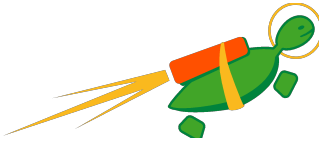


VRF Update

David Ahern

netconf 2016

Agenda

Status

What's Needed

Status

Basic support for IPv4 (v4.3) and IPv6 (v4.4)

- VRF device creates L3 domain
- Separate routing table for domain
- Interfaces enslaved to VRF device
- VRF device as VRF local loopback
- Cross-VRF routing

Option for VRF-global services with VRF-local connections (v4.5)

What's Needed

Usability Improvements

Missing Features

Hardware offload via switchdev

cumulus®

Usability Improvements

Missing Features

Hardware offload via switchdev

Usability - Simplifying VRF create / delete

cumulus®

FIB rules

- Rules required to direct lookups to proper table
- Option to add / remove rules in driver simplifies user overhead
- User sending rtnetlink message vs VRF driver sending message

cumulus

Loss of IPv6 addresses

- On enslave/release netdevice is cycled (down/up) if it is up
- Required to flush neighbor cache and routes for old VRF and move connected routes to new VRF
- Managed IPv6 addresses are lost

Usability - VRF management

iproute2 syntax is cumbersome

ip link add vrf-red type vrf table 123

ip {-6} rule add iif vrf-red table 123

ip {-6} rule add oif vrf-red table 123

ip link set dev vrf-red up

echo “123 vrf-red” > /etc/iproute2/rt_tables.d/vrf-red.conf

ip link set <dev> master vrf-red

ip route show table 123

ip route get oif vrf-red …

Usability - VRF management

vrf subcommand

- hides implementation details, provides more natural interface

  ip vrf add <vrf> table 123

  ip vrf <vrf> link add <dev>

  ip vrf <vrf> route show

  ip vrf <vrf> route get

  ip vrf <vrf> exec bash

Syntax simplification, re-using existing code

What's Needed

Usability Improvements

Missing Features

Hardware offload via switchdev

cumulus

Run task (and child tasks) in VRF context

- All AF_INET{6} sockets automatically bound to domain
- Inherit setting parent-child
- Run tasks as non-root and without NET_ADMIN
- Management VRF for example

cgroups fits the model

- RFC sent in January 2016

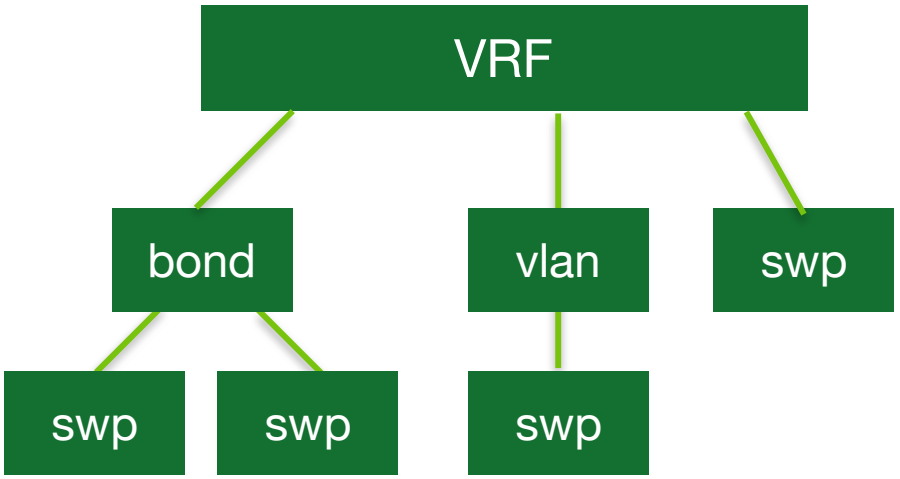
Netdevice index

{IP,IPV6}_PKTINFO interface

- rx_handler switches skb->dev to VRF device
- Need to save index of prior device; VRF easily derived
- Have a patch

link state protocols

- OSPF



cumulus

Rx rules on ingress device

- have a patch

Other hooks on Rx / Tx paths

cumulus

L3 Rx handler for an L3 master device

- Have a patch

Simplifies netfilter + netdev index

cumulus®

Allow socket binding to enslaved device

Multicast

- any changes required?

More testing with various setups / user needs

- requests from telecomm, networking companies

What's Needed

Usability Improvements

Missing Features

Hardware offload via switchdev

Hardware Offload via switchdev

switchdev disables L3 offload if any IP rules are installed, ever

Does not align with VRF devices

- FIB rules are required for it to work

Need to relax this overly cautious starting point

- “Simple” rules like those needed for VRFs
- Rules installed for non-hardware ports

cumulus

Options for FIB rules make it a challenge

- simple {i,o}if-to-table lookups
- table jumps
- fwmarks
- source/destination rules
- tos

Hardware Offload via switchdev

cumulus

Notifier over switchdev operation

- Lower layer device drivers register handlers
- All handlers are invoked for each rule add/delete

Allows driver to make decision if a rule is acceptable

- e.g., Simple VRF rules are ok

User flag to indicate no impact to offload

- e.g., DNS server rules to force lookups out mgmt interface

Unleashing the Power of Open Networking

Thank You!