A Distribution Network for 2023 & Beyond

Presented by Michael Hobl

FIELDsolutions rural regional remote

TELECOMMUNICATIONS + SOFTWARE + CLOUD

Who am I?

- Network Engineer for ~12yrs
- Currently @ FSG
- Previous travels included
  - Lots of Cisco ASR
  - Juniper MX, SRX
  - MikroTik
  - A Few Tears...
- A little something called B4P

  Not the Lead Admin, though...

A Little About FSG

- ASX-listed Telco
- Rural / Regional / Resource
- Main Arenas:
  - Regional Wireless
  - IP Transit
  - L2VPN/L3VPN
  - NBN EE & V-NNI

But nothing's wrong, right?

~~**But nothing's wrong, right?**~~

```
AR3.NXTB1#traceroute 10.61.0.42
traceroute to 10.61.0.42 (10.61.0.42), 30 hops max, 60 byte packets
1  10.61.2.238 (10.61.2.238)  0.537 ms  0.393 ms  0.341 ms
2  10.61.2.188 (10.61.2.188)  11.893 ms  11.926 ms  11.889 ms
3  10.61.2.29 (10.61.2.29)  11.954 ms  11.925 ms  11.871 ms
4  10.61.2.25 (10.61.2.25)  12.074 ms  12.037 ms  12.022 ms
5  10.61.2.21 (10.61.2.21)  12.035 ms  12.060 ms  12.048 ms
6  10.61.2.190 (10.61.2.190)  26.581 ms  26.504 ms  26.425 ms
7  10.61.0.42 (10.61.0.42)  26.536 ms  26.544 ms  26.546 ms
AR3.NXTB1#
```

Exhibit A: *Routing from B1-5A-05-07 RU38F to B1-5A-05-07 RU35R via the Distribution Network (2023, Colorized)*

16/01 15:07

damn this is some legible routing

```
                                                    P1    P1    S1    S1    S1
* >   10.24.255.4/30      10.61.0.1        -     100   0     65026 65018 65000 64098 65541 ?
*     10.24.255.4/30      10.61.0.1        -     100   0     65029 65006 65007 65008 65000 64098 65541 ?
*     10.24.255.4/30      10.61.0.6        -     100   0     65029 65006 65005 65102 65100 65101 65001 65000 64098
65541 ?
*     10.24.255.4/30      10.61.0.22       -     100   0     65029 65006 65005 65003 65021 64098 65000 64098 65541
                                                    SY3   SY4   SY4   GS    GS    GS    S1    S1    S1
?
```

Exhibit B: *Routing from a VM in S1 to the PE in S1 via the Distribution Network (2023, Colorized)*

The Shortcomings

- Lack of resiliency
- Not designed for scale
- Inconsistent base configs
- Problematic to assure
- Needed for management!

The Technical Goals

- Decouple in-band device management
- Standardise inter-POP pathways
- Provide 10G and 40G handoff availability everywhere
- Additional resiliency within each POP
- Provide diverse customer handoffs

The Business Goals

- More resilient management domain
- Consistent failure modalities for each location
- Standardized port availability for sales & provisioning
- Accommodate inter-POP customer handoffs
- Easier triage for BAU/Ops teams

The Constraints

- Use existing hardware SKUs
  - 7050SX-64
  - 7050SX-72
  - 7050QX-32
- Out of support, but lots of spares
- Must remediate current environment
- Limited ability to overbuild, no new racks

Okay, so what do we do?

The Plan

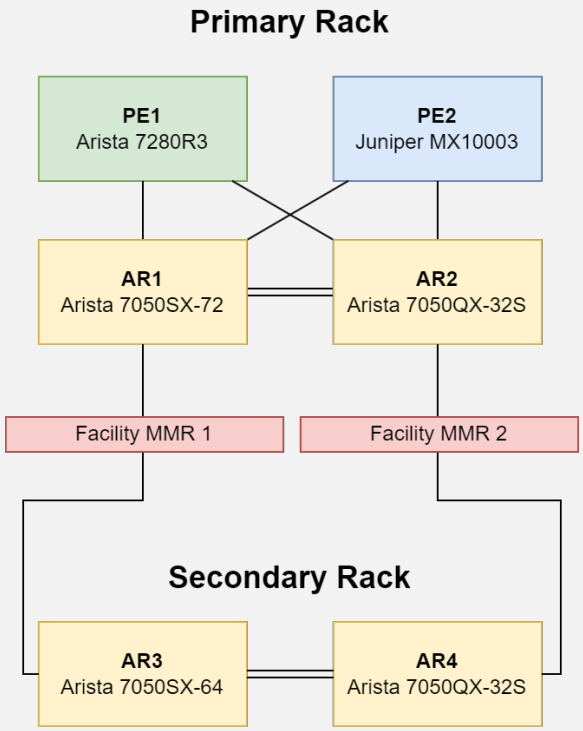
- … EVPN-VXLAN to the rescue!
- Already in production, so easy to realign
- Currently using Type 2 & Type 5 EVPN
- Moving to Type 2-only made sense
  - Limitations of Trident around VRF routing
  - L3VPNs existed in both stacks depending on alignment of Jupiter & Saturn at the time of initial provisioning

The Plan

- 7050SX-72 for 10G, 7050QX-32 for 40G
  - MXP ports suck, but boxes are cheap and plentiful
- Abstract L3 routing for Management VRF off-box
  - L1/L2 to be done on dedicated switching
  - L2 Haul can still be done via EVPN-VXLAN if needed
- Tour de POP for physicals

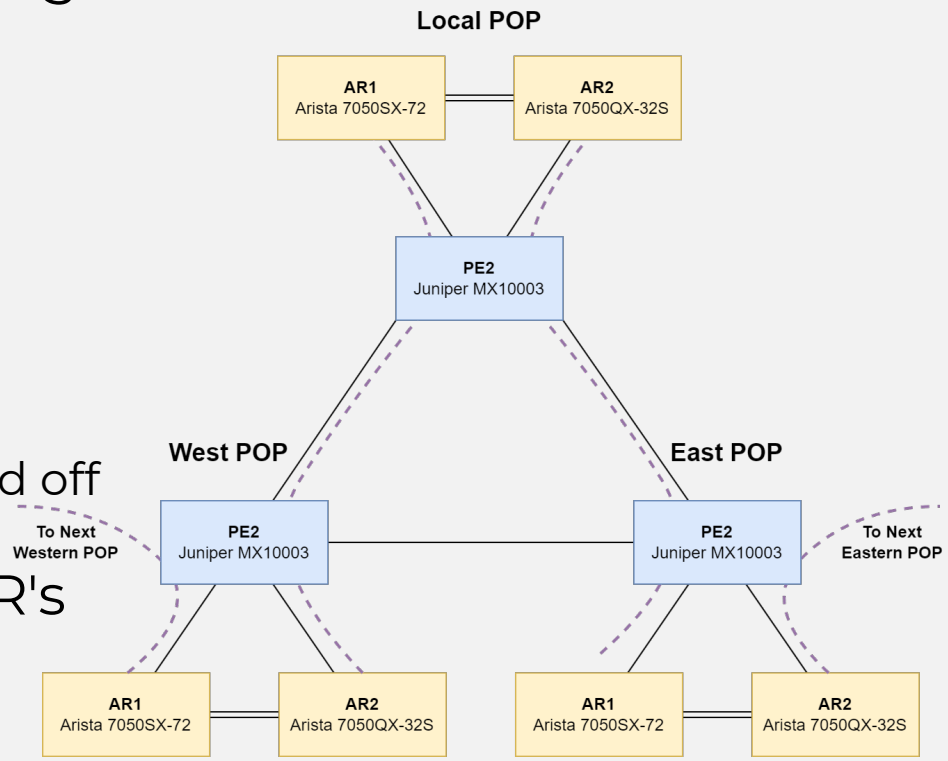
The Design

- Deploy pairs of switches
  - One 10G and one 40G
- Diverse uplinks from AR pairs
  - To each POP PE
- Interlink between each AR pair
- Daisy chain each additional pair
- AR1/2 pair as pseudo-spine



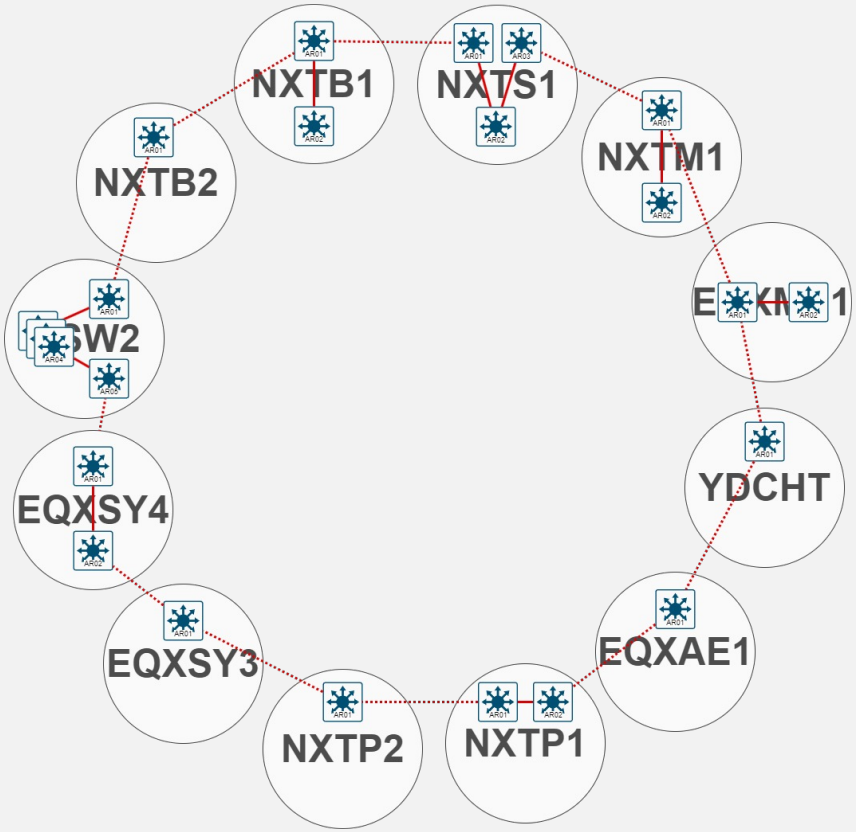
The Design

- EVPN-VXLAN A-A multihoming
  - Including towards PE
- East/West Pseudowires
  - via Routed SVIs on LAG
  - Pinned to each interface
  - Know when a physical port died off
- Routed interface between AR's
  - Exclusively utilise VXLAN



