DCI

Arnaud Fenioux

DataCenter Interconnection / Infrastructure

Attention!

The slides presented here after are more than inspired by orignal content listed at the end of this presentation.

Thanks to ipspace.net and ciscolive.com

What is DCI?

DataCenter ***Interconnection***

Or

DataCenter ***Infrastructure***?

From interconnection to infrastructure

- Interconnection
  - Dark fiber
  - Wavelength Division Multiplexing (WDM)
  - MPLS pseudowires : Virtual Private Wire Service (VPWS)
- Infrastructure
  - Intra DataCenter :
    - Multi-Chassis Link Aggregation (MLAG)
    - TRILL / SPB / Overlay transport virtualization (OTV)
  - L3-VPN : MPLS / VRF
  - L2-VPN :
    - Virtual Private LAN Service (VPLS) implemented with MPLS pseudowires
    - EVPN with VXLAN/MPLS (BGP Top Of Rack)

L1 interconnection - Dark Fiber

- Rent a dark fiber
- Check optical budget (-0.3dBm /km)
- Put proper SFP on both sides
  - Plug and Swap fibers (Tx/Rx)
  - Configure both port on same VLAN
- Done

L1 interconnection - WDM

- Add multiplexers to the previous setup
- Put coloured SFP on both sides or active WDM equipment

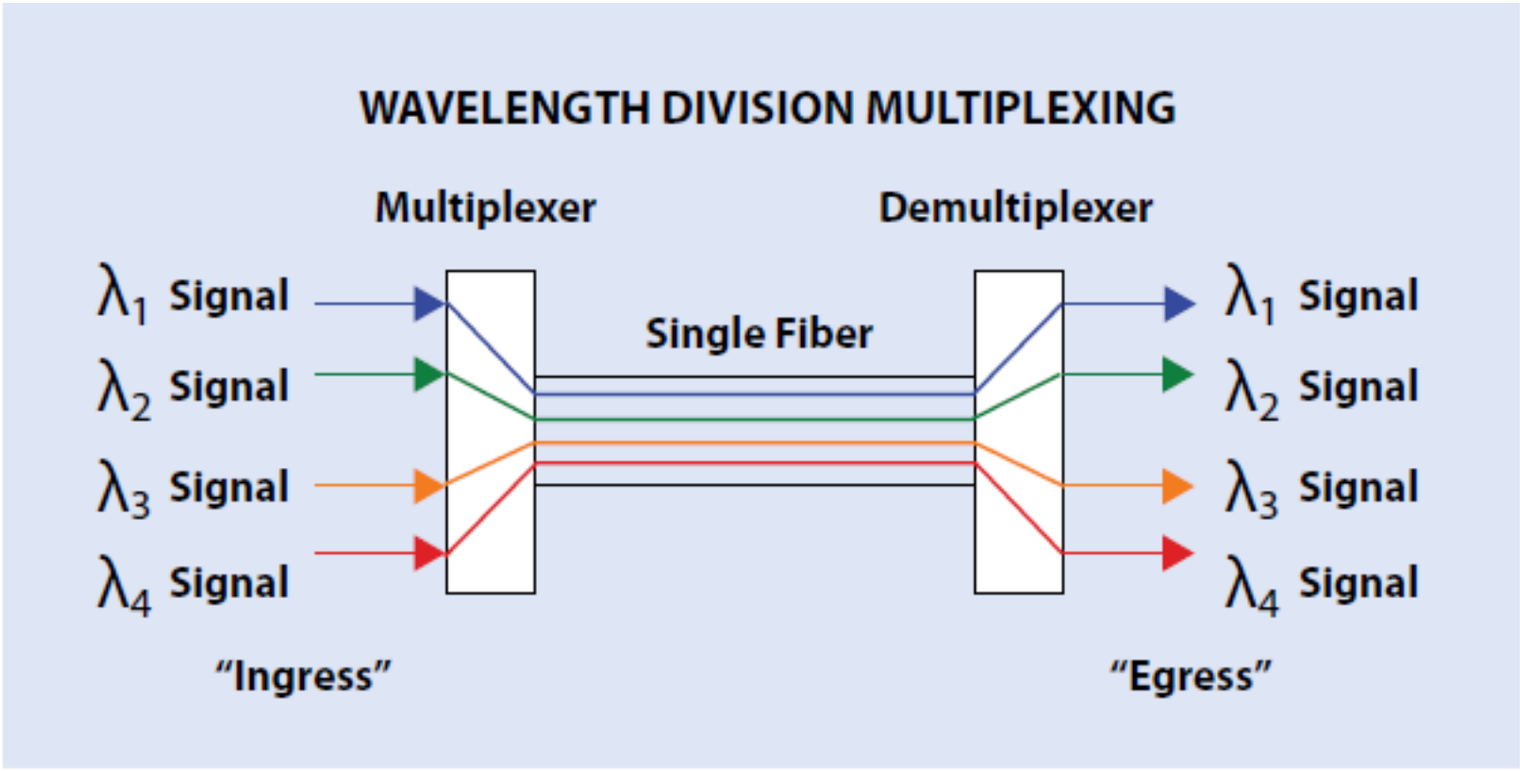


Figure 1: Basic WDM Technology Diagram

L1 interconnection

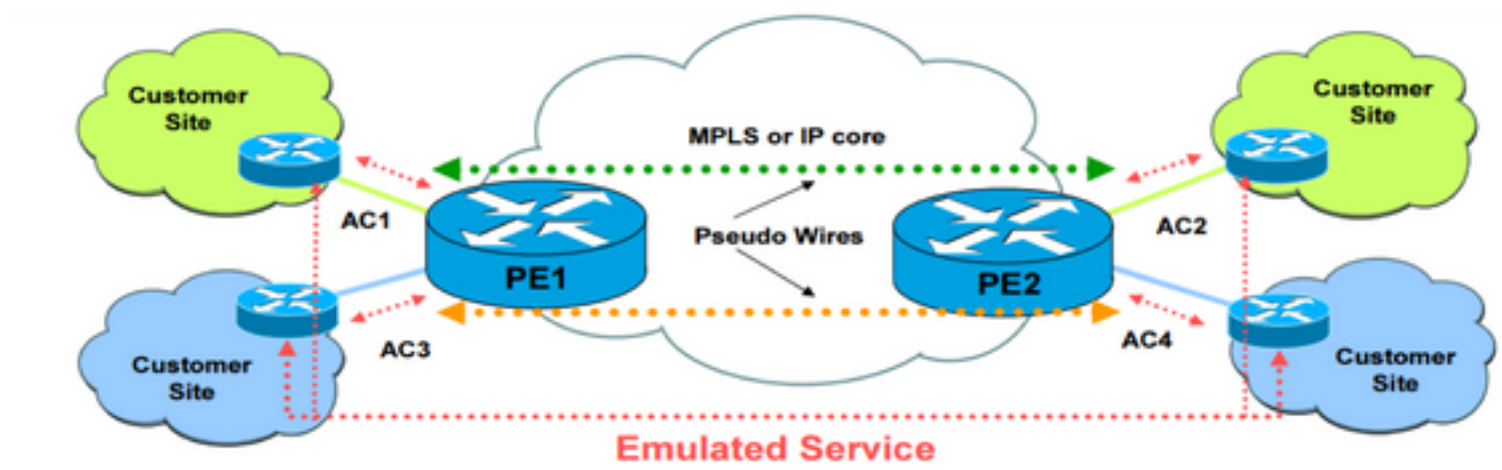
- What if the link breaks?
  - Service interruption
  - Potential split-brain of the Cluster
- Ooops…

It’s Time for some clarification

- L2 MPLS VPN's are divided in 2 main categories:
  - VPWS (Virtual Private Wire Service) also known as **point-to-point** VLL (virtual leased line) VPNs
  - and
- **point-to-multipoint** service like VPLS (Virtual Private Lan Service).

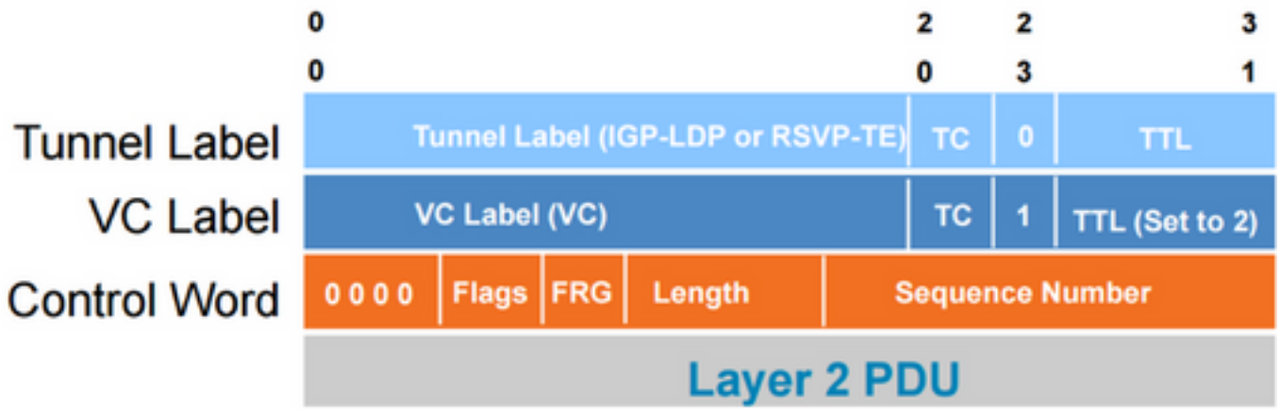
Virtual Private Wire Service (VPWS)

- Point-to-point : referred to as Pseudowires (PWs)
- PW is a connection between two PE devices which connects two ACs, carrying L2 frames
- Any Transport Over MPLS (AToM) or Ethernet Over MPLS (EoMPLS)
- Attachment Circuit (AC) is the physical or virtual circuit attaching a CE to a PE, can be Ethernet, ATM, Frame Relay, HDLC, PPP and so on.
- Frames that are received at the PE router on the AC are encapsulated and sent across the PSW to the remote PE router.
- The egress PE router receives the packet from the Pseudowire and removed their encapsulation.
- Customer Edge (CE) equipment perceives a PW as an unshared link or circuit



VPWS Traffic Encapsulation

- Three-level encapsulation
- Packets switched between PEs using Tunnel label
- VC label identifies PW
- VC label signalled between PEs
- Optional Control Word (CW) carries Layer 2 control bits and enables sequencing



From VPWS historically speaking we have:

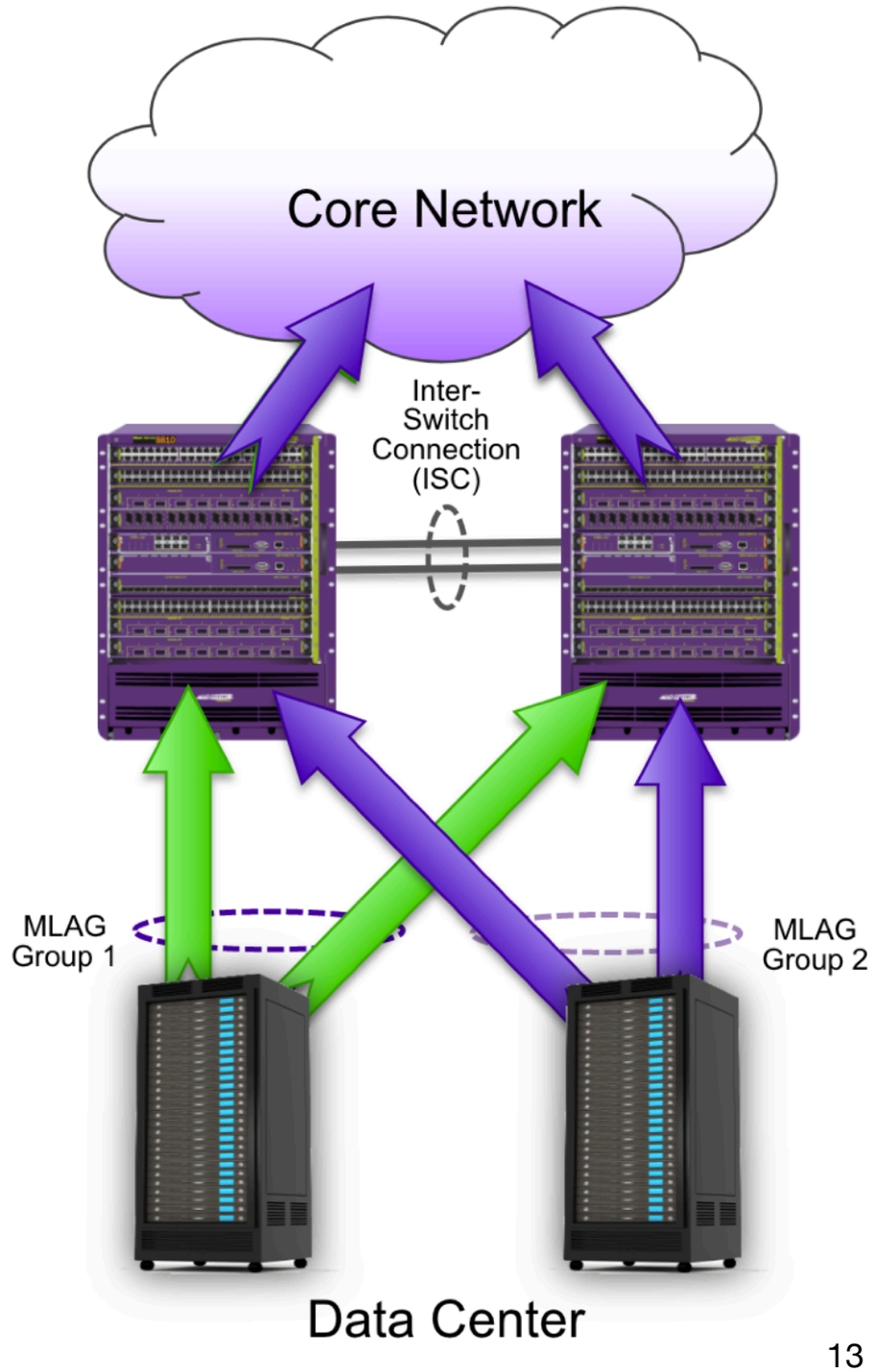
- **VLL in CCC mode - circuit cross connect**
  - In this type of VPN you don't have a inner label for customer service. You have only the outer label and this label is manually allocated by the operator when building the VPN.
- **SVC - static virtual circuit this can be called simplified Martini Mode**
  - In this mode you have 2 labels, one for LSP announced by LDP and one inner configured by the operator. The inner label defines the Virtual Circuit thus the customer Service.
- **VLL in Martini mode also known as EoMPLS service**
  - In Martini mode, the VC type and the VC ID are used to identify a VC between two CEs. You will need extended remote LDP sessions between PEs to transmit the VC message. To transmit the VC message, an FEC is added with the type as 128.
- **VLL in Kompella mode**
  - This mode adopts the MP-BGP as the signaling protocol for inner label.
  - This mode also adopts the RD and RT for transmitting the VC information and adopts the label block mode to assign the label to VC's.

