Scaleway approach to VXLAN EVPN Fabric

Pavel Lunin, Scaleway

plunin@scaleway.com

Scaleway

About Scaleway

- French company (ONLINE SAS), part of Iliad Group
- Europe’s key player on the cloud market
- Three main business lines
  - Virtual instances and surrounding ecosystem
  - Bare Metal Instances (Online Dedibox by Scaleway)
  - DC Buisiness in Paris region (Illiad-Datacenter by Scaleway)

Our Infrastructure

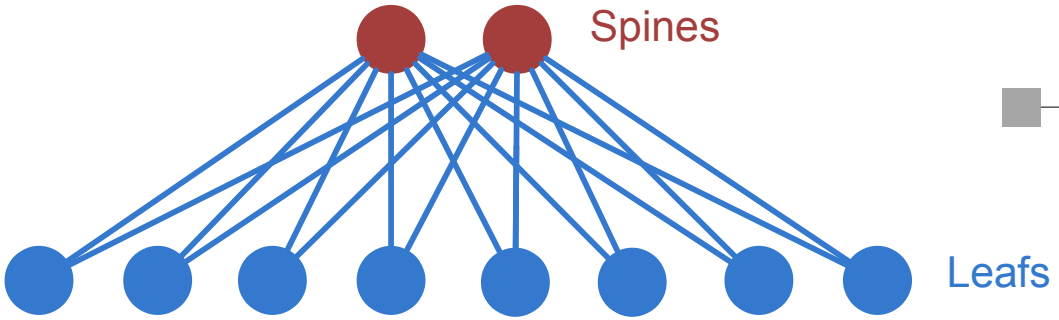
- ~100k physical server in 5 datacenters
- 3.5 Tbps of internet traffic
- Several platforms, 3-4 DC network architectures
- Thouthands of switches
  - Vendor-based as well as home-made
- Some platforms were already IP Fabric-based
- Some platforms needed redesign

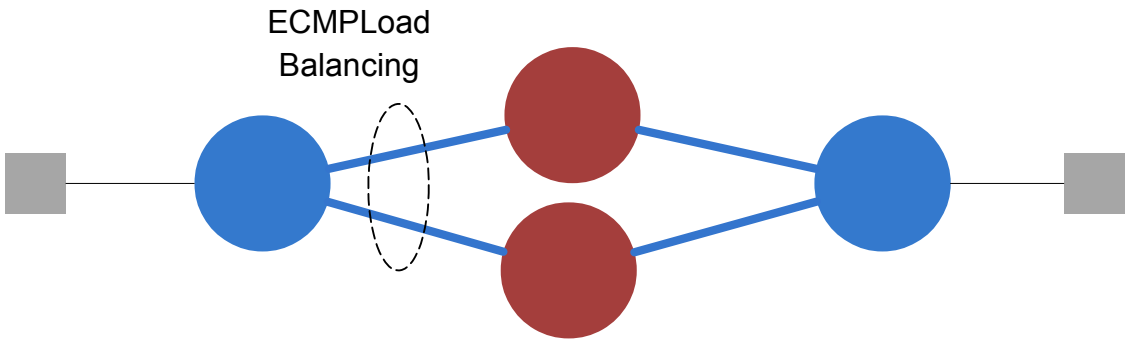
Hunt For for the Best Fabric

- Multi-service:
  - Internet
  - Block Storage for Hypervisiors
  - Infrastructure VRFs: IPMI, PXE etc
  - Should be easy to add more later
- Scalable upto at least 100k VMs per fabric
- Fast to design and deploy:
  - Needed an approch compatible with the available off-the-shelf products

