EVPN/VXLAN TO THE HOST

Journal, notes and observations

About:Me

Spyridon Kakaroukas

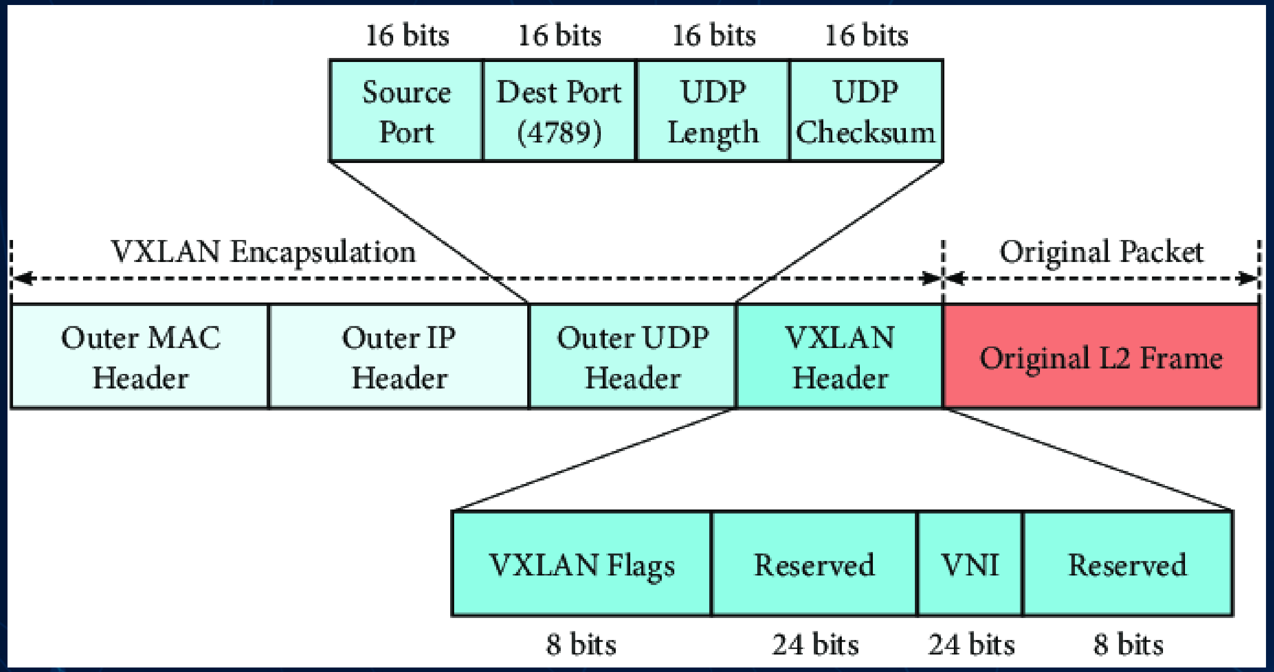
Network Planning and Design Engineer
ConsoleConnect / AS3491

Co-owner / Network engineer
ITMINDS

About:ThisPresentation

- What is EVPN/VXLAN
- Why use EVPN/VXLAN
- Our Journey
- Notable observations

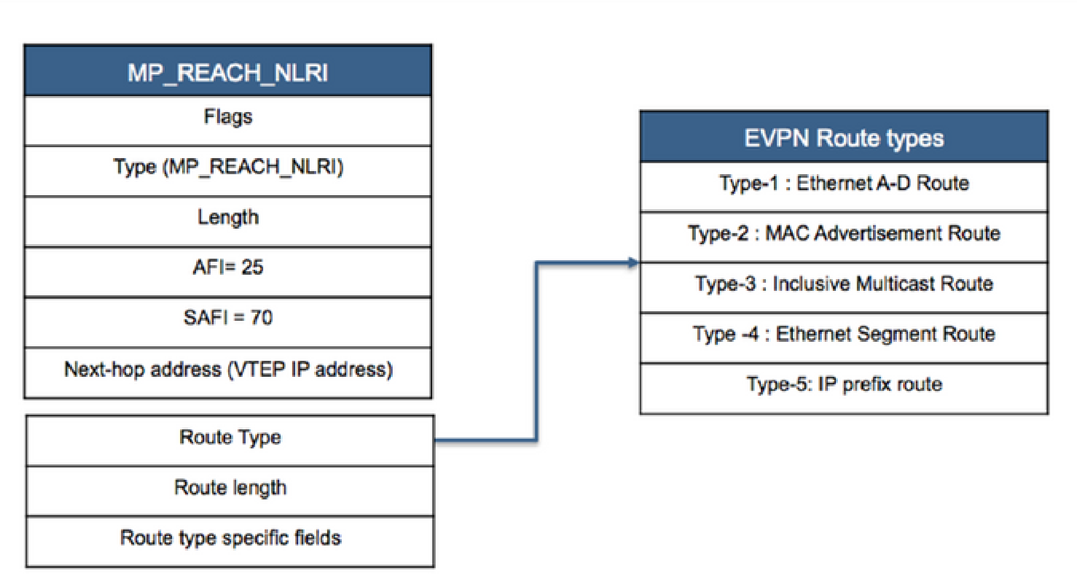
About:VXLAN



- VXLAN is the data plane protocol
- Ethernet tunnel over UDP
- DST Port == 4789
- SRC Port based on inner packet headers hash
- Easy load-balancing over the core

Image source: https://www.researchgate.net/figure/VXLAN-Packet-Encapsulation_fig1_354232739

About:EVPN



MP_REACH_NLRI

Flags

Type (MP_REACH_NLRI)

Length

AFI= 25

SAFI = 70

Next-hop address (VTEP IP address)