The Design

- East-West Pseudowires
  - Linking AR1/AR2 per-POP
  - Echoes underlying paths
- Resiliency from the ring
  - End-to-end resiliency
  - Tolerates N+1 link failures
  - Purposefully not 1:1 to core



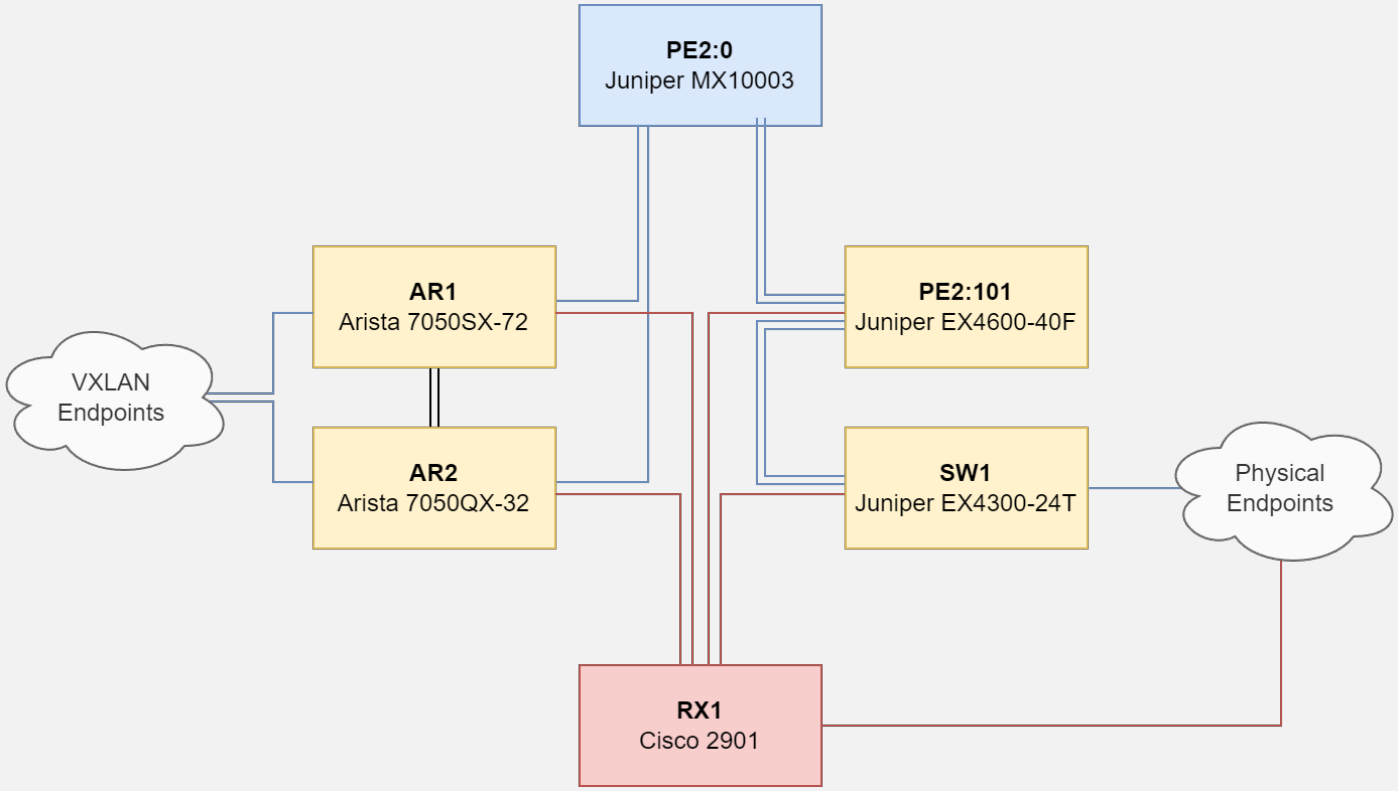
Management

- Utilised Type 5 EVPN
  - Combined both IB and OOB routes
  - Peered to Carrier Core via EBGP in S1
- Moving In-Band to L3VPN on PE's makes sense
  - Full resiliency of MPLS backplane
  - Consistent routing with other IB pathways
  - Broken out to devices via dedicated EX4300-24T switches

Management

- Moving OOB to L3VPN on dedicated routers
  - Using Cisco 29xx with HWIC-16A already
  - L3VPN provided by external third party
- Distribution uses EVPN-VXLAN fabric
  - Allocation of a VXLAN VNI per POP & per state
  - Allows portability of site addressing easily if needed
  - Means we don't need to burn a 40G port for in-band

Management


PE2:0
Juniper MX10003
AR1
Arista 7050SX-72
PE2:101
Juniper EX4600-40F
VXLAN
Endpoints
AR2
Arista 7050QX-32
SW1
Juniper EX4300-24T
Physical
Endpoints
RX1
Cisco 2901

So I get on a chopper.

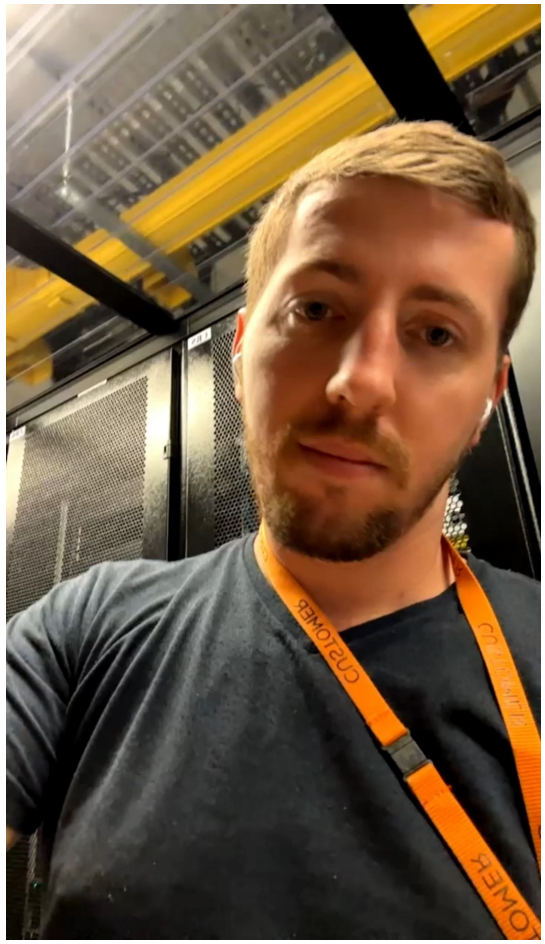
So I get on a ~~chopper~~.

So I get on a plane.

Part 1: The Mess

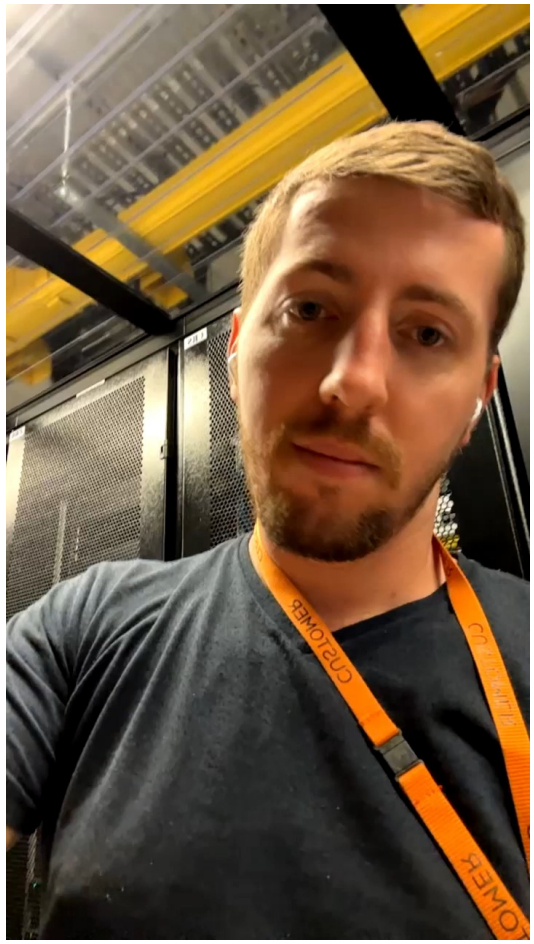
The Mess

- Unstandardised optics
- Inconsistent cabling
- Who even likes Uniboots?
- Overlength patch leads
- No cable management
- Lots of old cabling



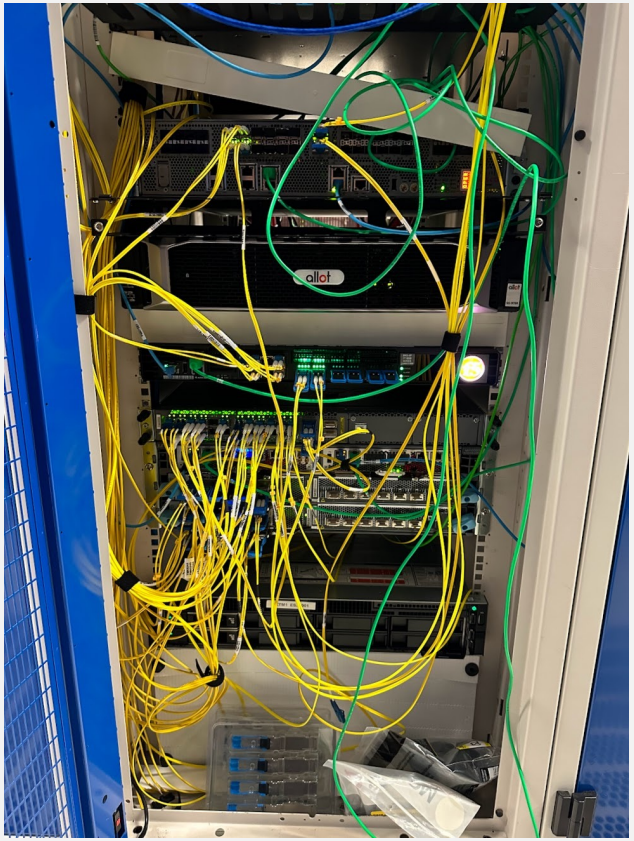
The Mess

- Unstandardised optics
- Inconsistent cabling
- Who even likes Uniboots?
- Overlength patch leads
- No cable management
- Lots of old cabling

