CERN Computer Centre(s) Network evolution

(Part II)

Vincent DUCRET – vincent.ducret@cern.ch

HEPiX Spring 2023 – Taipei

Agenda

Part I:

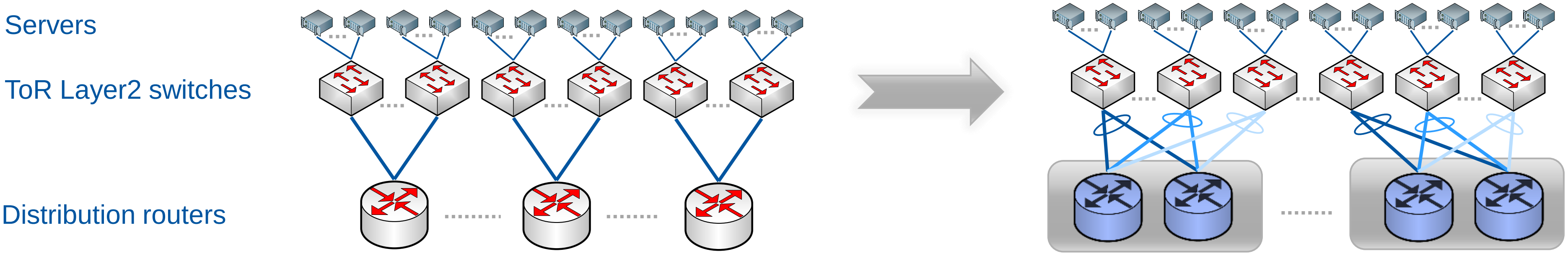
- Reminder about 2019 status
- Datacentre migration (during COVID19 lockdown - 2020)
- Overview of current Datacentre Network
- Evolution of links between Main datacentre and other CERN sites
- Plans for 2023 and new Prévessin Data Centre (PDC)

Part II:

- New tools and/or features we started to deploy
- Main issues faced with new Datacentre Network setup

New tools and/or features

- Dual router attachment for all switches
  - Based on EVPN/VxLAN ESI on the routers



Servers

ToR Layer2 switches

Distribution routers

Former setup

Current setup

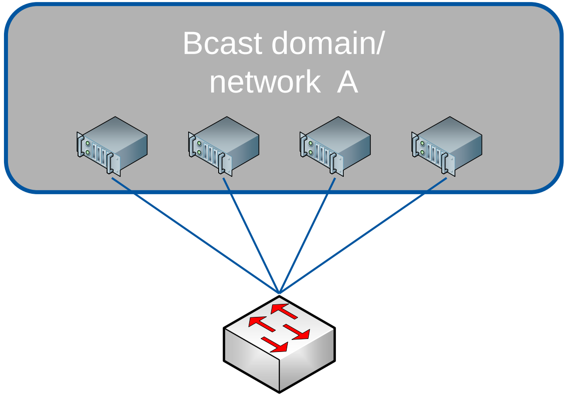
New tools and/or features

- Vlan on the ToR switch
  - Multi domain support on a single switch

- Before:
  - All servers on a switch must belong to the same Network/Bcast domain

➔ Needs and server allocation may evolve over time.

