Towards native OVN L3 fabric integration with BGP

Frode Nordahl
Martin Kalcok

OVN Engineering @ Canonical

Community effort

May 13, 2024 | OVN Community Meeting

Attendees:
Recording: https://youtu.be/k448ada9aFQ
Transcripts: https://drive.google.com/file/d/1VuF5l9wSR9z4rIH_LVF3PJBScj4Mb8JT
Agenda:

- Discussion: [Frode/Vladislav] explore tighter integration between OVN and BGP
  - Goals
    - Re-use existing BGP implementations as much as possible
      - BiRD, FRRouting, Holo Routing, others?
      - The running of a host level BGP daemon is up to the end user and outside the scope of this work.
    - It must be possible to construct a hardware offloaded data path.

Community effort

[ovs-dev] RFC OVN: fabric integration

Frode Nordahl fnordahl at ubuntu.com
Tue Jun 25 16:52:37 UTC 2024

- Next message (by thread): [ovs-dev] RFC OVN: fabric integration
- **Messages sorted by:** [date] [thread] [subject] [author]

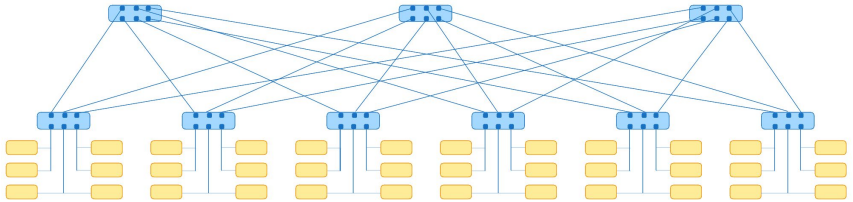
```
Hello,

We are increasingly seeing requests for integration between OVN
powered CMSs/workloads and the fabric.
```

⭐ 31 replies

9 participants ⭐

Motivation for native OVN BGP support

- Hardware acceleration.
- Build once and use in multiple projects.
- Enable the use of layer 3 only network designs [0].
- Resource location.
  - Route traffic directly to its intended destination.
  - Use the fabric as a load balancer — anycast.
- Avoid the many issues with large layer 2 broadcast domains.
- Avoid the configuration complexity of EVPN in fabric.



0: https://datatracker.ietf.org/doc/html/rfc7938

Requirements

- Data path must support **hardware acceleration**, i.e. the next hop address the peer resolves for announcements of OVN resources needs to be an OVN LRP IP.
- Minimize **address planning** and **configuration overhead** through the use of IPv6 LLAs for peering routing both IPv4 and IPv6 prefixes over a IPv6 BGP session [1][2] (aka. “BGP Unnumbered”).
- Support **ECMP out of the host**, i.e. use L3 interfaces potentially connecting to two different ToRs, instead of bonds, avoiding the additional complexity and potential vendor lock-in of multi-chassis bonds.
- Support **BGP authentication** [3][4], i.e. the source, destination address and ports in packet headers can not be changed.
- Implementation strategy that allow OVN to work with **multiple existing routing protocol suites** and **minimize duplication of effort**.

1: https://datatracker.ietf.org/doc/html/rfc5549
2: https://www.ietf.org/archive/id/draft-chroboczek-intarea-v4-via-v6-01.html
3: https://datatracker.ietf.org/doc/html/rfc2385
4: https://datatracker.ietf.org/doc/html/rfc5925

25.03 plans

- “BGP unnumbered” support.
  - Allow creation of LRP with no IPv4 address.
  - Allow LR to send RAs through localnet port.
  - Allow LR to send RAs even when there are no IPv6 prefixes other than a link-local address.
  - Routing IPv4 with IPv6 next hop (builds on Felix Huettner’s patches).
- OVS route-table library changes.
- OVN route exchange plugin framework.
  - For ovn-controller <-> routing protocol daemons IPC implementations.
  - In-tree.
- OVN route exchange Netlink VRF/NS(?) plugin.
  - Export routes to local NAT and Load Balancer VIP addresses attached to local gateway router.