Route Type

Route length

Route type specific fields

EVPN Route types

Type-1 : Ethernet A-D Route

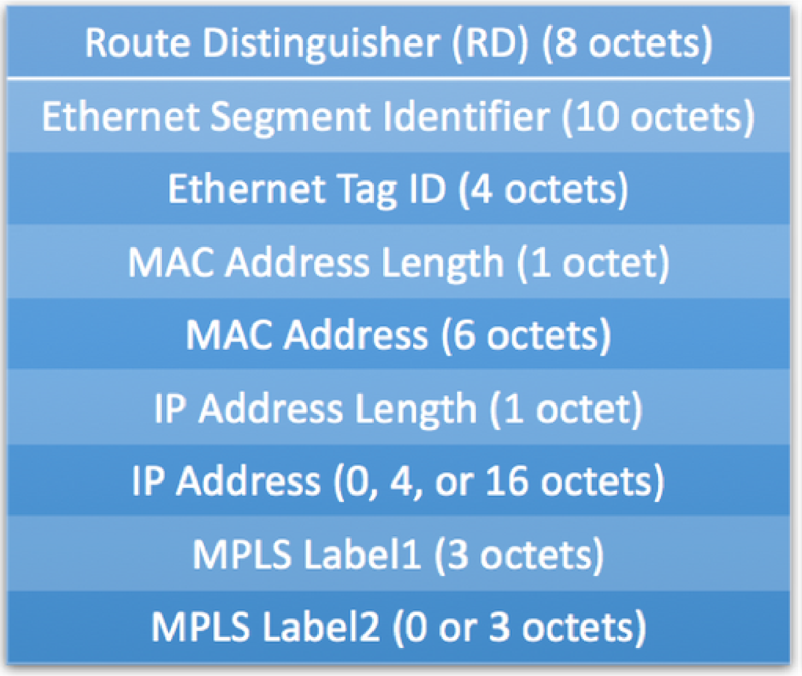
Type-2 : MAC Advertisement Route

Type-3 : Inclusive Multicast Route

Type -4 : Ethernet Segment Route

Type-5: IP prefix route

Type-2 route structure



Route Distinguisher (RD) (8 octets)
Ethernet Segment Identifier (10 octets)
Ethernet Tag ID (4 octets)
MAC Address Length (1 octet)
MAC Address (6 octets)
IP Address Length (1 octet)
IP Address (0, 4, or 16 octets)
MPLS Label1 (3 octets)
MPLS Label2 (0 or 3 octets)

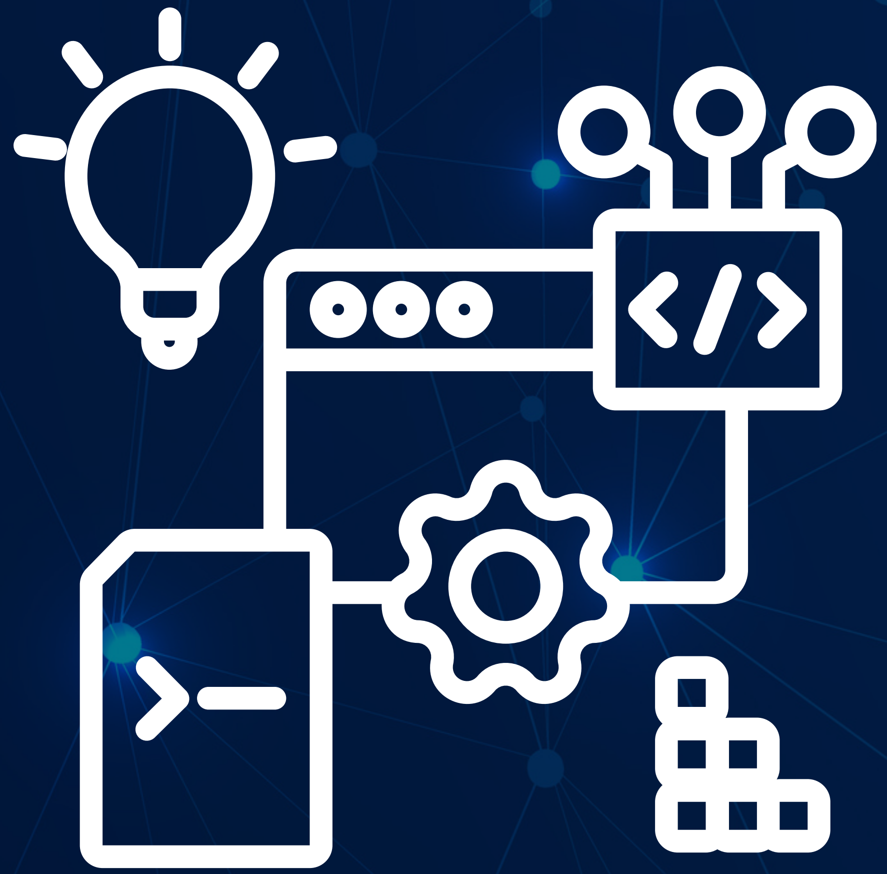
- EVPN is the data plane protocol
- BGP AFI/SAFI 25/70
- 5 Route types
- MAC/IP learning
- Unicast or Multicast BUM replication
- ARP suppression
- Bridging/Routing
- Multihoming
- Anycast Gateway

Image sources: https://www.researchgate.net/figure/VXLAN-Packet-Encapsulation_fig1_354232739
https://bgphelp.com/2017/04/03/evpn-type-2-macip-advertisement-route-explained