➔ If some servers need to be moved, we must connect them to a different switch

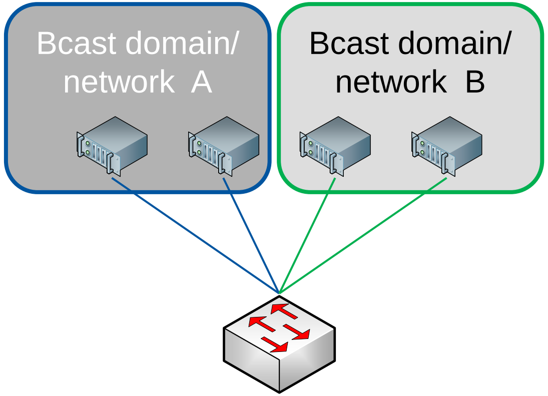


New tools and/or features

- Vlan on the ToR switch
  - Multi domain support on a single switch

- After:
  - Servers on a switch can belong to different Network/Bcast domains

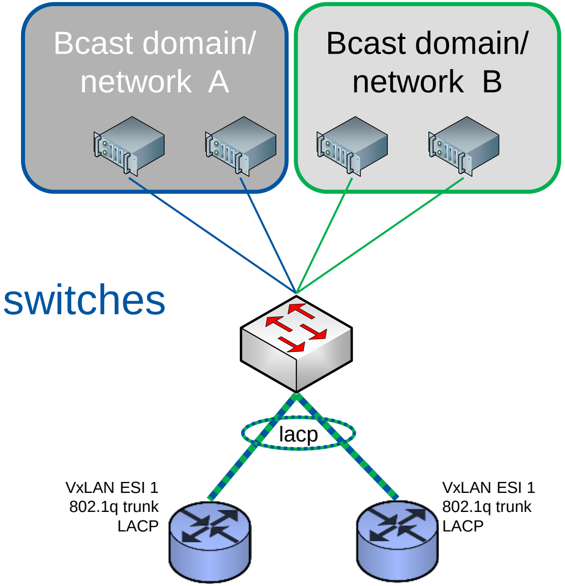
➔ Possible to “move” servers logically, without any additional hardware/cabling required



New tools and/or features

- Vlan on the ToR switch
  - Multi domain support on a single switch

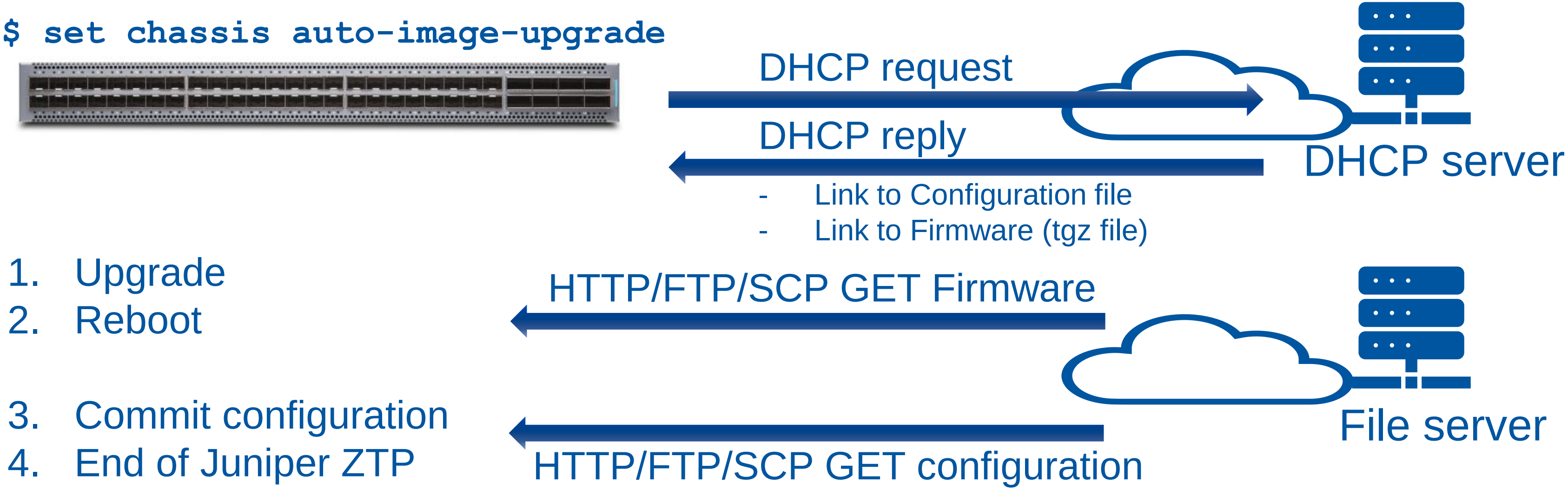
- After:
  - Need to configure 802.1q trunk between routers and switches
  - Migration implies small downtime
  - Adding Vlans afterwards is transparent



- Currently applied on a limited number of switches

New tools and/or features

- Zero Touch Provisioning (ZTP) for Juniper devices
  - ZTP is proposed by default on Juniper devices



```
$ set chassis auto-image-upgrade
```

DHCP request

DHCP reply

- Link to Configuration file
- Link to Firmware (tgz file)