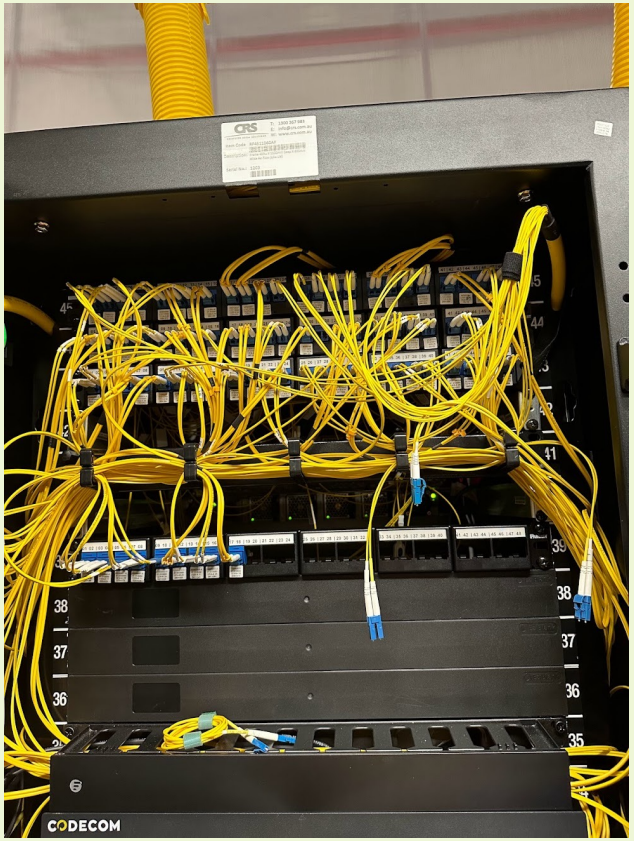


The Mess

Like here...



and here...

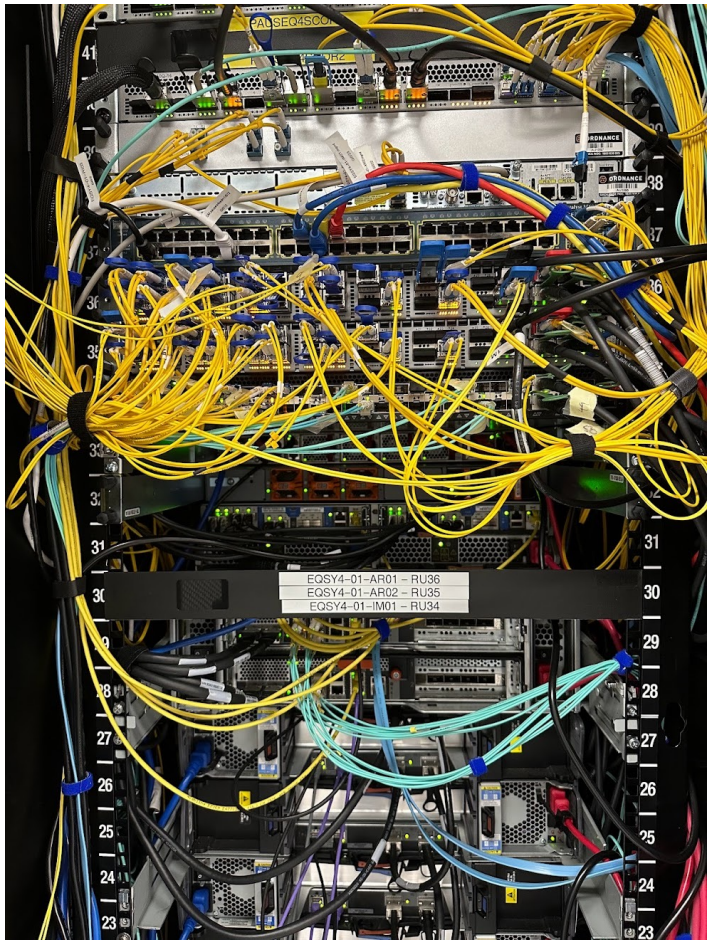


The Mess

- Like I said… lots of old
- Over 100kg of patching removed across all facilities
- Sydney POPs the worst offenders due to over-reliance on facility hands

The Mess

- Technical debt is real
  - Legacy infrastructure
  - Fast integration of acquisitions
  - Organic business growth
- Absolute nightmare for hands
- Service standup issues
- More break, less fix

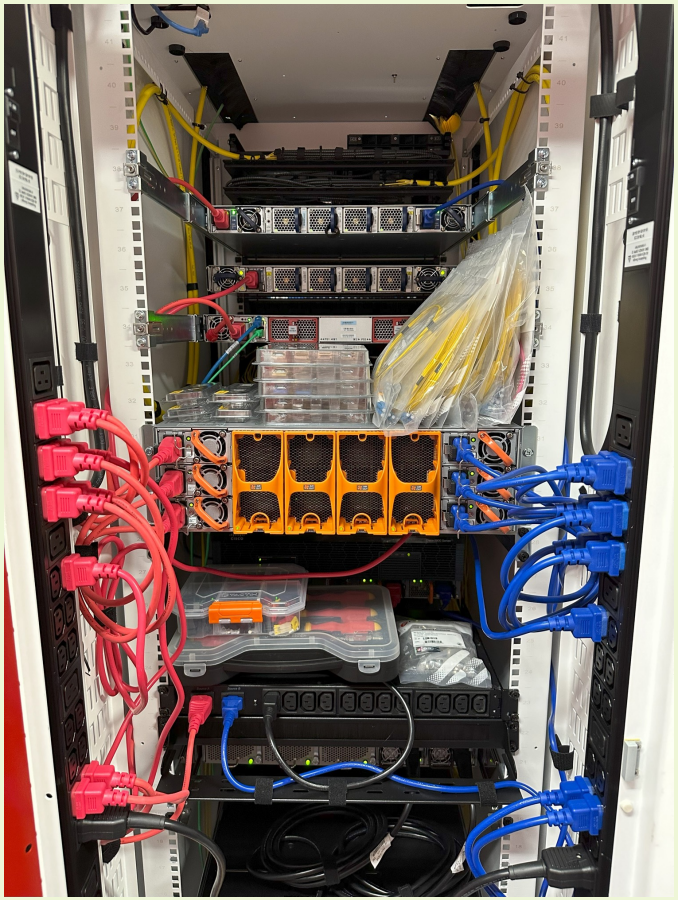


The Mess

- New rack layout designed
- More intuitive format
- Actually use cable managers
- Label all cable runs
- Front-to-Rear airflow only

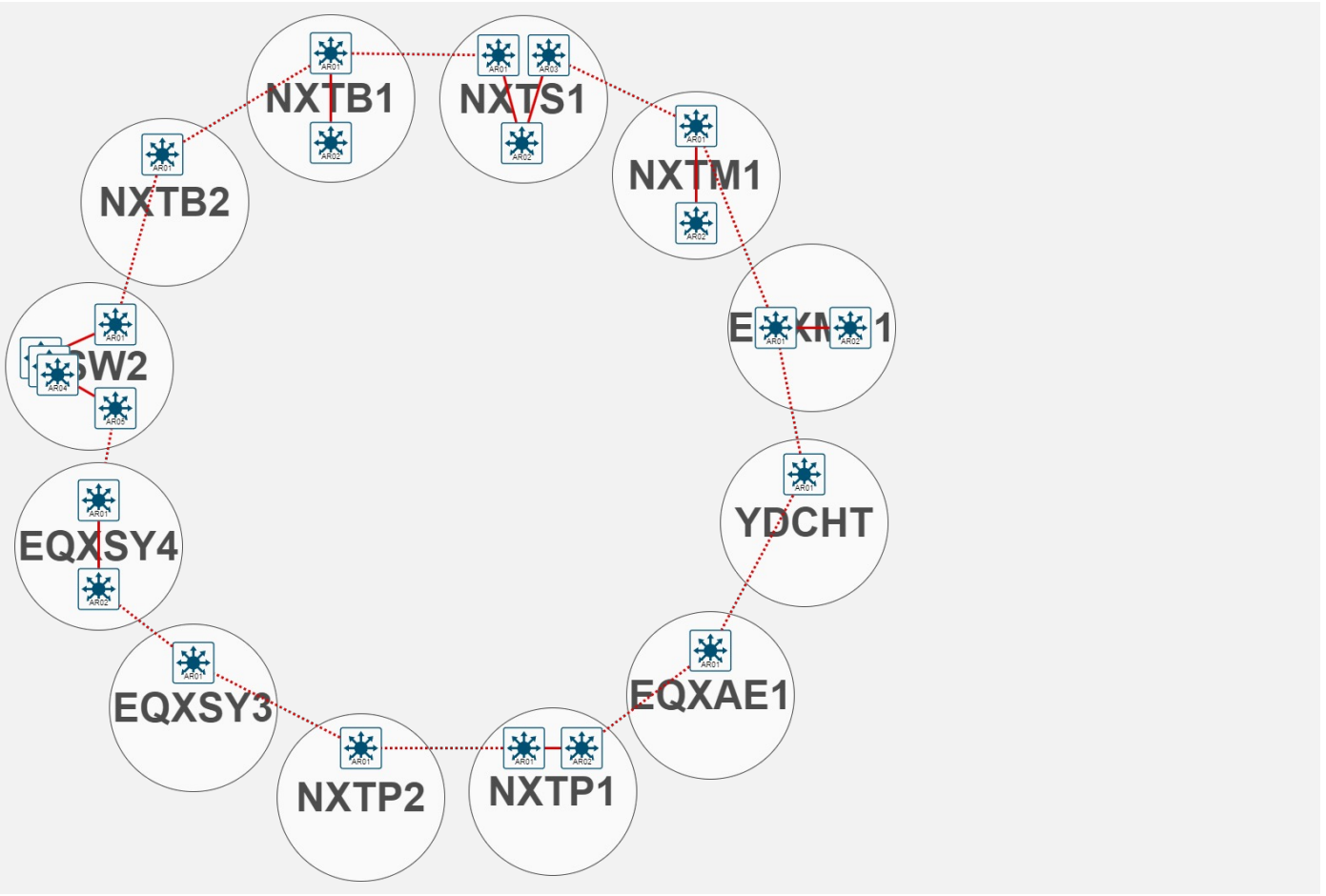
	CORE-01 - Front		CORE-01 - Rear
42	**Structured Cabling** - PP-01	42	
41	**Structured Cabling** - PP-02	41	
40	**Structured Cabling** - PP-03	40	
39	**Structured Cabling** - PP-04	39	
38	**Distribution Switch 01** - Arista 7050SX-64	38	
37	**Cable Management** - 0U Trays w/ Brush	37	**Cable Management** - 0U Trays w/ Brush
36	**Distribution Switch 02** - Arista 7050QX-32S	36	
35	**Cable Management** - 0U Trays w/ Brush	35	**Cable Management** - 0U Trays w/ Brush
34	**Provider Edge 01** - Arista 7280R3	34	
33	**Cable Management** - 0U Trays w/ Brush	33	
32	**Breakout Cabling** - PP05	32	
31	**Provider Edge 02** - Juniper MX10003	31	
30		30	
29		29	
28	**Fabric Switch 01** - Juniper EX4600-40F	28	
27	**Cable Management** - 0U Trays w/ Brush	27	
26	**OOB Router** - Cisco 2901	26	**Cable Management** - 0U Trays w/ Brush
25	**Cable Management** - Finger	25	**ATS** - APC AP4421
24	**Management Switch** - Juniper EX4300-24T	24	**Cable Management** - 0U Trays w/ Brush
23	**Cable Management** - 0U Trays w/ Brush	23	
22	**Provider Core 01** - Cisco NCS-55A1	22	
21	**Cable Management** - 0U Trays w/ Brush	21	
20	**Border Gateway 01** - Cisco ASR9901	20	
19		19	
18	**Cable Management** - 0U Trays w/ Brush	18	
17	**TBD** - Allot SG-9700	17	
16		16	
15	**Cable Management** - 0U Trays w/ Brush	15	
14	**CGNAT Router** - F5 BIG-IP i11800-DS	14	
13	**Cable Management** - 0U Trays w/ Brush	13	
12	**VM Host 01** - Dell R740	12	
11		11	
10	**Cable Management** - 0U Trays w/ Brush	10	
09	Legacy Infrastructure	09	
08		08	
07		07	
06		06	
05		05	
04		04	
03		03	
02		02	
01	**Cable Management** - 0U Trays w/ Brush	01	