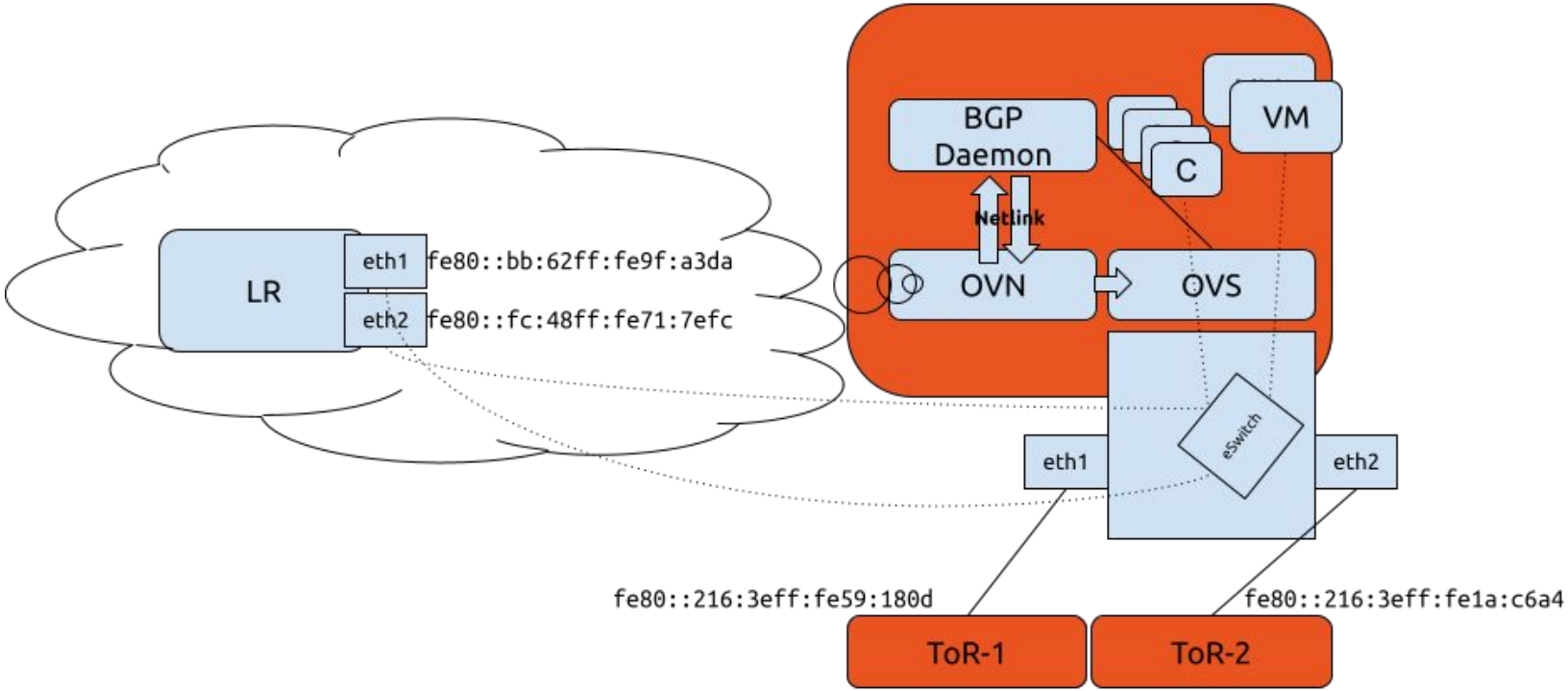
https://docs.google.com/document/d/1luzYUlzk0tur5OixJL5fTkP8nbhkf0PJrh5GS5rngB0

25.10 and beyond

- Export routes to local NAT and Load Balancer VIP addresses with local attachment through inter-LR routes?
  - LR with DGP connected to per chassis gateway router.
- Export other types of routes?
- Support export of non-local routes?

- Route learning.
  - LR learn default gateway from IPv6 Router Advertisements?
    - ECMP w/multiple LRPs.
  - OVN route exchange route learning from routing protocol daemon?

Overview of RFC implementation



https://mail.openvswitch.org/pipermail/ovs-dev/2024-August/416545.html
https://mail.openvswitch.org/pipermail/ovs-dev/2024-July/416039.html

BGP Control Plane Integration

Goals:

- Act like the BGP daemon listens on the Logical Router Port (and its IP)
  - Announced routes have automatically correct Next Hop address
  - BGP Authentication works out of the box

BGP Control Plane Integration

Goals:

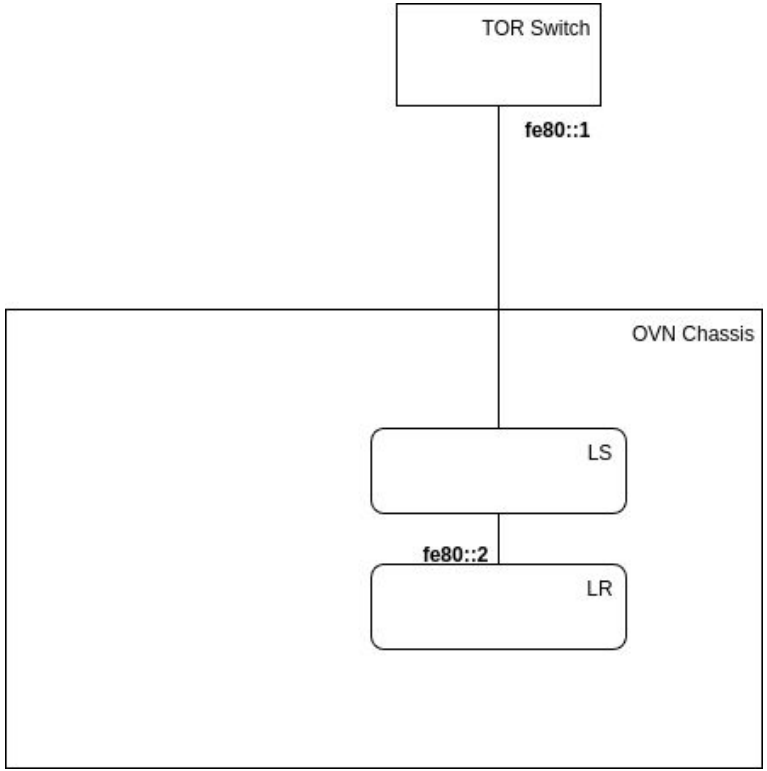
- Act like the BGP daemon listens on the Logical Router Port (and its IP)
  - Announced routes have automatically correct Next Hop address
  - BGP Authentication works out of the box
- Avoid implementing BGP daemon from scratch inside OVN

BGP Control Plane Integration

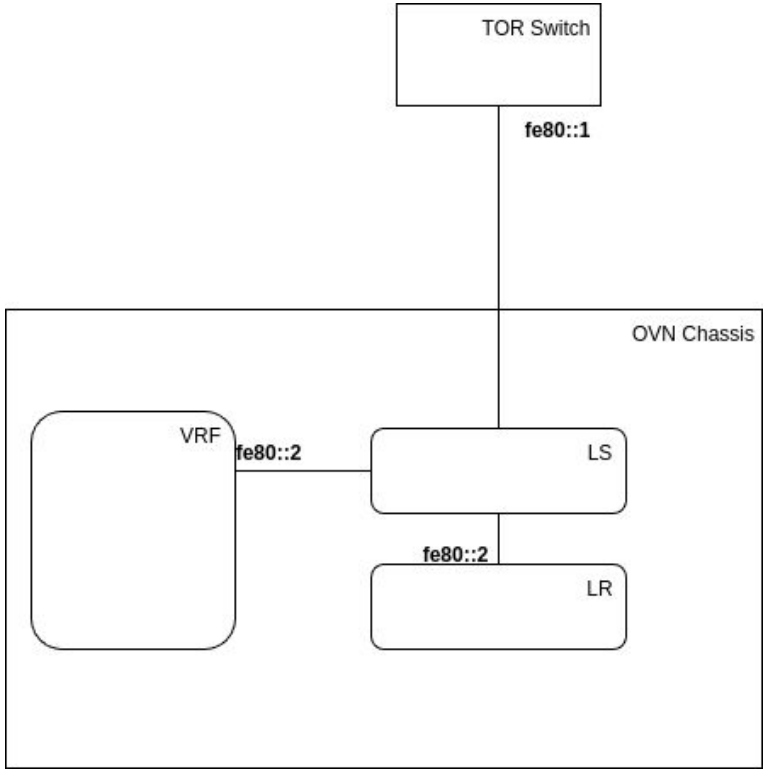
Goals:

- Act like the BGP daemon listens on the Logical Router Port (and its IP)
  - Announced routes have automatically correct Next Hop address
  - BGP Authentication works out of the box
- Avoid implementing BGP daemon from scratch inside OVN
- Don’t break ability to do Hardware Offloading