About:Requirements

- Network Immutability
- Scalability

About:Components



Hardware

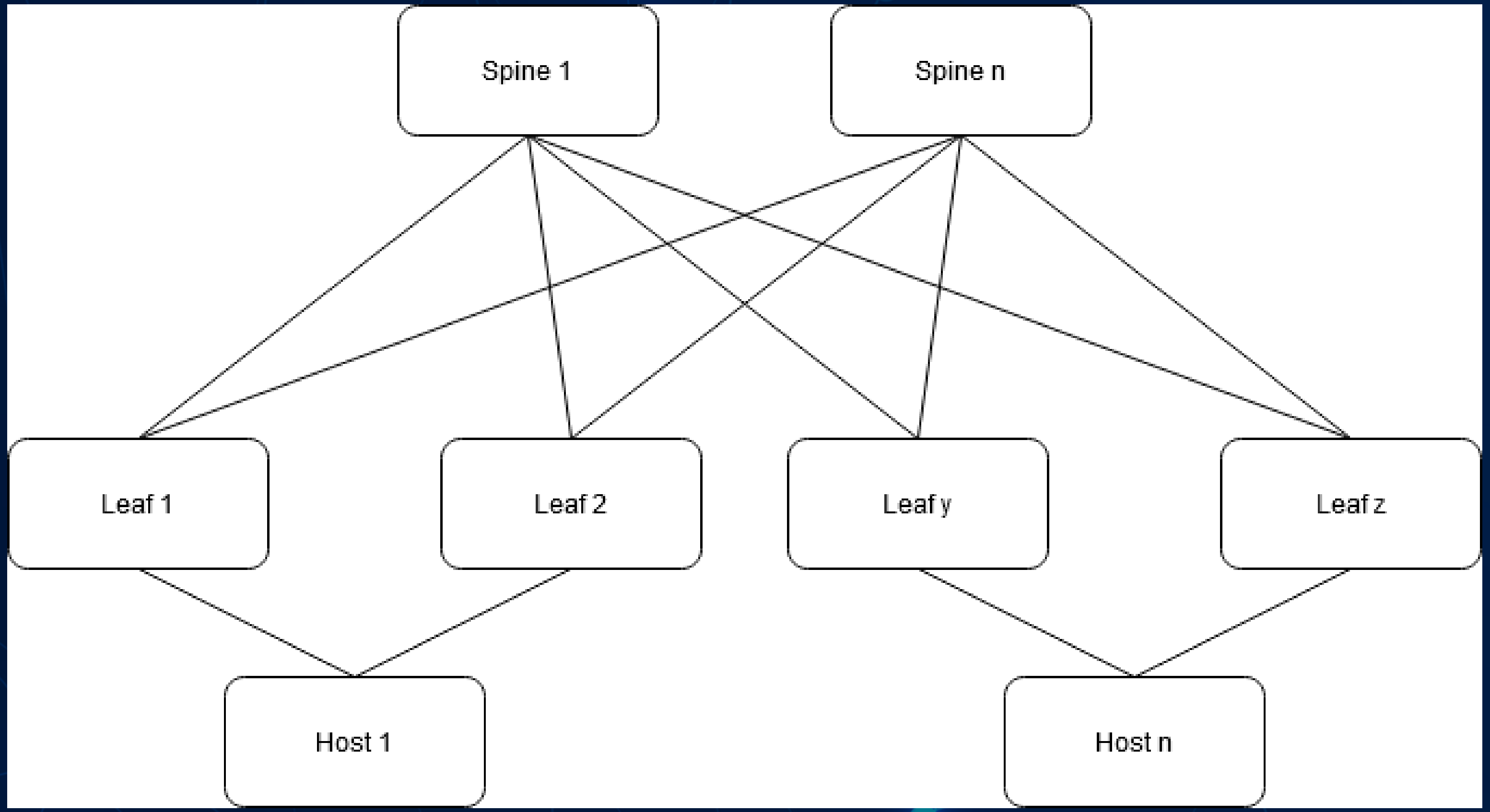
- Dell Poweredge
- Cisco Nexus

Software

- Ubuntu Server
- Linux Kernel
- KVM
- Cloudstack
- FRRouting

About:L1Design

Clos network



- Typical design for EVPN/VXLAN networks
- Provides equidistant host placement

About:IGPDesign

IGP Options

- eBGP - Requires vendor magic
- iBGP - Requires vendor magic
- OSPFv3 - No IPv4 support on our gear :(
- IS-IS it is!

About:BGPDesign

- iBGP for the fabric, using a private ASN.
- All leaves and hosts have sessions towards the BGP route-reflectors.
- Currently using the spine nodes as route-reflectors. When we are near the scaling limits, we will most likely spin up dedicated route-reflector VMs.
- Internet routers using the public ASN and peering with the fabric inside an internet VRF.