The Mess



Part 2: The Ring

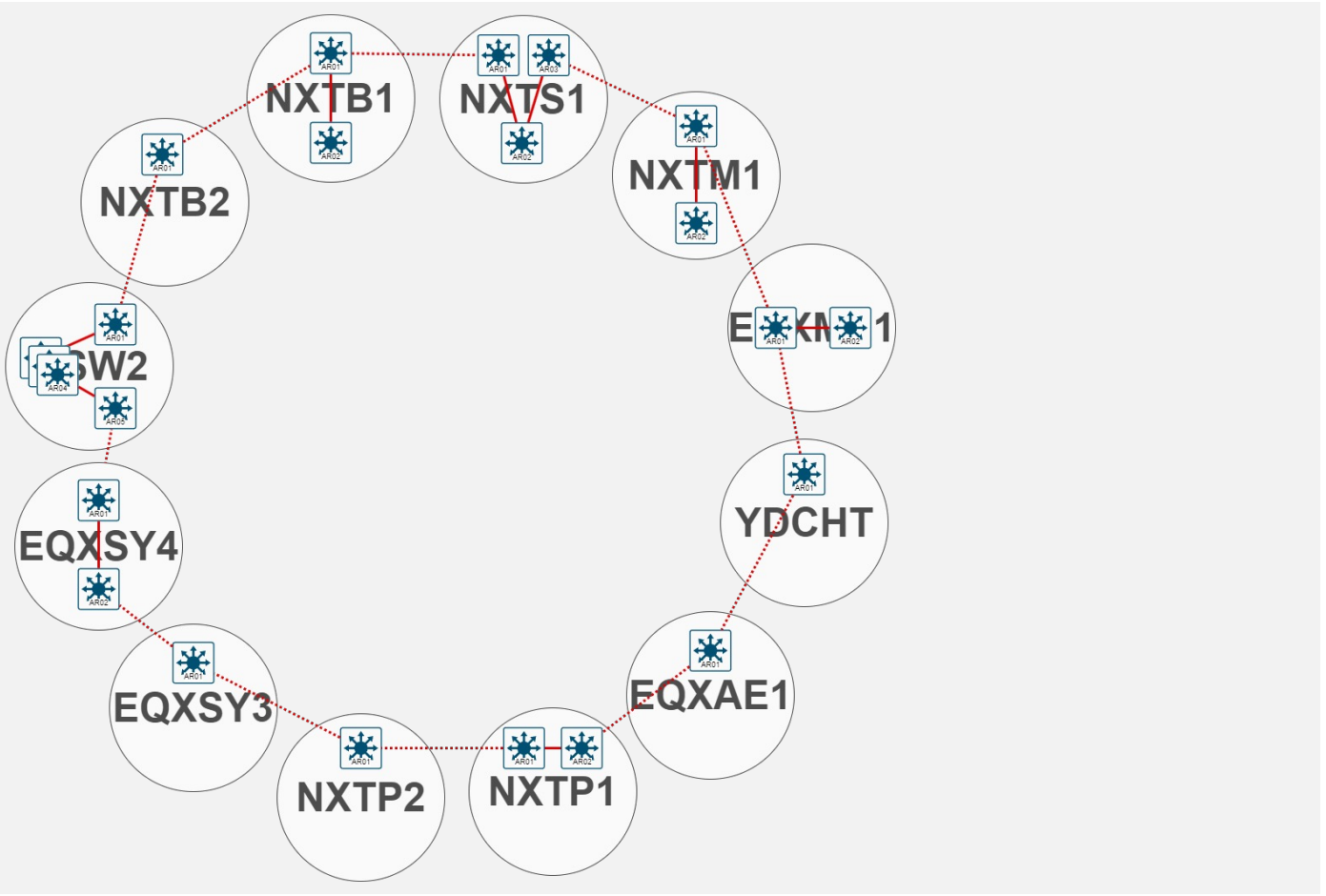
The Ring


NXTB1
NXTS1
NXTM1
NXTB2
YDCHT
EQXSY4
EQXSY3
EQXAE1
NXTP2
NXTP1

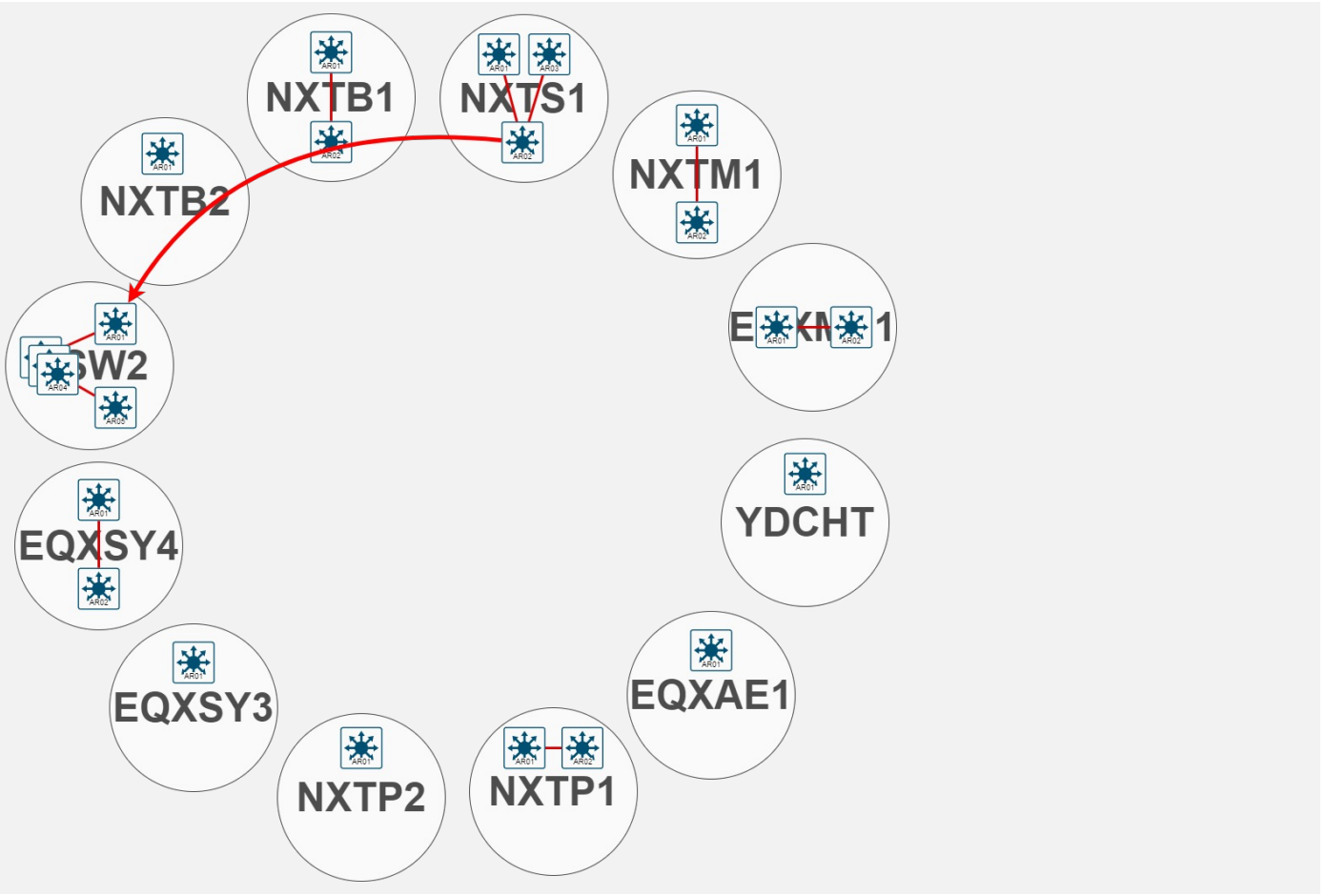
The Ring

- Staged windows to build pseudowires and light BGP
- Utilise peer groups for shared attributes on neighbors
- Enforce clean standards
  - Only VTEPs advertised by adjacencies
  - BFD on all links
- Rework B1, B2, S1, GS, SY3, SY4 POPs smoothly
- But then…