DHCP server

1. Upgrade
2. Reboot

3. Commit configuration
4. End of Juniper ZTP

HTTP/FTP/SCP GET Firmware

HTTP/FTP/SCP GET configuration

File server

New tools and/or features

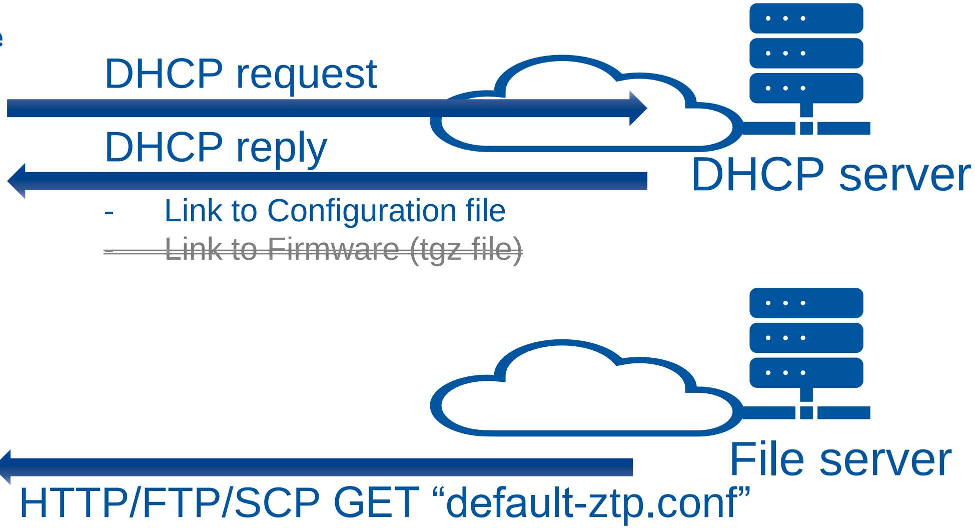
- Zero Touch Provisioning (ZTP) for Juniper devices
  - Juniper ZTP has some limitations:
    - Single firmware upgrade: not able to follow an upgrade path
    - Supporting switches with different version and/or different configuration file implies specific configuration in dhcpd.conf

➔ Need to tune the default ZTP process to have a cleaner process

New tools and/or features

- Zero Touch Provisioning (ZTP) for Juniper devices
  - Juniper ZTP + enhancement via SLAX scripts:

```
$ set chassis auto-image-upgrade
```



DHCP request

DHCP reply

- Link to Configuration file
- ~~Link to Firmware (tgz file)~~

DHCP server

File server

HTTP/FTP/SCP GET “default-ztp.conf”

1. Load “default-ztp.conf”
2. End of Juniper ZTP
3. Start “ztp-discovery.slax” script (called by the configuration file)

New tools and/or features

- Zero Touch Provisioning (ZTP) for Juniper devices
  - Juniper ZTP + enhancement via SLAX scripts:

ztp-discovery.slax

- Check our DB to get the hostname corresponding to the S/N
- Apply hostname to switch and run next script

ztp-upgrade.slax

- Based on the hostname or model type, upgrade the devices following a specific upgrade path (multiple upgrades if required)
- Once target version is reached, run next script

ztp-configure.slax

- Based on the hostname or model type, apply the configuration
- If required, apply licences
- If required, power-off device (ex: stock device preparation)

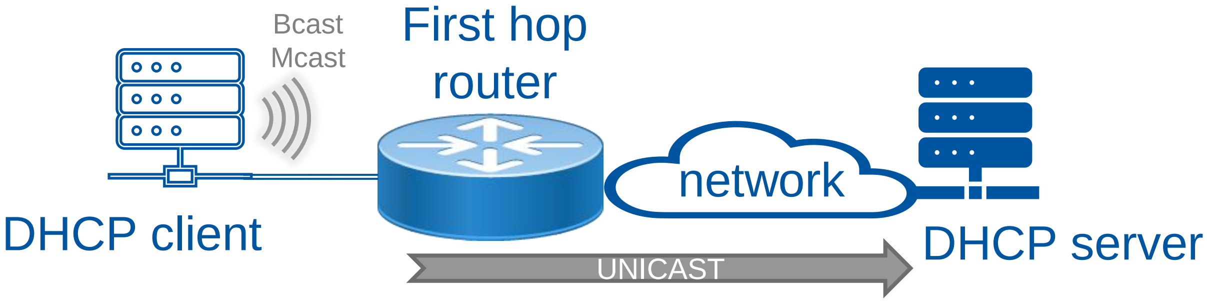
Special thanks to Carles Kishimoto (carles.kishimoto@cern.ch) who developed the SLAX scripts

Main issues

- High delivery time
  - Need to order ~1 year in advance (ex: PDC)
  - Rely on spare/lab devices for urgent request (Ex: ALICE O2 QFX chassis)