About:Cloudstack

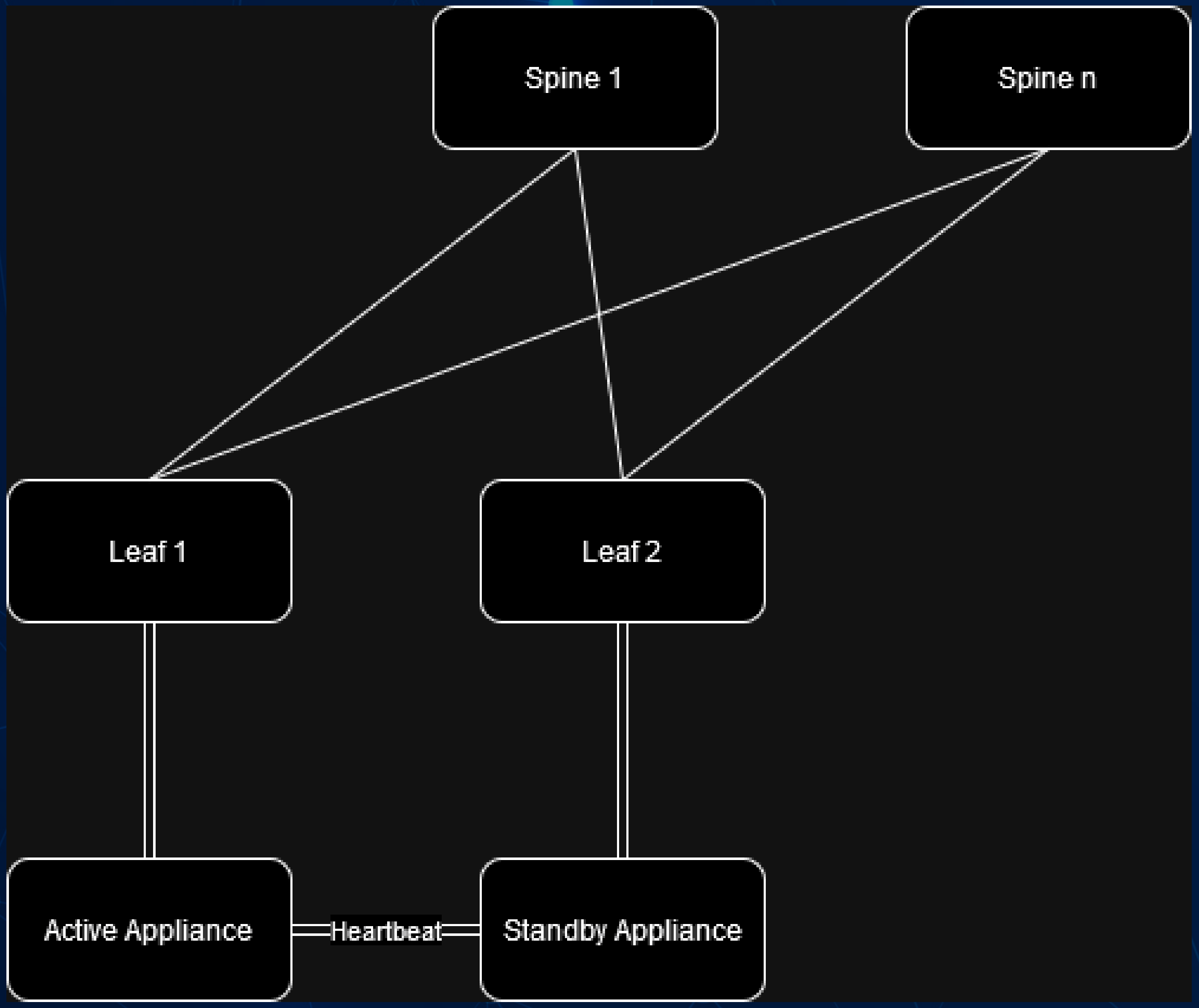
- Private cloudstack agent networks implemented as bridged VXLAN networks.
- Tenant networks use a slightly modified version of a publicly available script.
- Patch submitted (and merged) to allow VNI devices to work with cloudstack-agent.
- Internet implemented as a routed VXLAN network on the hosts, using symmetric IRB, in its own VRF.
- Internet access implemented mostly with cloudstack virtual routers.

Script source: https://gist.github.com/wido/51cb9880d86f08f73766634d7f6df3f4

About:FRRouting

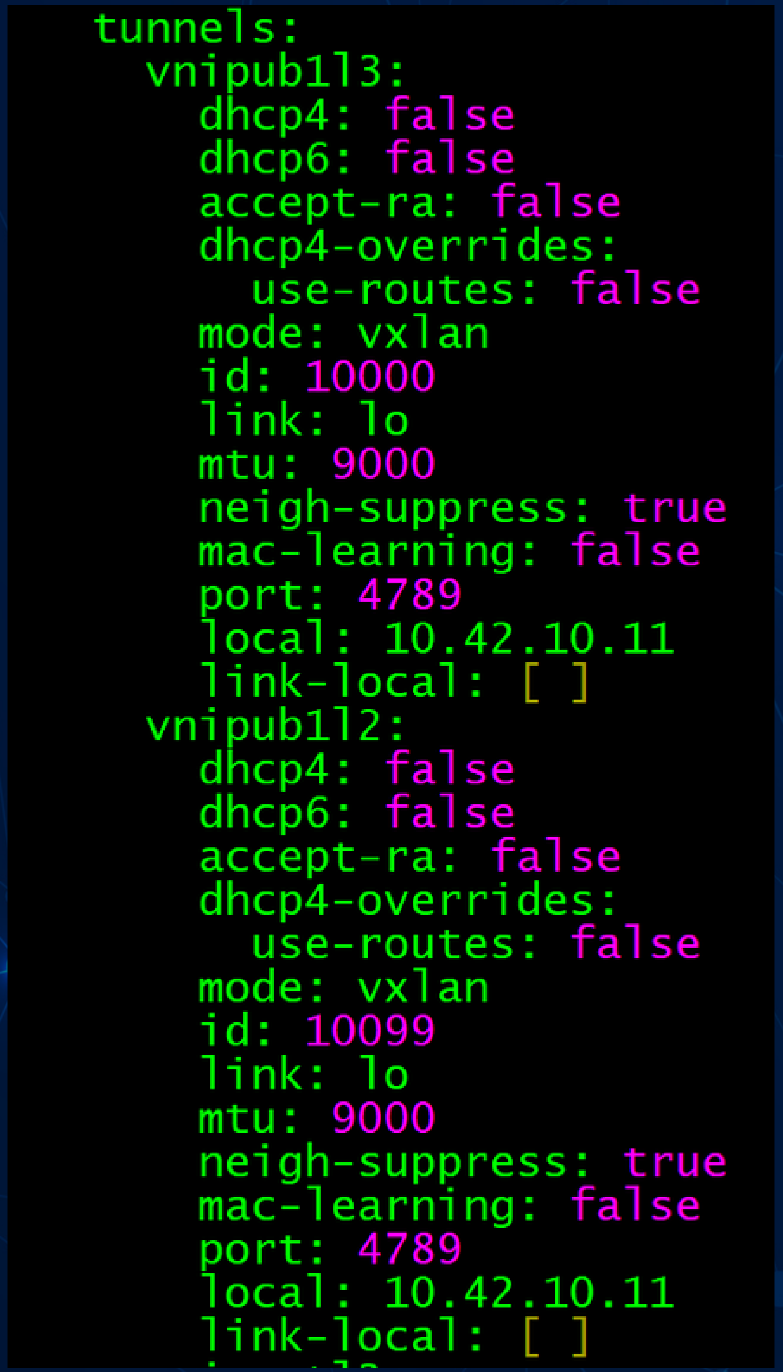
- Most issues we encounted had more to do with the interaction between FRR and the Linux kernel than FRR itself. We were able to solve those with a couple of scripts.