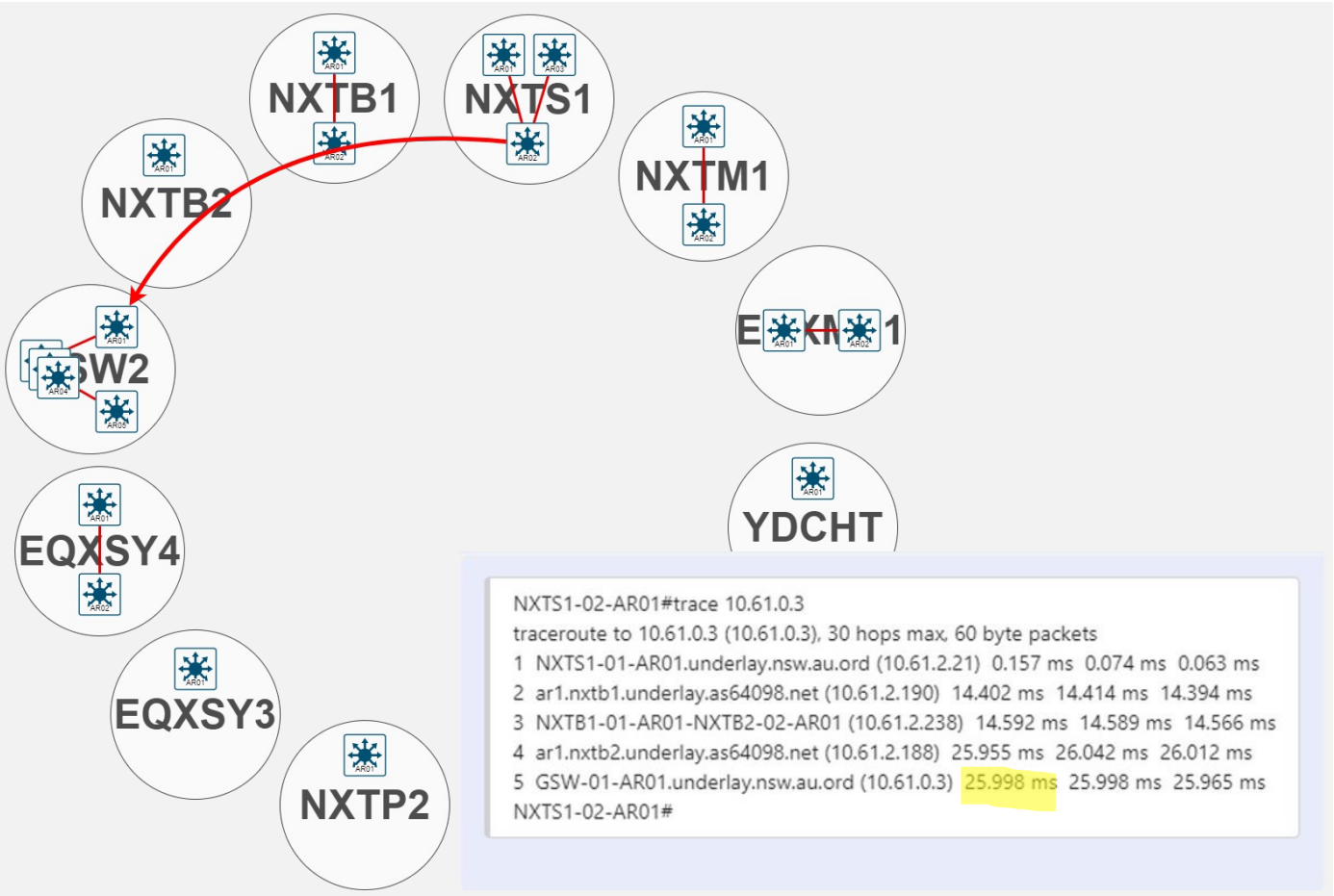
The Ring


NXTB1
NXTS1
NXTM1
NXTB2
YDCHT
EQXSY4
EQXSY3
EQXAE1
NXTP2
NXTP1

The Ring


NXTB1
NXTS1
NXTB2
NXTM1
EQXSY4
YDCHT
EQXSY3
EQXAE1
NXTP2
NXTP1

The Ring


NXTB1
NXTS1
NXTB2
NXTM1
EQXSY4
EQXSY3
NXTP2
YDCHT
NXTS1-02-AR01#trace 10.61.0.3
traceroute to 10.61.0.3 (10.61.0.3), 30 hops max, 60 byte packets
1 NXTS1-01-AR01.underlay.nsw.au.ord (10.61.2.21) 0.157 ms 0.074 ms 0.063 ms
2 ar1.nxtb1.underlay.as64098.net (10.61.2.190) 14.402 ms 14.414 ms 14.394 ms
3 NXTB1-01-AR01-NXTB2-02-AR01 (10.61.2.238) 14.592 ms 14.589 ms 14.566 ms
4 ar1.nxtb2.underlay.as64098.net (10.61.2.188) 25.955 ms 26.042 ms 26.012 ms
5 GSW-01-AR01.underlay.nsw.au.ord (10.61.0.3) 25.998 ms 25.998 ms 25.965 ms
NXTS1-02-AR01#

The Ring

- Rework topology to incorporate state full-mesh
- Complete deployment on M1, ME1, AE1, YourDC, P1, P2
- Now we're cooking with gas!
- Learnings:
  - Ensure the EVPN family is actually in the peer group
  - Arista can be funny with >3 L3 MTU's in config
  - Think of all traffic flows, not just fault domains

Part 3: Management

Management

- Now that's all cleaned up, time to rework our in-band
- L3VPN dropping an SVI off on each site's MX10003
  - Bridge-domain used to drop off via multiple ports
  - Most boxes will grab this via EX4600 fabric extension to a dedicated EX4300 management switch for copper handoff
  - Distribution switches get it via their uplink ports
  - Provide this as a subinterface alongside other services

Management

- Well, that's weird— it doesn't work on the Aristas
  - All other VXLAN traffic seems fine, only management
  - Pull the config back out, try to replicate in the lab
  - Confirm versions match, confirm hardware matches
  - Only thing missing in the lab topology is the pseudo
- So we went back into the datacenter to do it in prod!
  - Well, maybe not quite…

Management

- With a set of 4x unloaded AR's between B1 & B2, we isolated them from the EVPN-VXLAN topology and got to testing...
- Decide to use the NEXTDC B1 In-Band VLAN to test
  - B1-B1 traffic flows work fine
  - Stretch the subnet over to B2 via VNI, but no love
- Pivot to a test SVI as this testing isolated B1 in-band

Management

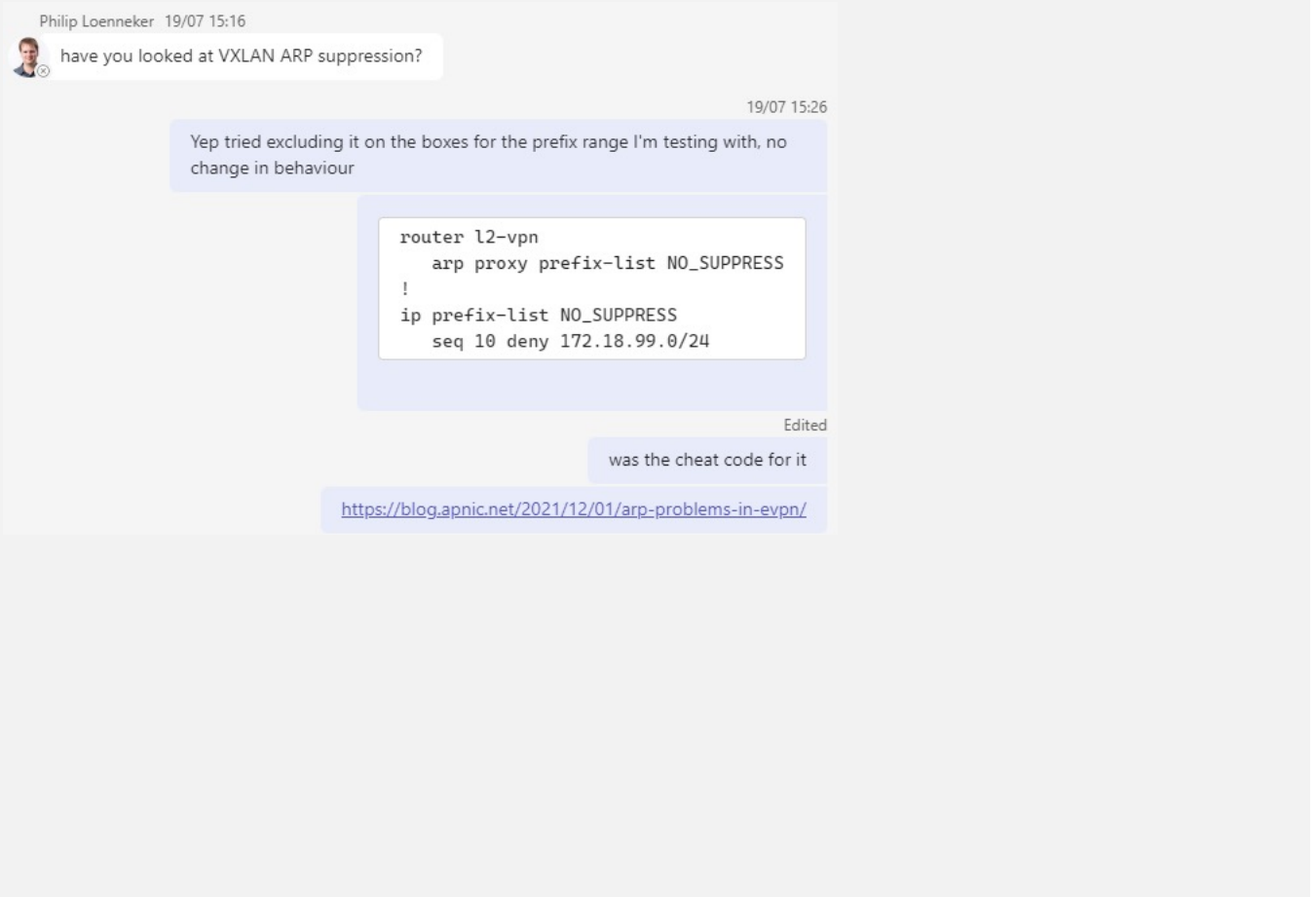
```
AR4.NXTB1#ping vrf FSG-MGMT-INB 172.18.99.21 repeat 10000 int 0.001
PING 172.18.99.21 (172.18.99.21) 72(100) bytes of data.

--- 172.18.99.21 ping statistics ---
10000 packets transmitted, 0 received, 100% packet loss, time 100894ms

AR4.NXTB1#
```

```
AR1.NXTB2#show int et47
Ethernet47 is up, line protocol is up (connected)
  Hardware is Ethernet, address is 444c.a803.8344 (bia 444c.a803.8344)
  Description: PTP: PE4.NXTB2 Xe-0/1/6
  Ethernet MTU 9214 bytes, BW 10000000 kbit
  Full-duplex, 10Gb/s, auto negotiation: off, uni-link: n/a
  Up 3 hours, 38 minutes, 20 seconds
  Loopback Mode : None
  0 link status changes since last clear
  Last clearing of "show interface" counters 0:02:05 ago
  5 minutes input rate 0 bps (0.0% with framing overhead), 0 packets/sec
  5 minutes output rate 0 bps (0.0% with framing overhead), 0 packets/sec
     11420 packets input, 1911985 bytes
     Received 0 broadcasts, 66 multicast
     0 runts, 0 giants
     0 input errors, 0 CRC, 0 alignment, 0 symbol, 0 input discards
     0 PAUSE input
     924 packets output, 74184 bytes
     Sent 17 broadcasts, 374 multicast
     0 output errors, 0 collisions
     0 late collision, 0 deferred, 0 output discards
     0 PAUSE output
AR1.NXTB2#
```