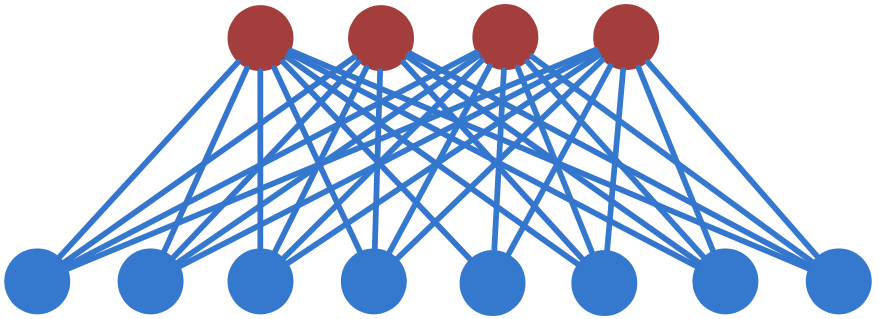
IP Fabric – the Basics

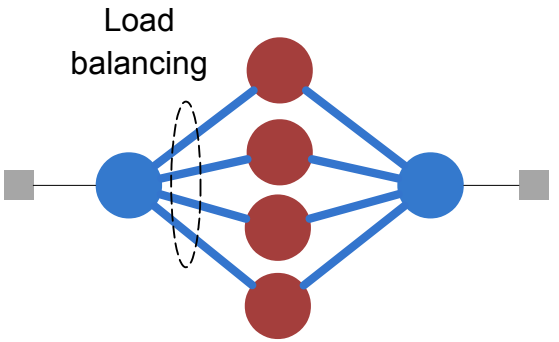
IP Fabric – 3 Stage Clos

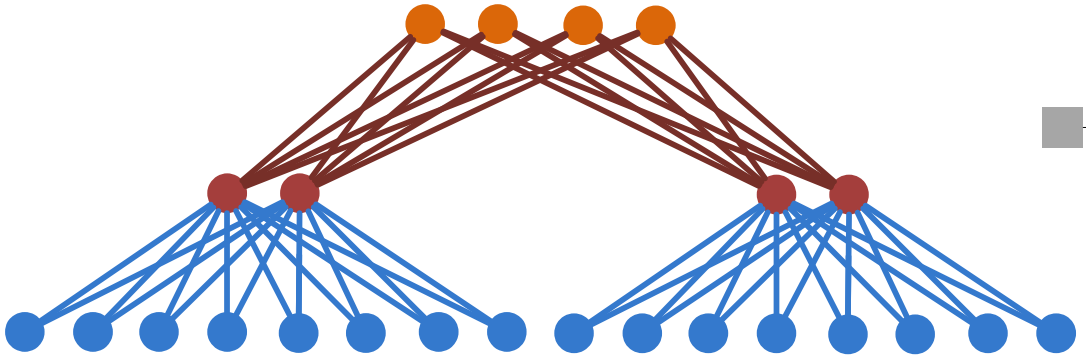




IP Fabric – Scaling Clos







Underlay

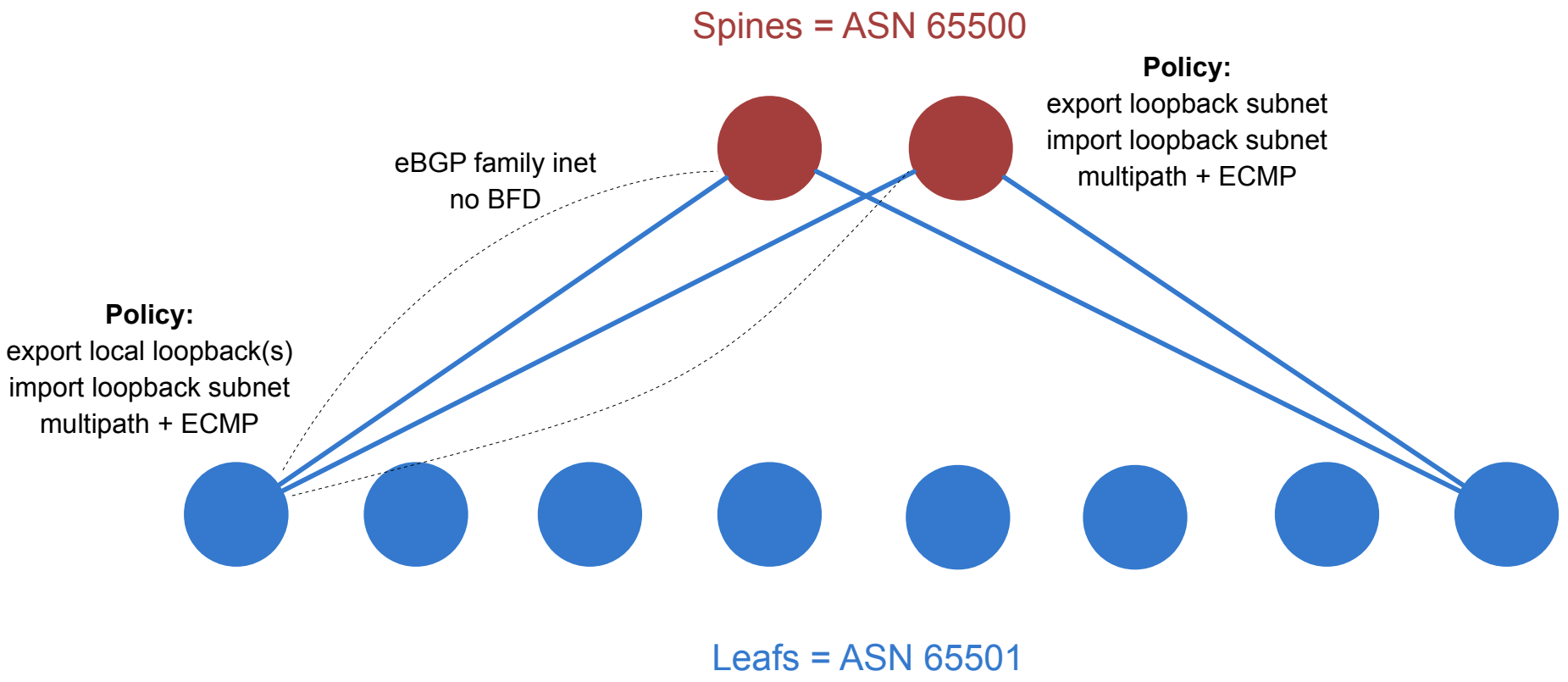
First Step - Underlay

- First we just need ping between hosts, connected to the fabric
- A very basic IP connectivity
- Known as Undelay

RFC7938 aka Draft Lapukhov

- RFC7938 is must read before you begin
- eBGP is the best IGP
- Internet-like routing, known to scale
- Challenges:
  - One ASN per level or one ASN per node
  - What to re-announce: loopbacks only or links as well
  - BFD?
  - Theese 3 questions are well discussed in the RFC

Underlay: What We Do



Underlay – IPv6?

- Options:
  - IPv4 only
  - IPv6 only
  - Dual-stack
- Reality:
  - Most available DC hardware still doesn't support IPv6 underlay for VXLAN
  - IPv6 underlay works just fine over IPv4 underlay

Underlay – Addressing

- *Yakov Rekhter's Law: «Addressing can follow topology or topology can follow addressing. Choose one»*
- In other words:
  - Allocate the next available prefix for each link (Internet-like)
  - Encode topology into the addressing (like phone numbers, post codes or office room numbering)