From interconnection to infrastructure

Let’s look technologies inside a Datacenter

Multi-Chassis Link Aggregation (MLAG)

- LAG allows combining of ports effectively increasing the bandwidth. Up to 64-128 ports in a LAG Group.
- M-LAG allows combining of ports on ‘two’ switches to form a single logical connection to another network device
- Active-active paths. No STP port blocking
- Fast Failover
- For both Layer-2 and Layer-3 deployments
- Works with servers, switches, storage, and other network appliances



TRILL / SPB in a nutshell

- **Same but different** :
  - They both use ISIS to announce MAC addresses
  - They allow all paths to be active with multiple equal cost paths
  - designed to virtually eliminate human error during configuration and preserves the plug-and-play nature of Ethernet
  - IS-IS runs directly at Layer 2, no IP addresses are needed and IS-IS can run with zero configuration
  - IS-IS uses a TLV (type, length, value) encoding which makes it easy to define and carry new types of data
- Cisco FabricPath is a proprietary implementation of TRILL that utilizes the TRILL control plane (including IS-IS for Layer 2), but a non-interoperable data plane.

OTV in a nutshell

- No VPLS on NX-OS
- Cisco needed an equivalent for the Nexus platform
- Use GRE header for encapsulation (hardware processed since a while in ASICS)
- Use FabricPath with ISIS to propagate MAC address
- But ISIS has not the scalability of BGP (with EVPN)

Summary

	TRILL	MLAG	SPB	VPLS
Standard Body	IETF	Vendor-specific	IEEE	IETF
Technology	New	Matured	New (Variant of PBB)	Matured
Minimal Configuration	Yes	Yes	Yes B-VID needs to be configured for each ECMP	No
ECMP	Yes 16 active links with true hop-by-hop ECMP decisions	Yes 2 active links	Yes 16 active links with ingress ECMP decisions	Yes 16 ECMP LSPs can be achieved
Loop Prevention	Yes TTL and RPC	Yes	Yes RPC only	Yes
Virtualization Scale	4K networks	4K networks	Higher scale with mac-in-mac	Higher scale with VPN ID

What’s next?

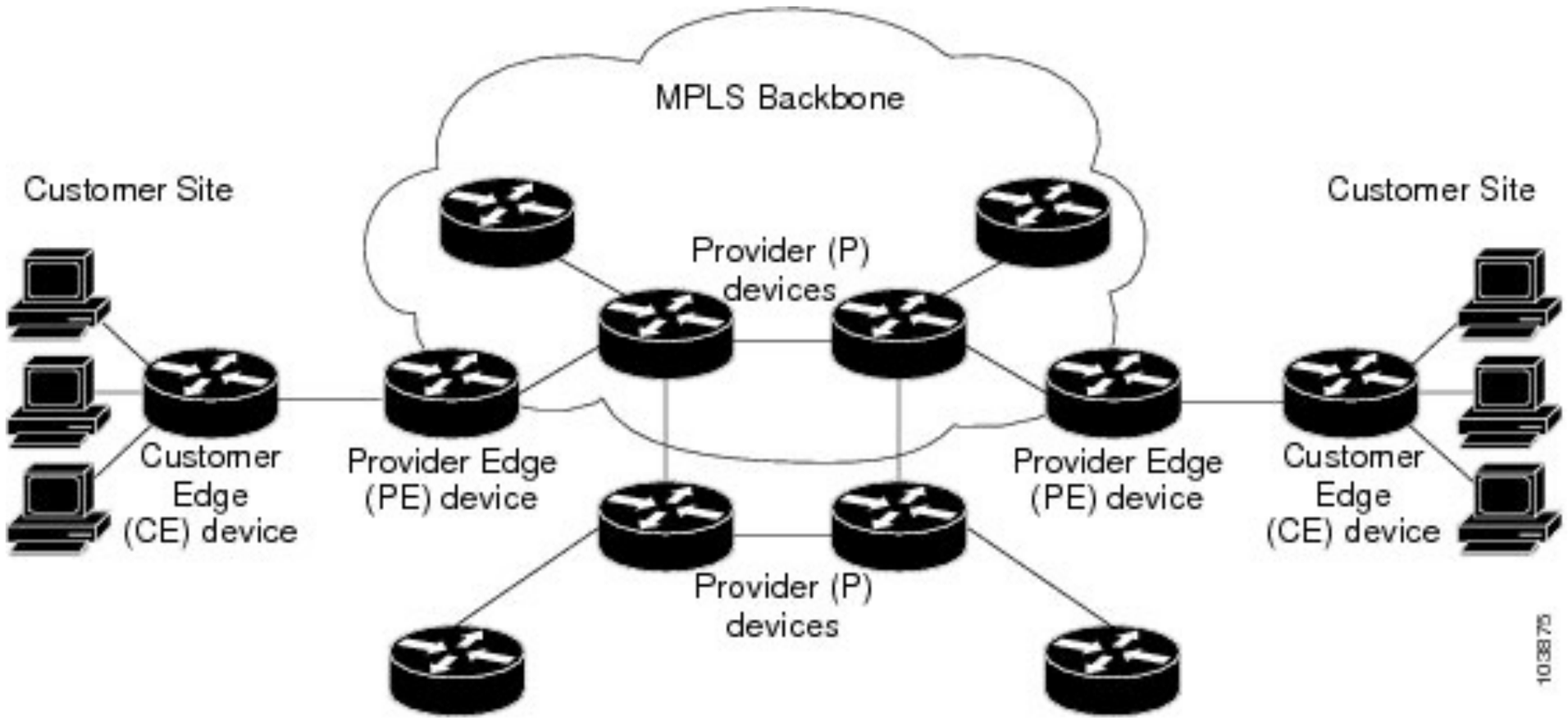
- L3-VPN
- The needs of L2 extension
- L2 services on top of L3 (IP) Networks (P2MP L2-VPN)

L3-VPN made simple

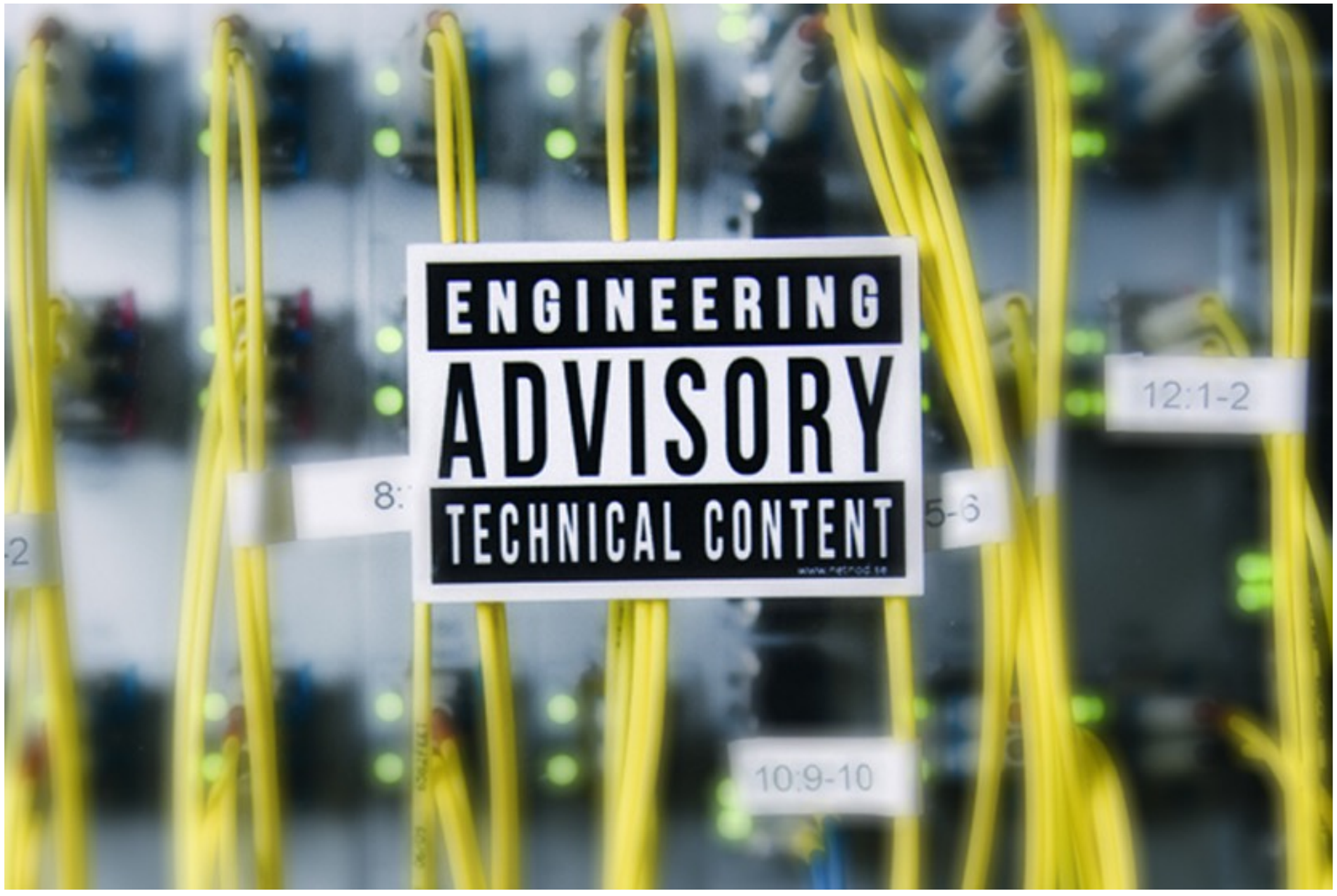
- Each **VRF** (virtual routing and forwarding) is a **distinct routing table**.
- **Each customer** (facing interface) is mapped **into a VRF**
- The **CE establish** BGP/OSPF/ISIS **session with the PE**
- The provider use **MP-BGP to propage VPNv4 / VPNv6** routing information into his network.
- **Routes** can be **leaked between VRF with RT** (route-target) **and RD** (route-distinguisher).
- **RT and RD** attributes **are** a kind of (extended) **BGP community**.

BGP/MPLS IP Virtual Private Networks (VPNs)

- https://tools.ietf.org/html/rfc4364



What’s next?



The Need for Stretched VLANs

- Stretched VLANs (or L2 extensions) are **used to solve a number of unrelated problems**.
  - **Subnet mobility**. You must move a subnet from one site to another during disaster recovery process.
  - start using automation or **configure the same subnet on VLAN interfaces or firewall** contexts **that are shutdown during the regular operation.**
  - **Most everyone solves this one by stretching a VLAN between datacenters** (because VMware consultants told them to do so) **and then experienced a dual-data-center meltdown before ever having the need to do a disaster recovery**.

The Need for Stretched VLANs

- Stretched VLANs (or L2 extensions) are used to solve a number of unrelated problems.

  - **IP multicast**. (stock exchange feeds and video streaming) it's easier to stretch a VLAN than to figure out how to **use Protocol-Independent Multicast**.

  - Some vendors "solve" this problem by requiring layer-2 connectivity between cluster members.

Who's Pushing Layer-2 VPN Services?

- Legitimate reason :
  - **Buying layer-2 services or IP transport services** from a service provider **because I know they can't mess them up too much**, and I'd still own the end-to-end routing.
  - **Then use the layer-2 transport** service **to build my own routed core**, or use IP transport service with a tunnel VPN on top of it.

L2 DCI

- L2-DCI are :
  - Bridging between datacenter
  - VLAN Extention (same VLAN and IP Subnet)
- Routing over a VPLS is *not* considered as L2 DCI.

What should a L2 DCI solution have?

- **Per-VLAN flooding control at data center edge.** Broadcasts/multicasts are usually not rate-limited within the data center, but should be tightly controlled at the data center edge

- **Broadcast reduction at data center edge.** Devices linking DC fabric to WAN core should implement features like ARP proxy.

- **Controlled unicast flooding.** It should be possible to disable flooding of unknown unicasts at DC-WAN boundary.

Stretch Cluster problems

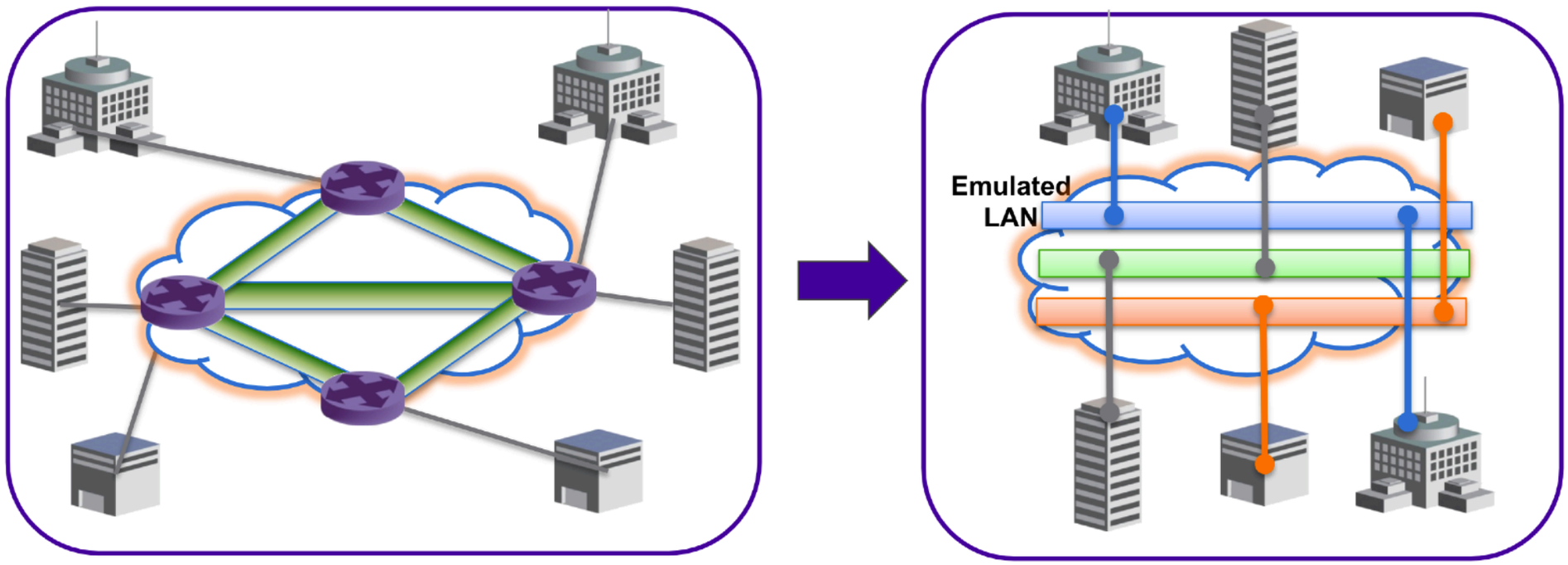
- **Trafic flows** accrocs the DCI are **unpredictable**
- **DCI failure splits the cluster** (half of the ressources are gone)
- DCI failure can cause **hard-to-recover split-brain problems**

What's next?

- L2 services on top of L3 (IP) Networks (P2MP L2-VPN)
  - Virtual Private LAN Service (VPLS)
  - Virtual extensible LANs (VXLANs)
  - Ethernet VPN (EVPN)

Virtual Private LAN Service (VPLS)

- L2 Ethernet VPN providing **point to multi-point communication** (P2MP)
- **All tenants** sites appear to be **on the same LAN regardless of location**
- **VPLS provides VLAN extensions over IP/MPLS networks**
  (L2TPv3 or even GRE)
- Each tenant **VLAN is mapped to a virtual switch instance** or VPN ID



Virtual Private LAN Service (VPLS)

- You can think that **your service provider appears as a big L2 switch** from the customer point of view.