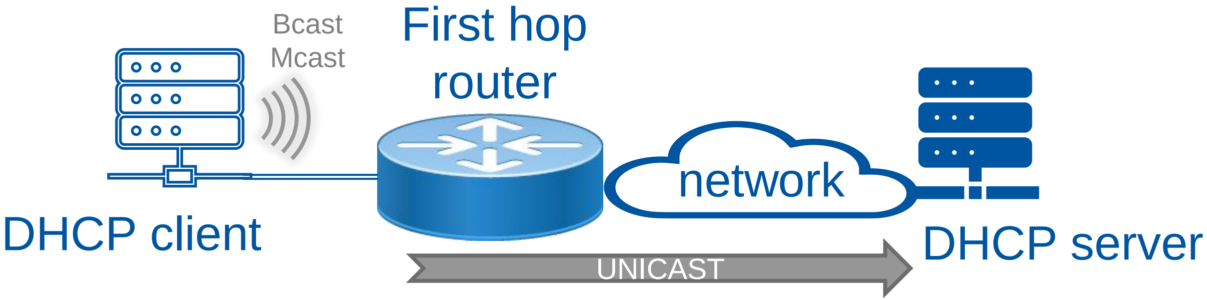
- Several IPv6 issues:
  - DHCPv6 issues
    - DHCPv6-relay option-79 in a wrong format
    - Packet loss with DHCPv6-relay binding and dual router setup
    - DHCPv6-relay option-79 not inserted for traffic crossing VxLAN tunnel
  - CEPH nodes not handling dual IPv6 default gateway dynamically

DHCPv6 relay issues: What is DHCP relay?



- DHCP client sends broadcast/multicast messages
- The first hop router intercepts the broadcast and transforms it to unicast to relay the message up to the DHCP server

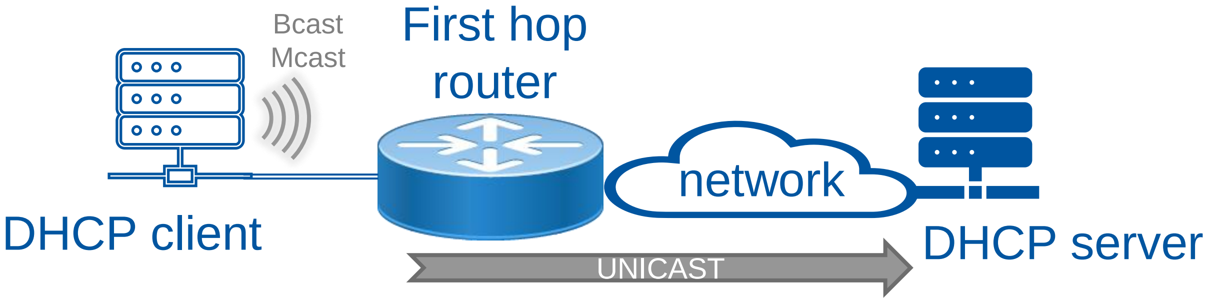
DHCPv6 relay issues: option-79 wrong format



- DHCPv6 option-79 is added by the relaying router
- It contains the client MAC-Address (which may not be part of the DUID)
- Juniper router in version 18.4 adds option-79 is a wrong format

RFC 6939:	8 Bytes = **`00:01:`**`<client-mac-address>`
JunOS 18.4:	6 Bytes = `<client-mac-address>`

DHCPv6 relay issues: option-79 wrong format



- We applied a workaround on the DHCP server so it accepts “6 Bytes” option-79 as sent by Juniper router
- This workaround is not supported on the new KEA DHCP servers, so we needed a long-term fix

DHCPv6 relay issues: DHCPv6 binding and dual router setup

- Juniper routers has two modes for DHCP relay:
  - Forward-only: the router simply relays the DHCP packets.
  - DHCP binding: the router relays the DHCP packets and keep a track (binding-table) of the `<mac-address><ip-address>` associations.

- Until version 21.2, adding option-79 for DHCPv6 was only possible with DHCPv6 binding

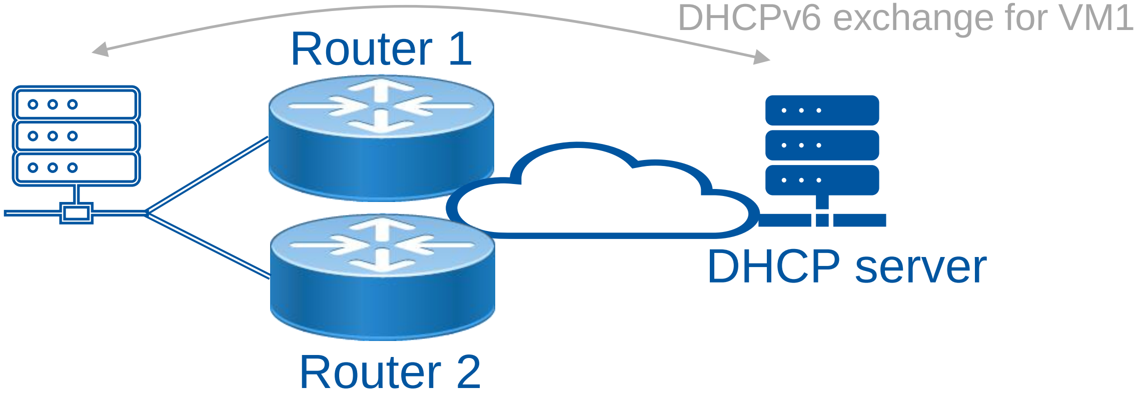
- When DHCPv6 binding is used, some traffic inspection is done (and cannot be disabled) and will drop packet if the `<mac-address><ip-address>` association is not correct.