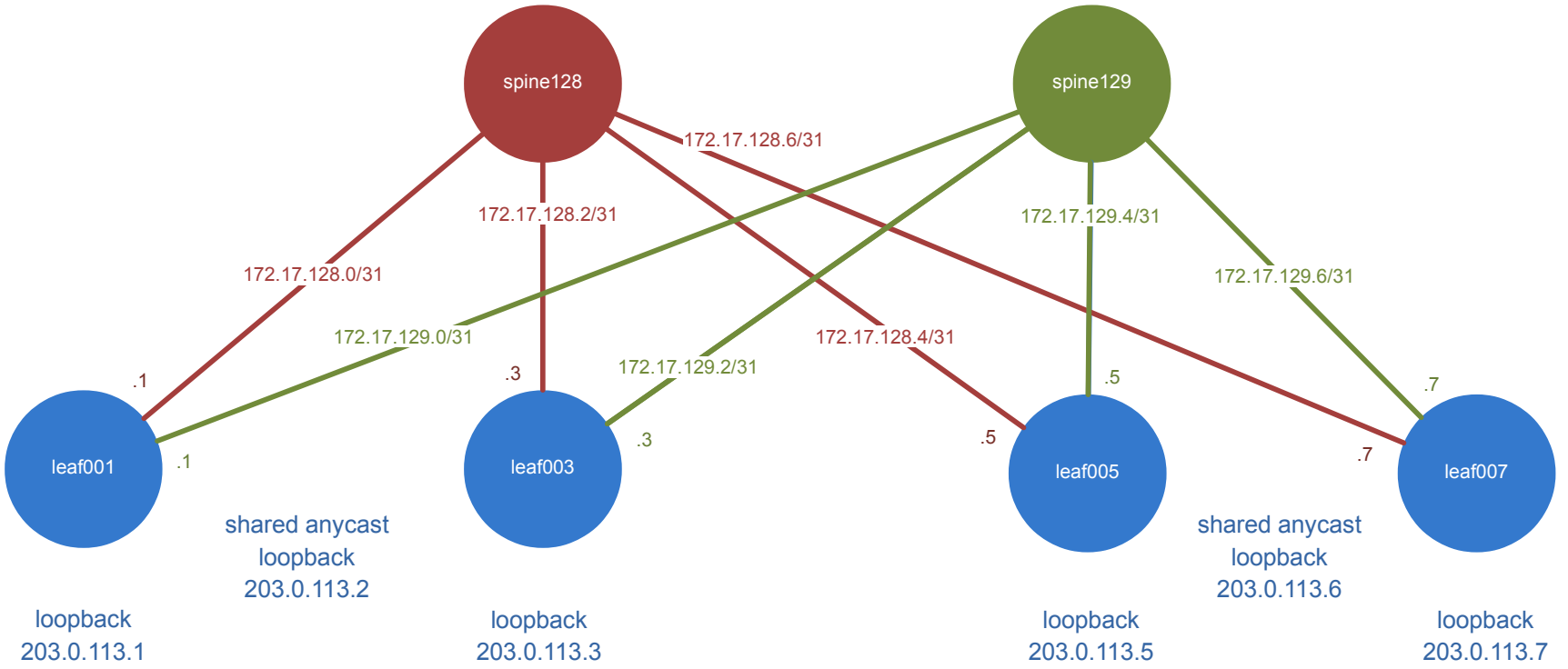
Internet-Like Addressing

- Pros:
  - In theory, up to 100% numbering space utilization
  - Works well for flexible/undefined topologies
  - What most IP people are used to
- Cons:
  - Requires strong integration with IPAM/registry
  - Not human-friendly (no encoded semantics), error prone
  - Practically address space is never really 100% utilized

Topology-Driven Addressing

- Pros:
  - Auto-defined addressing. You only need to choose a prefix for the whole fabric. The rest can be calculated locally
  - No need for an external registry to address each link
  - Human-friendly: node ID, stage ID, roles and other information are encoded into the address
- Cons:
  - Some numbering resources are wasted by design
  - Some people are confused when they see it for the first time
  - Don't even think of using this for the Internet!

Underlay Addressing: Conclusion



Underlay: How We Do It

- One ASN per stage per fabric
  - I. e. same ASN on all leafs in the same DC
- Only loopbacks are reannounced
  - Leafs don't see point-to-point links of other leafs
- No BFD
- Topology-driven addressing
- IPv4-only
- Public IPv4 for leaf loopbaks
  - To ease inter-fabric DCI
  - Not routed from/to the Internet (VXLAN is too simple to hijack)

Overlay

The Need for Overlay

- OK, we can forward IP packets from one network port to another:
  - Scalable, resilient to failures, easy to deploy and maintain
- But we need multiservice:
  - Internet, ADM, Block Storage