About:MiscIssues

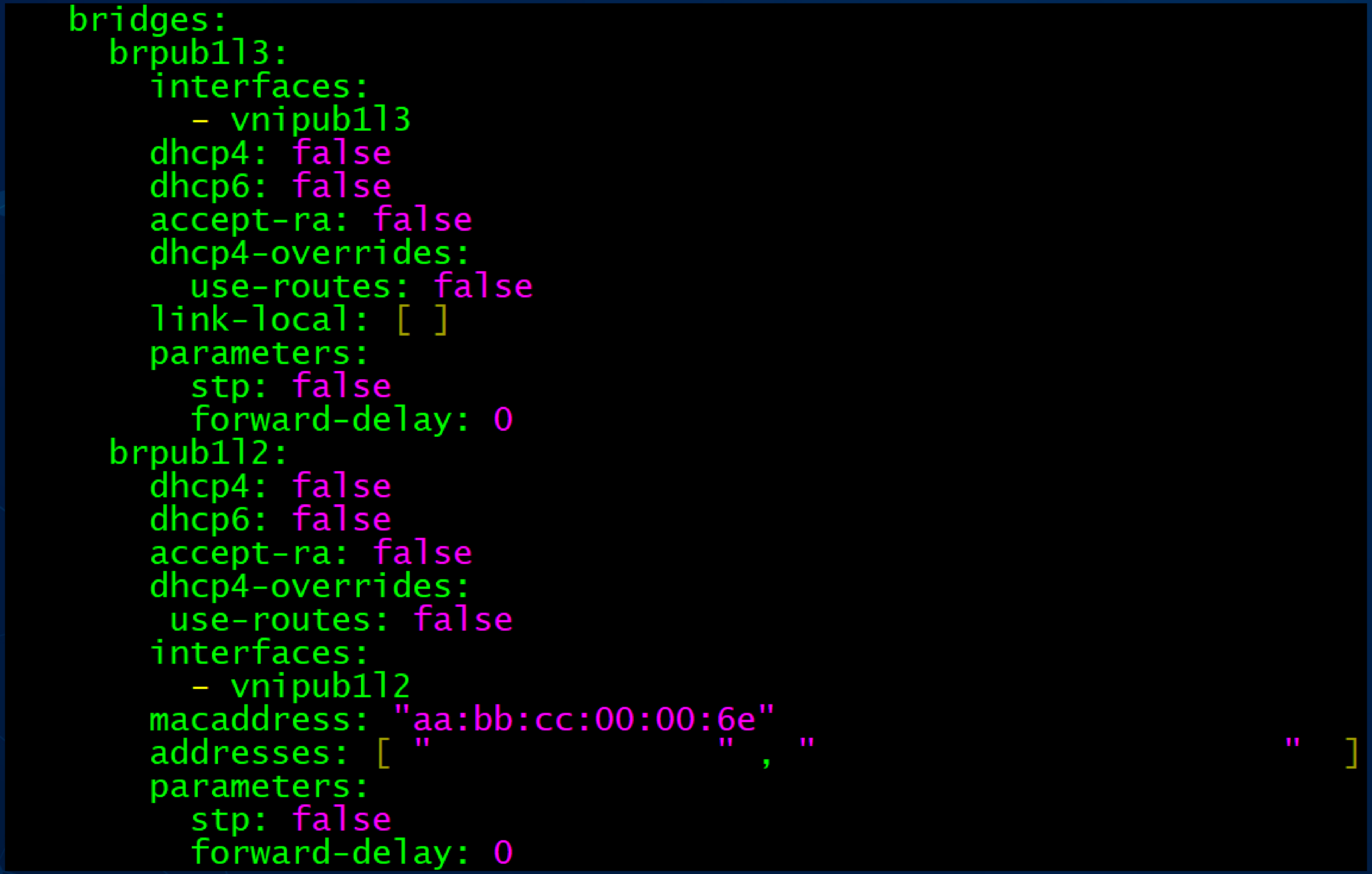


- **MLAG - Don’t do it, unless you really have to.**

About:VXLAN routing example - Netplan

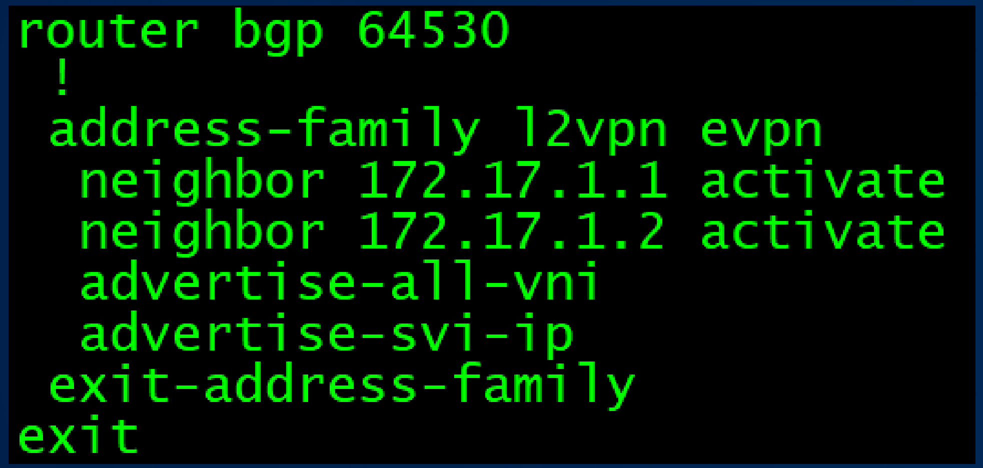


```
  tunnels:
    vnipub1l3:
      dhcp4: false
      dhcp6: false
      accept-ra: false
      dhcp4-overrides:
        use-routes: false
      mode: vxlan
      id: 10000
      link: lo
      mtu: 9000
      neigh-suppress: true
      mac-learning: false
      port: 4789
      local: 10.42.10.11
      link-local: [ ]
    vnipub1l2:
      dhcp4: false
      dhcp6: false
      accept-ra: false
      dhcp4-overrides:
        use-routes: false
      mode: vxlan
      id: 10099
      link: lo
      mtu: 9000
      neigh-suppress: true