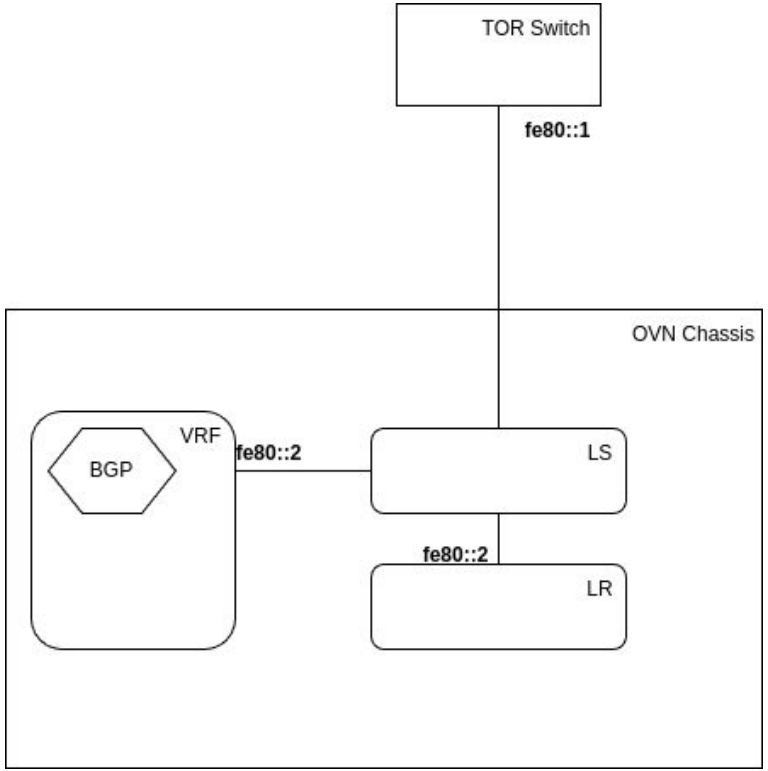
BGP protocol redirection



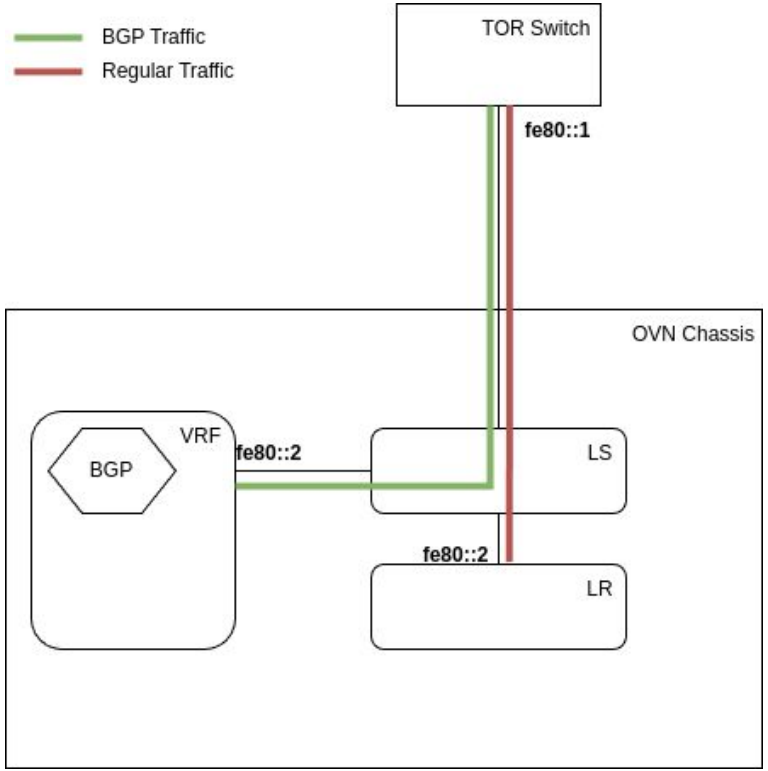
BGP protocol redirection



BGP protocol redirection

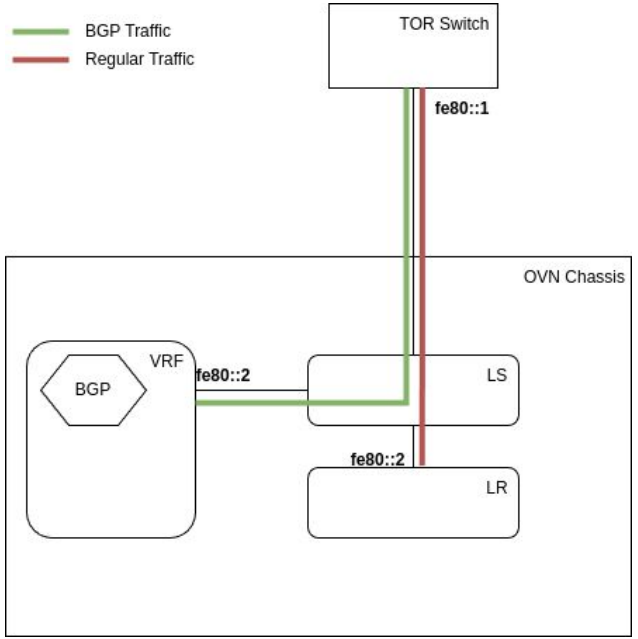


BGP protocol redirection



BGP Unnumbered - Hurdles Ahead

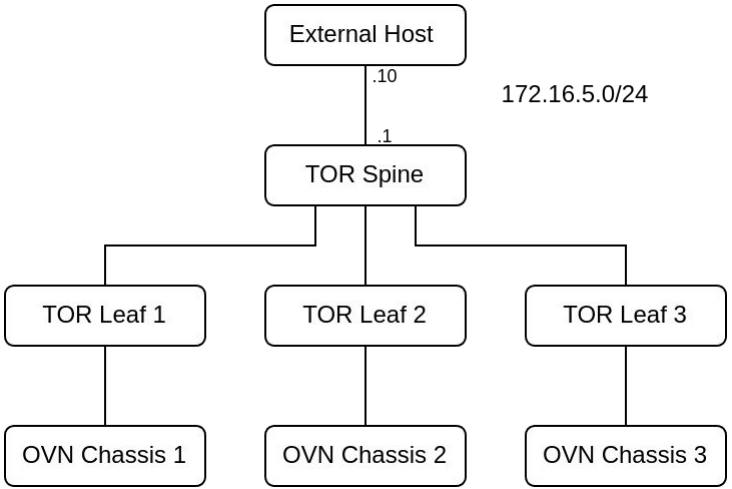
- Remove necessity to set explicit IP address on LRP
  - IPv4 LLA
  - No global IP at all
- IPv4 over IPv6 next hop
  - Promising series by Felix [0]
  - IPv6 encoded IPv4



[0] https://patchwork.ozlabs.org/project/ovn/list/?series=410497&state=%2A&archive=both

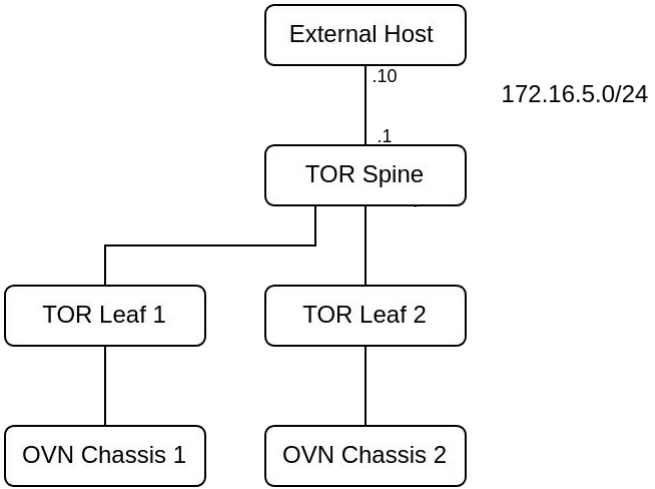
Demo time

- BGP Unnumbered
  - Minimal manual configuration
- Uses MicroOVN
  - Experimental branch
  - Configuration automation
  - OVN build with unreleased changes
- No CMS



Demo time

- BGP Unnumbered
  - Minimal manual configuration
- Uses MicroOVN
  - Experimental branch
  - Configuration automation
  - OVN build with unreleased changes
- No CMS
- TOR Leaf 3 & OVN Chassis 3 excluded



root@external-host: ~

```
root@external-host:~#
root@external-host:~#
root@external-host:~#
```

root@tor-spine: ~

```
root@tor-spine:~#
```

root@tor-leaf-1: ~

```
root@tor-leaf-1:~#
```

root@tor-leaf-2: ~

```
root@tor-leaf-2:~#
```

root@ovn-chassis-1: ~

```
root@ovn-chassis-1:~#
```

root@ovn-chassis-2: ~

```
root@ovn-chassis-2:~#
```