Services Need Isolation

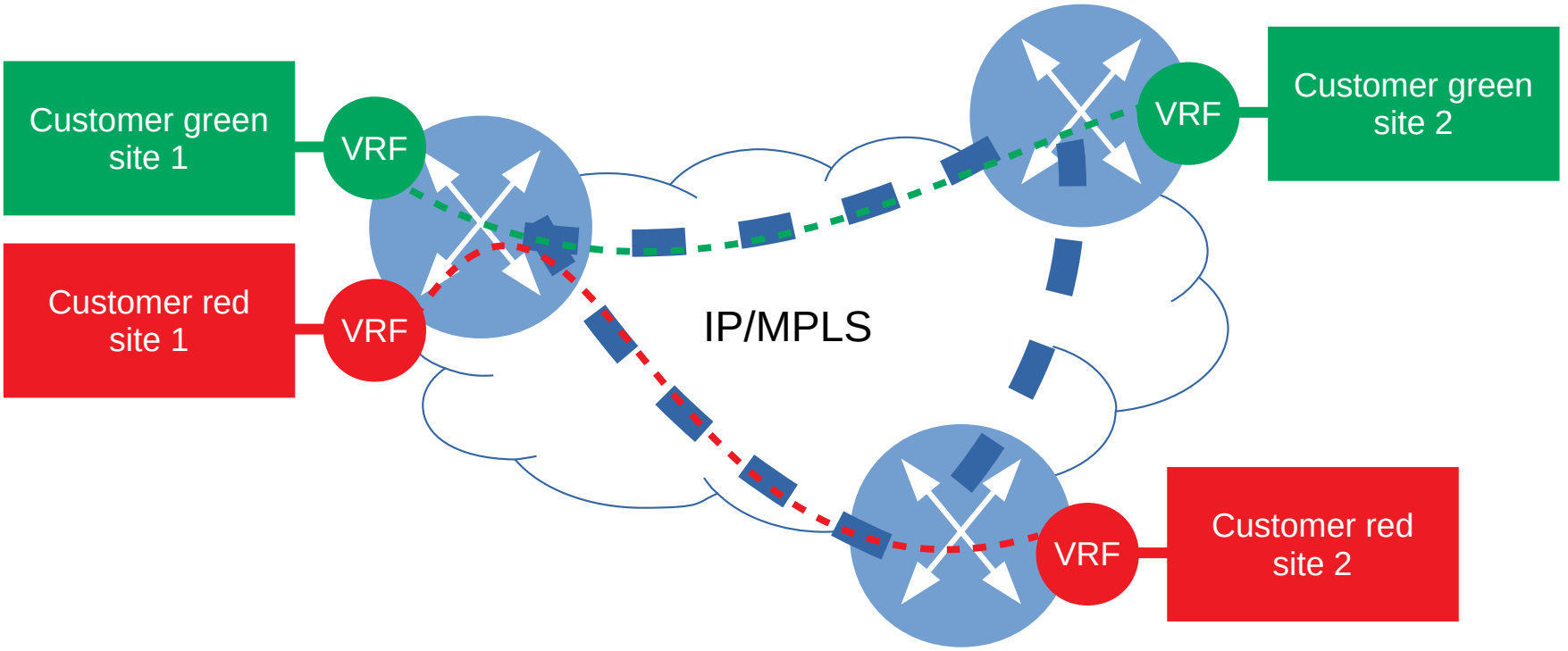
- Overlapping IP space
- Security: hosts in the Internet VRF should not have access to block storage
- Special needs: greater MTU for Storage
- Different scaling scope:
  - Some services need per VM forwarding state
  - Others only need per HV