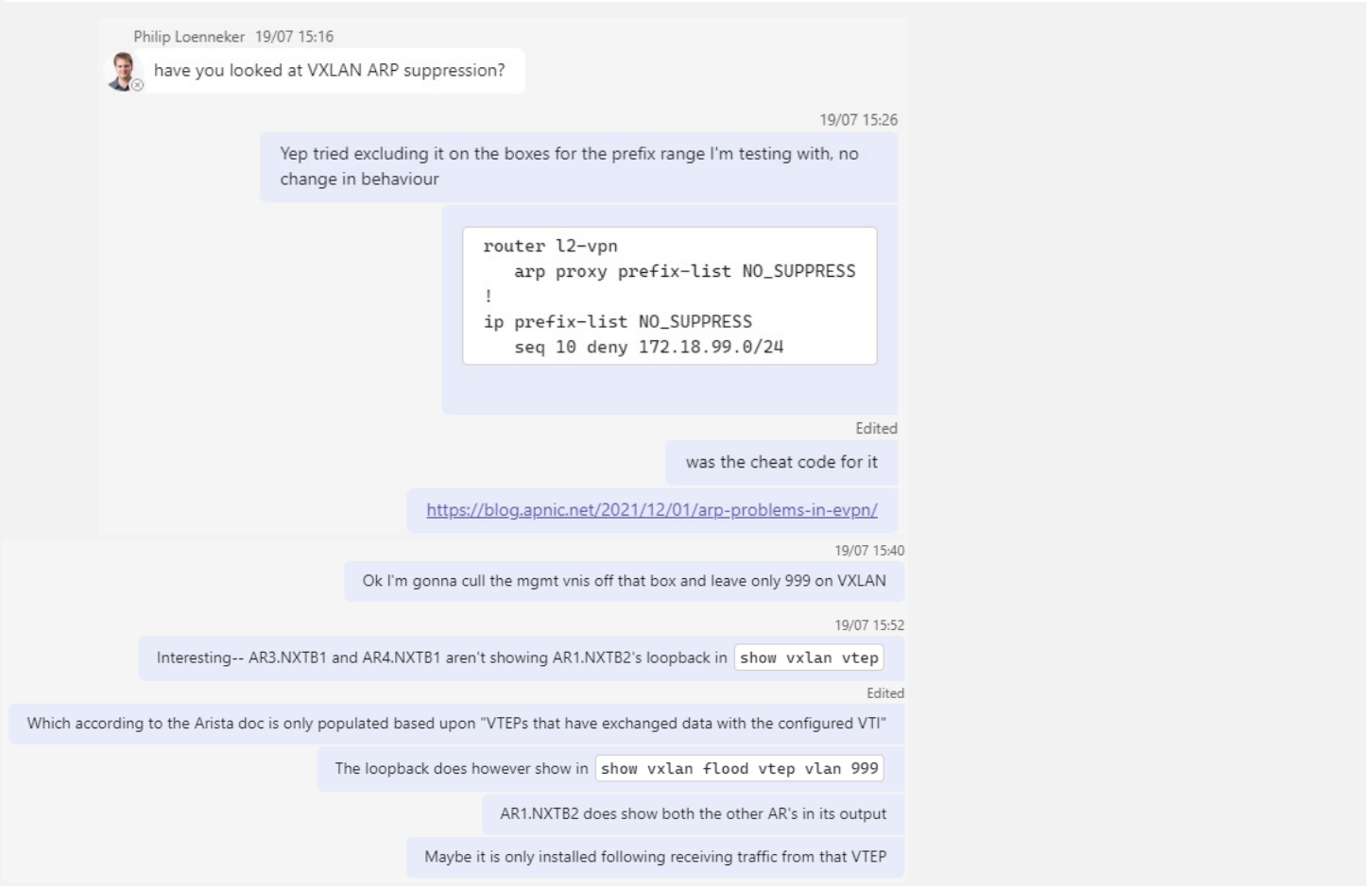
Looks like packets are getting to AR1.NXTB2

That's on a freshly cleared counter, also has BGP+BFD+VXLAN signalling traffic across it

Management

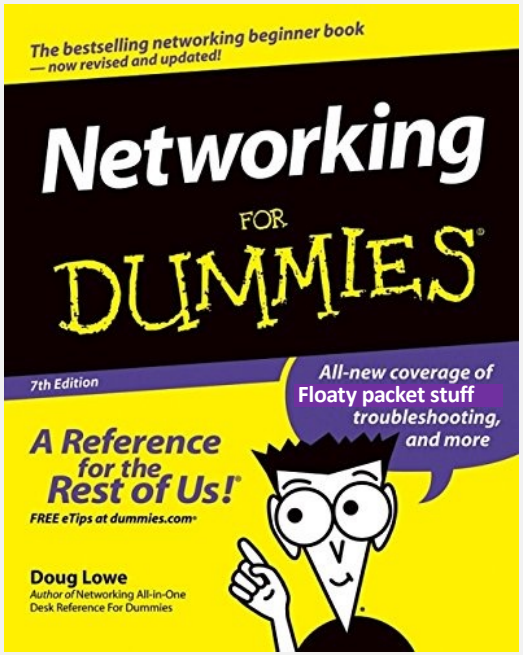

Philip Loenneker 19/07 15:16
have you looked at VXLAN ARP suppression?
19/07 15:26
Yep tried excluding it on the boxes for the prefix range I'm testing with, no change in behaviour
router l2-vpn
arp proxy prefix-list NO_SUPPRESS
!
ip prefix-list NO_SUPPRESS
seq 10 deny 172.18.99.0/24
Edited
was the cheat code for it
https://blog.apnic.net/2021/12/01/arp-problems-in-evpn/

Management


Philip Loenneker 19/07 15:16
have you looked at VXLAN ARP suppression?
19/07 15:26
Yep tried excluding it on the boxes for the prefix range I'm testing with, no change in behaviour
router l2-vpn
arp proxy prefix-list NO_SUPPRESS
!
ip prefix-list NO_SUPPRESS
seq 10 deny 172.18.99.0/24
Edited
was the cheat code for it
https://blog.apnic.net/2021/12/01/arp-problems-in-evpn/
19/07 15:40
Ok I'm gonna cull the mgmt vnis off that box and leave only 999 on VXLAN
19/07 15:52
Interesting-- AR3.NXTB1 and AR4.NXTB1 aren't showing AR1.NXTB2's loopback in show vxlan vtep
Edited
Which according to the Arista doc is only populated based upon "VTEPs that have exchanged data with the configured VTI"
The loopback does however show in show vxlan flood vtep vlan 999
AR1.NXTB2 does show both the other AR's in its output
Maybe it is only installed following receiving traffic from that VTEP

Management

- This is probably a good point to brush up on VXLAN
  - L2 frames are encapsulated with a VXLAN header
  - This header specifies a unique VNI
  - The VNI identifies the unique layer 2 domain which was present in the VLAN
  - VXLAN packet is then sent via IP/UDP to any participating endpoints
  - These endpoints are called VTEPs



Management

- Upon receiving, the VTEP decapsulates the packet and associates it to a VLAN
- This is done based upon its local VLAN:VNI binding table
- The traffic is then forwarded to the host as normal
- VTEPs can participate in a VNI either via static flooding or by utilising a control protocol such as EVPN
- Now that's done... what do we see?