DHCPv6 relay issues: DHCPv6 binding and dual router setup

Router 1 binding table

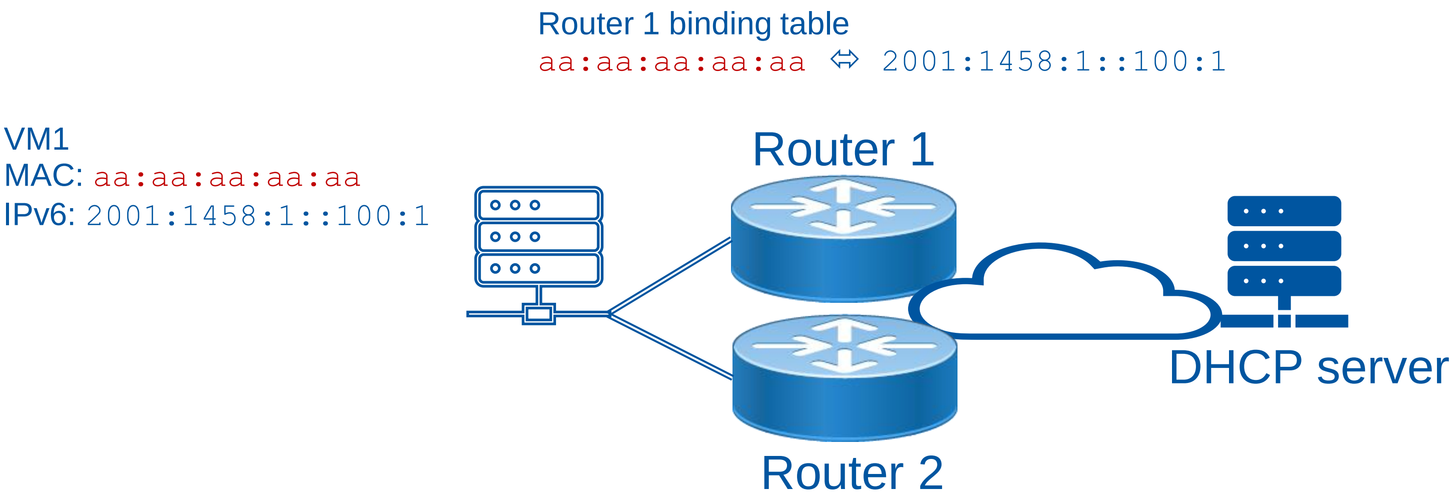
aa:aa:aa:aa:aa ⇔ 2001:1458:1::100:1

VM1
MAC: aa:aa:aa:aa:aa
IPv6: 2001:1458:1::100:1



- VM1 created and DHPCv6 relayed by Router 1
- VM1 traffic routed normally by the two routers

DHCPv6 relay issues: DHCPv6 binding and dual router setup



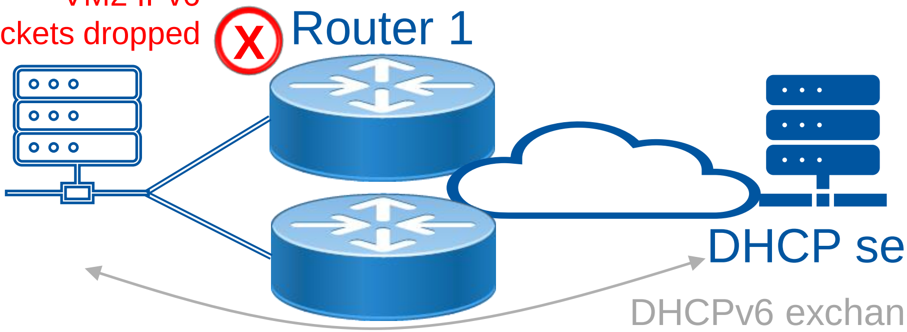
- VM1 destroyed
- Router Binding table entry stays (1 week idle timeout – based on DHCP lease duration)