- All **VPLS peers need to establish PW between them** (full-mesh)

- MAC addresses are learn with the **“Flood & Learn”** method

- **BUM** Broadcast, unknown unicast and multicast packet are **replicated at the source** (Head-End replication) and then **sent to each PW** —> CPU + link usage

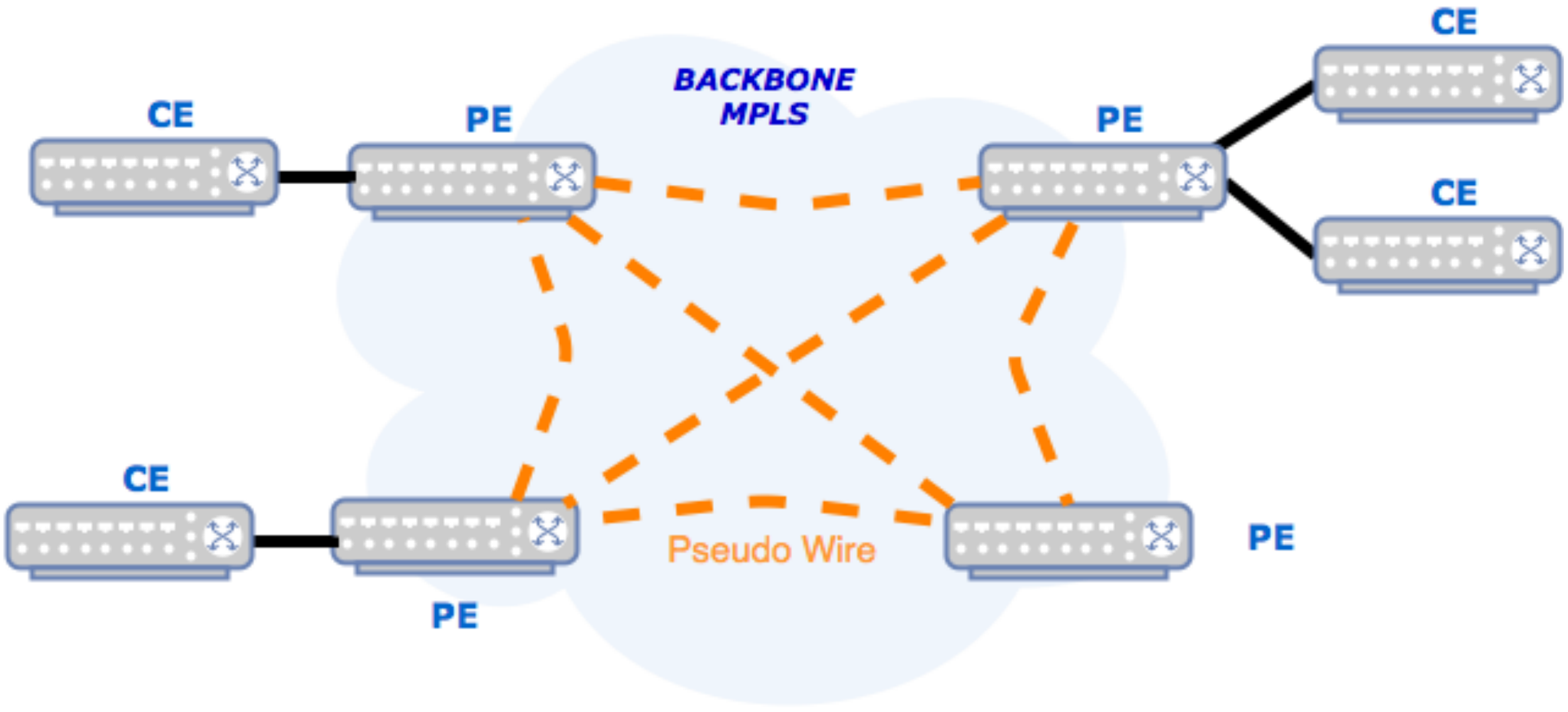
VPLS over IP/MPLS

Trame 802.1q issue du serveur ou encapsulée par un switch

@ mac src	@ mac dst	TPID 0x8100	PCP	DEI	VLAN ID	Ethertype 0x8000	DATA / IP PAYLOAD	FCS

Trame EoMPLS

@ mac src	@ mac dst	Ethertype 0x8847 0x8848	Tunnel Label	EXP	S	TTL	VC Label	EXP	S	TTL	@ mac src	@ mac dst	TPID 0x8100	PCP	DEI	VLAN ID	Ethertype 0x8000	DATA / IP PAYLOAD	FCS



Virtual Private LAN Service (VPLS)

- In VPLS you can adopt **LDP** as **signaling** protocol **or MP-BGP**.
- The **LDP** mode is preferable when the number of VPLS sites is relatively small (**no auto-discovery between peers**).
- The **BGP mode provides both auto-discovery and signalling** (with route reflectors).
- **If the scale of a VPLS network is large** (a great number of nodes or a wide geographical range), you can **use hierarchical VPLS (HVPLS)** that combines the two modes. In HVPLS, the **core** layer uses the **BGP** mode and the **access** layer uses the **LDP** mode.

Virtual extensible LANs (VXLANs)

- VXLANs introduced an overlay scheme that **expands the Layer 2 network address space from 4K to 16 million**.

- VXLAN tunneling protocol **encapsulates Layer 2 Ethernet frames in Layer 3 UDP packets**.

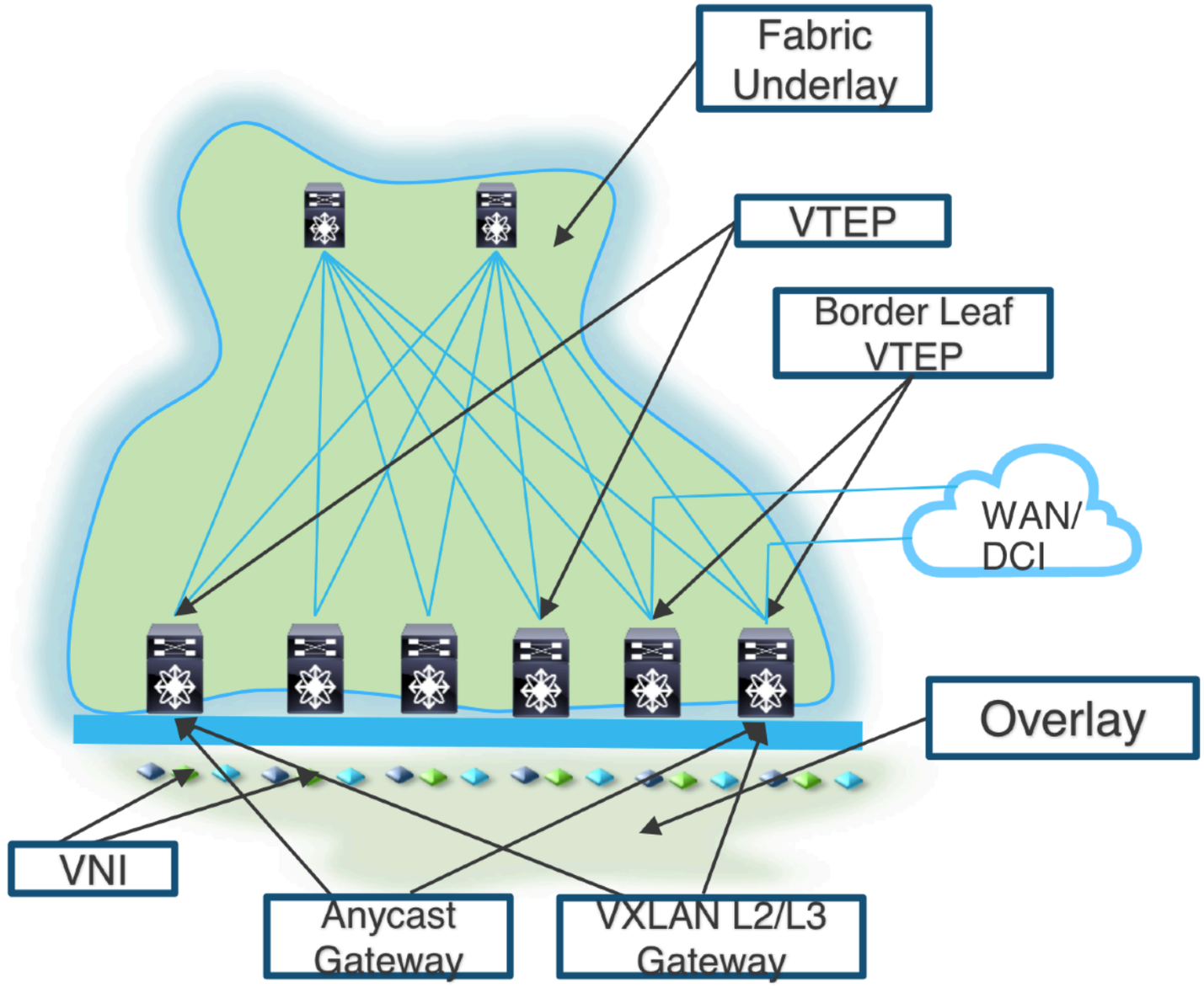
- In a VXLAN overlay network, a VXLAN network identifier **(VNI) uniquely identifies each Layer 2 subnet or segment**.

- The entity that performs the **encapsulation and de-encapsulation** is called a **VXLAN tunnel endpoint (VTEP)**.

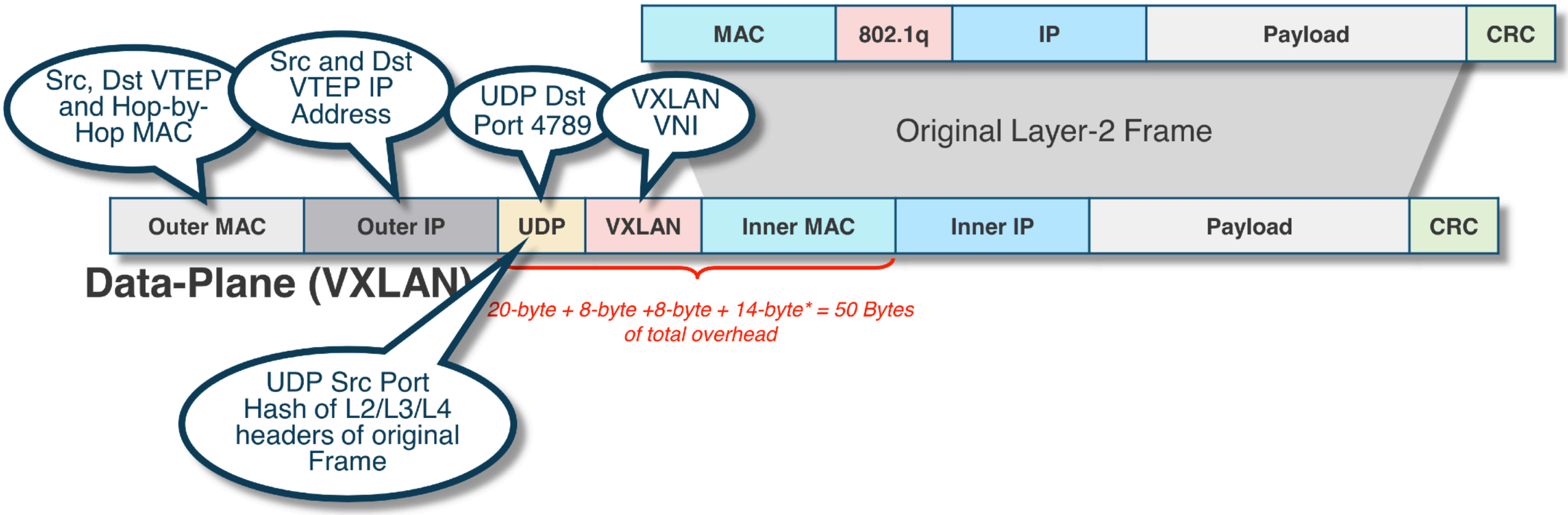
- **VTEPs** can reside **in hypervisor hosts**, such as kernel-based virtual machine (KVM) hosts **or on Networks devices** that functions as a Layer 2 or Layer 3 **VXLAN gateway**.

VXLAN Terminology

- **VTEP**: Hardware or software element at the edge for VXLAN encapsulation
- **VNI**: a logical network instance for layer 2 broadcast domain
- **VNID**: 24 bit segment ID
- **Anycast Gateway**: distributed default gateway function across al leaf nodes
- **VXLAN L2 Gateway**: gateway translate VLAN to VXLAN and VXLAN to VLAN in same BD
- **VXLAN L3 Gateway**: gateway translate VXLAN to VXLAN or VXLAN to VLAN in different BD



VXLAN Encapsulation

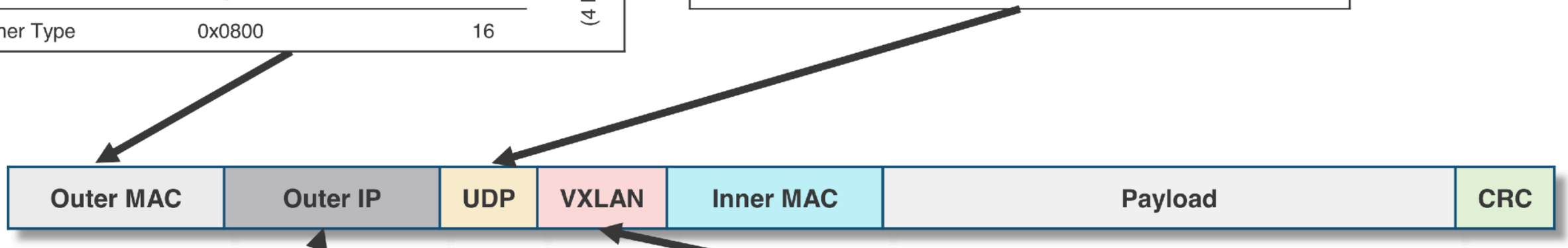


*plus 4 byte if IEEE 802.1q exists as part of Inner MAC Header

VXLAN Frame Format
–
MAC in IP Encapsulation

Field	Value	Bites	Total
Dest. MAC Address	Next-Hop MAC Address	48	14 Bytes (4 Bytes Optional)
Src. MAC Address	Next-Hop MAC Address	48	
VLAN Type	0x8100	16	
VLAN ID	Tag	16	
Ether Type	0x0800	16	

Field	Value	Bites	Total
Source Port	**L2/L3/L4 Hash**	16	8 Bytes
Destination Port	**4789 (UDP)**	16	
UDP Length		16	
Checksum	0x0000	16	

