss-fuzz-targets
```

The oss-fuzz build environment assumes that OvS build system respects compiler/linker flags defined via standard bash environment variables called *CFLAGS*, *CXXFLAGS* etc. The oss-fuzz builder bot defines these flags so that OvS fuzzing binaries are correctly instrumented.

oss-fuzz expects all fuzzing binaries, and optionally, configuration and seed inputs to be located in a hard-coded directory, referenced by the bash variable *$OUT*, in the root filesystem of the build box.

OvS source repo contains configuration for the oss-fuzz fuzzers in the *tests/oss-fuzz/config* sub-directory. There are two types of configuration files:

- <fuzzer_target_name>.options: Defines configuration options for the fuzzer that apply while fuzzing *fuzzer_target_name*
- <name>.dict: Defines a dictionary to be used for some *fuzzer_target_name*

Fuzzer configuration parameters of relevance to OvS are:

- dict: names the dictionary file to be used for fuzzing
- close_fd_mask: defines a file descriptor mask that filters console output generated by OvS fuzzing binaries

OVS Fuzzers

OvS fuzzer test harnesses define the libFuzzer fuzz API. In doing so, they define what is to be done with the input supplied by the fuzzer.

At a minimum, the libfuzzer API is defined as follows:

```
// input_ is a byte array, size is the length of said byte array
int
LLVMFuzzerTestOneInput(const uint8_t *input, size_t size)
{
    // Input processing
    process_input(input, size);

    // Must always return 0. Non-zero return codes are reserved by libFuzzer.
    return 0;
}
```

In certain scenarios, it may be necessary to constrain the input supplied by the fuzzer. One scenario is when *process_input* accepts a C string. One way to do this would be as follows:

```
// input_ is a byte array, size is the length of said byte array
int
LLVMFuzzerTestOneInput(const uint8_t *input, size_t size)
{
    // Constrain input
    // Check if input is null terminated
    const char *cstring = (const char*) input;
    if (cstring[size - 1] != '\0')
        return 0;

    // Input processing
    process_input(cstring);

    // Must always return 0. Non-zero return codes are reserved by libFuzzer.
    return 0;
}
```

OvS fuzzer test harnesses are located in the *tests/oss-fuzz* sub-directory. At the time of writing, there are a total of six harnesses:

- *flow_extract_target.c*
- *json_parser_target.c*
- *miniflow_target.c*
- *odp_target.c*
- *ofctl_parse_target.c*
- *ofp_print_target.c*

flow_extract_target

Extracts flow from and parses fuzzer supplied packet payload.

json_parser_target

Parses fuzzer supplied string as JSON, encoding the parsed JSON into a JSON RPC message, and finally decoding the encoded JSON RPC message back to JSON.

miniflow_target

Extracts flow from fuzzer supplied packet payload, converts flow to a miniflow and performs various miniflow operations.

odp_target

Parses fuzzer supplied string as an ODP flow, and the same string as an ODP action.

ofctl_parse_target

Treats fuzzer supplied input as a <flow_command> followed by a <flow_mod_string>, invoking the *parse_ofp_flow_mod_str* on the pair.

ofp_print_target

Parses fuzzer supplied data as an Open Flow Protocol buffer.

Security Analysis of OVS Fuzzers

OvS fuzzer harnesses test different components of OvS. This document performs a security analysis of these harnesses. The intention is to not only tabulate what is currently done but also to help expand on the set of harnesses.

Fuzzer harness	Interface	Input source
flow_extract_target	External	Untrusted
json_parser_target	OVS DB	Trusted
miniflow_target	External	Untrusted
odp_target	North-bound	Trusted
ofctl_parse_target	Management	Trusted
ofp_print_target	South/North-bound	Trusted

5.1.4 Integration Guide for Centralized Control

This document describes how to integrate Open vSwitch onto a new platform to expose the state of the switch and attached devices for centralized control. (If you are looking to port the switching components of Open vSwitch to a new platform, refer to *Porting Open vSwitch to New Software or Hardware*) The focus of this guide is on hypervisors, but many of the interfaces are useful for hardware switches, as well.

The externally visible interface to this integration is platform-agnostic. We encourage anyone who integrates Open vSwitch to use the same interface, because keeping a uniform interface means that controllers require less customization for individual platforms (and perhaps no customization at all).

Integration centers around the Open vSwitch database and mostly involves the `external_ids` columns in several of the tables. These columns are not interpreted by Open vSwitch itself. Instead, they provide information to a controller that permits it to associate a database record with a more meaningful entity. In contrast, the `other_config` column is used to configure behavior of the switch. The main job of the integrator, then, is to ensure that these values are correctly populated and maintained.

An integrator sets the columns in the database by talking to the ovsdb-server daemon. A few of the columns can be set during startup by calling the ovs-ctl tool from inside the startup scripts. The `rhel/etc_init.d_openvswitch` script provides examples of its use, and the ovs-ctl(8) manpage contains complete documentation. At runtime, ovs-vsctl can be used to set columns in the database.

Python and C bindings to the database are provided if deeper integration with a program are needed. More information on the python bindings is available at `python/ovs/db/idl.py`. Information on the C bindings is available at `lib/ovsdb-idl.h`.

The following diagram shows how integration scripts fit into the Open vSwitch architecture:

```
Diagram

         +---------------------------------------+
         |           Controller Cluster          +
         +---------------------------------------+
                              |
                              |
+-------------------------------------------------------------+
|                             |                               |
|             +---------------+---------------+               |
|             |                               |               |
|   +---------------------+           +-------------------+   |
|   |   ovsdb-server      |-----------|   ovs-vswitchd    |   |
|   +---------------------+           +-------------------+   |
|             |                                 |             |
|  +----------------------+                     |             |
|  | Integration scripts |                      |             |
|  +----------------------+                     |             |
|                                               |   Userspace |
|-------------------------------------------------------------|
|                                               |      Kernel |
|                                               |             |
|                                +--------------------+       |
|                                |  OVS Kernel Module |       |
|                                +--------------------+       |
+-------------------------------------------------------------+
```

A description of the most relevant fields for integration follows. By setting these values, controllers are able to understand the network and manage it more dynamically and precisely. For more details about the database and each

individual column, please refer to the ovs-vswitchd.conf.db(5) manpage.

`Open_vSwitch` table

The `Open_vSwitch` table describes the switch as a whole. The `system_type` and `system_version` columns identify the platform to the controller. The `external_ids:system-id` key uniquely identifies the physical host. This key allows controllers to distinguish between multiple hypervisors.

Most of this configuration can be done with the ovs-ctl command at startup. For example:

```
$ ovs-ctl --system-type="KVM" --system-version="4.18.el8_6" \
    --system-id="${UUID}" "${other_options}" start
```

Alternatively, the ovs-vsctl command may be used to set a particular value at runtime. For example:

```
$ ovs-vsctl set open_vswitch . external-ids:system-id='"${UUID}"'
```

The `other_config:enable-statistics` key may be set to `true` to have OVS populate the database with statistics (e.g., number of CPUs, memory, system load) for the controller's use.

Bridge table

The Bridge table describes individual bridges within an Open vSwitch instance. The `external-ids:bridge-id` key uniquely identifies a particular bridge.

For example, to set the identifier for bridge "br0", the following command can be used:

```
$ ovs-vsctl set Bridge br0 external-ids:bridge-id='"${UUID}"'
```

The MAC address of the bridge may be manually configured by setting it with the `other_config:hwaddr` key. For example:

```
$ ovs-vsctl set Bridge br0 other_config:hwaddr="12:34:56:78:90:ab"
```

Interface table

The Interface table describes an interface under the control of Open vSwitch. The `external_ids` column contains keys that are used to provide additional information about the interface:

attached-mac

> This field contains the MAC address of the device attached to the interface. On a hypervisor, this is the MAC address of the interface as seen inside a VM. It does not necessarily correlate to the host-side MAC address.

iface-id

> This field uniquely identifies the interface. In hypervisors, this allows the controller to follow VM network interfaces as VMs migrate. A well-chosen identifier should also allow an administrator or a controller to associate the interface with the corresponding object in the VM management system.

iface-status

> In a hypervisor, there are situations where there are multiple interface choices for a single virtual ethernet interface inside a VM. Valid values are "active" and "inactive". A complete description is available in the ovs-vswitchd.conf.db(5) manpage.

vm-id

This field uniquely identifies the VM to which this interface belongs. A single VM may have multiple interfaces attached to it.

As in the previous tables, the ovs-vsctl command may be used to configure the values. For example, to set the `iface-id` on eth0, the following command can be used:

```
$ ovs-vsctl set Interface eth0 external-ids:iface-id='"${UUID}"'
```

HA for OVN DB servers using pacemaker

The ovsdb servers can work in either active or backup mode. In backup mode, db server will be connected to an active server and replicate the active servers contents. At all times, the data can be transacted only from the active server. When the active server dies for some reason, entire OVN operations will be stalled.

Pacemaker is a cluster resource manager which can manage a defined set of resource across a set of clustered nodes. Pacemaker manages the resource with the help of the resource agents. One among the resource agent is OCF

OCF is nothing but a shell script which accepts a set of actions and returns an appropriate status code.

With the help of the OCF resource agent ovn/utilities/ovndb-servers.ocf, one can defined a resource for the pacemaker such that pacemaker will always maintain one running active server at any time.

After creating a pacemaker cluster, use the following commands to create one active and multiple backup servers for OVN databases:

```
$ pcs resource create ovndb_servers ocf:ovn:ovndb-servers \
    master_ip=x.x.x.x \
    ovn_ctl=<path of the ovn-ctl script> \
    op monitor interval="10s" \
    op monitor role=Master interval="15s"
$ pcs resource master ovndb_servers-master ovndb_servers \
   meta notify="true"
```

The *master_ip* and *ovn_ctl* are the parameters that will be used by the OCF script. *ovn_ctl* is optional, if not given, it assumes a default value of /usr/share/openvswitch/scripts/ovn-ctl. *master_ip* is the IP address on which the active database server is expected to be listening, the slave node uses it to connect to the master node. You can add the optional parameters 'nb_master_port', 'nb_master_protocol', 'sb_master_port', 'sb_master_protocol' to set the protocol and port.

Whenever the active server dies, pacemaker is responsible to promote one of the backup servers to be active. Both ovn-controller and ovn-northd needs the ip-address at which the active server is listening. With pacemaker changing the node at which the active server is run, it is not efficient to instruct all the ovn-controllers and the ovn-northd to listen to the latest active server's ip-address.

This problem can be solved by two ways:

1. By using a native ocf resource agent `ocf:heartbeat:IPaddr2`. The IPAddr2 resource agent is just a resource with an ip-address. When we colocate this resource with the active server, pacemaker will enable the active server to be connected with a single ip-address all the time. This is the ip-address that needs to be given as the parameter while creating the *ovndb_servers* resource.

Use the following command to create the IPAddr2 resource and colocate it with the active server:

```
$ pcs resource create VirtualIP ocf:heartbeat:IPaddr2 ip=x.x.x.x \
   op monitor interval=30s
$ pcs constraint order promote ovndb_servers-master then VirtualIP
$ pcs constraint colocation add VirtualIP with master ovndb_servers-master \
   score=INFINITY
```

2. Using load balancer vip ip as a master_ip. In order to use this feature, one needs to use listen_on_master_ip_only to no. Current code for load balancer have been tested to work with tcp protocol and needs to be tested/enhanced for ssl. Using load balancer, standby nodes will not listen on nb and sb db ports so that load balancer will always communicate to the active node and all the traffic will be sent to active node only. Standby will continue to sync using LB VIP IP in this case.

Use the following command to create pcs resource using LB VIP IP:

```
$ pcs resource create ovndb_servers ocf:ovn:ovndb-servers \
     master_ip="<load_balance_vip_ip>" \
     listen_on_master_ip_only="no" \
     ovn_ctl=<path of the ovn-ctl script> \
     op monitor interval="10s" \
     op monitor role=Master interval="15s"
$ pcs resource master ovndb_servers-master ovndb_servers \
    meta notify="true"
```

5.1.5 Porting Open vSwitch to New Software or Hardware

Open vSwitch (OVS) is intended to be easily ported to new software and hardware platforms. This document describes the types of changes that are most likely to be necessary in porting OVS to Unix-like platforms. (Porting OVS to other kinds of platforms is likely to be more difficult.)

Vocabulary

For historical reasons, different words are used for essentially the same concept in different areas of the Open vSwitch source tree. Here is a concordance, indexed by the area of the source tree:

```
datapath/       vport           ---
vswitchd/       iface           port
ofproto/        port            bundle
ofproto/bond.c  member          bond
lib/lacp.c      member          lacp
lib/netdev.c    netdev          ---
database        Interface       Port
```

Open vSwitch Architectural Overview

The following diagram shows the very high-level architecture of Open vSwitch from a porter's perspective.

```
+-------------------+
|    ovs-vswitchd   |<-->ovsdb-server
+-------------------+
|      ofproto      |<-->OpenFlow controllers
+--------+-+--------+
| netdev | | ofproto|
+--------+ |provider|
| netdev | +--------+
|provider|
+--------+
```

Some of the components are generic. Modulo bugs or inadequacies, these components should not need to be modified as part of a port:

ovs-vswitchd
: The main Open vSwitch userspace program, in vswitchd/. It reads the desired Open vSwitch configuration from

the ovsdb-server program over an IPC channel and passes this configuration down to the "ofproto" library. It also passes certain status and statistical information from ofproto back into the database.

ofproto
The Open vSwitch library, in ofproto/, that implements an OpenFlow switch. It talks to OpenFlow controllers over the network and to switch hardware or software through an "ofproto provider", explained further below.

netdev
The Open vSwitch library, in lib/netdev.c, that abstracts interacting with network devices, that is, Ethernet interfaces. The netdev library is a thin layer over "netdev provider" code, explained further below.

The other components may need attention during a port. You will almost certainly have to implement a "netdev provider". Depending on the type of port you are doing and the desired performance, you may also have to implement an "ofproto provider" or a lower-level component called a "dpif" provider.

The following sections talk about these components in more detail.

Writing a netdev Provider

A "netdev provider" implements an operating system and hardware specific interface to "network devices", e.g. eth0 on Linux. Open vSwitch must be able to open each port on a switch as a netdev, so you will need to implement a "netdev provider" that works with your switch hardware and software.

`struct netdev_class`, in `lib/netdev-provider.h`, defines the interfaces required to implement a netdev. That structure contains many function pointers, each of which has a comment that is meant to describe its behavior in detail. If the requirements are unclear, report this as a bug.

The netdev interface can be divided into a few rough categories:

- Functions required to properly implement OpenFlow features. For example, OpenFlow requires the ability to report the Ethernet hardware address of a port. These functions must be implemented for minimally correct operation.
- Functions required to implement optional Open vSwitch features. For example, the Open vSwitch support for in-band control requires netdev support for inspecting the TCP/IP stack's ARP table. These functions must be implemented if the corresponding OVS features are to work, but may be omitted initially.
- Functions needed in some implementations but not in others. For example, most kinds of ports (see below) do not need functionality to receive packets from a network device.

The existing netdev implementations may serve as useful examples during a port:

- lib/netdev-linux.c implements netdev functionality for Linux network devices, using Linux kernel calls. It may be a good place to start for full-featured netdev implementations.
- lib/netdev-vport.c provides support for "virtual ports" implemented by the Open vSwitch datapath module for the Linux kernel. This may serve as a model for minimal netdev implementations.
- lib/netdev-dummy.c is a fake netdev implementation useful only for testing.

Porting Strategies

After a netdev provider has been implemented for a system's network devices, you may choose among three basic porting strategies.

The lowest-effort strategy is to use the "userspace switch" implementation built into Open vSwitch. This ought to work, without writing any more code, as long as the netdev provider that you implemented supports receiving packets. It yields poor performance, however, because every packet passes through the ovs-vswitchd process. Refer to *Open vSwitch without Kernel Support* for instructions on how to configure a userspace switch.

If the userspace switch is not the right choice for your port, then you will have to write more code. You may implement either an "ofproto provider" or a "dpif provider". Which you should choose depends on a few different factors:

- Only an ofproto provider can take full advantage of hardware with built-in support for wildcards (e.g. an ACL table or a TCAM).
- A dpif provider can take advantage of the Open vSwitch built-in implementations of bonding, LACP, 802.1ag, 802.1Q VLANs, and other features. An ofproto provider has to provide its own implementations, if the hardware can support them at all.
- A dpif provider is usually easier to implement, but most appropriate for software switching. It "explodes" wildcard rules into exact-match entries (with an optional wildcard mask). This allows fast hash lookups in software, but makes inefficient use of TCAMs in hardware that support wildcarding.

The following sections describe how to implement each kind of port.

ofproto Providers

An "ofproto provider" is what ofproto uses to directly monitor and control an OpenFlow-capable switch. `struct ofproto_class`, in `ofproto/ofproto-provider.h`, defines the interfaces to implement an ofproto provider for new hardware or software. That structure contains many function pointers, each of which has a comment that is meant to describe its behavior in detail. If the requirements are unclear, report this as a bug.

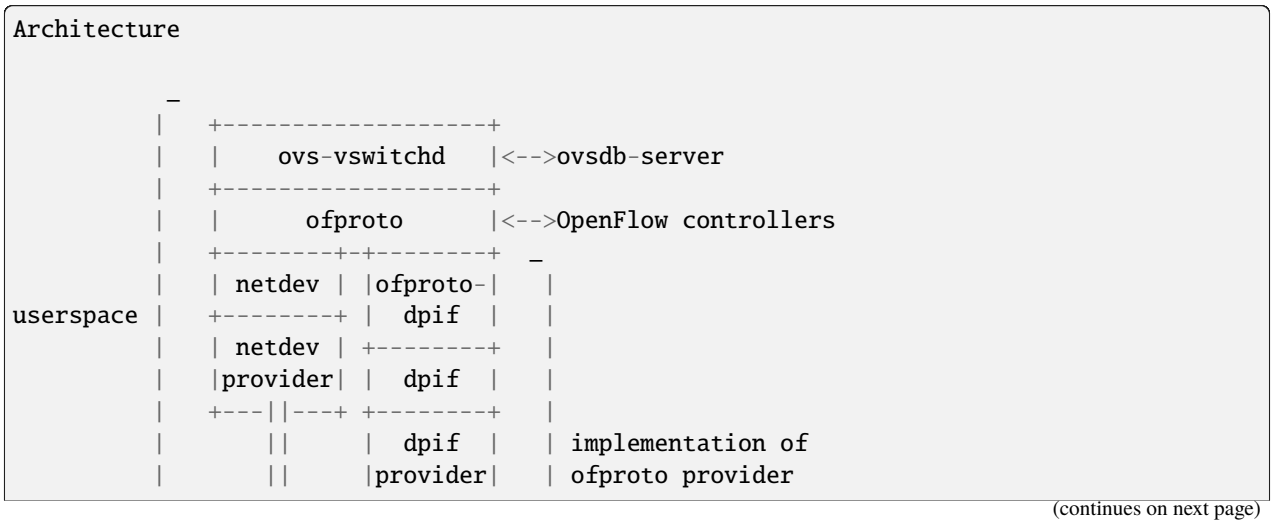
The ofproto provider interface is preliminary. Let us know if it seems unsuitable for your purpose. We will try to improve it.

Writing a dpif Provider

Open vSwitch has a built-in ofproto provider named "ofproto-dpif", which is built on top of a library for manipulating datapaths, called "dpif". A "datapath" is a simple flow table, one that is only required to support exact-match flows, that is, flows without wildcards. When a packet arrives on a network device, the datapath looks for it in this table. If there is a match, then it performs the associated actions. If there is no match, the datapath passes the packet up to ofproto-dpif, which maintains the full OpenFlow flow table. If the packet matches in this flow table, then ofproto-dpif executes its actions and inserts a new entry into the dpif flow table. (Otherwise, ofproto-dpif passes the packet up to ofproto to send the packet to the OpenFlow controller, if one is configured.)

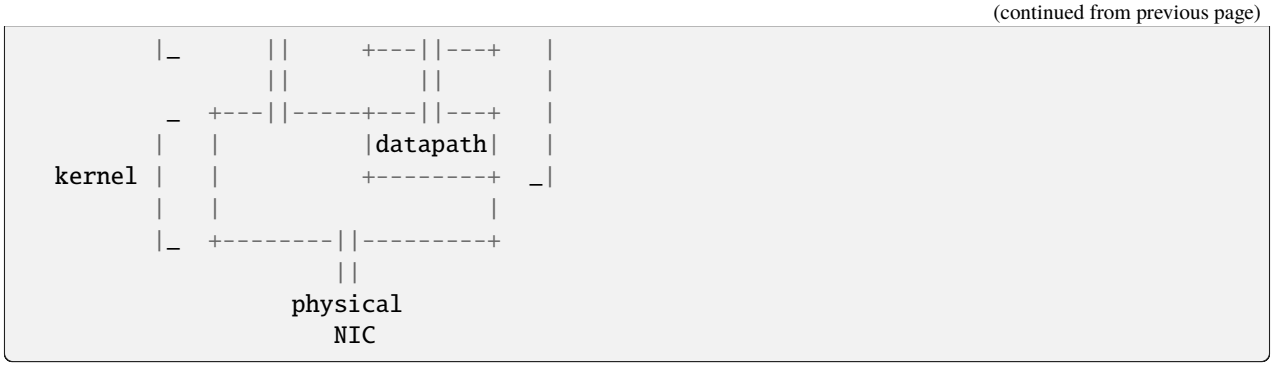
When calculating the dpif flow, ofproto-dpif generates an exact-match flow that describes the missed packet. It makes an effort to figure out what fields can be wildcarded based on the switch's configuration and OpenFlow flow table. The dpif is free to ignore the suggested wildcards and only support the exact-match entry. However, if the dpif supports wildcarding, then it can use the masks to match multiple flows with fewer entries and potentially significantly reduce the number of flow misses handled by ofproto-dpif.

The "dpif" library in turn delegates much of its functionality to a "dpif provider". The following diagram shows how dpif providers fit into the Open vSwitch architecture:



(continues on next page)

(continued from previous page)



struct `dpif_class`, in `lib/dpif-provider.h`, defines the interfaces required to implement a dpif provider for new hardware or software. That structure contains many function pointers, each of which has a comment that is meant to describe its behavior in detail. If the requirements are unclear, report this as a bug.

There are two existing dpif implementations that may serve as useful examples during a port:

- lib/dpif-netlink.c is a Linux-specific dpif implementation that talks to an Open vSwitch-specific kernel module (whose sources are in the "datapath" directory). The kernel module performs all of the switching work, passing packets that do not match any flow table entry up to userspace. This dpif implementation is essentially a wrapper around calls into the kernel module.
- lib/dpif-netdev.c is a generic dpif implementation that performs all switching internally. This is how the Open vSwitch userspace switch is implemented.

Miscellaneous Notes

Open vSwitch source code uses `uint16_t`, `uint32_t`, and `uint64_t` as fixed-width types in host byte order, and `ovs_be16`, `ovs_be32`, and `ovs_be64` as fixed-width types in network byte order. Each of the latter is equivalent to the one of the former, but the difference in name makes the intended use obvious.

The default "fail-mode" for Open vSwitch bridges is "standalone", meaning that, when the OpenFlow controllers cannot be contacted, Open vSwitch acts as a regular MAC-learning switch. This works well in virtualization environments where there is normally just one uplink (either a single physical interface or a bond). In a more general environment, it can create loops. So, if you are porting to a general-purpose switch platform, you should consider changing the default "fail-mode" to "secure", which does not behave this way. See documentation for the "fail-mode" column in the Bridge table in ovs-vswitchd.conf.db(5) for more information.

`lib/entropy.c` assumes that it can obtain high-quality random number seeds at startup by reading from /dev/urandom. You will need to modify it if this is not true on your platform.

`vswitchd/system-stats.c` only knows how to obtain some statistics on Linux. Optionally you may implement them for your platform as well.

Why OVS Does Not Support Hybrid Providers

The *porting strategies* section above describes the "ofproto provider" and "dpif provider" porting strategies. Only an ofproto provider can take advantage of hardware TCAM support, and only a dpif provider can take advantage of the OVS built-in implementations of various features. It is therefore tempting to suggest a hybrid approach that shares the advantages of both strategies.

However, Open vSwitch does not support a hybrid approach. Doing so may be possible, with a significant amount of extra development work, but it does not yet seem worthwhile, for the reasons explained below.

First, user surprise is likely when a switch supports a feature only with a high performance penalty. For example, one user questioned why adding a particular OpenFlow action to a flow caused a 1,058x slowdown on a hardware OpenFlow

implementation[1]. The action required the flow to be implemented in software.

Given that implementing a flow in software on the slow management CPU of a hardware switch causes a major slowdown, software-implemented flows would only make sense for very low-volume traffic. But many of the features built into the OVS software switch implementation would need to apply to every flow to be useful. There is no value, for example, in applying bonding or 802.1Q VLAN support only to low-volume traffic.

Besides supporting features of OpenFlow actions, a hybrid approach could also support forms of matching not supported by particular switching hardware, by sending all packets that might match a rule to software. But again this can cause an unacceptable slowdown by forcing bulk traffic through software in the hardware switch's slow management CPU. Consider, for example, a hardware switch that can match on the IPv6 Ethernet type but not on fields in IPv6 headers. An OpenFlow table that matched on the IPv6 Ethernet type would perform well, but adding a rule that matched only UDPv6 would force every IPv6 packet to software, slowing down not just UDPv6 but all IPv6 processing.

Questions

Direct porting questions to dev@openvswitch.org. We will try to use questions to improve this porting guide.

5.1.6 OpenFlow Support in Open vSwitch

Open vSwitch support for OpenFlow 1.1 and beyond is a work in progress. This file describes the work still to be done.

The Plan

OpenFlow version support is not a build-time option. A single build of Open vSwitch must be able to handle all supported versions of OpenFlow. Ideally, even at runtime it should be able to support all protocol versions at the same time on different OpenFlow bridges (and perhaps even on the same bridge).

At the same time, it would be a shame to litter the core of the OVS code with lots of ugly code concerned with the details of various OpenFlow protocol versions.

The primary approach to compatibility is to abstract most of the details of the differences from the core code, by adding a protocol layer that translates between OF1.x and a slightly higher-level abstract representation. The core of this approach is the many `struct ofputil_*` structures in `include/openvswitch/ofp-*.h`.

As a consequence of this approach, OVS cannot use OpenFlow protocol definitions that closely resemble those in the OpenFlow specification, because `openflow.h` in different versions of the OpenFlow specification defines the same identifier with different values. Instead, `openflow-common.h` contains definitions that are common to all the specifications and separate protocol version-specific headers contain protocol-specific definitions renamed so as not to conflict, e.g. `OFPAT10_ENQUEUE` and `OFPAT11_ENQUEUE` for the OpenFlow 1.0 and 1.1 values for `OFPAT_ENQUEUE`. Generally, in cases of conflict, the protocol layer will define a more abstract `OFPUTIL_*` or struct `ofputil_*`.

Here are the current approaches in a few tricky areas:

- Port numbering.

  OpenFlow 1.0 has 16-bit port numbers and later OpenFlow versions have 32-bit port numbers. For now, OVS support for later protocol versions requires all port numbers to fall into the 16-bit range, translating the reserved `OFPP_*` port numbers.

- Actions.

  OpenFlow 1.0 and later versions have very different ideas of actions. OVS reconciles by translating all the versions' actions (and instructions) to and from a common internal representation.

[1] Aaron Rosen, "Modify packet fields extremely slow", openflow-discuss mailing list, June 26, 2011, archived at https://www.mail-archive.com/openflow-discuss@lists.stanford.edu/msg00447.html

OpenFlow 1.1

OpenFlow 1.1 support is complete.

OpenFlow 1.2

OpenFlow 1.2 support is complete.

OpenFlow 1.3

OpenFlow 1.3 support requires OpenFlow 1.2 as a prerequisite, plus the following additional work. (This is based on the change log at the end of the OF1.3 spec, reusing most of the section titles directly. I didn't compare the specs carefully yet.)

- IPv6 extension header handling support.

  Fully implementing this requires kernel support. This likely will take some careful and probably time-consuming design work. The actual coding, once that is all done, is probably 2 or 3 days work.

  (optional for OF1.3+)

- Auxiliary connections.

  An implementation in generic code might be a week's worth of work. The value of an implementation in generic code is questionable, though, since much of the benefit of axuiliary connections is supposed to be to take advantage of hardware support. (We could make the kernel module somehow send packets across the auxiliary connections directly, for some kind of "hardware" support, if we judged it useful enough.)

  (optional for OF1.3+)

- Provider Backbone Bridge tagging.

  I don't plan to implement this (but we'd accept an implementation).

  (optional for OF1.3+)

- On-demand flow counters.

  I think this might be a real optimization in some cases for the software switch.

  (optional for OF1.3+)

OpenFlow 1.4 & ONF Extensions for 1.3.X Pack1

The following features are both defined as a set of ONF Extensions for 1.3 and integrated in 1.4.

When defined as an ONF Extension for 1.3, the feature is using the Experimenter mechanism with the ONF Experimenter ID.

When defined integrated in 1.4, the feature use the standard OpenFlow structures (for example defined in openflow-1.4.h).

The two definitions for each feature are independent and can exist in parallel in OVS.

- Flow entry notifications

  This seems to be modelled after OVS's NXST_FLOW_MONITOR.

  (EXT-187) (optional for OF1.4+)

- Flow entry eviction

  OVS has flow eviction functionality. `table_mod OFPTC_EVICTION`, `flow_mod 'importance'`, and `table_desc ofp_table_mod_prop_eviction` need to be implemented.

  (EXT-192-e)

(optional for OF1.4+)

- Vacancy events

  (EXT-192-v)

  (optional for OF1.4+)

- Table synchronisation

  Probably not so useful to the software switch.

  (EXT-232)

  (optional for OF1.4+)

- Group and Meter change notifications

  (EXT-235)

  (optional for OF1.4+)

- PBB UCA header field

  See comment on Provider Backbone Bridge in section about OpenFlow 1.3.

  (EXT-256)

  (optional for OF1.4+)

OpenFlow 1.4 only

Those features are those only available in OpenFlow 1.4, other OpenFlow 1.4 features are listed in the previous section.

- Optical port properties

  (EXT-154)

  (optional for OF1.4+)

OpenFlow 1.5 & ONF Extensions for 1.3.X Pack2

The following features are both defined as a set of ONF Extensions for 1.3 and integrated in 1.5. Note that this list is not definitive as those are not yet published.

When defined as an ONF Extension for 1.3, the feature is using the Experimenter mechanism with the ONF Experimenter ID. When defined integrated in 1.5, the feature use the standard OpenFlow structures (for example defined in openflow-1.5.h).

The two definitions for each feature are independent and can exist in parallel in OVS.

- Time scheduled bundles

  (EXT-340)

  (optional for OF1.5+)

OpenFlow 1.5 only

Those features are those only available in OpenFlow 1.5, other OpenFlow 1.5 features are listed in the previous section.

- Egress Tables

  (EXT-306)

  (optional for OF1.5+)

- Flow Entry Statistics Trigger

  (EXT-335)

  (optional for OF1.5+)

- Controller connection status

  Prototype for OVS was done during specification.

  (EXT-454)

  (optional for OF1.5+)

- Port properties for pipeline fields

  Prototype for OVS was done during specification.

  (EXT-388)

  (optional for OF1.5+)

- Port property for recirculation

  Prototype for OVS was done during specification.

  (EXT-399)

  (optional for OF1.5+)

General

- ovs-ofctl(8) often lists as Nicira extensions features that later OpenFlow versions support in standard ways.

How to contribute

If you plan to contribute code for a feature, please let everyone know on ovs-dev before you start work. This will help avoid duplicating work.

Consider the following:

- Testing.

  Please test your code.

- Unit tests.

  Consider writing some. The tests directory has many examples that you can use as a starting point.

- ovs-ofctl.

  If you add a feature that is useful for some ovs-ofctl command then you should add support for it there.

- Documentation.

  If you add a user-visible feature, then you should document it in the appropriate manpage and mention it in NEWS as well.

Refer to *Contributing to Open vSwitch* for more information.

5.1.7 Bonding

Bonding allows two or more interfaces, its "members", to share network traffic. From a high-level point of view, bonded interfaces act like a single port, but they have the bandwidth of multiple network devices, e.g. two 1 GB physical interfaces act like a single 2 GB interface. Bonds also increase robustness: the bonded port does not go down as long as at least one of its members is up.

In vswitchd, a bond always has at least two members (and may have more). If a configuration error, etc. would cause a bond to have only one member, the port becomes an ordinary port, not a bonded port, and none of the special features of bonded ports described in this section apply.

There are many forms of bonding of which ovs-vswitchd implements only a few. The most complex bond ovs-vswitchd implements is called "source load balancing" or SLB bonding. SLB bonding divides traffic among the members based on the Ethernet source address. This is useful only if the traffic over the bond has multiple Ethernet source addresses, for example if network traffic from multiple VMs are multiplexed over the bond.

> **Note**
>
> Most of the ovs-vswitchd implementation is in `vswitchd/bridge.c`, so code references below should be assumed to refer to that file except as otherwise specified.

Enabling and Disabling Members

When a bond is created, a member is initially enabled or disabled based on whether carrier is detected on the NIC (see `iface_create()`). After that, a member is disabled if its carrier goes down for a period of time longer than the downdelay, and it is enabled if carrier comes up for longer than the updelay (see `bond_link_status_update()`). There is one exception where the updelay is skipped: if no members at all are currently enabled, then the first member on which carrier comes up is enabled immediately.

The updelay should be set to a time longer than the STP forwarding delay of the physical switch to which the bond port is connected (if STP is enabled on that switch). Otherwise, the member will be enabled, and load may be shifted to it, before the physical switch starts forwarding packets on that port, which can cause some data to be dropped for a time. The exception for a single enabled member does not cause any problem in this regard because when no members are enabled all output packets are dropped anyway.

When a member becomes disabled, the vswitch immediately chooses a new output port for traffic that was destined for that member (see `bond_enable_member()`). It also sends a "gratuitous learning packet", specifically a RARP, on the bond port (on the newly chosen member) for each MAC address that the vswitch has learned on a port other than the bond (see `bundle_send_learning_packets()`), to teach the physical switch that the new member should be used in place of the one that is now disabled. (This behavior probably makes sense only for a vswitch that has only one port (the bond) connected to a physical switch; vswitchd should probably provide a way to disable or configure it in other scenarios.)

Bond Packet Input

Bonding accepts unicast packets on any member. This can occasionally cause packet duplication for the first few packets sent to a given MAC, if the physical switch attached to the bond is flooding packets to that MAC because it has not yet learned the correct member for that MAC.

Bonding only accepts multicast (and broadcast) packets on a single bond member (the "active member") at any given time. Multicast packets received on other members are dropped. Otherwise, every multicast packet would be duplicated, once for every bond member, because the physical switch attached to the bond will flood those packets.

Bonding also drops received packets when the vswitch has learned that the packet's MAC is on a port other than the bond port itself. This is because it is likely that the vswitch itself sent the packet out the bond port on a different member and is now receiving the packet back. This occurs when the packet is multicast or the physical switch has not yet learned the MAC and is flooding it. However, the vswitch makes an exception to this rule for broadcast ARP replies, which indicate that the MAC has moved to another switch, probably due to VM migration. (ARP replies are normally unicast, so this exception does not match normal ARP replies. It will match the learning packets sent on bond fail-over.)

The active member is simply the first member to be enabled after the bond is created (see `bond_choose_active_member()`). If the active member is disabled, then a new active member is chosen among the members that remain active. Currently due to the way that configuration works, this tends to be the remaining member whose interface name is first alphabetically, but this is by no means guaranteed.

Bond Packet Output

When a packet is sent out a bond port, the bond member actually used is selected based on the packet's source MAC and VLAN tag (see `bond_choose_output_member()`). In particular, the source MAC and VLAN tag are hashed into one of 256 values, and that value is looked up in a hash table (the "bond hash") kept in the `bond_hash` member of struct port. The hash table entry identifies a bond member. If no bond member has yet been chosen for that hash table entry, vswitchd chooses one arbitrarily.

Every 10 seconds, vswitchd rebalances the bond members (see `bond_rebalance()`). To rebalance, vswitchd examines the statistics for the number of bytes transmitted by each member over approximately the past minute, with data sent more recently weighted more heavily than data sent less recently. It considers each of the members in order from most-loaded to least-loaded. If highly loaded member H is significantly more heavily loaded than the least-loaded member L, and member H carries at least two hashes, then vswitchd shifts one of H's hashes to L. However, vswitchd will only shift a hash from H to L if it will decrease the ratio of the load between H and L by at least 0.1.

Currently, "significantly more loaded" means that H must carry at least 1 Mbps more traffic, and that traffic must be at least 3% greater than L's.

Bond Balance Modes

Each bond balancing mode has different considerations, described below.

LACP Bonding

LACP bonding requires the remote switch to implement LACP, but it is otherwise very simple in that, after LACP negotiation is complete, there is no need for special handling of received packets.

Several of the physical switches that support LACP block all traffic for ports that are configured to use LACP, until LACP is negotiated with the host. When configuring a LACP bond on a OVS host, this means that there will be an interruption of the network connectivity between the time the ports on the physical switch and the bond on the OVS host are configured. The interruption may be relatively long, if different people are responsible for managing the switches and the OVS host.

Such network connectivity failure can be avoided if LACP can be configured on the OVS host before configuring the physical switch, and having the OVS host fall back to a bond mode (active-backup) till the physical switch LACP configuration is complete. An option "lacp-fallback-ab" exists to provide such behavior on Open vSwitch.

Active Backup Bonding

Active Backup bonds send all traffic out one "active" member until that member becomes unavailable. Since they are significantly less complicated than SLB bonds, they are preferred when LACP is not an option. Additionally, they are the only bond mode which supports attaching each member to a different upstream switch.

SLB Bonding

SLB bonding allows a limited form of load balancing without the remote switch's knowledge or cooperation. The basics of SLB are simple. SLB assigns each source MAC+VLAN pair to a link and transmits all packets from that MAC+VLAN through that link. Learning in the remote switch causes it to send packets to that MAC+VLAN through the same link.

SLB bonding has the following complications:

0. When the remote switch has not learned the MAC for the destination of a unicast packet and hence floods the packet to all of the links on the SLB bond, Open vSwitch will forward duplicate packets, one per link, to each other switch port.

   Open vSwitch does not solve this problem.

1. When the remote switch receives a multicast or broadcast packet from a port not on the SLB bond, it will forward it to all of the links in the SLB bond. This would cause packet duplication if not handled specially.

   Open vSwitch avoids packet duplication by accepting multicast and broadcast packets on only the active member, and dropping multicast and broadcast packets on all other members.

2. When Open vSwitch forwards a multicast or broadcast packet to a link in the SLB bond other than the active member, the remote switch will forward it to all of the other links in the SLB bond, including the active member. Without special handling, this would mean that Open vSwitch would forward a second copy of the packet to each switch port (other than the bond), including the port that originated the packet.

   Open vSwitch deals with this case by dropping packets received on any SLB bonded link that have a source MAC+VLAN that has been learned on any other port. (This means that SLB as implemented in Open vSwitch relies critically on MAC learning. Notably, SLB is incompatible with the "flood_vlans" feature.)

3. Suppose that a MAC+VLAN moves to an SLB bond from another port (e.g. when a VM is migrated from this hypervisor to a different one). Without additional special handling, Open vSwitch will not notice until the MAC learning entry expires, up to 60 seconds later as a consequence of rule #2.

   Open vSwitch avoids a 60-second delay by listening for gratuitous ARPs, which VMs commonly emit upon migration. As an exception to rule #2, a gratuitous ARP received on an SLB bond is not dropped and updates the MAC learning table in the usual way. (If a move does not trigger a gratuitous ARP, or if the gratuitous ARP is lost in the network, then a 60-second delay still occurs.)

4. Suppose that a MAC+VLAN moves from an SLB bond to another port (e.g. when a VM is migrated from a different hypervisor to this one), that the MAC+VLAN emits a gratuitous ARP, and that Open vSwitch forwards that gratuitous ARP to a link in the SLB bond other than the active member. The remote switch will forward the gratuitous ARP to all of the other links in the SLB bond, including the active member. Without additional special handling, this would mean that Open vSwitch would learn that the MAC+VLAN was located on the SLB bond, as a consequence of rule #3.

   Open vSwitch avoids this problem by "locking" the MAC learning table entry for a MAC+VLAN from which a gratuitous ARP was received from a non-SLB bond port. For 5 seconds, a locked MAC learning table entry will not be updated based on a gratuitous ARP received on a SLB bond.

5.1.8 Open vSwitch Networking Namespaces on Linux

The Open vSwitch has networking namespaces basic support on Linux. That allows ovs-vswitchd daemon to continue tracking status and statistics after moving a port to another networking namespace.

How It Works

The daemon ovs-vswitchd runs on what is called parent network namespace. It listens to netlink event messages from all networking namespaces (netns) with an identifier on the parent. Each netlink message contains the network namespace identifier (netnsid) as ancillary data which is used to match the event to the corresponding port.

The ovs-vswitchd uses an extended openvswitch kernel API[1] to get the current netnsid (stored in struct netdev_linux) and statistics from a specific port. The netnsid remains cached in userspace until a changing event is received, for example, when the port is moved to another network namespace.

Using another extended kernel API[2], the daemon gets port's information such as flags, MTU, MAC address and ifindex from a port already in another namespace.

The upstream kernel 4.15 includes the necessary changes for the basic support. In case of the running kernel doesn't provide the APIs, the daemon falls back to the previous behavior.

[1] Request cmd: OVS_VPORT_CMD_GET, attribute: OVS_VPORT_ATTR_NETNSID
[2] Request cmd: RTM_GETLINK passing IFLA_IF_NETNSID attribute.

Limitations

Currently it is only possible to retrieve the information listed in the above section. Most of other operations, for example querying MII or setting MTU, lacks the proper API in the kernel, so they remain unsupported.

In most use cases that needs to move ports to another networking namespaces should use veth pairs instead because it offers a cleaner and more robust solution with no noticeable performance penalty.

5.1.9 Scaling OVSDB Access With Relay

Open vSwitch 2.16 introduced support for OVSDB Relay mode with the goal to increase database scalability for a big deployments. Mainly, OVN (Open Virtual Network) Southbound Database deployments. This document describes the main concept and provides the configuration examples.

What is OVSDB Relay?

Relay is a database service model in which one `ovsdb-server` (`relay`) connects to another standalone or clustered database server (`relay source`) and maintains in-memory copy of its data, receiving all the updates via this OVSDB connection. Relay server handles all the read-only requests (monitors and transactions) on its own and forwards all the transactions that requires database modifications to the relay source.

Why is this needed?

Some OVN deployment could have hundreds or even thousands of nodes. On each of these nodes there is an ovn-controller, which is connected to the OVN_Southbound database that is served by a standalone or clustered OVSDB. Standalone database is handled by a single ovsdb-server process and clustered could consist of 3 to 5 ovsdb-server processes. For the clustered database, higher number of servers may significantly increase transaction latency due to necessity for these servers to reach consensus. So, in the end limited number of ovsdb-server processes serves ever growing number of clients and this leads to performance issues.

Read-only access could be scaled up with OVSDB replication on top of active-backup service model, but ovn-controller is a read-mostly client, not a read-only, i.e. it needs to execute write transactions from time to time. Here relay service model comes into play.

2-Tier Deployment

Solution for the scaling issue could look like a 2-tier deployment, where a set of relay servers is connected to the main database cluster (OVN_Southbound) and clients (ovn-conrtoller) connected to these relay servers:

```
                                 172.16.0.1
+--------------------+   +-----+ ovsdb-relay-1 +--+---+ client-1
|                    |   |                          |
|    Clustered       |   |                          +---+ client-2
|     Database       |   |                            ...
|                    |   |                          +---+ client-N
|    10.0.0.2        |   |
|  ovsdb-server-2    |   |       172.16.0.2
|   +        +       |   +-----+ ovsdb-relay-2 +--+---+ client-N+1
|   |        |       |   |                          |
|   |        +       +---+                          +---+ client-N+2
|   |   10.0.0.1     |   |                            ...
|   | ovsdb-server-1 |   |                          +---+ client-2N
|   |        +       |   |
|   |        |       |   |
|   +        +       |   +      ... ... ... ... ...
|  ovsdb-server-3    |   |
```

(continues on next page)

(continued from previous page)

```
|    10.0.0.3        |   |                      +---+ client-KN-1
|                    |   |       172.16.0.K     |
+--------------------+   +----+ ovsdb-relay-K +--+---+ client-KN
```

In practice, the picture might look a bit more complex, because all relay servers might connect to any member of a main cluster and clients might connect to any relay server of their choice.

Assuming that servers of a main cluster started like this:

```
$ ovsdb-server --remote=ptcp:6642:10.0.0.1 ovn-sb-1.db
```

The same for other two servers. In this case relay servers could be started like this:

```
$ REMOTES=tcp:10.0.0.1:6642,tcp:10.0.0.2:6642,tcp:10.0.0.3:6642
$ ovsdb-server --remote=ptcp:6642:172.16.0.1 relay:OVN_Southbound:$REMOTES
$ ...
$ ovsdb-server --remote=ptcp:6642:172.16.0.K relay:OVN_Southbound:$REMOTES
```

Open vSwitch 3.3 introduced support for configuration files via `--config-file` command line option. The configuration file for relay database servers in this case may look like this:

```
{
    "remotes": { "ptcp:6642:172.16.0.X": {} },
    "databases": {
        "OVN_Southbound": {
            "service-model": "relay",
            "source": {
                "$REMOTES": {}
            }
        }
    }
}
```

See `ovsdb-server(1)` and `Relay Service Model` in `ovsdb(7)` for more configuration options.

Every relay server could connect to any of the cluster members of their choice, fairness of load distribution is achieved by shuffling remotes.

For the actual clients, they could be configured to connect to any of the relay servers. For ovn-controllers the configuration could look like this:

```
$ REMOTES=tcp:172.16.0.1:6642,...,tcp:172.16.0.K:6642
$ ovs-vsctl set Open_vSwitch . external-ids:ovn-remote=$REMOTES
```

Setup like this allows the system to serve `K * N` clients while having only `K` actual connections on the main clustered database keeping it in a stable state.

It's also possible to create multi-tier deployments by connecting one set of relay servers to another (smaller) set of relay servers, or even create tree-like structures with the cost of increased latency for write transactions, because they will be forwarded multiple times.

5.1.10 OVSDB Replication Implementation

Given two Open vSwitch databases with the same schema, OVSDB replication keeps these databases in the same state, i.e. each of the databases have the same contents at any given time even if they are not running in the same host. This document elaborates on the implementation details to provide this functionality.

Terminology

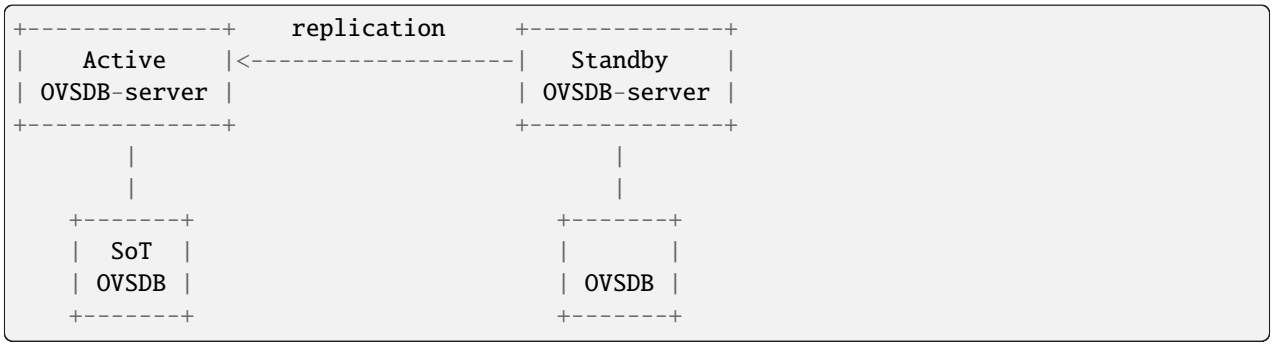
Source of truth database
: database whose content will be replicated to another database.

Active server
: ovsdb-server providing RPC interface to the source of truth database.

Standby server
: ovsdb-server providing RPC interface to the database that is not the source of truth.

Design

The overall design of replication consists of one ovsdb-server (active server) communicating the state of its databases to another ovsdb-server (standby server) so that the latter keep its own databases in that same state. To achieve this, the standby server acts as a client of the active server, in the sense that it sends a monitor request to keep up to date with the changes in the active server databases. When a notification from the active server arrives, the standby server executes the necessary set of operations so its databases reach the same state as the active server databases. Below is the design represented as a diagram.:



Setting Up The Replication

To initiate the replication process, the standby server must be executed indicating the location of the active server via the command line option `--sync-from=server`, where server can take any form described in the ovsdb-client manpage and it must specify an active connection type (tcp, unix, ssl). This option will cause the standby server to attempt to send a monitor request to the active server in every main loop iteration, until the active server responds.

When sending a monitor request the standby server is doing the following:

1. Erase the content of the databases for which it is providing a RPC interface.
2. Open the jsonrpc channel to communicate with the active server.
3. Fetch all the databases located in the active server.
4. For each database with the same schema in both the active and standby servers: construct and send a monitor request message specifying the tables that will be monitored (i.e all the tables on the database except the ones explicitly excluded [*]).
5. Set the standby database to the current state of the active database.

Once the monitor request message is sent, the standby server will continuously receive notifications of changes occurring to the tables specified in the request. The process of handling this notifications is detailed in the next section.

[*] A set of tables that will be excluded from replication can be configured via the command line option `--sync-exclude-tables=db:table[,db:table]...`, where db corresponds to the database where the table resides.

Replication Process

The replication process consists on handling the update notifications received in the standby server caused by the monitor request that was previously sent to the active server. In every loop iteration, the standby server attempts to receive a message from the active server which can be an error, an echo message (used to keep the connection alive) or an update notification. In case the message is a fatal error, the standby server will disconnect from the active without dropping the replicated data. If it is an echo message, the standby server will reply with an echo message as well. If the message is an update notification, the following process occurs:

1. Create a new transaction.
2. Get the <table-updates> object from the `params` member of the notification.
3. For each <table-update> in the <table-updates> object do:
    1. For each <row-update> in <table-update> check what kind of operation should be executed according to the following criteria about the presence of the object members:
        - If `old` member is not present, execute an insert operation using <row> from the `new` member.
        - If `old` member is present and `new` member is not present, execute a delete operation using <row> from the `old` member
        - If both `old` and `new` members are present, execute an update operation using <row> from the `new` member.
4. Commit the transaction.

   If an error occurs during the replication process, all replication is restarted by resending a new monitor request as described in the section "Setting up the replication".

Runtime Management Commands

Runtime management commands can be sent to a running standby server via ovs-appctl in order to configure the replication functionality. The available commands are the following.

ovsdb-server/set-remote-ovsdb-server {server}
: sets the name of the active server

ovsdb-server/get-remote-ovsdb-server
: gets the name of the active server

ovsdb-server/connect-remote-ovsdb-server
: causes the server to attempt to send a monitor request every main loop iteration

ovsdb-server/disconnect-remote-ovsdb-server
: closes the jsonrpc channel between the active server and frees the memory used for the replication configuration.

ovsdb-server/set-sync-exclude-tables {db:table,...}
: sets the tables list that will be excluded from being replicated

ovsdb-server/get-sync-excluded-tables
: gets the tables list that is currently excluded from replication

5.1.11 DPDK Support

5.1.12 Language Bindings

Bindings exist for Open vSwitch in a variety of languages.

Official Bindings

Python

The Python bindings are part of the Open vSwitch package. You can install the bindings using `pip`:

```
$ pip install ovs
```

The Python bindings include an optional flow parsing library. To install it's required dependencies, run:

```
$ pip install ovs[flow]
```

or install *python3-netaddr* and *python3-pyparsing*.

Third-Party Bindings

Lua

- LJIT2ovs: LuaJIT binding for Open vSwitch

Go

- go-odp: A Go library to control the Open vSwitch in-kernel datapath

5.1.13 Debugging with Record/Replay

The `ovs-replay` library provides a set of internal functions for recording certain events for later replay. This library is integrated into the `stream` and some other modules to record all incoming data across all streams (ssl, tcp, unixctl) of applications based on Open vSwitch libraries and play these streams later for debugging or performance testing purposes.

Support for this feature is currently integrated into the `ovsdb-server` and `ovsdb-client` applications. As a result, this allows to record lifecycle of the `ovsdb-server` process in large OVN deployments. Later, by using only the recorded data, the user can replay transactions and connections that occurred in a large deployment on their local PC. At the same time it is possible to tweak various log levels, run a process under debugger or tracer, measure performance with `perf`, and so on.

> **Note**
>
> The current version of record/replay engine does not work correctly with internal time-based events that lead to communications with other processes. For this reason it can not be used with clustered databases (RAFT implementation is heavily time dependent). In addition, recording automatically disables inactivity probes on JSONRPC connections and updates for the Manager status in a _Server database.

High-level feature overview was presented on Open vSwitch and OVN 2020 Fall Conference: Debugging OVSDB with stream record/replay

Recording ovsdb-server events

To start recording events for the `ovsdb-server` process, there is a special command line argument `--record`. Before starting the database server, make sure that you have a copy of a database file, so you can use it for replay later. Here are the general steps to take:

1. Create a directory where the replay files will be stored:

```
$ mkdir replay-dir
$ REPLAY_DIR=$(pwd)/replay-dir
```

2. Copy the current database file for later use:

```
$ cp my_database $REPLAY_DIR/
```

3. Run `ovsdb-server` with recording enabled:

```
$ ovsdb-server --record=$REPLAY_DIR <other arguments> my_database
```

4. Work with the database as usual.

5. Stop the `ovsdb-server` process at the end (it is important to send an `exit` command so that during replay the process will exit in the end too):

```
$ ovs-appctl -t ovsdb-server exit
```

After that `$REPLAY_DIR` should contain replay files with recorded data.

Replay of recorded session

During replay, the `ovsdb-server` will receive all the same connections, transactions and commands as it had at the time of recording, but it will not create any actual network/socket connections and will not communicate with any other process. Everything will be read from the replay files.

Since there is no need to wait for IPC, all events will be received one by one without any delays, so the application will process them as quickly as possible. This can be used as a performance test where the user can measure how quickly the `ovsdb-server` can handle some workload recorded in a real deployment.

The command line argument to start a replay session is `--replay`. The steps will look like this:

1. Restore the database file from a previous copy:

```
$ cp $REPLAY_DIR/my_database my_database
```

2. Start `ovsdb-server` with the same set of arguments as in the recording stage, except for `--record`:

```
$ ovsdb-server --replay=$REPLAY_DIR <other arguments> my_database
```

3. The process should exit in the end when the `exit` command is replayed.

On step 2 it is possible to add extra logging arguments to debug some recorded issue, or run the process under debugger. It's also possible to replay with a different version of `ovsdb-server` binary as long as this does not affect the data that goes in and out of the process, e.g. pure performance optimizations.

Limitations

The record/replay engine has the following limitations:

1. Record/Replay of clustered databases is not supported.

2. Inactivity probes on JSONRPC connections are suppressed.
3. Manager status updates suppressed in `ovsdb-server`.

To remove above limitations, it is necessary to implement correct handling of internally generated time-based events. (possibly by recording of time and subsequent time warping).

5.1.14 Testing

It is possible to test Open vSwitch using both tooling provided with Open vSwitch and using a variety of third party tooling.

Built-in Tooling

Open vSwitch provides a number of different test suites and other tooling for validating basic functionality of OVS. Before running any of the tests described here, you must bootstrap, configure and build Open vSwitch as described in *Open vSwitch on Linux, FreeBSD and NetBSD*. You do not need to install Open vSwitch or to build or load the kernel module to run these test suites. You do not need supervisor privilege to run these test suites.

Unit Tests

Open vSwitch includes a suite of self-tests. Before you submit patches upstream, we advise that you run the tests and ensure that they pass. If you add new features to Open vSwitch, then adding tests for those features will ensure your features don't break as developers modify other areas of Open vSwitch.

To run all the unit tests in Open vSwitch, one at a time, run:

```
$ make check
```

This takes under 5 minutes on a modern desktop system.

To run all the unit tests in Open vSwitch in parallel, run:

```
$ make check TESTSUITEFLAGS=-j8
```

You can run up to eight threads. This takes under a minute on a modern 4-core desktop system.

To see a list of all the available tests, run:

```
$ make check TESTSUITEFLAGS=--list
```

To run only a subset of tests, e.g. test 123 and tests 477 through 484, run:

```
$ make check TESTSUITEFLAGS='123 477-484'
```

Tests do not have inter-dependencies, so you may run any subset.

To run tests matching a keyword, e.g. `ovsdb`, run:

```
$ make check TESTSUITEFLAGS='-k ovsdb'
```

To see a complete list of test options, run:

```
$ make check TESTSUITEFLAGS=--help
```

The results of a testing run are reported in `tests/testsuite.log`. Report test failures as bugs and include the `testsuite.log` in your report.

> **Note**
>
> Sometimes a few tests may fail on some runs but not others. This is usually a bug in the testsuite, not a bug in Open vSwitch itself. If you find that a test fails intermittently, please report it, since the developers may not have noticed. You can make the testsuite automatically rerun tests that fail, by adding `RECHECK=yes` to the `make` command line, e.g.:
>
> ```
> $ make check TESTSUITEFLAGS=-j8 RECHECK=yes
> ```

Debugging unit tests

To initiate debugging from artifacts generated from *make check* run, set the `OVS_PAUSE_TEST` environment variable to 1. For example, to run test case 139 and pause on error:

```
$ OVS_PAUSE_TEST=1 make check TESTSUITEFLAGS='-v 139'
```

When error occurs, above command would display something like this:

```
Set environment variable to use various ovs utilities
export OVS_RUNDIR=<dir>/ovs/_build-gcc/tests/testsuite.dir/0139
Press ENTER to continue:
```

And from another window, one can execute ovs-xxx commands like:

```
export OVS_RUNDIR=/opt/vdasari/Developer/ovs/_build-gcc/tests/testsuite.dir/0139
$ ovs-ofctl dump-ports br0
.
.
```

Once done with investigation, press ENTER to perform cleanup operation.

Coverage

If the build was configured with `--enable-coverage` and the `lcov` utility is installed, you can run the testsuite and generate a code coverage report by using the `check-lcov` target:

```
$ make check-lcov
```

All the same options are available via TESTSUITEFLAGS. For example:

```
$ make check-lcov TESTSUITEFLAGS='-j8 -k ovsdb'
```

Valgrind

If you have `valgrind` installed, you can run the testsuite under valgrind by using the `check-valgrind` target:

```
$ make check-valgrind
```

When you do this, the "valgrind" results for test <N> are reported in files named `tests/testsuite.dir/<N>/valgrind.*`.

To test the testsuite of kernel datapath under valgrind, you can use the `check-kernel-valgrind` target and find the "valgrind" results under directory `tests/system-kmod-testsuite.dir/`.

All the same options are available via TESTSUITEFLAGS.

Hint

You may find that the valgrind results are easier to interpret if you put `-q` in `~/.valgrindrc`, since that reduces the amount of output.

OFTest

OFTest is an OpenFlow protocol testing suite. Open vSwitch includes a Makefile target to run OFTest with Open vSwitch in "dummy mode". In this mode of testing, no packets travel across physical or virtual networks. Instead, Unix domain sockets stand in as simulated networks. This simulation is imperfect, but it is much easier to set up, does not require extra physical or virtual hardware, and does not require supervisor privileges.

To run OFTest with Open vSwitch, you must obtain a copy of OFTest and install its prerequisites. You need a copy of OFTest that includes commit 406614846c5 (make ovs-dummy platform work again). This commit was merged into the OFTest repository on Feb 1, 2013, so any copy of OFTest more recent than that should work. Testing OVS in dummy mode does not require root privilege, so you may ignore that requirement.

Optionally, add the top-level OFTest directory (containing the `oft` program) to your `$PATH`. This slightly simplifies running OFTest later.

To run OFTest in dummy mode, run the following command from your Open vSwitch build directory:

```
$ make check-oftest OFT=<oft-binary>
```

where <`oft-binary`> is the absolute path to the `oft` program in OFTest. If you added "oft" to your $PATH, you may omit the OFT variable assignment

By default, `check-oftest` passes `oft` just enough options to enable dummy mode. You can use `OFTFLAGS` to pass additional options. For example, to run just the `basic.Echo` test instead of all tests (the default) and enable verbose logging, run:

```
$ make check-oftest OFT=<oft-binary> OFTFLAGS='--verbose -T basic.Echo'
```

If you use OFTest that does not include commit 4d1f3eb2c792 (oft: change default port to 6653), merged into the OFTest repository in October 2013, then you need to add an option to use the IETF-assigned controller port:

```
$ make check-oftest OFT=<oft-binary> OFTFLAGS='--port=6653'
```

Interpret OFTest results cautiously. Open vSwitch can fail a given test in OFTest for many reasons, including bugs in Open vSwitch, bugs in OFTest, bugs in the "dummy mode" integration, and differing interpretations of the OpenFlow standard and other standards.

Note

Open vSwitch has not been validated against OFTest. Report test failures that you believe to represent bugs in Open vSwitch. Include the precise versions of Open vSwitch and OFTest in your bug report, plus any other information needed to reproduce the problem.

Ryu

Ryu is an OpenFlow controller written in Python that includes an extensive OpenFlow testsuite. Open vSwitch includes a Makefile target to run Ryu in "dummy mode". See *OFTest* above for an explanation of dummy mode.

To run Ryu tests with Open vSwitch, first read and follow the instructions under **Testing** above. Second, obtain a copy of Ryu, install its prerequisites, and build it. You do not need to install Ryu (some of the tests do not get installed, so it

does not help).

To run Ryu tests, run the following command from your Open vSwitch build directory:

```
$ make check-ryu RYUDIR=<ryu-source-dir>
```

where `<ryu-source-dir>` is the absolute path to the root of the Ryu source distribution. The default `<ryu-source-dir>` is `$srcdir/../ryu` where `$srcdir` is your Open vSwitch source directory. If this is correct, omit `RYUDIR`

> **Note**
>
> Open vSwitch has not been validated against Ryu. Report test failures that you believe to represent bugs in Open vSwitch. Include the precise versions of Open vSwitch and Ryu in your bug report, plus any other information needed to reproduce the problem.

Datapath testing

Open vSwitch includes a suite of tests specifically for datapath functionality, which can be run against the userspace or kernel datapaths. If you are developing datapath features, it is recommended that you use these tests and build upon them to verify your implementation.

The datapath tests make some assumptions about the environment. They must be run under root privileges on a Linux system with support for network namespaces.

Make sure, no other Open vSwitch instance is running on the test suite. These tests may take several minutes to complete, and cannot be run in parallel.

Userspace datapath

To invoke the datapath testsuite with the userspace datapath, run:

```
$ make check-system-userspace
```

The results of the testsuite are in `tests/system-userspace-testsuite.dir`.

All the features documented under *Unit Tests* are available for the userspace datapath testsuite.

Userspace datapath with DPDK

To test *Open vSwitch with DPDK* (i.e., the build was configured with `--with-dpdk`, the DPDK is installed), run the testsuite and generate a report by using the `check-dpdk` target:

```
# make check-dpdk
```

or if you are not a root, but a sudo user:

```
$ sudo -E make check-dpdk
```

To see a list of all the available tests, run:

```
# make check-dpdk TESTSUITEFLAGS=--list
```

These tests support a DPDK supported NIC. The tests operate on a wider set of environments, for instance, when a virtual port is used. Moreover you need to have root privileges to load the required modules and to bind a PCI device to the DPDK-compatible driver.

The phy test will skip if no suitable PCI device is found. It is possible to select which PCI device is used for this test by setting the DPDK_PCI_ADDR environment variable, which is especially useful when testing with a mlx5 device:

```
# DPDK_PCI_ADDR=0000:82:00.0 make check-dpdk
```

All tests are skipped if no hugepages are configured. User must look into the DPDK manual to figure out how to Configure hugepages.

All the features documented under *Unit Tests* are available for the DPDK testsuite.

Userspace datapath with DPDK offload

To invoke the userspace datapath tests with DPDK and its rte_flow offload, the same prerequisites apply as above. In addition, six Virtual Function (VF) interfaces must be preconfigured on a single Physical Function (PF) that supports rte_flow hardware offload.

This is an example on how to set this up for an NVIDIA blade on port `ens2f0np0`:

```
OVS_PF_PCI=$(basename $(readlink /sys/class/net/ens2f0np0/device))
echo 0 > /sys/bus/pci/devices/$OVS_PF_PCI/sriov_numvfs
devlink dev eswitch set pci/$OVS_PF_PCI mode switchdev
echo 6 > /sys/bus/pci/devices/$OVS_PF_PCI/sriov_numvfs
```

This PF's PCI ID needs to be passed with the OVS_PF_PCI variable. To invoke the DPDK offloads testsuite with the userspace datapath, run:

```
make check-dpdk-offloads OVS_PF_PCI=0000:17:00.0
```

Note

This has only been tested on NVIDIA blades due to the limited availability of other blades that support rte_flow.

Kernel datapath

Make targets are also provided for testing the Linux kernel module. Note that these tests operate by inserting modules into the running Linux kernel, so if the tests are able to trigger a bug in the OVS kernel module or in the upstream kernel then the kernel may panic.

To run the testsuite against the kernel module which is currently installed on your system, run:

```
$ make check-kernel
```

All the features documented under *Unit Tests* are available for the kernel datapath testsuite.

Note

Many of the kernel tests are dependent on the utilities present in the iproute2 package, especially the 'ip' command. If there are many otherwise unexplained errors it may be necessary to update the iproute2 package utilities on the system. It is beyond the scope of this documentation to explain all that is necessary to build and install an updated iproute2 utilities package. The package is available from the Linux kernel organization open source git repositories.

https://git.kernel.org/pub/scm/network/iproute2/iproute2.git

It is also possible to run retis capture along with the *check-kernel* and *check-offloads* tests by setting *OVS_TEST_WITH_RETIS* environment variable to 'yes'. This can be useful for debugging the test cases. For example, the following command can be used to run the test 167 under *retis*:

```
$ make check-kernel OVS_TEST_WITH_RETIS=yes TESTSUITEFLAGS='167 -d'
```

After the test is completed, the following data will be available in the test directory:

- *retis.err* - standard error stream of the *retis collect*.
- *retis.log* - standard output of the *retis collect*, contains all captured events in the order they appeared.
- *retis.data* - raw events collected by retis, *retis sort* or other commands can be used on this file for further analysis.
- *retis.sorted* - text file containing the output of *retis sort* executed on the *retis.data*, for convenience.

Requires retis version 1.5 or newer and enabling support for *User Statically-Defined Tracing (USDT) probes*.

Static Code Analysis

Static Analysis is a method of debugging Software by examining code rather than actually executing it. This can be done through 'scan-build' commandline utility which internally uses clang (or) gcc to compile the code and also invokes a static analyzer to do the code analysis. At the end of the build, the reports are aggregated in to a common folder and can later be analyzed using 'scan-view'.

Open vSwitch includes a Makefile target to trigger static code analysis:

```
$ ./boot.sh
$ ./configure CC=clang  # clang
# or
$ ./configure CC=gcc CFLAGS="-std=gnu99"  # gcc
$ make clang-analyze
```

You should invoke scan-view to view analysis results. The last line of output from `clang-analyze` will list the command (containing results directory) that you should invoke to view the results on a browser.

ViNePerf

The ViNePerf project, formerly known as VswitchPerf or vsperf, aims to develop a vSwitch test framework that can be used to validate the suitability of different vSwitch implementations in a telco deployment environment. More information can be found on the Anuket project wiki.

5.1.15 Tracing packets inside Open vSwitch

Open vSwitch (OVS) is a programmable software switch that can execute actions at per packet level. This document explains how to use the tracing tool to know what is happening with packets as they go through the data plane processing.

The ovs-vswitchd(8) manpage describes basic usage of the ofproto/trace command used for tracing in Open vSwitch.

Packet Tracing

In order to understand the tool, let's use the following flows as an example:

```
table=3,ip,tcp,tcp_dst=80,action=output:2
table=2,ip,tcp,tcp_dst=22,action=output:1
table=0,in_port=3,ip,nw_src=192.0.2.0/24,action=resubmit(,2)
table=0,in_port=3,ip,nw_src=198.51.100.0/24,action=resubmit(,3)
```

> **Note**
>
> If you can't use a "real" OVS setup you can use `ovs-sandbox`, as described in *Open vSwitch Advanced Features*, which also provides additional tracing examples.

The first line adds a rule in table 3 matching on TCP/IP packet with destination port 80 (HTTP). If a packet matches, the action is to output the packet on OpenFlow port 2.

The second line is similar but matches on destination port 22. If a packet matches, the action is to output the packet on OpenFlow port 1.

The next two lines matches on source IP addresses. If there is a match, the packet is submitted to table indicated as parameter to the resubmit() action.

Now let's see if a packet from IP address 192.0.2.1 and destination port 22 would really go to OpenFlow port 1:

```
$ ovs-appctl ofproto/trace br0 in_port=3,tcp,nw_src=192.0.2.2,tcp_dst=22
Flow: tcp,in_port=3,vlan_tci=0x0000,dl_src=00:00:00:00:00:00,dl_dst=00:00:00:00:00:00,nw_
↪src=192.0.2.2,nw_dst=0.0.0.0,nw_tos=0,nw_ecn=0,nw_ttl=0,tp_src=0,tp_dst=22,tcp_flags=0

bridge("br0")
-------------
 0. ip,in_port=3,nw_src=192.0.2.0/24, priority 32768
    resubmit(,2)
 2. tcp,tp_dst=22, priority 32768
    output:1

Final flow: unchanged
Megaflow: recirc_id=0,tcp,in_port=3,nw_src=192.0.2.0/24,nw_frag=no,tp_dst=22
Datapath actions: 1
```

The first line is the trace command. The br0 is the bridge where the packet is going through. The next arguments describe the packet itself. For instance, the nw_src matches with the IP source address. All the packet fields are well documented in the ovs-fields(7) man-page.

The second line shows the flow extracted from the packet described in the command line. Unspecified packet fields are zeroed.

The second group of lines shows the packet's trip through bridge br0. We see, in table 0, the OpenFlow flow that the fields matched, along with its priority, followed by its actions, one per line. In this case, we see that this packet matches the flow that resubmit those packets to table 2. The "resubmit" causes a second lookup in OpenFlow table 2, described by the block of text that starts with "2.". In the second lookup we see that this packet matches the rule that outputs those packets to OpenFlow port #1.

In summary, it is possible to follow the flow entries and actions until the final decision is made. At the end, the trace tool shows the Megaflow which matches on all relevant fields followed by the data path actions.

Let's see what happens with the same packet but with another TCP destination port:

```
$ ovs-appctl ofproto/trace br0 in_port=3,tcp,nw_src=192.0.2.2,tcp_dst=80
Flow: tcp,in_port=3,vlan_tci=0x0000,dl_src=00:00:00:00:00:00,dl_dst=00:00:00:00:00:00,nw_
↪src=192.0.2.2,nw_dst=0.0.0.0,nw_tos=0,nw_ecn=0,nw_ttl=0,tp_src=0,tp_dst=80,tcp_flags=0

bridge("br0")
-------------
 0. ip,in_port=3,nw_src=192.0.2.0/24, priority 32768
```

(continues on next page)

(continued from previous page)

```
    resubmit(,2)
 2. No match.
    drop

Final flow: unchanged
Megaflow: recirc_id=0,tcp,in_port=3,nw_src=192.0.2.0/24,nw_frag=no,tp_dst=0x40/0xffc0
Datapath actions: drop
```

In the second group of lines, in table 0, you can see that the packet matches with the rule because of the source IP address, so it is resubmitted to the table 2 as before. However, it doesn't match any rule there. When the packet doesn't match any rule in the flow tables, it is called a table miss. The virtual switch table miss behavior can be configured and it depends on the OpenFlow version being used. In this example the default action was to drop the packet.

Credits

This document is heavily based on content from Flavio Bruno Leitner at Red Hat:

- https://developers.redhat.com/blog/2016/10/12/tracing-packets-inside-open-vswitch/

5.1.16 Userspace Datapath - TSO

Note: This feature is considered experimental.

TCP Segmentation Offload (TSO) enables a network stack to delegate segmentation of an oversized TCP segment to the underlying physical NIC. Offload of frame segmentation achieves computational savings in the core, freeing up CPU cycles for more useful work.

A common use case for TSO is when using virtualization, where traffic that's coming in from a VM can offload the TCP segmentation, thus avoiding the fragmentation in software. Additionally, if the traffic is headed to a VM within the same host further optimization can be expected. As the traffic never leaves the machine, no MTU needs to be accounted for, and thus no segmentation and checksum calculations are required, which saves yet more cycles. Only when the traffic actually leaves the host the segmentation needs to happen, in which case it will be performed by the egress NIC. Consult your controller's datasheet for compatibility. Secondly, the NIC must have an associated DPDK Poll Mode Driver (PMD) which supports *TSO*. For a list of features per PMD, refer to the DPDK documentation.

Enabling TSO

The TSO support may be enabled via a global config value `userspace-tso-enable`. Setting this to `true` enables TSO support for all ports.:

```
$ ovs-vsctl set Open_vSwitch . other_config:userspace-tso-enable=true
```

The default value is `false`.

Changing `userspace-tso-enable` requires restarting the daemon.

When using *vHost User ports*, TSO may be enabled as follows.

TSO is enabled in OvS by the DPDK vHost User backend; when a new guest connection is established, *TSO* is thus advertised to the guest as an available feature:

1. QEMU Command Line Parameter:

```
$ sudo $QEMU_DIR/x86_64-softmmu/qemu-system-x86_64 \
...
-device virtio-net-pci,mac=00:00:00:00:00:01,netdev=mynet1,\
```

(continues on next page)

(continued from previous page)

```
csum=on,guest_csum=on,guest_tso4=on,guest_tso6=on\
...
```

2. Ethtool. Assuming that the guest's OS also supports *TSO*, ethtool can be used to enable same:

```
$ ethtool -K eth0 sg on     # scatter-gather is a prerequisite for TSO
$ ethtool -K eth0 tso on
$ ethtool -k eth0
```

Note: Enabling this feature impacts the virtio features exposed by the DPDK vHost User backend to a guest. If a guest was already connected to OvS before enabling TSO and restarting OvS, this guest ports won't have TSO available:

```
$ ovs-vsctl get interface vhost0 status:tx_tcp_seg_offload
"false"
```

To help diagnose the issue, those ports have some additional information in their status field in ovsdb:

```
$ ovs-vsctl get interface vhost0 status:userspace-tso
disabled
```

To restore TSO for this guest ports, this guest QEMU process must be stopped, then started again. OvS will then report:

```
$ ovs-vsctl get interface vhost0 status:tx_tcp_seg_offload
"true"

$ ovs-vsctl get interface vhost0 status:userspace-tso
ovs-vsctl: no key "userspace-tso" in Interface record "vhost0" column status
```

Software segmentation fallback

In case a *TSO* packet is transmitted over a port that does not support segmentation, OvS userspace datapath segments the packet in software (see lib/dp-packet-gso.c).

This case is common for port not supporting TSO at all. A coverage counter exists to reflect when the software fallback helper is called.:

```
$ ovs-appctl coverage/read-counter netdev_soft_seg_good
178760
```

Another possibility is when the *TSO* packet involves a header encapsulation requested by OvS. When the tunnel encapsulation is UDP based, and checksum on the tunnel header is requested but the transmitting port does not support this combination, OvS does some partial segmentation based on the segmentation size coming from the receiving port. A coverage counter exists to reflect when this optimisation is called.:

```
$ ovs-appctl coverage/read-counter netdev_partial_seg_good
1345
```

Limitations

The current OvS userspace *TSO* implementation supports flat, VLAN networks, and some tunneled connections. Currently only VxLAN, Geneve and GRE tunnels are supported.

The NIC driver must support and advertise checksum offload for TCP and UDP. However, SCTP is not mandatory because very few drivers advertised support and it wasn't a widely used protocol at the moment this feature was intro-

duced in Open vSwitch. Currently, if the NIC supports that, then the feature is enabled, otherwise TSO can still be enabled but SCTP packets sent to the NIC will be dropped.

There is a limited software implementation of TSO when tunnels are used which only supports VxLAN, Geneve, and GRE. When these tunnels are used with TSO, not all ports attached to the datapath need to support hardware TSO. Guests using vhost-user in client mode will receive TSO packet regardless of TSO being enabled or disabled within the guest.

All kernel devices that use the raw socket interface (veth, for example) require the kernel commit 9d2f67e43b73 ("net/packet: fix packet drop as of virtio gso") in order to work properly. This commit was merged in upstream kernel 4.19-rc7, so make sure your kernel is either newer or contains the backport.

Performance Tuning

iperf is often used to test TSO performance. Care needs to be taken when configuring the environment in which the iperf server process is being run. Since the iperf server uses the NIC's kernel driver, IRQs will be generated. By default with some NICs eg. i40e, the IRQs will land on the same core as that which is being used by the server process, provided the number of NIC queues is greater or equal to that lcoreid. This causes contention between the iperf server process and the IRQs. For optimal performance, it is suggested to pin the IRQs to their own core. To change the affinity associated with a given IRQ number, you can 'echo' the desired coremask to the file /proc/irq/<number>/smp_affinity For more on SMP affinity, refer to the Linux kernel documentation.

5.1.17 C IDL Compound Indexes

Introduction

This document describes the design and usage of the C IDL Compound Indexes feature, which allows OVSDB client applications to efficiently search table contents using arbitrary sets of column values in a generic way.

This feature is implemented entirely in the client IDL, requiring no changes to the OVSDB Server, OVSDB Protocol (OVSDB RFC (RFC 7047)) or additional interaction with the OVSDB server.

Please note that in this document, the term "index" refers to the common database term defined as "a data structure that facilitates data retrieval". Unless stated otherwise, the definition for index from the OVSDB RFC (RFC 7047) is not used.

Typical Use Cases

Fast lookups

Depending on the topology, the route table of a network device could manage thousands of routes. Commands such as "show ip route <*specific route*>" would need to do a sequential lookup of the routing table to find the specific route. With an index created, the lookup time could be faster.

This same scenario could be applied to other features such as Access List rules and even interfaces lists.

Lexicographic order

There are a number of cases in which retrieving data in a particular lexicographic order is needed. For example, SNMP. When an administrator or even a NMS would like to retrieve data from a specific device, it's possible that they will request data from full tables instead of just specific values. Also, they would like to have this information displayed in lexicographic order. This operation could be done by the SNMP daemon or by the CLI, but it would be better if the database could provide the data ready for consumption. Also, duplicate efforts by different processes will be avoided. Another use case for requesting data in lexicographic order is for user interfaces (web or CLI) where it would be better and quicker if the DB sends the data sorted instead of letting each process to sort the data by itself.

Implementation Design

This feature maintains a collection of indexes per table. The application can create any number of indexes per table.

An index can be defined over any number of columns, and supports the following options:

- Add a column with type string, boolean, uuid, integer or real (using default comparators).
- Select ordering direction of a column (ascending or descending, must be selected when creating the index).
- Use a custom ordering comparator (eg: treat a string column like a IP, or sort by the value of the "config" key in a map column).

Indexes can be searched for matches based on the key. They can also be iterated across a range of keys or in full.

For lookups, the user needs to provide a key to be used for locating the specific rows that meet his criteria. This key could be an IP address, a MAC address, an ACL rule, etc. If several rows match the query then the user can easily iterate over all of the matches.

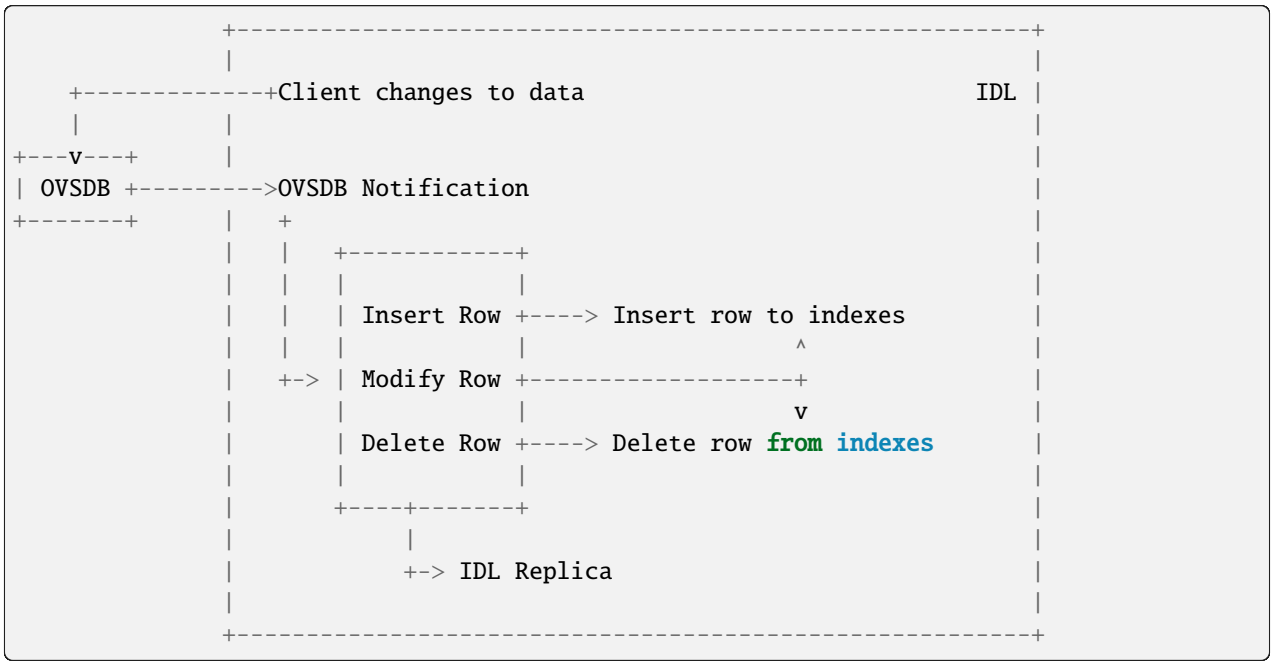
For accessing data in lexicographic order, the user can use the ranged iterators, which use "from" and "to" values to define a range.

The indexes maintain a pointer to the row in the local replica, avoiding the need to make additional copies of the data and thereby minimizing any additional memory and CPU overhead for their maintenance. It is intended that creating and maintaining indexes should be very cheap.

Another potential issue is the time needed to create the data structure and the time needed to add/remove elements. The indexes are always synchronized with the replica. For this reason is VERY IMPORTANT that the comparison functions (built-in and user provided) are FAST.

Skiplists are used as the primary data structure for the implementation of indexes. Indexes therefore have an expected `O(log(n))` cost when inserting, deleting or modifying a row, `O(log(n))` when retrieving a row by key, and O(1) when retrieving the first or next row.

Indexes are maintained incrementally in the replica as notifications of database changes are received from the OVSDB server, as shown in the following diagram.



C IDL API

Index Creation

Each index must be created with the function `ovsdb_idl_index_create()` or one of the simpler convenience functions `ovsdb_idl_index_create1()` or `ovsdb_idl_index_create2()`. All indexes must be created before the first call to ovsdb_idl_run().

Index Creation Example

```
/* Define a custom comparator for the column "stringField" in table
 * "Test". (Note that custom comparison functions are not often
 * necessary.)
 */
int stringField_comparator(const void *a, const void *b)
{
    struct ovsrec_test *AAA, *BBB;
    AAA = (struct ovsrec_test *)a;
    BBB = (struct ovsrec_test *)b;
    return strcmp(AAA->stringField, BBB->stringField);
}

void init_idl(struct ovsdb_idl **, char *remote)
{
    /* Add the columns to the IDL */
    *idl = ovsdb_idl_create(remote, &ovsrec_idl_class, false, true);
    ovsdb_idl_add_table(*idl, &ovsrec_table_test);
    ovsdb_idl_add_column(*idl, &ovsrec_test_col_stringField);
    ovsdb_idl_add_column(*idl, &ovsrec_test_col_numericField);
    ovsdb_idl_add_column(*idl, &ovsrec_test_col_enumField);
    ovsdb_idl_add_column(*idl, &ovsrec_test_col_boolField);

    struct ovsdb_idl_index_column columns[] = {
        { .column = &ovsrec_test_col_stringField,
          .comparer = stringField_comparator },
        { .column = &ovsrec_test_col_numericField,
          .order = OVSDB_INDEX_DESC },
    };
    struct ovsdb_idl_index *index = ovsdb_idl_create_index(
        *idl, columns, ARRAY_SIZE(columns));
    ...
}
```

Index Usage

Iterators

The recommended way to do queries is using a “ranged foreach”, an “equal foreach” or a “full foreach” over an index. The mechanism works as follows:

1. Create index row objects with index columns set to desired search key values (one is needed for equality iterators, two for range iterators, a search key is not needed for the full index iterator).
2. Pass the index, an iteration variable, and the index row object to the iterator.
3. Use the values within iterator loop.

The library implements three different iterators: a range iterator, an equality iterator and a full index iterator. The range iterator receives two values and iterates over all rows with values that are within that range (inclusive of the two values defining the range). The equality iterator iterates over all rows that exactly match the value passed. The full index iterator iterates over all rows in the index, in an order determined by the comparison function and configured direction (ascending or descending).

Note that indexes are *sorted by the "concatenation" of the values in all indexed columns*, so the ranged iterator returns all the values between "from.col1 from.col2 … from.coln" and "to.col1 to.col2 … to.coln", *NOT the rows with a value in column 1 between from.col1 and to.col1, and so on*.

The iterators are macros specific to each table. An example of the use of these iterators follows:

```
/*
 * Equality iterator; iterates over all the records equal to "value".
 */
struct ovsrec_test *target = ovsrec_test_index_init_row(index);
ovsrec_test_index_set_stringField(target, "hello world");
struct ovsrec_test *record;
OVSREC_TEST_FOR_EACH_EQUAL (record, target, index) {
    /* Can return zero, one or more records */
    assert(strcmp(record->stringField, "hello world") == 0);
    printf("Found one record with %s", record->stringField);
}
ovsrec_test_index_destroy_row(target);

/*
 * Range iterator; iterates over all records between two values
 * (inclusive).
 */
struct ovsrec_test *from = ovsrec_test_index_init_row(index);
struct ovsrec_test *to = ovsrec_test_index_init_row(index);

ovsrec_test_index_set_stringField(from, "aaa");
ovsrec_test_index_set_stringField(to, "mmm");
OVSREC_TEST_FOR_EACH_RANGE (record, from, to, index) {
    /* Can return zero, one or more records */
    assert(strcmp("aaa", record->stringField) <= 0);
    assert(strcmp(record->stringField, "mmm") <= 0);
    printf("Found one record with %s", record->stringField);
}
ovsrec_test_index_destroy_row(from);
ovsrec_test_index_destroy_row(to);

/*
 * Index iterator; iterates over all nodes in the index, in order
 * determined by comparison function and configured order (ascending
 * or descending).
 */
OVSREC_TEST_FOR_EACH_BYINDEX (record, index) {
    /* Can return zero, one or more records */
    printf("Found one record with %s", record->stringField);
}
```

General Index Access

While the currently defined iterators are suitable for many use cases, it is also possible to create custom iterators using the more general API on which the existing iterators have been built. See `ovsdb-idl.h` for the details.

5.1.18 Open vSwitch Extensions

Introduction

OpenFlow allows vendor extensions to the protocol. OVS implements many of its own extensions. These days, we typically refer to these as "Open vSwitch extensions" to OpenFlow. You might also see them called "Nicira extensions" after the company that initiated the Open vSwitch project. These extensions have been used to add additional functionality for the desired features not present in the standard OpenFlow protocol.

OpenFlow 1.0 refers to extensions as "vendor extensions", whereas OpenFlow 1.1 and later call them "experimenter extensions". These are different names for the same concept and we use them interchangeably.

OVS vendor extension messages in OpenFlow and OVS

Vendor/experimenter request and replies

OpenFlow supports vendor extensions for basic requests and replies. In OpenFlow 1.0, such messages carry `OFPT_VENDOR` in the `struct ofp_header` message type field, and in later versions `OFPT_EXPERIMENTER`. After the header of this message, there is a vendor id field which identifies the vendor. Everything after the vendor ID is vendor specific, but it would typically include a subtype field to identify a particular kind of vendor-specific message. Vendor ids are defined in `ovs/include/openflow/openflow-common.h`.

To see a list of all the vendor message subtypes that OVS understands, we can refer to `ovs/lib/ofp-msgs.h`. We can see the instances of `enum ofpraw` which has a comment containing the keyword `NXT`, e.g. `OFPRAW_NXT_FLOW_MOD` below:

```
/* NXT 1.0+ (13): struct nx_flow_mod, uint8_t[8][]. */
OFPRAW_NXT_FLOW_MOD,
```

which may be interpreted as follows:

NXT
: stands for Nicira extension message.

nx_flow_mod
: data that follow the OpenFlow header.

uint8_t[8][]
: multiple of 8 data.

13
: the subtype for the Flow Mod message when it is sent as a Open vSwitch extension message

OFPRAW_NXT_FLOW_MOD
: the Open vSwitch Flow Mod extension message.

For reference, the vendor message header is defined as `struct ofp_vendor_header` in `ovs/lib/ofp-msgs.c`.

The general structure of a message with a vendor message type is:

ofp_header(msg_type=VENDOR/EXPERIMENTER) / vendor id / vendor subtype / vendor defined additional data (e.g. nx_flow_mod structure for OFPRAW_NXT_FLOW_MOD message)

Multipart vendor requests and replies

OpenFlow supports "statistics" or "multipart" messages that consist of a sequence of shorter messages with associated content. In OpenFlow 1.0 through 1.2, these are `OFPT_STATS_REQUEST` requests and `OFPT_STATS_REPLY` replies, and in OpenFlow 1.3 and later, they are `OFPT_MULTIPART_REQUEST``and ``OFPT_MULTIPART_REPLY`.

A multipart message carries its own embedded type that denotes the kind of multipart data. Multipart vendor requests and replies use type `OFPT_VENDOR` in OpenFlow 1.0, `OFPST_EXPERIMENTER` in OpenFlow 1.1 and 1.2, and `OFPMP_EXPERIMENTER` in OpenFlow 1.3 and later.

Again if we refer to `ovs/lib/ofp-msgs.h`, we see the following lines:

```
/* NXST 1.0 (2): uint8_t[8][]. */
OFPRAW_NXST_FLOW_MONITOR_REQUEST,
```

NXST
: stands for Nicira extension statistics or multipart message.

uint8_t[8][]
: multiple of 8 data.

2
: the subtype for the Flow Monitor Request message when it is sent as a Flow Monitor Request message with extension vendor id.

OFPRAW_NXST_FLOW_MONITOR_REQUEST
: the OpenFlow Flow Monitor extension message.

For reference, the vendor extension stats message header is defined as `struct ofp11_vendor_stats_msg` in `ovs/lib/ofp-msgs.c`.

The general structure of a multipart/stats message with vendor type is:

ofp_header(msg_type=STATS/MULTIPART) / stats_msg(type=VENDOR/EXPERIMENTER) /
: vendor-id / subtype / vendor defined additional data

Extended Match

OpenFlow 1.0 uses a fixed size flow match structure (struct ofp_match) to define the fields to match in a packet. This is limiting and not extensible. To make the match structure extensible, OVS added as an extension `nx_match` structure, called NXM, as a series of TLV (type-length-value) entries or `nxm_entry`. OpenFlow 1.2 standardized NXM as OXM, with extensions of its own. OVS supports standard and extension OXM and NXM TLVs.

For a detailed description of NXM and OXM, please see the OVS fields documentation at https://www.openvswitch.org/support/dist-docs/ovs-fields.7.pdf.

Error Message Extension

In OpenFlow version 1.0 and 1.1, there is no provision to generate vendor specific error codes and does not even provide `generic` error codes that can apply to problems not anticipated by the OpenFlow specification authors. OVS added a generic "error vendor extension" which uses `NXET_VENDOR` as type and `NXVC_VENDOR_ERROR` as code, followed by `struct nx_vendor_error` with vendor-specific details, followed by at least 64 bytes of the failed request.

OpenFlow version 1.2+ added a `OFPET_EXPERIMENTER` error type to support vendor specific error codes.

Source files related to Open vSwitch extensions

```
ovs/include/openflow/nicira-ext.h
ovs/lib/ofp-msgs.inc
ovs/include/openvswitch/ofp-msgs.h
ovs/lib/ofp-msgs.c
```

5.1.19 Userspace Datapath - Checksum Offloading

This document explains the internals of Open vSwitch support for checksum offloading in the userspace datapath.

Design

Open vSwitch strives to forward packets as they arrive regardless of whether the checksum is correct or not. OVS is not responsible for fixing external checksum issues.

The interface (internally referred to as a netdev) can set flags indicating whether each packet's checksum is good or bad upon receipt. If this flag is not set, OVS will consider the validity of the packet's checksum to be unknown.

OVS will not re-calculate or update the packet's checksum if the checksum is already known to be correct, known to be explicitly incorrect, or destined for an egress interface that will recalculate the checksum anyways.

If OVS does invalidate the checksum, and the packet ingresses the datapath with a checksum that is not known to be incorrect, OVS postpones checksum updates until the packet egresses the datapath. This recalculation can either be performed by OVS or, be offloaded onto the NIC if the egress NIC supports checksum offloading.

When a packet egress the datapath, the packet flags and the egress interface flags are verified to make sure all required offload features to send out the packet are available on the egress interface. If not, the data path will fall back to equivalent software implementation.

Interface (a.k.a. Netdev)

When the interface initiates, it should set the flags to tell the datapath which offload features are supported. For example, if the driver supports IP checksum offloading, then `netdev->ol_flags` should set the flag `NETDEV_TX_OFFLOAD_IPV4_CKSUM`.

Rules

1) OVS should strive to forward all packets regardless of checksum.
2) OVS must not correct a known bad packet checksum.
3) Packet with only flag `DP_PACKET_OL_RX_IP_CKSUM_GOOD` means that the IP checksum is present in the packet and it is good.
4) Packet with only flag `DP_PACKET_OL_RX_IP_CKSUM_BAD` means that the IP checksum is present in the packet and it is bad. Extra care should be taken to not fix the packet during data path processing.
5) Packet with both `DP_PACKET_OL_RX_IP_CKSUM_GOOD` and `DP_PACKET_OL_RX_IP_CKSUM_BAD` means that the IP header is valid, but the checksum may not be set in the packet data. This is basically encountered with some virtual drivers, or after OVS modified the IPv4 header content of a not bad ("bad" as defined in 4)) packet.
6) Packet with neither `DP_PACKET_OL_RX_IP_CKSUM_GOOD` nor `DP_PACKET_OL_RX_IP_CKSUM_BAD` means that the IP header status is unknown. It may be a IPv4 packet, or not. It may have a valid IPv4 cheksum, or not. This situation is encountered with virtual drivers that provide no information about the IP header, and for IPv6 packets.

5.1.20 Userspace Tx packet steering

The userspace datapath supports two transmit packet steering modes.

Thread mode

This mode is automatically selected when the port's `tx-steering` option is set to `thread` or unset.

Thread mode enables static (1:1) thread-to-txq mapping when the number of Tx queues is greater than number of PMD threads, and dynamic (N:1) mapping if equal or lower. In this mode a single thread can not use more than 1 transmit queue of a given port.

This is the recommended mode for performance reasons if the number of Tx queues is greater than the number of PMD threads, because the Tx lock is not acquired.

If the number of Tx queues is greater than the number of threads (including the main thread), the remaining Tx queues will not be used.

This mode is enabled by default.

Hash mode

Hash-based Tx packet steering mode distributes the packets on all the port's transmit queues, whatever the number of PMD threads. The queue selection is based on the 5-tuples hash to build the flows batches, the selected queue being the modulo between the hash and the number of Tx queues of the port.

Hash mode may be used for example with vhost-user ports, when the number of vCPUs and queues of the guest are greater than the number of PMD threads. Without hash mode, the Tx queues used would be limited to the number of threads.

Hash-based Tx packet steering may have an impact on the performance, given the Tx lock acquisition is always required and a second level of batching is performed.

Usage

To enable hash mode:

```
$ ovs-vsctl set Interface <iface> other_config:tx-steering=hash
```

To disable hash mode:

```
$ ovs-vsctl set Interface <iface> other_config:tx-steering=thread
```

5.1.21 User Statically-Defined Tracing (USDT) probes

Sometimes it's desired to troubleshoot one of OVS's components in the field. One of the techniques used is to add dynamic tracepoints, for example using perf. However, the desired dynamic tracepoint and/or the desired variable, might not be available due to compiler optimizations.

In this case, a well-thought-off, static tracepoint could be permanently added, so it's always available. For OVS we use the DTrace probe macro's, which have little to no overhead when disabled. Various tools exist to enable them. See some examples below.

Compiling with USDT probes enabled

Since USDT probes are compiled out by default, a compile-time option is available to include them. To add the probes to the generated code, use the following configure option

```
$ ./configure --enable-usdt-probes
```

The following line should be seen in the configure output when the above option is used

```
checking whether USDT probes are enabled... yes
```

As USDT probes internally use the `DTRACE_PROBExx` macros, which are part of the SystemTap framework, you need to install the appropriate package for your Linux distribution. For example, on Fedora, you need to install the `systemtap-sdt-devel` package.

Listing available probes

There are various ways to display USDT probes available in a userspace application. Here we show three examples. All assuming ovs-vswitchd is in the search path with USDT probes enabled:

You can use the **perf** tool as follows

```
$ perf buildid-cache --add $(which ovs-vswitchd)
$ perf list | grep sdt_
  sdt_main:poll_block                                [SDT event]
  sdt_main:run_start                                 [SDT event]
```

You can use the bpftrace tool

```
# bpftrace -l "usdt:$(which ovs-vswitchd):*"
usdt:/usr/sbin/ovs-vswitchd:main:poll_block
usdt:/usr/sbin/ovs-vswitchd:main:run_start
```

> **Note**
>
> If you execute this on a running process, `bpftrace -lp $(pidof ovs-vswitchd) "usdt:*"` , it will list all USDT events, i.e., also the ones available in the used shared libraries.

Finally, you can use the **tplist** tool which is part of the bcc framework

```
$ /usr/share/bcc/tools/tplist -vv -l $(which ovs-vswitchd)
b'main':b'poll_block' [sema 0x0]
  location #1 b'/usr/sbin/ovs-vswitchd' 0x407fdc
b'main':b'run_start' [sema 0x0]
  location #1 b'/usr/sbin/ovs-vswitchd' 0x407ff6
```

Using probes

We will use the OVS sandbox environment in combination with the probes shown above to try out some of the available trace tools. To start up the virtual environment use the `make sandbox` command. In addition we have to create a bridge to kick of the main bridge loop

```
$ ovs-vsctl add-br test_bridge
$ ovs-vsctl show
055acdca-2f0c-4f6e-b542-f4b6d2c44e08
    Bridge test_bridge
        Port test_bridge
```

(continues on next page)

(continued from previous page)

```
            Interface test_bridge
                type: internal
```

perf

Perf is using Linux uprobe based event tracing to for capturing the events. To enable the main:* probes as displayed above and take an actual trace, you need to execute the following sequence of perf commands

```
# perf buildid-cache --add $(which ovs-vswitchd)

# perf list | grep sdt_
  sdt_main:poll_block                                [SDT event]
  sdt_main:run_start                                 [SDT event]

# perf probe --add=sdt_main:poll_block --add=sdt_main:run_start
Added new events:
  sdt_main:poll_block  (on %poll_block in /usr/sbin/ovs-vswitchd)
  sdt_main:run_start   (on %run_start in /usr/sbin/ovs-vswitchd)

You can now use it in all perf tools, such as:

  perf record -e sdt_main:run_start -aR sleep 1

# perf record -e sdt_main:run_start -e sdt_main:poll_block \
    -p $(pidof ovs-vswitchd) sleep 30
[ perf record: Woken up 1 times to write data ]
[ perf record: Captured and wrote 0.039 MB perf.data (132 samples) ]

# perf script
    ovs-vswitchd  8576 [011] 21031.340433:  sdt_main:run_start: (407ff6)
    ovs-vswitchd  8576 [011] 21031.340516: sdt_main:poll_block: (407fdc)
    ovs-vswitchd  8576 [011] 21031.841726:  sdt_main:run_start: (407ff6)
    ovs-vswitchd  8576 [011] 21031.842088: sdt_main:poll_block: (407fdc)
...
```

Note that the above examples works with the sandbox environment, so make sure you execute the above command while in the sandbox shell!

There are a lot more options available with perf, for example, the `--call-graph dwarf` option, which would give you a call graph in the `perf script` output. See the perf documentation for more information.

One other interesting feature is that the perf data can be converted for use by the trace visualizer Trace Compass. This can be done using the `--all --to-ctf` option to the `perf data convert` tool.

bpftrace

bpftrace is a high-level tracing language based on eBPF, which can be used to script USDT probes. Here we will show a simple one-liner to display the USDT probes being hit. However, the script section below reference some more advanced bpftrace scripts.

This is a simple bpftrace one-liner to show all `main:*` USDT probes

```
# bpftrace -p $(pidof ovs-vswitchd) -e \
    'usdt::main:* { printf("%s %u [%u] %u %s\n",
      comm, pid, cpu, elapsed, probe); }'
Attaching 2 probes...
ovs-vswitchd 8576 [11] 203833199 usdt:main:run_start
ovs-vswitchd 8576 [11] 204086854 usdt:main:poll_block
ovs-vswitchd 8576 [11] 221611985 usdt:main:run_start
ovs-vswitchd 8576 [11] 221892019 usdt:main:poll_block
```

bcc

The BPF Compiler Collection (BCC) is a set of tools and scripts that also use eBPF for tracing. The example below uses the `trace` tool to show the events while they are being generated

```
# /usr/share/bcc/tools/trace -T -p $(pidof ovs-vswitchd) \
    'u::main:run_start' 'u::main:poll_block'
TIME     PID     TID     COMM            FUNC
15:49:06 8576    8576    ovs-vswitchd    main:run_start
15:49:06 8576    8576    ovs-vswitchd    main:poll_block
15:49:06 8576    8576    ovs-vswitchd    main:run_start
15:49:06 8576    8576    ovs-vswitchd    main:poll_block
^C
```

Scripts

To not have to re-invent the wheel when trying to debug complex OVS issues, a set of scripts are provided in the source repository. They are located in the `utilities/usdt-scripts/` directory, and each script contains detailed information on how they should be used, and what information they provide.

Available probes

The next sections describes all the available probes, their use case, and if used in any script, which one. Any new probes being added to OVS should get their own section. See the below "Adding your own probes" section for the used naming convention.

Available probes in `ovs_vswitchd`:

- dpif_netlink_operate__:op_flow_del
- dpif_netlink_operate__:op_flow_execute
- dpif_netlink_operate__:op_flow_get
- dpif_netlink_operate__:op_flow_put
- dpif_recv:recv_upcall
- main:poll_block
- main:run_start
- revalidate:flow_result
- revalidate_ukey__:entry
- revalidate_ukey__:exit
- revalidator_sweep__:flow_result
- udpif_revalidator:start_dump

- udpif_revalidator:sweep_done

dpif_netlink_operate__:op_flow_del

Description:

This probe gets triggered when the Netlink datapath is about to execute the DPIF_OP_FLOW_DEL operation as part of the dpif `operate()` callback.

Arguments:

- *arg0*: `(struct dpif_netlink *) dpif`
- *arg1*: `(struct dpif_flow_del *) del`
- *arg2*: `(struct dpif_netlink_flow *) flow`
- *arg3*: `(struct ofpbuf *) aux->request`

Script references:

- `utilities/usdt-scripts/dpif_op_nl_monitor.py`

dpif_netlink_operate__:op_flow_execute

Description:

This probe gets triggered when the Netlink datapath is about to execute the DPIF_OP_FLOW_EXECUTE operation as part of the dpif `operate()` callback.

Arguments:

- *arg0*: `(struct dpif_netlink *) dpif`
- *arg1*: `(struct dpif_execute *) op->execute`
- *arg2*: `dp_packet_data(op->execute.packet)`
- *arg3*: `dp_packet_size(op->execute.packet)`
- *arg4*: `(struct ofpbuf *) aux->request`

Script references:

- `utilities/usdt-scripts/dpif_op_nl_monitor.py`
- `utilities/usdt-scripts/upcall_cost.py`

dpif_netlink_operate__:op_flow_get

Description:

This probe gets triggered when the Netlink datapath is about to execute the DPIF_OP_FLOW_GET operation as part of the dpif `operate()` callback.

Arguments:

- *arg0*: `(struct dpif_netlink *) dpif`
- *arg1*: `(struct dpif_flow_get *) get`
- *arg2*: `(struct dpif_netlink_flow *) flow`
- *arg3*: `(struct ofpbuf *) aux->request`

Script references:

- `utilities/usdt-scripts/dpif_op_nl_monitor.py`

dpif_netlink_operate__:op_flow_put

Description:

This probe gets triggered when the Netlink datapath is about to execute the DPIF_OP_FLOW_PUT operation as part of the dpif `operate()` callback.

Arguments:

- *arg0*: `(struct dpif_netlink *) dpif`
- *arg1*: `(struct dpif_flow_put *) put`
- *arg2*: `(struct dpif_netlink_flow *) flow`
- *arg3*: `(struct ofpbuf *) aux->request`

Script references:

- `utilities/usdt-scripts/dpif_op_nl_monitor.py`
- `utilities/usdt-scripts/upcall_cost.py`

probe dpif_recv:recv_upcall

Description:

This probe gets triggered when the datapath independent layer gets notified that a packet needs to be processed by userspace. This allows the probe to intercept all packets sent by the kernel to `ovs-vswitchd`. The `upcall_monitor.py` script uses this probe to display and capture all packets sent to `ovs-vswitchd`.

Arguments:

- *arg0*: `(struct dpif *)->full_name`
- *arg1*: `(struct dpif_upcall *)->type`
- *arg2*: `dp_packet_data((struct dpif_upcall *)->packet)`
- *arg3*: `dp_packet_size((struct dpif_upcall *)->packet)`
- *arg4*: `(struct dpif_upcall *)->key`
- *arg5*: `(struct dpif_upcall *)->key_len`

Script references:

- `utilities/usdt-scripts/upcall_cost.py`
- `utilities/usdt-scripts/upcall_monitor.py`

probe main:run_start

Description:

The ovs-vswitchd's main process contains a loop that runs every time some work needs to be done. This probe gets triggered every time the loop starts from the beginning. See also the `main:poll_block` probe below.

Arguments:

None

Script references:

- `utilities/usdt-scripts/bridge_loop.bt`

probe main:poll_block

Description:

The ovs-vswitchd's main process contains a loop that runs every time some work needs to be done. This probe gets triggered every time the loop is done, and it's about to wait for being re-started by a poll_block() call returning. See also the `main:run_start` probe above.

Arguments:

None

Script references:

- `utilities/usdt-scripts/bridge_loop.bt`

revalidate_ukey__:entry

Description:

This probe gets triggered on entry of the revalidate_ukey__() function.

Arguments:

- *arg0*: `(struct udpif *) udpif`
- *arg1*: `(struct udpif_key *) ukey`
- *arg2*: `(uint16_t) tcp_flags`
- *arg3*: `(struct ofpbuf *) odp_actions`
- *arg4*: `(struct recirc_refs *) recircs`
- *arg5*: `(struct xlate_cache *) xcache`

Script references:

- `utilities/usdt-scripts/reval_monitor.py`

revalidate_ukey__:exit

Description:

This probe gets triggered right before the revalidate_ukey__() function exits.

Arguments:

- *arg0*: `(struct udpif *) udpif`
- *arg1*: `(struct udpif_key *) ukey`
- *arg2*: `(enum reval_result) result`

Script references:

None

udpif_revalidator:start_dump

Description:

The ovs-vswitchd's revalidator process contains a loop that runs every time revalidation work is needed. This probe gets triggered every time the dump phase has started.

Arguments:

- *arg0*: `(struct udpif *) udpif`
- *arg1*: `(size_t) n_flows`

Script references:

- `utilities/usdt-scripts/reval_monitor.py`

udpif_revalidator:sweep_done

Description:

The ovs-vswitchd's revalidator process contains a loop that runs every time revalidation work is needed. This probe gets triggered every time the sweep phase was completed.

Arguments:

- *arg0*: `(struct udpif *) udpif`
- *arg1*: `(size_t) n_flows`
- *arg2*: `(unsigned) MIN(ofproto_max_idle, ofproto_max_revalidator)`

Script references:

- `utilities/usdt-scripts/reval_monitor.py`

probe revalidate:flow_result

Description: This probe is triggered when the revalidator has executed on a particular flow key to make a determination whether to evict a flow, and the cause for eviction. The revalidator runs periodically, and this probe will only be triggered when a flow is flagged for revalidation.

Arguments:

- *arg0*: `(struct udpif *) udpif`
- *arg1*: `(struct udpif_key *) ukey`
- *arg2*: `(enum reval_result) result`
- *arg3*: `(enum flow_del_reason) del_reason`

Script references:

- `utilities/usdt-scripts/flow_reval_monitor.py`

probe revalidator_sweep__:flow_result

Description: This probe is placed in the path of the revalidator sweep, and is executed under the condition that a flow entry is in an unexpected state, or the flows were asked to be purged due to a user action.

Arguments:

- *arg0*: `(struct udpif *) udpif`
- *arg1*: `(struct udpif_key *) ukey`

- *arg2*: `(enum reval_result) result`
- *arg3*: `(enum flow_del_reason) del_reason`

Script references:

- `utilities/usdt-scripts/flow_reval_monitor.py`

Adding your own probes

Adding your own probes is as simple as adding the `OVS_USDT_PROBE()` macro to the code. It's similar to the `DTRACE_PROBExx` macro's with the difference that it does automatically determine the number of optional arguments.

The macro requires at least two arguments. The first one being the *provider*, and the second one being the *name*. To keep some consistency with the probe naming, please use the following convention. The *provider* should be the function name, and the *name* should be the name of the tracepoint. If you do function entry and exit like probes, please use `entry` and `exit`.

If, for some reason, you do not like to use the function name as a *provider*, please prefix it with `__`, so we know it's not a function name.

The remaining parameters, up to 10, can be variables, pointers, etc., that might be of interest to capture at this point in the code. Note that the provided variables can cause the compiler to be less effective in optimizing.

5.1.22 Visualizing flows with ovs-flowviz

When troubleshooting networking issues with OVS, it's common to end up looking at OpenFlow or datapath flow dumps. These dumps tend to be quite dense and difficult to reason about.

`ovs-flowviz` is a utility script that helps visualizing OpenFlow and datapath flows to make it easier to understand what is going on.

The ovs-flowviz(8) manpage describes its basic usage. In this document a few of its advanced visualization formats will be expanded.

Installing ovs-flowviz

`ovs-flowviz` is part of the openvswitch python package but its extra dependencies have to be installed explicitly by running:

```
$ pip install openvswitch[flowviz]
```

Or, if you are working with the OVS tree:

```
$ cd python && pip install .[flowviz]
```

Visualizing OpenFlow logical block

When controllers such as OVN write OpenFlow flows, they typically organize flows in functional blocks. These blocks can expand to multiple flows that "look similar", in the sense that they match on the same fields and have similar actions.

However, looking at a flow dump the number of flows can make it difficult to perceive this logical functionality that the controller is trying to implement using OpenFlow.

`ovs-flowviz openflow logic` visualization can be used to understand an OVN flow dump a bit better.

On a particular flow dump table 0 contains 23 flows:

```
$ grep -c "table=0" flows.txt
23
```

Looking at the first few lines, the amount of information can be overwhelming and difficult our analysis:

```
$ head flows.txt
  cookie=0xf76b4b20, duration=765.107s, table=0, n_packets=0, n_bytes=0, priority=180,
↪vlan_tci=0x0000/0x1000 actions=conjunction(100,2/2)
  cookie=0xf76b4b20, duration=765.107s, table=0, n_packets=0, n_bytes=0, priority=180,
↪conj_id=100,in_port="patch-br-int-to",vlan_tci=0x0000/0x1000 actions=load:0xa->NXM_NX_
↪REG13[],load:0xc->NXM_NX_REG11[],load:0xb->NXM_NX_REG12[],load:0xb->OXM_OF_METADATA[],
↪load:0x1->NXM_NX_REG14[],mod_dl_src:02:42:ac:12:00:03,resubmit(,8)
  cookie=0x0, duration=765.388s, table=0, n_packets=0, n_bytes=0, priority=100,in_port=
↪"ovn-6bb3b3-0" actions=move:NXM_NX_TUN_ID[0..23]->OXM_OF_METADATA[0..23],move:NXM_NX_
↪TUN_METADATA0[16..30]->NXM_NX_REG14[0..14],move:NXM_NX_TUN_METADATA0[0..15]->NXM_NX_
↪REG15[0..15],resubmit(,40)
  cookie=0x0, duration=765.388s, table=0, n_packets=0, n_bytes=0, priority=100,in_port=
↪"ovn-a6ff98-0" actions=move:NXM_NX_TUN_ID[0..23]->OXM_OF_METADATA[0..23],move:NXM_NX_
↪TUN_METADATA0[16..30]->NXM_NX_REG14[0..14],move:NXM_NX_TUN_METADATA0[0..15]->NXM_NX_
↪REG15[0..15],resubmit(,40)
  cookie=0xf2ca6195, duration=765.107s, table=0, n_packets=6, n_bytes=636, priority=100,
↪in_port="ovn-k8s-mp0" actions=load:0x1->NXM_NX_REG13[],load:0x2->NXM_NX_REG11[],
↪load:0x7->NXM_NX_REG12[],load:0x4->OXM_OF_METADATA[],load:0x2->NXM_NX_REG14[],
↪resubmit(,8)
  cookie=0x236e941d, duration=408.874s, table=0, n_packets=11, n_bytes=846, priority=100,
↪in_port=aceac9829941d11 actions=load:0x11->NXM_NX_REG13[],load:0x2->NXM_NX_REG11[],
↪load:0x7->NXM_NX_REG12[],load:0x4->OXM_OF_METADATA[],load:0x3->NXM_NX_REG14[],
↪resubmit(,8)
  cookie=0x3facf689, duration=405.581s, table=0, n_packets=11, n_bytes=846, priority=100,
↪in_port="363ba22029cd92b" actions=load:0x12->NXM_NX_REG13[],load:0x2->NXM_NX_REG11[],
↪load:0x7->NXM_NX_REG12[],load:0x4->OXM_OF_METADATA[],load:0x4->NXM_NX_REG14[],
↪resubmit(,8)
  cookie=0xe7c8c4bb, duration=405.570s, table=0, n_packets=11, n_bytes=846, priority=100,
↪in_port="6a62cde0d50ef44" actions=load:0x13->NXM_NX_REG13[],load:0x2->NXM_NX_REG11[],
↪load:0x7->NXM_NX_REG12[],load:0x4->OXM_OF_METADATA[],load:0x5->NXM_NX_REG14[],
↪resubmit(,8)
  cookie=0x99a0ffc1, duration=59.391s, table=0, n_packets=8, n_bytes=636, priority=100,
↪in_port="5ff3bfaaa4eb622" actions=load:0x14->NXM_NX_REG13[],load:0x2->NXM_NX_REG11[],
↪load:0x7->NXM_NX_REG12[],load:0x4->OXM_OF_METADATA[],load:0x6->NXM_NX_REG14[],
↪resubmit(,8)
  cookie=0xe1b5c263, duration=59.365s, table=0, n_packets=8, n_bytes=636, priority=100,
↪in_port="8d9e0bc76347e59" actions=load:0x15->NXM_NX_REG13[],load:0x2->NXM_NX_REG11[],
↪load:0x7->NXM_NX_REG12[],load:0x4->OXM_OF_METADATA[],load:0x7->NXM_NX_REG14[],
↪resubmit(,8)
```

However, table 0 can be better understood by looking at its logical representation:

```
$ ovs-flowviz -i flows.txt -f "table=0" openflow logic
 Ofproto Flows (logical)
  └── ** TABLE 0 **
      ├── priority=180 priority,vlan_tci  --->  conjunction ( x 1 )
      ├── priority=180 priority,conj_id,in_port,vlan_tci  --->  load,load,load,load,load,
↪mod_dl_src resubmit(,8), ( x 1 )
      ├── priority=100 priority,in_port  --->  move,move,move resubmit(,40), ( x 2 )
      ├── priority=100 priority,in_port  --->  load,load,load,load,load resubmit(,8), ( x␣
↪16 )
      ├── priority=100 priority,in_port,vlan_tci  --->  load,load,load,load,load␣
```

(continues on next page)

(continued from previous page)

```
↪resubmit(,8), ( x 1 )
    ├── priority=100 priority,in_port,dl_vlan  --->  strip_vlan,load,load,load,load,
↪load resubmit(,8), ( x 1 )
    └── priority=0 priority  --->   drop, ( x 1 )
```

In only a few logical blocks, there is a good overview of what this table is doing. It looks like it's adding metadata based on input ports and vlan IDs and mainly sending traffic to table 8.

A possible next step might be to look at table 8, and in this case, filter out the flows that have not been hit by actual traffic. This is quite easy to do with the arithmetic filtering expressions:

```
$ ovs-flowviz -i flows.txt -f "table=8 and n_packets>0" openflow logic

 Ofproto Flows (logical)
 └── ** TABLE 8 **
     ├── priority=50 priority,reg14,metadata,dl_dst  --->  load resubmit(,9), ( x 3 )
     └── priority=50 priority,metadata  --->  load,move resubmit(,73),resubmit(,9), ( x␣
↪2 )
```

At this point, understanding the output might be difficult without relating it to the matadata OVN stored in the previous table. This is where `ovs-flowviz`'s OVN integration is useful:

```
$ export OVN_NB_DB=tcp:172.18.0.4:6641
$ export OVN_SB_DB=tcp:172.18.0.4:6642
$ ovs-flowviz -i flows.txt -f "table=8 and n_packets>0" openflow logic --ovn-detrace
Ofproto Flows (logical)
└── ** TABLE 8 **
    ├── cookie=0xe10c34ee priority=50 priority,reg14,metadata,dl_dst  --->  load␣
↪resubmit(,9), ( x 1 )
    │   └── OVN Info
    │       ├── *  Logical datapaths:
    │       ├── *      "ovn_cluster_router" (366e1c41-0f3d-4420-b796-10692b64e3e4)
    │       ├── *  Logical flow: table=0 (lr_in_admission), priority=50, match=(eth.
↪mcast && inport == "rtos-ovn-worker2), actions=(xreg0[0..47] = 0a:58:0a:f4:01:01; next;
↪)
    │       └── *  Logical Router Port: rtos-ovn-worker2 mac 0a:58:0a:f4:01:01 networks [
↪'10.244.1.1/24'] ipv6_ra_configs {}
    ├── cookie=0x11e1adbc priority=50 priority,reg14,metadata,dl_dst  --->  load␣
↪resubmit(,9), ( x 1 )
    │   └── OVN Info
    │       ├── *  Logical datapaths:
    │       ├── *      "GR_ovn-worker2" (c07f8387-6479-4e81-9304-9f8e54f81c56)
    │       ├── *  Logical flow: table=0 (lr_in_admission), priority=50, match=(eth.
↪mcast && inport == "rtoe-GR_ovn-worker2), actions=(xreg0[0..47] = 02:42:ac:12:00:03;␣
↪next;)
    │       └── *  Logical Router Port: rtoe-GR_ovn-worker2 mac 02:42:ac:12:00:03␣
↪networks ['172.18.0.3/16'] ipv6_ra_configs {}
    ├── cookie=0xf42133f  priority=50 priority,reg14,metadata,dl_dst  --->  load␣
↪resubmit(,9), ( x 1 )
    │   └── OVN Info
    │       ├── *  Logical datapaths:
    │       ├── *      "GR_ovn-worker2" (c07f8387-6479-4e81-9304-9f8e54f81c56)
    │       ├── *  Logical flow: table=0 (lr_in_admission), priority=50, match=(eth.dst␣
```

(continues on next page)

(continued from previous page)

```
↪== 02:42:ac:12:00:03 && inport == "rtoe-GR_ovn-worker2), actions=(xreg0[0..47] =␣
↪02:42:ac:12:00:03; next;)
    │       └── *  Logical Router Port: rtoe-GR_ovn-worker2 mac 02:42:ac:12:00:03␣
↪networks ['172.18.0.3/16'] ipv6_ra_configs {}
    └── cookie=0x43a0327  priority=50 priority,metadata  --->  load,move resubmit(,73),
↪resubmit(,9), ( x 2 )
        └── OVN Info
            ├── *  Logical datapaths:
            ├── *      "ovn-worker" (24280d0b-fee0-4f8e-ba4f-036a9b9af921)
            ├── *      "ovn-control-plane" (3262a782-8961-416b-805e-08233e8fda72)
            ├── *      "ext_ovn-worker2" (3f88dcd2-c56d-478f-a3b1-c7aee2efe967)
            ├── *      "ext_ovn-worker" (5facbaf0-485d-4cf5-8940-eff9678ef7bb)
            ├── *      "ext_ovn-control-plane" (8b0aecb6-b05a-48a7-ad09-72524bb91d40)
            ├── *      "join" (e2dc230e-2f2a-4b93-93fa-0fe495163514)
            ├── *      "ovn-worker2" (f7709fbf-d728-4cff-9b9b-150461cc75d2)
            └── *  Logical flow: table=0 (ls_in_check_port_sec), priority=50, match=(1),␣
↪actions=(reg0[15] = check_in_port_sec(); next;)
```

`ovs-flowviz` has automatically added the *cookie* to the logical block key so more blocks have been printed. In exchange, it has looked up each cookie on the running OVN databases and inserted the known information on each block.

The logical flow that generated each OpenFlow flow and the logical datapath it belongs to are now printed, making OVN's pipeline clearer.

Visualizing datapath flow trees

Another typical usecase that can lead to eyestrain is understanding datapath conntrack recirculations.

OVS makes heavy use of connection tracking and the `recirc()` action to build complex datapaths. Typically, OVS will insert a flow that, when matched, will send the packet through conntrack (using the `ct` action) and recirculate it with a particular recirculation id (`recirc_id`). Then, flows matching on that `recirc_id` will be matched and further process the packet. This can happen more than once for a given packet.

This sequential set of events is, however, difficult to visualize when you look at a datapath flow dump. Flows are unordered so recirculations need to be followed manually (typically, with heavy use of "grep").

For this use-case, `ovs-flowviz datapath tree` format can be extremely useful. It builds a hierarchical tree based on the `recirc_id`, `in_port` and `recirc()` actions.

Furthermore, it is common to end up with multiple flows that have the same list of actions. An example of this is a number flows that perform mac/vlan checks for a given port and send the traffic though the same conntrack zone. In order to better visualize this and reduce the amount of duplicated flows that are printed in this view, these flows are combined into a block, and the match keys that are equal for all flows are removed.

For example:

```
Datapath Flows (logical)
└── ──────────────────────────────────
    │ [recirc_id(0x0) in_port(eth0)] │
    ──────────────────────────────────
    └──␣
↪──────────────────────────────────────────────────────────────────────────────────
        │ recirc_id(0),dp_hash(0/0),skb_priority(0/0),in_port(eth0),skb_mark(0/0),ct_
↪state(0/0),ct_zone(0/0),ct_mark(0/0),ct_label(0/0),eth(src=0a:58:0a:84:00:07,
↪dst=22:a1:5d:dc:95:50),eth_type(0x0800),ipv4(src=10.132.0.7,dst=1 │
```

(continues on next page)

(continued from previous page)

```
        │ 0.128.0.0/255.128.0.0,proto=6,tos=0/0,ttl=0/0,frag=no),tcp(src=0/0,dst=0/0),
↪tcp_flags(0/0), packets:4924, bytes:468961,                                    ␣
↪                                                             │
        │ recirc_id(0),dp_hash(...),skb_priority(...),in_port(eth0),skb_mark(...),ct_
↪state(...),ct_zone(...),ct_mark(...),ct_label(...),eth(src=0a:58:0a:84:00:07,
↪dst=0a:58:0a:84:00:01),eth_type(......),ipv4(src=10.132.0.7,dst=1 │
        │ 0.0.0.0/255.255.128.0,proto=17,tos=0/0,ttl=0/0,frag=no),udp(src=32768/0x8000,
↪dst=0/0), packets:711, bytes:114236,                                           ␣
↪                                                            │
        │ recirc_id(0),dp_hash(...),skb_priority(...),in_port(eth0),skb_mark(...),ct_
↪state(...),ct_zone(...),ct_mark(...),ct_label(...),eth(src=0a:58:0a:84:00:07,
↪dst=0a:58:0a:84:00:14),eth_type(......),ipv4(src=10.132.0.7,dst=1 │
        │ 0.128.0.0/255.128.0.0,proto=17,tos=0/0,ttl=0/0,frag=no),udp(src=4096/0xf000,
↪dst=0/0), packets:140, bytes:114660,                                           ␣
↪                                                             │
        │ recirc_id(0),dp_hash(...),skb_priority(...),in_port(eth0),skb_mark(...),ct_
↪state(...),ct_zone(...),ct_mark(...),ct_label(...),eth(src=0a:58:0a:84:00:07,
↪dst=0a:58:0a:84:00:22),eth_type(......),ipv4(src=10.132.0.7,dst=1 │
        │ 0.128.0.0/255.128.0.0,proto=6,tos=0/0,ttl=0/0,frag=no),tcp(src=0/0,dst=0/0),
↪tcp_flags(0/0), packets:1, bytes:66,                                           ␣
↪                                                             │
        │ recirc_id(0),dp_hash(...),skb_priority(...),in_port(eth0),skb_mark(...),ct_
↪state(...),ct_zone(...),ct_mark(...),ct_label(...),eth(src=0a:58:0a:84:00:07,
↪dst=0a:58:0a:84:00:09),eth_type(......),ipv4(src=10.132.0.7,dst=1 │
        │ 0.128.0.0/255.128.0.0,proto=17,tos=0/0,ttl=0/0,frag=no),udp(src=4096/0xf000,
↪dst=0/0), packets:0, bytes:0,                                                  ␣
↪                                                             │
      ␣
↪────────────────────────────────────────────────────────────────────────────────────────────
        └── ──────────────────────────────────
            │ actions: ct(zone=32,nat),recirc(0xc1) │
            ──────────────────────────────────
            └── ─────────────────────────────
                │ [recirc_id(0xc1) in_port(eth0)] │
                ─────────────────────────────
                ├──␣
↪────────────────────────────────────────────────────────────────────────────────────────────
                │   │ recirc_id(0xc1),dp_hash(0/0),skb_priority(0/0),in_port(eth0),skb_
↪mark(0/0),ct_state(0x2a/0x3f),ct_zone(0/0),ct_mark(0/0xf),ct_label(0/0),
↪eth(src=0a:58:0a:84:00:07,dst=22:a1:5d:dc:95:50),eth_type(0x0800),ip │
                │   │ v4(src=0.0.0.0/0.0.0.0,dst=0.0.0.0/0.0.0.0,proto=6,tos=0/0,ttl=0/0,
↪frag=no),tcp(src=0/0,dst=0/0),tcp_flags(0/0), packets:4924, bytes:468961,      ␣
↪                                                          │
                │  ␣
↪────────────────────────────────────────────────────────────────────────────────────────────
                │   └── ──────────────────────────────────
                │       │ actions: ct(zone=14,nat),recirc(0xc2) │
                │       ──────────────────────────────────
                │       └── ─────────────────────────────
                │           │ [recirc_id(0xc2) in_port(eth0)] │
                │           ─────────────────────────────
                │           └──␣
```

(continues on next page)

(continued from previous page)

```
↪──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────
                │                 │ recirc_id(0xc2),dp_hash(0/0),skb_priority(0/0),in_
↪port(eth0),skb_mark(0/0),ct_state(0x2a/0x3f),ct_zone(0/0),ct_mark(0/0x1),ct_label(0/0),
↪eth(src=00:00:00:00:00:00/00:00:00:00:00:00,dst=00:00:00 │
                │                 │ :00:00:00/01:00:00:00:00:00),eth_type(0x0800),
↪ipv4(src=0.0.0.0/0.0.0.0,dst=0.0.0.0/0.0.0.0,proto=0/0,tos=0/0,ttl=0/0,frag=no),␣
↪packets:4924, bytes:468961,                                                    │
                │             ␣
↪──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────
                │                 └── ──────────────────────
                │                     │ actions: ovn-k8s-mp0 │
                │                     ──────────────────────
                │
                ├──␣
↪──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────
                │   │ recirc_id(0xc1),dp_hash(0/0),skb_priority(0/0),in_port(eth0),skb_
↪mark(0/0),ct_state(0x2a/0x3f),ct_zone(0/0),ct_mark(0/0xf),ct_label(0/0),
↪eth(src=0a:58:0a:84:00:07,dst=0a:58:0a:84:00:14),eth_type(0x0800),ip │
                │   │ v4(src=0.0.0.0/0.0.0.0,dst=0.0.0.0/0.0.0.0,proto=17,tos=0/0,ttl=0/
↪0,frag=no),udp(src=4096/0xf000,dst=0/0), packets:140, bytes:114660                ␣
↪                                                                  │
                │  ␣
↪──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────
```

The above shows a part of a bigger tree with an initial block of flows at `recirc_id(0)` which match on different destination Ethernet addresses and protocols, and send traffic through conntrack (zone 32).

Then some additional flows at `recirc_id(0xc1)` process each connection independently. One of them, shown in the example, sends packets through conntrack zone 14, and after another recirculation the packet is ultimately sent through a port.

This is a truly complex multi-zone conntrack pipeline that is now significantly clearer thanks to this visualization.

Also note, the flows in the block are conveniently sorted by sent packets.

This example shows only a single "subtree". Even though the combination of flows with the same action helps, if we use this command to display a large dump, the output can be verbose. There are two, combinable, mechanisms that can help.

Plotting datapath trees

By using the `ovs-flowviz datapath html` format, long datapath trees can be displayed in an interactive HTML table. The resulting web page allows subtrees to be expanded and collapsed, allowing focus on the desired information.

The `ovs-flowviz datapath graph` format generates a graphviz graph definition where blocks of flows with the same `recirc_id` match are arranged together, and edges are created to represent recirculations. This format comes with further features such as displaying the conntrack zones, which are key to understanding what the datapath is really doing with a packet.

The `html` and `graph` can also be combined. `ovs-flowviz datapath graph --html` command will output an interactive HTML table alongside a SVG graphical representation of the flows. Flows in the SVG representation link to the corresponding entry in the HTML table.

Filtering

As well as allowing expanding and collapsing subtrees, filtering can be used.

However, filtering works in a slightly different way than it does with OpenFlow flows. Instead of just removing non-matching flows, the output of a filtered datapath flow tree will show full sub-trees containing at least one flow that satisfies the filter.

For example, the following command allows understanding the flows in the above example in the context of traffic going out on port `ovn-k8s-mp0`:

```
$ ovs-appctl dpctl/dump-flows | ovs-flowviz -f "output.port=ovn-k8s-mp0" datapath tree
```

The resulting flow tree will contain all of the flows above, including those with `recirc_id(0)` and `recirc_id(0xc1)` that don't actually output traffic to port `ovn-k8s-mp0`. This is because they are part of a subtree that contains flows that output packets on port `ovn-k8s-mp0`

This provides a "full picture" of how traffic, ending up in a particular port, is being processed.

CHAPTER

SIX

REFERENCE GUIDE

6.1 Man Pages

The following man pages are written in rST and converted to roff at compile time:

6.1.1 ovs-actions

Introduction

This document aims to comprehensively document all of the OpenFlow actions and instructions, both standard and non-standard, supported by Open vSwitch, regardless of origin. The document includes information of interest to Open vSwitch users, such as the semantics of each supported action and the syntax used by Open vSwitch tools, and to developers seeking to build controllers and switches compatible with Open vSwitch, such as the wire format for each supported message.

Actions

In this document, we define an `action` as an OpenFlow action, which is a kind of command that specifies what to do with a packet. Actions are used in OpenFlow flows to describe what to do when the flow matches a packet, and in a few other places in OpenFlow. Each version of the OpenFlow specification defines standard actions, and beyond that many OpenFlow switches, including Open vSwitch, implement extensions to the standard.

OpenFlow groups actions in two ways: as an `action list` or an `action set`, described below.

Action Lists

An `action list`, a concept present in every version of OpenFlow, is simply an ordered sequence of actions. The OpenFlow specifications require a switch to execute actions within an action list in the order specified, and to refuse to execute an action list entirely if it cannot implement the actions in that order [OpenFlow 1.0, section 3.3], with one exception: when an action list outputs multiple packets, the switch may output the packets in an order different from that specified. Usually, this exception is not important, especially in the common case when the packets are output to different ports.

Action Sets

OpenFlow 1.1 introduced the concept of an `action set`. An action set is also a sequence of actions, but the switch reorders the actions and drops duplicates according to rules specified in the OpenFlow specifications. Because of these semantics, some standard OpenFlow actions cannot usefully be included in an action set. For some, but not all, Open vSwitch extension actions, Open vSwitch defines its own action set semantics and ordering.

The OpenFlow pipeline has an action set associated with it as a packet is processed. After pipeline processing is otherwise complete, the switch executes the actions in the action set.

Open vSwitch applies actions in an action set in the following order: Except as noted otherwise below, the action set only executes at most a single action of each type, and when more than one action of a given type is present, the one added to the set later replaces the earlier action:

1. `strip_vlan`
2. `pop_mpls`
3. `decap`
4. `encap`
5. `push_mpls`
6. `push_vlan`
7. `dec_ttl`
8. `dec_mpls_ttl`
9. `dec_nsh_ttl`
10. All of the following actions are executed in the order added to the action set, with cumulative effect. That is, when multiple actions modify the same part of a field, the later modification takes effect, and when they modify different parts of a field (or different fields), then both modifications are applied:
    - `load`
    - `move`
    - `mod_dl_dst`
    - `mod_dl_src`
    - `mod_nw_dst`
    - `mod_nw_src`
    - `mod_nw_tos`
    - `mod_nw_ecn`
    - `mod_nw_ttl`
    - `mod_tp_dst`
    - `mod_tp_src`
    - `mod_vlan_pcp`
    - `mod_vlan_vid`
    - `set_field`
    - `set_tunnel`
    - `set_tunnel64`
11. `set_queue`
12. `group`, `output`, `resubmit`, `ct_clear`, or `ct`. If more than one of these actions is present, then the one listed earliest above is executed and the others are ignored, regardless of the order in which they were added to the action set. (If none of these actions is present, the action set has no real effect, because the modified packet is not sent anywhere and thus the modifications are not visible.)

An action set may only contain the actions listed above.

Error Handling

Packet processing can encounter a variety of errors:

Bridge not found
Open vSwitch supports an extension to the standard OpenFlow `controller` action called a `continuation`, which allows the controller to interrupt and later resume the processing of a packet through the switch pipeline. This error occurs when such a packet's processing cannot be resumed, e.g. because the bridge processing it has been destroyed. Open vSwitch reports this error to the controller as Open vSwitch extension error `NXR_STALE`.

This error prevents packet processing entirely.

Recursion too deep
While processing a given packet, Open vSwitch limits the flow table recursion depth to 64, to ensure that packet processing uses a finite amount of time and space. Actions that count against the recursion limit include `resubmit` from a given OpenFlow table to the same or an earlier table, `group`, and `output` to patch ports.

A `resubmit` from one table to a later one (or, equivalently, a `goto_table` instruction) does not count against the depth limit because resubmits to strictly monotonically increasing tables will eventually terminate. OpenFlow tables are most commonly traversed in numerically increasing order, so this limit has little effect on conventionally designed OpenFlow pipelines.

This error terminates packet processing. Any previous side effects (e.g. output actions) are retained.

Usually this error indicates a loop or other bug in the OpenFlow flow tables. To assist debugging, when this error occurs, Open vSwitch 2.10 and later logs a trace of the packet execution, as if by `ovs-appctl ofproto/trace`, rate-limited to one per minute to reduce the log volume.

Too many resubmits
Open vSwitch limits the total number of `resubmit` actions that a given packet can execute to 4,096. For this purpose, `goto_table` instructions and output to the `table` port are treated like `resubmit`. This limits the amount of time to process a single packet.

Unlike the limit on recursion depth, the limit on resubmits counts all resubmits, regardless of direction.

This error has the same effect, including logging, as exceeding the recursion depth limit.

Stack too deep
Open vSwitch limits the amount of data that the `push` action can put onto the stack at one time to 64 kB of data.

This error terminates packet processing. Any previous side effects (e.g. output actions) are retained.

No recirculation context / Recirculation conflict
These errors indicate internal errors inside Open vSwitch and should generally not occur. If you notice recurring log messages about these errors, please report a bug.

Too many MPLS labels
Open vSwitch can process packets with any number of MPLS labels, but its ability to push and pop MPLS labels is limited, currently to 3 labels. Attempting to push more than the supported number of labels onto a packet, or to pop any number of labels from a packet with more than the supported number, raises this error.

This error terminates packet processing, retaining any previous side effects (e.g. output actions). When this error arises within the execution of a group bucket, it only terminates that bucket's execution, not packet processing overall.

Invalid tunnel metadata
Open vSwitch raises this error when it processes a Geneve packet that has TLV options with an invalid form, e.g. where the length in a TLV would extend past the end of the options.

This error prevents packet processing entirely.

Unsupported packet type
When a `encap` action encapsulates a packet, Open vSwitch raises this error if it does not support the combination

of the new encapsulation with the current packet. `encap(ethernet)` raises this error if the current packet is not an L3 packet, and `encap(nsh)` raises this error if the current packet is not Ethernet, IPv4, IPv6, or NSH.

The `decap` action is supported only for packet types ethernet, NSH and MPLS. Openvswitch raises this error for other packet types. When a `decap` action decapsulates a packet, Open vSwitch raises this error if it does not support the type of inner packet. `decap` of an Ethernet header raises this error if a VLAN header is present, `decap` of a NSH packet raises this error if the NSH inner packet is not Ethernet, IPv4, IPv6, or NSH.

This error terminates packet processing, retaining any previous side effects (e.g. output actions). When this error arises within the execution of a group bucket, it only terminates that bucket's execution, not packet processing overall.

Inconsistencies

OpenFlow 1.0 allows any action to be part of any flow, regardless of the flow's match. Some combinations do not make sense, e.g. an `set_nw_tos` action in a flow that matches only ARP packets or `strip_vlan` in a flow that matches packets without VLAN tags. Other combinations have varying results depending on the kind of packet that the flow processes, e.g. a `set_nw_src` action in a flow that does not match on Ethertype will be treated as a no-op when it processes a non-IPv4 packet. Nevertheless OVS allows all of the above in conformance with OpenFlow 1.0, that is, the following will succeed:

```
$ ovs-ofctl -O OpenFlow10 add-flow br0 arp,actions=mod_nw_tos:12
$ ovs-ofctl -O OpenFlow10 add-flow br0 dl_vlan=0xffff,actions=strip_vlan
$ ovs-ofctl -O OpenFlow10 add-flow br0 actions=mod_nw_src:1.2.3.4
```

Open vSwitch calls these kinds of combinations `inconsistencies` between match and actions. OpenFlow 1.1 and later forbid inconsistencies, and disallow the examples described above by preventing such flows from being added. All of the above, for example, will fail with an error message if one replaces `OpenFlow10` by `OpenFlow11`.

OpenFlow 1.1 and later cannot detect and disallow all inconsistencies. For example, the `write_actions` instruction arbitrarily delays execution of the actions inside it, which can even be canceled with `clear_actions`, so that there is no way to ensure that its actions are consistent with the packet at the time they execute. Thus, actions with `write_actions` and some other contexts are exempt from consistency requirements.

When OVS executes an action inconsistent with the packet, it treats it as a no-op.

Inter-Version Compatibility

Open vSwitch supports multiple OpenFlow versions simultaneously on a single switch. When actions are added with one OpenFlow version and then retrieved with another, Open vSwitch does its best to translate between them.

Inter-version compatibility issues can still arise when different connections use different OpenFlow versions. Backward compatibility is the most obvious case. Suppose, for example, that an OpenFlow 1.1 session adds a flow with a `push_vlan` action, for which there is no equivalent in OpenFlow 1.0. If an OpenFlow 1.0 session retrieves this flow, Open vSwitch must somehow represent the action.

Forward compatibility can also be an issue, because later OpenFlow versions sometimes remove functionality. The best example is the `enqueue` action from OpenFlow 1.0, which OpenFlow 1.1 removed.

In practice, Open vSwitch uses a variety of strategies for inter-version compatibility:

- Most standard OpenFlow actions, such as `output` actions, translate without compatibility issues.
- Open vSwitch supports its extension actions in every OpenFlow version, so they do not pose inter-version compatibility problems.
- Open vSwitch sometimes adds extension actions to ensure backward or forward compatibility. For example, for backward compatibility with the `group` action added in OpenFlow 1.1, Open vSwitch includes an OpenFlow 1.0 extension `group` action.

Perfect inter-version compatibility is not possible, so best results require OpenFlow connections to use a consistent version. One may enforce use of a particular version by setting the `protocols` column for a bridge, e.g. to force `br0` to use only OpenFlow 1.3:

```
ovs-vsctl set bridge br0 protocols=OpenFlow13
```

Field Specifications

Many Open vSwitch actions refer to fields. In such cases, fields may usually be referred to by their common names, such as `eth_dst` for the Ethernet destination field, or by their full OXM or NXM names, such as `NXM_OF_ETH_DST` or `OXM_OF_ETH_DST`. Before Open vSwitch 2.7, only OXM or NXM field names were accepted.

Many actions that act on fields can also act on `subfields`, that is, parts of fields, written as `field[start..end]`, where `start` is the first bit and `end` is the last bit to use in `field`, e.g. `vlan_tci[13..15]` for the VLAN PCP. A single-bit subfield may also be written as `field[offset]`, e.g. `vlan_tci[13]` for the least-significant bit of the VLAN PCP. Empty brackets may be used to explicitly designate an entire field, e.g. `vlan_tci[]` for the entire 16-bit VLAN TCI header. Before Open vSwitch 2.7, brackets were required in field specifications.

See `ovs-fields(7)` for a list of fields and their names.

Port Specifications

Many Open vSwitch actions refer to OpenFlow ports. In such cases, the port may be specified as a numeric port number in the range 0 to 65,535, although Open vSwitch only assigns port numbers in the range 1 through 65,279 to ports. OpenFlow 1.1 and later use 32-bit port numbers, but Open vSwitch never assigns a port number that requires more than 16 bits.

In most contexts, the name of a port may also be used. (The most obvious context where a port name may not be used is in an `ovs-ofctl` command along with the `--no-names` option.) When a port's name contains punctuation or could be ambiguous with other actions, the name may be enclosed in double quotes, with JSON-like string escapes supported (see [RFC 8259]).

Open vSwitch also supports the following standard OpenFlow port names (even in contexts where port names are not otherwise supported). The corresponding OpenFlow 1.0 and 1.1+ port numbers are listed alongside them but should not be used in flow syntax:

- `in_port` (65528 or 0xfff8; 0xfffffff8)
- `table` (65529 or 0xfff9; 0xfffffff9)
- `normal` (65530 or 0xfffa; 0xfffffffa)
- `flood` (65531 or 0xfffb; 0xfffffffb)
- `all` (65532 or 0xfffc; 0xfffffffc)
- `controller` (65533 or 0xfffd; 0xfffffffd)
- `local` (65534 or 0xfffe; 0xfffffffe)
- `any` or `none` (65535 or 0xffff; 0xffffffff)
- `unset` (not in OpenFlow 1.0; 0xfffffff7)

Output Actions

These actions send a packet to a physical port or a controller. A packet that never encounters an output action on its trip through the Open vSwitch pipeline is effectively dropped. Because actions are executed in order, a packet modification action that is not eventually followed by an output action will not have an externally visible effect.

The `output` action

Syntax:

port
`output:`*port*
`output:`*field*
`output(port=`*port*`, max_len=`*nbytes*`)`

Outputs the packet to an OpenFlow port most commonly specified as *port*. Alternatively, the output port may be read from *field*, a field or subfield in the syntax described under *Field Specifications* above. Either way, if the port is the packet's input port, the packet is not output.

The *port* may be one of the following standard OpenFlow ports:

`local`
Outputs the packet on the `local port` that corresponds to the network device that has the same name as the bridge, unless the packet was received on the local port. OpenFlow switch implementations are not required to have a local port, but Open vSwitch bridges always do.

`in_port`
Outputs the packet on the port on which it was received. This is the only standard way to output the packet to the input port (but see *Output to the Input port*, below).

The *port* may also be one of the following additional OpenFlow ports, unless `max_len` is specified:

`normal`
Subjects the packet to the device's normal L2/L3 processing. This action is not implemented by all OpenFlow switches, and each switch implements it differently. The section *The OVS Normal Pipeline* below documents the OVS implementation.

`flood`
Outputs the packet on all switch physical ports, except the port on which it was received and any ports on which flooding is disabled. Flooding can be disabled automatically on a port by Open vSwitch when IEEE 802.1D spanning tree (STP) or rapid spanning tree (RSTP) is enabled, or by a controller using an OpenFlow `OFPT_MOD_PORT` request to set the port's `OFPPC_NO_FLOOD` flag (`ovs-ofctl mod-port` provides a command-line interface to set this flag).

`all`
Outputs the packet on all switch physical ports except the port on which it was received.

`controller`
Sends the packet and its metadata to an OpenFlow controller or controllers encapsulated in an OpenFlow `packet-in` message. The separate `controller` action, described below, provides more options for output to a controller.

Open vSwitch rejects output to other standard OpenFlow ports, including `none`, `unset`, and port numbers reserved for future use as standard ports, with the error `OFPBAC_BAD_OUT_PORT`.

With `max_len`, the packet is truncated to at most *nbytes* bytes before being output. In this case, the output port may not be a patch port. Truncation is just for the single output action, so that later actions in the OpenFlow pipeline work with the complete packet. The truncation feature is meant for use in monitoring applications, e.g. for mirroring packets to a collector.

When an `output` action specifies the number of a port that does not currently exist (and is not in the range for standard ports), the OpenFlow specification allows but does not require OVS to reject the action. All versions of Open vSwitch treat such an action as a no-op. If a port with the number is created later, then the action will be honored at that point. (OpenFlow requires OVS to reject output to a port number that will never be valid, with `OFPBAC_BAD_OUT_PORT`, but this situation does not arise when OVS is a software switch, since the user can add or renumber ports at any time.)

A controller can suppress output to a port by setting its OFPPC_NO_FORWARD flag using an OpenFlow OFPT_MOD_PORT request (ovs-ofctl mod-port provides a command-line interface to set this flag). When output is disabled, output actions (and other actions that output to the port) are allowed but have no effect.

Open vSwitch allows output to a port that does not exist, although OpenFlow allows switches to reject such actions.

Conformance

All versions of OpenFlow and Open vSwitch support output to a literal port. Output to a register is an OpenFlow extension introduced in Open vSwitch 1.3. Output with truncation is an OpenFlow extension introduced in Open vSwitch 2.6.

Output to the Input Port

OpenFlow requires a switch to ignore attempts to send a packet out its ingress port in the most straightforward way. For example, output:234 has no effect if the packet has ingress port 234. The rationale is that dropping these packets makes it harder to loop the network. Sometimes this behavior can even be convenient, e.g. it is often the desired behavior in a flow that forwards a packet to several ports (floods the packet).

Sometimes one really needs to send a packet out its ingress port (hairpin). In this case, use in_port to explicitly output the packet to its input port, e.g.:

```
$ ovs-ofctl add-flow br0 in_port=2,actions=in_port
```

This also works in some circumstances where the flow doesn't match on the input port. For example, if you know that your switch has five ports numbered 2 through 6, then the following will send every received packet out every port, even its ingress port:

```
$ ovs-ofctl add-flow br0 actions=2,3,4,5,6,in_port
```

or, equivalently:

```
$ ovs-ofctl add-flow br0 actions=all,in_port
```

Sometimes, in complicated flow tables with multiple levels of resubmit actions, a flow needs to output to a particular port that may or may not be the ingress port. It's difficult to take advantage of output to in_port in this situation. To help, Open vSwitch provides, as an OpenFlow extension, the ability to modify the in_port field. Whatever value is currently in the in_port field is both the port to which output will be dropped and the destination for in_port. This means that the following adds flows that reliably output to port 2 or to ports 2 through 6, respectively:

```
$ ovs-ofctl add-flow br0 "in_port=2,actions=load:0->in_port,2"
$ ovs-ofctl add-flow br0 "actions=load:0->in_port,2,3,4,5,6"
```

If in_port is important for matching or other reasons, one may save and restore it on the stack:

```
$ ovs-ofctl add-flow br0 \
      actions="push:in_port,load:0->in_port,2,3,4,5,6,pop:in_port"
```

The OVS Normal Pipeline

This section documents how Open vSwitch implements output to the normal port. The OpenFlow specification places no requirements on how this port works, so all of this documentation is specific to Open vSwitch.

Open vSwitch uses the Open_vSwitch database, detailed in ovs-vswitchd.conf.db(5), to determine the details of the normal pipeline.

The normal pipeline executes the following ingress stages for each packet. Each stage either accepts the packet, in which case the packet goes on to the next stage, or drops the packet, which terminates the pipeline. The result of the ingress stages is a set of output ports, which is the empty set if some ingress stage drops the packet:

1. **Input port lookup**: Looks up the OpenFlow `in_port` field's value to the corresponding `Port` and `Interface` record in the database.

   The `in_port` is normally the OpenFlow port that the packet was received on. If `set_field` or another actions changes the `in_port`, the updated value is honored. Accept the packet if the lookup succeeds, which it normally will. If the lookup fails, for example because `in_port` was changed to an unknown value, drop the packet.

2. **Drop malformed packet**: If the packet is malformed enough that it contains only part of an 802.1Q header, then drop the packet with an error.

3. **Drop packets sent to a port reserved for mirroring**: If the packet was received on a port that is configured as the output port for a mirror (that is, it is the `output_port` in some `Mirror` record), then drop the packet.

4. **VLAN input processing**: This stage determines what VLAN the packet is in. It also verifies that this VLAN is valid for the port; if not, drop the packet. How the VLAN is determined and which ones are valid vary based on the `vlan-mode` in the input port's `Port` record:

   `trunk`
   : The packet is in the VLAN specified in its 802.1Q header, or in VLAN 0 if there is no 802.1Q header. The `trunks` column in the `Port` record lists the valid VLANs; if it is empty, all VLANs are valid.

   `access`
   : The packet is in the VLAN specified in the `tag` column of its `Port` record. The packet must not have an 802.1Q header with a nonzero VLAN ID; if it does, drop the packet.

   `native-tagged` / `native-untagged`
   : Same as `trunk` except that the VLAN of a packet without an 802.1Q header is not necessarily zero; instead, it is taken from the `tag` column.

   `dot1q-tunnel`
   : The packet is in the VLAN specified in the `tag` column of its `Port` record, which is a QinQ service VLAN with the Ethertype specified by the `Port`'s `other_config:qinq-ethtype`. If the packet has an 802.1Q header, then it specifies the customer VLAN. The `cvlans` column specifies the valid customer VLANs; if it is empty, all customer VLANs are valid.

5. **Drop reserved multicast addresses**: If the packet is addressed to a reserved Ethernet multicast address and the `Bridge` record does not have `other_config:forward-bpdu` set to `true`, drop the packet.

6. **LACP bond admissibility**: This step applies only if the input port is a member of a bond (a `Port` with more than one `Interface`) and that bond is configured to use LACP. Otherwise, skip to the next step.

   The behavior here depends on the state of LACP negotiation:

   - If LACP has been negotiated with the peer, accept the packet if the bond member is enabled (i.e. carrier is up and it hasn't been administratively disabled). Otherwise, drop the packet.
   - If LACP negotiation is incomplete, then drop the packet. There is one exception: if fallback to active-backup mode is enabled, continue with the next step, pretending that the active-backup balancing mode is in use.

7. **Non-LACP bond admissibility**: This step applies if the input port is a member of a bond without LACP configured, or if a LACP bond falls back to active-backup as described in the previous step. If neither of these applies, skip to the next step.

   If the packet is an Ethernet multicast or broadcast, and not received on the bond's active member, drop the packet.

   The remaining behavior depends on the bond's balancing mode:

L4 (aka TCP balancing)
Drop the packet (this balancing mode is only supported with LACP).

Active-backup
Accept the packet only if it was received on the active member.

SLB (Source Load Balancing)
Drop the packet if the bridge has not learned the packet's source address (in its VLAN) on the port that received it. Otherwise, accept the packet unless it is a gratuitous ARP. Otherwise, accept the packet if the MAC entry we found is ARP-locked. Otherwise, drop the packet. (See the `SLB Bonding` section in the OVS bonding document for more information and a rationale.)

8. **Learn source MAC**: If the source Ethernet address is not a multicast address, then insert a mapping from packet's source Ethernet address and VLAN to the input port in the bridge's MAC learning table. (This is skipped if the packet's VLAN is listed in the switch's `Bridge` record in the `flood_vlans` column, since there is no use for MAC learning when all packets are flooded.)

   When learning happens on a non-bond port, if the packet is a gratuitous ARP, the entry is marked as ARP-locked. The lock expires after 5 seconds. (See the `SLB Bonding` section in the OVS bonding document for more information and a rationale.)

9. **IP multicast path**: If multicast snooping is enabled on the bridge, and the packet is an Ethernet multicast but not an Ethernet broadcast, and the packet is an IP packet, then the packet takes a special processing path. This path is not yet documented here.

10. **Output port set**: Search the MAC learning table for the port corresponding to the packet's Ethernet destination and VLAN. If the search finds an entry, the output port set is just the learned port. Otherwise (including the case where the packet is an Ethernet multicast or in `flood_vlans`), the output port set is all of the ports in the bridge that belong to the packet's VLAN, except for any ports that were disabled for flooding via OpenFlow or that are configured in a `Mirror` record as a mirror destination port.

The following egress stages execute once for each element in the set of output ports. They execute (conceptually) in parallel, so that a decision or action taken for a given output port has no effect on those for another one:

1. **Drop loopback**: If the output port is the same as the input port, drop the packet.

2. **VLAN output processing**: This stage adjusts the packet to represent the VLAN in the correct way for the output port. Its behavior varies based on the `vlan-mode` in the output port's `Port` record:

   `trunk` / `native-tagged` / `native-untagged`
   If the packet is in VLAN 0 (for `native-untagged`, if the packet is in the native VLAN) drops any 802.1Q header. Otherwise, ensures that there is an 802.1Q header designating the VLAN.

   `access`
   Remove any 802.1Q header that was present.

   `dot1q-tunnel`
   Ensures that the packet has an outer 802.1Q header with the QinQ Ethertype and the specified configured tag, and an inner 802.1Q header with the packet's VLAN.

3. **VLAN priority tag processing**: If VLAN output processing discarded the 802.1Q headers, but priority tags are enabled with `other_config:priority-tags` in the output port's `Port` record, then a priority-only tag is added (perhaps only if the priority would be nonzero, depending on the configuration).

4. **Bond member choice**: If the output port is a bond, the code chooses a particular member. This step is skipped for non-bonded ports.

   If the bond is configured to use LACP, but LACP negotiation is incomplete, then normally the packet is dropped. The exception is that if fallback to active-backup mode is enabled, the egress pipeline continues choosing a bond member as if active-backup mode was in use.

For active-backup mode, the output member is the active member. Other modes hash appropriate header fields and use the hash value to choose one of the enabled members.

5. **Output**: The pipeline sends the packet to the output port.

The `controller` action

Syntax:

`controller`
`controller:`*max_len*
`controller(`*key*`[=`*value*`], ...)`

Sends the packet and its metadata to an OpenFlow controller or controllers encapsulated in an OpenFlow `packet-in` message. The supported options are:

`max_len=`*max_len*
Limit to *max_len* the number of bytes of the packet to send in the `packet-in`. A *max_len* of 0 prevents any of the packet from being sent (thus, only metadata is included). By default, the entire packet is sent, equivalent to a *max_len* of 65535. This option has no effect in Open vSwith 2.7 and later: the entire packet will always be sent.

`reason=`*reason*
Specify *reason* as the reason for sending the message in the `packet-in`. The supported reasons are `no_match`, `action`, `invalid_ttl`, `action_set`, `group`, and `packet_out`. The default reason is `action`.

`id=`*controller_id*
Specify *controller_id*, a 16-bit integer, as the connection ID of the OpenFlow controller or controllers to which the `packet-in` message should be sent. The default is zero. Zero is also the default connection ID for each controller connection, and a given controller connection will only have a nonzero connection ID if its controller uses the `NXT_SET_CONTROLLER_ID` Open vSwitch extension to OpenFlow.

`userdata=`*hh*`...`
Supplies the bytes represented as hex digits *hh* as additional data to the controller in the `packet-in` message. Pairs of hex digits may be separated by periods for readability.

`pause`
Causes the switch to freeze the packet's trip through Open vSwitch flow tables and serializes that state into the packet-in message as a `continuation`, an additional property in the `NXT_PACKET_IN2` message. The controller can later send the continuation back to the switch in an `NXT_RESUME` message, which will restart the packet's traversal from the point where it was interrupted. This permits an OpenFlow controller to interpose on a packet midway through processing in Open vSwitch.

Conformance
All versions of OpenFlow and Open vSwitch support `controller` action and its `max_len` option. The `userdata` and `pause` options require the Open vSwitch `NXAST_CONTROLLER2` extension action added in Open vSwitch 2.6. In the absence of these options, the `reason` (other than `reason=action`) and `controller_id` (option than `controller_id=0`) options require the Open vSwitch `NXAST_CONTROLLER` extension action added in Open vSwitch 1.6.

Open vSwitch 2.7 and later is configured to not buffer packets for the packet-in event. As a result, the full packet is always sent to controllers. This means that the `max_len` option has no effect on the `controller` action, and all values (even 0) are equivalent to the default value of 65535.

The `enqueue` action

Syntax:

`enqueue(`*port*`,`*queue*`)`
`enqueue:`*port*`:`*queue*

Enqueues the packet on the specified *queue* within port *port*.

port must be an OpenFlow port number or name as described under *Port Specifications* above. *port* may be `in_port` or `local` but the other standard OpenFlow ports are not allowed.

queue must be a number between 0 and 4294967294 (0xfffffffe), inclusive. The number of actually supported queues depends on the switch. Some OpenFlow implementations do not support queuing at all. In Open vSwitch, the supported queues vary depending on the operating system, datapath, and hardware in use. Use the `QoS` and `Queue` tables in the Open vSwitch database to configure queuing on individual OpenFlow ports (see `ovs-vswitchd.conf.db(5)` for more information).

Conformance

Only OpenFlow 1.0 supports `enqueue`. OpenFlow 1.1 added the `set_queue` action to use in its place along with `output`.

Open vSwitch translates `enqueue` to a sequence of three actions in OpenFlow 1.1 or later: `set_queue:`*queue*`,output:`*port*`,pop_queue`. This is equivalent in behavior as long as the flow table does not otherwise use `set_queue`, but it relies on the `pop_queue` Open vSwitch extension action.

The `bundle` and `bundle_load` actions

Syntax:

`bundle(`*fields*`,`*basis*`,`*algorithm*`,ofport,members:`*port*`...)`
`bundle_load(`*fields*`,`*basis*`,`*algorithm*`,ofport,`*dst*`,members:`*port*`...)`

These actions choose a port (a `member`) from a comma-separated OpenFlow *port* list. After selecting the port, `bundle` outputs to it, whereas `bundle_load` writes its port number to *dst*, which must be a 16-bit or wider field or subfield in the syntax described under *Field Specifications* above.

These actions hash a set of *fields* using *basis* as a universal hash parameter, then apply the bundle link selection *algorithm* to choose a *port*.

fields must be one of the following. For the options with `symmetric` in the name, reversing source and destination addresses yields the same hash:

`eth_src`
Ethernet source address.

`nw_src`
IPv4 or IPv6 source address.

`nw_dst`
IPv4 or IPv6 destination address.

`symmetric_l4`
Ethernet source and destination, Ethernet type, VLAN ID or IDs (if any), IPv4 or IPv6 source and destination, IP protocol, TCP or SCTP (but not UDP) source and destination.

`symmetric_l3l4`
IPv4 or IPv6 source and destination, IP protocol, TCP or SCTP (but not UDP) source and destination.

`symmetric_l3l4+udp`
Like `symmetric_l3l4` but include UDP ports.

algorithm must be one of the following:

active_backup
Chooses the first live port listed in `members`.

hrw (Highest Random Weight)
Computes the following, considering only the live ports in `members`:

```
for i in [1, n_members]:
    weights[i] = hash(flow, i)
member = { i such that weights[i] >= weights[j] for all j != i }
```

This algorithm is specified by RFC 2992.

The algorithms take port liveness into account when selecting members. The definition of whether a port is live is subject to change. It currently takes into account carrier status and link monitoring protocols such as BFD and CFM. If none of the members is live, `bundle` does not output the packet and `bundle_load` stores `OFPP_NONE` (65535) in the output field.

Example: `bundle(eth_src,0,hrw,ofport,members:4,8)` uses an Ethernet source hash with basis 0, to select between OpenFlow ports 4 and 8 using the Highest Random Weight algorithm.

Conformance
Open vSwitch 1.2 introduced the `bundle` and `bundle_load` OpenFlow extension actions.

The `group` action

Syntax:

`group:`*group*

Outputs the packet to the OpenFlow group *group*, which must be a number in the range 0 to 4294967040 (0xffffff00). The group must exist or Open vSwitch will refuse to add the flow. When a group is deleted, Open vSwitch also deletes all of the flows that output to it.

Groups contain action sets, whose semantics are described above in the section *Action Sets*. The semantics of action sets can be surprising to users who expect action list semantics, since action sets reorder and sometimes ignore actions.

A `group` action usually executes the action set or sets in one or more group buckets. Open vSwitch saves the packet and metadata before it executes each bucket, and then restores it afterward. Thus, when a group executes more than one bucket, this means that each bucket executes on the same packet and metadata. Moreover, regardless of the number of buckets executed, the packet and metadata are the same before and after executing the group.

Sometimes saving and restoring the packet and metadata can be undesirable. In these situations, workarounds are possible. For example, consider a pipeline design in which a `select` group bucket is to communicate to a later stage of processing a value based on which bucket was selected. An obvious design would be for the bucket to communicate the value via `set_field` on a register. This does not work because registers are part of the metadata that `group` saves and restores. The following alternative bucket designs do work:

- Recursively invoke the rest of the pipeline with `resubmit`.
- Use `resubmit` into a table that uses `push` to put the value on the stack for the caller to `pop` off. This works because `group` preserves only packet data and metadata, not the stack.

  (This design requires indirection through `resubmit` because actions sets may not contain `push` or `pop` actions.)

An `exit` action within a group bucket terminates only execution of that bucket, not other buckets or the overall pipeline.

Conformance
OpenFlow 1.1 introduced `group`. Open vSwitch 2.6 and later also supports `group` as an extension to OpenFlow 1.0.

Encapsulation and Decapsulation Actions

The `strip_vlan` and `pop` actions

Syntax:

`strip_vlan`
`pop_vlan`

Removes the outermost VLAN tag, if any, from the packet.

The two names for this action are synonyms with no semantic difference. The OpenFlow 1.0 specification uses the name `strip_vlan` and later versions use `pop_vlan`, but OVS accepts either name regardless of version.

In OpenFlow 1.1 and later, consistency rules allow `strip_vlan` only in a flow that matches only packets with a VLAN tag (or following an action that pushes a VLAN tag, such as `push_vlan`). See *Inconsistencies*, above, for more information.

Conformance
All versions of OpenFlow and Open vSwitch support this action.

The `push_vlan` action

Syntax:

`push_vlan:`*ethertype*

Pushes a new outermost VLAN onto the packet. Uses TPID *ethertype*, which must be `0x8100` for an 802.1Q C-tag or `0x88a8` for a 802.1ad S-tag.

Conformance
OpenFlow 1.1 and later supports this action. Open vSwitch 2.8 added support for multiple VLAN tags (with a limit of 2) and 802.1ad S-tags.

The `push_mpls` action

Syntax:

`push_mpls:`*ethertype*

Pushes a new outermost MPLS label stack entry (LSE) onto the packet and changes the packet's Ethertype to *ethertype*, which must be either `B0x8847` or `0x8848`. If the packet did not already contain any MPLS labels, initializes the new LSE as:

Label
2, if the packet contains IPv6, 0 otherwise.

TC
The low 3 bits of the packet's DSCP value, or 0 if the packet is not IP.

TTL
Copied from the IP TTL, or 64 if the packet is not IP.

If the packet did already contain an MPLS label, initializes the new outermost label as a copy of the existing outermost label.

OVS currently supports at most 3 MPLS labels.

This action applies only to Ethernet packets.

Conformance
Open vSwitch 1.11 introduced support for MPLS. OpenFlow 1.1 and later support `push_mpls`. Open vSwitch implements `push_mpls` as an extension to OpenFlow 1.0.

The pop_mpls action

Syntax:

pop_mpls:*ethertype*

Strips the outermost MPLS label stack entry and changes the packet's Ethertype to *ethertype*. This action applies only to Ethernet packets with at least one MPLS label. If there is more than one MPLS label, then *ethertype* should be an MPLS Ethertype (B0x8847 or 0x8848).

Conformance

Open vSwitch 1.11 introduced support for MPLS. OpenFlow 1.1 and later support pop_mpls. Open vSwitch implements pop_mpls as an extension to OpenFlow 1.0.

The encap action

Syntax:

encap(nsh([md_type=*md_type*], [tlv(*class*,*type*,*value*)]...))
encap(ethernet)
encap(mpls)
encap(mpls_mc)

The encap action encapsulates a packet with a specified header. It has variants for different kinds of encapsulation.

The encap(nsh(...)) variant encapsulates an Ethernet frame with NSH. The *md_type* may be 1 or 2 for metadata type 1 or 2, defaulting to 1. For metadata type 2, TLVs may be specified with *class* as a 16-bit hexadecimal integer beginning with 0x, *type* as an 8-bit decimal integer, and *value* a sequence of pairs of hex digits beginning with 0x. For example:

encap(nsh(md_type=1))
Encapsulates the packet with an NSH header with metadata type 1.

encap(nsh(md_type=2,tlv(0x1000,10,0x12345678)))
Encapsulates the packet with an NSH header, NSH metadata type 2, and an NSH TLV with class 0x1000, type 10, and the 4-byte value 0x12345678.

The encap(ethernet) variant encapsulate a bare L3 packet in an Ethernet frame. The Ethernet type is initialized to the L3 packet's type, e.g. 0x0800 if the L3 packet is IPv4. The Ethernet source and destination are initially zeroed.

The encap(mpls) variant adds a MPLS header at the start of the packet. When encap(ethernet) is applied after this action, the ethertype of ethernet header will be populated with MPLS unicast ethertype (0x8847).

The encap(mpls_mc) variant adds a MPLS header at the start of the packet. When encap(ethernet) is applied after this action, the ethertype of ethernet header will be populated with MPLS multicast ethertype (0x8848).

Conformance

This action is an Open vSwitch extension to OpenFlow 1.3 and later, introduced in Open vSwitch 2.8.

The MPLS support for this action is added in Open vSwitch 2.17.

The decap action

Syntax:

decap
decap(packet_type(ns=*namespace*,type=*type*))

Removes an outermost encapsulation from the packet:

- If the packet is an Ethernet packet, removes the Ethernet header, which changes the packet into a bare L3 packet. If the packet has VLAN tags, raises an unsupported packet type error (see *Error Handling*, above).

- Otherwise, if the packet is an NSH packet, removes the NSH header, revealing the inner packet. Open vSwitch supports Ethernet, IPv4, IPv6, and NSH inner packet types. Other types raise unsupported packet type errors.
- Otherwise, if the packet is encapsulated inside a MPLS header, removes the MPLS header and classifies the inner packet as mentioned in the packet type argument of the decap. The *packet_type* field specifies the type of the packet in the format specified in OpenFlow 1.5 chapter *7.2.3.11 Packet Type Match Field*. The inner packet will be incorrectly classified, if the inner packet is different from mentioned in the *packet_type* field.
- Otherwise, raises an unsupported packet type error.

Conformance

This action is an Open vSwitch extension to OpenFlow 1.3 and later, introduced in Open vSwitch 2.8.

The MPLS support for this action is added in Open vSwitch 2.17.

Field Modification Actions

These actions modify packet data and metadata fields.

The `set_field` and `load` actions

Syntax:

`set_field:`*value*`[/`*mask*`]->`*dst*
`load:`*value*`->`*dst*

These actions loads a literal value into a field or part of a field. The `set_field` action takes *value* in the customary syntax for field *dst*, e.g. `00:11:22:33:44:55` for an Ethernet address, and *dst* as the field's name. The optional *mask* allows part of a field to be set.

The `load` action takes *value* as an integer value (in decimal or prefixed by `0x` for hexadecimal) and *dst* as a field or subfield in the syntax described under *Field Specifications* above.

The following all set the Ethernet source address to 00:11:22:33:44:55:

- `set_field:00:11:22:33:44:55->eth_src`
- `load:0x001122334455->eth_src`
- `load:0x001122334455->OXM_OF_ETH_SRC[]`

The following all set the multicast bit in the Ethernet destination address:

- `set_field:01:00:00:00:00:00/01:00:00:00:00:00->eth_dst`
- `load:1->eth_dst[40]`

Open vSwitch prohibits a `set_field` or `load` action whose *dst* is not guaranteed to be part of the packet; for example, `set_field` of `nw_dst` is only allowed in a flow that matches on Ethernet type 0x800. In some cases, such as in an action set, Open vSwitch can't statically check that *dst* is part of the packet, and in that case if it is not then Open vSwitch treats the action as a no-op.

Conformance

Open vSwitch 1.1 introduced `NXAST_REG_LOAD` as a extension to OpenFlow 1.0 and used `load` to express it. Later, OpenFlow 1.2 introduced a standard `OFPAT_SET_FIELD` action that was restricted to loading entire fields, so Open vSwitch added the form `set_field` with this restriction. OpenFlow 1.5 extended `OFPAT_SET_FIELD` to the point that it became a superset of `NXAST_REG_LOAD`. Open vSwitch translates either syntax as necessary for the OpenFlow version in use: in OpenFlow 1.0 and 1.1, `NXAST_REG_LOAD`; in OpenFlow 1.2, 1.3, and 1.4, `NXAST_REG_LOAD` for `load` or for loading a subfield, `OFPAT_SET_FIELD` otherwise; and OpenFlow 1.5 and later, `OFPAT_SET_FIELD`.

The `move` action

Syntax:

`move:`*src*`->`*dst*

Copies the named bits from field or subfield *src* to field or subfield *dst*. *src* and *dst* should fields or subfields in the syntax described under *Field Specifications* above. The two fields or subfields must have the same width.

Examples:

- `move:reg0[0..5]->reg1[26..31]` copies the six bits numbered 0 through 5 in register 0 into bits 26 through 31 of register 1.
- `move:reg0[0..15]->vlan_tci` copies the least significant 16 bits of register 0 into the VLAN TCI field.

Conformance

In OpenFlow 1.0 through 1.4, `move` ordinarily uses an Open vSwitch extension to OpenFlow. In OpenFlow 1.5, `move` uses the OpenFlow 1.5 standard `OFPAT_COPY_FIELD` action. The ONF has also made `OFPAT_COPY_FIELD` available as an extension to OpenFlow 1.3. Open vSwitch 2.4 and later understands this extension and uses it if a controller uses it, but for backward compatibility with older versions of Open vSwitch, `ovs-ofctl` does not use it.

The `mod_dl_src` and `mod_dl_dst` actions

Syntax:

`mod_dl_src:`*mac*
`mod_dl_dst:`*mac*

Sets the Ethernet source or destination address, respectively, to *mac*, which should be expressed in the form `xx:xx:xx:xx:xx:xx`.

For L3-only packets, that is, those that lack an Ethernet header, this action has no effect.

Conformance

OpenFlow 1.0 and 1.1 have specialized actions for these purposes. OpenFlow 1.2 and later do not, so Open vSwitch translates them to appropriate `OFPAT_SET_FIELD` actions for those versions,

The `mod_nw_src` and `mod_nw_dst` actions

Syntax:

`mod_nw_src:`*ip*
`mod_nw_dst:`*ip*

Sets the IPv4 source or destination address, respectively, to *ip*, which should be expressed in the form `w.x.y.z`.

In OpenFlow 1.1 and later, consistency rules allow these actions only in a flow that matches only packets that contain an IPv4 header (or following an action that adds an IPv4 header, e.g. `pop_mpls:0x0800`). See *Inconsistencies*, above, for more information.

Conformance

OpenFlow 1.0 and 1.1 have specialized actions for these purposes. OpenFlow 1.2 and later do not, so Open vSwitch translates them to appropriate `OFPAT_SET_FIELD` actions for those versions,

The mod_nw_tos and mod_nw_ecn actions

Syntax:

`mod_nw_tos:`*tos*
`mod_nw_ecn:`*ecn*

The `mod_nw_tos` action sets the DSCP bits in the IPv4 ToS/DSCP or IPv6 traffic class field to *tos*, which must be a multiple of 4 between 0 and 255. This action does not modify the two least significant bits of the ToS field (the ECN bits).

The `mod_nw_ecn` action sets the ECN bits in the IPv4 ToS or IPv6 traffic class field to *ecn*, which must be a value between 0 and 3, inclusive. This action does not modify the six most significant bits of the field (the DSCP bits).

In OpenFlow 1.1 and later, consistency rules allow these actions only in a flow that matches only packets that contain an IPv4 or IPv6 header (or following an action that adds such a header). See *Inconsistencies*, above, for more information.

Conformance

OpenFlow 1.0 has a `mod_nw_tos` action but not `mod_nw_ecn`. Open vSwitch implements the latter in OpenFlow 1.0 as an extension using `NXAST_REG_LOAD`. OpenFlow 1.1 has specialized actions for these purposes. OpenFlow 1.2 and later do not, so Open vSwitch translates them to appropriate `OFPAT_SET_FIELD` actions for those versions.

The mod_tp_src and mod_tp_dst actions

Syntax:

`mod_tp_src:`*port*
`mod_tp_dst:`*port*

Sets the TCP or UDP or SCTP source or destination port, respectively, to *port*. Both IPv4 and IPv6 are supported.

In OpenFlow 1.1 and later, consistency rules allow these actions only in a flow that matches only packets that contain a TCP or UDP or SCTP header. See *Inconsistencies*, above, for more information.

Conformance

OpenFlow 1.0 and 1.1 have specialized actions for these purposes. OpenFlow 1.2 and later do not, so Open vSwitch translates them to appropriate `OFPAT_SET_FIELD` actions for those versions,

The dec_ttl action

Syntax:

`dec_ttl`
`dec_ttl(`*id1*`[,`*id2*`[, ...]])`

Decrement TTL of IPv4 packet or hop limit of IPv6 packet. If the TTL or hop limit is initially 0 or 1, no decrement occurs, as packets reaching TTL zero must be rejected. Instead, Open vSwitch sends a `packet-in` message with reason code `OFPR_INVALID_TTL` to each connected controller that has enabled receiving such messages, and stops processing the current set of actions. (However, if the current set of actions was reached through `resubmit`, the remaining actions in outer levels resume processing.)

As an Open vSwitch extension to OpenFlow, this action supports the ability to specify a list of controller IDs. Open vSwitch will only send the message to controllers with the given ID or IDs. Specifying no list is equivalent to specifying a single controller ID of zero.

In OpenFlow 1.1 and later, consistency rules allow these actions only in a flow that matches only packets that contain an IPv4 or IPv6 header. See *Inconsistencies*, above, for more information.

Conformance

All versions of OpenFlow and Open vSwitch support this action.

The `set_mpls_label`, `set_mpls_tc`, and `set_mpls_ttl` actions

Syntax:

`set_mpls_label:`*label*
`set_mpls_tc:`*tc*
`set_mpls_ttl:`*ttl*

The `set_mpls_label` action sets the label of the packet's outer MPLS label stack entry. *label* should be a 20-bit value that is decimal by default; use a `0x` prefix to specify the value in hexadecimal.

The `set_mpls_tc` action sets the traffic class of the packet's outer MPLS label stack entry. *tc* should be in the range 0 to 7, inclusive.

The `set_mpls_ttl` action sets the TTL of the packet's outer MPLS label stack entry. *ttl* should be in the range 0 to 255 inclusive. In OpenFlow 1.1 and later, consistency rules allow these actions only in a flow that matches only packets that contain an MPLS label (or following an action that adds an MPLS label, e.g. `push_mpls:0x8847`). See *Inconsistencies*, above, for more information.

Conformance

OpenFlow 1.0 does not support MPLS, but Open vSwitch implements these actions as extensions. OpenFlow 1.1 has specialized actions for these purposes. OpenFlow 1.2 and later do not, so Open vSwitch translates them to appropriate `OFPAT_SET_FIELD` actions for those versions,

The `dec_mpls_ttl` and `dec_nsh_ttl` actions

Syntax:

`dec_mpls_ttl`
`dec_nsh_ttl`

These actions decrement the TTL of the packet's outer MPLS label stack entry or its NSH header, respectively. If the TTL is initially 0 or 1, no decrement occurs. Instead, Open vSwitch sends a `packet-in` message with reason code `BOFPR_INVALID_TTL` to OpenFlow controllers with ID 0, if it has enabled receiving them. Processing the current set of actions then stops. (However, if the current set of actions was reached through `resubmit`, remaining actions in outer levels resume processing.)

In OpenFlow 1.1 and later, consistency rules allow this actions only in a flow that matches only packets that contain an MPLS label or an NSH header, respectively. See *Inconsistencies*, above, for more information.

Conformance

Open vSwitch 1.11 introduced support for MPLS. OpenFlow 1.1 and later support `dec_mpls_ttl`. Open vSwitch implements `dec_mpls_ttl` as an extension to OpenFlow 1.0.

Open vSwitch 2.8 introduced support for NSH, although the NSH draft changed after release so that only Open vSwitch 2.9 and later conform to the final protocol specification. The `dec_nsh_ttl` action and NSH support in general is an Open vSwitch extension not supported by any version of OpenFlow.

The `check_pkt_larger` action

Syntax:

`check_pkt_larger(`*pkt_len*`)->`*dst*

Checks if the packet is larger than the specified length in *pkt_len*. If so, stores 1 in *dst*, which should be a 1-bit field; if not, stores 0.

The packet length to check against the argument *pkt_len* includes the L2 header and L2 payload of the packet, but not the VLAN tag (if present).

Examples:

- `check_pkt_larger(1500)->reg0[0]`
- `check_pkt_larger(8000)->reg9[10]`

This action was added in Open vSwitch 2.12.

The `delete_field` action

Syntax:

`delete_field:`*field*

The `delete_field` action deletes a *field* in the syntax described under *Field Specifications* above. Currently, only the `tun_metadata` fields are supported.

This action was added in Open vSwitch 2.14.

Metadata Actions

The `set_tunnel` action

Syntax:

`set_tunnel:`*id*
`set_tunnel64:`*id*

Many kinds of tunnels support a tunnel ID, e.g. VXLAN and Geneve have a 24-bit VNI, and GRE has an optional 32-bit key. This action sets the value used for tunnel ID in such tunneled packets, although whether it is used for a particular tunnel depends on the tunnel's configuration. See the tunnel ID documentation in `ovs-fields(7)` for more information.

Conformance

These actions are OpenFlow extensions. `set_tunnel` was introduced in Open vSwitch 1.0. `set_tunnel64`, which is needed if *id* is wider than 32 bits, was added in Open vSwitch 1.1. Both actions always set the entire tunnel ID field. Open vSwitch supports these actions in all versions of OpenFlow, but in OpenFlow 1.2 and later it translates them to an appropriate standardized `OFPAT_SET_FIELD` action.

The `set_queue` and `pop_queue` actions

Syntax:

`set_queue:`*queue*
`pop_queue`

The `set_queue` action sets the queue ID to be used for subsequent output actions to *queue*, which must be a 32-bit integer. The range of meaningful values of *queue*, and their meanings, varies greatly from one OpenFlow implementation to another. Even within a single implementation, there is no guarantee that all OpenFlow ports have the same queues configured or that all OpenFlow ports in an implementation can be configured the same way queue-wise. For more information, see the documentation for the output queue field in `ovs-fields(7)`.

The `pop_queue` restores the output queue to the default that was set when the packet entered the switch (generally 0).

Four billion queues ought to be enough for anyone: https://mailman.stanford.edu/pipermail/openflow-spec/2009-August/000394.html

Conformance

OpenFlow 1.1 introduced the `set_queue` action. Open vSwitch also supports it as an extension in OpenFlow 1.0.

The `pop_queue` action is an Open vSwitch extension.

Firewalling Actions

Open vSwitch is often used to implement a firewall. The preferred way to implement a firewall is `connection tracking`, that is, to keep track of the connection state of individual TCP sessions. The `ct` action described in this section, added in Open vSwitch 2.5, implements connection tracking. For new deployments, it is the recommended way to implement firewalling with Open vSwitch.

Before `ct` was added, Open vSwitch did not have built-in support for connection tracking. Instead, Open vSwitch supported the `learn` action, which allows a received packet to add a flow to an OpenFlow flow table. This could be used to implement a primitive form of connection tracking: packets passing through the firewall in one direction could create flows that allowed response packets back through the firewall in the other direction. The additional `fin_timeout` action allowed the learned flows to expire quickly after TCP session termination.

The `ct` action

Syntax:

`ct([`*argument*`]...)`
`ct(commit[,`*argument*`]...)`

The action has two modes of operation, distinguished by whether `commit` is present. The following arguments may be present in either mode:

`zone=`*value*
A zone is a 16-bit id that isolates connections into separate domains, allowing overlapping network addresses in different zones. If a zone is not provided, then the default is 0. The *value* may be specified either as a 16-bit integer literal or a field or subfield in the syntax described under *Field Specifications* above.

Without `commit`, this action sends the packet through the connection tracker. The connection tracker keeps track of the state of TCP connections for packets passed through it. For each packet through a connection, it checks that it satisfies TCP invariants and signals the connection state to later actions using the `ct_state` metadata field, which is documented in `ovs-fields(7)`.

In this form, `ct` forks the OpenFlow pipeline:

- In one fork, `ct` passes the packet to the connection tracker. Afterward, it reinjects the packet into the OpenFlow pipeline with the connection tracking fields initialized. The `ct_state` field is initialized with connection state and `ct_zone` to the connection tracking zone specified on the `zone` argument. If the connection is one that is already tracked, `ct_mark` and `ct_label` to its existing mark and label, respectively; otherwise they are zeroed. In addition, `ct_nw_proto`, `ct_nw_src`, `ct_nw_dst`, `ct_ipv6_src`, `ct_ipv6_dst`, `ct_tp_src`, and `ct_tp_dst` are initialized appropriately for the original direction connection. See the `resubmit` action for a way to search the flow table with the connection tracking original direction fields swapped with the packet 5-tuple fields. See `ovs-fields(7)` for details on the connection tracking fields.
- In the other fork, the original instance of the packet continues independent processing following the `ct` action. The `ct_state` field and other connection tracking metadata are cleared.

Without `commit`, the `ct` action accepts the following arguments:

`table=`*table*
Sets the OpenFlow table where the packet is reinjected. The *table* must be a number between 0 and 254 inclusive, or a table's name. If *table* is not specified, then the packet is not reinjected.

`nat`

`nat(`*type=addrs*`[:`*ports*`][,`*flag*`]...)`
Specify address and port translation for the connection being tracked. The *type* must be `src`, for source address/port translation (SNAT), or `dst`, for destination address/port translation (DNAT).

Setting up address translation for a new connection takes effect only if the connection is later committed with `ct(commit ...)`.

The `src` and `dst` options take the following arguments:

addrs
The IP address `addr` or range `addr1-addr2` from which the translated address should be selected. If only one address is given, then that address will always be selected, otherwise the address selection can be informed by the optional persistent flag as described below. Either IPv4 or IPv6 addresses can be provided, but both addresses must be of the same type, and the datapath behavior is undefined in case of providing IPv4 address range for an IPv6 packet, or IPv6 address range for an IPv4 packet. IPv6 addresses must be bracketed with `[` and `]` if a port range is also given.

ports
The L4 `port` or range `port1-port2` from which the translated port should be selected. When a port range is specified, fallback to ephemeral ports does not happen, else, it will. The port number selection can be informed by the optional `random` and `hash` flags described below.

The optional *flags* are:

`random`
The selection of the port from the given range should be done using a fresh random number. This flag is mutually exclusive with `hash`.

`hash`
The selection of the port from the given range should be done using a datapath specific hash of the packet's IP addresses and the other, non-mapped port number. This flag is mutually exclusive with `random`.

`persistent`
The selection of the IP address from the given range should be done so that the same mapping can be provided after the system restarts.

If `alg` is specified for the committing `ct` action that also includes `nat` with a `src` or `dst` attribute, then the datapath tries to set up the helper to be NAT-aware. This functionality is datapath specific and may not be supported by all datapaths.

A `bare nat` argument with no options will only translate the packet being processed in the way the connection has been set up with an earlier, committed `ct` action. A `nat` action with `src` or `dst`, when applied to a packet belonging to an established (rather than new) connection, will behave the same as a bare `nat`.

For SNAT, there is a special case when the `src` IP address is configured as all 0's, i.e., `nat(src=0.0.0.0)`. In this case, when a source port collision is detected during the commit, the source port will be translated to an ephemeral port. If there is no collision, no SNAT is performed.

Open vSwitch 2.6 introduced `nat`. Linux 4.6 was the earliest upstream kernel that implemented `ct` support for `nat`.

With `commit`, the connection tracker commits the connection to the connection tracking module. The `commit` flag should only be used from the pipeline within the first fork of `ct` without `commit`. Information about the connection is stored beyond the lifetime of the packet in the pipeline. Some `ct_state` flags are only available for committed connections.

The following options are available only with `commit`:

`force`
A committed connection always has the directionality of the packet that caused the connection to be committed in the first place. This is the `original direction` of the connection, and the opposite

direction is the `reply direction`. If a connection is already committed, but it is in the wrong direction, `force` effectively terminates the existing connection and starts a new one in the current direction. This flag has no effect if the original direction of the connection is already the same as that of the current packet.

exec(*action***...)**
Perform each *action* within the context of connection tracking. Only actions which modify the `ct_mark` or `ct_label` fields are accepted within `exec` action, and these fields may only be modified with this option. For example:

set_field:*value***[/***mask***]->ct_mark**
Store a 32-bit metadata value with the connection. Subsequent lookups for packets in this connection will populate `ct_mark` when the packet is sent to the connection tracker with the table specified.

set_field:*value***[/***mask***]->ct_label**
Store a 128-bit metadata value with the connection. Subsequent lookups for packets in this connection will populate `ct_label` when the packet is sent to the connection tracker with the table specified.

alg=*alg*
Specify application layer gateway *alg* to track specific connection types. If subsequent related connections are sent through the `ct` action, then the `rel` flag in the `ct_state` field will be set. Supported types include:

ftp
Look for negotiation of FTP data connections. Specify this option for FTP control connections to detect related data connections and populate the `rel` flag for the data connections.

tftp
Look for negotiation of TFTP data connections. Specify this option for TFTP control connections to detect related data connections and populate the `rel` flag for the data connections.

Related connections inherit `ct_mark` from that stored with the original connection (i.e. the connection created by `ct(alg=...)`.

With the Linux datapath, global sysctl options affect `ct` behavior. In particular, if `net.netfilter.nf_conntrack_helper` is enabled, which it is by default until Linux 4.7, then application layer gateway helpers may be executed even if *alg* is not specified. For security reasons, the netfilter team recommends users disable this option. For further details, please see http://www.netfilter.org/news.html#2012-04-03 .

The `ct` action may be used as a primitive to construct stateful firewalls by selectively committing some traffic, then matching `ct_state` to allow established connections while denying new connections. The following flows provide an example of how to implement a simple firewall that allows new connections from port 1 to port 2, and only allows established connections to send traffic from port 2 to port 1:

```
table=0,priority=1,action=drop
table=0,priority=10,arp,action=normal
table=0,priority=100,ip,ct_state=-trk,action=ct(table=1)
table=1,in_port=1,ip,ct_state=+trk+new,action=ct(commit),2
table=1,in_port=1,ip,ct_state=+trk+est,action=2
table=1,in_port=2,ip,ct_state=+trk+new,action=drop
table=1,in_port=2,ip,ct_state=+trk+est,action=1
```

If `ct` is executed on IPv4 (or IPv6) fragments, then the message is implicitly reassembled before sending to the connection tracker and refragmented upon output, to the original maximum received fragment size. Reassembly occurs within the context of the zone, meaning that IP fragments in different zones are not assembled together. Pipeline processing for the initial fragments is halted. When the final fragment is received, the message is assembled and pipeline processing continues for that flow. Packet ordering is not guaranteed by IP protocols, so it is not possible to determine

which IP fragment will cause message reassembly (and therefore continue pipeline processing). As such, it is strongly recommended that multiple flows should not execute `ct` to reassemble fragments from the same IP message.

Conformance

The `ct` action was introduced in Open vSwitch 2.5. Some of its features were introduced later, noted individually above.

The `ct_clear` action

Syntax:

`ct_clear`

Clears connection tracking state from the flow, zeroing `ct_state`, `ct_zone`, `ct_mark`, and `ct_label`.

This action was introduced in Open vSwitch 2.7.

The `learn` action

Syntax:

`learn(`*argument*`...)`

The `learn` action adds or modifies a flow in an OpenFlow table, similar to `ovs-ofctl --strict mod-flows`. The arguments specify the match fields, actions, and other properties of the flow to be added or modified.

Match fields for the new flow are specified as follows. At least one match field should ordinarily be specified:

field*=*value

Specifies that *field*, in the new flow, must match the literal *value*, e.g. `dl_type=0x800`. Shorthand match syntax, such as `ip` in place of `dl_type=0x800`, is not supported.

field*=*src

Specifies that *field* in the new flow must match *src* taken from the packet currently being processed. For example, `udp_dst=udp_src`, applied to a UDP packet with source port 53, creates a flow which matches `udp_dst=53`. *field* and *src* must have the same width.

field

Shorthand for the previous form when *field* and *src* are the same. For example, `udp_dst`, applied to a UDP packet with destination port 53, creates a flow which matches `udp_dst=53`.

The *field* and *src* arguments above should be fields or subfields in the syntax described under *Field Specifications* above.

Match field specifications must honor prerequisites for both the flow with the `learn` and the new flow that it creates. Consider the following complete flow, in the syntax accepted by `ovs-ofctl`. If the flow's match on `udp` were omitted, then the flow would not satisfy the prerequisites for the `learn` action's use of `udp_src`. If `dl_type=0x800` or `nw_proto` were omitted from `learn`, then the new flow would not satisfy the prerequisite for its match on `udp_dst`. For more information on prerequisites, please refer to `ovs-fields(7)`:

```
udp, actions=learn(dl_type=0x800, nw_proto=17, udp_dst=udp_src)
```

Actions for the new flow are specified as follows. At least one action should ordinarily be specified:

`load:`*value*`->`*dst*

Adds a `load` action to the new flow that loads the literal *value* into *dst*. The syntax is the same as the `load` action explained in the *Field Modification Actions* section.

`load:`*src*`->`*dst*

Adds a `load` action to the new flow that loads *src*, a field or subfield from the packet being processed, into *dst*.

`output:`*field*
: Adds an `output` action to the new flow's actions that outputs to the OpenFlow port taken from *field*, which must be a field as described above.

`fin_idle_timeout=`*seconds* / `fin_hard_timeout=`*seconds*
: Adds a `fin_timeout` action with the specified arguments to the new flow. This feature was added in Open vSwitch 1.6.

The following additional arguments are optional:

`idle_timeout=`*seconds*

`hard_timeout=`*seconds*

`priority=`*value*

`cookie=`*value*

`send_flow_rem`
: These arguments have the same meaning as in the usual flow syntax documented in `ovs-ofctl(8)`.

`table=`*table*
: The table in which the new flow should be inserted. Specify a decimal number between 0 and 254 inclusive or the name of a table. The default, if table is unspecified, is table 1 (not 0).

`delete_learned`
: When this flag is specified, deleting the flow that contains the `learn` action will also delete the flows created by `learn`. Specifically, when the last `learn` action with this flag and particular `table` and `cookie` values is removed, the switch deletes all of the flows in the specified table with the specified cookie.

  This flag was added in Open vSwitch 2.4.

`limit=`*number*
: If the number of flows in the new flow's table with the same cookie exceeds *number*, the action will not add a new flow. By default, or with `limit=0`, there is no limit.

  This flag was added in Open vSwitch 2.8.

`result_dst=`*field*`[`*bit*`]`
: If learn fails (because the number of flows exceeds `limit`), the action sets *field*[*bit*] to 0, otherwise it will be set to 1. *field*[*bit*] must be a single bit.

  This flag was added in Open vSwitch 2.8.

By itself, the `learn` action can only put two kinds of actions into the flows that it creates: `load` and `output` actions. If `learn` is used in isolation, these are severe limits.

However, `learn` is not meant to be used in isolation. It is a primitive meant to be used together with other Open vSwitch features to accomplish a task. Its existing features are enough to accomplish most tasks.

Here is an outline of a typical pipeline structure that allows for versatile behavior using `learn`:

- Flows in table `A` contain a `learn` action, that populates flows in table `L`, that use a `load` action to populate register `R` with information about what was learned.
- Flows in table `B` contain two sequential resubmit actions: one to table `L` and another one to table `B + 1`.
- Flows in table `B + 1` match on register `R` and act differently depending on what the flows in table `L` loaded into it.

This approach can be used to implement many `learn`-based features. For example:

- Resubmit to a table selected based on learned information, e.g. see https://mail.openvswitch.org/pipermail/ovs-discuss/2016-June/021694.html .

- MAC learning in the middle of a pipeline, as described in the `Open vSwitch Advanced Features Tutorial` in the OVS documentation.
- TCP state based firewalling, by learning outgoing connections based on SYN packets and matching them up with incoming packets. (This is usually better implemented using the `ct` action.)
- At least some of the features described in T. A. Hoff, `Extending Open vSwitch to Facilitate Creation of Stateful SDN Applications`.

Conformance

The `learn` action is an Open vSwitch extension to OpenFlow added in Open vSwitch 1.3. Some features of `learn` were added in later versions, as noted individually above.

The `fin_timeout` action

Syntax:

`fin_timeout(`*key*`=`*value*`...)`

This action changes the idle timeout or hard timeout, or both, of the OpenFlow flow that contains it, when the flow matches a TCP packet with the FIN or RST flag. When such a packet is observed, the action reduces the rule's timeouts to those specified on the action. If the rule's existing timeout is already shorter than the one that the action specifies, then that timeout is unaffected.

The timeouts are specified as key-value pairs:

`idle_timeout=`*seconds*
Causes the flow to expire after the given number of seconds of inactivity.

`hard_timeout=`*seconds*
Causes the flow to expire after the given number of *seconds*, regardless of activity. (*seconds* specifies time since the flow's creation, not since the receipt of the FIN or RST.)

This action is normally added to a learned flow by the `learn` action. It is unlikely to be useful otherwise.

Conformance

This Open vSwitch extension action was added in Open vSwitch 1.6.

Programming and Control Flow Actions

The `resubmit` action

Syntax:

`resubmit:`*port*
`resubmit([`*port*`],[`*table*][,ct])``

Searches an OpenFlow flow table for a matching flow and executes the actions found, if any, before continuing to the following action in the current flow entry. Arguments can customize the search:

- If *port* is given as an OpenFlow port number or name, then it specifies a value to use for the input port metadata field as part of the search, in place of the input port currently in the flow. Specifying `in_port` as `port` is equivalent to omitting it.
- If *table* is given as an integer between 0 and 254 or a table name, it specifies the OpenFlow table to search. If it is not specified, the table from the current flow is used.
- If `ct` is specified, then the search is done with packet 5-tuple fields swapped with the corresponding conntrack original direction tuple fields. See the documentation for `ct` above, for more information about connection tracking, or `ovs-fields(7)` for details about the connection tracking fields.

This flag requires a valid connection tracking state as a match prerequisite in the flow where this action is placed. Examples of valid connection tracking state matches include `ct_state=+new`, `ct_state=+est`, `ct_state=+rel`, and `ct_state=+trk-inv`.

The changes, if any, to the input port and connection tracking fields are just for searching the flow table. The changes are not visible to actions or to later flow table lookups.

The most common use of `resubmit` is to visit another flow table without *port* or `ct`, like this: `resubmit(,`*table*`)`.

Recursive `resubmit` actions are permitted.

Conformance

The `resubmit` action is an Open vSwitch extension. However, the `goto_table` instruction in OpenFlow 1.1 and later can be viewed as a kind of restricted `resubmit`.

Open vSwitch 1.3 added `table`. Open vSwitch 2.7 added `ct`.

Open vSwitch imposes a limit on `resubmit` recursion that varies among version:

- Open vSwitch 1.0.1 and earlier did not support recursion.
- Open vSwitch 1.0.2 and 1.0.3 limited recursion to 8 levels.
- Open vSwitch 1.1 and 1.2 limited recursion to 16 levels.
- Open vSwitch 1.2 through 1.8 limited recursion to 32 levels.
- Open vSwitch 1.9 through 2.0 limited recursion to 64 levels.
- Open vSwitch 2.1 through 2.5 limited recursion to 64 levels and impose a total limit of 4,096 resubmits per flow translation (earlier versions did not impose any total limit).
- Open vSwitch 2.6 and later imposes the same limits as 2.5, with one exception: resubmit from table `x` to any table `y > x` does not count against the recursion depth limit.

The `clone` action

Syntax:

`clone(`*action*`...)`

Executes each nested *action*, saving much of the packet and pipeline state beforehand and then restoring it afterward. The state that is saved and restored includes all flow data and metadata (including, for example, `in_port` and `ct_state`), the stack accessed by `push` and `pop` actions, and the OpenFlow action set.

This action was added in Open vSwitch 2.7.

The `push` and `pop` actions

Syntax:

`push:`*src*
`pop:`*dst*

The `push` action pushes *src* on a general-purpose stack. The `pop` action pops an entry off the stack into *dst*. *src* and *dst* should be fields or subfields in the syntax described under *Field Specifications* above.

Controllers can use the stack for saving and restoring data or metadata around `resubmit` actions, for swapping or rearranging data and metadata, or for other purposes. Any data or metadata field, or part of one, may be pushed, and any modifiable field or subfield may be popped.

The number of bits pushed in a stack entry do not have to match the number of bits later popped from that entry. If more bits are popped from an entry than were pushed, then the entry is conceptually left-padded with 0-bits as needed. If fewer bits are popped than pushed, then bits are conceptually trimmed from the left side of the entry.

The stack's size is limited. The limit is intended to be high enough that `normal` use will not pose problems. Stack overflow or underflow is an error that stops action execution (see `Stack too deep` under *Error Handling*, above).

Examples:

- `push:reg2[0..5]` or `push:NXM_NX_REG2[0..5]` pushes on the stack the 6 bits in register 2 bits 0 through 5.
- `pop:reg2[0..5]` or `pop:NXM_NX_REG2[0..5]` pops the value from top of the stack and copy bits 0 through 5 of that value into bits 0 through 5 of register 2.

Conformance
Open vSwitch 1.2 introduced `push` and `pop` as OpenFlow extension actions.

The `exit` action

Syntax:

`exit`

This action causes Open vSwitch to immediately halt execution of further actions. Actions which have already been executed are unaffected. Any further actions, including those which may be in other tables, or different levels of the `resubmit` call stack, are ignored. However, an `exit` action within a group bucket terminates only execution of that bucket, not other buckets or the overall pipeline. Actions in the action set are still executed (specify `clear_actions` before `exit` to discard them).

The `multipath` action

Syntax:

`multipath(`*fields*`,`*basis*`,`*algorithm*`,`*n_links*`,`*arg*`,`*dst*`)`

Hashes *fields* using *basis* as a universal hash parameter, then the applies multipath link selection *algorithm* (with parameter *arg*) to choose one of *n_links* output links numbered 0 through *n_links* minus 1, and stores the link into *dst*, which must be a field or subfield in the syntax described under *Field Specifications* above.

The `bundle` or `bundle_load` actions are usually easier to use than `multipath`.

fields must be one of the following:

`eth_src`
Hashes Ethernet source address only.

`symmetric_l4`
Hashes Ethernet source, destination, and type, VLAN ID, IPv4/IPv6 source, destination, and protocol, and TCP or SCTP (but not UDP) ports. The hash is computed so that pairs of corresponding flows in each direction hash to the same value, in environments where L2 paths are the same in each direction. UDP ports are not included in the hash to support protocols such as VXLAN that use asymmetric ports in each direction.

`symmetric_l3l4`
Hashes IPv4/IPv6 source, destination, and protocol, and TCP or SCTP (but not UDP) ports. Like `symmetric_l4`, this is a symmetric hash, but by excluding L2 headers it is more effective in environments with asymmetric L2 paths (e.g. paths involving VRRP IP addresses on a router). Not an effective hash function for protocols other than IPv4 and IPv6, which hash to a constant zero.

`symmetric_l3l4+udp`
Like `symmetric_l3l4+udp`, but UDP ports are included in the hash. This is a more effective hash when asymmetric UDP protocols such as VXLAN are not a consideration.

`symmetric_l3`
Hashes network source address and network destination address.

nw_src
Hashes network source address only.

nw_dst
Hashes network destination address only.

The *algorithm* used to compute the final result `link` must be one of the following:

modulo_n
Computes `link = hash(flow) % n_links`.

This algorithm redistributes all traffic when `n_links` changes. It has `O(1)` performance.

Use 65535 for `max_link` to get a raw hash value.

This algorithm is specified by RFC 2992.

hash_threshold
Computes `link = hash(flow) / (MAX_HASH / n_links)`.

Redistributes between one-quarter and one-half of traffic when `n_links` changes. It has `O(1)` performance.

This algorithm is specified by RFC 2992.

hrw (Highest Random Weight)
Computes the following:

```
for i in [0, n_links]:
    weights[i] = hash(flow, i)
link = { i such that weights[i] >= weights[j] for all j != i }
```

Redistributes `1 / n_links` of traffic when `n_links` changes. It has `O(n_links)` performance. If `n_links` is greater than a threshold (currently 64, but subject to change), Open vSwitch will substitute another algorithm automatically.

This algorithm is specified by RFC 2992.

iter_hash (Iterative Hash)
Computes the following:

```
i = 0
repeat:
    i = i + 1
    link = hash(flow, i) % arg
while link > max_link
```

Redistributes `1 / n_links` of traffic when `n_links` changes. `O(1)` performance when `arg / max_link` is bounded by a constant.

Redistributes all traffic when `arg` changes.

arg must be greater than `max_link` and for best performance should be no more than approximately `max_link * 2`. If *arg* is outside the acceptable range, Open vSwitch will automatically substitute the least power of 2 greater than `max_link`.

This algorithm is specific to Open vSwitch.

Only the `iter_hash` algorithm uses *arg*.

It is an error if `max_link` is greater than or equal to `2**n_bits`.

Conformance
This is an OpenFlow extension added in Open vSwitch 1.1.

Other Actions

The `conjunction` action

Syntax:

`conjunction`(*id*, *k*/*n*)

This action allows for sophisticated `conjunctive match` flows. Refer to `Conjunctive Match Fields` in `ovs-fields(7)` for details.

A flow that has one or more `conjunction` actions may not have any other actions except for `note` actions.

Conformance
Open vSwitch 2.4 introduced the `conjunction` action and `conj_id` field. They are Open vSwitch extensions to OpenFlow.

The `note` action

Syntax:

`note:`[*hh*]`...`

This action does nothing at all. OpenFlow controllers may use it to annotate flows with more data than can fit in a flow cookie.

The action may include any number of bytes represented as hex digits *hh*. Periods may separate pairs of hex digits, for readability. The `note` action's format doesn't include an exact length for its payload, so the provided bytes will be padded on the right by enough bytes with value 0 to make the total number 6 more than a multiple of 8.

Conformance
This action is an extension to OpenFlow introduced in Open vSwitch 1.1.

The `sample` action

Syntax:

`sample`(*argument*`...`)

Samples packets and sends one sample for every sampled packet.

The following *argument* forms are accepted:

`probability`=*packets*
The number of sampled packets out of 65535. Must be greater or equal to 1.

`collector_set_id`=*id*
The unsigned 32-bit integer identifier of the set of sample collectors to send sampled packets to. Defaults to 0.

`obs_domain_id`=*value*
When sending samples to IPFIX collectors, the unsigned 32-bit integer Observation Domain ID sent in every IPFIX flow record. The *value* may be specified as a 32-bit integer or a field or subfield in the syntax described under *Field Specifications* above. Defaults to 0.

`obs_point_id`=*value*
When sending samples to IPFIX collectors, the unsigned 32-bit integer Observation Point ID sent in every IPFIX flow record. The *value* may be specified as a 32-bit integer or a field or subfield in the syntax described under *Field Specifications* above. Defaults to 0.

`sampling_port=`*port*
: Sample packets on *port*, which should be the ingress or egress port. This option, which was added in Open vSwitch 2.6, allows the IPFIX implementation to export egress tunnel information.

`ingress`

`egress`
: Specifies explicitly that the packet is being sampled on ingress to or egress from the switch. IPFIX reports sent by Open vSwitch before version 2.6 did not include a direction. From 2.6 until 2.7, IPFIX reports inferred a direction from `sampling_port`: if it was the packet's output port, then the direction was reported as egress, otherwise as ingress. Open vSwitch 2.7 introduced these options, which allow the inferred direction to be overridden. This is particularly useful when the ingress (or egress) port is not a tunnel.

Refer to `ovs-vswitchd.conf.db(5)` for more details on configuring sample collector sets.

Conformance
: This action is an OpenFlow extension added in Open vSwitch 2.4.

  Support for subfields in *obs_domain_id* and *obs_point_id* was added in Open vSwitch 3.4.

Instructions

Every version of OpenFlow includes actions. OpenFlow 1.1 introduced the higher-level, related concept of `instructions`. In OpenFlow 1.1 and later, actions within a flow are always encapsulated within an instruction. Each flow has at most one instruction of each kind, which are executed in the following fixed order defined in the OpenFlow specification:

1. `Meter`
2. `Apply-Actions`
3. `Clear-Actions`
4. `Write-Actions`
5. `Write-Metadata`
6. `Stat-Trigger` (not supported by Open vSwitch)
7. `Goto-Table`

The most important instruction is `Apply-Actions`. This instruction encapsulates any number of actions, which the instruction executes. Open vSwitch does not explicitly represent `Apply-Actions`. Instead, any action by itself is implicitly part of an `Apply-Actions` instructions.

Open vSwitch syntax requires other instructions, if present, to be in the order listed above. Otherwise it will flag an error.

The `meter` action and instruction

Syntax:

`meter:`*meter_id*

Apply meter *meter_id*. If a meter band rate is exceeded, the packet may be dropped, or modified, depending on the meter band type.

Conformance
: OpenFlow 1.3 introduced the `meter` instruction. OpenFlow 1.5 changes `meter` from an instruction to an action.

OpenFlow 1.5 allows implementations to restrict `meter` to be the first action in an action list and to exclude `meter` from action sets, for better compatibility with OpenFlow 1.3 and 1.4. Open vSwitch restricts the `meter` action both ways.

Open vSwitch 2.0 introduced OpenFlow protocol support for meters, but it did not include a datapath implementation. Open vSwitch 2.7 added meter support to the userspace datapath. Open vSwitch 2.10 added meter support to the kernel datapath. Open vSwitch 2.12 added support for meter as an action in OpenFlow 1.5.

The `clear_actions` instruction

Syntax:

`clear_actions`

Clears the action set. See *Action Sets*, above, for more information.

Conformance

OpenFlow 1.1 introduced `clear_actions`. Open vSwitch 2.1 added support for `clear_actions`.

The `write_actions` instruction

Syntax:

`write_actions(`*action*`...)`

Adds each *action* to the action set. The action set is carried between flow tables and then executed at the end of the pipeline. Only certain actions may be written to the action set. See *Action Sets*, above, for more information.

Conformance

OpenFlow 1.1 introduced `write_actions`. Open vSwitch 2.1 added support for `write_actions`.

The `write_metadata` instruction

Syntax:

`write_metadata:`*value*`[/`*mask*`]`

Updates the flow's `metadata` field. If *mask* is omitted, `metadata` is set exactly to *value*; if *mask* is specified, then a 1-bit in *mask* indicates that the corresponding bit in `metadata` will be replaced with the corresponding bit from *value*. Both *value* and *mask* are 64-bit values that are decimal by default; use a `0x` prefix to specify them in hexadecimal.

The `metadata` field can also be matched in the flow table and updated with actions such as `set_field` and `move`.

Conformance

OpenFlow 1.1 introduced `write_metadata`. Open vSwitch 2.1 added support for `write_metadata`.

The `goto_table` instruction

Syntax:

`goto_table:`*table*

Jumps to *table* as the next table in the process pipeline. The table may be a number between 0 and 254 or a table name.

It is an error if *table* is less than or equal to the table of the flow that contains it; that is, `goto_table` must move forward in the OpenFlow pipeline. Since `goto_table` must be the last instruction in a flow, it never leads to recursion. The `resubmit` extension action is more flexible.

Conformance

OpenFlow 1.1 introduced `goto_table`. Open vSwitch 2.1 added support for `goto_table`.

6.1.2 ovs-appctl

Synopsis

`ovs-appctl` [`--target`=*target* | `-t` *target*] [`--timeout`=*secs* | `-T` *secs*] [`--format`=*format* | `-f` *format*] [`--pretty`] *command* [*arg* `...`]

`ovs-appctl --help`

`ovs-appctl --version`

Description

Open vSwitch daemons accept certain commands at runtime to control their behavior and query their settings. Every daemon accepts a common set of commands documented under *Common Commands* below. Some daemons support additional commands documented in their own manpages. `ovs-vswitchd` in particular accepts a number of additional commands documented in `ovs-vswitchd(8)`.

The `ovs-appctl` program provides a simple way to invoke these commands. The command to be sent is specified on `ovs-appctl`'s command line as non-option arguments. `ovs-appctl` sends the command and prints the daemon's response on standard output.

In normal use only a single option is accepted:

- `-t` *target* or `--target`=*target*

  Tells `ovs-appctl` which daemon to contact.

  If *target* begins with `/` it must name a Unix domain socket on which an Open vSwitch daemon is listening for control channel connections. By default, each daemon listens on a Unix domain socket in the rundir (e.g. `/run`) named `<program>.<pid>.ctl`, where <program> is the program's name and <pid> is its process ID. For example, if `ovs-vswitchd` has PID 123, it would listen on `ovs-vswitchd.123.ctl`.

  Otherwise, `ovs-appctl` looks in the rundir for a pidfile, that is, a file whose contents are the process ID of a running process as a decimal number, named *target*`.pid`. (The `--pidfile` option makes an Open vSwitch daemon create a pidfile.) `ovs-appctl` reads the pidfile, then looks in the rundir for a Unix socket named *target*`.<pid>.ctl`, where <pid> is replaced by the process ID read from the pidfile, and uses that file as if it had been specified directly as the target.

- `-T` *secs* or `--timeout`=*secs*

  By default, or with a *secs* of `0`, `ovs-appctl` waits forever to connect to the daemon and receive a response. This option limits runtime to approximately *secs* seconds. If the timeout expires, `ovs-appctl` exits with a `SIGALRM` signal.

- `-f` *format* or `--format`=*format*

  Tells `ovs-appctl` which output format to use. By default, or with a *format* of `text`, `ovs-appctl` will print plain-text for humans. When *format* is `json`, `ovs-appctl` will return a JSON document. When `json` is requested, but a command has not implemented JSON output, the plain-text output will be wrapped in a provisional JSON document with the following structure:

  ```
  {"reply-format":"plain","reply":"$PLAIN_TEXT_HERE"}
  ```

- `--pretty`

  By default, JSON output is printed as compactly as possible. This option causes JSON in output to be printed in a more readable fashion. For example, members of objects and elements of arrays are printed one per line, with indentation. Requires `--format=json`.

Common Commands

Every Open vSwitch daemon supports a common set of commands, which are documented in this section.

General Commands

These commands display daemon-specific commands and the running version. Note that these commands are different from the `--help` and `--version` options that return information about the `ovs-appctl` utility itself.

- `list-commands`

  Lists the commands supported by the target.

- `version`

  Displays the version and compilation date of the target.

Logging Commands

Open vSwitch has several log levels. The highest-severity log level is:

- `off`

  No message is ever logged at this level, so setting a logging destination's log level to `off` disables logging to that destination.

The following log levels, in order of descending severity, are available:

- `emer`

  A major failure forced a process to abort.

- `err`

  A high-level operation or a subsystem failed. Attention is warranted.

- `warn`

  A low-level operation failed, but higher-level subsystems may be able to recover.

- `info`

  Information that may be useful in retrospect when investigating a problem.

- `dbg`

  Information useful only to someone with intricate knowledge of the system, or that would commonly cause too-voluminous log output. Log messages at this level are not logged by default.

Every Open vSwitch daemon supports the following commands for examining and adjusting log levels:

- `vlog/list`

  Lists the known logging modules and their current levels.

- `vlog/list-pattern`

  Lists logging pattern used for each destination.

- `vlog/set` [*spec*]

  Sets logging levels. Without any *spec*, sets the log level for every module and destination to `dbg`. Otherwise, *spec* is a list of words separated by spaces or commas or colons, up to one from each category below:

  - A valid module name, as displayed by the `vlog/list` command on `ovs-appctl(8)`, limits the log level change to the specified module.

  - `syslog`, `console`, or `file`, to limit the log level change to only to the system log, to the console, or to a file, respectively.
  - `off`, `emer`, `err`, `warn`, `info`, or `dbg`, to control the log level. Messages of the given severity or higher will be logged, and messages of lower severity will be filtered out. `off` filters out all messages.

  Case is not significant within *spec*.

  Regardless of the log levels set for `file`, logging to a file will not take place unless the target application was invoked with the `--log-file` option.

  For compatibility with older versions of OVS, `any` is accepted within *spec* but it has no effect.

- `vlog/set PATTERN:`*destination*`:`*pattern*

  Sets the log pattern for *destination* to *pattern*. Each time a message is logged to *destination*, *pattern* determines the message's formatting. Most characters in *pattern* are copied literally to the log, but special escapes beginning with `%` are expanded as follows:

  - `%A`

    The name of the application logging the message, e.g. `ovs-vswitchd`.

  - `%B`

    The RFC5424 syslog PRI of the message.

  - `%c`

    The name of the module (as shown by `ovs-appctl --list`) logging the message.

  - `%d`

    The current date and time in ISO 8601 format (`YYYY-MM-DD HH:MM:SS`).

  - `%d{`*format*`}`

    The current date and time in the specified *format*, which takes the same format as the `template` argument to `strftime(3)`. As an extension, any `#` characters in *format* will be replaced by fractional seconds, e.g. use `%H:%M:%S.###` for the time to the nearest millisecond. Sub-second times are only approximate and currently decimal places after the third will always be reported as zero.

  - `%D`

    The current UTC date and time in ISO 8601 format (`YYYY-MM-DD HH:MM:SS`).

  - `%D{`*format*`}`

    The current UTC date and time in the specified *format*, which takes the same format as the `template` argument to `strftime(3)`. Supports the same extension for sub-second resolution as `%d{...}`.

  - `%E`

    The hostname of the node running the application.

  - `%m`

    The message being logged.

  - `%N`

    A serial number for this message within this run of the program, as a decimal number. The first message a program logs has serial number 1, the second one has serial number 2, and so on.

  - `%n`

    A new-line.

- `%p`

  The level at which the message is logged, e.g. `DBG`.

- `%P`

  The program's process ID (pid), as a decimal number.

- `%r`

  The number of milliseconds elapsed from the start of the application to the time the message was logged.

- `%t`

  The subprogram name, that is, an identifying name for the process or thread that emitted the log message, such as `monitor` for the process used for `--monitor` or `main` for the primary process or thread in a program.

- `%T`

  The subprogram name enclosed in parentheses, e.g. `(monitor)`, or the empty string for the primary process or thread in a program.

- `%%`

  A literal `%`.

A few options may appear between the `%` and the format specifier character, in this order:

- `-`

  Left justify the escape's expansion within its field width. Right justification is the default.

- `0`

  Pad the field to the field width with `0` characters. Padding with spaces is the default.

- *width*

  A number specifies the minimum field width. If the escape expands to fewer characters than *width* then it is padded to fill the field width. (A field wider than *width* is not truncated to fit.)

The default pattern for console and file output is `%D{%Y-%m-%dT %H:%M:%SZ}|%05N|%c|%p|%m`; for syslog output, `%05N|%c|%p|%m`.

Daemons written in Python (e.g. `ovs-monitor-ipsec`) do not allow control over the log pattern.

- `vlog/set FACILITY:`*facility*

  Sets the RFC5424 facility of the log message. *facility* can be one of `kern`, `user`, `mail`, `daemon`, `auth`, `syslog`, `lpr`, `news`, `uucp`, `clock`, `ftp`, `ntp`, `audit`, `alert`, `clock2`, `local0`, `local1`, `local2`, `local3`, `local4`, `local5`, `local6` or `local7`.

- `vlog/close`

  Causes the daemon to close its log file, if it is open. (Use `vlog/reopen` to reopen it later.)

- `vlog/reopen`

  Causes the daemon to close its log file, if it is open, and then reopen it. (This is useful after rotating log files, to cause a new log file to be used.)

  This has no effect if the target application was not invoked with the `--log-file` option.

Options

-h, **--help**

Prints a brief help message to the console.

-V, **--version**

Prints version information to the console.

See Also

`ovs-appctl` can control all Open vSwitch daemons, including `ovs-vswitchd(8)` and `ovsdb-server(1)`.

6.1.3 ovs-ctl

Synopsis

```
ovs-ctl --system-id=random|<uuid> [<options>] start
ovs-ctl stop
ovs-ctl --system-id=random|<uuid> [<options>] restart
ovs-ctl status
ovs-ctl version
ovs-ctl [<options>] load-kmod
ovs-ctl --system-id=random|<uuid> [<options>] force-reload-kmod
ovs-ctl [--protocol=<protocol>] [--sport=<sport>] [--dport=<dport>] enable-protocol
ovs-ctl delete-transient-ports
ovs-ctl help | -h | --help
ovs-ctl --version
```

Description

The `ovs-ctl` program starts, stops, and checks the status of Open vSwitch daemons. It is not meant to be invoked directly by system administrators but to be called internally by system startup scripts.

Each `ovs-ctl` command is described separately below.

The `start` command

The `start` command starts Open vSwitch. It performs the following tasks:

1. Loads the Open vSwitch kernel module. If this fails, and the Linux bridge module is loaded but no bridges exist, it tries to unload the bridge module and tries loading the Open vSwitch kernel module again. (This is because the Open vSwitch kernel module cannot coexist with the Linux bridge module before 2.6.37.)

The `start` command skips the following steps if `ovsdb-server` is already running:

2. If the Open vSwitch database file does not exist, it creates it. If the database does exist, but it has an obsolete version, it upgrades it to the latest schema.
3. Starts `ovsdb-server`, unless the `--no-ovsdb-server` command option is given.
4. Initializes a few values inside the database.
5. If the `--delete-bridges` option was used, deletes all of the bridges from the database.

6. If the `--delete-transient-ports` option was used, deletes all ports that have `other_config:transient` set to true.

The `start` command skips the following step if `ovs-vswitchd` is already running, or if the `--no-ovs-vswitchd` command option is given:

7. Starts `ovs-vswitchd`.

Options

Several command-line options influence the `start` command's behavior. Some form of the following option should ordinarily be specified:

- `--system-id=<uuid>` or `--system-id=random`

  This specifies a unique system identifier to store into `external-ids:system-id` in the database's `Open_vSwitch` table. Remote managers that talk to the Open vSwitch database server over network protocols use this value to identify and distinguish Open vSwitch instances, so it should be unique (at least) within OVS instances that will connect to a single controller.

  When `random` is specified, `ovs-ctl` will generate a random ID that persists from one run to another (stored in a file). When another string is specified `ovs-ctl` uses it literally.

The following options should be specified if the defaults are not suitable:

- `--system-type=<type>` or `--system-version=<version>`

  Sets the value to store in the `system-type` and `system-version` columns, respectively, in the database's `Open_vSwitch` table. Remote managers may use these values too determine the kind of system to which they are connected (primarily for display to human administrators).

  When not specified, `ovs-ctl` uses values from the optional `system-type.conf` and `system-version.conf` files (see *Files*) or it uses the `lsb_release` program, if present, to provide reasonable defaults.

The following options are also likely to be useful:

- `--external-id="<name>=<value>"`

  Sets `external-ids:<name>` to <value> in the database's `Open_vSwitch` table. Specifying this option multiple times adds multiple key-value pairs.

- `--delete-bridges`

  Ordinarily Open vSwitch bridges persist from one system boot to the next, as long as the database is preserved. Some environments instead expect to re-create all of the bridges and other configuration state on every boot. This option supports that, by deleting all Open vSwitch bridges after starting `ovsdb-server` but before starting `ovs-vswitchd`.

- `--delete-transient-ports`

  Deletes all ports that have `other_config:transient` set to `true`. This is important on certain environments where some ports are going to be recreated after reboot, but other ports need to be persisted in the database.

- `--ovs-user=user[:group]`

  Ordinarily Open vSwitch daemons are started as the user invoking the ovs-ctl command. Some system administrators would prefer to have the various daemons spawn as different users in their environments. This option allows passing the `--user` option to the `ovsdb-server` and `ovs-vswitchd` daemons, allowing them to change their privilege levels.

The following options are less important:

- `--no-monitor`

  By default `ovs-ctl` passes `--monitor` to `ovs-vswitchd` and `ovsdb-server`, requesting that it spawn a process monitor which will restart the daemon if it crashes. This option suppresses that behavior.

- `--daemon-cwd=<directory>`

  Specifies the current working directory that the OVS daemons should run from. The default is `/` (the root directory) if this option is not specified. (This option is useful because most systems create core files in a process's current working directory and because a file system that is in use as a process's current working directory cannot be unmounted.)

- `--no-force-corefiles`

  By default, `ovs-ctl` enables core dumps for the OVS daemons. This option disables that behavior.

- `--no-mlockall`

  By default `ovs-ctl` passes `--mlockall` to `ovs-vswitchd`, requesting that it lock all of its virtual memory on page fault (on allocation, when running on Linux kernel 4.4 and older), preventing it from being paged to disk. This option suppresses that behavior.

- `--no-self-confinement`

  Disable self-confinement for `ovs-vswitchd` and `ovsdb-server` daemons. This flag may be used when, for example, OpenFlow controller creates its Unix Domain Socket outside OVS run directory and OVS needs to connect to it. It is better to stick with the default behavior and not to use this flag, unless:

  - You have Open vSwitch running under SELinux or AppArmor Mandatory Access Control that would prevent OVS from messing with sockets outside ordinary OVS directories.
  - You believe that relying on protocol handshakes (e.g. OpenFlow) is enough to prevent OVS to adversely interact with other daemons running on your system.
  - You don't have much worries of remote OVSDB exploits in the first place, because, perhaps, OVSDB manager is running on the same host as OVS and share similar attack vectors.

- `--oom-score=<score>`

  Sets the Linux Out-Of-Memory (OOM) killer score for the OVS daemon after it's been started.

- `--ulimit-core=<LIMIT>`

  Sets ulimit core file size for the OVS daemon after it's been started.

- `--ovsdb-server-priority=<niceness>` or `--ovs-vswitchd-priority=<niceness>`

  Sets the `nice(1)` level used for each daemon. All of them default to `-10`.

- `--ovsdb-server-wrapper=<wrapper>` or `--ovs-vswitchd-wrapper=<wrapper>`

  Configures the specified daemon to run under `<wrapper>`, which is one of the following:

  - `valgrind`: Run the daemon under `valgrind(1)`, if it is installed, logging to `<daemon>.valgrind.log.<pid>` in the log directory.
  - `strace`: Run the daemon under `strace(1)`, if it is installed, logging to `<daemon>.strace.log.<pid>` in the log directory.
  - `glibc`: Enable GNU C library features designed to find memory errors.

  By default, no wrapper is used.

  Each of the wrappers can expose bugs in Open vSwitch that lead to incorrect operation, including crashes. The `valgrind` and `strace` wrappers greatly slow daemon operations so they should not be used in production. They

also produce voluminous logs that can quickly fill small disk partitions. The `glibc` wrapper is less resource-intensive but still somewhat slows the daemons.

The following options control file locations. They should only be used if the default locations cannot be used. See `FILES`, below, for more information.

- `--db-file=<file>`

  Overrides the file name for the OVS database.

- `--db-sock=<socket>`

  Overrides the file name for the Unix domain socket used to connect to `ovsdb-server`.

- `--db-schema=<schema>`

  Overrides the file name for the OVS database schema.

- `--extra-dbs=<file>`

  Adds <file> as an extra database for `ovsdb-server` to serve out. Multiple space-separated file names may also be specified. <file> should begin with `/`; if it does not, then it will be taken as relative to <dbdir>.

The `stop` command

The `stop` command stops the `ovs-vswitchd` and `ovsdb-server` daemons. It does not unload the Open vSwitch kernel modules. It can take the same `--no-ovsdb-server` and `--no-ovs-vswitchd` options as that of the `start` command.

This command does nothing and finishes successfully if the OVS daemons aren't running.

The `restart` command

The `restart` command performs a `stop` followed by a `start` command. The command can take the same options as that of the `start` command. In addition, it saves and restores OpenFlow flows for each individual bridge.

The `status` command

The `status` command checks whether the OVS daemons `ovs-vswitchd` and `ovsdb-server` are running and prints messages with that information. It exits with status 0 if the daemons are running, 1 otherwise.

The `version` command

The `version` command runs `ovsdb-server --version` and `ovs-vswitchd --version`.

The `force-reload-kmod` command

The `force-reload-kmod` command allows upgrading the Open vSwitch kernel module without rebooting. It performs the following tasks:

1. Gets a list of OVS "internal" interfaces, that is, network devices implemented by Open vSwitch. The most common examples of these are bridge "local ports".
2. Saves the OpenFlow flows of each bridge.
3. Stops the Open vSwitch daemons, as if by a call to `ovs-ctl stop`.
4. Saves the kernel configuration state of the OVS internal interfaces listed in step 1, including IP and IPv6 addresses and routing table entries.
5. Unloads the Open vSwitch kernel module (including the bridge compatibility module if it is loaded).

6. Starts OVS back up, as if by a call to `ovs-ctl start`. This reloads the kernel module, restarts the OVS daemons and finally restores the saved OpenFlow flows.
7. Restores the kernel configuration state that was saved in step 4.
8. Checks for daemons that may need to be restarted because they have packet sockets that are listening on old instances of Open vSwitch kernel interfaces and, if it finds any, prints a warning on stdout. DHCP is a common example: if the ISC DHCP client is running on an OVS internal interface, then it will have to be restarted after completing the above procedure. (It would be nice if `ovs-ctl` could restart daemons automatically, but the details are far too specific to a particular distribution and installation.)

`force-kmod-reload` internally stops and starts OVS, so it accepts all of the options accepted by the `start` command except for the `--no-ovs-vswitchd` option.

The `load-kmod` command

The `load-kmod` command loads the openvswitch kernel modules if they are not already loaded. This operation also occurs as part of the `start` command. The motivation for providing the `load-kmod` command is to allow errors when loading modules to be handled separately from other errors that may occur when running the `start` command.

By default the `load-kmod` command attempts to load the `openvswitch` kernel module.

The `enable-protocol` command

The `enable-protocol` command checks for rules related to a specified protocol in the system's `iptables(8)` configuration. If there are no rules specifically related to that protocol, then it inserts a rule to accept the specified protocol.

More specifically:

- If `iptables` is not installed or not enabled, this command does nothing, assuming that lack of filtering means that the protocol is enabled.
- If the `INPUT` chain has a rule that matches the specified protocol, then this command does nothing, assuming that whatever rule is installed reflects the system administrator's decisions.
- Otherwise, this command installs a rule that accepts traffic of the specified protocol.

This command normally completes successfully, even if it does nothing. Only the failure of an attempt to insert a rule normally causes it to return an exit code other than 0.

The following options control the protocol to be enabled:

- `--protocol=<protocol>`

  The name of the IP protocol to be enabled, such as `gre` or `tcp`. The default is `gre`.

- `--sport=<sport>` or `--dport=<dport>`

  TCP or UDP source or destination port to match. These are optional and allowed only with `--protocol=tcp` or `--protocol=udp`.

The `delete-transient-ports` command

Deletes all ports that have the `other_config:transient` value set to true.

The `help` command

Prints a usage message and exits successfully.

Options

In addition to the options listed for each command above, these options control the behavior of several `ovs-ctl` commands.

By default, `ovs-ctl` controls the `ovsdb-server` and `ovs-vswitchd` daemons. The following options restrict that control to exclude one or the other:

- `--no-ovsdb-server`

  Specifies that the `ovs-ctl` commands `start`, `stop`, and `restart` should not modify the running status of `ovsdb-server`.

- `--no-ovs-vswitchd`

  Specifies that the `ovs-ctl` commands `start`, `stop`, and `restart` should not modify the running status of `ovs-vswitchd`. It is an error to include this option with the `force-reload-kmod` command.

Exit Status

`ovs-ctl` exits with status 0 on success and nonzero on failure. The `start` command is considered to succeed if OVS is already started; the `stop` command is considered to succeed if OVS is already stopped.

Environment

The following environment variables affect `ovs-ctl`:

- `PATH`

  `ovs-ctl` does not hardcode the location of any of the programs that it runs. `ovs-ctl` will add the <sbindir> and <bindir> that were specified at `configure` time to `PATH`, if they are not already present.

- `OVS_LOGDIR`, `OVS_RUNDIR`, `OVS_DBDIR`, `OVS_SYSCONFDIR`, `OVS_PKGDATADIR`, `OVS_BINDIR`, `OVS_SBINDIR`

  Setting one of these variables in the environment overrides the respective `configure` option, both for `ovs-ctl` itself and for the other Open vSwitch programs that it runs.

Files

`ovs-ctl` uses the following files:

- `ovs-lib`

  Shell function library used internally by `ovs-ctl`. It must be installed in the same directory as `ovs-ctl`.

- `<logdir>/<daemon>.log`

  Per-daemon logfiles.

- `<rundir>/<daemon>.pid`

  Per-daemon pidfiles to track whether a daemon is running and with what process ID.

- `<pkgdatadir>/vswitch.ovsschema`

  The OVS database schema used to initialize the database (use `--db-schema` to override this location).

- `<dbdir>/conf.db`

  The OVS database (use `--db-file` to override this location).

- `<rundir>/openvswitch/db.sock`

  The Unix domain socket used for local communication with `ovsdb-server` (use `--db-sock` to override this location).

- <sysconfdir>/openvswitch/system-id.conf

  The persistent system UUID created and read by --system-id=random.

- <sysconfdir>/openvswitch/system-type.conf and <sysconfdir>/openvswitch/system-version.conf

  The system-type and system-version values stored in the database's Open_vSwitch table when not specified as a command-line option.

Example

The file debian/openvswitch-switch.init in the Open vSwitch source distribution is a good example of how to use ovs-ctl.

See Also

README.rst, ovsdb-server(8), ovs-vswitchd(8).

6.1.4 ovs-flowviz

Synopsis

ovs-flowviz [-i [*alias*,]*file* | --input [*alias*,]*file*] [-c *file* | --config *file*] [-f *filter* | --filter *filter*] [-l *filter* | --highlight *filter*] [--style *style*] *flow-type format* [*args*...]

ovs-flowviz --help

Description

ovs-flowviz helps visualize OpenFlow and datapath flow dumps in different formats in order to make them more easily understood.

ovs-flowviz reads flows from stdin or from a *file* specified by the --input option, filters them, highlights them, and finally outputs them in one of the predefined *format*s.

Options

-h, **--help**

Print a brief help message to the console.

-i [<alias>,]<file>, **--input** [<alias>,]<file>

File to read flows from. If not provided, ovs-flowviz will read flows from stdin.

This option can be specified multiple times. The file path can prepended by an alias that will be shown in the output. For example: --input node1,/path/to/file1 --input node2,/path/to/file2

-c <file>, **--config** <file>

Style configuration file to use, overriding the default one. Styles defined in the style configuration file can be selected using the --style option.

For more details on the style configuration file, see the *Style Configuration File* section below.

-f <filter>, **--filter** <filter>

Flow filter expression. Only those flows matching the expression will be shown (although some formats implement filtering differently, see the *Datapath tree format* section below).

The filtering syntax is detailed in *Filtering Syntax*.

-l <filter>, **--highlight** <filter>

Highlight the flows that match the provided *filter* expression.

The filtering syntax is detailed in *Filtering Syntax*.

--style <style>

Style. The selected *style* must be defined in the style configuration file.

flow-type

openflow or **datapath**.

format

See the *Supported formats* section.

Supported formats

`ovs-flowviz` supports several visualization formats for both OpenFlow and datapath flows:

Flow Type	Format	Description
Both	console	Prints the flows in a configurable, colorful style in the console.
Both	json	Prints the flows in JSON format.
Both	html	Prints the flows in an HTML list.
OpenFlow	cookie	Prints the flows in the console sorted by cookie.
OpenFlow	logic	Prints the logical structure of flows in the console.
Datapath	tree	Prints the flows as a tree structure arranged by `recirc_id` and `in_port`.
Datapath	graph	Prints a graphviz graph of the flows arranged by `recirc_id` and `in_port`.

Console format

The `console` format works for both OpenFlow and datapath flow types, and prints flows in the terminal using the style determined by the `--style` option.

Arguments:

-h, **--heat-map**

Color of the packet and byte counters to reflect their relative size. The color gradient goes through the following colors: blue (coldest, lowest), cyan, green, yellow, red (hottest, highest)

Note filtering is applied before the range is calculated.

JSON format

The `json` format works for both OpenFlow and datapath flow types, and prints flows in JSON format. See the *JSON Syntax* section for more details.

HTML format

The `html` format works for both OpenFlow and datapath flows, and prints flows in an HTML table that offers some basic interactivity. OpenFlow flows are sorted in tables and datapath flows are arranged in flow trees (see *Datapath tree format* for more details).

Styles defined via Style Configuration File and selected via `--style` option also apply to the `html` format.

OpenFlow cookie format

The OpenFlow `cookie` format is similar to the `console` format but instead of arranging the flows by table, it arranges the flows by cookie.

OpenFlow logic format

The OpenFlow `logic` format helps visualize the logic structure of OpenFlow pipelines by arranging flows into *logical blocks*. A logical block is a set of flows that have:

- Same `priority`.
- Match on the same fields (regardless of the match value and mask).
- Execute the same actions (regardless of the actions' arguments, except for resubmit and output).
- Optionally, the `cookie` can be included as part of the logical flow.

Arguments:

-s, **--show-flows**
: Show all the flows under each logical block.

-d, **--ovn-detrace**
: Use ovn-detrace.py script to extract cookie information (implies '-c').

-c, **--cookie**
: Consider the cookie in the logical block.

--ovn-detrace-path <path>
: Use an alternative path to search for ovn_detrace.py.

--ovnnb-db <conn>
: OVN NB database connection method (implies '-d'). Default: "unix:/var/run/ovn/ovnnb_db.sock".

--ovnsb-db <conn>
: OVN SB database connection method (implies '-d'). Default: "unix:/var/run/ovn/ovnsb_db.sock".

--o <filter>, **--ovn-filter** <filter>
: Specify the filter to be run on the ovn-detrace information. Syntax: python regular expression (See https://docs.python.org/3/library/re.html).

-h, **--heat-map**
: Change the color of the packet and byte counters to reflect their relative size. The color gradient goes through the following colors: blue (coldest, lowest), cyan, green, yellow, red (hottest, highest)

  Note filtering is applied before the range is calculated.

Datapath tree format

The datapath `tree` format arranges datapath flows in a hierarchical tree. The tree is comprised of blocks with the same `recirc_id` and `in_port`. Within those blocks, flows with the same action are combined. And matches which are the same are omitted to reduce the visual noise.

When a flow's actions includes the `recirc()` action with a specific `recirc_id`, flows matching on that `recirc_id` and the same `in_port` are listed below. This is done recursively for all actions.

The result is a hierarchical representation that shows how actions are related to each other via recirculation. Note that flows with a specific non-zero `recirc_id` are listed below each group of flows that have a corresponding `recirc()` action. Therefore, the output contains duplicated flows and can be verbose.

Filtering works in a slightly different way for datapath flow trees. Unlike other formats where a filter simply removes non-matching flows, the output of a filtered datapath flow tree will show full sub-trees that contain at least one flow that satisfies the filter.

The `html` format prints this same tree as an interactive HTML table and the `graph` format shows the same tree as a graphviz graph.

Datapath graph format

The datapath `graph` generates a graphviz visual representation of the same tree-like flow hierarchy that the `tree` format prints.

Arguments:

-h, **--html**
: Print the graphviz format as an svg image alongside an interactive HTML table of flows.

JSON Syntax

Printing a single-file OpenFlow or datapath dump without PMD thread blocks in `json` format results in a list of JSON objects, each representing a flow.

This list can be found inside one or more levels of JSON dictionaries if multiple files are processed (filename used as key) or if PMD thread blocks are found in datapath flows (name of the thread used as key).

Each flow object includes the following keys:

orig
: Original flow string.

info
: Object with the flow information such as: cookie, duration, table, n_packets, n_bytes, etc.

match
: Object with the flow match. For each match, the object contains a key-value where the key is the name of the match as defined in `ovs-fields(7)` and `ovs-ofctl(8)`, and the value represents the match value. The way each value is represented depends on its type. See *Value representation*.

actions
: List of action objects. Each action is represented by an JSON object that has one key and one value. The key corresponds to the action name. The value represents the arguments of the key. See *Action representation*.

ufid
: The UFID (datapath flows only).

Value representation

Values are represented differently depending on their type:

- Flags: The value of flags is `true`.
- Decimal / Hexadecimal: Represented by their integer value. If they support masking, represented by a dictionary with two keys: `value` contains the field value and `mask` contains the mask. Both are integers.
- Ethernet: Represented by a string: `{address}[/{mask}]`
- IPv4 / IPv6: Represented by a string `{address}[/{mask}]`
- Registers: Represented by a dictionary with three keys: `field` contains the field value (string), `start`, and `end` represent the first and last bit of the register value.

For example, the register

```
NXM_NX_REG10[0..15]
```

is represented as

```
{
    "field": "NXM_NX_REG10",
    "start": 0,
    "end": 15
},
```

Action representation

Actions are generally represented by an object that has a single key and value. The key is the action name as defined `ovs-actions(7)`.

The value of actions that have no arguments (such as `drop`) is (boolean) `true`.

The value of actions that have a list of arguments (e.g: `resubmit([port],[table],[ct])`) is an object that has the name of the argument as key. The argument names for each action is defined in ovs-actions. For example, the action

```
resubmit(,10)
```

is represented as

```
{
    "resubmit": {
        "port": "",
        "table": 10
    }
}
```

The value of actions that have a key-word list as arguments (e.g: `ct([argument])`) is an object whose keys correspond to the keys defined in `ovs-actions(7)`. The way values are represented depends on the type of the argument. For example, the action

```
ct(table=14,zone=NXM_NX_REG12[0..15],nat)
```

is represented as

```
{
    "ct": {
        "table": 14,
        "zone": {
            "field": "NXM_NX_REG12",
            "start": 0,
            "end": 15
        },
        "nat": true
    }
}
```

Style Configuration File

The style configuration file is selected via the `--config` option and has INI syntax. It can define any number of styles to be used by both `console` and `html` formats. Once defined in the configuration file, formats are selected using the `--style` option.

INI sections are used to define styles, `[styles.mystyle]` defines a style called *mystle*. Within a section styles can be defined as:

```
[FORMAT].[PORTION].[SELECTOR].[ELEMENT] = [VALUE]
```

FORMAT
Either `console` or `html`

PORTION
Part of the key-value the style applies to: `key` to indicate the key part of a key-value, `value` to indicate the value part of a key-value, `flag` to indicate a single flag or `delim` to indicate delimiters such as parentheses, brackets, etc.

SELECTOR
Select the key-value the style applies to: `highlighted` to indicate highlighted key-values, `type.<type>` to indicate certain types such as `IPAddress` or `EthMask` or `<keyname>` to select a particular key name.

ELEMENT
Select the style element to modify: **color** or **underline** (only for `console` format).

VALUE
Ether a color hex, other color names defined in the rich python library (https://rich.readthedocs.io/en/stable/appendix/colors.html) or **true** if the element is `underline`.

A default configuration file is shipped with `ovs-flowviz` and its path is printed in the `--help` output. A detailed description of the syntax alongside some examples are available there.

Filtering syntax

`ovs-flowviz` provides rich highlighting and filtering. The special command `ovs-flowviz filter` dumps the filtering syntax:

```
$ ovs-flowviz filter
Filter Syntax
*************

   [! | not ] {key}[[.subkey]...] [OPERATOR] {value})] [LOGICAL OPERATOR] ...

  Comparison operators:
      =   equality
      <   less than
      >   more than
      ~=  masking (valid for IP and Ethernet fields)

  Logical operators:
      !{expr}:  NOT
      {expr} && {expr}: AND
      {expr} || {expr}: OR

  Matches and flow metadata:
      To compare against a match or info field, use the field directly, e.g:
```

(continues on next page)

(continued from previous page)

```
          priority=100
          n_bytes>10
      Use simple keywords for flags:
          tcp and ip_src=192.168.1.1

  Actions:
      Actions values might be dictionaries, use subkeys to access individual
      values, e.g:
          output.port=3
      Use simple keywords for flags
          drop

  Examples of valid filters:
      nw_addr~=192.168.1.1 && (tcp_dst=80 || tcp_dst=443)
      arp=true && !arp_tsa=192.168.1.1
      n_bytes>0 && drop=true
```

Example expressions:

```
n_bytes > 0 and drop
nw_src~=192.168.1.1 or arp.tsa=192.168.1.1
! tcp && output.port=2
```

Examples

Print OpenFlow flows sorted by cookie adding OVN data to each one:

```
$ ovs-flowviz -i flows.txt openflow cookie --ovn-detrace
```

Print OpenFlow logical structure, showing the flows and heat-map:

```
$ ovs-flowviz -i flows.txt openflow logic --show-flows --heat-map
```

Display OpenFlow flows in HTML format with “light” style and highlight drops:

```
$ ovs-flowviz -i flows.txt --style "light" --highlight "n_packets > 0 and drop" openflow␣
↪html > flows.html
```

Display the datapath flows in an interactive graphviz + HTML view:

```
$ ovs-flowviz -i flows.txt datapath graph --html > flows.html
```

Display the datapath flow trees that lead to packets being sent to port 10:

```
$ ovs-flowviz -i flows.txt --filter "output.port=10" datapath tree
```

6.1.5 ovs-l3ping

Synopsis

```
ovs-l3ping -s <TunnelRemoteIP>,<InnerIP>[/<mask>] -t <tunnelmode>
```

```
ovs-l3ping -s <TunnelRemoteIP>,<InnerIP>[/<mask>][:<ControlPort>] -t <tunnelmode>
```

```
ovs-l3ping -c <TunnelRemoteIP>,<InnerIP>[/<mask>],<RemoteInnerIP> -t <tunnelmode>
```

```
ovs-l3ping -c <TunnelRemoteIP>,<InnerIP>[/<mask>][:<ControlPort>[:<DataPort>]],
<RemoteInnerIP>[:<ControlPort>[:<DataPort>]] [-b <targetbandwidth>] [-i <testinterval>]
-t <tunnelmode>

ovs-l3ping -h | --help

ovs-l3ping -V | --version
```

Description

The `ovs-l3ping` program may be used to check for problems that could be caused by invalid routing policy, misconfigured firewall in the tunnel path or a bad NIC driver. On one of the nodes, run `ovs-l3ping` in server mode and on the other node run it in client mode. The client and server will establish L3 tunnel, over which client will give further testing instructions. The `ovs-l3ping` client will perform UDP and TCP tests. This tool is different from `ovs-test` that it encapsulates XML/RPC control connection over the tunnel, so there is no need to open special holes in firewall.

UDP tests can report packet loss and achieved bandwidth for various datagram sizes. By default target bandwidth for UDP tests is 1Mbit/s.

TCP tests report only achieved bandwidth, because kernel TCP stack takes care of flow control and packet loss.

Client Mode

An `ovs-l3ping` client will create a L3 tunnel and connect over it to the `ovs-l3ping` server to schedule the tests. <TunnelRemoteIP> is the peer's IP address, where tunnel will be terminated. <InnerIP> is the address that will be temporarily assigned during testing. All test traffic originating from this IP address to the <RemoteInnerIP> will be tunneled. It is possible to override default <ControlPort> and <DataPort>, if there is any other application that already listens on those two ports.

Server Mode

To conduct tests, `ovs-l3ping` server must be running. It is required that both client and server <InnerIP> addresses are in the same subnet. It is possible to specify <InnerIP> with netmask in CIDR format.

Options

One of `-s` or `-c` is required. The `-t` option is also required.

- `-s <TunnelRemoteIP>,<InnerIP>[/<mask>][:<ControlPort>]` or `--server <TunnelRemoteIP>, <InnerIP>[/<mask>][:<ControlPort>]`

  Run in server mode and create L3 tunnel with the client that will be accepting tunnel at <TunnelRemoteIP> address. The socket on `<InnerIP>[:<ControlPort>]` will be used to receive further instructions from the client.

- `-c <TunnelRemoteIP>,<InnerIP>[/<mask>][:<ControlPort>[:<DataPort>]], <RemoteInnerIP>[:<ControlPort>[:<DataPort>]]` or `--client <TunnelRemoteIP>,<InnerIP>[/ <mask>][:<ControlPort>[:<DataPort>]],<RemoteInnerIP>[:<ControlPort>[:<DataPort>]]`

  Run in client mode and create L3 tunnel with the server on <TunnelRemoteIP>. The client will use <InnerIP> to generate test traffic with the server's <RemoteInnerIP>.

- `-b <targetbandwidth>` or `--bandwidth <targetbandwidth>`

  Target bandwidth for UDP tests. The <targetbandwidth> must be given in bits per second. Use postfix M or K to alter the target bandwidth magnitude.

- `-i <testinterval>` or `--interval <testinterval>`

  How long each test should run. By default 5 seconds.

- `-t <tunnelmode>` or `--tunnel-mode <tunnelmode>`

  Specify the tunnel type. This option must match on server and client.

- `-h` or `--help`

  Prints a brief help message to the console.

- `-V` or `--version`

  Prints version information to the console.

Examples

On host 192.168.122.220 start `ovs-l3ping` in server mode. This command will create a temporary GRE tunnel with the host 192.168.122.236 and assign 10.1.1.1/28 as the inner IP address, where client will have to connect:

```
ovs-l3ping -s 192.168.122.236,10.1.1.1/28 -t gre
```

On host 192.168.122.236 start `ovs-l3ping` in client mode. This command will use 10.1.1.2/28 as the local inner IP address and will connect over the L3 tunnel to the server's inner IP address at 10.1.1.1:

```
ovs-l3ping -c 192.168.122.220,10.1.1.2/28,10.1.1.1 -t gre
```

See Also

`ovs-vswitchd(8)`, `ovs-ofctl(8)`, `ovs-vsctl(8)`, `ovs-vlan-test(8)`, `ovs-test(8)`, `ethtool(8)`, `uname(1)`.

6.1.6 ovs-pki

Synopsis

Each command takes the form:

`ovs-pki <options> <command> <args>...`

The implemented commands and their arguments are:

- `ovs-pki init`
- `ovs-pki req <name>`
- `ovs-pki sign <name> [<type>]`
- `ovs-pki req+sign <name> [<type>]`
- `ovs-pki verify <name> [<type>]`
- `ovs-pki fingerprint <file>`
- `ovs-pki self-sign <name>`

Each `<type>` above is a certificate type, either `switch` (default) or `controller`.

The available options are:

- `-k <type>` or `--key=<type>`
- `-B <nbits>` or `--bits=<nbits>`
- `-D <file>` or `--dsaparam=<file>`
- `-b` or `--batch`
- `-f` or `--force`

- `-d <dir>` or `--dir=<dir>`
- `-l <file>` or `--log=<file>`
- `-u` or `--unique`
- `-h` or `--help`

Description

The `ovs-pki` program sets up and manages a public key infrastructure for use with OpenFlow. It is intended to be a simple interface for organizations that do not have an established public key infrastructure. Other PKI tools can substitute for or supplement the use of `ovs-pki`.

`ovs-pki` uses `openssl(1)` for certificate management and key generation.

Offline Commands

The following `ovs-pki` commands support manual PKI administration:

- `init`

  Initializes a new PKI (by default in `/var/lib/openvswitch/pki`, although this default may be changed at Open vSwitch build time) and populates it with a pair of certificate authorities for controllers and switches.

  This command should ideally be run on a high-security machine separate from any OpenFlow controller or switch, called the CA machine. The files `pki/controllerca/cacert.pem` and `pki/switchca/cacert.pem` that it produces will need to be copied over to the OpenFlow switches and controllers, respectively. Their contents may safely be made public.

  By default, `ovs-pki` generates 2048-bit RSA keys. The `-B` or `--bits` option (see below) may be used to override the key length. The `-k dsa` or `--key=dsa` option may be used to use DSA in place of RSA. If DSA is selected, the `dsaparam.pem` file generated in the new PKI hierarchy must be copied to any machine on which the `req` command (see below) will be executed. Its contents may safely be made public.

  Other files generated by `init` may remain on the CA machine. The files `pki/controllerca/private/cakey.pem` and `pki/switchca/private/cakey.pem` have particularly sensitive contents that should not be exposed.

- `req <name>`

  Generates a new private key named `<name>-privkey.pem` and corresponding certificate request named `<name>-req.pem`. The private key can be intended for use by a switch or a controller.

  This command should ideally be run on the switch or controller that will use the private key to identify itself. The file `<name>-req.pem` must be copied to the CA machine for signing with the `sign` command (below).

  This command will output a fingerprint to stdout as its final step. Write down the fingerprint and take it to the CA machine before continuing with the `sign` step.

  When RSA keys are in use (as is the default), `req`, unlike the rest of the `ovs-pki` commands, does not need access to a PKI hierarchy created by `ovs-pki init`. The `-B` or `--bits` option (see below) may be used to specify the number of bits in the generated RSA key.

  When DSA keys are used (as specified with `--key=dsa`), `req` needs access to the `dsaparam.pem` file created as part of the PKI hierarchy (but not to other files in that tree). By default, `ovs-pki` looks for this file in the PKI directory as `dsaparam.pem`, but the `-D` or `--dsaparam` option (see below) may be used to specify an alternate location.

  `<name>-privkey.pem` has sensitive contents that should not be exposed. `<name>-req.pem` may be safely made public.

- `sign <name> [<type>]`

Signs the certificate request named <name>-req.pem that was produced in the previous step, producing a certificate named <name>-cert.pem. <type>, either switch (default) or controller, indicates the use for which the key is being certified.

This command must be run on the CA machine.

The command will output a fingerprint to stdout and request that you verify that it is the same fingerprint output by the req command. This ensures that the request being signed is the same one produced by req. (The -b or --batch option suppresses the verification step.)

The file <name>-cert.pem will need to be copied back to the switch or controller for which it is intended. Its contents may safely be made public.

- req+sign <name> [<type>]

  Combines the req and sign commands into a single step, outputting all the files produced by each. The <name>-privkey.pem and <name>-cert.pem files must be copied securely to the switch or controller. <name>-privkey.pem has sensitive contents and must not be exposed in transit. Afterward, it should be deleted from the CA machine.

  This combined method is, theoretically, less secure than the individual steps performed separately on two different machines, because there is additional potential for exposure of the private key. However, it is also more convenient.

- verify <name> [<type>]

  Verifies that <name>-cert.pem is a valid certificate for the given <type> of use, either switch (default) or controller. If the certificate is valid for this use, it prints the message <name>-cert.pem: OK; otherwise, it prints an error message.

- fingerprint <file>

  Prints the fingerprint for <file>. If <file> is a certificate, then this is the SHA-1 digest of the DER encoded version of the certificate; otherwise, it is the SHA-1 digest of the entire file.

- self-sign <name>

  Signs the certificate request named <name>-req.pem using the private key <name>-privkey.pem, producing a self-signed certificate named <name>-cert.pem. The input files should have been produced with ovs-pki req.

  Some controllers accept such self-signed certificates.

Options

- -k <type> or --key=<type>

  For the init command, sets the public key algorithm to use for the new PKI hierarchy. For the req and req+sign commands, sets the public key algorithm to use for the key to be generated, which must match the value specified on init. With other commands, the value has no effect.

  The <type> may be rsa (the default) or dsa.

- -B <nbits> or --bits=<nbits>

  Sets the number of bits in the key to be generated. When RSA keys are in use, this option affects only the init, req, and req+sign commands, and the same value should be given each time. With DSA keys are in use, this option affects only the init command.

  The value must be at least 1024. The default is 2048.

- -D <file> or --dsaparam=<file>

Specifies an alternate location for the `dsaparam.pem` file required by the `req` and `req+sign` commands. This option affects only these commands, and only when DSA keys are used.

The default is `dsaparam.pem` under the PKI hierarchy.

- `-b` or `--batch`

  Suppresses the interactive verification of fingerprints that the `sign` command by default requires.

- `-d <dir>` or `--dir=<dir>`

  Specifies the location of the PKI hierarchy to be used or created by the command. All commands, except `req`, need access to a PKI hierarchy.

  The default PKI hierarchy is `/var/lib/openvswitch/pki`, although this default may be changed at Open vSwitch build time

- `-f` or `--force`

  By default, `ovs-pki` will not overwrite existing files or directories. This option overrides this behavior.

- `-l <file>` or `--log=<file>`

  Sets the log file to <file>. The default is `ovs-pki.log` in the OVS log directory. The default OVS log directory is `/var/log/openvswitch`, although this default may be changed at Open vSwitch build time.

- `-u` or `--unique`

  Changes the format of the certificate's Common Name (CN) field. By default, this field has the format `<name> id:<uuid-or-date>`. This option causes the provided name to be treated as unique and changes the format of the CN field to be simply `<name>`.

- `-h` or `--help`

  Prints a help usage message and exits.

6.1.7 ovs-sim

Synopsis

`ovs-sim` [*option*]... [*script*]...

Description

`ovs-sim` provides a convenient environment for running one or more Open vSwitch instances and related software in a sandboxed simulation environment.

To use `ovs-sim`, first build Open vSwitch, then invoke it directly from the build directory, e.g.:

```
git clone https://github.com/openvswitch/ovs.git
cd ovs
./configure
make
utilities/ovs-sim
```

When invoked in the most ordinary way as shown above, ovs-sim does the following:

1. Creates a directory `sandbox` as a subdirectory of the current directory (first destroying such a directory if it already exists) and makes it the current directory.
2. Installs all of the Open vSwitch manpages into a `man` subdirectory of sandbox and adjusts the `MANPATH` environment variable so that `man` and other manpage viewers can find them.

3. Creates a simulated Open vSwitch named `main` and sets it up as the default target for OVS commands, as if the following `ovs-sim` commands had been run:

```
sim_add main
as main
```

   See *Commands*, below, for an explanation.

4. Runs any scripts specified on the command line (see *Options*, below). The scripts can use arbitrary Bash syntax, plus the additional commands described under *Commands*, below.

5. If no scripts were specified, or if `-i` or `--interactive` was specified, invokes an interactive Bash subshell. The user can use arbitrary Bash commands, plus the additional commands described under *Commands*, below.

`ovs-sim` and the sandbox environment that it creates does not require superuser or other special privileges. Generally, it should not be run with such privileges.

Options

script
: Runs *script*, which should be a Bash script, within a subshell after initializing. If multiple script arguments are given, then they are run in the order given. If any script exits with a nonzero exit code, then `ovs-sim` exits immediately with the same exit code.

`-i` or `--interactive`
: By default, if any script is specified, `ovs-sim` exits as soon as the scripts finish executing. With this option, or if no scripts are specified, `ovs-sim` instead starts an interactive Bash session.

Commands

Scripts and interactive usage may use the following commands implemented by `ovs-sim`. They are implemented as Bash shell functions exported to subshells.

Basic Commands

These are the basic commands for working with sandboxed Open vSwitch instances.

`sim_add` ***sandbox***
: Starts a new simulated Open vSwitch instance named *sandbox*. Files related to the instance, such as logs, databases, sockets, and pidfiles, are created in a subdirectory also named *sandbox*. Afterward, the `as` command (see below) can be used to run Open vSwitch utilities in the context of the new sandbox.

  The new sandbox starts out without any bridges. Use `ovs-vsctl` in the context of the new sandbox to create a bridge, e.g.:

```
sim_add hv0           # Create sandbox hv0.
as hv0                # Set hv0 as default sandbox.
ovs-vsctl add-br br0  # Add bridge br0 inside hv0.
```

  The Open vSwitch instances that `sim_add` creates enable `dummy` devices. This means that bridges and interfaces can be created with type `dummy` to indicate that they should be totally simulated, without any reference to system entities. In fact, `ovs-sim` also configures Open vSwitch so that the default system type of bridges and interfaces are replaced by dummy devices. Other types of devices, however, retain their usual functions, which means that, e.g., vxlan tunnels still act as tunnels (refer to the documentation).

`as` ***sandbox***
: Sets sandbox as the default simulation target for Open vSwitch commands (e.g. `ovs-vsctl`, `ovs-ofctl`, `ovs-appctl`).

This command updates the beginning of the shell prompt to indicate the new default target.

as ***sandbox command arg…***
Runs the given command with *sandbox* as the simulation target, e.g. `as hv0 ovs-vsctl add-br br0` runs `ovs-vsctl add-br br0` within sandbox `hv0`. The default target is unchanged.

Interconnection Network Commands

When multiple sandboxed Open vSwitch instances exist, one will inevitably want to connect them together. These commands allow for that. Conceptually, an interconnection network is a switch that `ovs-sim` makes it easy to plug into other switches in other sandboxed Open vSwitch instances. Interconnection networks are implemented as bridges in the `main` switch that `ovs-sim` creates by default, so to use interconnection networks please avoid working with `main` directly.

net_add ***network***
Creates a new interconnection network named *network*.

net_attach ***network bridge***
Adds a new port to *bridge* in the default sandbox (as set with `as`) and plugs it into interconnection network *network*, which must already have been created by a previous invocation of `net_add`. The default sandbox must not be `main`.

6.1.8 ovs-tcpdump

Synopsis

```
ovs-tcpdump -i <port> <tcpdump options>...
```

Description

`ovs-tcpdump` creates switch mirror ports in the `ovs-vswitchd` daemon and executes `tcpdump` to listen against those ports. When the `tcpdump` instance exits, it then cleans up the mirror port it created.

`ovs-tcpdump` will not allow multiple mirrors for the same port. It has some logic to parse the current configuration and prevent duplicate mirrors.

The `-i` option may not appear multiple times.

It is important to note that under Linux-based kernels, tap devices do not receive packets unless the specific tuntap device has been opened by an application. This requires `CAP_NET_ADMIN` privileges, so the `ovs-tcpdump` command must be run as a user with such permissions (this is usually a super-user).

Options

- `-h` or `--help`

  Prints a brief help message to the console.

- `-V` or `--version`

  Prints version information to the console.

- `--db-sock <socket>`

  The Open vSwitch database socket connection string. The default is `unix:<rundir>/db.sock`.

- `--dump-cmd <command>`

  The command to run instead of `tcpdump`.

- `-i` or `--interface`

  The interface for which a mirror port should be created, and packets should be dumped.

- `--mirror-to`

  The name of the interface which should be the destination of the mirrored packets. If the specified interface does not exist, it will be created as part of the setup process. If the interface already exists, it must be a port type that can be used with the `tcpdump` utility. Mirror ports cannot be used for normal traffic. The default value is `mi<port>`.

- `--span`

  If specified, mirror all ports (optional).

- `--filter <flow>`

  If specified, only mirror packets that match the provided OpenFlow filter. The available fields are documented in `ovs-fields(7)`.

See Also

`ovs-appctl(8)`, `ovs-vswitchd(8)`, `ovs-pcap(1)`, `ovs-fields(7)`, `ovs-tcpundump(1)`, `tcpdump(8)`, `wireshark(8)`.

6.1.9 ovs-tcpundump

Synopsis

`ovs-tcpundump < <file>`

`ovs-tcpundump -h | --help`

`ovs-tcpundump -V | --version`

The `ovs-tcpundump` program reads `tcpdump -xx output` from stdin, looking for hexadecimal packet data, and dumps each Ethernet as a single hexadecimal string on stdout. This format is suitable for use with the `ofproto/trace` command supported by `ovs-vswitchd(8)` via `ovs-appctl(8)`.

At least two `-x` or `-X` options must be given, otherwise the output will omit the Ethernet header, which prevents the output from being used with `ofproto/trace`.

Options

- `-h` or `--help`

  Prints a brief help message to the console.

- `-V` or `--version`

  Prints version information to the console.

See Also

`ovs-appctl(8)`, `ovs-vswitchd(8)`, `ovs-pcap(1)`, `tcpdump(8)`, `wireshark(8)`.

6.1.10 ovs-test

Synopsis

ovs-test -s *port*

ovs-test -c ***server1 server2*** **[-b** ***targetbandwidth***] **[-i** ***testinterval***] **[-d]**
[**-l** *vlantag*] [**-t** *tunnelmodes*]

Description

The **ovs-test** program may be used to check for problems sending 802.1Q or GRE traffic that Open vSwitch may uncover. These problems, for example, can occur when Open vSwitch is used to send 802.1Q traffic through physical interfaces running certain drivers of certain Linux kernel versions. To run a test, configure IP addresses on *server1* and *server2* for interfaces you intended to test. These interfaces could also be already configured OVS bridges that have a physical interface attached to them. Then, on one of the nodes, run **ovs-test** in server mode and on the other node run it in client mode. The client will connect to **ovs-test** server and schedule tests between both of them. The **ovs-test** client will perform UDP and TCP tests.

UDP tests can report packet loss and achieved bandwidth for various datagram sizes. By default target bandwidth for UDP tests is 1Mbit/s.

TCP tests report only achieved bandwidth, because kernel TCP stack takes care of flow control and packet loss. TCP tests are essential to detect potential TSO related issues.

To determine whether Open vSwitch is encountering any problems, the user must compare packet loss and achieved bandwidth in a setup where traffic is being directly sent and in one where it is not. If in the 802.1Q or L3 tunneled tests both **ovs-test** processes are unable to communicate or the achieved bandwidth is much lower compared to direct setup, then, most likely, Open vSwitch has encountered a pre-existing kernel or driver bug.

Some examples of the types of problems that may be encountered are:

- When NICs use VLAN stripping on receive they must pass a pointer to a *vlan_group* when reporting the stripped tag to the networking core. If no *vlan_group* is in use then some drivers just drop the extracted tag. Drivers are supposed to only enable stripping if a *vlan_group* is registered but not all of them do that.
- On receive, some drivers handle priority tagged packets specially and don't pass the tag onto the network stack at all, so Open vSwitch never has a chance to see it.
- Some drivers size their receive buffers based on whether a *vlan_group* is enabled, meaning that a maximum size packet with a VLAN tag will not fit if no *vlan_group* is configured.
- On transmit, some drivers expect that VLAN acceleration will be used if it is available, which can only be done if a *vlan_group* is configured. In these cases, the driver may fail to parse the packet and correctly setup checksum offloading or TSO.

Client Mode
: An **ovs-test** client will connect to two **ovs-test** servers and will ask them to exchange test traffic. It is also possible to spawn an **ovs-test** server automatically from the client.

Server Mode
: To conduct tests, two **ovs-test** servers must be running on two different hosts where the client can connect. The actual test traffic is exchanged only between both **ovs-test** servers. It is recommended that both servers have their IP addresses in the same subnet, otherwise one would have to make sure that routing is set up correctly.

Options

-s <port>, **--server** <port>
: Run in server mode and wait for the client to establish XML RPC Control Connection on this TCP port. It is recommended to have *ethtool(8)* installed on the server so that it could retrieve information about the NIC driver.

-c <server1> <server2>, **--client** <server1> <server2>
: Run in client mode and schedule tests between *server1* and *server2*, where each server must be given in the following format:

```
OuterIP[:OuterPort],InnerIP[/Mask][:InnerPort].
```

The *OuterIP* must be already assigned to the physical interface which is going to be tested. This is the IP address where client will try to establish XML RPC connection. If *OuterIP* is 127.0.0.1 then client will automatically spawn a local instance of **ovs-test** server. OuterPort is TCP port where server is listening for incoming XML/RPC control connections to schedule tests (by default it is 15531). The **ovs-test** will automatically assign *InnerIP[/Mask]* to the interfaces that will be created on the fly for testing purposes. It is important that *InnerIP[/Mask]* does not interfere with already existing IP addresses on both **ovs-test** servers and client. InnerPort is port which will be used by server to listen for test traffic that will be encapsulated (by default it is 15532).

-b <targetbandwidth>, **--bandwidth** <targetbandwidth>

Target bandwidth for UDP tests. The targetbandwidth must be given in bits per second. It is possible to use postfix *M* or *K* to alter the target bandwidth magnitude.

-i <testinterval>, **--interval** <testinterval>

How long each test should run. By default 5 seconds.

-h, **--help**

Prints a brief help message to the console.

-V, **--version**

Prints version information to the console.

The following test modes are supported by **ovs-test**. It is possible to combine multiple of them in a single **ovs-test** invocation.

-d, **--direct**

Perform direct tests between both OuterIP addresses. These tests could be used as a reference to compare 802.1Q or L3 tunneling test results.

-l <vlantag>, **--vlan-tag** <vlantag>

Perform 802.1Q tests between both servers. These tests will create a temporary OVS bridge, if necessary, and attach a VLAN tagged port to it for testing purposes.

-t <tunnelmodes>, **--tunnel-modes** <tunnelmodes>

Perform L3 tunneling tests. The given argument is a comma sepa- rated string that specifies all the L3 tunnel modes that should be tested (e.g. gre). The L3 tunnels are terminated on interface that has the OuterIP address assigned.

Examples

On host 1.2.3.4 start **ovs-test** in server mode:

```
ovs-test -s 15531
```

On host 1.2.3.5 start **ovs-test** in client mode and do direct, VLAN and GRE tests between both nodes:

```
ovs-test -c 127.0.0.1,1.1.1.1/30 1.2.3.4,1.1.1.2/30 -d -l 123 -t
gre
```

See Also

ovs-vswitchd(8), *ovs-ofctl(8)*, *ovs-vsctl(8)*, **ovs-vlan-test**, *ethtool(8)*, *uname(1)*

6.1.11 ovs-vlan-test

Synopsis

ovs-vlan-test [**-s** | **–server**] *control_ip vlan_ip*

Description

The **ovs-vlan-test** utility has some limitations, for example, it does not use TCP in its tests. Also it does not take into account MTU to detect potential edge cases. To overcome those limitations a new tool was developed - **ovs-test**. **ovs-test** is currently supported only on Debian so, if possible, try to use that on instead of **ovs-vlan-test**.

The **ovs-vlan-test** program may be used to check for problems sending 802.1Q traffic which may occur when running Open vSwitch. These problems can occur when Open vSwitch is used to send 802.1Q traffic through physical interfaces running certain drivers of certain Linux kernel versions. To run a test, configure Open vSwitch to tag traffic originating from *vlan_ip* and forward it out the target interface. Then run the **ovs-vlan-test** in client mode connecting to an **ovs-vlan-test** server. **ovs-vlan-test** will display "OK" if it did not detect problems.

Some examples of the types of problems that may be encountered are:

- When NICs use VLAN stripping on receive they must pass a pointer to a *vlan_group* when reporting the stripped tag to the networking core. If no *vlan_group* is in use then some drivers just drop the extracted tag. Drivers are supposed to only enable stripping if a *vlan_group* is registered but not all of them do that.
- On receive, some drivers handle priority tagged packets specially and don't pass the tag onto the network stack at all, so Open vSwitch never has a chance to see it.
- Some drivers size their receive buffers based on whether a *vlan_group* is enabled, meaning that a maximum size packet with a VLAN tag will not fit if no *vlan_group* is configured.
- On transmit, some drivers expect that VLAN acceleration will be used if it is available, which can only be done if a *vlan_group* is configured. In these cases, the driver may fail to parse the packet and correctly setup checksum offloading or TSO.

Client Mode
An **ovs-vlan-test** client may be run on a host to check for VLAN connectivity problems. The client must be able to establish HTTP connections with an **ovs-vlan-test** server located at the specified *control_ip* address. UDP traffic sourced at *vlan_ip* should be tagged and directed out the interface whose connectivity is being tested.

Server Mode
To conduct tests, an **ovs-vlan-test** server must be running on a host known not to have VLAN connectivity problems. The server must have a *control_ip* on a non-VLAN network which clients can establish connectivity with. It must also have a *vlan_ip* address on a VLAN network which clients will use to test their VLAN connectivity. Multiple clients may test against a single **ovs-vlan-test** server concurrently.

Options

-s, **--server**
Run in server mode.

-h, **--help**
Prints a brief help message to the console.

-V, **--version**
Prints version information to the console.

Examples

Display the Linux kernel version and driver of *eth1*:

```
uname -r
ethtool -i eth1
```

Set up a bridge which forwards traffic originating from *1.2.3.4* out *eth1* with VLAN tag 10:

```
ovs-vsctl -- add-br vlan-br \
  -- add-port vlan-br eth1 \
  -- add-port vlan-br vlan-br-tag tag=10 \
  -- set Interface vlan-br-tag type=internal
ip addr add 1.2.3.4/8 dev vlan-br-tag
ip link set vlan-br-tag up
```

Run an **ovs-vlan-test** server listening for client control traffic on *172.16.0.142* port *8080* and VLAN traffic on the default port of *1.2.3.3*:

```
ovs-vlan-test -s 172.16.0.142:8080 1.2.3.3
```

Run an **ovs-vlan-test** client with a control server located at *172.16.0.142* port *8080* and a local VLAN IP of *1.2.3.4*:

```
ovs-vlan-test 172.16.0.142:8080 1.2.3.4
```

See Also

ovs-vswitchd(8), *ovs-ofctl(8)*, *ovs-vsctl(8)*, **ovs-test**, *ethtool(8)*, *uname(1)*

6.1.12 ovsdb-server

Description

`ovsdb-server` implements the Open vSwitch Database (OVSDB) protocol specified in RFC 7047. This document provides clarifications for how `ovsdb-server` implements the protocol and describes the extensions that it provides beyond RFC 7047. Numbers in section headings refer to corresponding sections in RFC 7047.

3.1 JSON Usage

RFC 4627 says that names within a JSON object should be unique. The Open vSwitch JSON parser discards all but the last value for a name that is specified more than once.

The definition of <error> allows for implementation extensions. Currently `ovsdb-server` uses the following additional `error` strings (which might change in later releases):

`syntax error` or `unknown column`
: The request could not be parsed as an OVSDB request. An additional `syntax` member, whose value is a string that contains JSON, may narrow down the particular syntax that could not be parsed.

`internal error`
: The request triggered a bug in `ovsdb-server`.

`ovsdb error`
: A map or set contains a duplicate key.

`permission error`
: The request was denied by the role-based access control extension, introduced in version 2.8.

3.2 Schema Format

RFC 7047 requires the `version` field in <database-schema>. Current versions of `ovsdb-server` allow it to be omitted (future versions are likely to require it).

RFC 7047 allows columns that contain weak references to be immutable. This raises the issue of the behavior of the weak reference when the rows that it references are deleted. Since version 2.6, `ovsdb-server` forces columns that contain weak references to be mutable.

Since version 2.8, the table name `RBAC_Role` is used internally by the role-based access control extension to `ovsdb-server` and should not be used for purposes other than defining mappings of role names to table access permissions. This table has one row per role name and the following columns:

name
: The role name.

permissions
: A map of table name to a reference to a row in a separate permission table.

The separate RBAC permission table has one row per access control configuration and the following columns:

name
: The name of the table to which the row applies.

authorization
: The set of column names and column:key pairs to be compared with the client ID in order to determine the authorization status of the requested operation.

insert_delete
: A boolean value, true if authorized insertions and deletions are allowed, false if no insertions or deletions are allowed.

update
: The set of columns and column:key pairs for which authorized update and mutate operations should be permitted.

4 Wire Protocol

The original OVSDB specifications included the following reasons, omitted from RFC 7047, to operate JSON-RPC directly over a stream instead of over HTTP:

- JSON-RPC is a peer-to-peer protocol, but HTTP is a client-server protocol, which is a poor match. Thus, JSON-RPC over HTTP requires the client to periodically poll the server to receive server requests.
- HTTP is more complicated than stream connections and doesn't provide any corresponding advantage.
- The JSON-RPC specification for HTTP transport is incomplete.

4.1.3 Transact

Since version 2.8, role-based access controls can be applied to operations within a transaction that would modify the contents of the database (these operations include row insert, row delete, column update, and column mutate). Role-based access controls are applied when the database schema contains a table with the name `RBAC_Role` and the connection on which the transaction request was received has an associated role name (from the `role` column in the remote connection table). When role-based access controls are enabled, transactions that are otherwise well-formed may be rejected depending on the client's role, ID, and the contents of the `RBAC_Role` table and associated permissions table.

4.1.5 Monitor

For backward compatibility, `ovsdb-server` currently permits a single <monitor-request> to be used instead of an array; it is treated as a single-element array. Future versions of `ovsdb-server` might remove this compatibility feature.

Because the <json-value> parameter is used to match subsequent update notifications (see below) to the request, it must be unique among all active monitors. `ovsdb-server` rejects attempt to create two monitors with the same identifier.

When a given client sends a `transact` request that changes a table that the same client is monitoring, `ovsdb-server` always sends the `update` (or `update2` or `update3`) for these changes before it sends the reply to the `transact` request. Thus, when a client receives a `transact` reply, it can know immediately what changes (if any) the transaction made. (If ovsdb-server might use the other order, then a client that wishes to act on based on the results of its own transactions would not know when this was guaranteed to have taken place.)

4.1.7 Monitor Cancellation

When a database monitored by a session is removed, and database change awareness is enabled for the session (see Section 4.1.16), the database server spontaneously cancels all monitors (including conditional monitors described in Section 4.1.12) for the removed database. For each canceled monitor, it issues a notification in the following form:

```
"method": "monitor_canceled"
"params": [<json-value>]
"id": null
```

4.1.12 Monitor_cond

A new monitor method added in Open vSwitch version 2.6. The `monitor_cond` request enables a client to replicate subsets of tables within an OVSDB database by requesting notifications of changes to rows matching one of the conditions specified in `where` by receiving the specified contents of these rows when table updates occur. `monitor_cond` also allows a more efficient update notifications by receiving <table-updates2> notifications (described below).

The `monitor` method described in Section 4.1.5 also applies to `monitor_cond`, with the following exceptions:

- RPC request method becomes `monitor_cond`.
- Reply result follows <table-updates2>, described in Section 4.1.14.
- Subsequent changes are sent to the client using the `update2` monitor notification, described in Section 4.1.14
- Update notifications are being sent only for rows matching [<condition>*].

The request object has the following members:

```
"method": "monitor_cond"
"params": [<db-name>, <json-value>, <monitor-cond-requests>]
"id": <nonnull-json-value>
```

The <json-value> parameter is used to match subsequent update notifications (see below) to this request. The <monitor-cond-requests> object maps the name of the table to an array of <monitor-cond-request>.

Each <monitor-cond-request> is an object with the following members:

```
"columns": [<column>*]           optional
"where": [<condition>*]          optional
"select": <monitor-select>       optional
```

The `columns`, if present, define the columns within the table to be monitored that match conditions. If not present, all columns are monitored.

The `where`, if present, is a JSON array of <condition> and boolean values. If not present or condition is an empty array, implicit True will be considered and updates on all rows will be sent.

<monitor-select> is an object with the following members:

```
"initial": <boolean>              optional
"insert": <boolean>               optional
"delete": <boolean>               optional
"modify": <boolean>               optional
```

The contents of this object specify how the columns or table are to be monitored as explained in more detail below.

The response object has the following members:

```
"result": <table-updates2>
"error": null
"id": same "id" as request
```

The <table-updates2> object is described in detail in Section 4.1.14. It contains the contents of the tables for which initial rows are selected. If no tables initial contents are requested, then `result` is an empty object.

Subsequently, when changes to a specified table that match one of the conditions in <monitor-cond-request> are committed, the changes are automatically sent to the client using the `update2` monitor notification (see Section 4.1.14). This monitoring persists until the JSON-RPC session terminates or until the client sends a `monitor_cancel` JSON-RPC request.

Each <monitor-cond-request> specifies one or more conditions and the manner in which the rows that match the conditions are to be monitored. The circumstances in which an `update` notification is sent for a row within the table are determined by <monitor-select>:

- If `initial` is omitted or true, every row in the original table that matches one of the conditions is sent as part of the response to the `monitor_cond` request.
- If `insert` is omitted or true, update notifications are sent for rows newly inserted into the table that match conditions or for rows modified in the table so that their old version does not match the condition and new version does.
- If `delete` is omitted or true, update notifications are sent for rows deleted from the table that match conditions or for rows modified in the table so that their old version does match the conditions and new version does not.
- If `modify` is omitted or true, update notifications are sent whenever a row in the table that matches conditions in both old and new version is modified.

Both `monitor` and `monitor_cond` sessions can exist concurrently. However, `monitor` and `monitor_cond` shares the same <json-value> parameter space; it must be unique among all `monitor` and `monitor_cond` sessions.

4.1.13 Monitor_cond_change

The `monitor_cond_change` request enables a client to change an existing `monitor_cond` replication of the database by specifying a new condition and columns for each replicated table. Currently changing the columns set is not supported.

The request object has the following members:

```
"method": "monitor_cond_change"
"params": [<json-value>, <json-value>, <monitor-cond-update-requests>]
"id": <nonnull-json-value>
```

The <json-value> parameter should have a value of an existing conditional monitoring session from this client. The second <json-value> in params array is the requested value for this session. This value is valid only after

`monitor_cond_change` is committed. A user can use these values to distinguish between update messages before conditions update and after. The <monitor-cond-update-requests> object maps the name of the table to an array of <monitor-cond-update-request>. Monitored tables not included in <monitor-cond-update-requests> retain their current conditions.

Each <monitor-cond-update-request> is an object with the following members:

```
"columns": [<column>*]          optional
"where": [<condition>*]         optional
```

The `columns` specify a new array of columns to be monitored, although this feature is not yet supported.

The `where` specify a new array of conditions to be applied to this monitoring session.

The response object has the following members:

```
"result": {}
"error": null
"id": same "id" as request
```

Subsequent <table-updates2> notifications are described in detail in Section 4.1.14 in the RFC. If insert contents are requested by original monitor_cond request, <table-updates2> will contain rows that match the new condition and do not match the old condition. If deleted contents are requested by origin monitor request, <table-updates2> will contain any matched rows by old condition and not matched by the new condition.

Changes according to the new conditions are automatically sent to the client using the `update2` or `update3` monitor notification depending on the monitor method. An update, if any, as a result of a condition change, will be sent to the client before the reply to the `monitor_cond_change` request.

4.1.14 Update2 notification

The `update2` notification is sent by the server to the client to report changes in tables that are being monitored following a `monitor_cond` request as described above. The notification has the following members:

```
"method": "update2"
"params": [<json-value>, <table-updates2>]
"id": null
```

The <json-value> in `params` is the same as the value passed as the <json-value> in `params` for the corresponding `monitor` request. <table-updates2> is an object that maps from a table name to a <table-update2>. A <table-update2> is an object that maps from row's UUID to a <row-update2> object. A <row-update2> is an object with one of the following members:

"initial": <row>
: present for `initial` updates

"insert": <row>
: present for `insert` updates

"delete": <row>
: present for `delete` updates

"modify": <row>"
: present for `modify` updates

The format of <row> is described in Section 5.1.

<row> is always a null object for a `delete` update. In `initial` and `insert` updates, <row> omits columns whose values equal the default value of the column type.

For a `modify` update, <row> contains only the columns that are modified. <row> stores the difference between the old and new value for those columns, as described below.

For columns with single value, the difference is the value of the new column.

The difference between two sets are all elements that only belong to one of the sets.

The difference between two maps are all key-value pairs whose keys appears in only one of the maps, plus the key-value pairs whose keys appear in both maps but with different values. For the latter elements, <row> includes the value from the new column.

Initial views of rows are not presented in update2 notifications, but in the response object to the `monitor_cond` request. The formatting of the <table-updates2> object, however, is the same in either case.

4.1.15 Monitor_cond_since

A new monitor method added in Open vSwitch version 2.12. The `monitor_cond_since` request enables a client to request changes that happened after a specific transaction id. A client can use this feature to request only latest changes after a server connection reset instead of re-transfer all data from the server again.

The `monitor_cond` method described in Section 4.1.12 also applies to `monitor_cond_since`, with the following exceptions:

- RPC request method becomes `monitor_cond_since`.
- Reply result includes extra parameters.
- Subsequent changes are sent to the client using the `update3` monitor notification, described in Section 4.1.16

The request object has the following members:

```
"method": "monitor_cond_since"
"params": [<db-name>, <json-value>, <monitor-cond-requests>, <last-txn-id>]
"id": <nonnull-json-value>
```

The <last-txn-id> parameter is the transaction id that identifies the latest data the client already has, and it requests server to send changes AFTER this transaction (exclusive).

All other parameters are the same as `monitor_cond` method.

The response object has the following members:

```
"result": [<found>, <last-txn-id>, <table-updates2>]
"error": null
"id": same "id" as request
```

The <found> is a boolean value that tells if the <last-txn-id> requested by client is found in server's history or not. If true, the changes after that version up to current is sent. Otherwise, all data is sent.

The <last-txn-id> is the transaction id that identifies the latest transaction included in the changes in <table-updates2> of this response, so that client can keep tracking. If there is no change involved in this response, it is the same as the <last-txn-id> in the request if <found> is true, or zero uuid if <found> is false. If the server does not support transaction uuid, it will be zero uuid as well.

All other parameters are the same as in response object of `monitor_cond` method.

Like in `monitor_cond`, subsequent changes that match conditions in <monitor-cond-request> are automatically sent to the client, but using `update3` monitor notification (see Section 4.1.16), instead of `update2`.

4.1.16 Update3 notification

The `update3` notification is sent by the server to the client to report changes in tables that are being monitored following a `monitor_cond_since` request as described above. The notification has the following members:

```
"method": "update3"
"params": [<json-value>, <last-txn-id>, <table-updates2>]
"id": null
```

The <last-txn-id> is the same as described in the response object of `monitor_cond_since`.

All other parameters are the same as in `update2` monitor notification (see Section 4.1.14).

4.1.17 Get Server ID

A new RPC method added in Open vSwitch version 2.7. The request contains the following members:

```
"method": "get_server_id"
"params": null
"id": <nonnull-json-value>
```

The response object contains the following members:

```
"result": "<server_id>"
"error": null
"id": same "id" as request
```

<server_id> is JSON string that contains a UUID that uniquely identifies the running OVSDB server process. A fresh UUID is generated when the process restarts.

4.1.18 Database Change Awareness

RFC 7047 does not provide a way for a client to find out about some kinds of configuration changes, such as about databases added or removed while a client is connected to the server, or databases changing between read/write and read-only due to a transition between active and backup roles. Traditionally, `ovsdb-server` disconnects all of its clients when this happens, because this prompts a well-written client to reassess what is available from the server when it reconnects.

OVS 2.9 provides a way for clients to keep track of these kinds of changes, by monitoring the `Database` table in the `_Server` database introduced in this release (see `ovsdb-server(5)` for details). By itself, this does not suppress `ovsdb-server` disconnection behavior, because a client might monitor this database without understanding its special semantics. Instead, `ovsdb-server` provides a special request:

```
"method": "set_db_change_aware"
"params": [<boolean>]
"id": <nonnull-json-value>
```

If the boolean in the request is true, it suppresses the connection-closing behavior for the current connection, and false restores the default behavior. The reply is always the same:

```
"result": {}
"error": null
"id": same "id" as request
```

4.1.19 Schema Conversion

Open vSwitch 2.9 adds a new JSON-RPC request to convert an online database from one schema to another. The request contains the following members:

```
"method": "convert"
"params": [<db-name>, <database-schema>]
"id": <nonnull-json-value>
```

Upon receipt, the server converts database <db-name> to schema <database-schema>. The schema's name must be <db-name>. The conversion is atomic, consistent, isolated, and durable. The data in the database must be valid when interpreted under <database-schema>, with only one exception: data for tables and columns that do not exist in the new schema are ignored. Columns that exist in <database-schema> but not in the database are set to their default values. All of the new schema's constraints apply in full.

If the conversion is successful, the server notifies clients that use the `set_db_change_aware` RPC introduced in Open vSwitch 2.9 and cancels their outstanding transactions and monitors. The server disconnects other clients, enabling them to notice the change when they reconnect. The server sends the following reply:

```
"result": {}
"error": null
"id": same "id" as request
```

If the conversion fails, then the server sends an error reply in the following form:

```
"result": null
"error": [<error>]
"id": same "id" as request
```

5.1 Notation

For <condition>, RFC 7047 only allows the use of `!=`, `==`, `includes`, and `excludes` operators with set types. Open vSwitch 2.4 and later extend <condition> to allow the use of <, <=, >=, and > operators with a column with type "set of 0 or 1 integer" and an integer argument, and with "set of 0 or 1 real" and a real argument. These conditions evaluate to false when the column is empty, and otherwise as described in RFC 7047 for integer and real types.

<condition> is specified in Section 5.1 in the RFC with the following change: A condition can be either a 3-element JSON array as described in the RFC or a boolean value. In case of an empty array an implicit true boolean value will be considered.

5.2.1 Insert

As an extension, Open vSwitch 2.13 and later allow an optional `uuid` member to specify the UUID for the new row. The specified UUID must be unique within the table when it is inserted and not the UUID of a row previously deleted within the transaction. If the UUID violates these rules, then the operation fails with a `duplicate uuid` error.

5.2.6 Wait, 5.2.7 Commit, 5.2.9 Comment

RFC 7047 says that the `wait`, `commit`, and `comment` operations have no corresponding result object. This is not true. Instead, when such an operation is successful, it yields a result object with no members.

6.1.13 ovsdb

Description

OVSDB, the Open vSwitch Database, is a database system whose network protocol is specified by RFC 7047. The RFC does not specify an on-disk storage format. The OVSDB implementation in Open vSwitch implements two storage formats: one for standalone (and active-backup) databases, and the other for clustered databases. This manpage documents both of these formats.

Most users do not need to be concerned with this specification. Instead, to manipulate OVSDB files, refer to *ovsdb-tool(1)*. For an introduction to OVSDB as a whole, read *ovsdb(7)*.

OVSDB files explicitly record changes that are implied by the database schema. For example, the OVSDB "garbage collection" feature means that when a client removes the last reference to a garbage-collected row, the database server automatically removes that row. The database file explicitly records the deletion of the garbage-collected row, so that the reader does not need to infer it.

OVSDB files do not include the values of ephemeral columns.

Standalone and clustered database files share the common structure described here. They are text files encoded in UTF-8 with LF (U+000A) line ends, organized as append-only series of records. Each record consists of 2 lines of text.

The first line in each record has the format `OVSDB <magic> <length> <hash>`, where <magic> is `JSON` for standalone databases or `CLUSTER` for clustered databases, <length> is a positive decimal integer, and <hash> is a SHA-1 checksum expressed as 40 hexadecimal digits. Words in the first line must be separated by exactly one space.

The second line must be exactly *length* bytes long (including the LF) and its SHA-1 checksum (including the LF) must match *hash* exactly. The line's contents must be a valid JSON object as specified by RFC 4627. Strings in the JSON object must be valid UTF-8. To ensure that the second line is exactly one line of text, the OVSDB implementation expresses any LF characters within a JSON string as `\n`. For the same reason, and to save space, the OVSDB implementation does not "pretty print" the JSON object with spaces and LFs. (The OVSDB implementation tolerates LFs when reading an OVSDB database file, as long as *length* and *hash* are correct.)

JSON Notation

We use notation from RFC 7047 here to describe the JSON data in records. In addition to the notation defined there, we add the following:

<raw-uuid>
: A 36-character JSON string that contains a UUID in the format described by RFC 4122, e.g. `"550e8400-e29b-41d4-a716-446655440000"`

Standalone Format

The first record in a standalone database contains the JSON schema for the database, as specified in RFC 7047. Only this record is mandatory (a standalone file that contains only a schema represents an empty database).

The second and subsequent records in a standalone database are transaction records. Each record may have the following optional special members, which do not have any semantics but are often useful to administrators looking through a database log with `ovsdb-tool show-log`:

"_date": <integer>
: The time at which the transaction was committed, as an integer number of milliseconds since the Unix epoch. Early versions of OVSDB counted seconds instead of milliseconds; these can be detected by noticing that their values are less than 2**32.

  OVSDB always writes a `_date` member.

`"_comment": <string>`
: A JSON string that specifies the comment provided in a transaction `comment` operation. If a transaction has multiple `comment` operations, OVSDB concatenates them into a single `_comment` member, separated by a newline.

  OVSDB only writes a `_comment` member if it would be a nonempty string.

Each of these records also has one or more additional members, each of which maps from the name of a database table to a <table-txn>:

<table-txn>
: A JSON object that describes the effects of a transaction on a database table. Its names are <raw-uuid>s for rows in the table and its values are <row-txn>s.

<row-txn>
: Either `null`, which indicates that the transaction deleted this row, or a JSON object that describes how the transaction inserted or modified the row, whose names are the names of columns and whose values are <value>s that give the column's new value.

  For new rows, the OVSDB implementation omits columns whose values have the default values for their types defined in RFC 7047 section 5.2.1; for modified rows, the OVSDB implementation omits columns whose values are unchanged.

Clustered Format

The clustered format has the following additional notation:

<uint64>
: A JSON integer that represents a 64-bit unsigned integer. The OVS JSON implementation only supports integers in the range -2**63 through 2**63-1, so 64-bit unsigned integer values from 2**63 through 2**64-1 are expressed as negative numbers.

<address>
: A JSON string that represents a network address to support clustering, in the `<protocol>:<ip>:<port>` syntax described in `ovsdb-tool(1)`.

<servers>
: A JSON object whose names are <raw-uuid>s that identify servers and whose values are <address>es that specify those servers' addresses.

<cluster-txn>
: A JSON array with two elements:

  1. The first element is either a <database-schema> or `null`. A <database-schema> element is always present in the first record of a clustered database to indicate the database's initial schema. If it is not `null` in a later record, it indicates a change of schema for the database.
  2. The second element is either a transaction record in the format described under `Standalone Format` above, or `null`.

  When a schema is present, the transaction record is relative to an empty database. That is, a schema change effectively resets the database to empty and the transaction record represents the full database contents. This allows readers to be ignorant of the full semantics of schema change.

The first record in a clustered database contains the following members, all of which are required, except `prev_election_timer`:

`"server_id": <raw-uuid>`
: The server's own UUID, which must be unique within the cluster.

"local_address": <address>
The address on which the server listens for connections from other servers in the cluster.

"name": <id>
The database schema name. It is only important when a server is in the process of joining a cluster: a server will only join a cluster if the name matches. (If the database schema name were unique, then we would not also need a cluster ID.)

"cluster_id": <raw-uuid>
The cluster's UUID. The all-zeros UUID is not a valid cluster ID.

"prev_term": <uint64> and "prev_index": <uint64>
The Raft term and index just before the beginning of the log.

"prev_servers": <servers>
The set of one or more servers in the cluster at index "prev_index" and term "prev_term". It might not include this server, if it was not the initial server in the cluster.

"prev_election_timer": <uint64>
The election base time before the beginning of the log. If not exist, the default value 1000 ms is used as if it exists this record.

"prev_data": <json-value> and "prev_eid": <raw-uuid>
A snapshot of the data in the database at index "prev_index" and term "prev_term", and the entry ID for that data. The snapshot must contain a schema.

The second and subsequent records, if present, in a clustered database represent changes to the database, to the cluster state, or both. There are several types of these records. The most important types of records directly represent persistent state described in the Raft specification:

Entry
A Raft log entry.

Term
The start of a new term.

Vote
The server's vote for a leader in the current term.

The following additional types of records aid debugging and troubleshooting, but they do not affect correctness.

Leader
Identifies a newly elected leader for the current term.

Commit Index
An update to the server's `commit_index`.

Note
A human-readable description of some event.

The table below identifies the members that each type of record contains. "yes" indicates that a member is required, "?" that it is optional, blank that it is forbidden, and [1] that `data` and `eid` must be either both present or both absent.

member	Entry	Term	Vote	Leader	Commit Index	Note
comment	?	?	?	?	?	?
term	yes	yes	yes	yes		
index	yes					
servers	?					
election_timer	?					
data	[1]					
eid	[1]					
vote			yes			
leader				yes		
commit_index					yes	
note						yes

The members are:

"comment": <string>
A human-readable string giving an administrator more information about the reason a record was emitted.

"term": <uint64>
The term in which the activity occurred.

"index": <uint64>
The index of a log entry.

"servers": <servers>
Server configuration in a log entry.

"election_timer": <uint64>
Leader election timeout base value in a log entry.

"data": <json-value>
The data in a log entry.

"eid": <raw-uuid>
Entry ID in a log entry.

"vote": <raw-uuid>
The server ID for which this server voted.

"leader": <raw-uuid>
The server ID of the server. Emitted by both leaders and followers when a leader is elected.

"commit_index": <uint64>
Updated `commit_index` value.

"note": <string>
One of a few special strings indicating important events. The currently defined strings are:

"transfer leadership"
This server transferred leadership to a different server (with details included in `comment`).

"left"
This server finished leaving the cluster. (This lets subsequent readers know that the server is not part of the cluster and should not attempt to connect to it.)

Joining a Cluster

In addition to general format for a clustered database, there is also a special case for a database file created by `ovsdb-tool join-cluster`. Such a file contains exactly one record, which conveys the information passed to the `join-cluster` command. It has the following members:

"server_id": <raw-uuid> and "local_address": <address> and "name": <id>
: These have the same semantics described above in the general description of the format.

"cluster_id": <raw-uuid>
: This is provided only if the user gave the `--cid` option to `join-cluster`. It has the same semantics described above.

"remote_addresses"; [<address>*]
: One or more remote servers to contact for joining the cluster.

When the server successfully joins the cluster, the database file is replaced by one described in *Clustered Format*.

6.1.14 ovsdb

Description

OVSDB, the Open vSwitch Database, is a network-accessible database system. Schemas in OVSDB specify the tables in a database and their columns' types and can include data, uniqueness, and referential integrity constraints. OVSDB offers atomic, consistent, isolated, durable transactions. RFC 7047 specifies the JSON-RPC based protocol that OVSDB clients and servers use to communicate.

The OVSDB protocol is well suited for state synchronization because it allows each client to monitor the contents of a whole database or a subset of it. Whenever a monitored portion of the database changes, the server tells the client what rows were added or modified (including the new contents) or deleted. Thus, OVSDB clients can easily keep track of the newest contents of any part of the database.

While OVSDB is general-purpose and not particularly specialized for use with Open vSwitch, Open vSwitch does use it for multiple purposes. The leading use of OVSDB is for configuring and monitoring `ovs-vswitchd(8)`, the Open vSwitch switch daemon, using the schema documented in `ovs-vswitchd.conf.db(5)`. The Open Virtual Network (OVN) project uses two OVSDB schemas, documented as part of that project. Finally, Open vSwitch includes the "VTEP" schema, documented in `vtep(5)` that many third-party hardware switches support for configuring VXLAN, although OVS itself does not directly use this schema.

The OVSDB protocol specification allows independent, interoperable implementations of OVSDB to be developed. Open vSwitch includes an OVSDB server implementation named `ovsdb-server(1)`, which supports several protocol extensions documented in its manpage, and a basic command-line OVSDB client named `ovsdb-client(1)`, as well as OVSDB client libraries for C and for Python. Open vSwitch documentation often speaks of these OVSDB implementations in Open vSwitch as simply "OVSDB," even though that is distinct from the OVSDB protocol; we make the distinction explicit only when it might otherwise be unclear from the context.

In addition to these generic OVSDB server and client tools, Open vSwitch includes tools for working with databases that have specific schemas: `ovs-vsctl` works with the `ovs-vswitchd` configuration database and `vtep-ctl` works with the VTEP database.

RFC 7047 specifies the OVSDB protocol but it does not specify an on-disk storage format. Open vSwitch includes `ovsdb-tool(1)` for working with its own on-disk database formats. The most notable feature of this format is that `ovsdb-tool(1)` makes it easy for users to print the transactions that have changed a database since the last time it was compacted. This feature is often useful for troubleshooting.

Schemas

Schemas in OVSDB have a JSON format that is specified in RFC 7047. They are often stored in files with an extension `.ovsschema`. An on-disk database in OVSDB includes a schema and data, embedding both into a single file. The Open vSwitch utility `ovsdb-tool` has commands that work with schema files and with the schemas embedded in database files.

An Open vSwitch schema has three important identifiers. The first is its name, which is also the name used in JSON-RPC calls to identify a database based on that schema. For example, the schema used to configure Open vSwitch has the name `Open_vSwitch`. Schema names begin with a letter or an underscore, followed by any number of letters, underscores, or digits. The `ovsdb-tool` commands `schema-name` and `db-name` extract the schema name from a schema or database file, respectively.

An OVSDB schema also has a version of the form `x.y.z` e.g. `1.2.3`. Schemas managed within the Open vSwitch project manage version numbering in the following way (but OVSDB does not mandate this approach). Whenever we change the database schema in a non-backward compatible way (e.g. when we delete a column or a table), we increment <x> and set <y> and <z> to 0. When we change the database schema in a backward compatible way (e.g. when we add a new column), we increment <y> and set <z> to 0. When we change the database schema cosmetically (e.g. we reindent its syntax), we increment <z>. The `ovsdb-tool` commands `schema-version` and `db-version` extract the schema version from a schema or database file, respectively.

Very old OVSDB schemas do not have a version, but RFC 7047 mandates it.

An OVSDB schema optionally has a "checksum." RFC 7047 does not specify the use of the checksum and recommends that clients ignore it. Open vSwitch uses the checksum to remind developers to update the version: at build time, if the schema's embedded checksum, ignoring the checksum field itself, does not match the schema's content, then it fails the build with a recommendation to update the version and the checksum. Thus, a developer who changes the schema, but does not update the version, receives an automatic reminder. In practice this has been an effective way to ensure compliance with the version number policy. The `ovsdb-tool` commands `schema-cksum` and `db-cksum` extract the schema checksum from a schema or database file, respectively.

Service Models

OVSDB supports four service models for databases: **standalone**, **active-backup**, **relay** and **clustered**. The service models provide different compromises among consistency, availability, and partition tolerance. They also differ in the number of servers required and in terms of performance. The standalone and active-backup database service models share one on-disk format, and clustered databases use a different format, but the OVSDB programs work with both formats. `ovsdb(5)` documents these file formats. Relay databases have no on-disk storage.

RFC 7047, which specifies the OVSDB protocol, does not mandate or specify any particular service model.

The following sections describe the individual service models.

Standalone Database Service Model

A **standalone** database runs a single server. If the server stops running, the database becomes inaccessible, and if the server's storage is lost or corrupted, the database's content is lost. This service model is appropriate when the database controls a process or activity to which it is linked via "fate-sharing." For example, an OVSDB instance that controls an Open vSwitch virtual switch daemon, `ovs-vswitchd`, is a standalone database because a server failure would take out both the database and the virtual switch.

To set up a standalone database, use `ovsdb-tool create` to create a database file, then run `ovsdb-server` to start the database service.

To configure a client, such as `ovs-vswitchd` or `ovs-vsctl`, to use a standalone database, configure the server to listen on a "connection method" that the client can reach, then point the client to that connection method. See *Connection Methods* below for information about connection methods.

Open vSwitch 3.3 introduced support for configuration files via `--config-file` command line option. The configuration file for a server with a **standalone** database may look like this:

```
{
    "remotes": { "<connection method>": {} },
    "databases": { "<database file>": {} }
}
```

`ovsdb-server` will infer the service model from the database file itself. However, if additional verification is desired, an optional `"service-model": "standalone"` can be provided for the database file inside the inner curly braces. If the specified `service-model` will not match the content of the database file, `ovsdb-server` will refuse to open this database.

Active-Backup Database Service Model

An **active-backup** database runs two servers (on different hosts). At any given time, one of the servers is designated with the **active** role and the other the **backup** role. An active server behaves just like a standalone server. A backup server makes an OVSDB connection to the active server and uses it to continuously replicate its content as it changes in real time. OVSDB clients can connect to either server but only the active server allows data modification or lock transactions.

Setup for an active-backup database starts from a working standalone database service, which is initially the active server. On another node, to set up a backup server, create a database file with the same schema as the active server. The initial contents of the database file do not matter, as long as the schema is correct, so `ovsdb-tool create` will work, as will copying the database file from the active server. Then use `ovsdb-server --sync-from=<active>` to start the backup server, where <active> is an OVSDB connection method (see *Connection Methods* below) that connects to the active server. At that point, the backup server will fetch a copy of the active database and keep it up-to-date until it is killed.

Open vSwitch 3.3 introduced support for configuration files via `--config-file` command line option. The configuration file for a backup server in this case may look like this:

```
{
    "remotes": { "<connection method>": {} },
    "databases": {
        "<database file>": {
            "service-model": "active-backup",
            "backup": true,
            "source": {
                "<active>": {
                    "inactivity-probe": <integer>,
                    "max-backoff": <integer>
                }
            }
        }
    }
}
```

All the fields in the `"<database file>"` description above are required. Options for the `"<active>"` connection method (`"inactivity-probe"`, etc.) can be omitted.

When the active server in an active-backup server pair fails, an administrator can switch the backup server to an active role with the `ovs-appctl` command `ovsdb-server/disconnect-active-ovsdb-server`. Clients then have read/write access to the now-active server. When the `--config-file` is in use, the same can be achieved by changing the `"backup"` value in the file and running `ovsdb-server/reload` command. Of course, administrators are slow to respond compared to software, so in practice external management software detects the active server's failure and

changes the backup server's role. For example, the "Integration Guide for Centralized Control" in the OVN documentation describes how to use Pacemaker for this purpose in OVN.

Suppose an active server fails and its backup is promoted to active. If the failed server is revived, it must be started as a backup server. Otherwise, if both servers are active, then they may start out of sync, if the database changed while the server was down, and they will continue to diverge over time. This also happens if the software managing the database servers cannot reach the active server and therefore switches the backup to active, but other hosts can reach both servers. These "split-brain" problems are unsolvable in general for server pairs.

Compared to a standalone server, the active-backup service model somewhat increases availability, at a risk of split-brain. It adds generally insignificant performance overhead. On the other hand, the clustered service model, discussed below, requires at least 3 servers and has greater performance overhead, but it avoids the need for external management software and eliminates the possibility of split-brain.

Open vSwitch 2.6 introduced support for the active-backup service model.

Important

There was a change of a database file format in version 2.15. To upgrade/downgrade the `ovsdb-server` processes across this version follow the instructions described under *Upgrading from version 2.14 and earlier to 2.15 and later* and *Downgrading from version 2.15 and later to 2.14 and earlier*.

Another change happened in version 3.2. To upgrade/downgrade the `ovsdb-server` processes across this version follow the instructions described under *Upgrading from version 3.1 and earlier to 3.2 and later* and *Downgrading from version 3.2 and later to 3.1 and earlier*.

Clustered Database Service Model

A **clustered** database runs across 3 or 5 or more database servers (the **cluster**) on different hosts. Servers in a cluster automatically synchronize writes within the cluster. A 3-server cluster can remain available in the face of at most 1 server failure; a 5-server cluster tolerates up to 2 failures. Clusters larger than 5 servers will also work, with every 2 added servers allowing the cluster to tolerate 1 more failure, but write performance decreases. The number of servers should be odd: a 4- or 6-server cluster cannot tolerate more failures than a 3- or 5-server cluster, respectively.

To set up a clustered database, first initialize it on a single node by running `ovsdb-tool create-cluster`, then start `ovsdb-server`. Depending on its arguments, the `create-cluster` command can create an empty database or copy a standalone database's contents into the new database.

Open vSwitch 3.3 introduced support for configuration files via `--config-file` command line option. The configuration file for a server with a **clustered** database may look like this:

```
{
    "remotes": { "<connection method>": {} },
    "databases": { "<database file>": {} }
}
```

`ovsdb-server` will infer the service model from the database file itself. However, if additional verification is desired, an optional `"service-model": "clustered"` can be provided for the database file inside the inner curly braces. If the specified `service-model` will not match the content of the database file, `ovsdb-server` will refuse to open this database.

To configure a client to use a clustered database, first configure all of the servers to listen on a connection method that the client can reach, then point the client to all of the servers' connection methods, comma-separated. See *Connection Methods*, below, for more detail.

Open vSwitch 2.9 introduced support for the clustered service model.

How to Maintain a Clustered Database

To add a server to a cluster, run `ovsdb-tool join-cluster` on the new server and start `ovsdb-server`. To remove a running server from a cluster, use `ovs-appctl` to invoke the `cluster/leave` command. When a server fails and cannot be recovered, e.g. because its hard disk crashed, or to otherwise remove a server that is down from a cluster, use `ovs-appctl` to invoke `cluster/kick` to make the remaining servers kick it out of the cluster.

The above methods for adding and removing servers only work for healthy clusters, that is, for clusters with no more failures than their maximum tolerance. For example, in a 3-server cluster, the failure of 2 servers prevents servers joining or leaving the cluster (as well as database access).

To prevent data loss or inconsistency, the preferred solution to this problem is to bring up enough of the failed servers to make the cluster healthy again, then if necessary remove any remaining failed servers and add new ones. If this is not an option, see the next section for *Manual cluster recovery*.

Once a server leaves a cluster, it may never rejoin it. Instead, create a new server and join it to the cluster.

The servers in a cluster synchronize data over a cluster management protocol that is specific to Open vSwitch; it is not the same as the OVSDB protocol specified in RFC 7047. For this purpose, a server in a cluster is tied to a particular IP address and TCP port, which is specified in the `ovsdb-tool` command that creates or joins the cluster. The TCP port used for clustering must be different from that used for OVSDB clients. To change the port or address of a server in a cluster, first remove it from the cluster, then add it back with the new address.

To upgrade the `ovsdb-server` processes in a cluster from one version of Open vSwitch to another, upgrading them one at a time will keep the cluster healthy during the upgrade process. (This is different from upgrading a database schema, which is covered later under *Upgrading or Downgrading a Database*.)

> **Important**
>
> There was a change of a database file format in version 2.15. To upgrade/downgrade the `ovsdb-server` processes across this version follow the instructions described under *Upgrading from version 2.14 and earlier to 2.15 and later* and *Downgrading from version 2.15 and later to 2.14 and earlier*.
>
> Another change happened in version 3.2. To upgrade/downgrade the `ovsdb-server` processes across this version follow the instructions described under *Upgrading from version 3.1 and earlier to 3.2 and later* and *Downgrading from version 3.2 and later to 3.1 and earlier*.

Clustered OVSDB does not support the OVSDB "ephemeral columns" feature. `ovsdb-tool` and `ovsdb-client` change ephemeral columns into persistent ones when they work with schemas for clustered databases. Future versions of OVSDB might add support for this feature.

Manual cluster recovery

> **Important**
>
> The procedure below will result in `cid` and `sid` change. A *new* cluster will be initialized.

To recover a clustered database after a failure:

1. Stop *all* old cluster `ovsdb-server` instances before proceeding.
2. Pick one of the old members which will serve as a bootstrap member of the to-be-recovered cluster.
3. Convert its database file to the standalone format using `ovsdb-tool cluster-to-standalone`.
4. Backup the standalone database file.

5. Create a new single-node cluster with `ovsdb-tool create-cluster` using the previously saved standalone database file, then start `ovsdb-server`.
6. Once the single-node cluster is up and running and serves the restored data, new members should be created and added to the cluster, as usual, with `ovsdb-tool join-cluster`.

> **Note**
>
> The data in the new cluster may be inconsistent with the former cluster: transactions not yet replicated to the server chosen in step 2 will be lost, and transactions not yet applied to the cluster may be committed.

Upgrading from version 2.14 and earlier to 2.15 and later

There is a change of a database file format in version 2.15 that doesn't allow older versions of `ovsdb-server` to read the database file modified by the `ovsdb-server` version 2.15 or later. This also affects runtime communications between servers in **active-backup** and **cluster** service models. To upgrade the `ovsdb-server` processes from one version of Open vSwitch (2.14 or earlier) to another (2.15 or higher) instructions below should be followed. (This is different from upgrading a database schema, which is covered later under *Upgrading or Downgrading a Database*.)

In case of **standalone** service model no special handling during upgrade is required.

For the **active-backup** service model, administrator needs to update backup `ovsdb-server` first and the active one after that, or shut down both servers and upgrade at the same time.

For the **cluster** service model recommended upgrade strategy is following:

1. Upgrade processes one at a time. Each `ovsdb-server` process after upgrade should be started with `--disable-file-column-diff` command line argument.
2. When all `ovsdb-server` processes upgraded, use `ovs-appctl` to invoke `ovsdb/file/column-diff-enable` command on each of them or restart all `ovsdb-server` processes one at a time without `--disable-file-column-diff` command line option.

Downgrading from version 2.15 and later to 2.14 and earlier

Similar to upgrading covered under *Upgrading from version 2.14 and earlier to 2.15 and later*, downgrading from the `ovsdb-server` version 2.15 and later to 2.14 and earlier requires additional steps. (This is different from upgrading a database schema, which is covered later under *Upgrading or Downgrading a Database*.)

For all service models it's required to:

1. Stop all `ovsdb-server` processes (single process for **standalone** service model, all involved processes for **active-backup** and **cluster** service models).
2. Compact all database files with `ovsdb-tool compact` command.
3. Downgrade and restart `ovsdb-server` processes.

Upgrading from version 3.1 and earlier to 3.2 and later

There is another change of a database file format in version 3.2 that doesn't allow older versions of `ovsdb-server` to read the database file modified by the `ovsdb-server` version 3.2 or later. This also affects runtime communications between servers in **cluster** service models. To upgrade the `ovsdb-server` processes from one version of Open vSwitch (3.1 or earlier) to another (3.2 or higher) instructions below should be followed. (This is different from upgrading a database schema, which is covered later under *Upgrading or Downgrading a Database*.)

In case of **standalone** or **active-backup** service model no special handling during upgrade is required.

For the **cluster** service model recommended upgrade strategy is following:

1. Upgrade processes one at a time. Each `ovsdb-server` process after upgrade should be started with `--disable-file-no-data-conversion` command line argument.
2. When all `ovsdb-server` processes upgraded, use `ovs-appctl` to invoke `ovsdb/file/no-data-conversion-enable` command on each of them or restart all `ovsdb-server` processes one at a time without `--disable-file-no-data-conversion` command line option.

Downgrading from version 3.2 and later to 3.1 and earlier

Similar to upgrading covered under *Upgrading from version 3.1 and earlier to 3.2 and later*, downgrading from the `ovsdb-server` version 3.2 and later to 3.1 and earlier requires additional steps. (This is different from upgrading a database schema, which is covered later under *Upgrading or Downgrading a Database*.)

For all service models it's required to:

1. Compact all database files via `ovsdb-server/compact` command with `ovs-appctl` utility. This should be done for each involved `ovsdb-server` process separately (single process for **standalone** service model, all involved processes for **active-backup** and **cluster** service models).
2. Stop all `ovsdb-server` processes. Make sure that no database schema conversion operations were performed between steps 1 and 2. For **standalone** and **active-backup** service models, the database compaction can be performed after stopping all the processes instead with the `ovsdb-tool compact` command.
3. Downgrade and restart `ovsdb-server` processes.

Understanding Cluster Consistency

To ensure consistency, clustered OVSDB uses the Raft algorithm described in Diego Ongaro's Ph.D. thesis, "Consensus: Bridging Theory and Practice". In an operational Raft cluster, at any given time a single server is the "leader" and the other nodes are "followers". Only the leader processes transactions, but a transaction is only committed when a majority of the servers confirm to the leader that they have written it to persistent storage.

In most database systems, read and write access to the database happens through transactions. In such a system, Raft allows a cluster to present a strongly consistent transactional interface. OVSDB uses conventional transactions for writes, but clients often effectively do reads a different way, by asking the server to "monitor" a database or a subset of one on the client's behalf. Whenever monitored data changes, the server automatically tells the client what changed, which allows the client to maintain an accurate snapshot of the database in its memory. Of course, at any given time, the snapshot may be somewhat dated since some of it could have changed without the change notification yet being received and processed by the client.

Given this unconventional usage model, OVSDB also adopts an unconventional clustering model. Each server in a cluster acts independently for the purpose of monitors and read-only transactions, without verifying that data is up-to-date with the leader. Servers forward transactions that write to the database to the leader for execution, ensuring consistency. This has the following consequences:

- Transactions that involve writes, against any server in the cluster, are linearizable if clients take care to use correct prerequisites, which is the same condition required for linearizability in a standalone OVSDB. (Actually, "at-least-once" consistency, because OVSDB does not have a session mechanism to drop duplicate transactions if a connection drops after the server commits it but before the client receives the result.)
- Read-only transactions can yield results based on a stale version of the database, if they are executed against a follower. Transactions on the leader always yield fresh results. (With monitors, as explained above, a client can always see stale data even without clustering, so clustering does not change the consistency model for monitors.)
- Monitor-based (or read-heavy) workloads scale well across a cluster, because clustering OVSDB adds no additional work or communication for reads and monitors.
- A write-heavy client should connect to the leader, to avoid the overhead of followers forwarding transactions to the leader.

- When a client conducts a mix of read and write transactions across more than one server in a cluster, it can see inconsistent results because a read transaction might read stale data whose updates have not yet propagated from the leader. By default, utilities such as `ovn-sbctl` (in OVN) connect to the cluster leader to avoid this issue.

  The same might occur for transactions against a single follower except that the OVSDB server ensures that the results of a write forwarded to the leader by a given server are visible at that server before it replies to the requesting client.

- If a client uses a database on one server in a cluster, then another server in the cluster (perhaps because the first server failed), the client could observe stale data. Clustered OVSDB clients, however, can use a column in the `_Server` database to detect that data on a server is older than data that the client previously read. The OVSDB client library in Open vSwitch uses this feature to avoid servers with stale data.

Relay Service Model

A **relay** database is a way to scale out read-mostly access to the existing database working in any service model including relay.

Relay database creates and maintains an OVSDB connection with another OVSDB server. It uses this connection to maintain an in-memory copy of the remote database (a.k.a. the `relay source`) keeping the copy up-to-date as the database content changes on the relay source in the real time.

The purpose of relay server is to scale out the number of database clients. Read-only transactions and monitor requests are fully handled by the relay server itself. For the transactions that request database modifications, relay works as a proxy between the client and the relay source, i.e. it forwards transactions and replies between them.

Compared to the clustered and active-backup models, relay service model provides read and write access to the database similarly to a clustered database (and even more scalable), but with generally insignificant performance overhead of an active-backup model. At the same time it doesn't increase availability that needs to be covered by the service model of the relay source.

Relay database has no on-disk storage and therefore cannot be converted to any other service model.

If there is already a database started in any service model, to start a relay database server use `ovsdb-server relay:<DB_NAME>:<relay source>`, where `<DB_NAME>` is the database name as specified in the schema of the database that existing server runs, and `<relay source>` is an OVSDB connection method (see *Connection Methods* below) that connects to the existing database server. `<relay source>` could contain a comma-separated list of connection methods, e.g. to connect to any server of the clustered database. Multiple relay servers could be started for the same relay source.

Open vSwitch 3.3 introduced support for configuration files via `--config-file` command line option. The configuration file for a relay database server in this case may look like this:

```
{
    "remotes": { "<connection method>": {} },
    "databases": {
        "<DB_NAME>": {
            "service-model": "relay",
            "source": {
                "<relay source>": {
                    "inactivity-probe": <integer>,
                    "max-backoff": <integer>
                }
            }
        }
    }
}
```

Both the "service-model" and the "source" are required. Options for the "<relay source>" connection method ("inactivity-probe", etc.) can be omitted.

Since the way relays handle read and write transactions is very similar to the clustered model where "cluster" means "set of relay servers connected to the same relay source", "follower" means "relay server" and the "leader" means "relay source", same consistency consequences as for the clustered model applies to relay as well (See *Understanding Cluster Consistency* above).

Open vSwitch 2.16 introduced support for relay service model.

Database Replication

OVSDB can layer **replication** on top of any of its service models. Replication, in this context, means to make, and keep up-to-date, a read-only copy of the contents of a database (the replica). One use of replication is to keep an up-to-date backup of a database. A replica used solely for backup would not need to support clients of its own. A set of replicas that do serve clients could be used to scale out read access to the primary database, however *Relay Service Model* is more suitable for that purpose.

A database replica is set up in the same way as a backup server in an active-backup pair, with the difference that the replica is never promoted to an active role.

A database can have multiple replicas.

Open vSwitch 2.6 introduced support for database replication.

Connection Methods

An OVSDB **connection method** is a string that specifies how to make a JSON-RPC connection between an OVSDB client and server. Connection methods are part of the Open vSwitch implementation of OVSDB and not specified by RFC 7047. ovsdb-server uses connection methods to specify how it should listen for connections from clients and ovsdb-client uses them to specify how it should connect to a server. Connections in the opposite direction, where ovsdb-server connects to a client that is configured to listen for an incoming connection, are also possible.

Connection methods are classified as **active** or **passive**. An active connection method makes an outgoing connection to a remote host; a passive connection method listens for connections from remote hosts. The most common arrangement is to configure an OVSDB server with passive connection methods and clients with active ones, but the OVSDB implementation in Open vSwitch supports the opposite arrangement as well.

OVSDB supports the following active connection methods:

ssl:<host>:<port>
: The specified SSL/TLS <port> on the given <host>.

tcp:<host>:<port>
: The specified TCP <port> on the given <host>.

unix:<file>
: On Unix-like systems, connect to the Unix domain server socket named <file>.

<method1>,<method2>,...,<methodN>
: For a clustered database service to be highly available, a client must be able to connect to any of the servers in the cluster. To do so, specify connection methods for each of the servers separated by commas (and optional spaces).

  In theory, if machines go up and down and IP addresses change in the right way, a client could talk to the wrong instance of a database. To avoid this possibility, add cid:<uuid> to the list of methods, where <uuid> is the cluster ID of the desired database cluster, as printed by ovsdb-tool db-cid. This feature is optional.

OVSDB supports the following passive connection methods:

pssl:<port>[:<ip>]
Listen on the given TCP <port> for SSL/TLS connections. By default, connections are not bound to a particular local IP address. Specifying <ip> limits connections to those from the given IP.

ptcp:<port>[:<ip>]
Listen on the given TCP <port>. By default, connections are not bound to a particular local IP address. Specifying <ip> limits connections to those from the given IP.

punix:<file>
On Unix-like systems, listens for connections on the Unix domain socket named <file>.

pfd:<fd>
Listen on a pre-opened file descriptor <fd>. The file descriptor must refer to a bound, listening Unix domain stream socket. This is intended for use with systemd socket activation, where systemd opens the socket and passes it to the service.

For security, `pfd:` may only be specified on the command line (`--remote=pfd:<fd>`). It is rejected if added at runtime via `ovsdb-server/add-remote` or through the database.

All IP-based connection methods accept IPv4 and IPv6 addresses. To specify an IPv6 address, wrap it in square brackets, e.g. `ssl:[::1]:6640`. Passive IP-based connection methods by default listen for IPv4 connections only; use `[::]` as the address to accept both IPv4 and IPv6 connections, e.g. `pssl:6640:[::]`. DNS names are also accepted if built with unbound library. On Linux, use `%<device>` to designate a scope for IPv6 link-level addresses, e.g. `ssl:[fe80::1234%eth0]:6653`.

The <port> may be omitted from connection methods that use a port number. The default <port> for TCP-based connection methods is 6640, e.g. `pssl:` is equivalent to `pssl:6640`. In Open vSwitch prior to version 2.4.0, the default port was 6632. To avoid incompatibility between older and newer versions, we encourage users to specify a port number.

The `ssl` and `pssl` connection methods requires additional configuration through `--private-key`, `--certificate`, and `--ca-cert` command line options. Open vSwitch can be built without SSL/TLS support, in which case these connection methods are not supported.

Database Life Cycle

This section describes how to handle various events in the life cycle of a database using the Open vSwitch implementation of OVSDB.

Creating a Database

Creating and starting up the service for a new database was covered separately for each database service model in the *Service Models* section, above. A single `ovsdb-server` process may serve any number of databases with different service models at the same time.

Backing Up and Restoring a Database

OVSDB is often used in contexts where the database contents are not particularly valuable. For example, in many systems, the database for configuring `ovs-vswitchd` is essentially rebuilt from scratch at boot time. It is not worthwhile to back up these databases.

When OVSDB is used for valuable data, a backup strategy is worth considering. One way is to use database replication, discussed above in *Database Replication* which keeps an online, up-to-date copy of a database, possibly on a remote system. This works with all OVSDB service models.

A more common backup strategy is to periodically take and store a snapshot. For the standalone and active-backup service models, making a copy of the database file, e.g. using `cp`, effectively makes a snapshot, and because OVSDB database files are append-only, it works even if the database is being modified when the snapshot takes place. This approach does not work for clustered databases.

Another way to make a backup, which works with all OVSDB service models, is to use `ovsdb-client backup`, which connects to a running database server and outputs an atomic snapshot of its schema and content, in the same format used for standalone and active-backup databases.

Multiple options are also available when the time comes to restore a database from a backup. For the standalone and active-backup service models, one option is to stop the database server or servers, overwrite the database file with the backup (e.g. with `cp`), and then restart the servers. Another way, which works with any service model, is to use `ovsdb-client restore`, which connects to a running database server and replaces the data in one of its databases by a provided snapshot. The advantage of `ovsdb-client restore` is that it causes zero downtime for the database and its server. It has the downside that UUIDs of rows in the restored database will differ from those in the snapshot, because the OVSDB protocol does not allow clients to specify row UUIDs.

None of these approaches saves and restores data in columns that the schema designates as ephemeral. This is by design: the designer of a schema only marks a column as ephemeral if it is acceptable for its data to be lost when a database server restarts.

Clustering and backup serve different purposes. Clustering increases availability, but it does not protect against data loss if, for example, a malicious or malfunctioning OVSDB client deletes or tampers with data.

Changing Database Service Model

Use `ovsdb-tool create-cluster` to create a clustered database from the contents of a standalone database. Use `ovsdb-client backup` to create a standalone database from the contents of a running clustered database. When the cluster is down and cannot be revived, `ovsdb-client backup` will not work.

Use `ovsdb-tool cluster-to-standalone` to convert clustered database to standalone database when the cluster is down and cannot be revived.

Upgrading or Downgrading a Database

The evolution of a piece of software can require changes to the schemas of the databases that it uses. For example, new features might require new tables or new columns in existing tables, or conceptual changes might require a database to be reorganized in other ways. In some cases, the easiest way to deal with a change in a database schema is to delete the existing database and start fresh with the new schema, especially if the data in the database is easy to reconstruct. But in many other cases, it is better to convert the database from one schema to another.

The OVSDB implementation in Open vSwitch has built-in support for some simple cases of converting a database from one schema to another. This support can handle changes that add or remove database columns or tables or that eliminate constraints (for example, changing a column that must have exactly one value into one that has one or more values). It can also handle changes that add constraints or make them stricter, but only if the existing data in the database satisfies the new constraints (for example, changing a column that has one or more values into a column with exactly one value, if every row in the column has exactly one value). The built-in conversion can cause data loss in obvious ways, for example if the new schema removes tables or columns, or indirectly, for example by deleting unreferenced rows in tables that the new schema marks for garbage collection.

Converting a database can lose data, so it is wise to make a backup beforehand.

To use OVSDB's built-in support for schema conversion with a standalone or active-backup database, first stop the database server or servers, then use `ovsdb-tool convert` to convert it to the new schema, and then restart the database server.

OVSDB also supports online database schema conversion for any of its database service models. To convert a database online, use `ovsdb-client convert`. The conversion is atomic, consistent, isolated, and durable. `ovsdb-server` disconnects any clients connected when the conversion takes place (except clients that use the `set_db_change_aware` Open vSwitch extension RPC). Upon reconnection, clients will discover that the schema has changed.

Schema versions and checksums (see *Schemas* above) can give hints about whether a database needs to be converted to a new schema. If there is any question, though, the `needs-conversion` command on `ovsdb-tool` and `ovsdb-client` can provide a definitive answer.

Working with Database History

Both on-disk database formats that OVSDB supports are organized as a stream of transaction records. Each record describes a change to the database as a list of rows that were inserted or deleted or modified, along with the details. Therefore, in normal operation, a database file only grows, as each change causes another record to be appended at the end. Usually, a user has no need to understand this file structure. This section covers some exceptions.

Compacting Databases

If OVSDB database files were truly append-only, then over time they would grow without bound. To avoid this problem, OVSDB can **compact** a database file, that is, replace it by a new version that contains only the current database contents, as if it had been inserted by a single transaction. From time to time, `ovsdb-server` automatically compacts a database that grows much larger than its minimum size.

Because `ovsdb-server` automatically compacts databases, it is usually not necessary to compact them manually, but OVSDB still offers a few ways to do it. First, `ovsdb-tool compact` can compact a standalone or active-backup database that is not currently being served by `ovsdb-server` (or otherwise locked for writing by another process). To compact any database that is currently being served by `ovsdb-server`, use `ovs-appctl` to send the `ovsdb-server/compact` command. Each server in an active-backup or clustered database maintains its database file independently, so to compact all of them, issue this command separately on each server.

Viewing History

The `ovsdb-tool` utility's `show-log` command displays the transaction records in an OVSDB database file in a human-readable format. By default, it shows minimal detail, but adding the option `-m` once or twice increases the level of detail. In addition to the transaction data, it shows the time and date of each transaction and any "comment" added to the transaction by the client. The comments can be helpful for quickly understanding a transaction; for example, `ovs-vsctl` adds its command line to the transactions that it makes.

The `show-log` command works with both OVSDB file formats, but the details of the output format differ. For active-backup and clustered databases, the sequence of transactions in each server's log will differ, even at points when they reflect the same data.

Truncating History

It may occasionally be useful to "roll back" a database file to an earlier point. Because of the organization of OVSDB records, this is easy to do. Start by noting the record number <i> of the first record to delete in `ovsdb-tool show-log` output. Each record is two lines of plain text, so trimming the log is as simple as running `head -n` <j>, where <j> = 2 * <i>.

Corruption

When `ovsdb-server` opens an OVSDB database file, of any kind, it reads as many transaction records as it can from the file until it reaches the end of the file or it encounters a corrupted record. At that point it stops reading and regards the data that it has read to this point as the full contents of the database file, effectively rolling the database back to an earlier point.

Each transaction record contains an embedded SHA-1 checksum, which the server verifies as it reads a database file. It detects corruption when a checksum fails to verify. Even though SHA-1 is no longer considered secure for use in cryptography, it is acceptable for this purpose because it is not used to defend against malicious attackers.

The first record in a standalone or active-backup database file specifies the schema. `ovsdb-server` will refuse to work with a database where this record is corrupted, or with a clustered database file with corruption in the first few records. Delete and recreate such a database, or restore it from a backup.

When `ovsdb-server` adds records to a database file in which it detected corruption, it first truncates the file just after the last good record.

See Also

RFC 7047, "The Open vSwitch Database Management Protocol."

Open vSwitch implementations of generic OVSDB functionality: `ovsdb-server(1)`, `ovsdb-client(1)`, `ovsdb-tool(1)`.

Tools for working with databases that have specific OVSDB schemas: `ovs-vsctl(8)`, `vtep-ctl(8)`, and (in OVN) `ovn-nbctl(8)`, `ovn-sbctl(8)`.

OVSDB schemas for Open vSwitch and related functionality: `ovs-vswitchd.conf.db(5)`, `vtep(5)`, and (in OVN) `ovn-nb(5)`, `ovn-sb(5)`.

The remainder are still in roff format can be found below:

ovs-bugtool(8)	(pdf)	(html)	(plain text)
ovsdb-client(1)	(pdf)	(html)	(plain text)
ovsdb-server(1)	(pdf)	(html)	(plain text)
ovsdb-tool(1)	(pdf)	(html)	(plain text)
ovs-dpctl(8)	(pdf)	(html)	(plain text)
ovs-dpctl-top(8)	(pdf)	(html)	(plain text)
ovs-fields(7)	(pdf)	(html)	(plain text)
ovs-ofctl(8)	(pdf)	(html)	(plain text)
ovs-pcap(1)	(pdf)	(html)	(plain text)
ovs-test(8)	(pdf)	(html)	(plain text)
ovs-testcontroller(8)	(pdf)	(html)	(plain text)
ovs-vlan-test(8)	(pdf)	(html)	(plain text)
ovs-vsctl(8)	(pdf)	(html)	(plain text)
ovs-vswitchd(8)	(pdf)	(html)	(plain text)
ovs-vswitchd.conf.db(5)	(pdf)	(html)	(plain text)
vtep(5)	(pdf)	(html)	(plain text)
vtep-ctl(8)	(pdf)	(html)	(plain text)

CHAPTER

SEVEN

OPEN VSWITCH INTERNALS

Information for people who want to know more about the Open vSwitch project itself and how they might involved.

7.1 Contributing to Open vSwitch

The below guides provide information on contributing to Open vSwitch itself.

7.1.1 Submitting Patches

Send changes to Open vSwitch as patches to dev@openvswitch.org. One patch per email. More details are included below.

If you are using Git, then *git format-patch* takes care of most of the mechanics described below for you.

Before You Start

Before you send patches at all, make sure that each patch makes sense. In particular:

- A given patch should not break anything, even if later patches fix the problems that it causes. The source tree should still build and work after each patch is applied. (This enables *git bisect* to work best.)
- A patch should make one logical change. Don't make multiple, logically unconnected changes to disparate subsystems in a single patch.
- A patch that adds or removes user-visible features should also update the appropriate user documentation or manpages. Consider adding an item to NEWS for nontrivial changes. Check "Feature Deprecation Guidelines" section in this document if you intend to remove user-visible feature.

Testing is also important:

- Test a patch that modifies existing code with `make check` before submission. Refer to the "Unit Tests" in *Testing*, for more information. We also encourage running the kernel and userspace system tests.
- Consider testing a patch that adds or deletes files with `make distcheck` before submission.
- A patch that modifies Linux kernel code should be at least build-tested on various Linux kernel versions before submission. I suggest versions 3.10 and whatever the current latest release version is at the time.
- A patch that adds a new feature should add appropriate tests for the feature. A bug fix patch should preferably add a test that would fail if the bug recurs.

If you are using GitHub, then you may utilize the GitHub Actions CI build systems. They will run some of the above tests automatically when you push changes to your repository.

Email Subject

The subject line of your email should be in the following format:

> [PATCH <n>/<m>] <area>: <summary>

Where:

[PATCH <n>/<m>]:
> indicates that this is the nth of a series of m patches. It helps reviewers to read patches in the correct order. You may omit this prefix if you are sending only one patch.

<area>:
> indicates the area of the Open vSwitch to which the change applies (often the name of a source file or a directory). You may omit it if the change crosses multiple distinct pieces of code.

<summary>:

> briefly describes the change. Use the imperative form, e.g. "Force SNAT for multiple gateway routers." or "Fix daemon exit for bad datapaths or flows." Try to keep the summary short, about 50 characters wide.

The subject, minus the [PATCH <n>/<m>] prefix, becomes the first line of the commit's change log message.

Description

The body of the email should start with a more thorough description of the change. This becomes the body of the commit message, following the subject. There is no need to duplicate the summary given in the subject.

Please limit lines in the description to 75 characters in width. That allows the description to format properly even when indented (e.g. by "git log" or in email quotations).

The description should include:

- The rationale for the change.
- Design description and rationale (but this might be better added as code comments).
- Testing that you performed (or testing that should be done but you could not for whatever reason).
- Tags (see below).

There is no need to describe what the patch actually changed, if the reader can see it for himself.

If the patch refers to a commit already in the Open vSwitch repository, please include both the commit number and the subject of the patch, e.g. 'commit 632d136c (vswitch: Remove restriction on datapath names.)'.

If you, the person sending the patch, did not write the patch yourself, then the very first line of the body should take the form From: <author name> <author email>, followed by a blank line. This will automatically cause the named author to be credited with authorship in the repository.

Tags

The description ends with a series of tags, written one to a line as the last paragraph of the email. Each tag indicates some property of the patch in an easily machine-parseable manner.

Please don't wrap a tag across multiple lines. If necessary, it's OK to have a tag extend beyond the customary maximum width of a commit message.

Examples of common tags follow.

Signed-off-by: Author Name <author.name@email.address...>

> Informally, this indicates that Author Name is the author or submitter of a patch and has the authority to submit it under the terms of the license. The formal meaning is to agree to the Developer's Certificate of Origin (see below).

If the author and submitter are different, each must sign off. If the patch has more than one author, all must sign off.

Signed-off-by tags should be the last tags in the commit message. If the author (or authors) and submitter are different, the author tags should come first. More generally, occasionally a patch might pass through a chain of submitters, and in such a case the sign-offs should be arranged in chronological order.

```
Signed-off-by: Author Name <author.name@email.address...>
Signed-off-by: Submitter Name <submitter.name@email.address...>
```

`Co-authored-by: Author Name <author.name@email.address...>`

Git can only record a single person as the author of a given patch. In the rare event that a patch has multiple authors, one must be given the credit in Git and the others must be credited via Co-authored-by: tags. (All co-authors must also sign off.)

`Acked-by: Reviewer Name <reviewer.name@email.address...>`

Reviewers will often give an `Acked-by:` tag to code of which they approve. It is polite for the submitter to add the tag before posting the next version of the patch or applying the patch to the repository. Quality reviewing is hard work, so this gives a small amount of credit to the reviewer.

Not all reviewers give `Acked-by:` tags when they provide positive reviews. It's customary only to add tags from reviewers who actually provide them explicitly.

`Tested-by: Tester Name <reviewer.name@email.address...>`

When someone tests a patch, it is customary to add a Tested-by: tag indicating that. It's rare for a tester to actually provide the tag; usually the patch submitter makes the tag himself in response to an email indicating successful testing results.

`Tested-at: <URL>`

When a test report is publicly available, this provides a way to reference it. Typical <URL>s would be build logs from autobuilders or references to mailing list archives.

Some autobuilders only retain their logs for a limited amount of time. It is less useful to cite these because they may be dead links for a developer reading the commit message months or years later.

`Reported-by: Reporter Name <reporter.name@email.address...>`

When a patch fixes a bug reported by some person, please credit the reporter in the commit log in this fashion. Please also add the reporter's name and email address to the list of people who provided helpful bug reports in the AUTHORS file at the top of the source tree.

Fairly often, the reporter of a bug also tests the fix. Occasionally one sees a combined "Reported-and-tested-by:" tag used to indicate this. It is also acceptable, and more common, to include both tags separately.

(If a bug report is received privately, it might not always be appropriate to publicly credit the reporter. If in doubt, please ask the reporter.)

`Requested-by: Requester Name <requester.name@email.address...>`

When a patch implements a request or a suggestion made by some person, please credit that person in the commit log in this fashion. For a helpful suggestion, please also add the person's name and email address to the list of people who provided suggestions in the AUTHORS file at the top of the source tree.

(If a suggestion or a request is received privately, it might not always be appropriate to publicly give credit. If in doubt, please ask.)

`Suggested-by: Suggester Name <suggester.name@email.address...>`

See `Requested-by:`.

`CC: Person <name@email>`

This is a way to tag a patch for the attention of a person when no more specific tag is appropriate. One use is to request a review from a particular person. It doesn't make sense to include the same person in CC and another tag, so e.g. if someone who is CCed later provides an Acked-by, add the Acked-by and remove the CC at the same time.

`Reported-at: <URL>`

If a patch fixes or is otherwise related to a bug reported in a public bug tracker, please include a reference to the bug in the form of a URL to the specific bug, e.g.:

```
Reported-at: https://bugs.debian.org/743635
```

This is also an appropriate way to refer to bug report emails in public email archives, e.g.:

```
Reported-at: https://mail.openvswitch.org/pipermail/ovs-dev/2014-June/284495.
↪html
```

`Submitted-at: <URL>`

If a patch was submitted somewhere other than the Open vSwitch development mailing list, such as a GitHub pull request, this header can be used to reference the source.

```
Submitted-at: https://github.com/openvswitch/ovs/pull/92
```

`VMware-BZ: #1234567`

If a patch fixes or is otherwise related to a bug reported in a private bug tracker, you may include some tracking ID for the bug for your own reference. Please include some identifier to make the origin clear, e.g. "VMware-BZ" refers to VMware's internal Bugzilla instance and "ONF-JIRA" refers to the Open Networking Foundation's JIRA bug tracker.

`ONF-JIRA: EXT-12345`

See `VMware-BZ:`.

`Bug #1234567.`

These are obsolete forms of VMware-BZ: that can still be seen in old change log entries. (They are obsolete because they do not tell the reader what bug tracker is referred to.)

`Issue: 1234567`

See `Bug:`.

`Fixes: 63bc9fb1c69f ("packets: Reorder CS_* flags to remove gap.")`

If you would like to record which commit introduced a bug being fixed, you may do that with a "Fixes" header. This assists in determining which OVS releases have the bug, so the patch can be applied to all affected versions. The easiest way to generate the header in the proper format is with this git command. This command also CCs the author of the commit being fixed, which makes sense unless the author also made the fix or is already named in another tag:

```
$ git log -1 --pretty=format:"CC: %an <%ae>%nFixes: %h (\"%s\")" \
  --abbrev=12 COMMIT_REF
```

`Vulnerability: CVE-2016-2074`

Specifies that the patch fixes or is otherwise related to a security vulnerability with the given CVE identifier. Other identifiers in public vulnerability databases are also suitable.

If the vulnerability was reported publicly, then it is also appropriate to cite the URL to the report in a Reported-at tag. Use a Reported-by tag to acknowledge the reporters.

`Assisted-by: Name of model, and/or AI Code Assistant`

When a patch has been created with the assistance of an AI tool, this tag should be used to disclose that fact. Provide the name of the underlying model used, if known. For the name of the tool, only include the name, not an email address. For example:

```
Assisted-by: model-name-42.0, OVS-Code-Assistant-Pro-9.0
```

The author of the patch remains fully responsible for the content and must ensure it complies with the *Developer's Certificate of Origin*. See the *AI-assisted Contributions* section for more information.

Developer's Certificate of Origin

To help track the author of a patch as well as the submission chain, and be clear that the developer has authority to submit a patch for inclusion in Open vSwitch please sign off your work. The sign off certifies the following:

```
Developer's Certificate of Origin 1.1

By making a contribution to this project, I certify that:

(a) The contribution was created in whole or in part by me and I
    have the right to submit it under the open source license
    indicated in the file; or

(b) The contribution is based upon previous work that, to the best
    of my knowledge, is covered under an appropriate open source
    license and I have the right under that license to submit that
    work with modifications, whether created in whole or in part
    by me, under the same open source license (unless I am
    permitted to submit under a different license), as indicated
    in the file; or

(c) The contribution was provided directly to me by some other
    person who certified (a), (b) or (c) and I have not modified
    it.