      mac-learning: false
      port: 4789
      local: 10.42.10.11
      link-local: [ ]
```

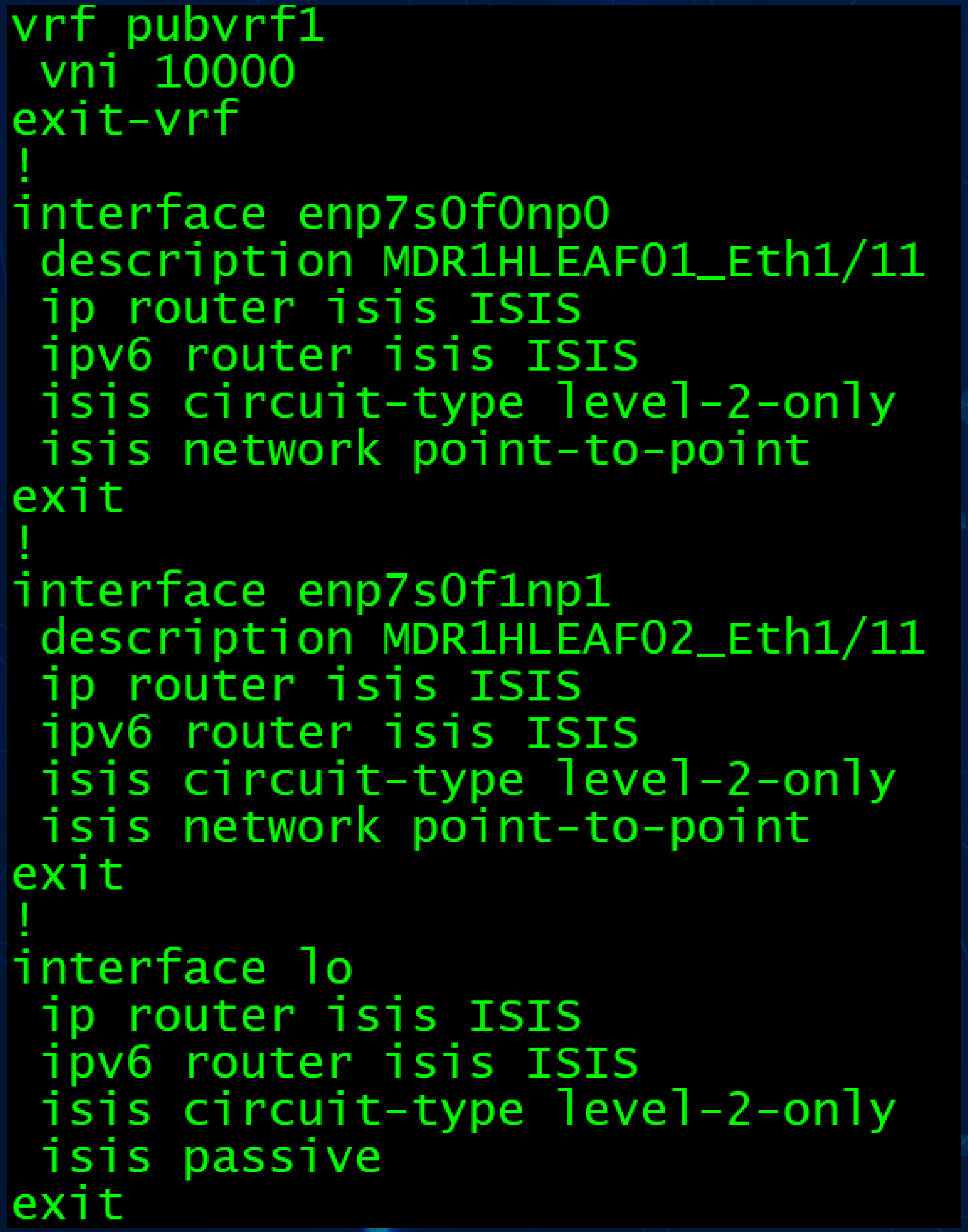


```
  bridges:
    brpub1l3:
      interfaces:
        - vnipub1l3
      dhcp4: false
      dhcp6: false
      accept-ra: false
      dhcp4-overrides:
        use-routes: false
      link-local: [ ]
      parameters:
        stp: false
        forward-delay: 0
    brpub1l2:
      dhcp4: false
      dhcp6: false
      accept-ra: false
      dhcp4-overrides:
       use-routes: false
      interfaces:
        - vnipub1l2
      macaddress: "aa:bb:cc:00:00:6e"
      addresses: [ "[redacted]" , "[redacted]" ]
      parameters:
        stp: false
        forward-delay: 0
```