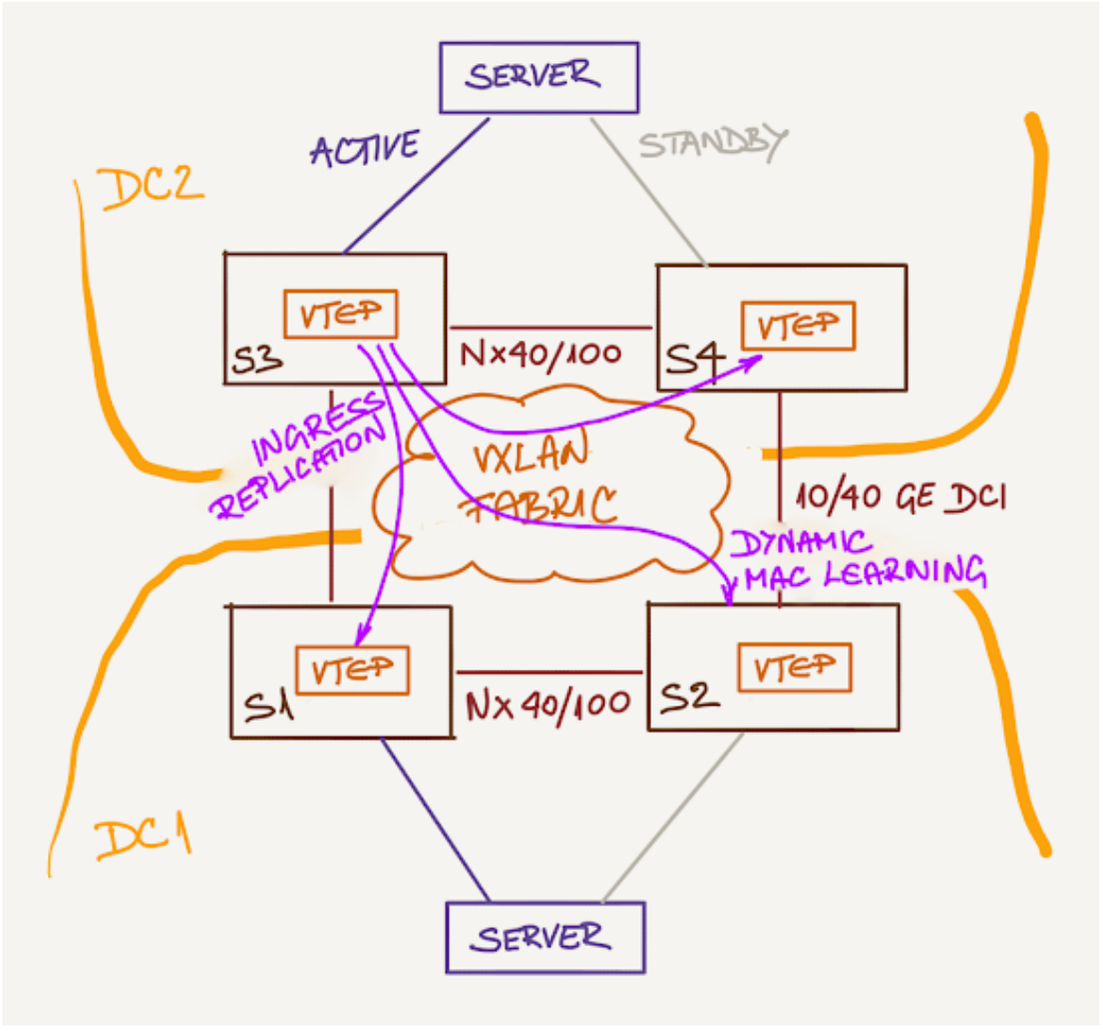


Field	Value	Bites	Total
IP Header	Misc. Data	72	20 Bytes
Protocol	0x11 (UDP)	8	
Header Checksum	Various	16	
Source IP	**Src, VTEP IP**	32	
Destination IP	**Dest. VTEP IP**	32	

Field	Value	Bites	Total
VXLAN Flags	RRRRIRRR	8	8 Bytes
Reserved		24	
VNI	**16M Possible Segments**	24	
Reserved		8	

Basic VXLAN setup

- pure IP routing within the fabric
- VXLAN encapsulation to transport Ethernet traffic across IP fabric
- **Ingress replication** instead of IP multicast to implement VXLAN flooding of BUM frames
- **Dynamic MAC learning relying on BUM** flooding to populate MAC-to-VTEP tables
- the list of remote VTEPs to which the traffic needs to be flooded **is statically configured** on every switch.



Ethernet VPN (EVPN)

- **BGP control-plane for Ethernet Segment and MAC distribution and learning** over MPLS or VXLAN core.

- EVPN **replaces flood-and-learn behavior** of traditional Ethernet bridges (of VPLS or simpler VXLAN implementations) **with BGP control plane** :
  MAC addresses are propagated as BGP prefixes within the **EVPN address family**.

- **Same principles** and operational experience **of IP VPNs**.

- EVPN implementations could use dynamic IP address discovery using DHCP reply snooping, ARP request snooping, or IP packet header gleaning, and **advertise IP-to-MAC bindings in EVPN BGP updates**.

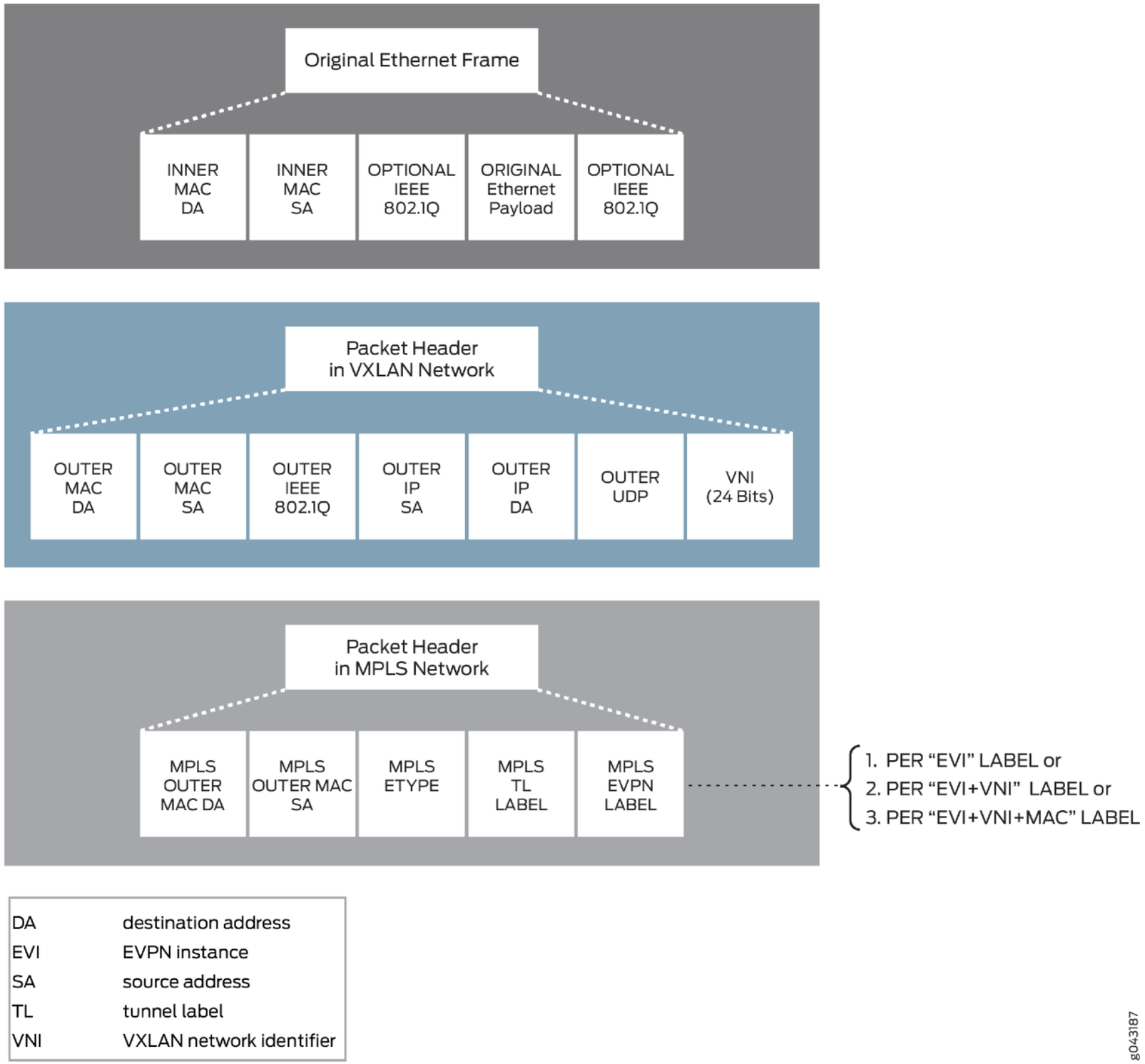
- **Combining L2 and L3 forwarding information.** You can use EVPN to implement end-to-end bridging, integrated bridging and routing, or routing-only fabrics.

- Decent **EVPN implementations** can **reduce flooding** by turning off unknown unicast flooding, **and eliminating ARP flooding** with local ARP proxy.

EVPN

- **No** use of **Pseudowires**.
- Uses **MP2P tunnels** for unicast.
- Multi-destination frame delivery via **ingress replication** (via MP2P tunnels) or LSM.
- EVPN has **built-in support** for **edge multihoming** based and **edge load balancing**.
- Multi-vendor solutions under **IETF standardization**.
- EVPN was **designed** to **be used with MPLS data plane to replace VPLS** in service provider networks, **and got adopted (with VXLAN encapsulation)** as the control plane used by the Network Virtualization Overlays (NVO) IETF workgroup.

EVPN-VXLAN Packet Format



DA	destination address
EVI	EVPN instance
SA	source address
TL	tunnel label
VNI	VXLAN network identifier

g043187

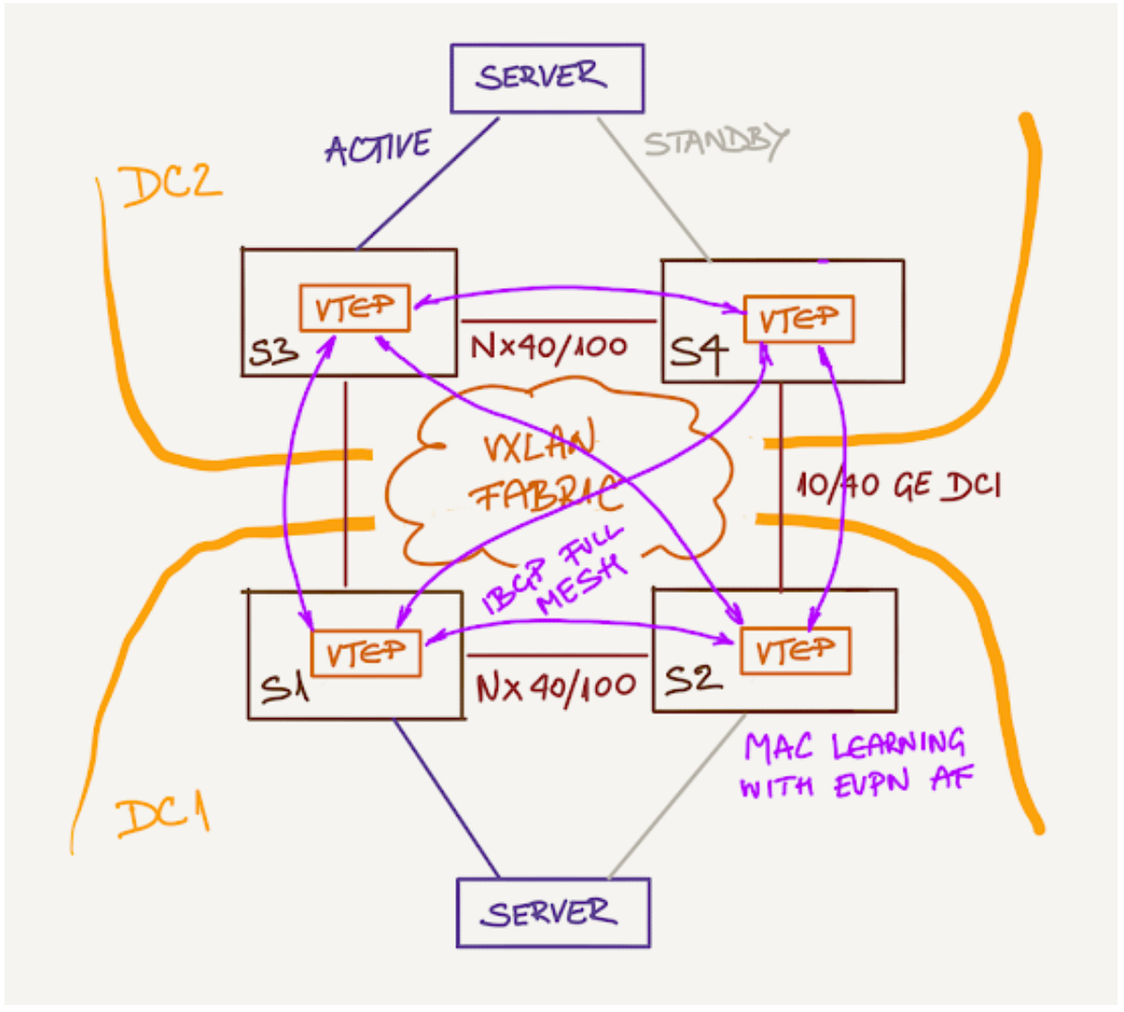
VXLAN & EVPN setup

- use of **EVPN control plane** between the four switches.

- **full mesh of IBGP sessions** (RR if more switches)

- EVPN **automatically builds the flood lists**, removing the need for manual configuration, and **propagate customer MAC addresses using BGP**.