(d) I understand and agree that this project and the contribution
    are public and that a record of the contribution (including all
    personal information I submit with it, including my sign-off) is
    maintained indefinitely and may be redistributed consistent with
    this project or the open source license(s) involved.
```

See also http://developercertificate.org/.

AI-assisted Contributions

OVS is a Linux Foundation Collaborative Project, and the Linux Foundation's policy for AI-generated patches can be found here: https://www.linuxfoundation.org/legal/generative-ai .

OVS allows the use of AI assistants in producing patches. Contributions which have used an AI assistant should disclose the use of the assistant by using the "Assisted-by" tag. While AI-assisted patches are allowed, the author of the patch is ultimately responsible for ensuring that the AI-generated code has not violated any terms of the Developer's Certificate of Origin.

Feature Deprecation Guidelines

Open vSwitch is intended to be user friendly. This means that under normal circumstances we don't abruptly remove features from OVS that some users might still be using. Otherwise, if we would, then we would possibly break our user setup when they upgrade and would receive bug reports.

Typical process to deprecate a feature in Open vSwitch is to:

(a) Mention deprecation of a feature in the NEWS file. Also, mention expected release or absolute time when this feature would be removed from OVS altogether. Don't use relative time (e.g. "in 6 months") because that is not clearly interpretable.

(b) If Open vSwitch is configured to use deprecated feature it should print a warning message to the log files clearly indicating that feature is deprecated and that use of it should be avoided.

(c) If this feature is mentioned in man pages, then add "Deprecated" keyword to it.

Also, if there is alternative feature to the one that is about to be marked as deprecated, then mention it in (a), (b) and (c) as well.

Remember to follow-up and actually remove the feature from OVS codebase once deprecation grace period has expired and users had opportunity to use at least one OVS release that would have informed them about feature deprecation!

Comments

If you want to include any comments in your email that should not be part of the commit's change log message, put them after the description, separated by a line that contains just `---`. It may be helpful to include a diffstat here for changes that touch multiple files.

Patch

The patch should be in the body of the email following the description, separated by a blank line.

Patches should be in `diff -up` format. We recommend that you use Git to produce your patches, in which case you should use the `-M -C` options to `git diff` (or other Git tools) if your patch renames or copies files. Quilt might be useful if you do not want to use Git.

Patches should be inline in the email message. Some email clients corrupt white space or wrap lines in patches. There are hints on how to configure many email clients to avoid this problem on kernel.org. If you cannot convince your email client not to mangle patches, then sending the patch as an attachment is a second choice.

Follow the style used in the code that you are modifying. *Coding Style* file describes the coding style used in most of Open vSwitch. Use Linux kernel coding style for Linux kernel code.

If your code is non-datapath code, you may use the `utilities/checkpatch.py` utility as a quick check for certain commonly occurring mistakes (improper leading/trailing whitespace, missing signoffs, some improper formatted patch files). For Linux datapath code, it is a good idea to use the Linux script `checkpatch.pl`.

Example

```
From fa29a1c2c17682879e79a21bb0cdd5bbe67fa7c0 Mon Sep 17 00:00:00 2001
From: Jesse Gross <jesse@nicira.com>
Date: Thu, 8 Dec 2011 13:17:24 -0800
Subject: [PATCH] datapath: Alphabetize include/net/ipv6.h compat header.

Signed-off-by: Jesse Gross <jesse@nicira.com>
---
 datapath/linux/Modules.mk |    2 +-
 1 files changed, 1 insertions(+), 1 deletions(-)
```

(continues on next page)

(continued from previous page)

```
diff --git a/datapath/linux/Modules.mk b/datapath/linux/Modules.mk
index fdd952e..f6cb88e 100644
--- a/datapath/linux/Modules.mk
+++ b/datapath/linux/Modules.mk
@@ -56,11 +56,11 @@ openvswitch_headers += \
    linux/compat/include/net/dst.h \
    linux/compat/include/net/genetlink.h \
    linux/compat/include/net/ip.h \
+   linux/compat/include/net/ipv6.h \
    linux/compat/include/net/net_namespace.h \
    linux/compat/include/net/netlink.h \
    linux/compat/include/net/protocol.h \
    linux/compat/include/net/route.h \
-   linux/compat/include/net/ipv6.h \
    linux/compat/genetlink.inc

 both_modules += brcompat
--
1.7.7.3
```

7.1.2 Backporting patches

> **Note**
>
> This is an advanced topic for developers and maintainers. Readers should familiarize themselves with building and running Open vSwitch, with the git tool, and with the Open vSwitch patch submission process.

The backporting of patches from one git tree to another takes multiple forms within Open vSwitch, but is broadly applied in the following fashion:

- Contributors submit their proposed changes to the latest development branch
- Contributors and maintainers provide feedback on the patches
- When the change is satisfactory, maintainers apply the patch to the development branch.
- Maintainers backport changes from a development branch to release branches.

With regards to Open vSwitch user space code and code that does not comprise the Linux datapath and compat code, the development branch is *main* in the Open vSwitch repository. Patches are applied first to this branch, then to the most recent *branch-X.Y*, then earlier *branch-X.Z*, and so on. The most common kind of patch in this category is a bugfix which affects main and other branches.

Changes to userspace components

Patches which are fixing bugs should be considered for backporting from *main* to release branches. Open vSwitch contributors submit their patches targeted to the *main* branch, using the `Fixes` tag described in *Submitting Patches*. The maintainer first applies the patch to *main*, then backports the patch to each older affected tree, as far back as it goes or at least to all currently supported branches. This is usually each branch back to the oldest maintained LTS release branch or the last 4 release branches if the oldest LTS is newer.

If the fix only affects a particular branch and not *main*, contributors should submit the change with the target branch listed in the subject line of the patch. Contributors should list all versions that the bug affects. The `git format-patch`

argument `--subject-prefix` may be used when posting the patch, for example:

```
$ git format-patch HEAD --subject-prefix="PATCH branch-2.7"
```

If a maintainer is backporting a change to older branches and the backport is not a trivial cherry-pick, then the maintainer may opt to submit the backport for the older branch on the mailing list for further review. This should be done in the same manner as described above.

Changes to Linux kernel components

Changes to the Linux kernel components in Open vSwitch go through review in the upstream Linux Netdev community. The Netdev Maintainer Handbook describes the general process for merging patches to the upstream Linux kernel networking subsystem.

Backports to older kernel versions are handled via the Stable tree mechanism.

Backports for Linux datapath code are no longer accepted into the Open vSwitch tree as that code is not present in the Open vSwitch distribution since Open vSwitch 3.0.

7.1.3 Coding Style

This file describes the coding style used in most C files in the Open vSwitch distribution. However, Linux kernel code datapath directory follows the Linux kernel's established coding conventions.

The following GNU indent options approximate this style.

```
-npro -bad -bap -bbb -br -blf -brs -cdw -ce -fca -cli0 -npcs -i4 -l79 \
-lc79 -nbfda -nut -saf -sai -saw -sbi4 -sc -sob -st -ncdb -pi4 -cs -bs \
-di1 -lp -il0 -hnl
```

Basics

- Limit lines to 79 characters.
- Use form feeds (control+L) to divide long source files into logical pieces. A form feed should appear as the only character on a line.
- Do not use tabs for indentation.
- Avoid trailing spaces on lines.

Naming

- Use names that explain the purpose of a function or object.
- Use underscores to separate words in an identifier: `multi_word_name`.
- Use lowercase for most names. Use uppercase for macros, macro parameters, and members of enumerations.
- Give arrays names that are plural.
- Pick a unique name prefix (ending with an underscore) for each module, and apply that prefix to all of that module's externally visible names. Names of macro parameters, struct and union members, and parameters in function prototypes are not considered externally visible for this purpose.
- Do not use names that begin with _. If you need a name for "internal use only", use __ as a suffix instead of a prefix.
- Avoid negative names: `found` is a better name than `not_found`.
- In names, a `size` is a count of bytes, a `length` is a count of characters. A buffer has size, but a string has length. The length of a string does not include the null terminator, but the size of the buffer that contains the string does.

Comments

Comments should be written as full sentences that start with a capital letter and end with a period. Put two spaces between sentences.

Write block comments as shown below. You may put the `/*` and `*/` on the same line as comment text if you prefer.

```
/*
 * We redirect stderr to /dev/null because we often want to remove all
 * traffic control configuration on a port so its in a known state.  If
 * this done when there is no such configuration, tc complains, so we just
 * always ignore it.
 */
```

Each function and each variable declared outside a function, and each struct, union, and typedef declaration should be preceded by a comment. See *functions* below for function comment guidelines.

Each struct and union member should each have an inline comment that explains its meaning. structs and unions with many members should be additionally divided into logical groups of members by block comments, e.g.:

```
/* An event that will wake the following call to poll_block(). */
struct poll_waiter {
    /* Set when the waiter is created. */
    struct ovs_list node;       /* Element in global waiters list. */
    int fd;                     /* File descriptor. */
    short int events;           /* Events to wait for (POLLIN, POLLOUT). */
    poll_fd_func *function;     /* Callback function, if any, or null. */
    void *aux;                  /* Argument to callback function. */
    struct backtrace *backtrace; /* Event that created waiter, or null. */

    /* Set only when poll_block() is called. */
    struct pollfd *pollfd;      /* Pointer to element of the pollfds array
                                   (null if added from a callback). */
};
```

Use `XXX` or `FIXME` comments to mark code that needs work.

Don't use `//` comments.

Don't comment out or `#if 0` out code. Just remove it. The code that was there will still be in version control history.

Functions

Put the return type, function name, and the braces that surround the function's code on separate lines, all starting in column 0.

Before each function definition, write a comment that describes the function's purpose, including each parameter, the return value, and side effects. References to argument names should be given in single-quotes, e.g. `'arg'`. The comment should not include the function name, nor need it follow any formal structure. The comment does not need to describe how a function does its work, unless this information is needed to use the function correctly (this is often better done with comments *inside* the function).

Simple static functions do not need a comment.

Within a file, non-static functions should come first, in the order that they are declared in the header file, followed by static functions. Static functions should be in one or more separate pages (separated by form feed characters) in logical groups. A commonly useful way to divide groups is by "level", with high-level functions first, followed by groups of progressively lower-level functions. This makes it easy for the program's reader to see the top-down structure by reading from top to bottom.

All function declarations and definitions should include a prototype. Empty parentheses, e.g. `int foo();`, do not include a prototype (they state that the function's parameters are unknown); write `void` in parentheses instead, e.g. `int foo(void);`.

Prototypes for static functions should either all go at the top of the file, separated into groups by blank lines, or they should appear at the top of each page of functions. Don't comment individual prototypes, but a comment on each group of prototypes is often appropriate.

In the absence of good reasons for another order, the following parameter order is preferred. One notable exception is that data parameters and their corresponding size parameters should be paired.

1. The primary object being manipulated, if any (equivalent to the `this` pointer in C++).
2. Input-only parameters.
3. Input/output parameters.
4. Output-only parameters.
5. Status parameter.

Example:

```
/* Stores the features supported by 'netdev' into each of '*current',
 * '*advertised', '*supported', and '*peer' that are non-null.  Each value
 * is a bitmap of "enum ofp_port_features" bits, in host byte order.
 * Returns 0 if successful, otherwise a positive errno value.  On failure,
 * all of the passed-in values are set to 0. */
int
netdev_get_features(struct netdev *netdev,
                    uint32_t *current, uint32_t *advertised,
                    uint32_t *supported, uint32_t *peer)
{
    ...
}
```

Functions that destroy an instance of a dynamically-allocated type should accept and ignore a null pointer argument. Code that calls such a function (including the C standard library function `free()`) should omit a null-pointer check. We find that this usually makes code easier to read.

Functions in `.c` files should not normally be marked `inline`, because it does not usually help code generation and it does suppress compiler warnings about unused functions. (Functions defined in `.h` usually should be marked `inline`.)

Function Prototypes

Put the return type and function name on the same line in a function prototype:

```
static const struct option_class *get_option_class(int code);
```

Omit parameter names from function prototypes when the names do not give useful information, e.g.:

```
int netdev_get_mtu(const struct netdev *, int *mtup);
```

Statements

Indent each level of code with 4 spaces. Use BSD-style brace placement:

```
if (a()) {
    b();
    d();
}
```

Put a space between `if`, `while`, `for`, etc. and the expressions that follow them.

Enclose single statements in braces:

```
if (a > b) {
    return a;
} else {
    return b;
}
```

Use comments and blank lines to divide long functions into logical groups of statements.

Avoid assignments inside `if` and `while` conditions.

Do not put gratuitous parentheses around the expression in a return statement, that is, write `return 0;` and not `return(0);`

Write only one statement per line.

Indent `switch` statements like this:

```
switch (conn->state) {
case S_RECV:
    error = run_connection_input(conn);
    break;

case S_PROCESS:
    error = 0;
    break;

case S_SEND:
    error = run_connection_output(conn);
    break;

default:
    OVS_NOT_REACHED();
}
```

`switch` statements with very short, uniform cases may use an abbreviated style:

```
switch (code) {
case 200: return "OK";
case 201: return "Created";
case 202: return "Accepted";
case 204: return "No Content";
default: return "Unknown";
}
```

Use `for (;;)` to write an infinite loop.

In an `if/else` construct where one branch is the "normal" or "common" case and the other branch is the "uncommon" or "error" case, put the common case after the `if`, not the `else`. This is a form of documentation. It also places the

most important code in sequential order without forcing the reader to visually skip past less important details. (Some compilers also assume that the `if` branch is the more common case, so this can be a real form of optimization as well.)

Return Values

For functions that return a success or failure indication, prefer one of the following return value conventions:

- An `int` where `0` indicates success and a positive errno value indicates a reason for failure.
- A `bool` where `true` indicates success and `false` indicates failure.

Macros

Don't define an object-like macro if an enum can be used instead.

Don't define a function-like macro if a `static inline` function can be used instead.

If a macro's definition contains multiple statements, enclose them with `do { ... } while (0)` to allow them to work properly in all syntactic circumstances.

Do use macros to eliminate the need to update different parts of a single file in parallel, e.g. a list of enums and an array that gives the name of each enum. For example:

```
/* Logging importance levels. */
#define VLOG_LEVELS                             \
    VLOG_LEVEL(EMER, LOG_ALERT)                 \
    VLOG_LEVEL(ERR, LOG_ERR)                    \
    VLOG_LEVEL(WARN, LOG_WARNING)               \
    VLOG_LEVEL(INFO, LOG_NOTICE)                \
    VLOG_LEVEL(DBG, LOG_DEBUG)
enum vlog_level {
#define VLOG_LEVEL(NAME, SYSLOG_LEVEL) VLL_##NAME,
    VLOG_LEVELS
#undef VLOG_LEVEL
    VLL_N_LEVELS
};

/* Name for each logging level. */
static const char *level_names[VLL_N_LEVELS] = {
#define VLOG_LEVEL(NAME, SYSLOG_LEVEL) #NAME,
    VLOG_LEVELS
#undef VLOG_LEVEL
};
```

Thread Safety Annotations

Use the macros in `lib/compiler.h` to annotate locking requirements. For example:

```
static struct ovs_mutex mutex = OVS_MUTEX_INITIALIZER;
static struct ovs_rwlock rwlock = OVS_RWLOCK_INITIALIZER;

void function_require_plain_mutex(void) OVS_REQUIRES(mutex);
void function_require_rwlock(void) OVS_REQ_RDLOCK(rwlock);
```

Pass lock objects, not their addresses, to the annotation macros. (Thus we have `OVS_REQUIRES(mutex)` above, not `OVS_REQUIRES(&mutex)`.)

Source Files

Each source file should state its license in a comment at the very top, followed by a comment explaining the purpose of the code that is in that file. The comment should explain how the code in the file relates to code in other files. The goal is to allow a programmer to quickly figure out where a given module fits into the larger system.

The first non-comment line in a `.c` source file should be:

```
#include <config.h>
```

`#include` directives should appear in the following order:

1. `#include <config.h>`
2. The module's own headers, if any. Including this before any other header (besides) ensures that the module's header file is self-contained (see *header files* below).
3. Standard C library headers and other system headers, preferably in alphabetical order. (Occasionally one encounters a set of system headers that must be included in a particular order, in which case that order must take precedence.)
4. Open vSwitch headers, in alphabetical order. Use "", not <>, to specify Open vSwitch header names.

Header Files

Each header file should start with its license, as described under *source files* above, followed by a "header guard" to make the header file idempotent, like so:

```
#ifndef NETDEV_H
#define NETDEV_H 1

...

#endif /* netdev.h */
```

Header files should be self-contained; that is, they should `#include` whatever additional headers are required, without requiring the client to `#include` them for it.

Don't define the members of a struct or union in a header file, unless client code is actually intended to access them directly or if the definition is otherwise actually needed (e.g. inline functions defined in the header need them).

Similarly, don't `#include` a header file just for the declaration of a struct or union tag (e.g. just for `struct ;`). Just declare the tag yourself. This reduces the number of header file dependencies.

Types

Use typedefs sparingly. Code is clearer if the actual type is visible at the point of declaration. Do not, in general, declare a typedef for a `struct`, `union`, or `enum`. Do not declare a typedef for a pointer type, because this can be very confusing to the reader.

A function type is a good use for a typedef because it can clarify code. The type should be a function type, not a pointer-to-function type. That way, the typedef name can be used to declare function prototypes. (It cannot be used for function definitions, because that is explicitly prohibited by C89 and C99.)

You may assume that `char` is exactly 8 bits and that `int` and `long` are at least 32 bits.

Don't assume that `long` is big enough to hold a pointer. If you need to cast a pointer to an integer, use `intptr_t` or `uintptr_t` from .

Use the `int_t` and `uint_t` types from for exact-width integer types. Use the `PRId`, `PRIu`, and `PRIx` macros from for formatting them with `printf()` and related functions.

For compatibility with antique `printf()` implementations:

- Instead of `"%zu"`, use `"%"PRIuSIZE`.
- Instead of `"%td"`, use `"%"PRIdPTR`.
- Instead of `"%ju"`, use `"%"PRIuMAX`.

Other variants exist for different radixes. For example, use `"%"PRIxSIZE` instead of `"%zx"` or `"%x"` instead of `"%hhx"`.

Also, instead of `"%hhd"`, use `"%d"`. Be cautious substituting `"%u"`, `"%x"`, and `"%o"` for the corresponding versions with `"hh"`: cast the argument to unsigned char if necessary, because `printf("%hhu", -1)` prints `255` but `printf("%u", -1)` prints `4294967295`.

Use bit-fields sparingly. Do not use bit-fields for layout of network protocol fields or in other circumstances where the exact format is important.

Declare bit-fields to be signed or unsigned integer types or `_Bool` (aka `bool`). Do *not* declare bit-fields of type `int`: C99 allows these to be either signed or unsigned according to the compiler's whim. (A 1-bit bit-field of type `int` may have a range of `-1...0`!)

Try to order structure members such that they pack well on a system with 2-byte `short`, 4-byte `int`, and 4- or 8-byte `long` and pointer types. Prefer clear organization over size optimization unless you are convinced there is a size or speed benefit.

Pointer declarators bind to the variable name, not the type name. Write `int *x`, not `int* x` and definitely not `int * x`.

Expressions

Put one space on each side of infix binary and ternary operators:

```
* / %
+ -
<< >>
< <= > >=
== !=
&
^
|
&&
||
?:
= += -= *= /= %= &= ^= |= <<= >>=
```

Avoid comma operators.

Do not put any white space around postfix, prefix, or grouping operators:

```
() [] -> .
! ~ ++ -- + - * &
```

Exception 1: Put a space after (but not before) the "sizeof" keyword.

Exception 2: Put a space between the `()` used in a cast and the expression whose type is cast: `(void *) 0`.

Break long lines before the ternary operators `?` and `:`, rather than after them, e.g.

```
return (out_port != VIGP_CONTROL_PATH
        ? alpheus_output_port(dp, skb, out_port)
```

(continues on next page)

(continued from previous page)

```
        : alpheus_output_control(dp, skb, fwd_save_skb(skb),
                                 VIGR_ACTION));
```

Parenthesize the operands of `&&` and `||` if operator precedence makes it necessary, or if the operands are themselves expressions that use `&&` and `||`, but not otherwise. Thus:

```
if (rule && (!best || rule->priority > best->priority)) {
    best = rule;
}
```

but:

```
if (!isdigit((unsigned char)s[0]) ||
    !isdigit((unsigned char)s[1]) ||
    !isdigit((unsigned char)s[2])) {
    printf("string %s does not start with 3-digit code\n", s);
}
```

Do parenthesize a subexpression that must be split across more than one line, e.g.:

```
*idxp = ((l1_idx << PORT_ARRAY_L1_SHIFT) |
         (l2_idx << PORT_ARRAY_L2_SHIFT) |
         (l3_idx << PORT_ARRAY_L3_SHIFT));
```

Breaking a long line after a binary operator gives its operands a more consistent look, since each operand has the same horizontal position. This makes the end-of-line position a good choice when the operands naturally resemble each other, as in the previous two examples. On the other hand, breaking before a binary operator better draws the eye to the operator, which can help clarify code by making it more obvious what's happening, such as in the following example:

```
if (!ctx.freezing
    && xbridge->has_in_band
    && in_band_must_output_to_local_port(flow)
    && !actions_output_to_local_port(&ctx)) {
```

Thus, decide whether to break before or after a binary operator separately in each situation, based on which of these factors appear to be more important.

Try to avoid casts. Don't cast the return value of malloc().

The `sizeof` operator is unique among C operators in that it accepts two very different kinds of operands: an expression or a type. In general, prefer to specify an expression, e.g. `int *x = xmalloc(sizeof *x);`. When the operand of `sizeof` is an expression, there is no need to parenthesize that operand, and please don't.

Use the `ARRAY_SIZE` macro from `lib/util.h` to calculate the number of elements in an array.

When using a relational operator like < or ==, put an expression or variable argument on the left and a constant argument on the right, e.g. `x == 0`, *not* `0 == x`.

Blank Lines

Put one blank line between top-level definitions of functions and global variables.

C DIALECT

Most C99 features are OK because they are widely implemented:

- Flexible array members (e.g. `struct { int foo[]; }`).
- `static inline` functions (but no other forms of `inline`, for which GCC and C99 have differing interpretations).
- `long long`
- `bool` and <stdbool.h>, but don't assume that `bool` or `_Bool` can only take on the values `0` or `1`, because this behavior can't be simulated on C89 compilers.

  Also, don't assume that a conversion to `bool` or `_Bool` follows C99 semantics, i.e. use `(bool) (some_value != 0)` rather than `(bool) some_value`. The latter might produce unexpected results on non-C99 environments. For example, if `bool` is implemented as a typedef of char and `some_value = 0x10000000`.
- Designated initializers (e.g. `struct foo foo = { .a = 1 };` and `int a[] = { [2] = 5 };`).
- Mixing of declarations and code within a block. Favor positioning that allows variables to be initialized at their point of declaration.
- Use of declarations in iteration statements (e.g. `for (int i = 0; i < 10; i++)`).
- Use of a trailing comma in an enum declaration (e.g. `enum { x = 1, };`).

As a matter of style, avoid `//` comments.

Avoid using GCC or Clang extensions unless you also add a fallback for other compilers. You can, however, use C99 features or GCC extensions also supported by Clang in code that compiles only on GNU/Linux (such as `lib/netdev-linux.c`), because GCC is the system compiler there.

Python

When introducing new Python code, try to follow Python's PEP 8 style. Consider running the `pep8` or `flake8` tool against your code to find issues.

Libraries

When introducing a new library, follow *Open vSwitch Library ABI guide*

7.1.4 Documentation Style

This file describes the documentation style used in all documentation found in Open vSwitch. Documentation includes any documents found in `Documentation` along with any `README`, `MAINTAINERS`, or generally `rst` suffixed documents found in the project tree.

> **Note**
>
> This guide only applies to documentation for Open vSwitch v2.7. or greater. Previous versions of Open vSwitch used a combination of Markdown and raw plain text, and guidelines for these are not detailed here.

reStructuredText vs. Sphinx

reStructuredText (rST) is the syntax, while Sphinx is a documentation generator. Sphinx introduces a number of extensions to rST, like the `:ref:` role, which can and should be used in documentation, but these will not work correctly on GitHub. As such, these extensions should not be used in any documentation in the root level, such as the `README`.

rST Conventions

Basics

Many of the basic documentation guidelines match those of the *Coding Style*.

- Use reStructuredText (rST) for all documentation.

  Sphinx extensions can be used, but only for documentation in the `Documentation` folder.

- Limit lines at 79 characters.

  > **Note**
  >
  > An exception to this rule is text within code-block elements that cannot be wrapped and links within references.

- Use spaces for indentation.
- Match indentation levels.

  A change in indentation level usually signifies a change in content nesting, by either closing the existing level or introducing a new level.

- Avoid trailing spaces on lines.
- Include a license (see this file) in all docs.
- Most importantly, always build and display documentation before submitting changes! Docs aren't unit testable, so visible inspection is necessary.

File Names

- Use hyphens as space delimiters. For example: `my-readme-document.rst`

  > **Note**
  >
  > An exception to this rule is any man pages, which take an trailing number corresponding to the number of arguments required. This number is preceded by an underscore.

- Use lowercase filenames.

  > **Note**
  >
  > An exception to this rule is any documents found in the root-level of the project.

Titles

- Use the following headers levels.

  ======= Heading 0 (reserved for the title in a document)
  ------- Heading 1
  ~~~~~~~ Heading 2
  +++++++ Heading 3
  ~~~~~~~

''''''' Heading 4

> **Note**
>
> Avoid using lower heading levels by rewriting and reorganizing the information.

- Under- and overlines should be of the same length as that of the heading text.
- Use "title case" for headers.

Code

- Use `::` to prefix code.
- Don't use syntax highlighting such as `.. highlight:: <syntax>` or `code-block:: <syntax>` because it depends on external `pygments` library.
- Prefix commands with `$`.
- Where possible, include fully-working snippets of code. If there pre-requisites, explain what they are and how to achieve them.

Admonitions

- Use admonitions to call attention to important information.:

```
.. note::

  This is a sample callout for some useful tip or trick.
```

  Example admonitions include: `warning`, `important`, `note`, `tip` or `seealso`.

- Use notes sparingly. Avoid having more than one per subsection.

Tables

- Use either graphic tables, list tables or CSV tables.

Graphic tables

```
.. table:: OVS-Linux kernel compatibility

  ============ ==============
  Open vSwitch Linux kernel
  ============ ==============
  1.4.x        2.6.18 to 3.2
  1.5.x        2.6.18 to 3.2
  1.6.x        2.6.18 to 3.2
  ============ ==============
```

```
.. table:: OVS-Linux kernel compatibility

  +--------------+---------------+
```

(continues on next page)

(continued from previous page)

```
| Open vSwitch | Linux kernel  |
+==============+===============+
| 1.4.x        | 2.6.18 to 3.2 |
+--------------+---------------+
| 1.5.x        | 2.6.18 to 3.2 |
+--------------+---------------+
| 1.6.x        | 2.6.18 to 3.2 |
+--------------+---------------+
```

Note

The `table` role - `.. table:: <name>` - can be safely omitted.

List tables

```
.. list-table:: OVS-Linux kernel compatibility
   :widths: 10 15
   :header-rows: 1

   * - Open vSwitch
     - Linux kernel
   * - 1.4.x
     - 2.6.18 to 3.2
   * - 1.5.x
     - 2.6.18 to 3.2
   * - 1.6.x
     - 2.6.18 to 3.2
```

CSV tables

```
.. csv-table:: OVS-Linux kernel compatibility
   :header: Open vSwitch, Linux kernel
   :widths: 10 15

   1.4.x, 2.6.18 to 3.2
   1.5.x, 2.6.18 to 3.2
   1.6.x, 2.6.18 to 3.2
```

Cross-referencing

- To link to an external file or document, include as a link.:

  ```
  Here's a `link <http://openvswitch.org>`__ to the Open vSwitch website.


  Here's a `link`_ in reference style.

  .. _link: http://openvswitch.org
  ```

- You can also use citations.:

```
Refer to the Open vSwitch documentation [1]_.

References
----------

.. [1]: http://openvswitch.org
```

- To cross-reference another doc, use the `doc` role.:

```
Here is a link to the :doc:`/README.rst`
```

> **Note**
>
> This is a Sphinx extension. Do not use this in any top-level documents.

- To cross-reference an arbitrary location in a doc, use the `ref` role.:

```
.. _sample-crossref

Title
~~~~~

Hello, world.

Another Title
~~~~~~~~~~~~~

Here is a cross-reference to :ref:`sample-crossref`.
```

> **Note**
>
> This is a Sphinx extension. Do not use this in any top-level documents.

Figures and Other Media

- All images should be in PNG format and compressed where possible. For PNG files, use OptiPNG and AdvanceCOMP's `advpng`:

```
$ optipng -o7 -zm1-9 -i0 -strip all <path_to_png>
$ advpng -z4 <path_to_png>
```

- Any ASCII text "images" should be included in code-blocks to preserve formatting
- Include other reStructuredText verbatim in a current document

Comments

- Comments are indicated by means of the `..` marker.:

```
.. TODO(stephenfin) This section needs some work. This TODO will not
   appear in the final generated document, however.
```

Man Pages

In addition to the above, man pages have some specific requirements:

- You **must** define the following sections:
  - Synopsis
  - Description
  - Options

  Note that *NAME* is not included - this is automatically generated by Sphinx and should not be manually defined. Also note that these do not need to be uppercase - Sphinx will do this automatically.

  Additional sections are allowed. Refer to *man-pages(8)* for information on the sections generally allowed.

- You **must not** define a *NAME* section.

  See above.

- The *OPTIONS* section must describe arguments and options using the program and option directives.

  This ensures the output is formatted correctly and that you can cross-reference various programs and commands from the documentation. For example:

```
.. program:: ovs-do-something

.. option:: -f, --force

   Force the operation

.. option:: -b <bridge>, --bridge <bridge>

   Name or ID of bridge
```

> **Important**
>
> Option argument names should be enclosed in angle brackets, as above.

- Any references to the application or any other Open vSwitch application must be marked up using the *program* role.

  This allows for easy linking in the HTML output and correct formatting in the man page output. For example:

```
To do something, run :program:`ovs-do-something`.
```

- The man page must be included in the list of man page documents found in conf.py

Refer to existing man pages, such as *ovs-vlan-test* for a worked example.

Writing Style

Follow these guidelines to ensure readability and consistency of the Open vSwitch documentation. These guidelines are based on the /*IBM Style Guide/*.

- Use standard US English

  Use a spelling and grammar checking tool as necessary.

- Expand initialisms and acronyms on first usage.

  Commonly used terms like CPU or RAM are allowed.

Do not use	Do use
OVS is a virtual switch. OVS has…	Open vSwitch (OVS) is a virtual switch. OVS has…
The VTEP emulator is…	The Virtual Tunnel Endpoint (VTEP) emulator is…

- Write in the active voice

  The subject should do the verb's action, rather than be acted upon.

Do not use	Do use
A bridge is created by you	Create a bridge

- Write in the present tense

Do not use	Do use
Once the bridge is created, you can create a port	Once the bridge is created, create a port

- Write in second person

Do not use	Do use
To create a bridge, the user runs:	To create a bridge, run:

- Keep sentences short and concise
- Eliminate needless politeness

  Avoid "please" and "thank you"

Helpful Tools

There are a number of tools, online and offline, which can be used to preview documents are you edit them:

- ReText

  A simple but powerful editor for Markdown and reStructuredText. ReText is written in Python.

- restview

  A viewer for ReStructuredText documents that renders them on the fly.

Useful Links

- Quick reStructuredText
- Sphinx Documentation

7.1.5 Inclusive Language

In order to help facilitate an inclusive environment in the Open vSwitch community we recognise the role of language in framing our communication with each other. It is important that terms that may exclude people through racial, cultural or other bias, are avoided as they may make people feel excluded.

We recognise that this is subjective, and to some extent is a journey. But we also recognise that we cannot begin that journey without taking positive action. To this end Open vSwitch is adopting the practice of an inclusive word list, which helps to guide the use of language within the project.

Word List

The intent of this document is to formally document the acceptance of a inclusive word list by Open vSwitch. Accordingly, this document specifies use of the use the Inclusive Naming Word List v1.0 (the word list) for Open vSwitch.

The adoption of the word list intended that this act as a guide for developers creating patches to the Open vSwitch repository, including both source code and documentation. And to aid maintainers in their role of shepherding changes into the repository.

Further steps to align usage of language in Open vSwitch, including clarification of application of the word list, to new and existing work, may follow.

7.1.6 Open vSwitch Library ABI Updates

This file describes the manner in which the Open vSwitch shared library manages different ABI and API revisions. This document aims to describe the background, goals, and concrete mechanisms used to export code-space functionality so that it may be shared between multiple applications.

Definitions

Table 1: Definitions for terms appearing in this document

Term	Definition
ABI	Abbreviation of Application Binary Interface
API	Abbreviation of Application Programming Interface
Application Binary Interface	The low-level runtime interface exposed by an object file.
Application Programming Interface	The source-code interface descriptions intended for use in multiple translation units when compiling.
Code library	A collection of function implementations and definitions intended to be exported and called through a well-defined interface.
Shared Library	A code library which is imported at run time.

Overview

C and C++ applications often use 'external' functionality, such as printing specialized data types or parsing messages, which has been exported for common use. There are many possible ways for applications to call such external functionality, for instance by including an appropriate inline definition which the compiler can emit as code in each function it appears. One such way of exporting and importing such functionality is through the use of a library of code.

When a compiler builds object code from source files to produce object code, the results are binary data arranged with specific calling conventions, alignments, and order suitable for a run-time environment or linker. This result defines a specific ABI.

As library of code develops and its exported interfaces change over time, the resulting ABI may change as well. Therefore, care must be taken to ensure the changes made to libraries of code are effectively communicated to applications which use them. This includes informing the applications when incompatible changes are made.

The Open vSwitch project exports much of its functionality through multiple such libraries of code. These libraries are intended for multiple applications to import and use. As the Open vSwitch project continues to evolve and change, its exported code will evolve as well. To ensure that applications linking to these libraries are aware of these changes, Open vSwitch employs libtool version stamps.

ABI Policy

Open vSwitch will export the ABI version at the time of release, such that the library name will be the major.minor version, and the rest of the release version information will be conveyed with a libtool interface version.

The intent is for Open vSwitch to maintain an ABI stability for each minor revision only (so that Open vSwitch release 2.5 carries a guarantee for all 2.5.ZZ micro-releases). This means that any porting effort to stable branches must take not to disrupt the existing ABI.

In the event that a bug must be fixed in a backwards-incompatible way, developers must bump the libtool 'current' version to inform the linker of the ABI breakage. This will signal that libraries exposed by the subsequent release will not maintain ABI stability with the previous version.

Coding

At build time, if building shared libraries by passing the *–enable-shared* arguments to *./configure*, version information is extracted from the `$PACKAGE_VERSION` automake variable and formatted into the appropriate arguments. These get exported for use in Makefiles as `$OVS_LTINFO`, and passed to each exported library along with other `LDFLAGS`.

Therefore, when adding a new library to the build system, these version flags should be included with the `$LDFLAGS` variable. Nothing else needs to be done.

Changing an exported function definition (from a file in, for instance *lib/*.h*) is only permitted from minor release to minor release. Likewise changes to library data structures should only occur from minor release to minor release.

7.2 Mailing Lists

> **Important**
>
> Report security issues **only** to security@openvswitch.org. For more information, refer to our *security policies*.

7.2.1 ovs-announce

The ovs-announce mailing list is used to announce new versions of Open vSwitch and is extremely low-volume. (subscribe) (archives)

7.2.2 ovs-discuss

The ovs-discuss mailing list is used to discuss plans and design decisions for Open vSwitch. It is also an appropriate place for user questions. (subscribe) (archives)

7.2.3 ovs-dev

The ovs-dev mailing list is used to discuss development and review code before being committed. (subscribe) (archives)

7.2.4 ovs-git

The ovs-git mailing list hooks into Open vSwitch's version control system to receive commits. (subscribe) (archives)

7.2.5 ovs-build

The ovs-build mailing list hooks into Open vSwitch's continuous integration system to receive build reports. (subscribe) (archives)

7.2.6 bugs

The bugs mailing list is an alias for the discuss mailing list.

7.2.7 security

The security mailing list is for submitting security vulnerabilities to the security team.

7.3 Patchwork

Open vSwitch uses Patchwork to track the status of patches sent to the *ovs-dev mailing list*. The Open vSwitch Patchwork instance can be found on ozlabs.org.

Patchwork provides a number of useful features for developers working on Open vSwitch:

- Tracking the lifecycle of patches (accepted, rejected, under-review, ...)
- Assigning reviewers (delegates) to patches
- Downloading/applying patches, series, and bundles via the web UI or the REST API (see *git-pw*)
- A usable UI for viewing patch discussions

7.3.1 git-pw

The *git-pw* tool provides a way to download and apply patches, series, and bundles. You can install *git-pw* from PyPi like so:

```
$ pip install --user git-pw
```

To actually use *git-pw*, you must configure it with the Patchwork instance URL, Patchwork project, and your Patchwork user authentication token. The URL and project are provided below, but you must obtain your authentication token from your Patchwork User Profile page. If you do not already have a Patchwork user account, you should create one now.

Once your token is obtained, configure *git-pw* as below. Note that this must be run from within the Open vSwitch Git repository:

```
$ git config pw.server https://patchwork.ozlabs.org/
$ git config pw.project openvswitch
$ git config pw.token $PW_TOKEN  # using the token obtained earlier
```

Once configured, run the following to get information about available commands:

```
$ git pw --help
```

7.3.2 pwclient

The *pwclient* is a legacy tool that provides some of the functionality of *git-pw* but uses the legacy XML-RPC API. It is considered deprecated in its current form and *git-pw* should be used instead.

7.4 Release Process

This document describes the process ordinarily used for Open vSwitch development and release. Exceptions are sometimes necessary, so all of the statements here should be taken as subject to change through rough consensus of Open vSwitch contributors, obtained through public discussion on, e.g., ovs-dev or the #openvswitch IRC channel.

7.4.1 Release Strategy

Open vSwitch feature development takes place on the "main" branch. Ordinarily, new features are rebased against main and applied directly. For features that take significant development, sometimes it is more appropriate to merge a separate branch into main; please discuss this on ovs-dev in advance.

The process of making a release has the following stages. See *Release Scheduling* for the timing of each stage:

1. "Soft freeze" of the main branch.

   During the freeze, we ask committers to refrain from applying patches that add new features unless those patches were already being publicly discussed and reviewed before the freeze began. Bug fixes are welcome at any time. Please propose and discuss exceptions on ovs-dev.

2. Fork a release branch from main, named for the expected release number, e.g. "branch-2.3" for the branch that will yield Open vSwitch 2.3.x.

   Release branches are intended for testing and stabilization. At this stage and in later stages, they should receive only bug fixes, not new features. Bug fixes applied to release branches should be backports of corresponding bug fixes to the main branch, except for bugs present only on release branches (which are rare in practice).

   At this stage, sometimes there can be exceptions to the rule that a release branch receives only bug fixes. Like bug fixes, new features on release branches should be backports of the corresponding commits on the main branch. Features to be added to release branches should be limited in scope and risk and discussed on ovs-dev before creating the branch.

3. When committers come to rough consensus that the release is ready, they release the .0 release on its branch, e.g. 2.3.0 for branch-2.3. To make the actual release, a committer pushes a signed tag named, e.g. v2.3.0, to the Open vSwitch repository, makes a release tarball available on openvswitch.org, and posts a release announcement to ovs-announce.

4. As bug fixes accumulate, or after important bugs or vulnerabilities are fixed, committers may make additional releases from a branch: 2.3.1, 2.3.2, and so on. The process is the same for these additional release as for a .0 release.

At most three release branches are formally maintained at any given time: the latest release, the latest release designed as LTS and a previous LTS release during the transition period. An LTS release is one that the OVS project has designated as being maintained for a longer period of time. Currently, an LTS release is maintained until the next major release after the new LTS is chosen. This one release time frame is a transition period which is intended for users to upgrade from old LTS to new one.

New LTS release is chosen every 2 years. The process is that current latest stable release becomes an LTS release at the same time the next major release is out. That could change based on the current state of OVS development. For example, we do not want to designate a new release as LTS that includes disruptive internal changes, as that may make it harder to support for a longer period of time. Discussion about skipping designation of the next LTS release occurs on the OVS development mailing list.

LTS designation schedule example (depends on current state of development):

Version	Release Date	Actions
2.17	Feb 2022	2.17 - new latest stable
3.0	Aug 2022	3.0 - new latest stable, 2.17 stable new LTS
3.1	Feb 2023	3.1 - new latest stable, 2.13 LTS EOL
3.2	Aug 2023	3.2 - new latest stable
3.3	Feb 2024	3.3 - new latest stable
3.4	Aug 2024	3.4 - new latest stable, 3.3 stable new LTS
3.5	Feb 2025	3.5 - new latest stable, 2.17 LTS EOL
3.6	Aug 2025	3.6 - new latest stable

While branches other than LTS and the latest release are not formally maintained, the OVS project usually provides stable releases for these branches for at least 2 years, i.e. stable releases are provided for the last 4 release branches. However, these branches may not include all the fixes that LTS has in case backporting is not straightforward and developers are not willing to spend their time on that (this mostly affects branches that are older than the LTS, because backporting to LTS implies backporting to all intermediate branches).

7.4.2 Release Numbering

The version number on main should normally end in .90. This indicates that the Open vSwitch version is "almost" the next version to branch.

Forking main into branch-x.y requires two commits to main. The first is titled "Prepare for x.y.0" and increments the version number to x.y. This is the initial commit on branch-x.y. The second is titled "Prepare for post-x.y.0 (x.y.90)" and increments the version number to x.y.90.

The version number on a release branch is x.y.z, where z is initially 0. Making a release requires two commits. The first is titled *Set release dates for x.y.z.* and updates NEWS and debian/changelog to specify the release date of the new release. This commit is the one made into a tarball and tagged. The second is titled *Prepare for x.y.(z+1).* and increments the version number and adds a blank item to NEWS with an unspecified date.

7.4.3 Release Scheduling

Open vSwitch makes releases at the following six-month cadence. All dates are approximate:

Time (months)	Dates	Stage
T	Mar 1, Sep 1	Begin x.y release cycle
T + 4	Jul 1, Jan 1	"Soft freeze" main for x.y release
T + 4.5	Jul 15, Jan 15	Fork branch-x.y from main
T + 5.5	Aug 15, Feb 15	Release version x.y.0

7.4.4 How to Branch

To branch "main" for the eventual release of OVS version x.y.0, prepare two patches against main:

1. "Prepare for x.y.0." following the model of commit 836d1973c56e ("Prepare for 2.11.0.").
2. "Prepare for post-x.y.0 (x.y.90)." following the model of commit fe2870c574db ("Prepare for post-2.11.0 (2.11.90).")

Post both patches to ovs-dev. Get them reviewed in the usual way.

Apply both patches to main, and create branch-x.y by pushing only the first patch. The following command illustrates how to do both of these at once assuming the local repository HEAD points to the "Prepare for post-x.y.0" commit:

```
git push origin HEAD:main HEAD^:refs/heads/branch-x.y
```

Branching should be announced on ovs-dev.

7.4.5 How to Release

Follow these steps to release version x.y.z of OVS from branch-x.y.

1. Prepare two patches against branch-x.y:

   a. "Set release date for x.y.z". For z = 0, follow the model of commit d11f4cbbfe05 ("Set release date for 2.12.0."); for z > 0, follow the model of commit 53d5c18118b0 ("Set release date for 2.11.3.").

   b. "Prepare for x.y.(z+1)." following the model of commit db02dd23e48a ("Prepare for 2.11.1.").

3. Post the patches to ovs-dev. Get them reviewed in the usual way.
4. Apply the patches to branch-x.y.
5. If z = 0, apply the first patch (only) to main.
6. Sign a tag vx.y.z "Open vSwitch version x.y.z" and push it to the repo.
7. Update http://www.openvswitch.org/download/. See commit 31eaa72cafac ("Add 2.12.0 and older release announcements.") in the website repo (https://github.com/openvswitch/openvswitch.github.io) for an example.
8. Consider updating the Wikipedia page for Open vSwitch at https://en.wikipedia.org/wiki/Open_vSwitch
9. Tweet.

7.4.6 Contact

Use dev@openvswitch.org to discuss the Open vSwitch development and release process.

7.5 Reporting Bugs

We are eager to hear from users about problems that they have encountered with Open vSwitch. This file documents how best to report bugs so as to ensure that they can be fixed as quickly as possible.

Please report bugs by sending email to bugs@openvswitch.org.

For reporting security vulnerabilities, please read *Security Process*.

The most important parts of your bug report are the following:

- What you did that make the problem appear.
- What you expected to happen.
- What actually happened.

Please also include the following information:

- The Open vSwitch version number (as output by `ovs-vswitchd --version`).
- The Git commit number (as output by `git rev-parse HEAD`), if you built from a Git snapshot.
- Any local patches or changes you have applied (if any).

The following are also handy sometimes:

- The kernel version on which Open vSwitch is running (from `/proc/version`) and the distribution and version number of your OS (e.g. "Centos 5.0").
- The contents of the vswitchd configuration database (usually `/etc/openvswitch/conf.db`).
- The output of `ovs-dpctl show`.

- If you have Open vSwitch configured to connect to an OpenFlow controller, the output of `ovs-ofctl show <bridge>` for each <bridge> configured in the vswitchd configuration database.
- A fix or workaround, if you have one.
- Any other information that you think might be relevant.

Important

bugs@openvswitch.org is a public mailing list, to which anyone can subscribe, so do not include confidential information in your bug report.

7.6 Security Process

This is a proposed security vulnerability reporting and handling process for Open vSwitch. It is based on the OpenStack vulnerability management process described at https://wiki.openstack.org/wiki/Vulnerability_Management.

The OVS security team coordinates vulnerability management using the ovs-security mailing list. Membership in the security team and subscription to its mailing list consists of a small number of trustworthy people, as determined by rough consensus of the Open vSwitch committers on the ovs-committers mailing list. The Open vSwitch security team should include Open vSwitch committers, to ensure prompt and accurate vulnerability assessments and patch review.

We encourage everyone involved in the security process to GPG-sign their emails. We additionally encourage GPG-encrypting one-on-one conversations as part of the security process.

7.6.1 What is a vulnerability?

All vulnerabilities are bugs, but not every bug is a vulnerability. Vulnerabilities compromise one or more of:

- Confidentiality (personal or corporate confidential data).
- Integrity (trustworthiness and correctness).
- Availability (uptime and service).

Here are some examples of vulnerabilities to which one would expect to apply this process:

- A crafted packet that causes a kernel or userspace crash (Availability).
- A flow translation bug that misforwards traffic in a way likely to hop over security boundaries (Integrity).
- An OpenFlow protocol bug that allows a controller to read arbitrary files from the file system (Confidentiality).
- Misuse of the OpenSSL library that allows bypassing certificate checks (Integrity).
- A bug (memory corruption, overflow, ...) that allows one to modify the behaviour of OVS through external configuration interfaces such as OVSDB (Integrity).
- Privileged information is exposed to unprivileged users (Confidentiality).

If in doubt, please do use the vulnerability management process. At worst, the response will be to report the bug through the usual channels.

7.6.2 Step 1: Reception

To report an Open vSwitch vulnerability, send an email to the ovs-security mailing list (see *contact* at the end of this document). A security team member should reply to the reporter acknowledging that the report has been received.

Consider reporting the information mentioned in *Reporting Bugs*, where relevant.

Reporters may ask for a GPG key while initiating contact with the security team to deliver more sensitive reports.

The Linux kernel has its own vulnerability management process. Handling of vulnerabilities that affect both the Open vSwitch tree and the upstream Linux kernel should be reported through both processes. Send your report as a single email to both the kernel and OVS security teams to allow those teams to most easily coordinate among themselves.

7.6.3 Step 2: Assessment

The security team should discuss the vulnerability. The reporter should be included in the discussion (via "CC") to an appropriate degree.

The assessment should determine which Open vSwitch versions are affected (e.g. every version, only the latest release, only unreleased versions), the privilege required to take advantage of the vulnerability (e.g. any network user, any local L2 network user, any local system user, connected OpenFlow controllers), the severity of the vulnerability, and how the vulnerability may be mitigated (e.g. by disabling a feature).

The treatment of the vulnerability could end here if the team determines that it is not a realistic vulnerability.

7.6.4 Step 3a: Document

The security team develops a security advisory document. The security team may, at its discretion, include the reporter (via "CC") in developing the security advisory document, but in any case should accept feedback from the reporter before finalizing the document. When the document is final, the security team should obtain a CVE for the vulnerability from a CNA (https://cve.mitre.org/cve/cna.html).

The document credits the reporter and describes the vulnerability, including all of the relevant information from the assessment in step 2. Suitable sections for the document include:

```
* Title: The CVE identifier, a short description of the
  vulnerability.  The title should mention Open vSwitch.

  In email, the title becomes the subject.  Pre-release advisories
  are often passed around in encrypted email, which have plaintext
  subjects, so the title should not be too specific.

* Description: A few paragraphs describing the general
  characteristics of the vulnerability, including the versions of
  Open vSwitch that are vulnerable, the kind of attack that
  exposes the vulnerability, and potential consequences of the
  attack.

  The description should re-state the CVE identifier, in case the
  subject is lost when an advisory is sent over email.

* Mitigation: How an Open vSwitch administrator can minimize the
  potential for exploitation of the vulnerability, before applying
  a fix.  If no mitigation is possible or recommended, explain
  why, to reduce the chance that at-risk users believe they are
  not at risk.

* Fix: Describe how to fix the vulnerability, perhaps in terms of
  applying a source patch.  The patch or patches themselves, if
  included in the email, should be at the very end of the advisory
  to reduce the risk that a reader would stop reading at this
  point.

* Recommendation: A concise description of the security team's
```

(continues on next page)

(continued from previous page)

```
  recommendation to users.

* Acknowledgments: Thank the reporters.

* Vulnerability Check: A step-by-step procedure by which a user
  can determine whether an installed copy of Open vSwitch is
  vulnerable.

  The procedure should clearly describe how to interpret the
  results, including expected results in vulnerable and
  not-vulnerable cases.  Thus, procedures that produce clear and
  easily distinguished results are preferred.

  The procedure should assume as little understanding of Open
  vSwitch as possible, to make it more likely that a competent
  administrator who does not specialize in Open vSwitch can
  perform it successfully.

  The procedure should have minimal dependencies on tools that are
  not widely installed.

  Given a choice, the procedure should be one that takes at least
  some work to turn into a useful exploit.  For example, a
  procedure based on "ovs-appctl" commands, which require local
  administrator access, is preferred to one that sends test
  packets to a machine, which only requires network connectivity.

  The section should say which operating systems it is designed
  for.  If the procedure is likely to be specific to particular
  architectures (e.g. x86-64, i386), it should state on which ones
  it has been tested.

  This section should state the risks of the procedure.  For
  example, if it can crash Open vSwitch or disrupt packet
  forwarding, say so.

  It is more useful to explain how to check an installed and
  running Open vSwitch than one built locally from source, but if
  it is easy to use the procedure from a sandbox environment, it
  can be helpful to explain how to do so.

* Patch: If a patch or patches are available, and it is practical
  to include them in the email, put them at the end.  Format them
  as described in :doc:`contributing/submitting-patches`, that is, as
  output by "git format-patch".

  The patch subjects should include the version for which they are
  suited, e.g. "[PATCH branch-2.3]" for a patch against Open
  vSwitch 2.3.x.  If there are multiple patches for multiple
  versions of Open vSwitch, put them in separate sections with
  clear titles.
```

(continues on next page)

(continued from previous page)

```
Multiple patches for a single version of Open vSwitch, that must
be stacked on top of each other to fix a single vulnerability,
are undesirable because users are less likely to apply all of
them correctly and in the correct order.

Each patch should include a Vulnerability tag with the CVE
identifier, a Reported-by tag or tags to credit the reporters,
and a Signed-off-by tag to acknowledge the Developer's
Certificate of Origin.  It should also include other appropriate
tags, such as Acked-by tags obtained during review.
```

CVE-2016-2074 is an example advisory document.

7.6.5 Step 3b: Fix

Steps 3a and 3b may proceed in parallel.

The security team develops and obtains (private) reviews for patches that fix the vulnerability. If necessary, the security team pulls in additional developers, who must agree to maintain confidentiality.

7.6.6 Step 4: Embargoed Disclosure

The security advisory and patches are sent to downstream stakeholders, with an embargo date and time set from the time sent. Downstream stakeholders are expected not to deploy or disclose patches until the embargo is passed.

A disclosure date is negotiated by the security team working with the bug submitter as well as vendors. However, the Open vSwitch security team holds the final say when setting a disclosure date. The timeframe for disclosure is from immediate (esp. if it's already publicly known) to a few weeks. As a basic default policy, we expect report date to disclosure date to be 10 to 15 business days.

Operating system vendors are obvious downstream stakeholders, however, any major Open vSwitch user who is interested and can be considered trustworthy enough could be included. To request being added to the Downstream mailing list, email the ovs-security mailing list. Please include a few sentences on how your organization uses Open vSwitch. If possible, please provide a security-related email alias rather than a direct end-user address.

If the vulnerability is already public, skip this step.

7.6.7 Step 5: Public Disclosure

When the embargo expires, push the (reviewed) patches to appropriate branches, post the patches to the ovs-dev mailing list (noting that they have already been reviewed and applied), post the security advisory to appropriate mailing lists (ovs-announce, ovs-discuss), and post the security advisory on the Open vSwitch webpage.

When the patch is applied to LTS (long-term support) branches, a new version should be released.

The security advisory should be GPG-signed by a security team member with a key that is in a public web of trust.

Contact

Report security vulnerabilities to the ovs-security mailing list: security@openvswitch.org

Report problems with this document to the ovs-bugs mailing list: bugs@openvswitch.org

7.7 The Linux Foundation Open vSwitch Project Charter

Effective August 9, 2016

1. Mission of Open vSwitch Project ("OVS").

   The mission of OVS is to:

   a. create an open source, production quality virtual networking platform, including a software switch, control plane, and related components, that supports standard management interfaces and opens the forwarding functions to programmatic extension and control; and

   b. host the infrastructure for an OVS community, establishing a neutral home for community assets, infrastructure, meetings, events and collaborative discussions.

2. Technical Steering Committee ("TSC")

   a. A TSC shall be composed of the Committers for OVS. The list of Committers on the TSC are available at *Committers*.

   b. TSC projects generally will involve Committers and Contributors:

      i. Contributors: anyone in the technical community that contributes code, documentation or other technical artifacts to the OVS codebase.

      ii. Committers: Contributors who have the ability to commit directly to a project's main branch or repository on an OVS project.

   c. Participation in as a Contributor and/or Committer is open to anyone under the terms of this Charter. The TSC may:

      i. establish work flows and procedures for the submission, approval and closure or archiving of projects,

      ii. establish criteria and processes for the promotion of Contributors to Committer status, available at *OVS Committer Grant/Revocation Policy*. and

      iii. amend, adjust and refine the roles of Contributors and Committers listed in Section 2.b., create new roles and publicly document responsibilities and expectations for such roles, as it sees fit, available at *Expectations for Developers with Open vSwitch Repo Access*.

   d. Responsibilities: The TSC is responsible for overseeing OVS activities and making decisions that impact the mission of OVS, including:

      i. coordinating the technical direction of OVS;

      ii. approving project proposals (including, but not limited to, incubation, deprecation and changes to a project's charter or scope);

      iii. creating sub-committees or working groups to focus on cross-project technical issues and requirements;

      iv. communicating with external and industry organizations concerning OVS technical matters;

      v. appointing representatives to work with other open source or standards communities;

      vi. establishing community norms, workflows or policies including processes for contributing (available at *Contributing to Open vSwitch*), issuing releases, and security issue reporting policies;

      vii. discussing, seeking consensus, and where necessary, voting on technical matters relating to the code base that affect multiple projects; and

      viii. coordinate any marketing, events or communications with The Linux Foundation.

3. TSC Voting

a. While it is the goal of OVS to operate as a consensus based community, if any TSC decision requires a vote to move forward, the Committers shall vote on a one vote per Committer basis.

b. TSC votes should be conducted by email. In the case of a TSC meeting where a valid vote is taken, the details of the vote and any discussion should be subsequently documented for the community (e.g. to the appropriate email mailing list).

c. Quorum for TSC meetings shall require two-thirds of the TSC representatives. The TSC may continue to meet if quorum is not met, but shall be prevented from making any decisions requiring a vote at the meeting.

d. Except as provided in Section 8.d. and 9.a., decisions by electronic vote (e.g. email) shall require a majority of all voting TSC representatives. Decisions by electronic vote shall be made timely, and unless specified otherwise, within three (3) business days. Except as provided in Section 8.d. and 9.a., decisions by vote at a meeting shall require a majority vote, provided quorum is met.

e. In the event of a tied vote with respect to an action that cannot be resolved by the TSC, any TSC representative shall be entitled to refer the matter to the Linux Foundation for assistance in reaching a decision.

4. Antitrust Guidelines

   a. All participants in OVS shall abide by The Linux Foundation Antitrust Policy available at http://www.linuxfoundation.org/antitrust-policy.

   b. All members shall encourage open participation from any organization able to meet the participation requirements, regardless of competitive interests. Put another way, the community shall not seek to exclude any participant based on any criteria, requirements or reasons other than those that are reasonable and applied on a non-discriminatory basis to all participants.

5. Code of Conduct

   a. The TSC may adopt a specific OVS Project code of conduct, with approval from the LF.

6. Budget and Funding

   a. The TSC shall coordinate any budget or funding needs with The Linux Foundation. Companies participating may be solicited to sponsor OVS activities and infrastructure needs on a voluntary basis.

   b. The Linux Foundation shall have custody of and final authority over the usage of any fees, funds and other cash receipts.

   c. A General & Administrative (G&A) fee will be applied by the Linux Foundation to funds raised to cover Finance, Accounting, and operations. The G&A fee shall equal 9% of OVS's first $1,000,000 of gross receipts and 6% of OVS's gross receipts over $1,000,000.

   d. Under no circumstances shall The Linux Foundation be expected or required to undertake any action on behalf of OVS that is inconsistent with the tax exempt purpose of The Linux Foundation.

7. General Rules and Operations.

   The OVS project shall be conducted so as to:

   a. engage in the work of the project in a professional manner consistent with maintaining a cohesive community, while also maintaining the goodwill and esteem of The Linux Foundation in the open source software community;

   b. respect the rights of all trademark owners, including any branding and usage guidelines;

   c. engage The Linux Foundation for all OVS press and analyst relations activities;

   d. upon request, provide information regarding Project participation, including information regarding attendance at Project-sponsored events, to The Linux Foundation; and

   e. coordinate with The Linux Foundation in relation to any websites created directly for OVS.

8. Intellectual Property Policy

a. Members agree that all new inbound code contributions to OVS shall be made under the Apache License, Version 2.0 (available at http://www.apache.org/licenses/LICENSE-2.0). All contributions shall be accompanied by a Developer Certificate of Origin sign-off (http://developercertificate.org) that is submitted through a TSC and LF-approved contribution process.

b. All outbound code will be made available under the Apache License, Version 2.0.

c. All documentation will be contributed to and made available by OVS under the Apache License, Version 2.0.

d. For any new project source code, if an alternative inbound or outbound license is required for compliance with the license for a leveraged open source project (e.g. GPLv2 for Linux kernel) or is otherwise required to achieve OVS's mission, the TSC may approve the use of an alternative license for specific inbound or outbound contributions on an exception basis. Any exceptions must be approved by a majority vote of the entire TSC and must be limited in scope to what is required for such purpose. Please email tsc@openvswitch.org to obtain exception approval.

e. Subject to available funds, OVS may engage The Linux Foundation to determine the availability of, and register, trademarks, service marks, which shall be owned by the LF.

9. Amendments

a. This charter may be amended by a two-thirds vote of the entire TSC, subject to approval by The Linux Foundation.

7.8 Emeritus Status for OVS Committers

OVS committers are nominated and elected based on their impact on the Open vSwitch project. Over time, as committers' responsibilities change, some may become unable or uninterested in actively participating in project governance. Committer "emeritus" status provides a way for committers to take a leave of absence from OVS governance responsibilities. The following guidelines clarify the process around the emeritus status for committers:

- A committer may choose to transition from active to emeritus, or from emeritus to active, by sending an email to the committers mailing list.
- If a committer hasn't been heard from in 6 months, and does not respond to reasonable attempts to contact him or her, the other committers can vote as a majority to transition the committer from active to emeritus. (If the committer resurfaces, he or she can transition back to active by sending an email to the committers mailing list.)
- Emeritus committers may stay on the committers mailing list to continue to follow any discussions there.
- Emeritus committers do not nominate or vote in committer elections. From a governance perspective, they are equivalent to a non-committer.
- Emeritus committers cannot merge patches to the OVS repository.
- Emeritus committers will be listed in a separate section in the MAINTAINERS.rst file to continue to recognize their contributions to the project.

Emeritus status does not replace the procedures for forcibly removing a committer.

Note that just because a committer is not able to work on the project on a day-to-day basis, we feel they are still capable of providing input on the direction of the project. No committer should feel pressured to move themselves to this status. Again, it's just an option for those that do not currently have the time or interest.

7.9 Expectations for Developers with Open vSwitch Repo Access

7.9.1 Pre-requisites

Be familiar with the guidelines and standards defined in *Contributing to Open vSwitch*.

7.9.2 Review

Code (yours or others') must be reviewed publicly (by you or others) before you push it to the repository. With one exception (see below), every change needs at least one review.

If one or more people know an area of code particularly well, code that affects that area should ordinarily get a review from one of them.

The riskier, more subtle, or more complicated the change, the more careful the review required. When a change needs careful review, use good judgment regarding the quality of reviews. If a change adds 1000 lines of new code, and a review posted 5 minutes later says just "Looks good," then this is probably not a quality review.

(The size of a change is correlated with the amount of care needed in review, but it is not strictly tied to it. A search and replace across many files may not need much review, but one-line optimization changes can have widespread implications.)

Your own small changes to fix a recently broken build ("make") or tests ("make check"), that you believe to be visible to a large number of developers, may be checked in without review. If you are not sure, ask for review. If you do push a build fix without review, send the patch to ovs-dev afterward as usual, indicating in the email that you have already pushed it.

Regularly review submitted code in areas where you have expertise. Consider reviewing other code as well.

7.9.3 Git conventions

Do not push merge commits to the Git repository without prior discussion on ovs-dev.

If you apply a change (yours or another's) then it is your responsibility to handle any resulting problems, especially broken builds and other regressions. If it is someone else's change, then you can ask the original submitter to address it. Regardless, you need to ensure that the problem is fixed in a timely way. The definition of "timely" depends on the severity of the problem.

If a bug is present on main and other branches, fix it on main first, then backport the fix to other branches. Straightforward backports do not require additional review (beyond that for the fix on main).

Feature development should be done only on main. Occasionally it makes sense to add a feature to the most recent release branch, before the first actual release of that branch. These should be handled in the same way as bug fixes, that is, first implemented on main and then backported.

Keep the authorship of a commit clear by maintaining a correct list of "Signed-off-by:"s. If a confusing situation comes up, as it occasionally does, bring it up on the mailing list. If you explain the use of "Signed-off-by:" to a new developer, explain not just how but why, since the intended meaning of "Signed-off-by:" is more important than the syntax. As part of your explanation, quote or provide a URL to the Developer's Certificate of Origin in *Submitting Patches*.

Use Reported-by: and Tested-by: tags in commit messages to indicate the source of a bug report.

Keep the `AUTHORS.rst` file up to date.

7.9.4 Pre-Push Hook

The following script can be helpful because it provides an extra chance to check for mistakes while pushing to the main branch of OVS or OVN. If you would like to use it, install it as `hooks/pre-push` in your `.git` directory and make sure to mark it as executable with `chmod +x`. For maximum utility, make sure `checkpatch.py` is in `$PATH`:

```
#! /bin/bash

remote=$1

case $remote in
    ovs|ovn|origin) ;;
    *) exit 0 ;;
esac

while read local_ref local_sha1 remote_ref remote_sha1; do
    case $remote_ref in
        refs/heads/main)
            n=0
            while read sha
            do
                n=$(expr $n + 1)
                git log -1 $sha
                echo
                checkpatch.py -1 $sha
            done <<EOF
$(git --no-pager log --pretty=%H $local_sha1...$remote_sha1)
EOF

            b=${remote_ref#refs/heads/}
            echo "You're about to push $n commits to protected branch $b on $remote."