Things Get Complicated

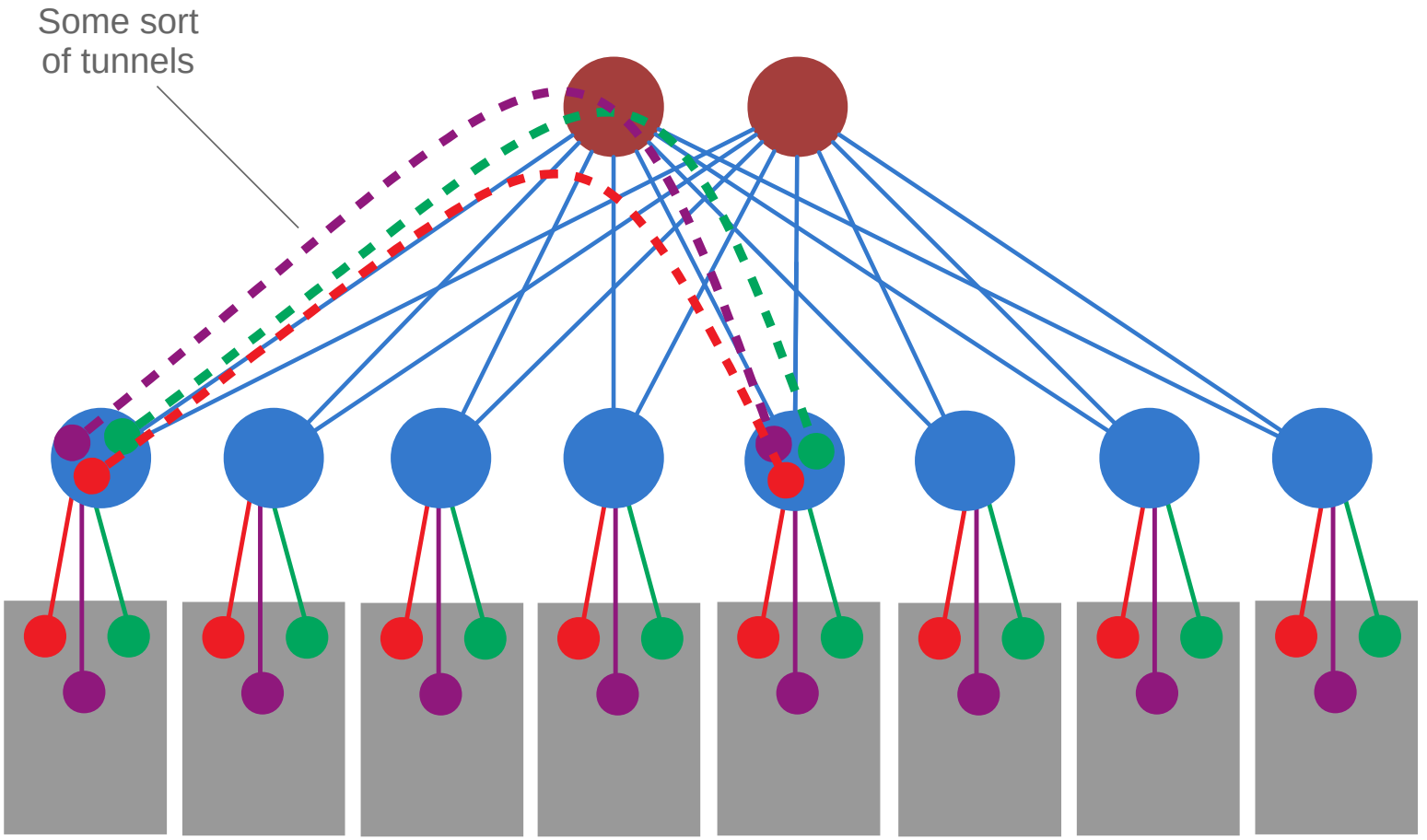
- Overlay: well known concept used in Carrier Networks
- As vendors (hello Cisco and Juniper) didn't want MPLS in the DC, the industry has reinvented the weel
- VXLAN dataplane:
  - Ugly:

    IP over Ethernet over VXLAN over UDP over IP over Ethernet
  - But works
  - https://tools.ietf.org/html/rfc7348

Overlay Principle



Overlay on Top of the DC Fabric



Beyond VXLAN

- The network needs a control plane technique to know how to forward traffic over VXLAN tunnels
- Initially proposed multicast-based VXLAN control plane just doesn't scale
- The industry adopted EVPN BGP for VXLAN

It's All About BGP

- Carrier MPLS networks have all the needed service signalling control plane tools:
  - L3VPN, L2VPN / pseudowires, VPLS, MVPN, many more
- In production for decades
- A lot of literature available, a lot of people know how it works
- It's all about BGP (iBGP, to be precise)

Lemma #1
(The Main Truth)

A good DC fabric starts by learning the MPLS VPN theory

BGP EVPN

- Yet another BGP family
- Initially developed for MPLS carrier services, later adopted for VXLAN
- Initially developed for bridged L2VPN applications
  - «VPLS on steroids»
- Mainly known for its «MAC addresses announced with BGP» feature

EVPN is Not VPLS on Steroids

- Well, not only...
- It can also signal L3VPN (both IPv4 and IPv6)
  - https://tools.ietf.org/html/draft-ietf-bess-evpn-prefix-advertisement-11
- In contrast to classic L3VPN or L2VPN BGP families, it's not strictly limited to MPLS data-plane
- VXLAN is one of the possible transports

VXLAN is not Virtual Extensible LAN

- Not anymore…
- It's just a tunnel over IP encapsulation. Much like GRE, GENEVE, L2TP and others
  - You still need ethernet inside but it's just a matter of encapsulation details
- In theory, you can use any control plane to signal the data-plane state. For example EVPN BGP
- You can route IPv4/v6 over VXLAN. No need to bother with BOM replication etc