- **Active / Active attachement** of the server is possible (leveraging BGP ECMP)



EVPN : Host reachability BGP update

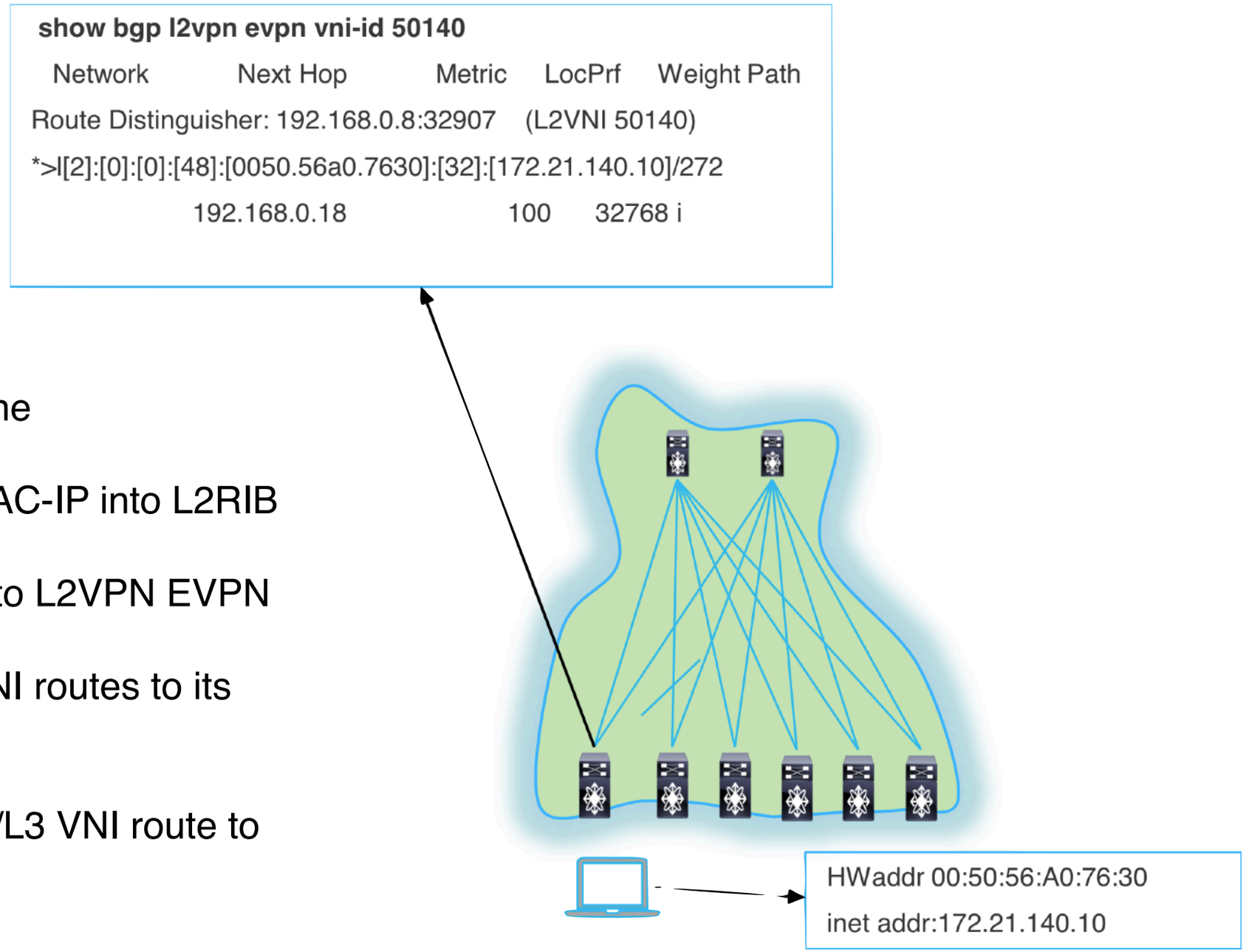
1. Host 172.24.140.10 comes online

2. VTEP leaf-1 install MAC and MAC-IP into L2RIB

3.VTEP leaf-1 installs host mac-ipto L2VPN EVPN

4.VTEP leaf-1 advertises L2/L3 VNI routes to its EVPN neighbors

5.VTEP Spine nodes advertise L2/L3 VNI route to all other leaf nodes



EVPN : route types

- **Host MAC (Route Type 2)**
  - MAC only, Single VNI, Single Route Target (for MAC-VRF)
  - MAC attributes are Mandatory
- **Host MAC+IP (Route Type 2)**
  - MAC and IP, Two VNI, Two Route Target (MAC-VRF and IP-VRF), Router MAC
  - IP Attributes are Optional, Populated through ARP/ND
- **Internal and External Subnet Prefixes (Route Type 5)**
  - IP Subnet Prefix, Single VNI, Single Route Target (for IP-VRF)
  - Populated through External Routing Protocol

EVPN : Host reachability BGP update

```
V2# show bgp l2vpn evpn

BGP routing tabl     n f   VRF d fault   PN EVPN
Route Distinguishe     0.10 :3277
BGP routing table entry for [2]:[0]:[0]:[48]:[0000.3001.1101]:[32]:[192.168.10.101]/272,
version 4
Paths: (1 available, best #1)
Flags: (0x000202) on xmit-list, is not in l2rib/evpn,

     sed path-id 1
  Type: inte          est path, no labeled nexthop
  AS-Path: NONE,          to AS
    10.200.200.101 (me    c 3)    om 10.10.10.201 (10.10
      Origin IGP, MED n t set, localpref 100, weight 0
      Received label 3001 5000
      Extcommunity:  RT:65500:3001 RT:65500:5000 ENCAP:8 Router MAC:0200.0ade.de01
      Originator: 10.10.10.1 Cluster l t: 10.10.10.201
```

Route Type: MAC/IP

Ethernet Segment Identifier (ESI)

Ethernet Tag Identifier (Ethtag)

MAC Address Length

MAC Address

IP Address Length

IP Address

Next-Hop IP Address

L2VNI (MPLS Label1)

L3VNI (MPLS Label2)

Encap:8 VXLAN

L2VNI Route Target

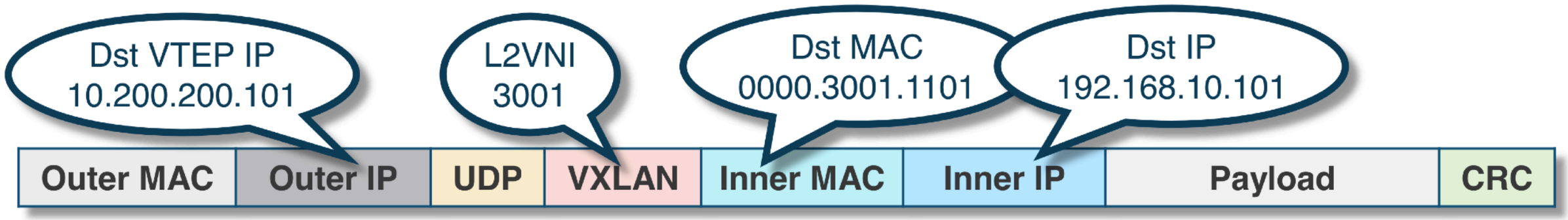
L3VNI Route Target

Router MAC

VXLAN and BGP EVPN
–
Putting it Together

Control-Plane (BGP EVPN)

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.1101/48	3001 65500:3001	192.168.10.101/32	5000 65500:5000	10.200.200.101	



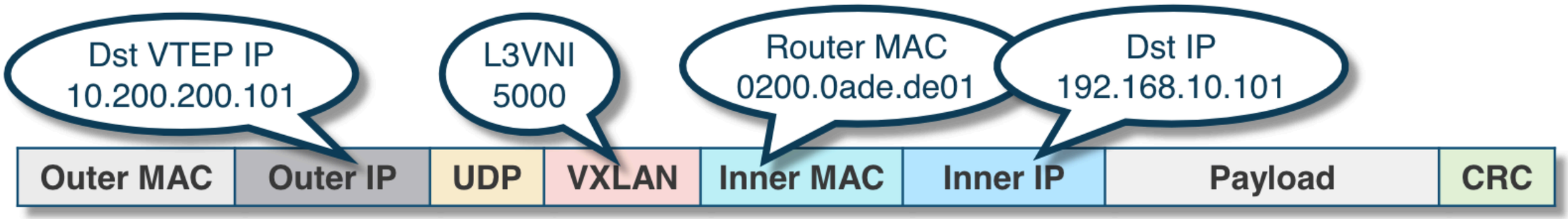
Data-Plane (VXLAN)

VXLAN and BGP EVPN
–
Putting it Together

Control-Plane (BGP EVPN)

Extended Community
Router MAC
0200.0ade.de01

Type	MAC / Length	L2VNI / RT	IP / Length	L3VNI / RT	Next-Hop	Seq.
2	0000.3001.1101/48	3001 65500:3001	192.168.10.101/32	5000 65500:5000	10.200.200.101	



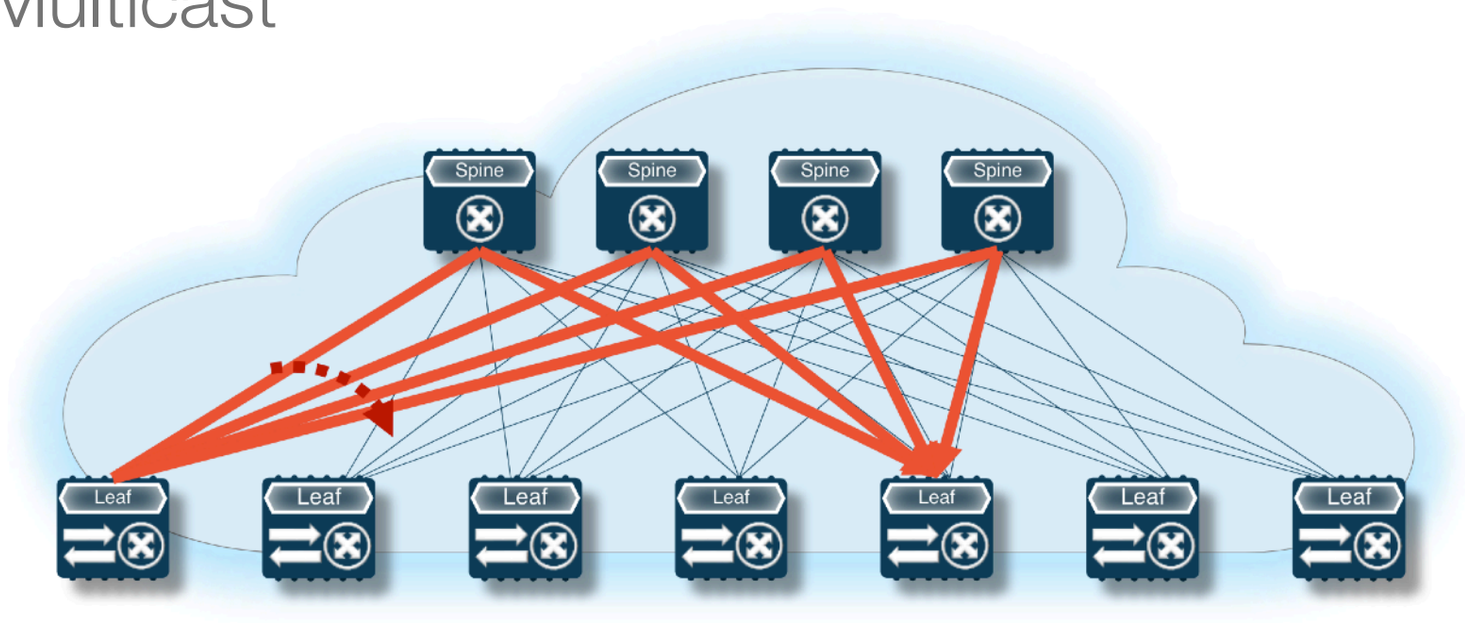
Data-Plane (VXLAN)

BGP Top Of Rack (TOR)

Spine & Leaf design

Spine & Leaf Topology

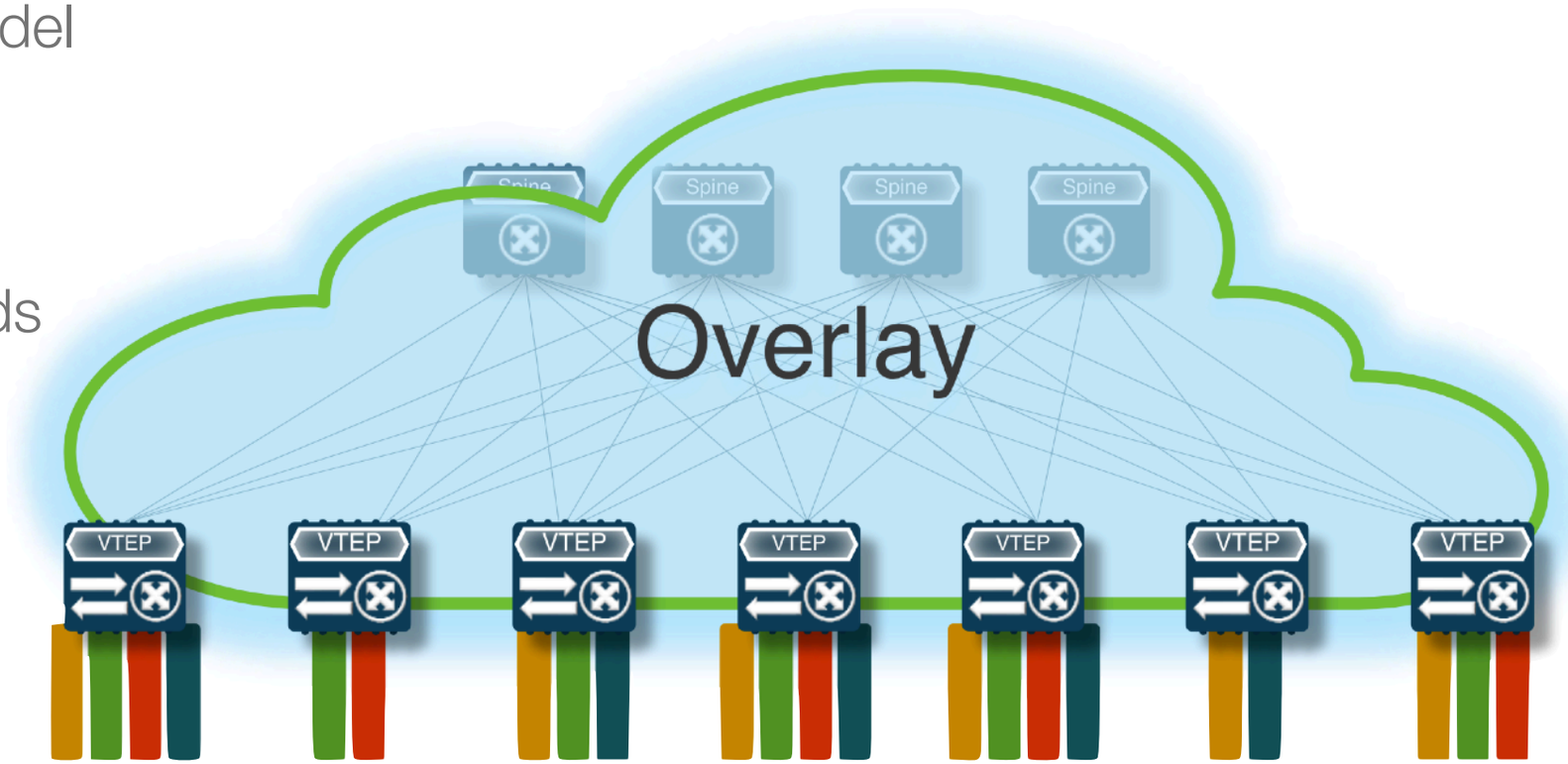
- Wide ECMP: Unicast or Multicast
- Uniform Reachability
- Deterministic Latency
- High Redundancy
  (On Node or Link Failure)



- Not Limited to Port Density and Not Limited to Port Density
- Also known as **Clos Network**
  - Charles Clos (1953) "A study of non-blocking switching networks"

Data Center Fabric Properties

- Any Subnet, Anywhere, Rapidly
- Reduced Failure Domain
- Extensible Scale and Resiliency
- Mobility / Segmentation / Scale
- Abstracted Consumption Model
- Layer-2 and Layer-3 Service
- Physical and Virtual Workloads



Summary of L2 - IP Fabric protocols

Transport protocol	Underlay for IP reachability	Overlay for EVPN
VXLAN or MPLS (GRE,L2TP, …)	ISIS/OSPF	iBGP
	eBGP	eBGP
	eBGP	iBGP

Spine & Leaf

- **If IGP as the fabric routing protocol**,
- and BGP to carry endpoint reachability information (EVPN addresses/routes or MPLS/VPN routes)
- **use iBGP to simplify the network design** and device configuration.
  - **you don’t need to run BGP on the spine switches unless** you use the spine switches as **iBGP route reflectors.**

Spine & Leaf

- **In larger fabrics that use BGP as the internal routing protocol** for scalability reasons, **use eBGP**.

- If you decide to carry global IP prefixes in BGP then **it's easier to implement BGP everywhere design** (with BGP running on spine switches) **than using BGP-free MPLS core design.**

- **When running eBGP as the sole routing protocol** in a data center fabric **use the same AS number on all spine switches** to **prevent path hunting** (prefixes advertised between spine switches through leaf switches).