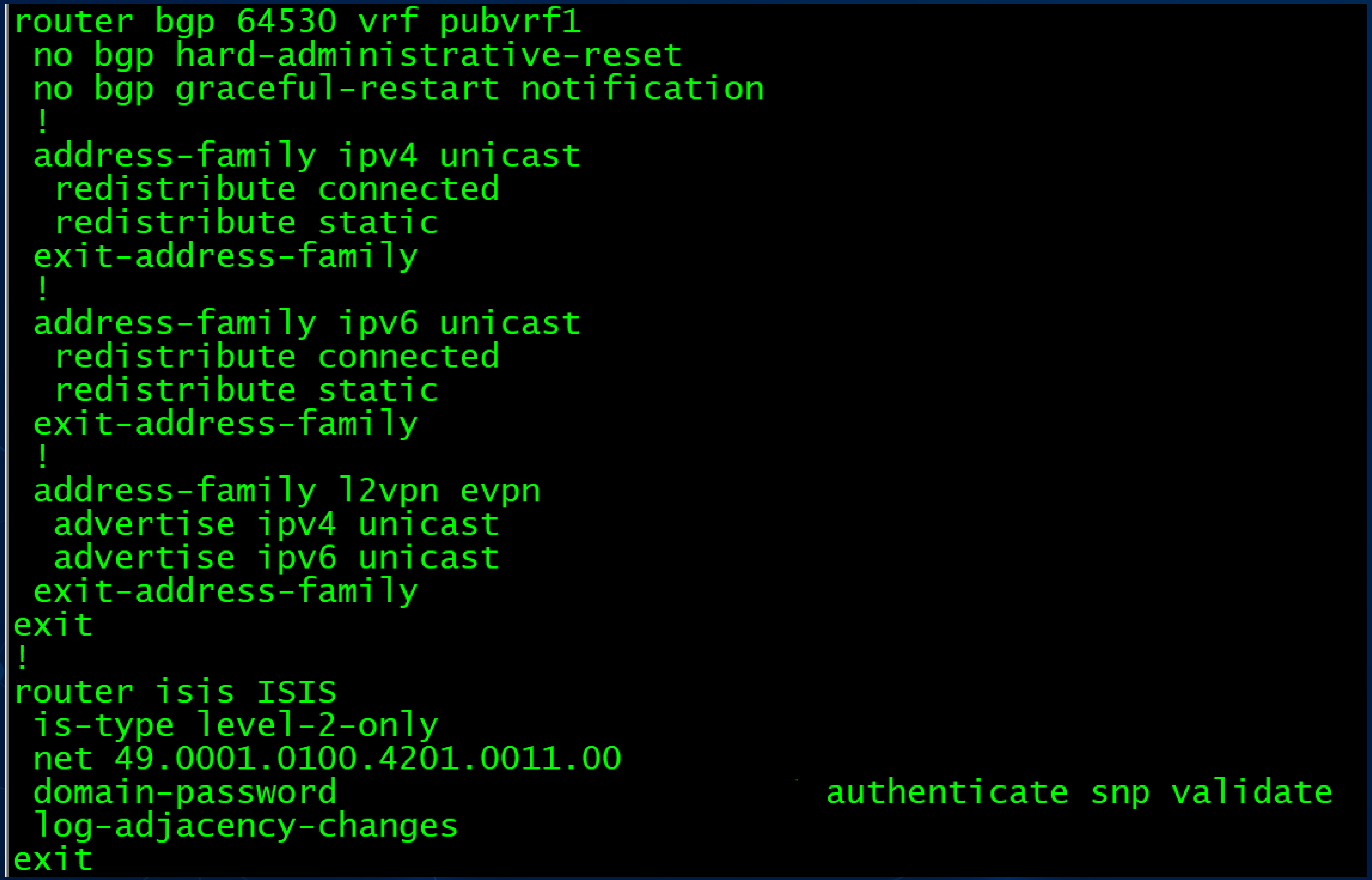
About:VXLAN routing example - FRR



```
router bgp 64530
 !
 address-family l2vpn evpn
  neighbor 172.17.1.1 activate
  neighbor 172.17.1.2 activate
  advertise-all-vni
  advertise-svi-ip
 exit-address-family
exit
```