EVPN VXLAN Toolset

- The most popular data-plane/control-plane options in the modern DC
- Most vendors support it
- Open source as well: FRR, GoBGP etc
- VXLAN terminating device is known as VTEP
  - It's like MPLS PE but VXLAN-based
  - Also known as NVE (Network Virtualization Edge)

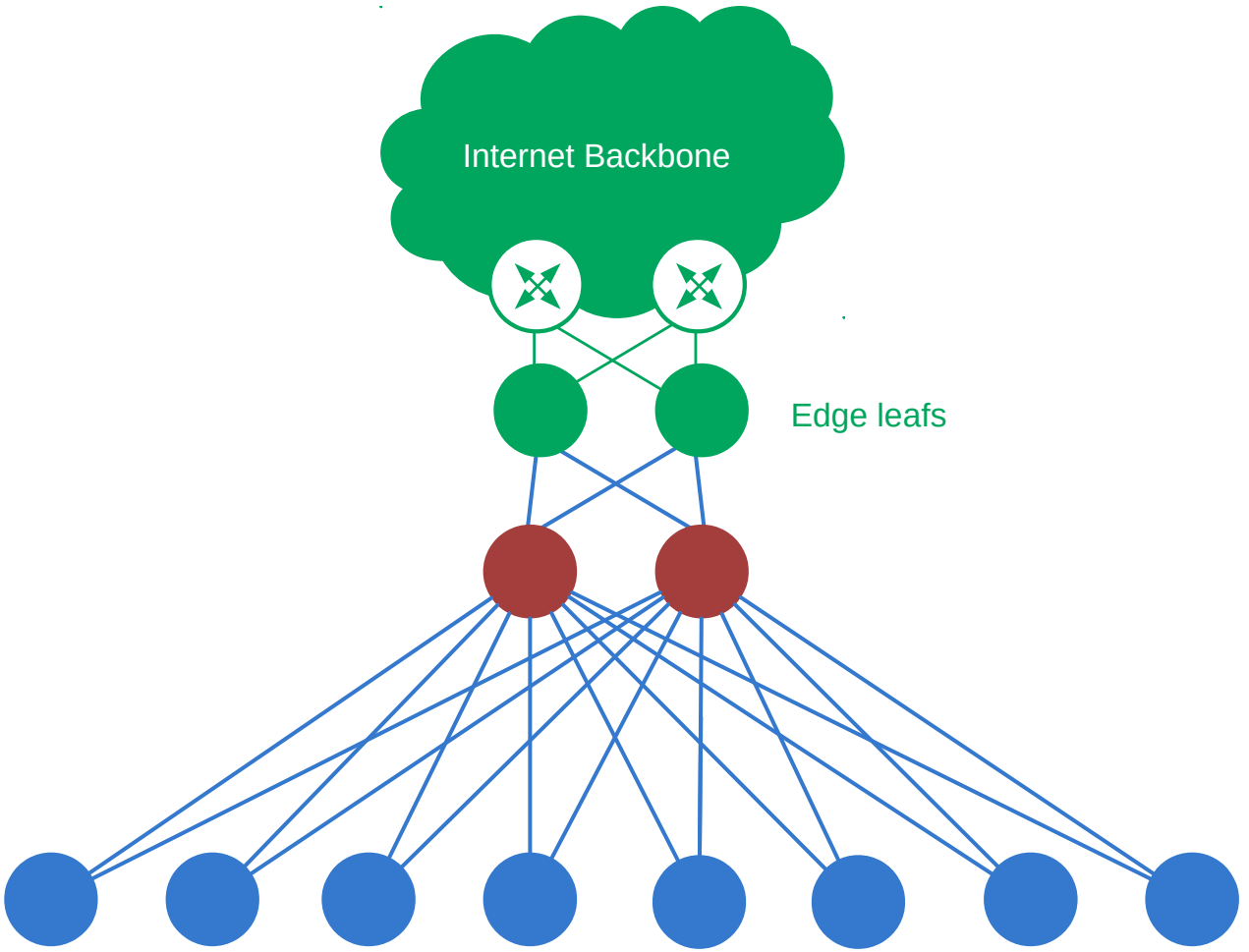
VXLAN L3VPN, the Hardware Challenge

- In order to do L3VPN with VXLAN EVPN, a VTEP must be able to perform an LPM lookup against the internal packet header
- Not all chipsets can do it (e. g. Broadcom Trident 2 can't)
- Some chipsets can only do it from/to a plain IP next-hop
  - They can't route between two VXLAN tunnels
  - E. g. Broadcom Trident 2+
  - It should be acceptable for ToR VTEP in most cases
- Challenge your vendor and PoC everything in advance