            read -p "Do you want to proceed? [y|n] " reply < /dev/tty
            if echo $reply | grep -E '^[Yy]$' > /dev/null; then
                :
            else
                exit 1
            fi
            ;;
    esac
done

exit 0
```

7.10 OVS Committer Grant/Revocation Policy

An OVS committer is a participant in the project with the ability to commit code directly to the main repository. Commit access grants a broad ability to affect the progress of the project as presented by its most important artifact, the code and related resources that produce working binaries of Open vSwitch. As such it represents a significant level of trust in an individual's commitment to working with other committers and the community at large for the benefit of the project. It can not be granted lightly and, in the worst case, must be revocable if the trust placed in an individual was inappropriate.

This document suggests guidelines for granting and revoking commit access. It is intended to provide a framework for evaluation of such decisions without specifying deterministic rules that wouldn't be sensitive to the nuance of specific situations. In the end the decision to grant or revoke committer privileges is a judgment call made by the existing set of committers.

7.10.1 Granting Commit Access

Granting commit access should be considered when a candidate has demonstrated the following in their interaction with the project:

- Contribution of significant new features through the patch submission process where:
  - Submissions are free of obvious critical defects
  - Submissions do not typically require many iterations of improvement to be accepted
- Consistent participation in code review of other's patches, including existing committers, with comments consistent with the overall project standards
- Assistance to those in the community who are less knowledgeable through active participation in project forums such as the ovs-discuss mailing list.
- Plans for sustained contribution to the project compatible with the project's direction as viewed by current committers.
- Commitment to meet the expectations described in the "Expectations of Developer's with Open vSwitch Access"

The process to grant commit access to a candidate is simple:

- An existing committer nominates the candidate by sending an email to all existing committers with information substantiating the contributions of the candidate in the areas described above.
- All existing committers discuss the pros and cons of granting commit access to the candidate in the email thread.
- When the discussion has converged or a reasonable time has elapsed without discussion developing (e.g. a few business days) the nominator calls for a final decision on the candidate with a followup email to the thread.
- Each committer may vote yes, no, or abstain by replying to the email thread. A failure to reply is an implicit abstention.
- After votes from all existing committers have been collected or a reasonable time has elapsed for them to be provided (e.g. a couple of business days) the votes are evaluated. To be granted commit access the candidate must receive yes votes from a majority of the existing committers and zero no votes. Since a no vote is effectively a veto of the candidate it should be accompanied by a reason for the vote.
- The nominator summarizes the result of the vote in an email to all existing committers.
- If the vote to grant commit access passed, the candidate is contacted with an invitation to become a committer to the project which asks them to agree to the committer expectations documented on the project web site.
- If the candidate agrees access is granted by setting up commit access to the repos on github.

7.10.2 Revoking Commit Access

When a committer behaves in a manner that other committers view as detrimental to the future of the project, it raises a delicate situation with the potential for the creation of division within the greater community. These situations should be handled with care. The process in this case is:

- Discuss the behavior of concern with the individual privately and explain why you believe it is detrimental to the project. Stick to the facts and keep the email professional. Avoid personal attacks and the temptation to hypothesize about unknowable information such as the other's motivations. Make it clear that you would prefer not to discuss the behavior more widely but will have to raise it with other contributors if it does not change. Ideally the behavior is eliminated and no further action is required. If not,
- Start an email thread with all committers, including the source of the behavior, describing the behavior and the reason it is detrimental to the project. The message should have the same tone as the private discussion and should generally repeat the same points covered in that discussion. The person whose behavior is being questioned should not be surprised by anything presented in this discussion. Ideally the wider discussion provides more perspective to all participants and the issue is resolved. If not,

- Start an email thread with all committers except the source of the detrimental behavior requesting a vote on revocation of commit rights. Cite the discussion among all committers and describe all the reasons why it was not resolved satisfactorily. This email should be carefully written with the knowledge that the reasoning it contains may be published to the larger community to justify the decision.
- Each committer may vote yes, no, or abstain by replying to the email thread. A failure to reply is an implicit abstention.
- After all votes have been collected or a reasonable time has elapsed for them to be provided (e.g. a couple of business days) the votes are evaluated. For the request to revoke commit access for the candidate to pass it must receive yes votes from two thirds of the existing committers.
- anyone that votes no must provide their reasoning, and
- if the proposal passes then counter-arguments for the reasoning in no votes should also be documented along with the initial reasons the revocation was proposed. Ideally there should be no new counter-arguments supplied in a no vote as all concerns should have surfaced in the discussion before the vote.
- The original person to propose revocation summarizes the result of the vote in an email to all existing committers excepting the candidate for removal.
- If the vote to revoke commit access passes, access is removed and the candidate for revocation is informed of that fact and the reasons for it as documented in the email requesting the revocation vote.
- Ideally the revoked committer peacefully leaves the community and no further action is required. However, there is a distinct possibility that he/she will try to generate support for his/her point of view within the larger community. In this case the reasoning for removing commit access as described in the request for a vote will be published to the community.

7.10.3 Changing the Policy

The process for changing the policy is:

- Propose the changes to the policy in an email to all current committers and request discussion.
- After an appropriate period of discussion (a few days) update the proposal based on feedback if required and resend it to all current committers with a request for a formal vote.
- After all votes have been collected or a reasonable time has elapsed for them to be provided (e.g. a couple of business days) the votes are evaluated. For the request to modify the policy to pass it must receive yes votes from two thirds of the existing committers.

Template Emails

7.10.4 Nomination to Grant Commit Access

> I would like to nominate *[candidate]* for commit access. I believe *[he/she]* has met the conditions for commit access described in the committer grant policy on the project web site in the following ways:
>
> *[list of requirements & evidence]*
>
> Please reply to all in this message thread with your comments and questions. If that discussion concludes favorably I will request a formal vote on the nomination in a few days.

7.10.5 Vote to Grant Commit Access

> I nominated *[candidate]* for commit access on *[date]*. Having allowed sufficient time for discussion it's now time to formally vote on the proposal.

Please reply to all in this thread with your vote of: YES, NO, or ABSTAIN. A failure to reply will be counted as an abstention. If you vote NO, by our policy you must include the reasons for that vote in your reply. The deadline for votes is *[date and time]*.

If a majority of committers vote YES and there are zero NO votes commit access will be granted.

7.10.6 Vote Results for Grant of Commit Access

The voting period for granting to commit access to *[candidate]* initiated at *[date and time]* is now closed with the following results:

YES: *[count of yes votes]* (*[% of voters]*)

NO: *[count of no votes]* (*[% of voters]*)

ABSTAIN: *[count of abstentions]* (*[% of voters]*)

Based on these results commit access *[is/is NOT]* granted.

7.10.7 Invitation to Accepted Committer

Due to your sustained contributions to the Open vSwitch (OVS) project we would like to provide you with commit access to the project repository. Developers with commit access must agree to fulfill specific responsibilities described in the source repository:

/Documentation/internals/committer-responsibilities.rst

Please let us know if you would like to accept commit access and if so that you agree to fulfill these responsibilities. Once we receive your response we'll set up access. We're looking forward continuing to work together to advance the Open vSwitch project.

7.10.8 Proposal to Revoke Commit Access for Detrimental Behavior

I regret that I feel compelled to propose revocation of commit access for *[candidate]*. I have privately discussed with *[him/her]* the following reasons I believe *[his/her]* actions are detrimental to the project and we have failed to come to a mutual understanding:

[List of reasons and supporting evidence]

Please reply to all in this thread with your thoughts on this proposal. I plan to formally propose a vote on the proposal on or after *[date and time]*.

It is important to get all discussion points both for and against the proposal on the table during the discussion period prior to the vote. Please make it a high priority to respond to this proposal with your thoughts.

7.10.9 Vote to Revoke Commit Access

I nominated *[candidate]* for revocation of commit access on *[date]*. Having allowed sufficient time for discussion it's now time to formally vote on the proposal.

Please reply to all in this thread with your vote of: YES, NO, or ABSTAIN. A failure to reply will be counted as an abstention. If you vote NO, by our policy you must include the reasons for that vote in your reply. The deadline for votes is *[date and time]*.

If 2/3rds of committers vote YES commit access will be revoked.

The following reasons for revocation have been given in the original proposal or during discussion:

[list of reasons to remove access]

The following reasons for retaining access were discussed:

[list of reasons to retain access]

The counter-argument for each reason for retaining access is:

[list of counter-arguments for retaining access]

7.10.10 Vote Results for Revocation of Commit Access

The voting period for revoking the commit access of *[candidate]* initiated at *[date and time]* is now closed with the following results:

- YES: *[count of yes votes]* (*[% of voters]*)
- NO: *[count of no votes]* (*[% of voters]*)
- ABSTAIN: *[count of abstentions]* (*[% of voters]*)

Based on these results commit access *[is/is NOT]* revoked. The following reasons for retaining commit access were proposed in NO votes:

[list of reasons]

The counter-arguments for each of these reasons are:

[list of counter-arguments]

7.10.11 Notification of Commit Revocation for Detrimental Behavior

After private discussion with you and careful consideration of the situation, the other committers to the Open vSwitch (OVS) project have concluded that it is in the best interest of the project that your commit access to the project repositories be revoked and this has now occurred.

The reasons for this decision are:

[list of reasons for removing access]

While your goals and those of the project no longer appear to be aligned we greatly appreciate all the work you have done for the project and wish you continued success in your future work.

7.11 Authors

The following people authored or signed off on commits in the Open vSwitch source code or webpage version control repository.

Name	Email
Aaron Conole	aconole@redhat.com
Aaron Rosen	arosen@clemson.edu
Abhiram R N	abhiramrn@gmail.com
Adrian Guzowski	adrian.guzowski@exatel.pl
Adrian Moreno	amorenoz@redhat.com
Aidan Shribman	aidan.shribman@gmail.com
Alan Pevec	alan.pevec@redhat.com
Aleksandr Smirnov	alekssmirnov@k2.cloud
Ales Musil	amusil@redhat.com
Alessandro Pilotti	apilotti@cloudbasesolutions.com
Alexander Duyck	alexander.h.duyck@redhat.com
Alexandra Rukomoinikova	arukomoinikova@k2.cloud
Alexandru Copot	alex.mihai.c@gmail.com

continues on next page

Table 2 – continued from previous page

Name	Email
Alexei Starovoitov	ast@plumgrid.com
Alexey I. Froloff	raorn@raorn.name
Alexey Roytman	roytman@il.ibm.com
Alex Wang	ee07b291@gmail.com
Alfredo Finelli	alf@computationes.de
Alin Balutoiu	abalutoiu@cloudbasesolutions.com
Alin Serdean	aserdean@ovn.org
Allen Chen	allen.chen@jaguarmicro.com
Amber Kumar	kumar.amber@intel.com
Ambika Arora	ambika.arora@tcs.com
Amit Bose	bose@noironetworks.com
Amit Prakash Shukla	amitprakashs@marvell.com
Amitabha Biswas	azbiswas@gmail.com
Anand Kumar	kumaranand@vmware.com
Andrea Kao	eirinikos@gmail.com
Andreas Karis	akaris@redhat.com
Andreas Stieger	andreas.stieger@gmx.de
Andrew Evans	
Andrew Beekhof	abeekhof@redhat.com
Andrew Kampjes	a.kampjes@gmail.com
Andrew Lambeth	alambeth@vmware.com
Andrew Rybchenko	andrew.rybchenko@oktetlabs.ru
Andre McCurdy	armccurdy@gmail.com
Andy Hill	hillad@gmail.com
Andy Southgate	andy.southgate@citrix.com
Andy Zhou	azhou@ovn.org
Ankur Sharma	ankursharma@vmware.com
Anoob Soman	anoob.soman@citrix.com
Ansis Atteka	aatteka@vmware.com
Anton Ivanov	anton.ivanov@cambridgegreys.com
Antonio Fischetti	antonio.fischetti@intel.com
Anupam Chanda	
Ariel Levkovich	lariel@nvidia.com
Ariel Tubaltsev	atubaltsev@vmware.com
Arnoldo Lutz	arnoldo.lutz.guevara@hpe.com
Arun Sharma	arun.sharma@calsoftinc.com
Aryan TaheriMonfared	aryan.taherimonfared@uis.no
Asaf Penso	asafp@mellanox.com
Ashish Varma	ashishvarma.ovs@gmail.com
Ashwin Swaminathan	ashwinds@arista.com
Babu Shanmugam	bschanmu@redhat.com
Bala Sankaran	bsankara@redhat.com
Balazs Nemeth	bnemeth@redhat.com
Ben Pfaff	blp@ovn.org
Ben Warren	ben@skyportsystems.com
Benli Ye	daniely@vmware.com
Bert Vermeulen	bert@biot.com
Bhanuprakash Bodireddy	bhanuprakash.bodireddy@intel.com
Billy O'Mahony	billy.o.mahony@intel.com
Binbin Xu	xu.binbin1@zte.com.cn

continues on next page

Table 2 – continued from previous page

Name	Email
Bodo Petermann	b.petermann@syseleven.de
Boleslaw Tokarski	boleslaw.tokarski@jollamobile.com
Brad Cowie	brad@faucet.nz
Brian Haley	haleyb.dev@gmail.com
Brian Kruger	bkruger+ovsdev@gmail.com
Bruce Davie	bdavie@vmware.com
Bryan Phillippe	bp@toroki.com
Carlo Andreotti	c.andreotti@m3s.it
Casey Barker	crbarker@google.com
Chandan Somani	csomani@redhat.com
Chandler Wu	chandler0149@gmail.com
Chandra Sekhar Vejendla	csvejend@us.ibm.com
Changliang Wu	changliang.wu@smartx.com
Chris Riches	chris.riches@nutanix.com
Chris Wright	chrisw@sous-sol.org
Christoph Jaeger	cj@linux.com
Christophe Fontaine	cfontain@redhat.com
Christopher Aubut	christopher@aubut.me
Chuck Short	zulcss@ubuntu.com
Cian Ferriter	cian.ferriter@intel.com
Ciara Loftus	ciara.loftus@intel.com
Clint Byrum	clint@fewbar.com
Colin Watson	cjwatson@ubuntu.com
Cong Wang	amwang@redhat.com
Conner Herriges	conner.herriges@ibm.com
Damien Millescamps	damien.millescamps@6wind.com
Damijan Skvarc	damjan.skvarc@gmail.com
Dan Carpenter	dan.carpenter@oracle.com
Dan McGregor	dan.mcgregor@usask.ca
Dan Wendlandt	
Dan Williams	dcbw@redhat.com
Daniel Alvarez	dalvarez@redhat.com
Daniel Borkmann	dborkman@redhat.com
Daniel Ding	zhihui.ding@easystack.cn
Daniel Hiltgen	daniel@netkine.com
Daniel Roman	
Daniele Di Proietto	daniele.di.proietto@gmail.com
Daniele Venturino	venturino.daniele+ovs@gmail.com
Danny Kukawka	danny.kukawka@bisect.de
Darrell Ball	dlu998@gmail.com
Dave Tucker	dave@dtucker.co.uk
David Erickson	derickso@stanford.edu
David Hill	dhill@redhat.com
David Marchand	david.marchand@redhat.com
David S. Miller	davem@davemloft.net
David Wilder	dwilder@us.ibm.com
David Yang	davidy@vmware.com
Dennis Sam	dsam@arista.com
Devendra Naga	devendra.aaru@gmail.com
Dexia Li	dexia.li@jaguarmicro.com

continues on next page

Table 2 – continued from previous page

Name	Email
Dima Chumak	dchumak@nvidia.com
Dincer Beken	dbeken@blackned.de
Dmitry Krivenok	krivenok.dmitry@gmail.com
Dmitry Mityugov	dmitry.mityugov@gmail.com
Dmitry Porokh	dporokh@nvidia.com
Dominic Curran	dominic.curran@citrix.com
Dongdong	dongdong1@huawei.com
Dongjun	dongj@dtdream.com
Duan Jiong	djduanjiong@gmail.com
Duffie Cooley	
Dujie	dujie@didiglobal.com
Dumitru Ceara	dceara@redhat.com
Dustin Lundquist	dustin@null-ptr.net
Ed Maste	emaste@freebsd.org
Ed Swierk	eswierk@skyportsystems.com
Edouard Bourguignon	madko@linuxed.net
Eelco Chaudron	echaudro@redhat.com
Eiichi Tsukata	eiichi.tsukata@nutanix.com
Eli Britstein	elibr@nvidia.com
Eli Oliver	eoliver@redhat.com
Emeel Hakim	ehakim@nvidia.com
Emma Finn	emma.finn@intel.com
Eric Lapointe	elapointe@corsa.com
Esteban Rodriguez Betancourt	estebarb@hpe.com
Aymerich Edward	edward.aymerich@hpe.com
Edward Tomasz Napierała	trasz@freebsd.org
Eitan Eliahu	eliahue@vmware.com
Eohyung Lee	liquidnuker@gmail.com
Eric Dumazet	edumazet@google.com
Eric Garver	e@erig.me
Eric Sesterhenn	eric.sesterhenn@lsexperts.de
Ethan J. Jackson	ejj@eecs.berkeley.edu
Ethan Rahn	erahn@arista.com
Eziz Durdyyev	ezizdurdy@gmail.com
Fabrizio D'Angelo	fdangelo@redhat.com
Faicker Mo	faicker.mo@ucloud.cn
fang	fangjiannan@cmss.chinamobile.com
Fangrui Song	maskray@google.com
Felix Huettner	felix.huettner@digits.schwarz
Felix Moebius	felix.moebius@digits.schwarz
Fengqi Li	lifengqi@inspur.com
Flavio Fernandes	flavio@flaviof.com
Flavio Leitner	fbl@redhat.com
Francesco Fusco	ffusco@redhat.com
Frank Wagner	frank.wagner@dbosoft.eu
François Rigault	frigo@amadeus.com
Frédéric Tobias Christ	fchrist@live.de
Frode Nordahl	frode.nordahl@gmail.com
FUJITA Tomonori	fujita.tomonori@lab.ntt.co.jp
Gabe Beged-Dov	gabe@begeddov.com

continues on next page

Table 2 – continued from previous page

Name	Email
Gaetan Rivet	grive@u256.net
Gaetano Catalli	gaetano.catalli@gmail.com
Gal Sagie	gal.sagie@gmail.com
Genevieve LEsperance	glesperance@pivotal.io
Geoffrey Wossum	gwossum@acm.org
Gianluca Merlo	gianluca.merlo@gmail.com
Giuseppe Lettieri	g.lettieri@iet.unipi.it
Glen Gibb	grg@stanford.edu
Gowrishankar Muthukrishnan	gmuthukr@redhat.com
Guoshuai Li	ligs@dtdream.com
Guolin Yang	gyang@vmware.com
Guru Chaitanya Perakam	gperakam@Brocade.com
Gurucharan Shetty	guru@ovn.org
Han Ding	handing@chinatelecom.cn
Han Zhou	zhouhan@gmail.com
Hao Zheng	
Hariprasad Govindharajan	hariprasad.govindharajan@intel.com
Harold Huang	baymaxhuang@gmail.com
Harry Van Haaren	harry.van.haaren@intel.com
Helmut Schaa	helmut.schaa@googlemail.com
Henry Mai	
Hiteshi Kalra	hiteshi.kalra@tcs.com
Hongzhi Guo	guohongzhi1@huawei.com
Huanle Han	hanxueluo@gmail.com
Hui Kang	kangh@us.ibm.com
Hyong Youb Kim	hyonkim@cisco.com
Ian Campbell	Ian.Campbell@citrix.com
Ian Stokes	ian.stokes.oss@gmail.com
Ihar Hrachyshka	ihar.hrachyshka@gmail.com
Ilya Maximets	i.maximets@ovn.org
Iman Tabrizian	tabrizian@outlook.com
Isaku Yamahata	yamahata@valinux.co.jp
Ivan Burnin	iburnin@k2.cloud
Ivan Dyukov	i.dyukov@samsung.com
Ivan Malov	ivan.malov@arknetworks.am
IWASE Yusuke	iwase.yusuke@gmail.com
Jaime Caamaño Ruiz	jcaamano@suse.com
Jakob Meng	code@jakobmeng.de
Jakub Libosvar	libosvar@redhat.com
Jakub Sitnicki	jsitnicki@gmail.com
James P.	roampune@gmail.com
James Page	james.page@ubuntu.com
James Raphael Tiovalen	jamestiotio@gmail.com
Jamie Lennox	jamielennox@gmail.com
Jan Scheurich	jan.scheurich@ericsson.com
Jan Vansteenkiste	jan@vstone.eu
Jarno Rajahalme	jarno@ovn.org
Jason Kölker	jason@koelker.net
Jason Wessel	jason.wessel@windriver.com
Jasper Capel	jasper@capel.tv

continues on next page

Table 2 – continued from previous page

Name	Email
Jay Ding	jay.ding@broadcom.com
Jean Tourrilhes	jt@hpl.hp.com
Jeff Squyres	jsquyres@cisco.com
Jeffrey Walton	noloader@gmail.com
Jeremy Stribling	
Jeroen van Bemmel	jvb127@gmail.com
Jesse Gross	jesse@kernel.org
Jian Li	lijian@ooclab.com
Jiang Lidong	jianglidong3@jd.com
Jianbo Liu	jianbol@mellanox.com
Jing Ai	jinga@google.com
Jinjun Gao	gjinjun@gmail.com
Jiri Benc	jbenc@redhat.com
Joe Perches	joe@perches.com
Joe Stringer	joe@ovn.org
Jon Kohler	jon@nutanix.com
Jonathan Davies	jonathan.davies@nutanix.com
Jonathan Vestin	jonavest@kau.se
Jorge Arturo Sauma Vargas	jorge.sauma@hpe.com
Jun Gu	jun.gu@easystack.cn
Jun Nakajima	jun.nakajima@intel.com
Jun Wang	junwang01@cestc.cn
JunhanYan	juyan@redhat.com
JunoZhu	zhunatuzi@gmail.com
Justin Pettit	jpettit@ovn.org
Kaige Fu	fukaige@huawei.com
Keith Amidon	
Ken Ajiro	ajiro@mxw.nes.nec.co.jp
Ken Sanislo	ken@intherack.com
Kenneth Duda	kduda@arista.com
Kentaro Ebisawa	ebiken.g@gmail.com
Keshav Gupta	keshav.gupta@ericsson.com
Kevin Lo	kevlo@FreeBSD.org
Kevin Sprague	ksprague0711@gmail.com
Kevin Traynor	ktraynor@redhat.com
Khem Raj	raj.khem@gmail.com
Kmindg G	kmindg@gmail.com
Kris Murphy	kriskend@linux.vnet.ibm.com
Krishna Kolakaluri	kkolakaluri@plume.com
Krishna Kondaka	kkondaka@vmware.com
Kyle Mestery	mestery@mestery.com
Kyle Simpson	kyleandrew.simpson@gmail.com
Kyle Upton	kupton@baymicrosystems.com
Lance Yang	lance.yang@arm.com
Lance Richardson	lance.richardson@broadcom.com
Lars Kellogg-Stedman	lars@redhat.com
Lei Huang	huang.f.lei@gmail.com
Leif Madsen	lmadsen@redhat.com
Leo Alterman	
Li RongQing	lirongqing@baidu.com

continues on next page

Table 2 – continued from previous page

Name	Email
Lian-min Wang	liang-min.wang@intel.com
Liang Mancang	liangmc1@chinatelecom.cn
Lin Huang	linhuang@ruijie.com.cn
Liu Chang	liuchang@cmss.chinamobile.com
Lilijun	jerry.lilijun@huawei.com
Lili Huang	huanglili.huang@huawei.com
Liliia Butorina	l.butorina@partner.samsung.com
Linda Sun	lsun@vmware.com
Linda Wang	linda.wang@jaguarmicro.com
Lior Neudorfer	lior@guardicore.com
Liu Chang	txfh2007@aliyun.com
Liu Yulong	liuyulong.xa@gmail.com
Lorand Jakab	lojakab@cisco.com
Lorenzo Bianconi	lorenzo.bianconi@redhat.com
Lubomir Rintel	lkundrak@v3.sk
Luca Giraudo	
Lucas Alvares Gomes	lucasagomes@gmail.com
Lucian Petrut	lpetrut@cloudbasesolutions.com
Luigi Rizzo	rizzo@iet.unipi.it
Luis E. P.	l31g@hotmail.com
Luca Czesla	luca.czesla@digits.schwarz
Lukasz Pawlik	lukaszx.pawlik@intel.com
Lukasz Rzasik	lukasz.rzasik@gmail.com
MJ Ponsonby	mj.ponsonby@canonical.com
Maciej Józefczyk	mjozefcz@redhat.com
Madhu Challa	challa@noironetworks.com
Manohar K C	manukc@gmail.com
Marcin Mirecki	mmirecki@redhat.com
Mario Cabrera	mario.cabrera@hpe.com
Mark D. Gray	mark.d.gray@redhat.com
Mark Hamilton	
Mark Kavanagh	mark.b.kavanagh81@gmail.com
Mark Maglana	mmaglana@gmail.com
Mark Michelson	mmichels@redhat.com
Markos Chandras	mchandras@suse.de
Markus Linnala	markus.linnala@gmail.com
Martin Casado	casado@cs.stanford.edu
Martin Fong	mwfong@csl.sri.com
Martin Kalcok	martin.kalcok@gmail.com
Martin Morgenstern	martin.morgenstern@cloudandheat.com
Martin Varghese	martin.varghese@nokia.com
Martin Xu	martinxu9.ovs@gmail.com
Martin Zhang	martinbj2008@gmail.com
Martino Fornasa	mf@fornasa.it
Maryam Tahhan	maryam.tahhan@intel.com
Matteo Croce	mcroce@redhat.com
Matteo Perin	matteo.perin@canonical.com
Matthias May	matthias.may@neratec.com
Mauricio Vásquez	mauricio.vasquezbernal@studenti.polito.it
Max Lamprecht	max.lamprecht@digits.schwarz

continues on next page

Table 2 – continued from previous page

Name	Email
Maxime Coquelin	maxime.coquelin@redhat.com
Mehak Mahajan	
Michael Arnaldi	arnaldimichael@gmail.com
Michael Santana	msantana@redhat.com
Michael Phelan	michael.phelan@intel.com
Michal Kazior	michal@plume.com
Michal Weglicki	michalx.weglicki@intel.com
Michele Baldessari	michele@acksyn.org
Mickey Spiegel	mickeys.dev@gmail.com
Miguel Angel Ajo	majopela@redhat.com
Miika Petäjäniemi	miika.petajaniemi@solita.fi
Mijo Safradin	mijo@linux.vnet.ibm.com
Mika Vaisanen	mika.vaisanen@gmail.com
Mike Ovsiannikov	mike.ovsiannikov@nutanix.com
Mike Pattrick	mkp@redhat.com
Mikhail Dmitrichenko	m.dmitrichenko222@gmail.com
Minoru TAKAHASHI	takahashi.minoru7@gmail.com
Miro Tomaska	mtomaska@redhat.com
Mohammad Heib	mheib@redhat.com
Moshe Levi	moshele@mellanox.com
Murphy McCauley	murphy.mccauley@gmail.com
Mykola Yurchenko	myurchenko@nvidia.com
Natasha Gude	
Naveen Yerramneni	naveen.yerramneni@nutanix.com
Neal Shrader	neal@digitalocean.com
Neil McKee	neil.mckee@inmon.com
Neil Zhu	zhuj@centecnetworks.com
Nicolas J. Bouliane	nbouliane@digitalocean.com
Nimay Desai	nimaydesai1@gmail.com
Nir Anteby	nanteby@nvidia.com
Nithin Raju	nithin@vmware.com
Niti Rohilla	niti.rohilla@tcs.com
Nitin Katiyar	nitin.katiyar@ericsson.com
Nobuhiro MIKI	nmiki@yahoo-corp.jp
Numan Siddique	nusiddiq@redhat.com
Ofer Ben-Yacov	ofer.benyacov@gmail.com
Ophir Munk	ophirmu@mellanox.com
Or Gerlitz	ogerlitz@mellanox.com
Ori Shoshan	ori.shoshan@guardicore.com
Padmanabhan Krishnan	kprad1@yahoo.com
Panu Matilainen	pmatilai@redhat.com
Paolo Valerio	pvalerio@redhat.com
Paraneetharan Chandrasekaran	paraneetharanc@gmail.com
Paul Boca	pboca@cloudbasesolutions.com
Paul Fazzone	pfazzone@vmware.com
Paul Ingram	
Paul-Emmanuel Raoul	skyper@skyplabs.net
Pavithra Ramesh	paramesh@vmware.com
Peng He	hepeng.0320@bytedance.com
Pengfei Sun	sunpengfei16@huawei.com

continues on next page

Table 2 – continued from previous page

Name	Email
Peter Downs	padowns@gmail.com
Philippe Jung	phil.jung@free.fr
Pim van den Berg	pim@nethuis.nl
pritesh	pritesh.kothari@cisco.com
Pravin B Shelar	pshelar@ovn.org
Przemyslaw Szczerbik	przemyslawx.szczerbik@intel.com
Qian Chen	cq674350529@163.com
Qiuyu Xiao	qiuyu.xiao.qyx@gmail.com
Quentin Monnet	quentin.monnet@6wind.com
Raju Subramanian	
Rami Rosen	ramirose@gmail.com
Ramu Ramamurthy	ramu.ramamurthy@us.ibm.com
Randall Sharo	andall.sharo@navy.mil
Ravi Kerur	Ravi.Kerur@telekom.com
Raymond Burkholder	ray@oneunified.net
Reid Price	
Remi Jouannet	remi.jouannet@outscale.com
Remko Tronçon	git@el-tramo.be
Renat Nurgaliyev	impleman@gmail.com
Rich Lane	rlane@bigswitch.com
Richard Oliver	richard@richard-oliver.co.uk
Rishi Bamba	rishi.bamba@tcs.com
Rob Adams	readams@readams.net
Rob Hoes	rob.hoes@citrix.com
Robert Wojciechowicz	robertx.wojciechowicz@intel.com
Robert Åkerblom-Andersson	Robert.nr1@gmail.com
Roberto Bartzen Acosta	roberto.acosta@luizalabs.com
Robin Jarry	rjarry@redhat.com
Rohith Basavaraja	rohith.basavaraja@gmail.com
Roi Dayan	roid@nvidia.com
Róbert Mulik	robert.mulik@ericsson.com
Romain Lenglet	romain.lenglet@berabera.info
Rosemarie O'Riorden	rosemarie@redhat.com
Roni Bar Yanai	roniba@mellanox.com
Russell Bryant	russell@ovn.org
RYAN D. MOATS	rmoats@us.ibm.com
Ryan Wilson	
Sairam Venugopal	vsairam@vmware.com
Sajjad Lateef	
Salem Sol	salems@nvidia.com
Saloni Jain	saloni.jain@tcs.com
Salvatore Daniele	sdaniele@redhat.com
Samuel Ghinet	sghinet@cloudbasesolutions.com
Sanjay Sane	
Saurabh Mohan	saurabh@cplanenetworks.com
Saurabh Shah	
Saurabh Shrivastava	saurabh.shrivastava@nuagenetworks.net
Sayali Naval	sanaval@cisco.com
Scott Cheloha	scottcheloha@gmail.com
Scott Lowe	scott.lowe@scottlowe.org

continues on next page

Table 2 – continued from previous page

Name	Email
Scott Mann	sdmnix@gmail.com
Seamus Ryan	seamus.ryan@intel.com
Selvamuthukumar	smkumar@merunetworks.com
Sergey Madaminov	sergey.madaminov@gmail.com
Sha Zhang	zhangsha.zhang@huawei.com
Shad Ansari	shad.ansari@hpe.com
Shahar Klein	sklein@nvidia.com
Shan Wei	davidshan@tencent.com
Shaohua Wu	wushaohua@chinatelecom.cn
Sharon Krendel	thekafkaf@gmail.com
Shashank Ram	rams@vmware.com
Shashwat Srivastava	shashwat.srivastava@tcs.com
Shih-Hao Li	shihli@vmware.com
Shu Shen	shu.shen@radisys.com
Simon Horman	horms@ovn.org
Simon Jones	batmanustc@gmail.com
Sivaprasad Tummala	sivaprasad.tummala@intel.com
Somnath Chatterjee	somnath.b.chatterjee@ericsson.com
Songtao Zhan	zhanst1@chinatelecom.cn
Sorin Vinturis	svinturis@cloudbasesolutions.com
Sriharsha Basavapatna	sriharsha.basavapatna@broadcom.com
Stefan Hoffmann	stefan.hoffmann@cloudandheat.com
Steffen Gebert	steffen.gebert@informatik.uni-wuerzburg.de
Sten Spans	sten@blinkenlights.nl
Stephane A. Sezer	sas@cd80.net
Stephen Finucane	stephen@that.guru
Steve Ruan	ruansx@cn.ibm.com
Stuart Cardall	developer@it-offshore.co.uk
Sugesh Chandran	sugesh.chandran@intel.com
SUGYO Kazushi	sugyo.org@gmail.com
Sunyang Wu	sunyang.wu@jaguarmicro.com
Surya Rudra	rudrasurya.r@altencalsoftlabs.com
Tadaaki Nagao	nagao@stratosphere.co.jp
Tao Liu	thomas.liu@ucloud.cn
Tao YunXiang	taoyunxiang@cmss.chinamobile.com
Terry Wilson	twilson@redhat.com
Tetsuo NAKAGAWA	nakagawa@mxc.nes.nec.co.jp
Thadeu Lima de Souza Cascardo	cascardo@cascardo.eti.br
Thilak Raj Surendra Babu	thilakraj.sb@nutanix.com
Thomas F. Herbert	thomasfherbert@gmail.com
Thomas Goirand	zigo@debian.org
Thomas Graf	tgraf@noironetworks.com
Thomas Lacroix	thomas.lacroix@citrix.com
Timo Puha	timox.puha@intel.com
Timothy Redaelli	tredaelli@redhat.com
Todd Deshane	deshantm@gmail.com
Tom Everman	teverman@google.com
Tomasz Konieczny	tomaszx.konieczny@intel.com
Toms Atteka	cpp.code.lv@gmail.com
Tony van der Peet	tony.vanderpeet@alliedtelesis.co.nz

continues on next page

Table 2 – continued from previous page

Name	Email
Torgny Lindberg	torgny.lindberg@ericsson.com
Tsvi Slonim	tsvi@toroki.com
Tuan Nguyen	tuan.nguyen@veriksystems.com
Tyler Coumbes	coumbes@gmail.com
Tony van der Peet	tony.vanderpeet@alliedtelesis.co.nz
Tonghao Zhang	xiangxia.m.yue@gmail.com
Usman Ansari	ua1422@gmail.com
Valient Gough	vgough@pobox.com
Vasu Dasari	vdasari@gmail.com
Vasyl Saienko	vsaienko@mirantis.com
Venkata Anil Kommaddi	vkommadi@redhat.com
Viacheslav Galaktionov	viacheslav.galaktionov@arknetworks.am
Ville Skyttä	ville.skytta@upcloud.com
Vishal Deep Ajmera	vishal.deep.ajmera@ericsson.com
Vivien Bernet-Rollande	vbr@soprive.net
Vlad Buslov	vladbu@nvidia.com
Vladislav Odintsov	odivlad@gmail.com
Volkan Atlı	volkan.atli@b-ulltech.com
Wan Junjie	wanjunjie@bytedance.com
Wang Li	wangli39@baidu.com
Wang Liang	wangliangrt@didiglobal.com
Wang Sheng-Hui	shhuiw@gmail.com
Wang Yibo	bobxxwang@126.com
Wang Zhike	wangzhike@jd.com
wangqianyu	wang.qianyu@zte.com.cn
Wei Li	liw@dtdream.com
Wei Yongjun	yjwei@cn.fujitsu.com
Wenyu Zhang	wenyuz@vmware.com
William Fulton	
William Tu	u9012063@gmail.com
Wilson Peng	pweisong@vmware.com
Xavier Simonart	xsimonar@redhat.com
Xiao Liang	shaw.leon@gmail.com
Xiaojie Chen	jackchanx@163.com
xu rong	xu.rong@zte.com.cn
YAMAMOTO Takashi	yamamoto@midokura.com
Yalei Li	liyl43@chinatelecom.cn
Yang Yang	yangyang92@baidu.com
Yanqin Wei	Yanqin.Wei@arm.com
Yasuhito Takamiya	yasuhito@gmail.com
Yi Li	yili@winhong.com
Yi Yang	yangyi01@inspur.com
Yi-Hung Wei	yihung.wei@gmail.com
Yifeng Sun	pkusunyifeng@gmail.com
Yin Lin	linyi@vmware.com
Yu Zhiguo	yuzg@cn.fujitsu.com
Yuanhan Liu	yuanhan.liu@linux.intel.com
Yunjian Wang	wangyunjian@huawei.com
Yousong Zhou	yszhou4tech@gmail.com
Zak Whittington	zwhitt.vmware@gmail.com

continues on next page

Table 2 – continued from previous page

Name	Email
Zang MingJie	zealot0630@gmail.com
Zengyuan Wang	wangzengyuan@huawei.com
ZhengLingyun	konghuarukhr@163.com
Zhenyu Gao	sysugaozhenyu@gmail.com
Zhi Yong Wu	zwu.kernel@gmail.com
ZhiPeng Lu	luzhipeng@uniudc.com
Zhiqi Chen	chenzhiqi.123@bytedance.com
Zhou Yangchao	1028519445@qq.com
Zoltan Kiss	zoltan.kiss@citrix.com
Zoltán Balogh	zoltan.balogh.eth@gmail.com
Zongkai LI	zealokii@gmail.com
aginwala	amginwal@gmail.com
gordonwwang	gordonwwang@tencent.com
lic121	lic121@chinatelecom.cn
lzhecheng	lzhecheng@vmware.com
parameswaran krishnamurthy	parkrish@gmail.com
solomon	liwei.solomon@gmail.com
wangchuanlei	wangchuanlei@inspur.com
wenxu	wenxu@ucloud.cn
wisd0me	ak47izatool@gmail.com
xushengping	shengping.xu@huawei.com
yangchang	yangchang@chinatelecom.cn
yaolingfei	543981924@qq.com
yinpeijun	yinpeijun@huawei.com
zangchuanqiang	zangchuanqiang@huawei.com
zhaojingjing	zhao.jingjing1@zte.com.cn
zhongbaisong	zhongbaisong@huawei.com
zhaozhanxu	zhaozhanxu@163.com

The following additional people are mentioned in commit logs as having provided helpful bug reports or suggestions.

Name	Email
Aaron M. Ucko	ucko@debian.org
Abhinav Singhal	Abhinav.Singhal@spirent.com
Adam Heath	doogie@brainfood.com
Ahmed Bilal	numan252@gmail.com
Alan Kayahan	hsykay@gmail.com
Alan Shieh	
Alban Browaeys	prahal@yahoo.com
Alex Yip	
Alexey I. Froloff	raorn@altlinux.org
Amar Padmanabhan	
Amey Bhide	
Amre Shakimov	ashakimov@vmware.com
André Ruß	andre.russ@hybris.com
Andreas Beckmann	debian@abeckmann.de
Andrei Andone	andrei.andone@softvision.ro
Andrey Korolyov	andrey@xdel.ru
Anil Jangam	anilj.mailing@gmail.com

continues on next page

Table 3 – continued from previous page

Name	Email
Anshuman Manral	anshuman.manral@outlook.com
Anton Matsiuk	anton.matsiuk@gmail.com
Anup Khadka	khadka.py@gmail.com
Anuprem Chalvadi	achalvadi@vmware.com
Ariel Tubaltsev	atubaltsev@vmware.com
Arkajit Ghosh	arkajit.ghosh@tcs.com
Atzm Watanabe	atzm@stratosphere.co.jp
Aurélien Poulain	aurepoulain@viacesi.fr
Bastian Blank	waldi@debian.org
Ben Basler	
Bhargava Shastry	bshastry@sec.t-labs.tu-berlin.de
Bob Ball	bob.ball@citrix.com
Brad Hall	
Brad Cowie	brad@wand.net.nz
Brailey Josh	josh@faucet.nz
Brandon Heller	brandonh@stanford.edu
Brendan Kelley	
Brent Salisbury	brent.salisbury@gmail.com
Brian Field	Brian_Field@cable.comcast.com
Bryan Fulton	
Bryan Osoro	
Cedric Hobbs	
Chris Hydon	chydon@aristanetworks.com
Christian Stigen Larsen	cslarsen@gmail.com
Christopher Paggen	cpaggen@cisco.com
Chunhe Li	lichunhe@huawei.com
Daniel Badea	daniel.badea@windriver.com
Darragh O'Reilly	darragh.oreilly@hpe.com
Dave Walker	DaveWalker@ubuntu.com
David Evans	davidjoshuaevans@gmail.com
David Palma	palma@onesource.pt
David van Moolenbroek	dvmoolenbroek@aimvalley.nl
Derek Cormier	derek.cormier@lab.ntt.co.jp
Derrick Lim	derrick.lim@rakuten.com
Dhaval Badiani	dbadiani@vmware.com
DK Moon	
Ding Zhi	zhi.ding@6wind.com
Dong Jun	dongj@dtdream.com
Dustin Spinhirne	dspinhirne@vmware.com
Edwin Chiu	echiu@vmware.com
Eivind Bulie Haanaes	
Enas Ahmad	enas.ahmad@kaust.edu.sa
Eric Lopez	
Frank Wang ()	wangpeihui@inspur.com
Frido Roose	fr.roose@gmail.com
Gaetano Catalli	gaetano.catalli@gmail.com
Gavin Remaley	gavin_remaley@selinc.com
Georg Schmuecking	georg.schmuecking@ericsson.com
George Shuklin	amarao@desunote.ru
Gerald Rogers	gerald.rogers@intel.com

continues on next page

Table 3 – continued from previous page

Name	Email
Ghanem Bahri	bahri.ghanem@gmail.com
Giuseppe de Candia	giuseppe.decandia@gmail.com
Gordon Good	ggood@vmware.com
Greg Dahlman	gdahlman@hotmail.com
Greg Rose	gvrose8192@gmail.com
Gregor Schaffrath	grsch@net.t-labs.tu-berlin.de
Gregory Smith	gasmith@nutanix.com
Guolin Yang	gyang@vmware.com
Gur Stavi	gstavi@mrv.com
Harish Kanakaraju	hkanakaraju@vmware.com
Hari Sasank Bhamidipalli	hbhamidi@cisco.com
Hassan Khan	hassan.khan@seecs.edu.pk
Hector Oron	hector.oron@gmail.com
Hemanth Kumar Mantri	mantri@nutanix.com
Henrik Amren	
Hiroshi Tanaka	
Hiroshi Miyata	miyahiro.dazu@gmail.com
Hsin-Yi Shen	shenh@vmware.com
Hui Xiang	xianghuir@gmail.com
Hyojoon Kim	joonk@gatech.edu
Igor Ganichev	
Igor Sever	igor@xorops.com
Jacob Cherkas	cherkasj@vmware.com
Jad Naous	jnaous@gmail.com
Jamal Hadi Salim	hadi@cyberus.ca
James Schmidt	jschmidt@vmware.com
Jan Medved	jmedved@juniper.net
Janis Hamme	janis.hamme@student.kit.edu
Jari Sundell	sundell.software@gmail.com
Javier Albornoz	javier.albornoz@hpe.com
Jed Daniels	openvswitch@jeddaniels.com
Jeff Merrick	jmerrick@vmware.com
Jeongkeun Lee	jklee@hp.com
Jian Qiu	swordqiu@gmail.com
Joan Cirer	joan@ev0.net
John Darrington	john@darrington.wattle.id.au
John Galgay	john@galgay.net
John Hurley	john.hurley@netronome.com
John Reumann	nofutznetworks@gmail.com
Karthik Sundaravel	ksundara@redhat.com
Kashyap Thimmaraju	kashyap.thimmaraju@sec.t-labs.tu-berlin.de
Keith Holleman	hollemanietf@gmail.com
Kevin Lin	kevinlin@berkeley.edu
K	k940545@hotmail.com
Kevin Mancuso	kevin.mancuso@rackspace.com
Kiran Shanbhog	kiran@vmware.com
Kirill Kabardin	
Kirkland Spector	kspector@salesforce.com
Klemens Nanni	klemens@posteo.de
Koichi Yagishita	yagishita.koichi@jrc.co.jp

continues on next page

Table 3 – continued from previous page

Name	Email
Konstantin Khorenko	khorenko@openvz.org
Kris zhang	zhang.kris@gmail.com
Krishna Miriyala	miriyalak@vmware.com
Krishna Mohan Elluru	elluru.kri.mohan@hpe.com
László Sürü	laszlo.suru@ericsson.com
Len Gao	leng@vmware.com
Linhaifeng	haifeng.lin@huawei.com
Logan Rosen	logatronico@gmail.com
Luca Falavigna	dktrkranz@debian.org
Lucas Nussbaum	lucas@debian.org
Luiz Henrique Ozaki	luiz.ozaki@gmail.com
Madhu Venugopal	mavenugo@gmail.com
Malvika Gupta	malvika.gupta@arm.com
Manpreet Singh	er.manpreet25@gmail.com
Mao YingMing	maoyingming@baidu.com
Marco d'Itri	md@Linux.IT
Martin Vizvary	vizvary@ics.muni.cz
Marvin Pascual	marvin@pascual.com.ph
Maxime Brun	m.brun@alphalink.fr
Michael A. Collins	mike.a.collins@ark-net.org
Michael Ben-Ami	mbenami@digitalocean.com
Michael Hu	humichael@vmware.com
Michael J. Smalley	michaeljsmalley@gmail.com
Michael Mao	
Michael Shigorin	mike@osdn.org.ua
Michael Stapelberg	stapelberg@debian.org
Mihir Gangar	gangarm@vmware.com
Mike Bursell	mike.bursell@citrix.com
Mike Kruze	
Mike Qing	mqing@vmware.com
Min Chen	ustcer.tonychan@gmail.com
Mikael Doverhag	
Mircea Ulinic	ping@mirceaulinic.net
Mrinmoy Das	mrdas@ixiacom.com
Muhammad Shahbaz	mshahbaz@cs.princeton.edu
Murali R	muralirdev@gmail.com
Nagi Reddy Jonnala	njonnala@Brocade.com
Niels van Adrichem	N.L.M.vanAdrichem@tudelft.nl
Niklas Andersson	
Oscar Wilde	xdxiaobin@gmail.com
Pankaj Thakkar	pthakkar@vmware.com
Pasi Kärkkäinen	pasik@iki.fi
Patrik Andersson R	patrik.r.andersson@ericsson.com
Paul Greenberg	
Paulo Cravero	pcravero@as2594.net
Pawan Shukla	shuklap@vmware.com
Periyasamy Palanisamy	periyasamy.palanisamy@ericsson.com
Peter Amidon	peter@picnicpark.org
Peter Balland	
Peter Phaal	peter.phaal@inmon.com

continues on next page

Table 3 – continued from previous page

Name	Email
Prabina Pattnaik	Prabina.Pattnaik@nechclst.in
Pratap Reddy	
Ralf Heiringhoff	ralf@frosty-geek.net
Ram Jothikumar	
Ramana Reddy	gtvrreddy@gmail.com
Ray Li	rayli1107@gmail.com
Richard Theis	rtheis@us.ibm.com
RishiRaj Maulick	rishi.raj2509@gmail.com
Rob Sherwood	rob.sherwood@bigswitch.com
Robert Strickler	anomalyst@gmail.com
Roger Leigh	rleigh@codelibre.net
Rogério Vinhal Nunes	
Roman Sokolkov	rsokolkov@gmail.com
Ronaldo A. Ferreira	ronaldof@CS.Princeton.EDU
Ronny L. Bull	bullrl@clarkson.edu
Sandeep Kumar	sandeep.kumar16@tcs.com
Sander Eikelenboom	linux@eikelenboom.it
Saul St. John	sstjohn@cs.wisc.edu
Scott Hendricks	
Sean Brady	sbrady@gtfservices.com
Sebastian Andrzej Siewior	sebastian@breakpoint.cc
Sébastien RICCIO	sr@swisscenter.com
Seiji Sakurai	Seiji.Sakurai@outlook.com
Shweta Seth	shwseth@cisco.com
Simon Jouet	simon.jouet@gmail.com
Spiro Kourtessis	spiro@vmware.com
Sridhar Samudrala	samudrala.sridhar@gmail.com
Srini Seetharaman	seethara@stanford.edu
Sabyasachi Sengupta	Sabyasachi.Sengupta@alcatel-lucent.com
Salvatore Cambria	salvatore.cambria@citrix.com
Soner Sevinc	sevincs@vmware.com
Stepan Andrushko	stepanx.andrushko@intel.com
Stephen Hemminger	shemminger@vyatta.com
Stuart Cardall	developer@it-offshore.co.uk
Suganya Ramachandran	suganyar@vmware.com
Sundar Nadathur	undar.nadathur@intel.com
Taekho Nam	thnam@smartx.kr
Takayuki HAMA	t-hama@cb.jp.nec.com
Teemu Koponen	
Thomas Morin	thomas.morin@orange.com
Timothy Chen	
Torbjorn Tornkvist	kruskakli@gmail.com
Tulio Ribeiro	tribeiro@lasige.di.fc.ul.pt
Tytus Kurek	Tytus.Kurek@pega.com
Valentin Bud	valentin@hackaserver.com
Vasiliy Tolstov	v.tolstov@selfip.ru
Vinllen Chen	cvinllen@gmail.com
Vipul Ashri	vipul.ashri@ericsson.com
Vishal Swarankar	vishal.swarnkar@gmail.com
Vjekoslav Brajkovic	balkan@cs.washington.edu

continues on next page

Table 3 – continued from previous page

Name	Email
Voravit T.	voravit@kth.se
Yeming Zhao	zhaoyeming@gmail.com
Yi Ba	yby.developer@yahoo.com
Ying Chen	yingchen@vmware.com
Yongqiang Liu	liuyq7809@gmail.com
ZHANG Zhiming	zhangzhiming@yunshan.net.cn
Zhangguanghui	zhang.guanghui@h3c.com
Zheng Jingzhou	glovejmm@163.com
Ziyou Wang	ziyouw@vmware.com
ankur dwivedi	ankurengg2003@gmail.com
chen zhang	3zhangchen9211@gmail.com
james hopper	jameshopper@email.com
kk yap	yapkke@stanford.edu
likunyun	kunyunli@hotmail.com
meishengxin	meishengxin@huawei.com
neeraj mehta	mehtaneeraj07@gmail.com
rahim entezari	rahim.entezari@gmail.com
shaoke xi	xishaoke.xsk@gmail.com
shivani dommeti	shivani.dommeti@gmail.com
Wentao Jia	wentao.jia@easystack.cn
weizj	34965317@qq.com
	zhaojun12@outlook.com
(Crab)	fqs888@126.com
	fortitude.zhang@gmail.com
	hujingfei914@msn.com
	zhangwqh@126.com
	zhangqiang@meizu.com

Thanks to all Open vSwitch contributors. If you are not listed above but believe that you should be, please write to dev@openvswitch.org.

7.12 Committers

Open vSwitch committers are the people who have been granted access to push changes to the Open vSwitch git repository.

The responsibilities of an Open vSwitch committer are documented here: *Expectations for Developers with Open vSwitch Repo Access*.

The process for adding or removing committers is documented here: *OVS Committer Grant/Revocation Policy*.

This is the current list of active Open vSwitch committers:

Table 4: OVS Maintainers

Name	Email
Aaron Conole	aconole@redhat.com
Alin Serdean	aserdean@ovn.org
Ansis Atteka	ansisatteka@gmail.com
Eelco Chaudron	echaudro@redhat.com
Ian Stokes	istokes@ovn.org
Ilya Maximets	i.maximets@ovn.org
Kevin Traynor	ktraynor@redhat.com
Simon Horman	horms@ovn.org
William Tu	u9012063@gmail.com

The project also maintains a list of Emeritus Committers (or Maintainers). More information about Emeritus Committers can be found here: *Emeritus Status for OVS Committers*.

Table 5: OVS Emeritus Maintainers

Name	Email
Alex Wang	ee07b291@gmail.com
Andy Zhou	azhou@ovn.org
Ben Pfaff	blp@ovn.org
Daniele Di Proietto	daniele.di.proietto@gmail.com
Ethan J. Jackson	ejj@eecs.berkeley.edu
Gurucharan Shetty	guru@ovn.org
Jarno Rajahalme	jarno@ovn.org
Jesse Gross	jesse@kernel.org
Joe Stringer	joe@ovn.org
Justin Pettit	jpettit@ovn.org
Pravin B Shelar	pshelar@ovn.org
Russell Bryant	russell@ovn.org
Thomas Graf	tgraf@tgraf.ch
YAMAMOTO Takashi	yamamoto@midokura.com

7.13 How Open vSwitch's Documentation Works

This document provides a brief overview on how the documentation build system within Open vSwitch works. This is intended to maximize the "bus factor" and share best practices with other projects.

7.13.1 reStructuredText and Sphinx

Nearly all of Open vSwitch's documentation is written in reStructuredText, with man pages being the sole exception. Of this documentation, most of it is fed into Sphinx, which provides not only the ability to convert rST to a variety of other output formats but also allows for things like cross-referencing and indexing. for more information on the two, refer to the *Documentation Style*.

7.13.2 ovs-sphinx-theme

The documentation uses its own theme, *ovs-sphinx-theme*, which can be found on GitHub and is published on pypi. This is packaged separately from Open vSwitch itself to ensure all documentation gets the latest version of the theme (assuming there are no major version bumps in that package). If building locally and the package is installed, it will be used. If the package is not installed, Sphinx will fallback to the default theme.

The package is currently maintained by Stephen Finucane and Russell Bryant.

7.13.3 Read the Docs

The documentation is hosted on readthedocs.org and a CNAME redirect is in place to allow access from docs.openvswitch.org. *Read the Docs* provides a couple of nifty features for us, such as automatic building of docs whenever there are changes and versioning of documentation.

The *Read the Docs* project is currently maintained by Stephen Finucane, Russell Bryant and Ben Pfaff.

7.13.4 openvswitch.org

The sources for openvswitch.org are maintained separately from docs.openvswitch.org. For modifications to this site, refer to the GitHub project.

CHAPTER

EIGHT

FAQ

8.1 Bareudp

Q: What is Bareudp?

A: There are various L3 encapsulation standards using UDP being discussed
to leverage the UDP based load balancing capability of different networks. MPLSoUDP (__ https://tools.ietf.org/html/rfc7510) is one among them.

The Bareudp tunnel provides a generic L3 encapsulation support for tunnelling different L3 protocols like MPLS, IP, NSH etc. inside a UDP tunnel.

An example to create bareudp device to tunnel MPLS unicast traffic is given below.:

```
$ ovs-vsctl add-port br0 mpls_udp_port -- set interface udp_port \
  type=bareudp options:remote_ip=2.1.1.3 options:local_ip=2.1.1.2 \
  options:payload_type=0x8847 options:dst_port=6635
```

The option payload_type specifies the ethertype of the l3 protocol which the bareudp device will be tunnelling.

The bareudp device supports special handling for MPLS & IP as they can have multiple ethertypes. MPLS procotcol can have ethertypes ETH_P_MPLS_UC (unicast) & ETH_P_MPLS_MC (multicast). IP protocol can have ethertypes ETH_P_IP (v4) & ETH_P_IPV6 (v6).

The bareudp device to tunnel L3 traffic with multiple ethertypes (MPLS & IP) can be created by passing the L3 protocol name as string in the field payload_type.

An example to create bareudp device to tunnel MPLS unicast & multicast traffic is given below.:

```
$ ovs-vsctl add-port  br0 mpls_udp_port -- set interface udp_port \
  type=bareudp options:remote_ip=2.1.1.3 options:local_ip=2.1.1.2 \
  options:payload_type=mpls options:dst_port=6635
```

The below example ovs rule shows how a bareudp tunnel port is used to tunnel an MPLS packet inside a UDP tunnel.:

```
$ ovs-ofctl -O OpenFlow13 add-flow br0 "in_port=10,dl_type=0x0800,\
  actions=push_mpls:0x8847,set_field:3->mpls_label,\
  output:mpls_udp_port"
```

This rule does MPLS encapsulation on IP packets and sends the l3 MPLS packets on a bareudp tunnel port which has its payload_type configured to 0x8847.

An example to create bareudp device to tunnel IPv4 & IPv6 traffic is given below.:

```
$ ovs-vsctl add-port  br0 ip_udp_port -- set interface udp_port \
  type=bareudp options:remote_ip=2.1.1.3 options:local_ip=2.1.1.2 \
  options:payload_type=ip options:dst_port=6636
```

8.2 Basic Configuration

Q: How do I configure a port as an access port?

A. Add `tag=VLAN` to your `ovs-vsctl add-port` command. For example, the following commands configure br0 with eth0 as a trunk port (the default) and tap0 as an access port for VLAN 9:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 eth0
$ ovs-vsctl add-port br0 tap0 tag=9
```

If you want to configure an already added port as an access port, use `ovs-vsctl set`, e.g.:

```
$ ovs-vsctl set port tap0 tag=9
```

Q: How do I configure a port as a SPAN port, that is, enable mirroring of all traffic to that port?

A. The following commands configure br0 with eth0 and tap0 as trunk ports. All traffic coming in or going out on eth0 or tap0 is also mirrored to tap1; any traffic arriving on tap1 is dropped:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 eth0
$ ovs-vsctl add-port br0 tap0
$ ovs-vsctl add-port br0 tap1 \
    -- --id=@p get port tap1 \
    -- --id=@m create mirror name=m0 select-all=true output-port=@p \
    -- set bridge br0 mirrors=@m
```

To later disable mirroring, run:

```
$ ovs-vsctl clear bridge br0 mirrors
```

Q: Does Open vSwitch support configuring a port in promiscuous mode?

A: Yes. How you configure it depends on what you mean by "promiscuous mode":

- Conventionally, "promiscuous mode" is a feature of a network interface card. Ordinarily, a NIC passes to the CPU only the packets actually destined to its host machine. It discards the rest to avoid wasting memory and CPU cycles. When promiscuous mode is enabled, however, it passes every packet to the CPU. On an old-style shared-media or hub-based network, this allows the host to spy on all packets on the network. But in the switched networks that are almost everywhere these days, promiscuous mode doesn't have much effect, because few packets not destined to a host are delivered to the host's NIC.

  This form of promiscuous mode is configured in the guest OS of the VMs on your bridge, e.g. with "ip link set <device> promisc".

- The VMware vSwitch uses a different definition of "promiscuous mode". When you configure promiscuous mode on a VMware vNIC, the vSwitch sends a copy of every packet received by the vSwitch to that vNIC. That has a much bigger effect than just enabling promiscuous mode in a guest OS. Rather than getting a few stray packets for which the switch does not yet know the correct destination, the vNIC gets every packet. The effect is similar to replacing the vSwitch by a virtual hub.

This "promiscuous mode" is what switches normally call "port mirroring" or "SPAN". For information on how to configure SPAN, see "How do I configure a port as a SPAN port, that is, enable mirroring of all traffic to that port?"

Q: How do I configure a DPDK port as an access port?

A: Firstly, you must have a DPDK-enabled version of Open vSwitch.

If your version is DPDK-enabled it may support the dpdk_version and dpdk_initialized keys in the configuration database. Earlier versions of Open vSwitch only supported the other-config:dpdk-init key in the configuration in the database. All versions will display lines with "EAL:..." during startup when other_config:dpdk-init is set to 'true'.

Secondly, when adding a DPDK port, unlike a system port, the type for the interface and valid dpdk-devargs must be specified. For example:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 myportname -- set Interface myportname \
    type=dpdk options:dpdk-devargs=0000:06:00.0
```

Refer to *Open vSwitch with DPDK* for more information on enabling and using DPDK with Open vSwitch.

Q: How do I configure a VLAN as an RSPAN VLAN, that is, enable mirroring of all traffic to that VLAN?

A: The following commands configure br0 with eth0 as a trunk port and tap0 as an access port for VLAN 10. All traffic coming in or going out on tap0, as well as traffic coming in or going out on eth0 in VLAN 10, is also mirrored to VLAN 15 on eth0. The original tag for VLAN 10, in cases where one is present, is dropped as part of mirroring:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 eth0
$ ovs-vsctl add-port br0 tap0 tag=10
$ ovs-vsctl \
    -- --id=@m create mirror name=m0 select-all=true select-vlan=10 \
       output-vlan=15 \
    -- set bridge br0 mirrors=@m
```

To later disable mirroring, run:

```
$ ovs-vsctl clear bridge br0 mirrors
```

Mirroring to a VLAN can disrupt a network that contains unmanaged switches. See ovs-vswitchd.conf.db(5) for details. Mirroring to a GRE tunnel has fewer caveats than mirroring to a VLAN and should generally be preferred.

Q: Can I mirror more than one input VLAN to an RSPAN VLAN?

A: Yes, but mirroring to a VLAN strips the original VLAN tag in favor of the specified output-vlan. This loss of information may make the mirrored traffic too hard to interpret.

To mirror multiple VLANs, use the commands above, but specify a comma-separated list of VLANs as the value for select-vlan. To mirror every VLAN, use the commands above, but omit select-vlan and its value entirely.

When a packet arrives on a VLAN that is used as a mirror output VLAN, the mirror is disregarded. Instead, in standalone mode, OVS floods the packet across all the ports for which the mirror output VLAN is configured. (If an OpenFlow controller is in use, then it can override this behavior through the flow table.) If OVS is used as an intermediate switch, rather than an edge switch, this ensures that the RSPAN traffic is distributed through the network.

Mirroring to a VLAN can disrupt a network that contains unmanaged switches. See ovs-vswitchd.conf.db(5) for details. Mirroring to a GRE tunnel has fewer caveats than mirroring to a VLAN and should generally be preferred.

Q: How do I configure mirroring of all traffic to a GRE tunnel?

A: The following commands configure br0 with eth0 and tap0 as trunk ports. All traffic coming in or going out on eth0 or tap0 is also mirrored to gre0, a GRE tunnel to the remote host 192.168.1.10; any traffic arriving on gre0 is dropped:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 eth0
$ ovs-vsctl add-port br0 tap0
$ ovs-vsctl add-port br0 gre0 \
    -- set interface gre0 type=gre options:remote_ip=192.168.1.10 \
    -- --id=@p get port gre0 \
    -- --id=@m create mirror name=m0 select-all=true output-port=@p \
    -- set bridge br0 mirrors=@m
```

To later disable mirroring and destroy the GRE tunnel:

```
$ ovs-vsctl clear bridge br0 mirrors
$ ovs-vsctl del-port br0 gre0
```

Q: Does Open vSwitch support ERSPAN?

A: Yes. ERSPAN version I and version II over IPv4 GRE and IPv6 GRE tunnel are supported. See ovs-fields(7) for matching and setting ERSPAN fields.

```
$ ovs-vsctl add-br br0
$ #For ERSPAN type 2 (version I)
$ ovs-vsctl add-port br0 at_erspan0 -- \
       set int at_erspan0 type=erspan options:key=1 \
       options:remote_ip=172.31.1.1 \
       options:erspan_ver=1 options:erspan_idx=1
$ #For ERSPAN type 3 (version II)
$ ovs-vsctl add-port br0 at_erspan0 -- \
       set int at_erspan0 type=erspan options:key=1 \
       options:remote_ip=172.31.1.1 \
       options:erspan_ver=2 options:erspan_dir=1 \
       options:erspan_hwid=4
```

Q: Does Open vSwitch support IPv6 GRE?

A: Yes. L2 tunnel interface GRE over IPv6 is supported. L3 GRE tunnel over IPv6 is not supported.

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 at_gre0 -- \
       set int at_gre0 type=ip6gre \
       options:remote_ip=fc00:100::1 \
       options:packet_type=legacy_l2
```

Q: Does Open vSwitch support GTP-U?

A: Yes. Starting with version 2.13, the Open vSwitch userspace datapath supports GTP-U (GPRS Tunnelling Protocol User Plane (GTPv1-U)). TEID is set by using tunnel key field.

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 gtpu0 -- \
       set int gtpu0 type=gtpu options:key=<teid> \
       options:remote_ip=172.31.1.1
```

Q: Does Open vSwitch support SRv6?

A: Yes. Starting with version 3.2, the Open vSwitch userspace datapath supports SRv6 (Segment Routing over IPv6). The following example shows tunneling to fc00:300::1 via fc00:100::1 and fc00:200::1. In the current implementation, if "IPv6 in IPv6" or "IPv4 in IPv6" packets are routed to this interface, and these packets are not SRv6 packets, they may be dropped, so be careful in workloads with a mix of these tunnels. Also note the following restrictions:

- Segment list length is limited to 6.
- SRv6 packets with other than segments_left = 0 are simply dropped.

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 srv6_0 -- \
       set int srv6_0 type=srv6  \
       options:remote_ip=fc00:100::1 \
       options:srv6_segs="fc00:100::1,fc00:200::1,fc00:300::1"
```

Q: How do I connect two bridges?

A: First, why do you want to do this? Two connected bridges are not much different from a single bridge, so you might as well just have a single bridge with all your ports on it.

If you still want to connect two bridges, you can use a pair of patch ports. The following example creates bridges br0 and br1, adds eth0 and tap0 to br0, adds tap1 to br1, and then connects br0 and br1 with a pair of patch ports.

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 eth0
$ ovs-vsctl add-port br0 tap0
$ ovs-vsctl add-br br1
$ ovs-vsctl add-port br1 tap1
$ ovs-vsctl \
    -- add-port br0 patch0 \
    -- set interface patch0 type=patch options:peer=patch1 \
    -- add-port br1 patch1 \
    -- set interface patch1 type=patch options:peer=patch0
```

Bridges connected with patch ports are much like a single bridge. For instance, if the example above also added eth1 to br1, and both eth0 and eth1 happened to be connected to the same next-hop switch, then you could loop your network just as you would if you added eth0 and eth1 to the same bridge (see the "Configuration Problems" section below for more information).

Q: How do I configure a bridge without an OpenFlow local port? (Local port in the sense of OFPP_LOCAL)

A: Open vSwitch does not support such a configuration. Bridges always have their local ports.

Q: Why does OVS pick its default datapath ID the way it does?

A: The default OpenFlow datapath ID for a bridge is the minimum non-local MAC address among all of the ports in a bridge. This means that a bridge with a given set of physical ports will always have the same datapath ID. This is useful for virtualization systems, which typically put a single physical port (or a single bond of multiple ports) on a given bridge alongside the virtual ports for running VMs. In such a setup, the

IP address for the NIC associated with a physical port gets migrated from the physical NIC to the bridge port. The bridge port should have the same MAC address as the physical NIC, so that the host doesn't suddenly start using a different MAC, and taking the minimum MAC address does this automatically and, if there is bond, consistently. Virtual ports for running VMs do not affect the situation because these normally have the "local" bit set, which OVS ignores.

If you want a stable MAC and datapath ID, you could set your own MAC by `hwaddr` in `other_config` of bridge.

```
ovs-vsctl set bridge br-int other_config:hwaddr=3a:4d:a7:05:2a:45
```

8.3 Development

Q: How do I implement a new OpenFlow message?

A: Add your new message to `enum ofpraw` and `enum ofptype` in `include/openvswitch/ofp-msgs.h`, following the existing pattern. Then recompile and fix all of the new warnings, implementing new functionality for the new message as needed. (If you configure with `--enable-Werror`, as described in *Open vSwitch on Linux, FreeBSD and NetBSD*, then it is impossible to miss any warnings.)

To add an OpenFlow vendor extension message (aka experimenter message) for a vendor that doesn't yet have any extension messages, you will also need to edit `build-aux/extract-ofp-msgs` and at least `ofphdrs_decode()` and `ofpraw_put__()` in `lib/ofp-msgs.c`. OpenFlow doesn't standardize vendor extensions very well, so it's hard to make the process simpler than that. (If you have a choice of how to design your vendor extension messages, it will be easier if you make them resemble the ONF and OVS extension messages.)

Q: How do I add support for a new field or header?

A: Add new members for your field to `struct flow` in `include/openvswitch/flow.h`, and add new enumerations for your new field to `enum mf_field_id` in `include/openvswitch/meta-flow.h`, following the existing pattern. If the field uses a new OXM class, add it to OXM_CLASSES in `build-aux/extract-ofp-fields`. Also, add support to `miniflow_extract()` in `lib/flow.c` for extracting your new field from a packet into struct miniflow, and to `nx_put_raw()` in `lib/nx-match.c` to output your new field in OXM matches. Then recompile and fix all of the new warnings, implementing new functionality for the new field or header as needed. (If you configure with `--enable-Werror`, as described in *Open vSwitch on Linux, FreeBSD and NetBSD*, then it is impossible to miss any warnings.)

If you want kernel datapath support for your new field, you also need to modify the kernel module for the operating systems you are interested in. This isn't mandatory, since fields understood only by userspace work too (with a performance penalty), so it's reasonable to start development without it. If you implement kernel module support for Linux, then the Linux kernel "netdev" mailing list is the place to submit that support first; please read up on the Linux kernel development process separately.

Q: How do I add support for a new OpenFlow action?

A: Add your new action to `enum ofp_raw_action_type` in `lib/ofp-actions.c`, following the existing pattern. Then recompile and fix all of the new warnings, implementing new functionality for the new action as needed. (If you configure with `--enable-Werror`, as described in the *Open vSwitch on Linux, FreeBSD and NetBSD*, then it is impossible to miss any warnings.)

If you need to add an OpenFlow vendor extension action for a vendor that doesn't yet have any extension actions, then you will also need to add the vendor to `vendor_map` in `build-aux/extract-ofp-actions`. Also, you will need to add support for the vendor to `ofpact_decode_raw()` and `ofpact_put_raw()` in `lib/ofp-actions.c`. (If you have a choice of how to design your vendor extension actions, it will be easier if you make them resemble the ONF and OVS extension actions.)

Q: How do I add support for a new OpenFlow error message?

A: Add your new error to `enum ofperr` in `include/openvswitch/ofp-errors.h`. Read the large comment at the top of the file for details. If you need to add an OpenFlow vendor extension error for a vendor that doesn't yet have any, first add the vendor ID to the `<name>_VENDOR_ID` list in `include/openflow/openflow-common.h`.

Q: What's a Signed-off-by and how do I provide one?

A: Free and open source software projects usually require a contributor to provide some assurance that they're entitled to contribute the code that they provide. Some projects, for example, do this with a Contributor License Agreement (CLA) or a copyright assignment that is signed on paper or electronically.

For this purpose, Open vSwitch has adopted something called the Developer's Certificate of Origin (DCO), which is also used by the Linux kernel and originated there. Informally stated, agreeing to the DCO is the developer's way of attesting that a particular commit that they are contributing is one that they are allowed to contribute. You should visit https://developercertificate.org/ to read the full statement of the DCO, which is less than 200 words long.

To certify compliance with the Developer's Certificate of Origin for a particular commit, just add the following line to the end of your commit message, properly substituting your name and email address:

Signed-off-by: Firstname Lastname <email@example.org>

Git has special support for adding a Signed-off-by line to a commit message: when you run "git commit", just add the -s option, as in "git commit -s". If you use the "git citool" GUI for commits, you can add a Signed-off-by line to the commit message by pressing Control+S. Other Git user interfaces may provide similar support.

Q: How do I apply patches from email?

A: You can use `git am` on raw email contents, either from a file saved by or piped from an email client. In `mutt`, for example, when you are viewing a patch, you can apply it to the tree in ~/ovs by issuing the command `|cd ~/ovs && git am`. If you are an OVS committer, you might want to add `-s` to sign off on the patch as part of applying it. If you do this often, then you can make the keystrokes `,a` shorthand for it by adding the following line to your `.muttrc`:

macro index,pager ,a "<pipe-message>cd ~/ovs && git am -s" "apply patch"

`git am` has a problem with some email messages from the ovs-dev list for which the mailing list manager edits the From: address, replacing it by the list's own address. The mailing list manager must do this for messages whose sender's email domain has DMARC configured, because receivers will otherwise discard these messages when they do not come directly from the sender's email domain. This editing makes the patches look like they come from the mailing list instead of the author. To work around this problem, one can use the following wrapper script for `git am`:

```
#! /bin/sh
tmp=$(mktemp)
cat >$tmp
if grep '^From:.*via dev.*' "$tmp" >/dev/null 2>&1; then
   sed '/^From:.*via dev.*/d
        s/^[Rr]eply-[tT]o:/From:/' $tmp
else
   cat "$tmp"
fi | git am "$@"
rm "$tmp"
```

Another way to apply emailed patches is to use the `pwclient` program, which can obtain patches from patchwork and apply them directly. Download `pwclient` at https://patchwork.ozlabs.org/project/openvswitch/. You probably want to set up a `.pwclientrc` that looks something like this:

```
[options]
default=openvswitch
signoff=true

[openvswitch]
url=https://patchwork.ozlabs.org/xmlrpc/
```

After you install `pwclient`, you can apply a patch from patchwork with `pwclient git-am #`, where # is the patch's number. (This fails with certain patches that contain form-feeds, due to a limitation of the protocol underlying `pwclient`.)

Another way to apply patches directly from patchwork which supports applying patch series is to use the `git-pw` program. It can be obtained with `pip install git-pw`. Alternative installation instructions and general documentation can be found at https://patchwork.readthedocs.io/projects/git-pw/en/latest/. You need to use your openvswitch patchwork login or create one at https://patchwork.ozlabs.org/register/. The following can then be set on the command line with `git config` or through a `.gitconfig` like this:

```
[pw]
server=https://patchwork.ozlabs.org/api/1.0
project=openvswitch
username=<username>
password=<password>
```

Patch series can be listed with `git-pw series list` and applied with `git-pw series apply #`, where # is the series number. Individual patches can be applied with `git-pw patch apply #`, where # is the patch number.

8.4 Implementation Details

Q: I hear OVS has a couple of kinds of flows. Can you tell me about them?