DHCPv6 relay issues: DHCPv6 binding and dual router setup

Router 1 binding table

`aa:aa:aa:aa:aa ⇔ 2001:1458:1::100:1`

VM2
MAC: `bb:bb:bb:bb:bb`
IPv6: `2001:1458:1::100:1`



Router 2 binding table

`bb:bb:bb:bb:bb ⇔ 2001:1458:1::100:1`

- VM2 created:
  - DHCPv6 relayed by Router 2
  - It reuses VM1 IPv6 address
- Router1 will drop packet from VM2 due to DHPCv6 binding old entry

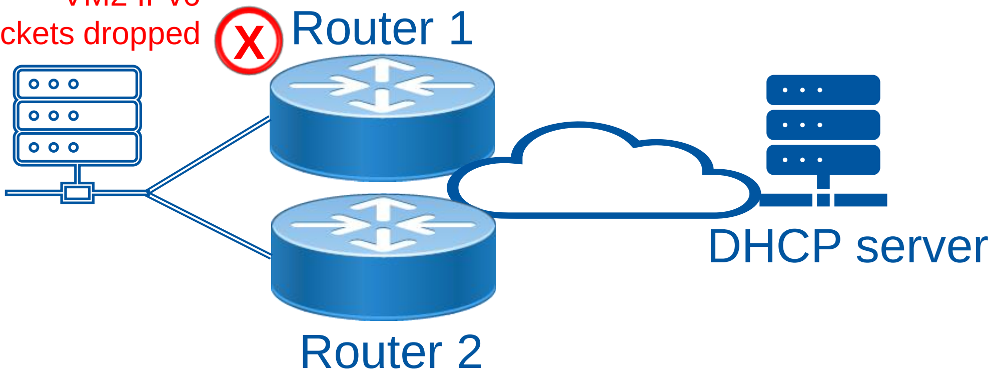
DHCPv6 relay issues: DHCPv6 binding and dual router setup

Router 1 binding table

`aa:aa:aa:aa:aa` ⇔ `2001:1458:1::100:1`

VM2
MAC: `bb:bb:bb:bb:bb`
IPv6: `2001:1458:1::100:1`

VM2 IPv6
packets dropped



Router 2 binding table

`bb:bb:bb:bb:bb` ⇔ `2001:1458:1::100:1`

- VM2 has ~50% IPv6 packet drop
- Workaround: manually clear Router1 old DHPCv6 binding entry

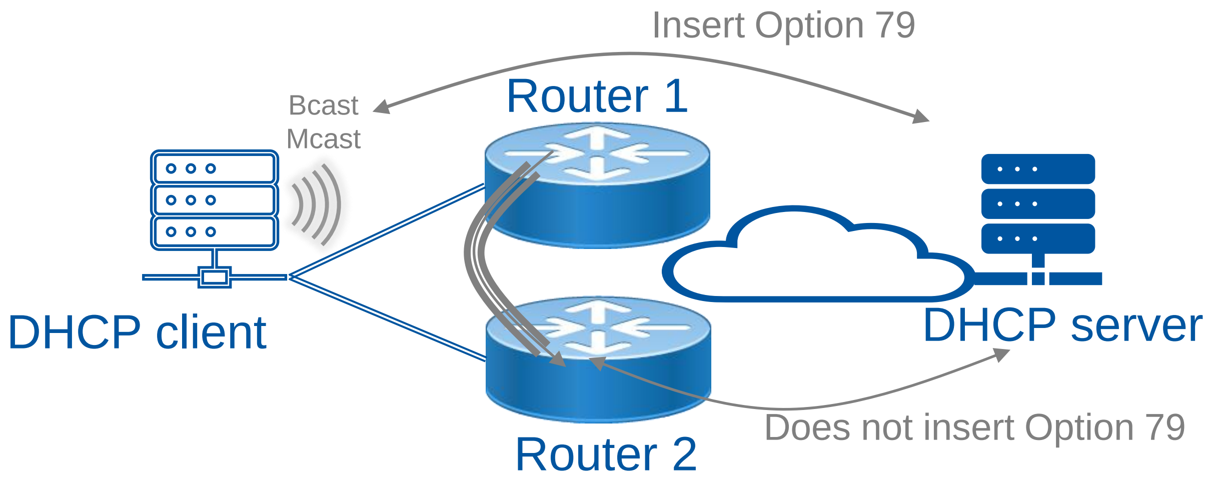
DHCPv6 relay issues

- Version 21.2R1-S2:
  - Fixed the issue with Option-79 “wrong format”
  - Support Option-79 insertion with DHPCv6 “forward only “ (no more binding table)