- You could also use traditional BGP policy tools like AS-path filters to **prevent leaf switches from becoming transit** switches, but **it's much easier to solve this problem with good AS numbering scheme**.

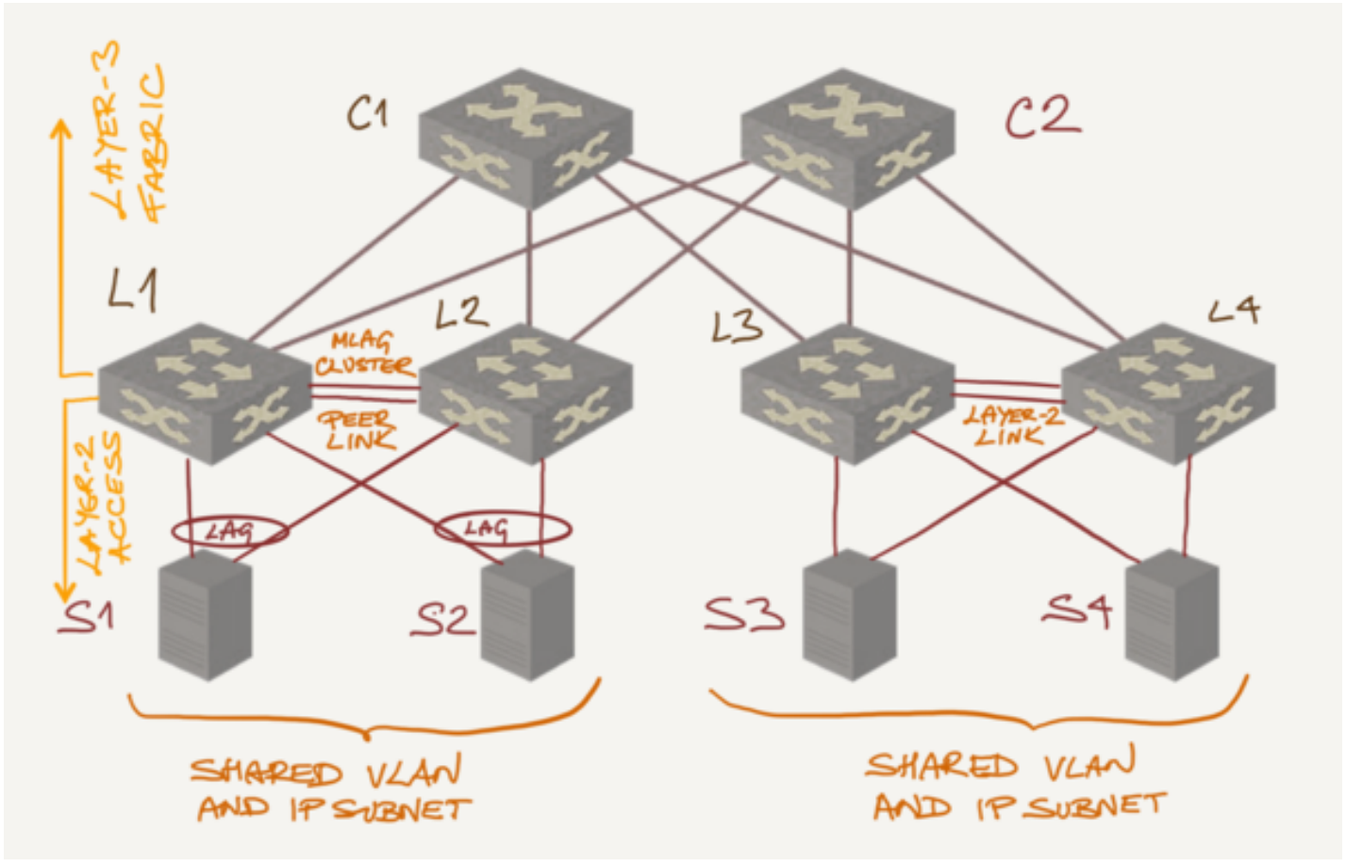
AS Numbers on Leaf Switches

- When using eBGP in your data center fabric **use a different AS number on every leaf switch**.
  - Eliminates BGP tweaks like allowas-in
  - **Makes troubleshooting** easier due to easy attribution of prefixes to leaf switches.
- **You might be forced to use the same AS number on all leaf switches** when working with equipment from vendors that never took the effort to simplify BGP configuration for data center fabrics use case

MLAG Clusters

- Layer-3-only fabric designs need **special provisions for multi-homed servers**.

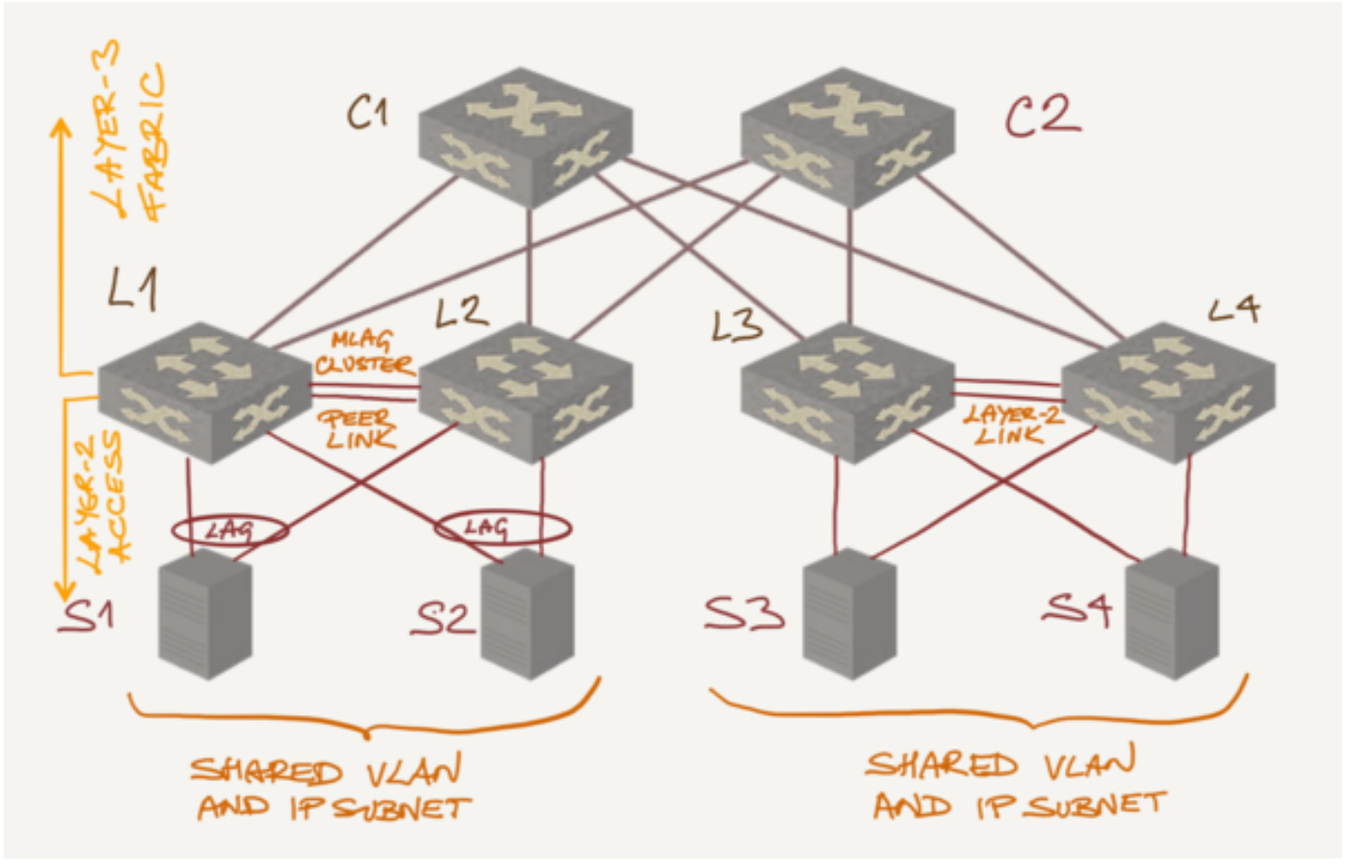
- **Leaf switches** to which the redundantly-connected servers are attached **must share a VLAN and an IP subnet**.



MLAG Clusters

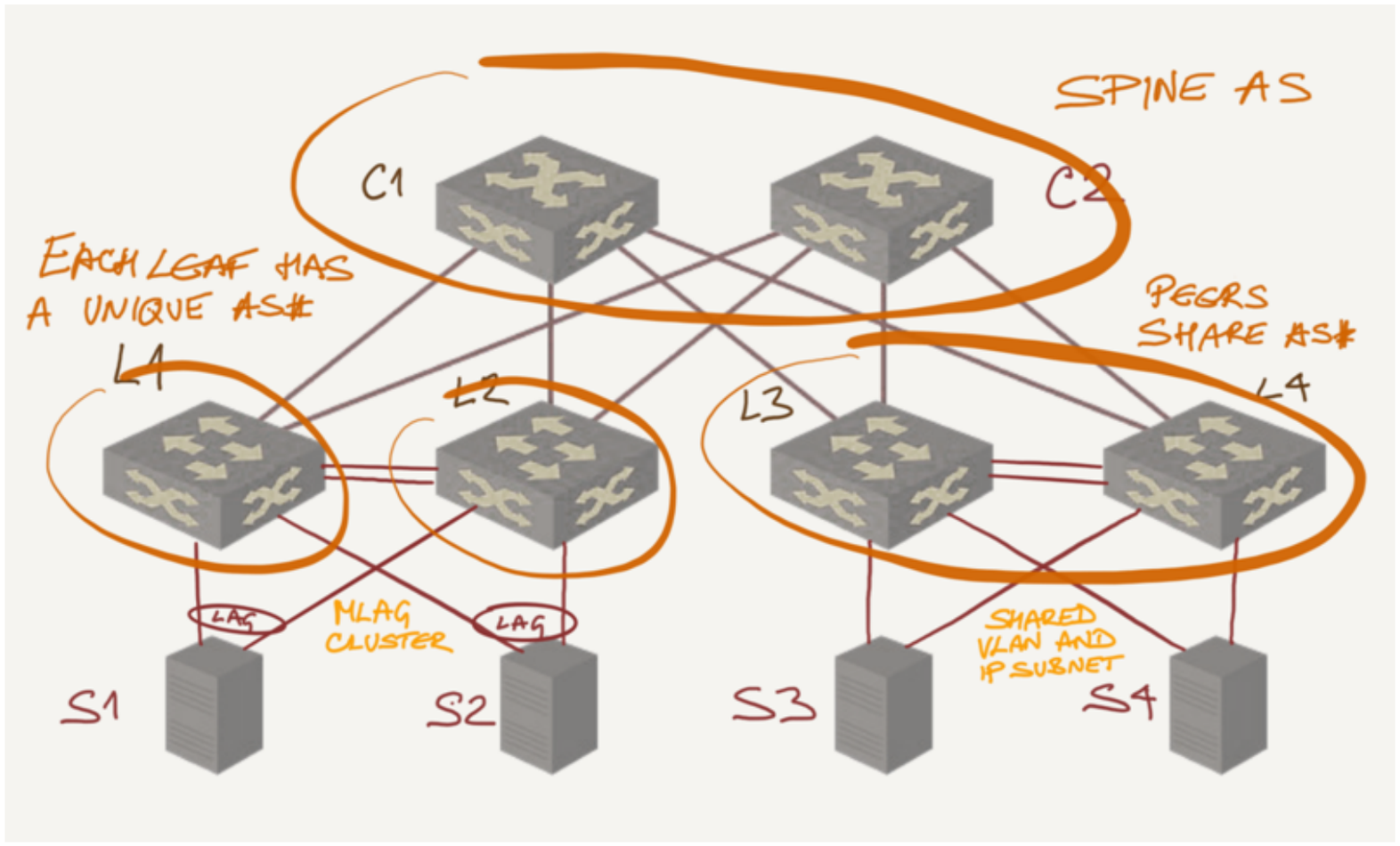
- **Layer-2 fabrics using EVPN or layer-2 routing technologies** (like TRILL or SPB) **don't have the same limitations** from the traffic forwarding perspective.

- But might **still need a direct inter-leaf link** to support **MLAG functionality.**



AS Numbers for MLAG Clusters

- **Traditional BGP designers** might be inclined to use the **same BGP AS number on all members** of an MLAG cluster (or a leaf pair sharing a VLAN) as these switches advertise the same IP prefix into the data center fabric.



AS Numbers for MLAG Clusters

- **Large-scale** data center fabric designers **prefer to use a unique BGP AS number on every leaf switch** to:

  - **make route advertisement attribution easier**
    the AS path in a BGP update uniquely identifies the leaf switch originating the advertisement;

  - **simplify automation scripts**
    leaf switch BGP configuration is consistent regardless of whether the leaf switch provides redundant or non-redundant server connectivity.

BGP in EVPN-Based Data Center Fabrics

- Similar to MPLS/VPN, **EVPN uses an additional BGP address family** that can be used **with iBGP or eBGP** sessions.

- However, **EVPN technology assumes the traditional use of BGP** as endpoint reachability distribution protocol **working in combination with an underlying routing protocol**:

- **BGP next hops** advertised in EVPN updates **point to egress Label Switch Router (LSR) or VXLAN Tunnel Endpoint (VTEP)**;

- **Underlying routing protocol computes the best path(s) to BGP next hops;**

- A BGP **router** that **changes the BGP next hop** in an EVPN update must **also perform data plane decapsulation and re-encapsulation**.

VXLAN-to-VXLAN forwarding

- **Numerous ASICs** used in high-speed data center switches **cannot perform VXLAN-to-VXLAN forwarding**.

- This **functionality is usually known as RIOT** – Routing In and Out of the Tunnels;

- **VXLAN-to-VXLAN bridging** requires even more **complex setup** due to split-horizon rules controlling BUM flooding.

Which underlay for EVPN?

- If you decided to use an IGP routing protocol in your data center fabric, **use iBGP on top of an IGP underlay**.