Management

```
AR1.NXTB2#show vxlan counters vni 10999
                                                                  Decap Drop Or
                              Decap Known             Decap BUM       Exception
VNI          Decap Bytes      Unicast Packets           Packets         Packets
---------- ---------------- ------------------- ------------------- -------------
10999                   0                   0                   0             0

                                                      Encap BUM      Encap Drop
VNI          Encap Bytes      Encap Packets             Packets         Packets
---------- ---------------- ------------------- ------------------- -------------
10999                   0                   0                   3             6

AR1.NXTB2#
```

```
AR4.NXTB1#show vxlan count vni 10999
                                                                  Decap Drop Or
                              Decap Known             Decap BUM       Exception
VNI          Decap Bytes      Unicast Packets           Packets         Packets
---------- ---------------- ------------------- ------------------- -------------
10999                   0                   0                   0             3

                                                      Encap BUM      Encap Drop
VNI          Encap Bytes      Encap Packets             Packets         Packets
---------- ---------------- ------------------- ------------------- -------------
10999                   0                   0                   0             0

AR4.NXTB1#
```

So packets do get encapped & decapped at both sides going AR1.NXTB2 -> AR4.NXTB1, but the reverse direction is only encapped.. but never decapped

Management

```
AR4.NXTB1#clear vxlan count vni
AR4.NXTB1#show vxlan count vni 10999
                                                                  Decap Drop Or
                                  Decap Known            Decap BUM       Exception
VNI          Decap Bytes      Unicast Packets              Packets         Packets
---------- ---------------- -------------------- -------------------- -------------
10999                  0                    0                    0             0

                                                        Encap BUM      Encap Drop
VNI          Encap Bytes      Encap Packets              Packets         Packets
---------- ---------------- -------------------- -------------------- -------------
10999                  0                    0                    0             0

AR4.NXTB1#ping vrf FSG-MGMT-INB 172.18.99.21
PING 172.18.99.21 (172.18.99.21) 72(100) bytes of data.

--- 172.18.99.21 ping statistics ---
5 packets transmitted, 0 received, 100% packet loss, time 40ms

AR4.NXTB1#show vxlan count vni 10999
                                                                  Decap Drop Or
                                  Decap Known            Decap BUM       Exception
VNI          Decap Bytes      Unicast Packets              Packets         Packets
---------- ---------------- -------------------- -------------------- -------------
10999                  0                    0                    0             0

                                                        Encap BUM      Encap Drop
VNI          Encap Bytes      Encap Packets              Packets         Packets
---------- ---------------- -------------------- -------------------- -------------
10999               1720                   10                    5             0

AR4.NXTB1#
```

Which I guess is consistent with the packets in that direction incrementing counters on the pseudo both sides towards AR1.NXTB2, and incrementing the physical int counter on AR1.NXTB2 Et47

But seems to just.. never hit the VXLAN interface there

Nothing being caught in the 'unlearned' counter

Management

- So I decide to dig into exactly what the drops mean
- And I find this…

> ***"VxLAN encap counters count packets as successfully encapsulated even if their size is larger than the egress MTU size. If the packet's final size (i.e. with the encapsulated header) is larger than the maximum jumbo frame size (9214 bytes) the packet gets counted as an encap drop or exception packet on DCS-7050X,DCS-7250X and DCS-7300X series switches."***
>
> - *Some guy at Arista, probably*

- Great, just a stupid error in MTU!

Management

- So we double check everything. *EVERYTHING*.
- All physical interfaces set to 9214 L2 MTU
- Pseudowires set to 9100 MTU
- East-West logical interfaces set to 9000 IP MTU
- SVI on all devices set to 1500 bytes
- What the hell?

Management

- Okay, time to ask for help
- One problem— no support on these boxes
  - Luckily, we're not dealing with Cisco
  - So we go to Turbo and plead our case
  - He calls in Aaron Chan from the ASE team
- 35 emails, five meetings and a 3hr TAC call later...

Management

Security Advisory 0055

PDF

Date:December 16th, 2020

Version: 1.0

Revision	Date	Changes
1.0	December 16th, 2020	Initial Release

The CVE-ID tracking this issue: CVE-2020-26568
CVSSv3.1 Base Score: 5.3 (CVSS:3.1/AV:N/AC:L/PR:N/UI:N/S:U/C:N/I:L/A:N)

Description

This advisory documents the impact of a vulnerability in Arista's EOS for device configurations leveraging VxLAN Routing and VRFs. To evaluate if a VxLAN enabled device is vulnerable, please see the "Symptoms" section below for details.

On impacted devices, malformed packets could be incorrectly forwarded across VRF boundaries when non-default VRFs are configured. This issue affects UDP traffic, and will fail to complete the three-way handshake for TCP traffic.

Please note that this advisory does not refer to the crossing of VRF boundaries as a result of the configuration of inter-VRF routing (which would be the expected behavior).

This issue was discovered internally and Arista is not aware of any malicious uses of this issue in customer networks.

Exhibit A: *TAC being TAC, surely this isn't relevant (Michael Hobl, 2023)*

Management

For customers whose network design leverages VxLAN decapsulation on an interface that carries traffic for multiple VRFs, the following additional steps may be required post EOS upgrade.

Once an upgrade to a release with the fix has been completed, the following warnings may be logged under "show logging all":

```
%VXLAN-4-DECAPSULATION_DISABLED: VXLAN decapsulation has been disabled on
Ethernet48 because it carries both default VRF and non-default VRF traffic
```

```
%VXLAN-4-DECAPSULATION_DISABLED: VXLAN decapsulation has been disabled on Ethernet48 because it carries non-default VRF traffic
To allow VXLAN decapsulation on interfaces that carry both default VRF and non-default VRF traffic issue the command: 'vxlan decapsulation filt
er interface multiple-vrf disabled'.
To entirely disable VRF-based VXLAN decapsulation filtering on this switch/router, configure 'vxlan decapsulation filter disabled'.
```

If the above warnings have been observed, it indicates that VxLAN decapsulation has been disabled on the listed interfaces (for example, in the above case VxLAN decapsulation has been disabled on ethernet48).

- So I check the boxes, nothing.
- But the test VLAN does drop off in a VRF, so we put the config on anyway. No change. Reboot just incase.

Management

AR1.NXTB1#show version | i uptime

Uptime: 6 minutes

AR1.NXTB1#show logging | grep -i vxlan-

Aug 10 15:39:42 AR1 StrataL2: %VXLAN-4-DECAPSULATION_DISABLED: VXLAN decapsulation has been disabled on Ethernet49/9 because it carries non-default VRF traffic

Exhibit B: *Wait, what? (Michael Hobl, 2023)*

Management

Oh, I configured the wrong interface.

Management

AR1.NXTB1(config-if-Vx1)#show active

interface Vxlan1

description VXLAN Interface

vxlan source-interface Loopback0

vxlan udp-port 4789

vxlan vlan 999 vni 10999

vxlan decapsulation filter interface multiple-vrf disabled Ethernet49/1

vxlan vlan 999 flood vtep 10.61.0.41 10.61.0.42 10.61.0.43

AR1.NXTB1(config-if-Vx1)#vxlan decapsulation filter interface multiple-vrf disabled Ethernet49/9

AR1.NXTB1(config-if-Vx1)#show logging | grep -i vxlan-

Aug 10 15:39:42 AR1 StrataL2: %VXLAN-4-DECAPSULATION_DISABLED: VXLAN decapsulation has been disabled on Ethernet49/9 because it carries non-default VRF traffic

Aug 10 15:42:00 AR1 StrataL2: %VXLAN-6-DECAPSULATION_ENABLED: VXLAN decapsulation has been enabled on Ethernet49/9

Management

```
AR1.NXTB1#show arp
Address         Age (sec)  Hardware Addr   Interface
10.61.3.7         0:00:13  001c.73b1.29ff  Ethernet47
10.61.2.238       0:00:36  444c.a803.8315  Vlan50, Port-Channel31
10.61.2.191       0:00:48  001c.73a2.440d  Vlan51, Port-Channel31
10.61.3.241       0:00:25  001c.73b2.89e9  Vlan434, Port-Channel31
AR1.NXTB1#ping 172.18.99.12
PING 172.18.99.12 (172.18.99.12) 72(100) bytes of data.
80 bytes from 172.18.99.12: icmp_seq=1 ttl=64 time=1.02 ms
80 bytes from 172.18.99.12: icmp_seq=2 ttl=64 time=0.201 ms
80 bytes from 172.18.99.12: icmp_seq=3 ttl=64 time=0.308 ms
80 bytes from 172.18.99.12: icmp_seq=4 ttl=64 time=0.294 ms
^C
```

Management

Management

- So, it turns out this log entry is only added once
- It hadn't shown up as the boxes were configured a month earlier, triggering this output then
- Logging fail by FSG, but documentation fail by Arista
- Nonetheless, they got us out of a real pickle

Management

- Rip out test SVI, reincorporate per-POP in-band VNI
- B1 works as expected, test haul B1 to B2 via VXLAN
- Works as anticipated, deploy changes to all AR's
- A HUGE thanks to the team at Arista for the help
- One piece left, multihoming.

Part 4: Multihoming

EVPN ESI

- All AR-PE adjacencies are still single-homed on physical interfaces, with the redundant pairs ready
- Cutover windows booked to migrate to LACP bonds
- First site up is NEXTDC B2, with a new AR pair used internally for a new compute cluster— perfect.
  - Hot Tip: Customers are way fussier than the server guys!
- Light up first AR pair with a week's gap to validate

EVPN ESI

- Signs look good, ESI has propagated correctly
- DF election has taken place, fine on default timers

```
AR1.NXTB1#show bgp evpn instance vlan 100
EVPN instance: VLAN 100
  Route distinguisher: 0:0
  Route target import: Route-Target-AS:1234:10001017
  Route target export: Route-Target-AS:1234:10001017
  Service interface: VLAN-based
  Local IP address: 10.61.0.41
  Encapsulation type: VXLAN
  Local ethernet segment:
    ESI: 00aa:bbcc:ddee:ff00:3100
      Interface: Ethernet16
      Mode: single-active
      State: up
      ES-Import RT: aa:bb:cc:dd:ee:ff
      DF election algorithm: modulus
      Designated forwarder: 10.61.0.42
      Non-Designated forwarder: 10.61.0.41
AR1.NXTB1#
```

```
AR1.NXTB1#show bgp evpn route-type mac-ip 172.18.17.1
BGP routing table information for VRF default
Router identifier 10.61.0.41, local AS number 65044
Route status codes: s - suppressed, * - valid, > - active, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP, b - backup
                    % - Pending BGP convergence
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST - Cluster List, LL Nexthop - Link Local
Nexthop

          Network                Next Hop              Metric  LocPref Weight  Path
* >     RD: 10.61.0.41:100 mac-ip c8e7.f0f9.1c81 172.18.17.1
                                 -                     -       -       0       i
* >Ec   RD: 10.61.0.42:100 mac-ip c8e7.f0f9.1c81 172.18.17.1
                                 10.61.0.42            -       100     0       65045 i
*  ec   RD: 10.61.0.42:100 mac-ip c8e7.f0f9.1c81 172.18.17.1
                                 10.61.0.42            -       100     0       65045 i
AR1.NXTB1#
```