A: Open vSwitch uses different kinds of flows for different purposes:

- OpenFlow flows are the most important kind of flow. OpenFlow controllers use these flows to define a switch's policy. OpenFlow flows support wildcards, priorities, and multiple tables.

  When in-band control is in use, Open vSwitch sets up a few "hidden" flows, with priority higher than a controller or the user can configure, that are not visible via OpenFlow. (See the "Controller" section of the FAQ for more information about hidden flows.)

- The Open vSwitch software switch implementation uses a second kind of flow internally. These flows, called "datapath" or "kernel" flows, do not support priorities and comprise only a single table, which makes them suitable for caching. (Like OpenFlow flows, datapath flows do support wildcarding, in Open vSwitch 1.11 and later.) OpenFlow flows and datapath flows also support different actions and number ports differently.

  Datapath flows are an implementation detail that is subject to change in future versions of Open vSwitch. Even with the current version of Open vSwitch, hardware switch implementations do not necessarily use this architecture.

Users and controllers directly control only the OpenFlow flow table. Open vSwitch manages the datapath flow table itself, so users should not normally be concerned with it.

Q: Why are there so many different ways to dump flows?

A: Open vSwitch has two kinds of flows (see the previous question), so it has commands with different purposes for dumping each kind of flow:

- `ovs-ofctl dump-flows <br>` dumps OpenFlow flows, excluding hidden flows. This is the most commonly useful form of flow dump. (Unlike the other commands, this should work with any OpenFlow switch, not just Open vSwitch.)
- `ovs-appctl bridge/dump-flows <br>` dumps OpenFlow flows, including hidden flows. This is occasionally useful for troubleshooting suspected issues with in-band control.
- `ovs-dpctl dump-flows [dp]` dumps the datapath flow table entries for a Linux kernel-based datapath. In Open vSwitch 1.10 and later, ovs-vswitchd merges multiple switches into a single datapath, so it will show all the flows on all your kernel-based switches. This command can occasionally be useful for debugging. It doesn't dump flows that was offloaded to hardware.
- `ovs-appctl dpif/dump-flows <br>`, new in Open vSwitch 1.10, dumps datapath flows for only the specified bridge, regardless of the type. Supports dumping of HW offloaded flows. See ovs-vswitchd(8) for details.

Q: How does multicast snooping works with VLANs?

A: Open vSwitch maintains snooping tables for each VLAN.

Q: Can OVS populate the kernel flow table in advance instead of in reaction to packets?

A: No. There are several reasons:

- Kernel flows are not as sophisticated as OpenFlow flows, which means that some OpenFlow policies could require a large number of kernel flows. The "conjunctive match" feature is an extreme example: the number of kernel flows it requires is the product of the number of flows in each dimension.
- With multiple OpenFlow flow tables and simple sets of actions, the number of kernel flows required can be as large as the product of the number of flows in each dimension. With more sophisticated actions, the number of kernel flows could be even larger.
- Open vSwitch is designed so that any version of OVS userspace interoperates with any version of the OVS kernel module. This forward and backward compatibility requires that userspace observe how the kernel module parses received packets. This is only possible in a straightforward way when userspace adds kernel flows in reaction to received packets.

For more relevant information on the architecture of Open vSwitch, please read "The Design and Implementation of Open vSwitch", published in USENIX NSDI 2015.

Q: How many packets does OVS buffer?

A: Open vSwitch fast path packet processing uses a "run to completion" model in which every packet is completely handled in a single pass. Therefore, in the common case where a packet just passes through the fast path, Open vSwitch does not buffer packets itself. The operating system and the network drivers involved in receiving and later in transmitting the packet do often include buffering. Open vSwitch is only a middleman between these and does not have direct access or influence over their buffers.

Outside the common case, Open vSwitch does sometimes buffer packets. When the OVS fast path processes a packet that does not match any of the flows in its megaflow cache, it passes that packet to the Open vSwitch slow path. This procedure queues a copy of the packet to the Open vSwitch userspace which processes it and, if necessary, passes it back to the kernel module. Queuing the packet to userspace as part of this process involves buffering. (Going the opposite direction does not, because the kernel actually processes the request synchronously.) A few other exceptional cases also queue packets to userspace for processing; most of these are due to OpenFlow actions that the fast path cannot handle and that must therefore be handled by the slow path instead.

OpenFlow also has a concept of packet buffering. When an OpenFlow switch sends a packet to a controller, it may opt to retain a copy of the packet in an OpenFlow "packet buffer". Later, if the controller wants to tell the switch to forward a copy of that packet, it can refer to the packet through its assigned buffer,

instead of sending the whole packet back to the switch, thereby saving bandwidth in the OpenFlow control channel. Before Open vSwitch 2.7, OVS implemented such buffering; Open vSwitch 2.7 and later do not.

8.5 General

Q: What is Open vSwitch?

A: Open vSwitch is a production quality open source software switch designed to be used as a vswitch in virtualized server environments. A vswitch forwards traffic between different VMs on the same physical host and also forwards traffic between VMs and the physical network. Open vSwitch supports standard management interfaces (e.g. sFlow, NetFlow, IPFIX, RSPAN, CLI), and is open to programmatic extension and control using OpenFlow and the OVSDB management protocol.

Open vSwitch as designed to be compatible with modern switching chipsets. This means that it can be ported to existing high-fanout switches allowing the same flexible control of the physical infrastructure as the virtual infrastructure. It also means that Open vSwitch will be able to take advantage of on-NIC switching chipsets as their functionality matures.

Q: What virtualization platforms can use Open vSwitch?

A: Open vSwitch can currently run on any Linux-based virtualization platform (kernel 3.10 and newer), including: KVM and VirtualBox. As of Linux 3.3 it is part of the mainline kernel. The bulk of the code is written in platform- independent C and is easily ported to other environments. We welcome inquires about integrating Open vSwitch with other virtualization platforms.

Q: How can I try Open vSwitch?

A: The Open vSwitch source code can be built on a Linux system. You can build and experiment with Open vSwitch on any Linux machine. Packages for various Linux distributions are available on many platforms, including: Debian, Ubuntu, Fedora.

You may also download and run a virtualization platform that already has Open vSwitch integrated. Be aware that the version integrated with a particular platform may not be the most recent Open vSwitch release.

Q: Does Open vSwitch only work on Linux?

A: No, Open vSwitch has been ported to a number of different operating systems and hardware platforms. Most of the development work occurs on Linux, but the code should be portable to any POSIX system. We've seen Open vSwitch ported to a number of different platforms, including FreeBSD, Windows, and even non-POSIX embedded systems.

By definition, the Open vSwitch Linux kernel module only works on Linux and will provide the highest performance. However, a userspace datapath is available that should be very portable.

Q: What's involved with porting Open vSwitch to a new platform or switching ASIC?

A: *Porting Open vSwitch to New Software or Hardware* describes how one would go about porting Open vSwitch to a new operating system or hardware platform.

Q: Why would I use Open vSwitch instead of the Linux bridge?

A: Open vSwitch is specially designed to make it easier to manage VM network configuration and monitor state spread across many physical hosts in dynamic virtualized environments. Refer to *Why Open vSwitch?* for a more detailed description of how Open vSwitch relates to the Linux Bridge.

Q: How is Open vSwitch related to distributed virtual switches like the VMware vNetwork distributed switch or the Cisco Nexus 1000V?

A: Distributed vswitch applications (e.g., VMware vNetwork distributed switch, Cisco Nexus 1000V) provide a centralized way to configure and monitor the network state of VMs that are spread across many

physical hosts. Open vSwitch is not a distributed vswitch itself, rather it runs on each physical host and supports remote management in a way that makes it easier for developers of virtualization/cloud management platforms to offer distributed vswitch capabilities.

To aid in distribution, Open vSwitch provides two open protocols that are specially designed for remote management in virtualized network environments: OpenFlow, which exposes flow-based forwarding state, and the OVSDB management protocol, which exposes switch port state. In addition to the switch implementation itself, Open vSwitch includes tools (ovs-ofctl, ovs-vsctl) that developers can script and extend to provide distributed vswitch capabilities that are closely integrated with their virtualization management platform.

Q: Why doesn't Open vSwitch support distribution?

A: Open vSwitch is intended to be a useful component for building flexible network infrastructure. There are many different approaches to distribution which balance trade-offs between simplicity, scalability, hardware compatibility, convergence times, logical forwarding model, etc. The goal of Open vSwitch is to be able to support all as a primitive building block rather than choose a particular point in the distributed design space.

Q: How can I contribute to the Open vSwitch Community?

A: You can start by joining the mailing lists and helping to answer questions. You can also suggest improvements to documentation. If you have a feature or bug you would like to work on, send a mail to one of the *mailing lists*.

Q: Why can I no longer connect to my OpenFlow controller or OVSDB manager?

A: Starting in OVS 2.4, we switched the default ports to the IANA-specified port numbers for OpenFlow (6633->6653) and OVSDB (6632->6640). We recommend using these port numbers, but if you cannot, all the programs allow overriding the default port. See the appropriate man page.

8.6 Common Configuration Issues

Q: I created a bridge and added my Ethernet port to it, using commands like these:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 eth0
```

and as soon as I ran the "add-port" command I lost all connectivity through eth0. Help!

A: A physical Ethernet device that is part of an Open vSwitch bridge should not have an IP address. If one does, then that IP address will not be fully functional.

You can restore functionality by moving the IP address to an Open vSwitch "internal" device, such as the network device named after the bridge itself. For example, assuming that eth0's IP address is 192.168.128.5, you could run the commands below to fix up the situation:

```
$ ip addr flush dev eth0
$ ip addr add 192.168.128.5/24 dev br0
$ ip link set br0 up
```

(If your only connection to the machine running OVS is through the IP address in question, then you would want to run all of these commands on a single command line, or put them into a script.) If there were any additional routes assigned to eth0, then you would also want to use commands to adjust these routes to go through br0.

If you use DHCP to obtain an IP address, then you should kill the DHCP client that was listening on the physical Ethernet interface (e.g. eth0) and start one listening on the internal interface (e.g. br0). You might still need to manually clear the IP address from the physical interface (e.g. with "ip addr flush dev eth0").

There is no compelling reason why Open vSwitch must work this way. However, this is the way that the Linux kernel bridge module has always worked, so it's a model that those accustomed to Linux bridging are already used to. Also, the model that most people expect is not implementable without kernel changes on all the versions of Linux that Open vSwitch supports.

By the way, this issue is not specific to physical Ethernet devices. It applies to all network devices except Open vSwitch "internal" devices.

Q: I created a bridge and added a couple of Ethernet ports to it, using commands like these:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 eth0
$ ovs-vsctl add-port br0 eth1
```

and now my network seems to have melted: connectivity is unreliable (even connectivity that doesn't go through Open vSwitch), all the LEDs on my physical switches are blinking, wireshark shows duplicated packets, and CPU usage is very high.

A: More than likely, you've looped your network. Probably, eth0 and eth1 are connected to the same physical Ethernet switch. This yields a scenario where OVS receives a broadcast packet on eth0 and sends it out on eth1, then the physical switch connected to eth1 sends the packet back on eth0, and so on forever. More complicated scenarios, involving a loop through multiple switches, are possible too.

The solution depends on what you are trying to do:

- If you added eth0 and eth1 to get higher bandwidth or higher reliability between OVS and your physical Ethernet switch, use a bond. The following commands create br0 and then add eth0 and eth1 as a bond:

  ```
  $ ovs-vsctl add-br br0
  $ ovs-vsctl add-bond br0 bond0 eth0 eth1
  ```

  Bonds have tons of configuration options. Please read the documentation on the Port table in ovs-vswitchd.conf.db(5) for all the details.

  Configuration for DPDK-enabled interfaces is slightly less straightforward. Refer to *Open vSwitch with DPDK* for more information.

- Perhaps you don't actually need eth0 and eth1 to be on the same bridge. For example, if you simply want to be able to connect each of them to virtual machines, then you can put each of them on a bridge of its own:

  ```
  $ ovs-vsctl add-br br0
  $ ovs-vsctl add-port br0 eth0

  $ ovs-vsctl add-br br1
  $ ovs-vsctl add-port br1 eth1
  ```

  and then connect VMs to br0 and br1. (A potential disadvantage is that traffic cannot directly pass between br0 and br1. Instead, it will go out eth0 and come back in eth1, or vice versa.)

- If you have a redundant or complex network topology and you want to prevent loops, turn on spanning tree protocol (STP). The following commands create br0, enable STP, and add eth0 and eth1 to the bridge. The order is important because you don't want have to have a loop in your network even transiently:

  ```
  $ ovs-vsctl add-br br0
  $ ovs-vsctl set bridge br0 stp_enable=true
  ```

(continues on next page)

(continued from previous page)

```
$ ovs-vsctl add-port br0 eth0
$ ovs-vsctl add-port br0 eth1
```

The Open vSwitch implementation of STP is not well tested. Report any bugs you observe, but if you'd rather avoid acting as a beta tester then another option might be your best shot.

Q: I can't seem to use Open vSwitch in a wireless network.

A: Wireless base stations generally only allow packets with the source MAC address of NIC that completed the initial handshake. Therefore, without MAC rewriting, only a single device can communicate over a single wireless link.

This isn't specific to Open vSwitch, it's enforced by the access point, so the same problems will show up with the Linux bridge or any other way to do bridging.

Q: I can't seem to add my PPP interface to an Open vSwitch bridge.

A: PPP most commonly carries IP packets, but Open vSwitch works only with Ethernet frames. The correct way to interface PPP to an Ethernet network is usually to use routing instead of switching.

Q: Is there any documentation on the database tables and fields?

A: Yes. ovs-vswitchd.conf.db(5) is a comprehensive reference.

Q: When I run ovs-dpctl I no longer see the bridges I created. Instead, I only see a datapath called "ovs-system". How can I see datapath information about a particular bridge?

A: In version 1.9.0, OVS switched to using a single datapath that is shared by all bridges of that type. The `ovs-appctl dpif/*` commands provide similar functionality that is scoped by the bridge.

Q: I created a GRE port using ovs-vsctl so why can't I send traffic or see the port in the datapath?

A: On Linux kernels before 3.11, the OVS GRE module and Linux GRE module cannot be loaded at the same time. It is likely that on your system the Linux GRE module is already loaded and blocking OVS (to confirm, check dmesg for errors regarding GRE registration). To fix this, unload all GRE modules that appear in lsmod as well as the OVS kernel module. You can then reload the OVS module following the directions in *Open vSwitch on Linux, FreeBSD and NetBSD* , which will ensure that dependencies are satisfied.

Q: Open vSwitch does not seem to obey my packet filter rules.

A: It depends on mechanisms and configurations you want to use.

You cannot usefully use typical packet filters, like iptables, on physical Ethernet ports that you add to an Open vSwitch bridge. This is because Open vSwitch captures packets from the interface at a layer lower below where typical packet-filter implementations install their hooks. (This actually applies to any interface of type "system" that you might add to an Open vSwitch bridge.)

You can usefully use typical packet filters on Open vSwitch internal ports as they are mostly ordinary interfaces from the point of view of packet filters.

For example, suppose you create a bridge br0 and add Ethernet port eth0 to it. Then you can usefully add iptables rules to affect the internal interface br0, but not the physical interface eth0. (br0 is also where you would add an IP address, as discussed elsewhere in the FAQ.)

For simple filtering rules, it might be possible to achieve similar results by installing appropriate OpenFlow flows instead. The OVS conntrack feature (see the "ct" action in ovs-actions(7)) can implement a stateful firewall.

If the use of a particular packet filter setup is essential, Open vSwitch might not be the best choice for you. On Linux, you might want to consider using the Linux Bridge. (This is the only choice if you want to use ebtables rules.) On NetBSD, you might want to consider using the bridge(4) with BRIDGE_IPF option.

Q: It seems that Open vSwitch does nothing when I removed a port and then immediately put it back. For example, consider that p1 is a port of `type=internal`:

```
$ ovs-vsctl del-port br0 p1 -- \
    add-port br0 p1 -- \
    set interface p1 type=internal
```

Any other type of port gets the same effect.

A: It's an expected behaviour.

If del-port and add-port happen in a single OVSDB transaction as your example, Open vSwitch always "skips" the intermediate steps. Even if they are done in multiple transactions, it's still allowed for Open vSwitch to skip the intermediate steps and just implement the overall effect. In both cases, your example would be turned into a no-op.

If you want to make Open vSwitch actually destroy and then re-create the port for some side effects like re-setting kernel setting for the corresponding interface, you need to separate operations into multiple OVSDB transactions and ensure that at least the first one does not have `--no-wait`. In the following example, the first ovs-vsctl will block until Open vSwitch reloads the new configuration and removes the port:

```
$ ovs-vsctl del-port br0 p1
$ ovs-vsctl add-port br0 p1 -- \
    set interface p1 type=internal
```

Q: I want to add thousands of ports to an Open vSwitch bridge, but it takes too long (minutes or hours) to do it with ovs-vsctl. How can I do it faster?

A: If you add them one at a time with ovs-vsctl, it can take a long time to add thousands of ports to an Open vSwitch bridge. This is because every invocation of ovs-vsctl first reads the current configuration from OVSDB. As the number of ports grows, this starts to take an appreciable amount of time, and when it is repeated thousands of times the total time becomes significant.

The solution is to add the ports in one invocation of ovs-vsctl (or a small number of them). For example, using bash:

```
$ ovs-vsctl add-br br0
$ cmds=; for i in {1..5000}; do cmds+=" -- add-port br0 p$i"; done
$ ovs-vsctl $cmds
```

takes seconds, not minutes or hours, in the OVS sandbox environment.

Q: I created a bridge named br0. My bridge shows up in "ovs-vsctl show", but "ovs-ofctl show br0" just prints "br0 is not a bridge or a socket".

A: Open vSwitch wasn't able to create the bridge. Check the ovs-vswitchd log for details (Debian and Red Hat packaging for Open vSwitch put it in /var/log/openvswitch/ovs-vswitchd.log).

In general, the Open vSwitch database reflects the desired configuration state. ovs-vswitchd monitors the database and, when it changes, reconfigures the system to reflect the new desired state. This normally happens very quickly. Thus, a discrepancy between the database and the actual state indicates that ovs-vswitchd could not implement the configuration, and so one should check the log to find out why. (Another possible cause is that ovs-vswitchd is not running. This will make ovs-vsctl commands hang, if they change the configuration, unless one specifies `--no-wait`.)

Q: I have a bridge br0. I added a new port vif1.0, and it shows up in "ovs-vsctl show", but "ovs-vsctl list port" says that it has OpenFlow port ("ofport") -1, and "ovs-ofctl show br0" doesn't show vif1.0 at all.

A: Open vSwitch wasn't able to create the port. Check the ovs-vswitchd log for details (Debian and Red Hat packaging for Open vSwitch put it in /var/log/openvswitch/ovs-vswitchd.log). Please see the previous question for more information.

You may want to upgrade to Open vSwitch 2.3 (or later), in which ovs-vsctl will immediately report when there is an issue creating a port.

Q: I created a tap device tap0, configured an IP address on it, and added it to a bridge, like this:

```
$ tunctl -t tap0
$ ip addr add 192.168.0.123/24 dev tap0
$ ip link set tap0 up
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 tap0
```

I expected that I could then use this IP address to contact other hosts on the network, but it doesn't work. Why not?

A: The short answer is that this is a misuse of a "tap" device. Use an "internal" device implemented by Open vSwitch, which works differently and is designed for this use. To solve this problem with an internal device, instead run:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 int0 -- set Interface int0 type=internal
$ ip addr add 192.168.0.123/24 dev int0
$ ip link set int0 up
```

Even more simply, you can take advantage of the internal port that every bridge has under the name of the bridge:

```
$ ovs-vsctl add-br br0
$ ip addr add 192.168.0.123/24 dev br0
$ ip link set br0 up
```

In more detail, a "tap" device is an interface between the Linux (or BSD) network stack and a user program that opens it as a socket. When the "tap" device transmits a packet, it appears in the socket opened by the userspace program. Conversely, when the userspace program writes to the "tap" socket, the kernel TCP/IP stack processes the packet as if it had been received by the "tap" device.

Consider the configuration above. Given this configuration, if you "ping" an IP address in the 192.168.0.x subnet, the Linux kernel routing stack will transmit an ARP on the tap0 device. Open vSwitch userspace treats "tap" devices just like any other network device; that is, it doesn't open them as "tap" sockets. That means that the ARP packet will simply get dropped.

You might wonder why the Open vSwitch kernel module doesn't intercept the ARP packet and bridge it. After all, Open vSwitch intercepts packets on other devices. The answer is that Open vSwitch only intercepts *received* packets, but this is a packet being transmitted. The same thing happens for all other types of network devices, except for Open vSwitch "internal" ports. If you, for example, add a physical Ethernet port to an OVS bridge, configure an IP address on a physical Ethernet port, and then issue a "ping" to an address in that subnet, the same thing happens: an ARP gets transmitted on the physical Ethernet port and Open vSwitch never sees it. (You should not do that, as documented at the beginning of this section.)

It can make sense to add a "tap" device to an Open vSwitch bridge, if some userspace program (other than Open vSwitch) has opened the tap socket. This is the case, for example, if the "tap" device was created by KVM (or QEMU) to simulate a virtual NIC. In such a case, when OVS bridges a packet to the "tap" device, the kernel forwards that packet to KVM in userspace, which passes it along to the VM, and in the other direction, when the VM sends a packet, KVM writes it to the "tap" socket, which causes OVS to receive it and bridge it to the other OVS ports. Please note that in such a case no IP address is configured

on the "tap" device (there is normally an IP address configured in the virtual NIC inside the VM, but this is not visible to the host Linux kernel or to Open vSwitch).

There is one special case in which Open vSwitch does directly read and write "tap" sockets. This is an implementation detail of the Open vSwitch userspace switch, which implements its "internal" ports as Linux (or BSD) "tap" sockets. In such a userspace switch, OVS receives packets sent on the "tap" device used to implement an "internal" port by reading the associated "tap" socket, and bridges them to the rest of the switch. In the other direction, OVS transmits packets bridged to the "internal" port by writing them to the "tap" socket, causing them to be processed by the kernel TCP/IP stack as if they had been received on the "tap" device. Users should not need to be concerned with this implementation detail.

Open vSwitch has a network device type called "tap". This is intended only for implementing "internal" ports in the OVS userspace switch and should not be used otherwise. In particular, users should not configure KVM "tap" devices as type "tap" (use type "system", the default, instead).

Q: I observe packet loss at the beginning of RFC2544 tests on a server running few hundred container apps bridged to OVS with traffic generated by HW traffic generator. How can I fix this?

A: This is expected behavior on virtual switches. RFC2544 tests were designed for hardware switches, which don't have caches on the fastpath that need to be heated. Traffic generators in order to prime the switch use learning phase to heat the caches before sending the actual traffic in test phase. In case of OVS the cache is flushed quickly and to accommodate the traffic generator's delay between learning and test phase, the max-idle timeout settings should be changed to 50000 ms.:

```
$ ovs-vsctl --no-wait set Open_vSwitch . other_config:max-idle=50000
```

Q: How can I configure the bridge internal interface MTU? Why does Open vSwitch keep changing internal ports MTU?

A: By default Open vSwitch overrides the internal interfaces (e.g. br0) MTU. If you have just an internal interface (e.g. br0) and a physical interface (e.g. eth0), then every change in MTU to eth0 will be reflected to br0. Any manual MTU configuration using *ip* on internal interfaces is going to be overridden by Open vSwitch to match the current bridge minimum.

Sometimes this behavior is not desirable, for example with tunnels. The MTU of an internal interface can be explicitly set using the following command:

```
$ ovs-vsctl set int br0 mtu_request=1450
```

After this, Open vSwitch will configure br0 MTU to 1450. Since this setting is in the database it will be persistent (compared to what happens with *ip*).

The MTU configuration can be removed to restore the default behavior with:

```
$ ovs-vsctl set int br0 mtu_request=[]
```

The `mtu_request` column can be used to configure MTU even for physical interfaces (e.g. eth0).

Q: I just upgraded and I see a performance drop. Why?

A: The OVS kernel datapath may have been updated to a new version without updating OVS userspace components. Sometimes new versions of OVS kernel module add functionality that is backwards compatible with older userspace components but may cause a drop in performance with them. Especially, if a kernel module from OVS 2.1 or newer is paired with OVS userspace 1.10 or older, there will be a performance drop for TCP traffic.

Updating the OVS userspace components to the latest released version should fix the performance degradation.

The OVS kernel modules has been distributed with the upstream Linux kernel since Linux 3.3. As of OVS 2.15 the use of the kernel module distributed with OVS is deprecated in favour of the kernel module distributed with the upstream Linux kernel. And as of OVS 3.0 the kernel module is no longer part of the OVS distribution.

Accordingly, for OVS 2.15 and newer, to get the best possible performance and functionality it is recommended to pair the latest released version of OVS with the latest released version of the Linux kernel. While, for releases of OVS prior to 2.15, is recommended to pair the same versions of the kernel module and OVS userspace.

Q: I can't unload the openvswitch kernel module. Why?

A: The userspace might still hold the reference count. So `rmmod openvswitch` does not work, for example:

```
$ lsmod | grep openvswitch
openvswitch       155648 4
nf_conncount      24576  1 openvswitch
```

Use the command below to drop the refcnt:

```
$ ovs-dpctl del-dp system@ovs-system
$ rmmod openvswitch
```

8.7 Using OpenFlow

Q: What versions of OpenFlow does Open vSwitch support?

A: The following table lists the versions of OpenFlow supported by each version of Open vSwitch:

Open vSwitch	OF1.0	OF1.1	OF1.2	OF1.3	OF1.4	OF1.5
1.9 and earlier	yes	—	—	—	—	—
1.10, 1.11	yes	—	(*)	(*)	—	—
2.0, 2.1	yes	(*)	(*)	(*)	—	—
2.2	yes	(*)	(*)	(*)	(%)	(*)
2.3, 2.4	yes	yes	yes	yes	(*)	(*)
2.5, 2.6, 2.7	yes	yes	yes	yes	(*)	(*)
2.8, 2.9, 2.10, 2.11	yes	yes	yes	yes	yes	(*)
2.12	yes	yes	yes	yes	yes	yes

—Not supported. yes Supported and enabled by default (*) Supported, but missing features, and must be enabled by user. (%) Experimental, unsafe implementation.

In any case, the user may override the default:

- To enable OpenFlow 1.0, 1.1, 1.2, and 1.3 on bridge br0:

```
$ ovs-vsctl set bridge br0 \
    protocols=OpenFlow10,OpenFlow11,OpenFlow12,OpenFlow13
```

- To enable OpenFlow 1.0, 1.1, 1.2, 1.3, 1.4, and 1.5 on bridge br0:

```
$ ovs-vsctl set bridge br0 \
    protocols=OpenFlow10,OpenFlow11,OpenFlow12,OpenFlow13,OpenFlow14,
↪OpenFlow15
```

- To enable only OpenFlow 1.0 on bridge br0:

```
$ ovs-vsctl set bridge br0 protocols=OpenFlow10
```

All current versions of ovs-ofctl enable only OpenFlow 1.0 by default. Use the -O option to enable support for later versions of OpenFlow in ovs-ofctl. For example:

```
$ ovs-ofctl -O OpenFlow13 dump-flows br0
```

(Open vSwitch 2.2 had an experimental implementation of OpenFlow 1.4 that could cause crashes. We don't recommend enabling it.)

OpenFlow Support in Open vSwitch tracks support for OpenFlow 1.1 and later features.

Q: Does Open vSwitch support MPLS?

A: Before version 1.11, Open vSwitch did not support MPLS. That is, these versions can match on MPLS Ethernet types, but they cannot match, push, or pop MPLS labels, nor can they look past MPLS labels into the encapsulated packet.

Open vSwitch versions 1.11, 2.0, and 2.1 have very minimal support for MPLS. With the userspace datapath only, these versions can match, push, or pop a single MPLS label, but they still cannot look past MPLS labels (even after popping them) into the encapsulated packet. Kernel datapath support is unchanged from earlier versions.

Open vSwitch version 2.3 can match, push, or pop a single MPLS label and look past the MPLS label into the encapsulated packet. Both userspace and kernel datapaths will be supported, but MPLS processing always happens in userspace either way, so kernel datapath performance will be disappointing.

Open vSwitch version 2.4 can match, push, or pop up to 3 MPLS labels and look past the MPLS label into the encapsulated packet. It will have kernel support for MPLS, yielding improved performance.

Q: I'm getting "error type 45250 code 0". What's that?

A: This is a Open vSwitch extension to OpenFlow error codes. Open vSwitch uses this extension when it must report an error to an OpenFlow controller but no standard OpenFlow error code is suitable.

Open vSwitch logs the errors that it sends to controllers, so the easiest thing to do is probably to look at the ovs-vswitchd log to find out what the error was.

If you want to dissect the extended error message yourself, the format is documented in include/openflow/nicira-ext.h in the Open vSwitch source distribution. The extended error codes are documented in include/openvswitch/ofp-errors.h.

Q: Some of the traffic that I'd expect my OpenFlow controller to see doesn't actually appear through the OpenFlow connection, even though I know that it's going through.

A: By default, Open vSwitch assumes that OpenFlow controllers are connected "in-band", that is, that the controllers are actually part of the network that is being controlled. In in-band mode, Open vSwitch sets up special "hidden" flows to make sure that traffic can make it back and forth between OVS and the controllers. These hidden flows are higher priority than any flows that can be set up through OpenFlow, and they are not visible through normal OpenFlow flow table dumps.

Usually, the hidden flows are desirable and helpful, but occasionally they can cause unexpected behavior. You can view the full OpenFlow flow table, including hidden flows, on bridge br0 with the command:

```
$ ovs-appctl bridge/dump-flows br0
```

to help you debug. The hidden flows are those with priorities greater than 65535 (the maximum priority that can be set with OpenFlow).

The `Documentation/topics/design` doc describes the in-band model in detail.

If your controllers are not actually in-band (e.g. they are on localhost via 127.0.0.1, or on a separate network), then you should configure your controllers in "out-of-band" mode. If you have one controller on bridge br0, then you can configure out-of-band mode on it with:

```
$ ovs-vsctl set controller br0 connection-mode=out-of-band
```

Q: Some of the OpenFlow flows that my controller sets up don't seem to apply to certain traffic, especially traffic between OVS and the controller itself.

A: See above.

Q: I configured all my controllers for out-of-band control mode but "ovs-appctl bridge/dump-flows" still shows some hidden flows.

A: You probably have a remote manager configured (e.g. with "ovs-vsctl set-manager"). By default, Open vSwitch assumes that managers need in-band rules set up on every bridge. You can disable these rules on bridge br0 with:

```
$ ovs-vsctl set bridge br0 other-config:disable-in-band=true
```

This actually disables in-band control entirely for the bridge, as if all the bridge's controllers were configured for out-of-band control.

Q: My OpenFlow controller doesn't see the VLANs that I expect.

A: See answer under "VLANs", above.

Q: I ran `ovs-ofctl add-flow br0 nw_dst=192.168.0.1,actions=drop` but I got a funny message like this:

```
ofp_util|INFO|normalization changed ofp_match, details:
ofp_util|INFO| pre: nw_dst=192.168.0.1
ofp_util|INFO|post:
```

and when I ran `ovs-ofctl dump-flows br0` I saw that my nw_dst match had disappeared, so that the flow ends up matching every packet.

A: The term "normalization" in the log message means that a flow cannot match on an L3 field without saying what L3 protocol is in use. The "ovs-ofctl" command above didn't specify an L3 protocol, so the L3 field match was dropped.

In this case, the L3 protocol could be IP or ARP. A correct command for each possibility is, respectively:

```
$ ovs-ofctl add-flow br0 ip,nw_dst=192.168.0.1,actions=drop
```

and:

```
$ ovs-ofctl add-flow br0 arp,nw_dst=192.168.0.1,actions=drop
```

Similarly, a flow cannot match on an L4 field without saying what L4 protocol is in use. For example, the flow match `tp_src=1234` is, by itself, meaningless and will be ignored. Instead, to match TCP source port 1234, write `tcp,tp_src=1234`, or to match UDP source port 1234, write `udp,tp_src=1234`.

Q: How can I figure out the OpenFlow port number for a given port?

A: The `OFPT_FEATURES_REQUEST` message requests an OpenFlow switch to respond with an `OFPT_FEATURES_REPLY` that, among other information, includes a mapping between OpenFlow port names and numbers. From a command prompt, `ovs-ofctl show br0` makes such a request and prints the response for switch br0.

The Interface table in the Open vSwitch database also maps OpenFlow port names to numbers. To print the OpenFlow port number associated with interface eth0, run:

```
$ ovs-vsctl get Interface eth0 ofport
```

You can print the entire mapping with:

```
$ ovs-vsctl -- --columns=name,ofport list Interface
```

but the output mixes together interfaces from all bridges in the database, so it may be confusing if more than one bridge exists.

In the Open vSwitch database, ofport value `-1` means that the interface could not be created due to an error. (The Open vSwitch log should indicate the reason.) ofport value `[]` (the empty set) means that the interface hasn't been created yet. The latter is normally an intermittent condition (unless ovs-vswitchd is not running).

Q: I added some flows with my controller or with ovs-ofctl, but when I run "ovs-dpctl dump-flows" I don't see them.

A: ovs-dpctl queries a kernel datapath, not an OpenFlow switch. It won't display the information that you want. You want to use `ovs-ofctl dump-flows` instead.

Q: It looks like each of the interfaces in my bonded port shows up as an individual OpenFlow port. Is that right?

A: Yes, Open vSwitch makes individual bond interfaces visible as OpenFlow ports, rather than the bond as a whole. The interfaces are treated together as a bond for only a few purposes:

- Sending a packet to the OFPP_NORMAL port. (When an OpenFlow controller is not configured, this happens implicitly to every packet.)
- Mirrors configured for output to a bonded port.

It would make a lot of sense for Open vSwitch to present a bond as a single OpenFlow port. If you want to contribute an implementation of such a feature, please bring it up on the Open vSwitch development mailing list at dev@openvswitch.org.

Q: I have a sophisticated network setup involving Open vSwitch, VMs or multiple hosts, and other components. The behavior isn't what I expect. Help!

A: To debug network behavior problems, trace the path of a packet, hop-by-hop, from its origin in one host to a remote host. If that's correct, then trace the path of the response packet back to the origin.

The open source tool called `plotnetcfg` can help to understand the relationship between the networking devices on a single host.

Usually a simple ICMP echo request and reply (`ping`) packet is good enough. Start by initiating an ongoing `ping` from the origin host to a remote host. If you are tracking down a connectivity problem, the "ping" will not display any successful output, but packets are still being sent. (In this case the packets being sent are likely ARP rather than ICMP.)

Tools available for tracing include the following:

- `tcpdump` and `wireshark` for observing hops across network devices, such as Open vSwitch internal devices and physical wires.
- `ovs-appctl dpif/dump-flows <br>` in Open vSwitch 1.10 and later or `ovs-dpctl dump-flows <br>` in earlier versions. These tools allow one to observe the actions being taken on packets in ongoing flows.

  See ovs-vswitchd(8) for `ovs-appctl dpif/dump-flows` documentation, ovs-dpctl(8) for `ovs-dpctl dump-flows` documentation, and "Why are there so many different ways to dump flows?" above for some background.
- `ovs-appctl ofproto/trace` to observe the logic behind how ovs-vswitchd treats packets. See ovs-vswitchd(8) for documentation. You can out more details about a given flow that `ovs-dpctl`

dump-flows displays, by cutting and pasting a flow from the output into an ovs-appctl ofproto/trace command.

- SPAN, RSPAN, and ERSPAN features of physical switches, to observe what goes on at these physical hops.

Starting at the origin of a given packet, observe the packet at each hop in turn. For example, in one plausible scenario, you might:

1. tcpdump the eth interface through which an ARP egresses a VM, from inside the VM.
2. tcpdump the vif or tap interface through which the ARP ingresses the host machine.
3. Use ovs-dpctl dump-flows to spot the ARP flow and observe the host interface through which the ARP egresses the physical machine. You may need to use ovs-dpctl show to interpret the port numbers. If the output seems surprising, you can use ovs-appctl ofproto/trace to observe details of how ovs-vswitchd determined the actions in the ovs-dpctl dump-flows output.
4. tcpdump the eth interface through which the ARP egresses the physical machine.
5. tcpdump the eth interface through which the ARP ingresses the physical machine, at the remote host that receives the ARP.
6. Use ovs-dpctl dump-flows to spot the ARP flow on the remote host remote host that receives the ARP and observe the VM vif or tap interface to which the flow is directed. Again, ovs-dpctl show and ovs-appctl ofproto/trace might help.
7. tcpdump the vif or tap interface to which the ARP is directed.
8. tcpdump the eth interface through which the ARP ingresses a VM, from inside the VM.

It is likely that during one of these steps you will figure out the problem. If not, then follow the ARP reply back to the origin, in reverse.

Q: How do I make a flow drop packets?

A: To drop a packet is to receive it without forwarding it. OpenFlow explicitly specifies forwarding actions. Thus, a flow with an empty set of actions does not forward packets anywhere, causing them to be dropped. You can specify an empty set of actions with actions= on the ovs-ofctl command line. For example:

```
$ ovs-ofctl add-flow br0 priority=65535,actions=
```

would cause every packet entering switch br0 to be dropped.

You can write "drop" explicitly if you like. The effect is the same. Thus, the following command also causes every packet entering switch br0 to be dropped:

```
$ ovs-ofctl add-flow br0 priority=65535,actions=drop
```

drop is not an action, either in OpenFlow or Open vSwitch. Rather, it is only a way to say that there are no actions.

Q: I added a flow to send packets out the ingress port, like this:

```
$ ovs-ofctl add-flow br0 in_port=2,actions=2
```

but OVS drops the packets instead.

A: Yes, OpenFlow requires a switch to ignore attempts to send a packet out its ingress port. The rationale is that dropping these packets makes it harder to loop the network. Sometimes this behavior can even be convenient, e.g. it is often the desired behavior in a flow that forwards a packet to several ports ("floods" the packet).

Sometimes one really needs to send a packet out its ingress port ("hairpin"). In this case, output to OFPP_IN_PORT, which in ovs-ofctl syntax is expressed as just in_port, e.g.:

```
$ ovs-ofctl add-flow br0 in_port=2,actions=in_port
```

This also works in some circumstances where the flow doesn't match on the input port. For example, if you know that your switch has five ports numbered 2 through 6, then the following will send every received packet out every port, even its ingress port:

```
$ ovs-ofctl add-flow br0 actions=2,3,4,5,6,in_port
```

or, equivalently:

```
$ ovs-ofctl add-flow br0 actions=all,in_port
```

Sometimes, in complicated flow tables with multiple levels of resubmit actions, a flow needs to output to a particular port that may or may not be the ingress port. It's difficult to take advantage of OFPP_IN_PORT in this situation. To help, Open vSwitch provides, as an OpenFlow extension, the ability to modify the in_port field. Whatever value is currently in the in_port field is the port to which outputs will be dropped, as well as the destination for OFPP_IN_PORT. This means that the following will reliably output to port 2 or to ports 2 through 6, respectively:

```
$ ovs-ofctl add-flow br0 in_port=2,actions=load:0->NXM_OF_IN_PORT[],2
$ ovs-ofctl add-flow br0 actions=load:0->NXM_OF_IN_PORT[],2,3,4,5,6
```

If the input port is important, then one may save and restore it on the stack:

```
$ ovs-ofctl add-flow br0 actions=push:NXM_OF_IN_PORT[],\
    load:0->NXM_OF_IN_PORT[],\
    2,3,4,5,6,\
    pop:NXM_OF_IN_PORT[]
```

Q: My bridge br0 has host 192.168.0.1 on port 1 and host 192.168.0.2 on port 2. I set up flows to forward only traffic destined to the other host and drop other traffic, like this:

```
priority=5,in_port=1,ip,nw_dst=192.168.0.2,actions=2
priority=5,in_port=2,ip,nw_dst=192.168.0.1,actions=1
priority=0,actions=drop
```

But it doesn't work–I don't get any connectivity when I do this. Why?

A: These flows drop the ARP packets that IP hosts use to establish IP connectivity over Ethernet. To solve the problem, add flows to allow ARP to pass between the hosts:

```
priority=5,in_port=1,arp,actions=2
priority=5,in_port=2,arp,actions=1
```

This issue can manifest other ways, too. The following flows that match on Ethernet addresses instead of IP addresses will also drop ARP packets, because ARP requests are broadcast instead of being directed to a specific host:

```
priority=5,in_port=1,dl_dst=54:00:00:00:00:02,actions=2
priority=5,in_port=2,dl_dst=54:00:00:00:00:01,actions=1
priority=0,actions=drop
```

The solution already described above will also work in this case. It may be better to add flows to allow all multicast and broadcast traffic:

```
priority=5,in_port=1,dl_dst=01:00:00:00:00:00/01:00:00:00:00:00,actions=2
priority=5,in_port=2,dl_dst=01:00:00:00:00:00/01:00:00:00:00:00,actions=1
```

Q: My bridge disconnects from my controller on add-port/del-port.

A: Reconfiguring your bridge can change your bridge's datapath-id because Open vSwitch generates datapath-id from the MAC address of one of its ports. In that case, Open vSwitch disconnects from controllers because there's no graceful way to notify controllers about the change of datapath-id.

To avoid the behaviour, you can configure datapath-id manually.:

```
$ ovs-vsctl set bridge br0 other-config:datapath-id=0123456789abcdef
```

Q: My controller complains that OVS is not buffering packets. What's going on?

A: "Packet buffering" is an optional OpenFlow feature, and controllers should detect how many "buffers" an OpenFlow switch implements. It was recently noticed that OVS implementation of the buffering feature was not compliant to OpenFlow specifications. Rather than fix it and risk controller incompatibility, the buffering feature is removed as of OVS 2.7. Controllers are already expected to work properly in cases where the switch can not buffer packets, but sends full packets in "packet-in" messages instead, so this change should not affect existing users. After the change OVS always sends the `buffer_id` as `0xffffffff` in "packet-in" messages and will send an error response if any other value of this field is included in a "packet-out" or a "flow mod" sent by a controller.

Packet buffers have limited usefulness in any case. Table-miss packet-in messages most commonly pass the first packet in a microflow to the OpenFlow controller, which then sets up an OpenFlow flow that handles remaining traffic in the microflow without further controller intervention. In such a case, the packet that initiates the microflow is in practice usually small (certainly for TCP), which means that the switch sends the entire packet to the controller and the buffer only saves a small number of bytes in the reverse direction.

Q: How does OVS divide flows among buckets in an OpenFlow "select" group?

A: In Open vSwitch 2.3 and earlier, Open vSwitch used the destination Ethernet address to choose a bucket in a select group.

Open vSwitch 2.4 and later by default hashes the source and destination Ethernet address, VLAN ID, Ethernet type, IPv4/v6 source and destination address and protocol, and for TCP and SCTP only, the source and destination ports. The hash is "symmetric", meaning that exchanging source and destination addresses does not change the bucket selection.

Select groups in Open vSwitch 2.4 and later can be configured to use a different hash function, using a Netronome extension to the OpenFlow 1.5+ group_mod message. For more information, see Documentation/group-selection-method-property.txt in the Open vSwitch source tree.

Q: An OpenFlow "select" group isn't dividing packets evenly among the buckets.

A: When a packet passes through a "select" group, Open vSwitch hashes a subset of the fields in the packet, then it maps the hash value to a bucket. This means that packets whose hashed fields are the same will always go to the same bucket[*]. More specifically, if you test with a single traffic flow, only one bucket will receive any traffic[**]. Furthermore, statistics and probability mean that testing with a small number of flows may still yield an uneven distribution.

[*] Unless its bucket has a watch port or group whose liveness changes during the test.

[**] Unless the hash includes fields that vary within a traffic flow, such as tcp_flags.

Q: I added a flow to accept packets on VLAN 123 and output them on VLAN 456, like so:

```
$ ovs-ofctl add-flow br0 dl_vlan=123,actions=output:1,mod_vlan_vid:456
```

but the packets are actually being output in VLAN 123. Why?

A: OpenFlow actions are executed in the order specified. Thus, the actions above first output the packet, then change its VLAN. Since the output occurs before changing the VLAN, the change in VLAN will have no visible effect.

To solve this and similar problems, order actions so that changes to headers happen before output, e.g.:

```
$ ovs-ofctl add-flow br0 dl_vlan=123,actions=mod_vlan_vid:456,output:1
```

See also the following question.

Q: I added a flow to a redirect packets for TCP port 80 to port 443, like so:

```
$ ovs-ofctl add-flow br0 tcp,tcp_dst=123,actions=mod_tp_dst:443
```

but the packets are getting dropped instead. Why?

A: This set of actions does change the TCP destination port to 443, but then it does nothing more. It doesn't, for example, say to continue to another flow table or to output the packet. Therefore, the packet is dropped.

To solve the problem, add an action that does something with the modified packet. For example:

```
$ ovs-ofctl add-flow br0 tcp,tcp_dst=123,actions=mod_tp_dst:443,normal
```

See also the preceding question.

Q: When using the "ct" action with FTP connections, it doesn't seem to matter if I set the "alg=ftp" parameter in the action. Is this required?

A: It is advisable to use this option. Some platforms may automatically detect and apply ALGs in the "ct" action regardless of the parameters you provide, however this is not consistent across all implementations. The ovs-ofctl(8) man pages contain further details in the description of the ALG parameter.

8.8 Quality of Service (QoS)

Q: Does OVS support Quality of Service (QoS)?

A: Yes. For traffic that egresses from a switch, OVS supports traffic shaping; for traffic that ingresses into a switch, OVS support policing. Policing is a simple form of quality-of-service that simply drops packets received in excess of the configured rate. Due to its simplicity, policing is usually less accurate and less effective than egress traffic shaping, which queues packets.

Keep in mind that ingress and egress are from the perspective of the switch. That means that egress shaping limits the rate at which traffic is allowed to transmit from a physical interface, but not the rate at which traffic will be received on a virtual machine's VIF. For ingress policing, the behavior is the opposite.

Q: How do I configure egress traffic shaping?

A: Suppose that you want to set up bridge br0 connected to physical Ethernet port eth0 (a 1 Gbps device) and virtual machine interfaces vif1.0 and vif2.0, and that you want to limit traffic from vif1.0 to eth0 to 10 Mbps and from vif2.0 to eth0 to 20 Mbps. Then, you could configure the bridge this way:

```
$ ovs-vsctl -- \
  add-br br0 -- \
```

(continues on next page)

(continued from previous page)

```
add-port br0 eth0 -- \
add-port br0 vif1.0 -- set interface vif1.0 ofport_request=5 -- \
add-port br0 vif2.0 -- set interface vif2.0 ofport_request=6 -- \
set port eth0 qos=@newqos -- \
--id=@newqos create qos type=linux-htb \
    other-config:max-rate=1000000000 \
    queues:123=@vif10queue \
    queues:234=@vif20queue -- \
--id=@vif10queue create queue other-config:max-rate=10000000 -- \
--id=@vif20queue create queue other-config:max-rate=20000000
```

At this point, bridge br0 is configured with the ports and eth0 is configured with the queues that you need for QoS, but nothing is actually directing packets from vif1.0 or vif2.0 to the queues that we have set up for them. That means that all of the packets to eth0 are going to the "default queue", which is not what we want.

We use OpenFlow to direct packets from vif1.0 and vif2.0 to the queues reserved for them:

```
$ ovs-ofctl add-flow br0 in_port=5,actions=set_queue:123,normal
$ ovs-ofctl add-flow br0 in_port=6,actions=set_queue:234,normal
```

Each of the above flows matches on the input port, sets up the appropriate queue (123 for vif1.0, 234 for vif2.0), and then executes the "normal" action, which performs the same switching that Open vSwitch would have done without any OpenFlow flows being present. (We know that vif1.0 and vif2.0 have OpenFlow port numbers 5 and 6, respectively, because we set their ofport_request columns above. If we had not done that, then we would have needed to find out their port numbers before setting up these flows.)

Now traffic going from vif1.0 or vif2.0 to eth0 should be rate-limited.

By the way, if you delete the bridge created by the above commands, with:

```
$ ovs-vsctl del-br br0
```

then that will leave one unreferenced QoS record and two unreferenced Queue records in the Open vSwich database. One way to clear them out, assuming you don't have other QoS or Queue records that you want to keep, is:

```
$ ovs-vsctl -- --all destroy QoS -- --all destroy Queue
```

If you do want to keep some QoS or Queue records, or the Open vSwitch you are using is older than version 1.8 (which added the `--all` option), then you will have to destroy QoS and Queue records individually.

Q: How do I configure ingress policing?

A: A policing policy can be configured on an interface to drop packets that arrive at a higher rate than the configured value. For example, the following commands will rate-limit traffic that vif1.0 may generate to 10Mbps:

```
$ ovs-vsctl set interface vif1.0 ingress_policing_rate=10000
$ ovs-vsctl set interface vif1.0 ingress_policing_burst=8000
```

Traffic policing can interact poorly with some network protocols and can have surprising results. The "Ingress Policing" section of ovs-vswitchd.conf.db(5) discusses the issues in greater detail.

Q: I configured Quality of Service (QoS) in my OpenFlow network by adding records to the QoS and Queue table, but the results aren't what I expect.

A: Did you install OpenFlow flows that use your queues? This is the primary way to tell Open vSwitch which queues you want to use. If you don't do this, then the default queue will be used, which will probably not have the effect you want.

Refer to the previous question for an example.

Q: I'd like to take advantage of some QoS feature that Open vSwitch doesn't yet support. How do I do that?

A: Open vSwitch does not implement QoS itself. Instead, it can configure some, but not all, of the QoS features built into the Linux kernel. If you need some QoS feature that OVS cannot configure itself, then the first step is to figure out whether Linux QoS supports that feature. If it does, then you can submit a patch to support Open vSwitch configuration for that feature, or you can use "tc" directly to configure the feature in Linux. (If Linux QoS doesn't support the feature you want, then first you have to add that support to Linux.)

Q: I configured QoS, correctly, but my measurements show that it isn't working as well as I expect.

A: With the Linux kernel, the Open vSwitch implementation of QoS has two aspects:

- Open vSwitch configures a subset of Linux kernel QoS features, according to what is in OVSDB. It is possible that this code has bugs. If you believe that this is so, then you can configure the Linux traffic control (QoS) stack directly with the "tc" program. If you get better results that way, you can send a detailed bug report to bugs@openvswitch.org.

  It is certain that Open vSwitch cannot configure every Linux kernel QoS feature. If you need some feature that OVS cannot configure, then you can also use "tc" directly (or add that feature to OVS).

- The Open vSwitch implementation of OpenFlow allows flows to be directed to particular queues. This is pretty simple and unlikely to have serious bugs at this point.

However, most problems with QoS on Linux are not bugs in Open vSwitch at all. They tend to be either configuration errors (please see the earlier questions in this section) or issues with the traffic control (QoS) stack in Linux. The Open vSwitch developers are not experts on Linux traffic control. We suggest that, if you believe you are encountering a problem with Linux traffic control, that you consult the tc manpages (e.g. tc(8), tc-htb(8), tc-hfsc(8)), web resources (e.g. http://lartc.org/), or mailing lists (e.g. http://vger.kernel.org/vger-lists.html#netdev).

Q: Does Open vSwitch support OpenFlow meters?

A: Open vSwitch 2.0 added OpenFlow protocol support for OpenFlow meters. Open vSwitch 2.7 implemented meters in the userspace datapath. Open vSwitch 2.10 implemented meters in the Linux kernel datapath.

8.9 Releases

Q: What does it mean for an Open vSwitch release to be LTS (long-term support)?

A: All official releases have been through a comprehensive testing process and are suitable for production use. Planned releases occur twice a year. If a significant bug is identified in an LTS release, we will provide an updated release that includes the fix. Releases that are not LTS may not be fixed and may just be supplanted by the next major release. The current LTS release is 3.3.x.

For more information on the Open vSwitch release process, refer to *Release Process*.

Q: What Linux kernel versions does each Open vSwitch release work with?

A: Open vSwitch userspace works with the kernel module shipped with Linux upstream 3.3 and later.

Building the Linux kernel module from the Open vSwitch source tree was deprecated starting with Open vSwitch 2.15. And the kernel module source code was completely removed from the Open vSwitch source tree in 3.0 release.

Q: Does Open vSwitch support running on Windows (Hyper-V)?

A: Support for the Windows datapath, a.k.a. Hyper-V, was deprecated starting with Open vSwitch 3.7, and the source code was completely removed from the Open vSwitch source tree in the next release.

Q: Are all features available with all datapaths?

A: Open vSwitch supports different datapaths on different platforms. Each datapath has a different feature set: the following tables try to summarize the status.

Supported datapaths:

Linux kernel
: The datapath implemented by the module shipped with upstream Linux kernel. Since features have been gradually introduced into the kernel, the table mentions the first Linux release whose OVS module supports the feature.

Userspace
: This datapath supports conventional system devices as well as DPDK and AF_XDP devices when support for those is built. This is the only datapath that works on NetBSD, FreeBSD and Mac OSX.

The following table lists the datapath supported features from an Open vSwitch user's perspective. The "Linux kernel" column lists the upstream Linux kernel version that introduced a given feature into its kernel module. The "Userspace" column lists the Open vSwitch release versions that introduced a given feature into the built-in userspace datapath.

Feature	Linux kernel	Userspace
Connection tracking	4.3	2.6
Connection tracking-IPv6	YES	YES
Conntrack Fragment Reass.	4.3	2.12
Conntrack IPv6 Fragment	4.3	2.12
Conntrack Timeout Policies	5.2	2.14
Conntrack Zone Limit	4.18	2.13
Conntrack NAT	4.6	2.8
Conntrack NAT6	4.6	2.8
Conntrack Helper Persist.	YES	3.3
Tunnel - GRE	3.11	2.4
Tunnel - VXLAN	3.12	2.4
Tunnel - Geneve	3.18	2.4
Tunnel - GRE-IPv6	4.18	2.6
Tunnel - VXLAN-IPv6	4.3	2.6
Tunnel - Geneve-IPv6	4.4	2.6
Tunnel - ERSPAN	4.18	2.10
Tunnel - ERSPAN-IPv6	4.18	2.10
Tunnel - GTP-U	NO	2.14
Tunnel - SRv6	NO	3.2
Tunnel - Bareudp	5.7	NO
QoS - Policing	YES	2.6
QoS - Shaping	YES	NO
sFlow	YES	1.0
IPFIX	3.10	1.11
Set action	YES	1.0
NIC Bonding	YES	1.0
Multiple VTEPs	YES	1.10
Meter action	4.15	2.7
check_pkt_len action	5.2	2.12

Do note, however:

- Userspace datapath support, in some cases, is dependent on the associated interface types. For example, DPDK interfaces support ingress and egress policing, but not shaping.

The following table lists features that do not *directly* impact an Open vSwitch user, e.g. because their absence can be hidden by the ofproto layer (usually this comes with a performance penalty).

Feature	Linux kernel	Userspace
SCTP flows	3.12	YES
MPLS	3.19	YES
UFID	4.0	YES
Megaflows	3.12	YES
Masked set action	4.0	YES
Recirculation	3.19	YES
TCP flags matching	3.13	YES
Validate flow actions	YES	N/A
Multiple datapaths	YES	YES

Q: What DPDK version does each Open vSwitch release work with?

A: The following table lists the DPDK version against which the given versions of Open vSwitch will successfully build.

Open vSwitch	DPDK
2.2.x	1.6
2.3.x	1.6
2.4.x	2.0
2.5.x	2.2
2.6.x	16.07.2
2.7.x	16.11.9
2.8.x	17.05.2
2.9.x	17.11.10
2.10.x	17.11.10
2.11.x	18.11.9
2.12.x	18.11.9
2.13.x	19.11.13
2.14.x	19.11.13
2.15.x	20.11.6
2.16.x	20.11.6
2.17.x	21.11.9
3.0.x	21.11.9
3.1.x	22.11.7
3.2.x	22.11.7
3.3.x	23.11.7
3.4.x	23.11.7
3.5.x	24.11.6
3.6.x	24.11.6
3.7.x	25.11.2

Q: Are all the DPDK releases that OVS versions work with maintained?

No. DPDK follows YY.MM.n (Year.Month.Number) versioning.

Typically, all DPDK releases get a stable YY.MM.1 update with bugfixes 3 months after the YY.MM.0 release. In some cases there may also be a YY.MM.2 release.

DPDK LTS releases start once a year at YY.11.0 and are maintained for two years, with YY.MM.n+1 releases around every 3 months.

The latest information about DPDK stable and LTS releases can be found at DPDK stable.

Q: What features are not available in the Open vSwitch kernel datapath that ships as part of the upstream Linux kernel?

A: Certain features require kernel support to function or to have reasonable performance. If the ovs-vswitchd log file indicates that a feature is not supported, consider upgrading to a newer upstream Linux release.

Q: Why do tunnels not work when using a Linux kernel module?

A: Support for tunnels was added to the upstream Linux kernel module after the rest of Open vSwitch. As a result, some kernels may contain support for Open vSwitch but not tunnels. The minimum kernel version that supports each tunnel protocol is:

Protocol	Linux Kernel
GRE	3.11
VXLAN	3.12
Geneve	3.18
ERSPAN	4.18

Q: Why are UDP tunnel checksums not computed for VXLAN or Geneve?

A: Generating outer UDP checksums requires kernel support that was not part of the initial implementation of these protocols. The kernel modules shipped with upstream Linux 4.0 and later support UDP checksums.

Q: What features are not available when using the userspace datapath?

A: Tunnel virtual ports are not supported, as described in the previous answer. It is also not possible to use queue-related actions. On Linux kernels before 2.6.39, maximum-sized VLAN packets may not be transmitted.

Q: Should userspace or kernel be upgraded first to minimize downtime?

A. In general, the Open vSwitch userspace should work with any kernel version from upstream Linux. However, when upgrading between two releases of Open vSwitch it is best to migrate userspace first to reduce the possibility of incompatibilities.

Q: What happened to the bridge compatibility feature?

A: Bridge compatibility was a feature of Open vSwitch 1.9 and earlier. When it was enabled, Open vSwitch imitated the interface of the Linux kernel "bridge" module. This allowed users to drop Open vSwitch into environments designed to use the Linux kernel bridge module without adapting the environment to use Open vSwitch.

Open vSwitch 1.10 and later do not support bridge compatibility. The feature was dropped because version 1.10 adopted a new internal architecture that made bridge compatibility difficult to maintain. Now that many environments use OVS directly, it would be rarely useful in any case.

To use bridge compatibility, install OVS 1.9 or earlier, including the accompanying kernel modules (both the main and bridge compatibility modules), following the instructions that come with the release. Be sure to start the ovs-brcompatd daemon.

8.10 Terminology

Q: I thought Open vSwitch was a virtual Ethernet switch, but the documentation keeps talking about bridges. What's a bridge?

A: In networking, the terms "bridge" and "switch" are synonyms. Open vSwitch implements an Ethernet switch, which means that it is also an Ethernet bridge.

Q: What's a VLAN?

A: See *VLANs*.

8.11 VLANs

Q: What's a VLAN?

A: At the simplest level, a VLAN (short for "virtual LAN") is a way to partition a single switch into multiple switches. Suppose, for example, that you have two groups of machines, group A and group B. You want the machines in group A to be able to talk to each other, and you want the machine in group B to be able to talk to each other, but you don't want the machines in group A to be able to talk to the machines in group B. You can do this with two switches, by plugging the machines in group A into one switch and the machines in group B into the other switch.

If you only have one switch, then you can use VLANs to do the same thing, by configuring the ports for machines in group A as VLAN "access ports" for one VLAN and the ports for group B as "access ports" for a different VLAN. The switch will only forward packets between ports that are assigned to the same VLAN, so this effectively subdivides your single switch into two independent switches, one for each group of machines.

So far we haven't said anything about VLAN headers. With access ports, like we've described so far, no VLAN header is present in the Ethernet frame. This means that the machines (or switches) connected to access ports need not be aware that VLANs are involved, just like in the case where we use two different physical switches.

Now suppose that you have a whole bunch of switches in your network, instead of just one, and that some machines in group A are connected directly to both switches 1 and 2. To allow these machines to talk to each other, you could add an access port for group A's VLAN to switch 1 and another to switch 2, and then connect an Ethernet cable between those ports. That works fine, but it doesn't scale well as the number of switches and the number of VLANs increases, because you use up a lot of valuable switch ports just connecting together your VLANs.

This is where VLAN headers come in. Instead of using one cable and two ports per VLAN to connect a pair of switches, we configure a port on each switch as a VLAN "trunk port". Packets sent and received on a trunk port carry a VLAN header that says what VLAN the packet belongs to, so that only two ports total are required to connect the switches, regardless of the number of VLANs in use. Normally, only switches (either physical or virtual) are connected to a trunk port, not individual hosts, because individual hosts don't expect to see a VLAN header in the traffic that they receive.

None of the above discussion says anything about particular VLAN numbers. This is because VLAN numbers are completely arbitrary. One must only ensure that a given VLAN is numbered consistently throughout a network and that different VLANs are given different numbers. (That said, VLAN 0 is usually synonymous with a packet that has no VLAN header, and VLAN 4095 is reserved.)

Q: VLANs don't work.

A: Do you have VLANs enabled on the physical switch that OVS is attached to? Make sure that the port is configured to trunk the VLAN or VLANs that you are using with OVS.

Q: Outgoing VLAN-tagged traffic goes through OVS to my physical switch and to its destination host, but OVS seems to drop incoming return traffic.

A: It's possible that you have the VLAN configured on your physical switch as the "native" VLAN. In this mode, the switch treats incoming packets either tagged with the native VLAN or untagged as part of the native VLAN. It may also send outgoing packets in the native VLAN without a VLAN tag.

If this is the case, you have two choices:

- Change the physical switch port configuration to tag packets it forwards to OVS with the native VLAN instead of forwarding them untagged.
- Change the OVS configuration for the physical port to a native VLAN mode. For example, the following sets up a bridge with port eth0 in "native-tagged" mode in VLAN 9:

```
$ ovs-vsctl add-br br0 $ ovs-vsctl add-port br0 eth0 tag=9
vlan_mode=native-tagged
```

In this situation, "native-untagged" mode will probably work equally well. Refer to the documentation for the Port table in ovs-vswitchd.conf.db(5) for more information.

Q: I added a pair of VMs on different VLANs, like this:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 eth0
$ ovs-vsctl add-port br0 tap0 tag=9
$ ovs-vsctl add-port br0 tap1 tag=10
```

but the VMs can't access each other, the external network, or the Internet.

A: It is to be expected that the VMs can't access each other. VLANs are a means to partition a network. When you configured tap0 and tap1 as access ports for different VLANs, you indicated that they should be isolated from each other.

As for the external network and the Internet, it seems likely that the machines you are trying to access are not on VLAN 9 (or 10) and that the Internet is not available on VLAN 9 (or 10).

Q: I added a pair of VMs on the same VLAN, like this:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 eth0
$ ovs-vsctl add-port br0 tap0 tag=9
$ ovs-vsctl add-port br0 tap1 tag=9
```

The VMs can access each other, but not the external network or the Internet.

A: It seems likely that the machines you are trying to access in the external network are not on VLAN 9 and that the Internet is not available on VLAN 9. Also, ensure VLAN 9 is set up as an allowed trunk VLAN on the upstream switch port to which eth0 is connected.

Q: Can I configure an IP address on a VLAN?

A: Yes. Use an "internal port" configured as an access port. For example, the following configures IP address 192.168.0.7 on VLAN 9. That is, OVS will forward packets from eth0 to 192.168.0.7 only if they have an 802.1Q header with VLAN 9. Conversely, traffic forwarded from 192.168.0.7 to eth0 will be tagged with an 802.1Q header with VLAN 9:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 eth0
```

(continues on next page)

(continued from previous page)

```
$ ovs-vsctl add-port br0 vlan9 tag=9 \
    -- set interface vlan9 type=internal
$ ip addr add 192.168.0.7/24 dev vlan9
$ ip link set vlan9 up
```

See also the following question.

Q: I configured one IP address on VLAN 0 and another on VLAN 9, like this:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 eth0
$ ip addr add 192.168.0.5/24 dev br0
$ ip link set br0 up
$ ovs-vsctl add-port br0 vlan9 tag=9 -- set interface vlan9 type=internal
$ ip addr add 192.168.0.9/24 dev vlan9
$ ip link set vlan9 up
```

but other hosts that are only on VLAN 0 can reach the IP address configured on VLAN 9. What's going on?

A: RFC 1122 section 3.3.4.2 "Multihoming Requirements" describes two approaches to IP address handling in Internet hosts:

- In the "Strong ES Model", where an ES is a host ("End System"), an IP address is primarily associated with a particular interface. The host discards packets that arrive on interface A if they are destined for an IP address that is configured on interface B. The host never sends packets from interface A using a source address configured on interface B.
- In the "Weak ES Model", an IP address is primarily associated with a host. The host accepts packets that arrive on any interface if they are destined for any of the host's IP addresses, even if the address is configured on some interface other than the one on which it arrived. The host does not restrict itself to sending packets from an IP address associated with the originating interface.

Linux uses the weak ES model. That means that when packets destined to the VLAN 9 IP address arrive on eth0 and are bridged to br0, the kernel IP stack accepts them there for the VLAN 9 IP address, even though they were not received on vlan9, the network device for vlan9.

To simulate the strong ES model on Linux, one may add iptables rule to filter packets based on source and destination address and adjust ARP configuration with sysctls.

BSD uses the strong ES model.

Q: My OpenFlow controller doesn't see the VLANs that I expect.

A: The configuration for VLANs in the Open vSwitch database (e.g. via ovs-vsctl) only affects traffic that goes through Open vSwitch's implementation of the OpenFlow "normal switching" action. By default, when Open vSwitch isn't connected to a controller and nothing has been manually configured in the flow table, all traffic goes through the "normal switching" action. But, if you set up OpenFlow flows on your own, through a controller or using ovs-ofctl or through other means, then you have to implement VLAN handling yourself.

You can use "normal switching" as a component of your OpenFlow actions, e.g. by putting "normal" into the lists of actions on ovs-ofctl or by outputting to OFPP_NORMAL from an OpenFlow controller. In situations where this is not suitable, you can implement VLAN handling yourself, e.g.:

- If a packet comes in on an access port, and the flow table needs to send it out on a trunk port, then the flow can add the appropriate VLAN tag with the "mod_vlan_vid" action.
- If a packet comes in on a trunk port, and the flow table needs to send it out on an access port, then the flow can strip the VLAN tag with the "strip_vlan" action.

Q: I configured ports on a bridge as access ports with different VLAN tags, like this:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl set-controller br0 tcp:192.168.0.10:6653
$ ovs-vsctl add-port br0 eth0
$ ovs-vsctl add-port br0 tap0 tag=9
$ ovs-vsctl add-port br0 tap1 tag=10
```

but the VMs running behind tap0 and tap1 can still communicate, that is, they are not isolated from each other even though they are on different VLANs.

A: Do you have a controller configured on br0 (as the commands above do)? If so, then this is a variant on the previous question, "My OpenFlow controller doesn't see the VLANs that I expect," and you can refer to the answer there for more information.

Q: How MAC learning works with VLANs?

A: Open vSwitch implements Independent VLAN Learning (IVL) for `OFPP_NORMAL` action, e.g. it logically has separate learning tables for each VLANs.

8.12 VXLANs

Q: What's a VXLAN?

A: VXLAN stands for Virtual eXtensible Local Area Network, and is a means to solve the scaling challenges of VLAN networks in a multi-tenant environment. VXLAN is an overlay network which transports an L2 network over an existing L3 network. For more information on VXLAN, please see RFC 7348.

Q: How much of the VXLAN protocol does Open vSwitch currently support?

A: Open vSwitch currently supports the framing format for packets on the wire. There is currently no support for the multicast aspects of VXLAN. To get around the lack of multicast support, it is possible to pre-provision MAC to IP address mappings either manually or from a controller.

Q: What destination UDP port does the VXLAN implementation in Open vSwitch use?

A: By default, Open vSwitch will use the assigned IANA port for VXLAN, which is 4789. However, it is possible to configure the destination UDP port manually on a per-VXLAN tunnel basis. An example of this configuration is provided below.:

```
$ ovs-vsctl add-br br0
$ ovs-vsctl add-port br0 vxlan1 -- set interface vxlan1 type=vxlan \
    options:remote_ip=192.168.1.2 options:key=flow options:dst_port=8472
```

INDEX

Symbols

C

F

O