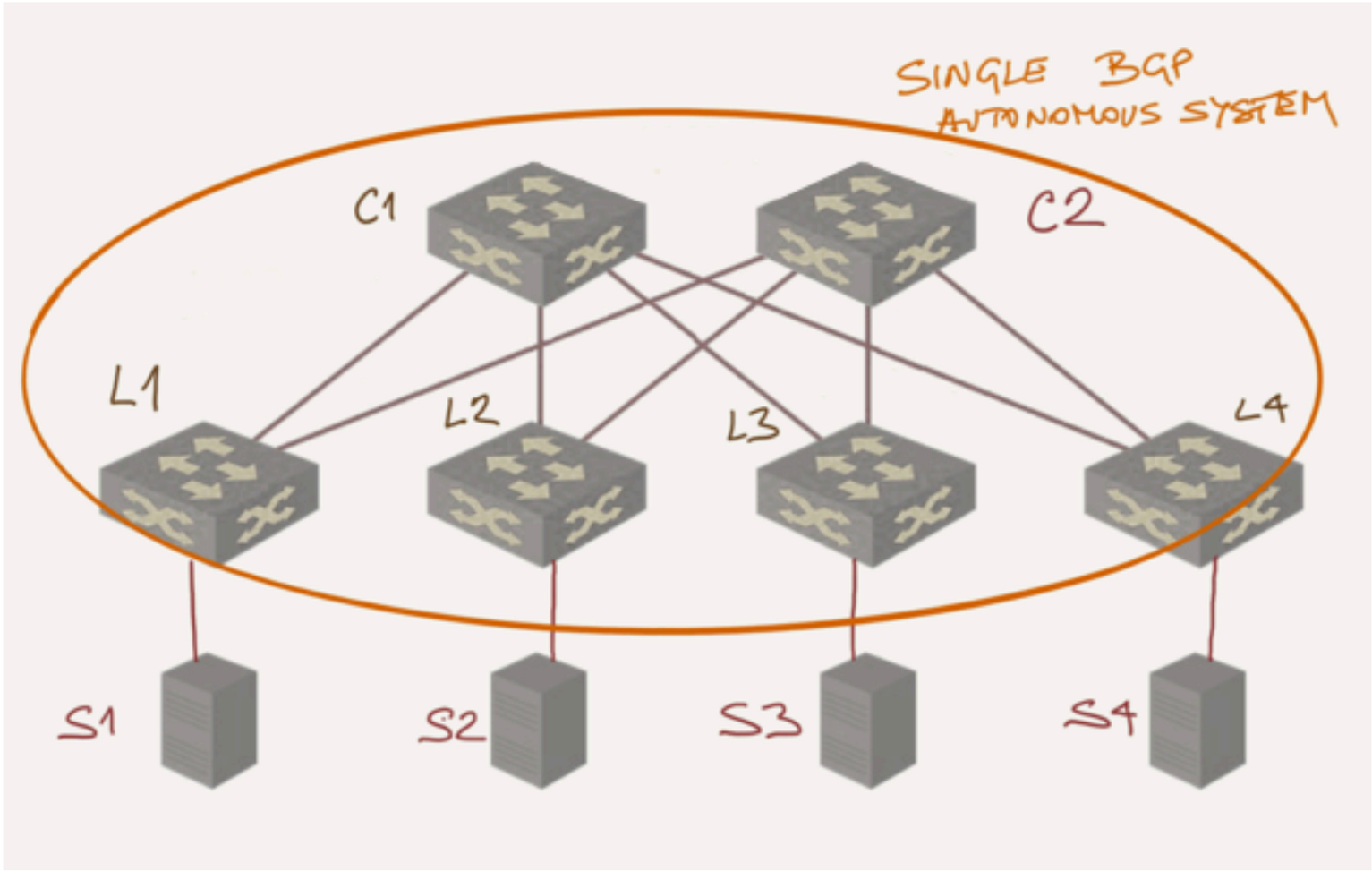
- If you decided to use **BGP as the underlay routing protocol,** and your chosen vendor provides a robust and easy-to-use implementation of EVPN over EBGP, **then use EBGP-only EVPN design**.

- **If you cannot use IGP routing protocol** in your data center fabric, and your chosen **vendor discourages the use of EVPN route servers on spine switches**, you might have to **use a combination of iBGP-based EVPN on top of eBGP-based underlay routing**;

- If your fabrics requires **extremely large number of prefixes that cannot be handled by data center switches acting as route reflectors** or route servers, **use VM-based route reflectors** or route servers. **IBGP-based EVPN on top of EBGP-based underlay might be your only option**.

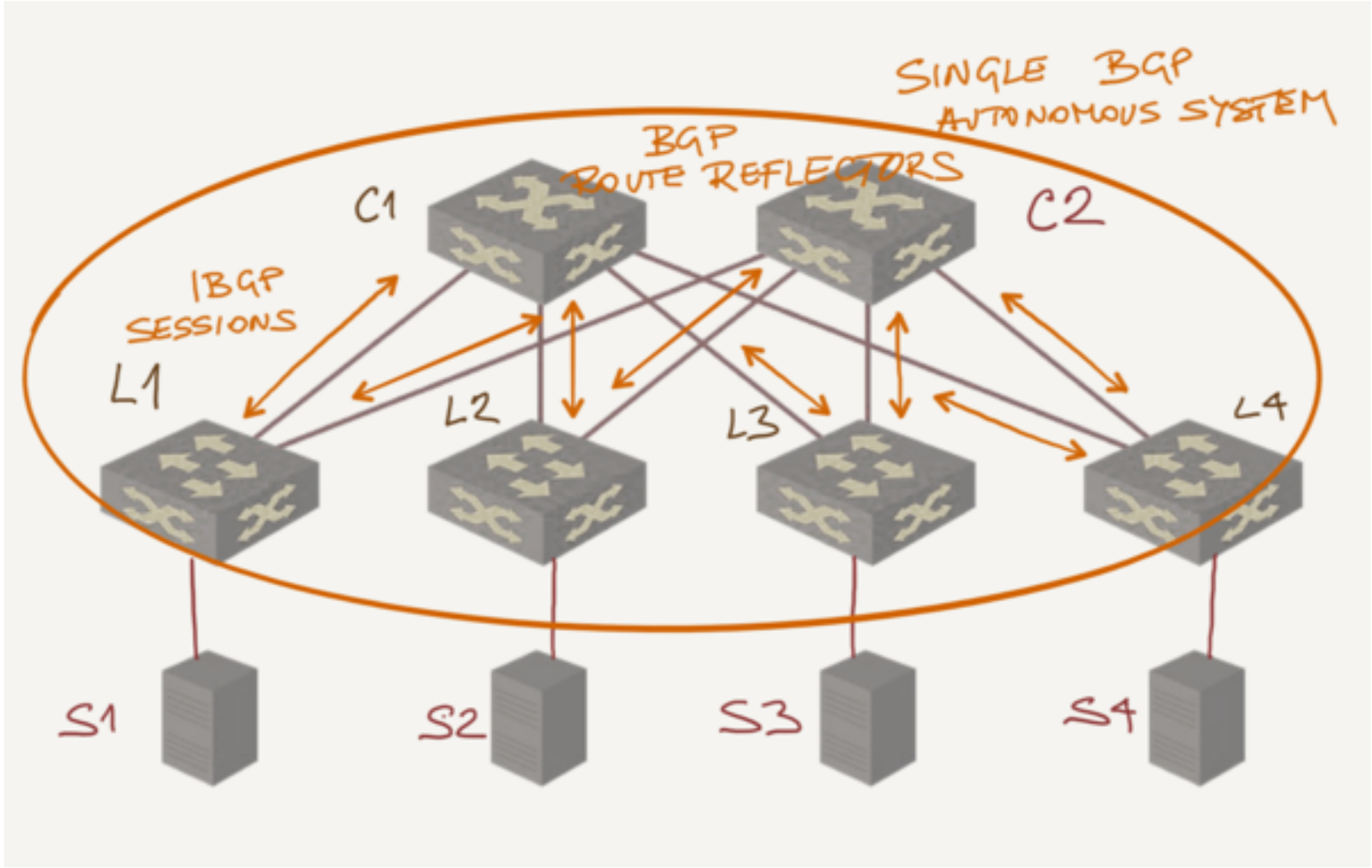
EVPN Using IBGP with IGP

- The **easiest way to build an EVPN-based** data center **fabric** is to:
  - **Use an IGP** (OSPF or IS-IS) **as the underlay** fabric routing protocol;
  - **Use iBGP between leaf switches** to exchange EVPN updates.



EVPN Using IBGP with IGP

- The **easiest way to build an EVPN-based** data center **fabric** is to:
  - **Use an IGP** (OSPF or IS-IS) **as the underlay** fabric routing protocol;
  - **Use iBGP between leaf switches** to exchange EVPN updates.
  - In fabrics **larger than a few switches** you should **deploy BGP route reflectors.**



Is OSPF or IS-IS Good Enough for My Data Center?

- **Short answer**: most probably yes
  - If it isn't, I sincerely hope you have an architecture/ design team in place and don't design your data center fabrics based on free information floating around the 'net
- **Few hundred devices in a single area should be OK**
  - **64 leaves + 4 spines** could get you up to **3072 ports**.
  - **30 VM per port** and you get **~90K hosts** in the fabric.

Why Is Everyone So Focused on BGP Then?

- We're telling you BGP is good for you **because Petr (and RFC 7938) said so**. Some vendor SEs doing exactly that;

- I always wanted to play with BGP and **now I have an excuse to do so**;

- **I want my network to be as cool as Microsoft's** (that's where Petr started using BGP as better IGP).

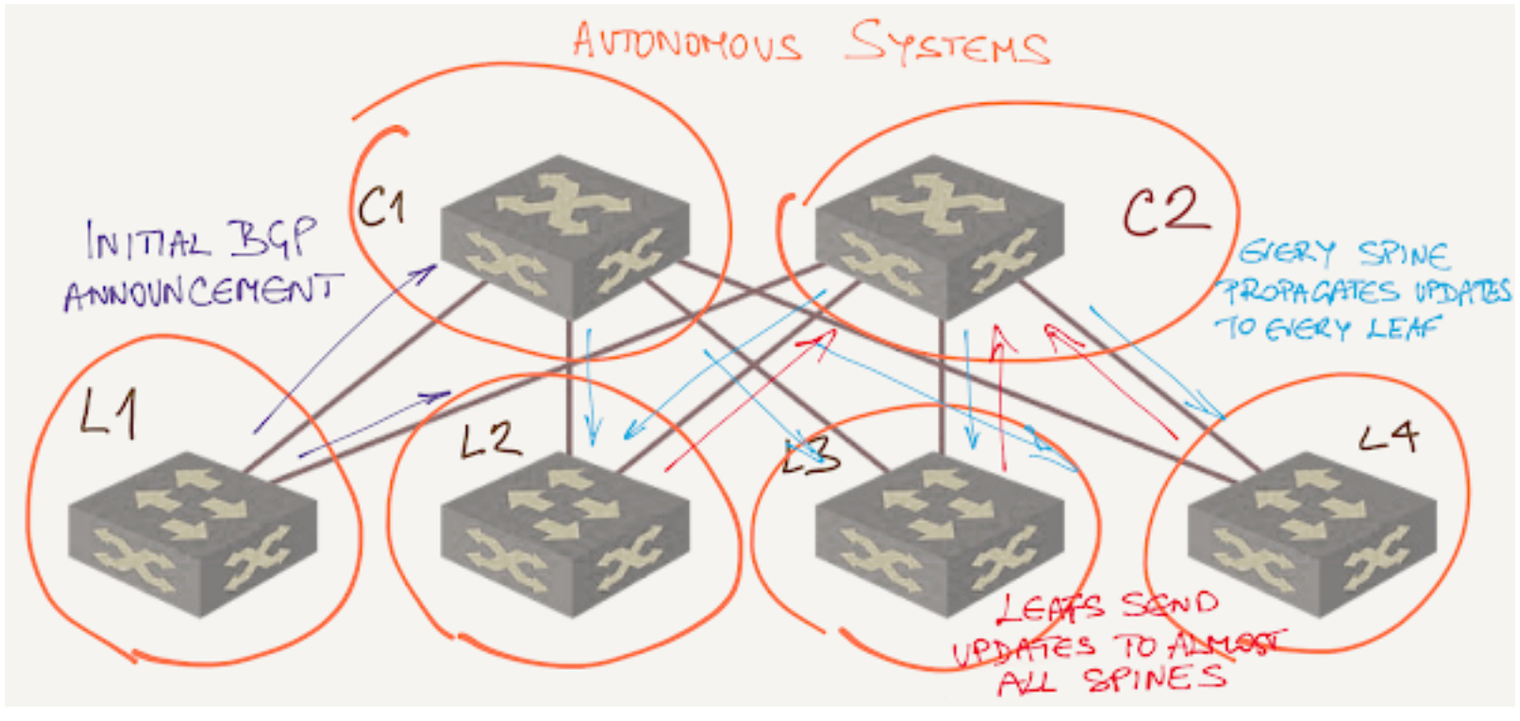
EVPN Using eBGP without an Additional IGP

- If you decided to use **eBGP as your underlay fabric routing protocol** due to large number of switches in your fabric, you could **add EVPN as an additional address family** over existing eBGP sessions (service disruption).

- **However**, as the **spine switches should not be involved in intra-fabric customer traffic forwarding** regardless of whether your implementation uses MPLS or VXLAN encapsulation, **the BGP next hop in an EVPN update must not be changed on the path between egress and ingress switch** – the BGP next hop should always point to the egress fabric edge switch.

Running EVPN over eBGP

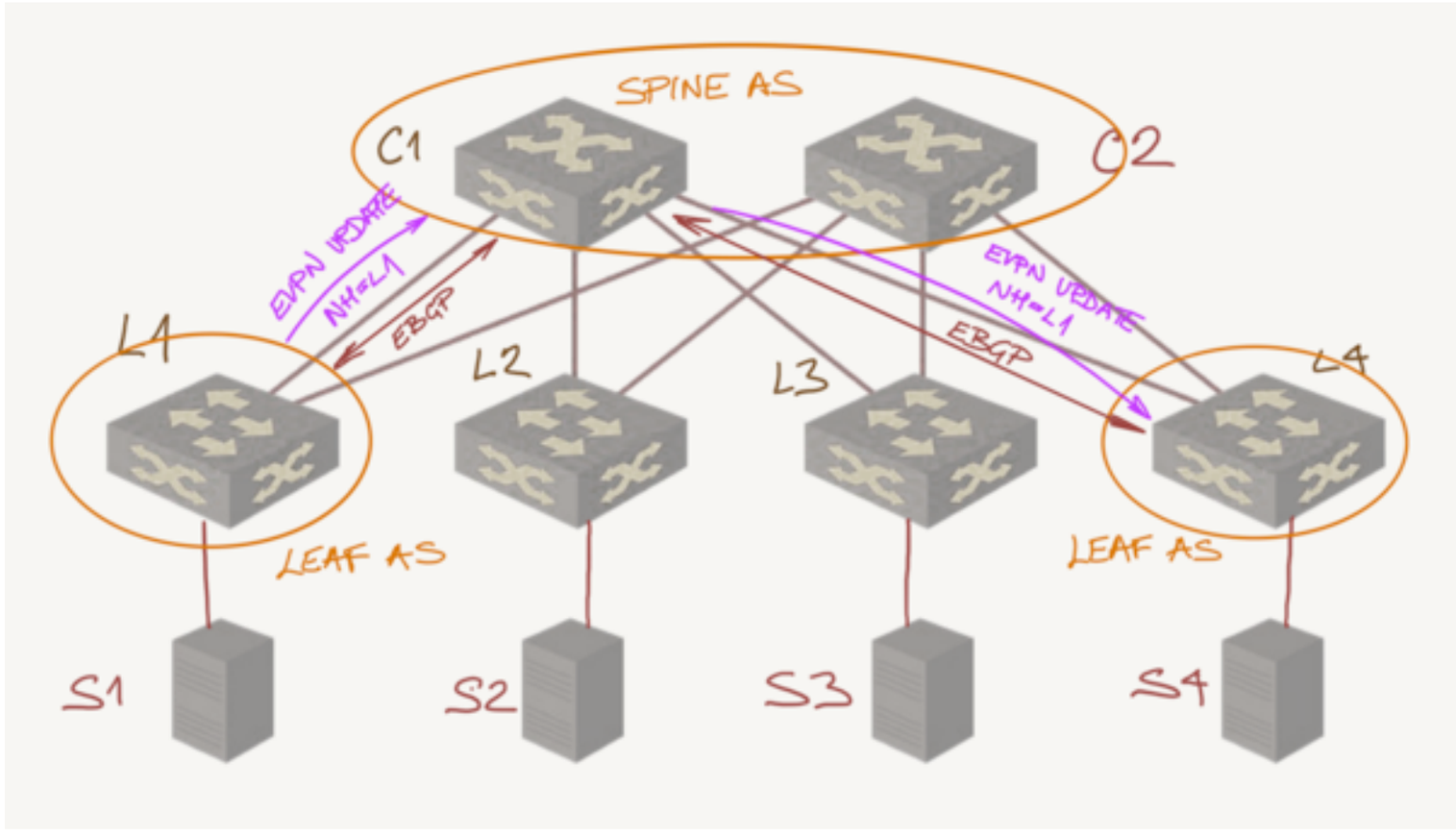
- Imagine **a leaf switch advertising a prefix**:
  - It **advertises the prefix to all spine** switches;
  - **Spine** switches **advertise the prefix to all other leaf switches**;
  - **Leaf switches advertise** their best BGP path **to all spine switches**.
  - Every single **spine switch installs all the alternate BGP paths** received from all leaf switches in BGP table…
- on most spine switches **you'll see N entries for every single prefix in the BGP table** (where N is the number of leaf switches)



Running EVPN over EBGP

- Now you know why smart network architects **use the same AS number on all spine switches** and why RFC 7938 recommends it.