- However it introduced a new bug…

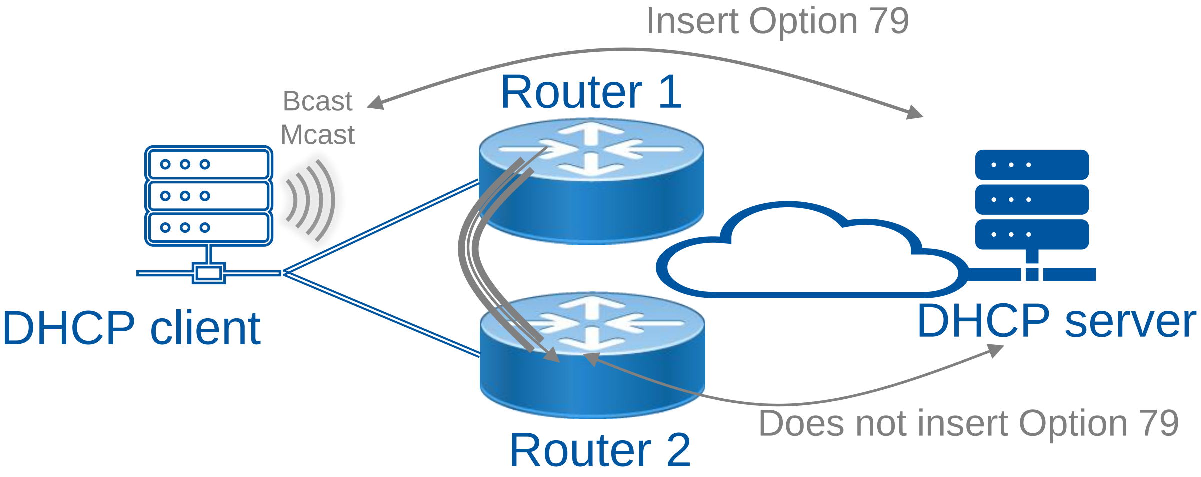
By Francisco Welter-Schultes - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=103292729

DHCPv6 relay issues: Option 79 not inserted via VxLAN tunnel



- On version 21.2, DHCPv6 Bcast/Macst packets are replicated to the 2nd router via VxLAN tunnel
- DHCPv6 packets crossing VxLAN are relayed by 2nd router without Option 79
- DHCP server sees two requests coming, one with Option 79, the second without

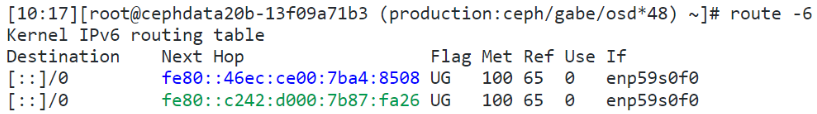
DHCPv6 relay issues: Option 79 not inserted via VxLAN tunnel



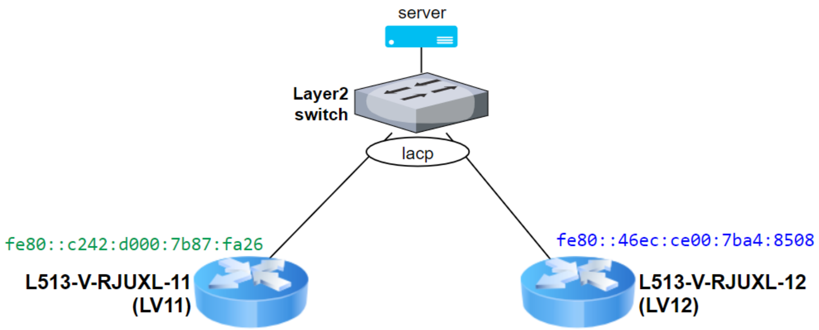
- DHPCv6 server replies twice (with or without IP address allocation)
- It is transparent for most of the DHCP clients
- Depending on the client/OS version, it may lead to clients being unable to get an IPv6 address…
- Version 21.2R3-S2 fixed this issue

CEPH Client and IPv6 default route handling: issue

- By default, servers learn their IPv6 default gateway via RA (Router Advertisement), and use “link-local” IPv6 address
- Servers have two default routes, each pointing to one of the router Link-local address
- Depending on the OS/system, the server may not detect correctly that one router is down.
- Behaviour tested on lab was OK, but production CEPH nodes were not…

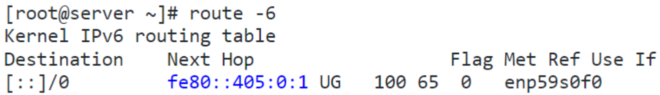


```
[10:17][root@cephdata20b-13f09a71b3 (production:ceph/gabe/osd*48) ~]# route -6
Kernel IPv6 routing table
Destination    Next Hop                   Flag Met Ref Use If
[::]/0         fe80::46ec:ce00:7ba4:8508 UG   100 65  0   enp59s0f0
[::]/0         fe80::c242:d000:7b87:fa26 UG   100 65  0   enp59s0f0
```