EVPN ESI

- Signs look good, ESI has propagated correctly
- DF election has taken place, fine on default timers

```
AR1.NXTB1#show bgp evpn instance vlan 100
EVPN instance: VLAN 100
  Route distinguisher: 0:0
  Route target import: Route-Target-AS:1234:10001017
  Route target export: Route-Target-AS:1234:10001017
  Service interface: VLAN-based
  Local IP address: 10.61.0.41
  Encapsulation type: VXLAN
  Local ethernet segment:
    ESI: 00aa:bbcc:ddee:ff00:3100
      Interface: Ethernet16
      Mode: single-active
      State: up
      ES-Import RT: aa:bb:cc:dd:ee:ff
      DF election algorithm: modulus
      Designated forwarder: 10.61.0.42
      Non-Designated forwarder: 10.61.0.41
AR1.NXTB1#
```

```
AR1.NXTB1#show bgp evpn route-type mac-ip 172.18.17.1
BGP routing table information for VRF default
Router identifier 10.61.0.41, local AS number 65044
Route status codes: s - suppressed, * - valid, > - active, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP, b - backup
                    % - Pending BGP convergence
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST - Cluster List, LL Nexthop - Link Local
Nexthop

          Network                Next Hop              Metric  LocPref Weight  Path
* >     RD: 10.61.0.41:100 mac-ip c8e7.f0f9.1c81 172.18.17.1
                                 -                     -       -       0       i
* >Ec   RD: 10.61.0.42:100 mac-ip c8e7.f0f9.1c81 172.18.17.1
                                 10.61.0.42            -       100     0       65045 i
*  ec   RD: 10.61.0.42:100 mac-ip c8e7.f0f9.1c81 172.18.17.1
                                 10.61.0.42            -       100     0       65045 i
AR1.NXTB1#
```

EVPN ESI

- Signs look good, ESI has propagated correctly
- DF election has taken place, fine on default timers

```
AR1.NXTB1#show bgp evpn instance vlan 100
EVPN instance: VLAN 100
  Route distinguisher: 0:0
  Route target import: Route-Target-AS:1234:10001017
  Route target export: Route-Target-AS:1234:10001017
  Service interface: VLAN-based
  Local IP address: 10.61.0.41
  Encapsulation type: VXLAN
  Local ethernet segment:
    ESI: 00aa:bbcc:ddee:ff00:3100
      Interface: Ethernet16
      Mode: single-active
      State: up
      ES-Import RT: aa:bb:cc:dd:ee:ff
      DF election algorithm: modulus
      Designated forwarder: 10.61.0.42
      Non-Designated forwarder: 10.61.0.41
AR1.NXTB1#
```

```
AR1.NXTB1#show bgp evpn route-type mac-ip 172.18.17.1
BGP routing table information for VRF default
Router identifier 10.61.0.41, local AS number 65044
Route status codes: s - suppressed, * - valid, > - active, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP, b - backup
                    % - Pending BGP convergence
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST - Cluster List, LL Nexthop - Link Local
Nexthop

          Network                Next Hop              Metric  LocPref Weight  Path
* >     RD: 10.61.0.41:100 mac-ip c8e7.f0f9.1c81 172.18.17.1
                                 -                     -       -       0       i
* >Ec   RD: 10.61.0.42:100 mac-ip c8e7.f0f9.1c81 172.18.17.1
                                 10.61.0.42            -       100     0       65045 i
*  ec   RD: 10.61.0.42:100 mac-ip c8e7.f0f9.1c81 172.18.17.1
                                 10.61.0.42            -       100     0       65045 i
AR1.NXTB1#
```

- But you know where this is going...

EVPN ESI

Jul 3 04:17:38 AR1 PortSec: %ETH-4-HOST_FLAPPING: Host 38:68:dd:4b:ca:85 in VLAN 100 is flapping between interface Port-Channel31 and interface Ethernet46 (message repeated 1 times in 131.631 secs)
Jul 3 04:18:05 AR1 PortSec: %ETH-4-HOST_FLAPPING: Host 38:68:dd:4b:c5:b5 in VLAN 100 is flapping between interface Port-Channel31 and interface Ethernet46 (message repeated 1 times in 26.8495 secs)
Jul 3 04:20:02 AR1 PortSec: %ETH-4-HOST_FLAPPING: Host 38:68:dd:4b:c5:b5 in VLAN 100 is flapping between interface Port-Channel31 and interface Ethernet46 (message repeated 1 times in 117 secs)
Jul 3 04:23:51 AR1 PortSec: %ETH-4-HOST_FLAPPING: Host 38:68:dd:4b:c5:b5 in VLAN 100 is flapping between interface Port-Channel31 and interface Ethernet46 (message repeated 1 times in 229.26 secs)
Jul 3 04:25:52 AR1 PortSec: %ETH-4-HOST_FLAPPING: Host 38:68:dd:4b:c5:b5 in VLAN 100 is flapping between interface Port-Channel31 and interface Ethernet46 (message repeated 3 times in 120.45 secs)
Jul 3 04:26:38 AR1 PortSec: %ETH-4-HOST_FLAPPING: Host 38:68:dd:4b:ca:85 in VLAN 100 is flapping between interface Port-Channel31 and interface Ethernet46 (message repeated 1 times in 46.1398 secs)

Tim Hynard 03/07 14:46

not sure why that mac is flapping, its ilo should only be on sw2.nxtb2

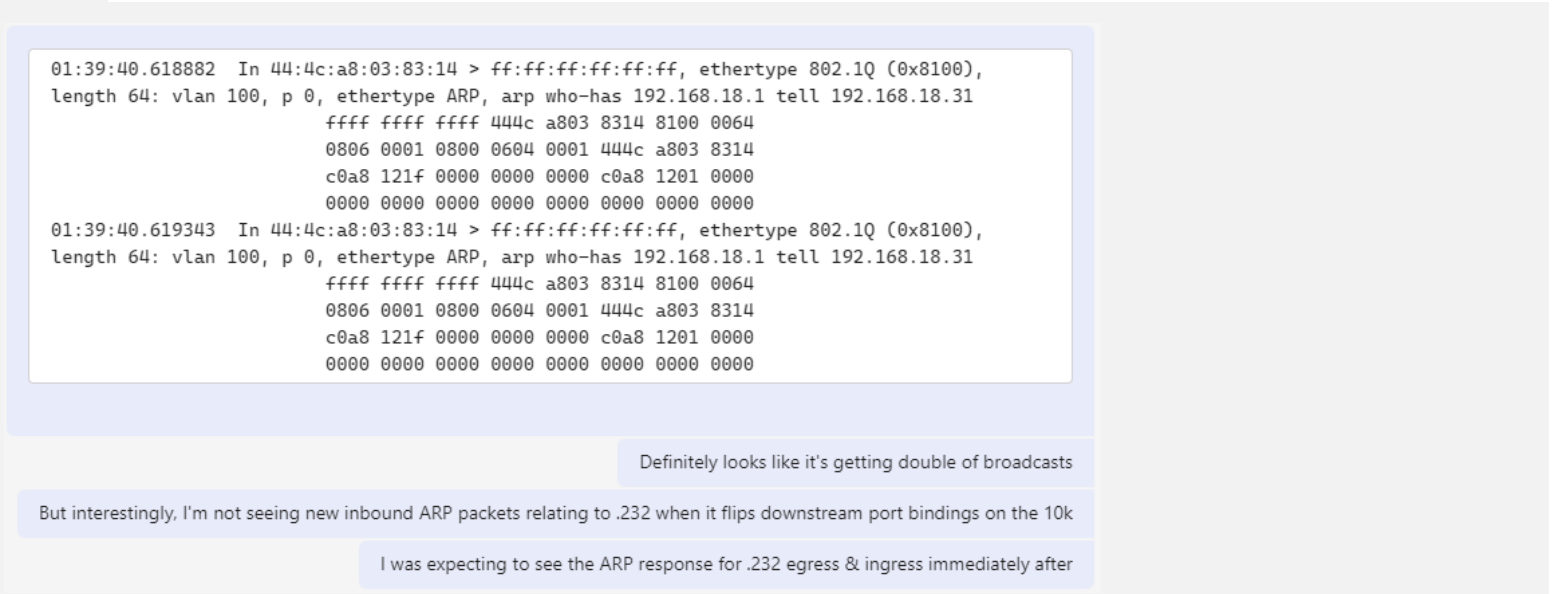
03/07 14:46

Yeah I didn't think there was another path on sw2

Tim Hynard 03/07 14:56

I reckon the macs are being sent back up the interface from ar3/4

EVPN ESI



- Only affecting BUM traffic
- But the DF election was successful?
- Time to hit the documentation...

EVPN ESI

- And there was nothing.
- Every resource, config excerpts, nothing.

EVPN ESI

- Now, we knew from the start that this was unsupported functionality on our platform
- We chose to forge ahead because it looked like it worked fine in the lab
- We had a couple of back-up options to consider…
  - MLAG via VTEP VIPs
  - EVPN Active-Passive

EVPN ESI

EVPN Active-Passive

Pros:

- Partially implemented in most locations via EVPN signaling
- Simple migration strategy to EVPN A-A multihoming when hardware is upgraded
- Able to failover cross-POP easily
- Very little added complexity to assure solutions

Cons:

- Unable to aggregate bandwidth across member ports
- Not supported formally by Arista in A-A, uncertain about A-P

MLAG / VTEP-VIP

Pros:

- Able to aggregate bandwidth across member ports
- Formally supported by Arista

Cons:

- Cannot failover cross-POP due to use of peer links
- Additional complexity required for configuration as every switch pair must have a VTEP VIP created and introduced into underlay fabric
- Requires rework of BGP deployment to implement underlay/overlay functions

EVPN ESI

- EVPN Active-Active is dead, long live the king
- We lab EVPN Active-Passive, to see if it works
- We keep almost all desired business functionality
- One-liner change required in base aggregate config
- No ESI configuration needed towards PE's
- All jets are go.

So What's Left?

What's Left?

- We still have migrations from legacy gear in progress
  - Change control is slow
  - Customers are even slower
  - Hoping to have this sorted by 2024
- While all POPs have 10G & 40G, not all boxes are final
  - Some service migrations needed from old racks to new
  - This is mostly legacy 40G handoffs, or 4x10G PLR optics

What's Left?

- Out-of-Band deployment is still in progress
  - Provider had an allergy to providing LOAs
  - ETA end September
- One final Tour de POP needed to finalize cabling
  - Labelling inconsistencies due to different hardware
  - Some cabling inconsistencies due to delays with lacing bars
  - ETA end October

Special Thanks

To	From
Murray Southwell	Field Solutions Group
Rowan Sakul	Field Solutions Group
Tim Hynard	Field Solutions Group
Philip Loenneker	Field Solutions Group
Matthew Thurbon	Arista Networks
Aaron Chan	Arista Networks
Nikhitha Bandlamudi	Arista Networks
Prateek Holla	Arista Networks
Lei Xu	Arista Networks
The BBL Crew	Various

Questions