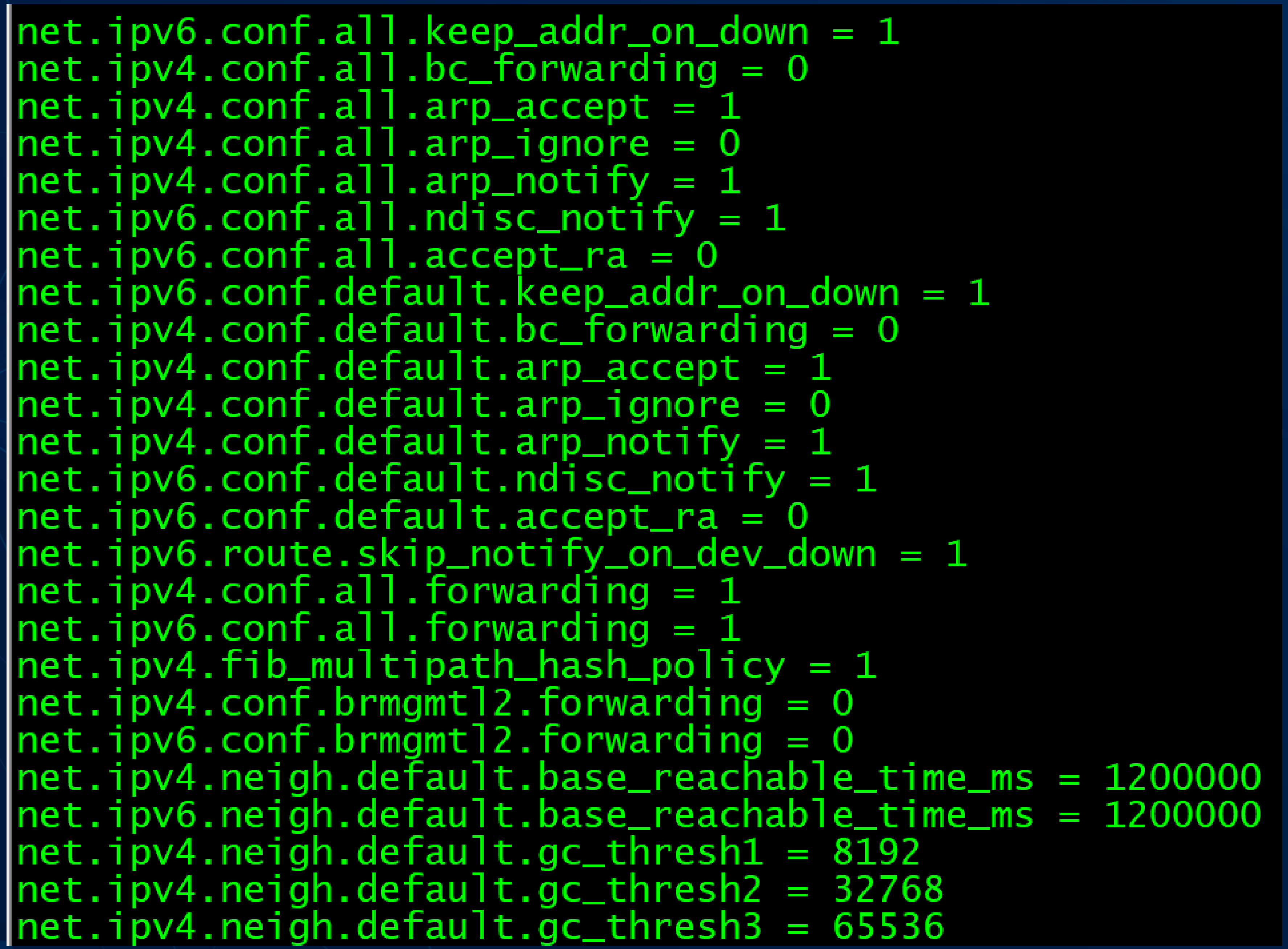


```
vrf pubvrf1
 vni 10000
exit-vrf
!
interface enp7s0f0np0
 description MDR1HLEAF01_Eth1/11
 ip router isis ISIS
 ipv6 router isis ISIS
 isis circuit-type level-2-only
 isis network point-to-point
exit
!
interface enp7s0f1np1
 description MDR1HLEAF02_Eth1/11
 ip router isis ISIS
 ipv6 router isis ISIS
 isis circuit-type level-2-only
 isis network point-to-point
exit
!
interface lo
 ip router isis ISIS
 ipv6 router isis ISIS
 isis circuit-type level-2-only
 isis passive
exit
```



```
router bgp 64530 vrf pubvrf1
 no bgp hard-administrative-reset
 no bgp graceful-restart notification
 !
 address-family ipv4 unicast
  redistribute connected
  redistribute static
 exit-address-family
 !
 address-family ipv6 unicast
  redistribute connected
  redistribute static
 exit-address-family
 !
 address-family l2vpn evpn
  advertise ipv4 unicast
  advertise ipv6 unicast
 exit-address-family
exit
!
router isis ISIS
 is-type level-2-only
 net 49.0001.0100.4201.0011.00
 domain-password [redacted] authenticate snp validate
 log-adjacency-changes
exit
```

About:sysctl variables



```
net.ipv6.conf.all.keep_addr_on_down = 1
net.ipv4.conf.all.bc_forwarding = 0
net.ipv4.conf.all.arp_accept = 1
net.ipv4.conf.all.arp_ignore = 0
net.ipv4.conf.all.arp_notify = 1
net.ipv6.conf.all.ndisc_notify = 1
net.ipv6.conf.all.accept_ra = 0
net.ipv6.conf.default.keep_addr_on_down = 1
net.ipv4.conf.default.bc_forwarding = 0
net.ipv4.conf.default.arp_accept = 1
net.ipv4.conf.default.arp_ignore = 0
net.ipv4.conf.default.arp_notify = 1
net.ipv6.conf.default.ndisc_notify = 1
net.ipv6.conf.default.accept_ra = 0
net.ipv6.route.skip_notify_on_dev_down = 1
net.ipv4.conf.all.forwarding = 1
net.ipv6.conf.all.forwarding = 1
net.ipv4.fib_multipath_hash_policy = 1
net.ipv4.conf.brmgmtl2.forwarding = 0
net.ipv6.conf.brmgmtl2.forwarding = 0
net.ipv4.neigh.default.base_reachable_time_ms = 1200000
net.ipv6.neigh.default.base_reachable_time_ms = 1200000
net.ipv4.neigh.default.gc_thresh1 = 8192
net.ipv4.neigh.default.gc_thresh2 = 32768
net.ipv4.neigh.default.gc_thresh3 = 65536
```

Questions ???