- Finally, it's **interesting to note that using iBGP without IGP**, with spine switches being BGP route reflectors (and some additional configuration tweaks), **results in exactly the same behaviour**.



Running EVPN over eBGP

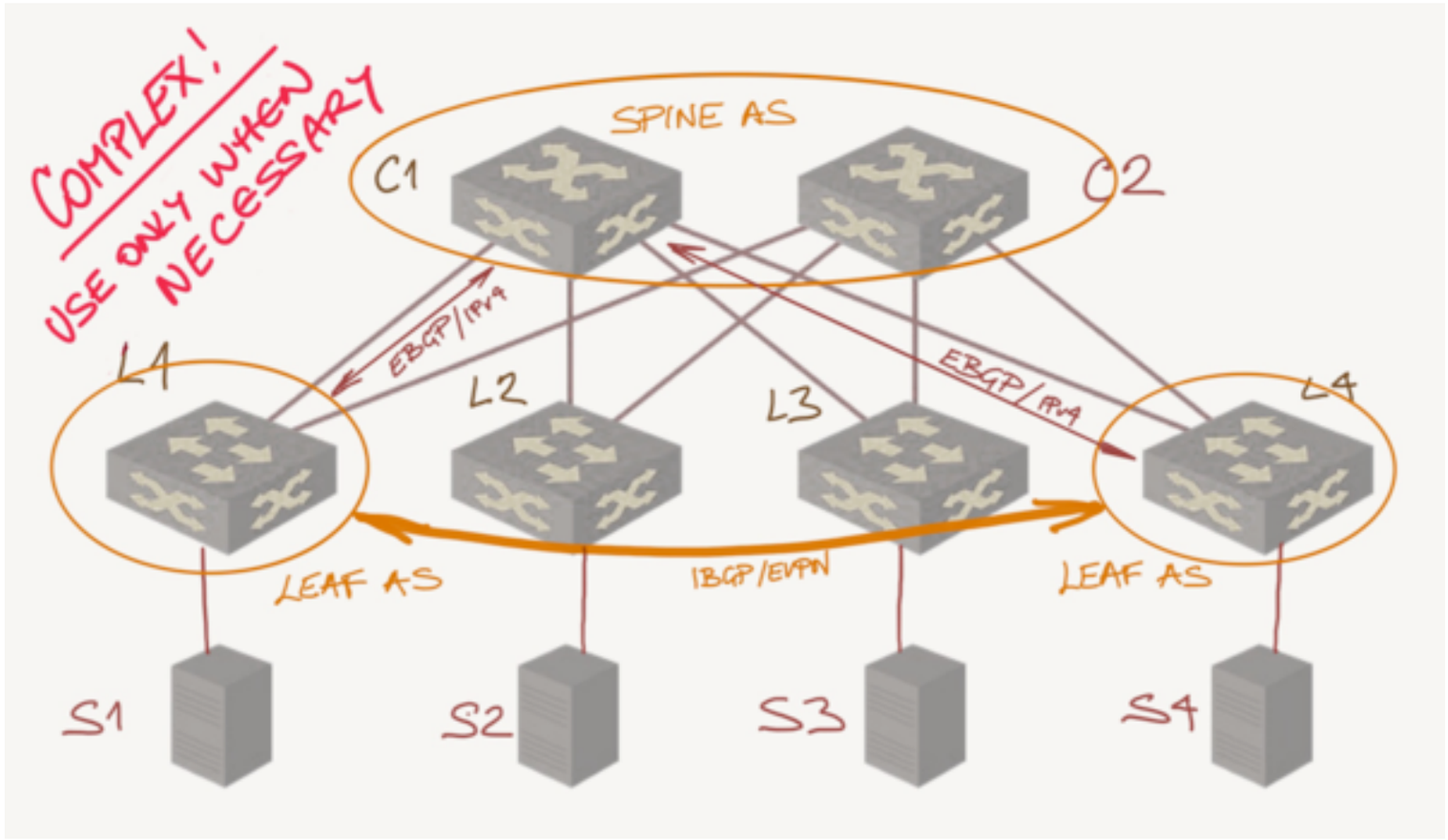
- if you want to exchange EVPN updates across eBGP sessions within a data center fabric, the **EVPN implementation must support retaining the original value of BGP next hop** on outgoing eBGP updates

- **eBGP only design should be avoided** for one simple reason - **why bother your spine switches with EVPN routes from all connected leaf switches?** (and there are a LOT of routes in EVPN).

- there is **no clear separation of underlay and overlay routing in this design** – only one BGP protocol which uses single session for transport of both address-families.

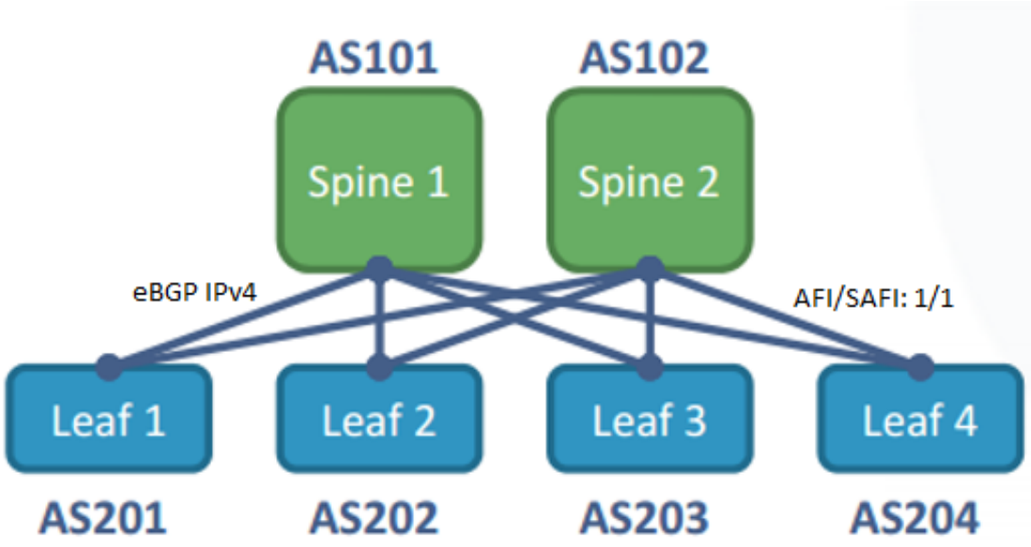
iBGP-Based EVPN on Top of eBGP-Based Fabric Routing

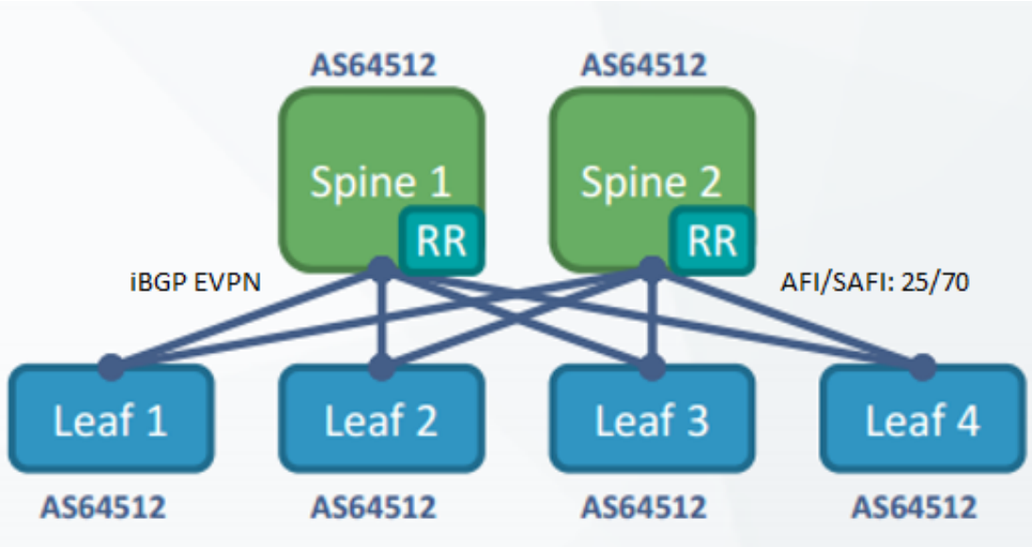
- **These designs are more complex and harder to troubleshoot** than simpler IBGP+IGP or EBGP-only designs but **appears to provide the most scale and flexibility**.
  - **These designs should be avoided unless** you have a **very large fabric** (making IGP impractical) **or a very large number of EVPN prefixes** that would exceed the **control-plane (CPU/main memory) capability of the spine** switches.



iBGP-Based EVPN on Top of eBGP-Based Fabric Routing

- The **main complexity** of this design is the fact that **every leaf switch need "to be" in two different BGP AS simultaneously** (or, better said, present different AS numbers to different BGP peers) – **unique-per-switch underlay AS number to eBGP underlay peers** and **common iBGP AS number to iBGP overlay peers**.
  - you need to use "local-as" option





BGP as a Data Center Fabric Routing Protocol

- **Very large data center operators** (including Microsoft and Facebook) build their data center fabrics **using BGP as the only routing protocol** (documented in RFC 7938),

- resulting in a fashionable trend of using BGP in a leaf-and-spine fabric **regardless of whether it makes sense or not.**

What Problem would BGP Solve?

- **BGP** is still the only answer for large-scale data center operators, but **might be overkill for smaller deployments**.

- **You might need BGP to implement EVPN** or MPLS/VPN and run a BGP-only fabric or use BGP in combination with IGP

- If your fabric won't grow beyond **a few dozen switches**, it **doesn't matter whether you use BGP, OSPF or IS-IS** as the fabric routing protocol – **all of them will work well.**

  - **If your fabric is big enough** that you'd have to consider OSPF areas or multi-level IS-IS, **go with BGP**.

BGP Configuration Complexity

- **Some vendor's** EVPN implementation **does not support the necessary eBGP features** and does **result in overly complex configuration**.

- Cumulus Networks engineers did a wonderful job minimizing the complexity of BGP configuration in FRRouting protocol suite. Their improvements include:

  - **Running BGP across unnumbered IPv4 interfaces** (using IPv6 link-local addresses to establish BGP sessions);

  - **Configuring BGP neighbors using interface names instead of neighbor IP addresses**;

  - **Configuring BGP neighbors as internal or external, and learning remote AS number** during BGP session establishment time;

  - **Advertising BGP router's name during BGP session establishment**, and using neighbor's name (instead of IP address) in all printouts.

- **Most other vendors still use traditional configuration** methods from the days when BGP was used solely to implement complex routing policies between service providers.

Using MPLS+EVPN in Data Center Fabrics

- There's a **fundamental difference between MPLS- and VXLAN-based transport**: the amount of coupling between edge and core devices.

- **MPLS-based VPN solutions require an end-to-end LSP** (virtual circuit) between edge devices **that has to be set up on every hop of the way** and coordinated between edge and core devices using whatever control-plane protocol you use for MPLS.

- **The LSP also has to be kept operational throughout various network failures**, and the changes signaled to the edge devices.

- **You'll have to deal with two encapsulations (IP and MPLS)**, two sets of forwarding tables (FIB and LFIB), and additional control-plane protocols – LDP, MPLS-TE, BGP IPv4+labels or segment routing.

- Admittedly, **the MPLS encapsulation introduces lower overhead** than whatever-over-IP encapsulation. **That overhead becomes relevant when the bandwidth becomes expensive: in WAN networks, not in data centers**.

Using VXLAN+EVPN in Data Center Fabrics

- **VXLAN-based VPN solutions require nothing more than IP connectivity** between edge devices. The edge devices don't have to participate in the core control-plane protocol (apart from using ARP) and the **changes in the transport core are not signaled to the edge**.

- There is almost **no state sharing between edge and core nodes** in VXLAN, and no per-edge-node state in the core, **making VXLAN more robust and easier to scale**.

- Finally, the **tight coupling of edge and core nodes in MPLS** gives you the **ability to do traffic engineering between fabric edges**.

Going Further (French)

- MISCmag 84 : EXTENSION DE LAN
  https://afenioux.fr/doc/presentations/MISCmag84.pdf

- Routage L3 jusqu'à l'hyperviseur avec BGP
  https://vincent.bernat.ch/fr/blog/2018-routage-l3-hyperviseur

- VXLAN: BGP EVPN avec FRR
  https://vincent.bernat.ch/fr/blog/2017-vxlan-bgp-evpn

Going Further

- https://www.ipspace.net/Data_Center_BGP/Autonomous_Systems_and_AS_Numbers
- https://blog.ipspace.net/2018/11/using-mplsevpn-in-data-center-fabrics.html
- https://www.ipspace.net/Data_Center_BGP/BGP_Fabric_Routing_Protocol
- https://www.ipspace.net/Data_Center_BGP/BGP_in_EVPN-Based_Data_Center_Fabrics
- https://blog.ipspace.net/2018/05/is-ospf-or-is-is-good-enough-for-my.html
- https://blog.ipspace.net/2018/02/using-evpn-in-very-small-data-center.html

Going Further

- https://www.juniper.net/us/en/training/jnbooks/day-one/data-center-technologies/data-center-deployment-evpn-vxlan/
- http://jncie.tech/2018/01/28/bgp-design-options-for-evpn-in-data-center-fabrics/
- http://www.trex.fi/2014/xtrm-trill-vs-spb.pdf
- https://learningnetwork.cisco.com/thread/68193
- https://www.juniper.net/documentation/en_US/junos/topics/concept/evpn-vxlan-data-plane-encapsulation.html

Going Further

- Cisco Live!
  VXLAN EVPN Fabric and automation using Ansible
  https://clnv.s3.amazonaws.com/2018/eur/pdf/LTRDCN-1572.pdf

- Cisco Live!
  VXLAN EVPN Day2 Operation
  https://clnv.s3.amazonaws.com/2017/usa/pdf/BRKDCN-2450.pdf

- Cisco Live!
  Building Data Center Networks with VXLAN EVPN Overlays
  https://clnv.s3.amazonaws.com/2018/eur/pdf/BRKDCT-2949.pdf