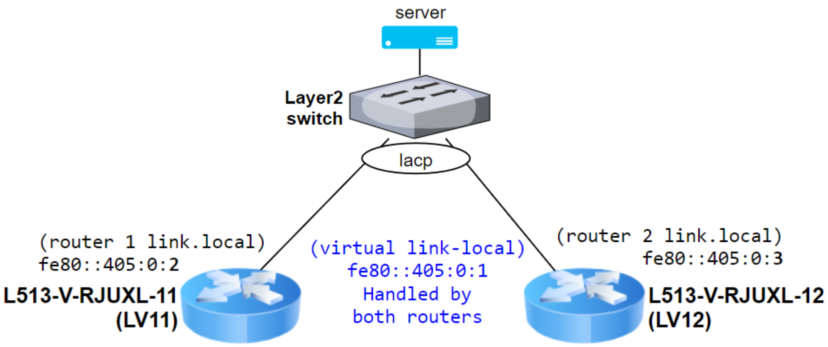


CEPH Client and IPv6 default route handling: fix

- This was fixed by changing the router configuration to use a virtual link-local address shared by both routers
- Servers will have only one default gateway pointing to this virtual link-local address (similar to IPv4)
- Behaviour is no more OS/System dependent
- Virtual addresses shared by routers were already configured for IPv4 and all IPv6 addresses, except link-local…

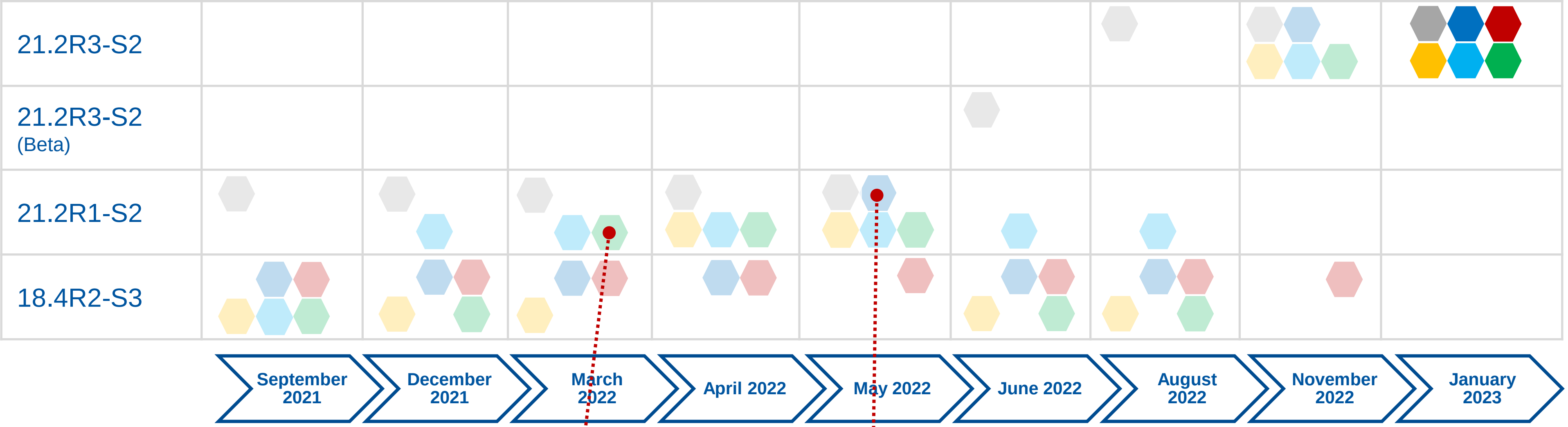


```
[root@server ~]# route -6
Kernel IPv6 routing table
Destination     Next Hop                    Flag Met Ref Use If
[::]/0          fe80::405:0:1 UG    100 65  0    enp59s0f0
```



Router Upgrade path

LAB
773(Prévessin)
513 Main Room
513 Main Room 100G
513 Barn (Critical)
513 Vault



Prod running with “known bugs”
Option-79 wrong format
DHCPv6-binding conflict

OK

Problem on CEPH IPv6
During router reboot
(default gw via RA)

Configuration fixed
OK

New (frequent) DHPCv6 issues in Main room
(option 79 missing via VxLAN)

Rollback Prod
OK

LAB extensive testing

PROD Upgrade
OK

PROD “critical” Upgrade
OK

Q&A

CERN