Edge Leaf or Not Edge Leaf

- To access the Internet, packets need to exit the overlay and go to plain IP world
- You need some sort of gateway between the fabric overlay and the Internet backbone
  - It can be spines (vendors often propose this option)
  - Or so called edge-leafs, which
    - Have uplinks towards the backbone
    - Announce the aggregate routes to the backbone
    - Announce the default route to the underlay
    - We use this option as we want to keep spines simple

Edge Leaf



It's all iBGP

Just like in the MPLS networks:

- Underlay provides connectivity between leaf loopbacks
- Each leaf has a pair of sessions with iBGP RRs to send and receive EVPN routes
- EVPN routes are used to forward end-point traffic
  - Using leaf loopbacks as iBGP next-hops and VXLAN src/dst

Route Reflectors on the Spines

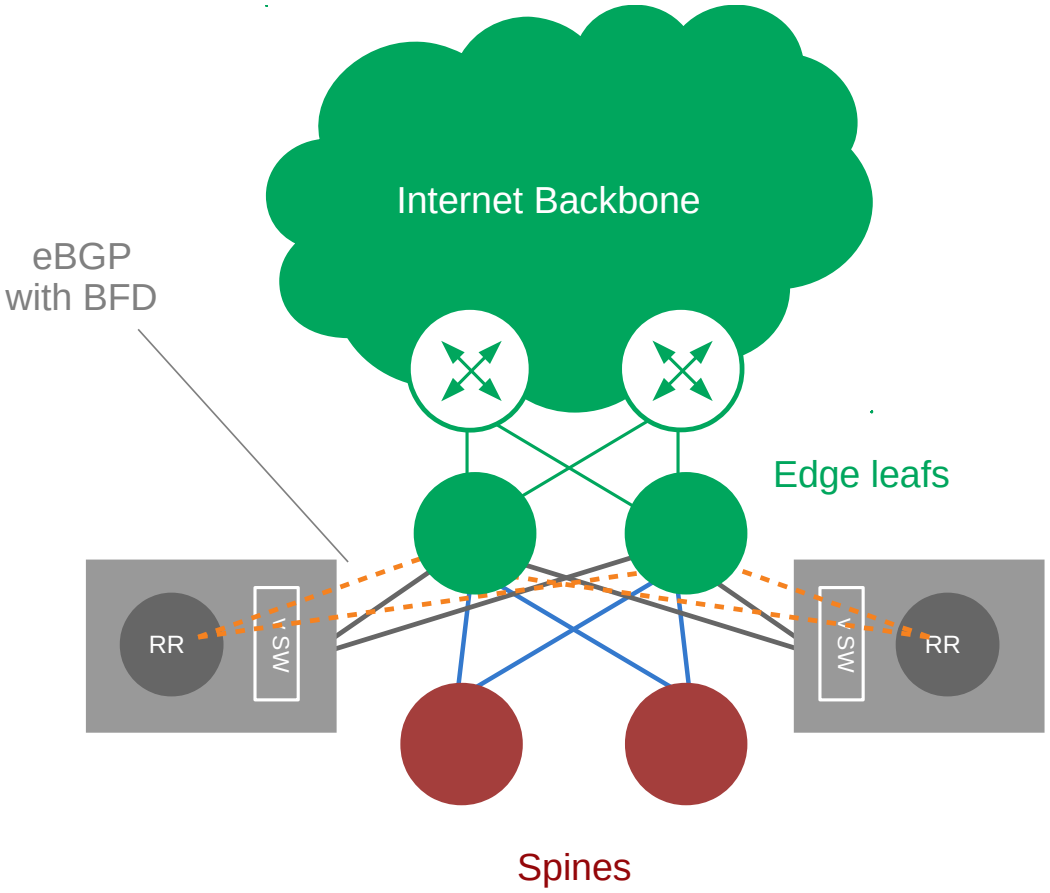
Vendors often recommend this:

- No need for additional hardware/software
- Easy to deploy
- Complexifies spines
- Limits the choice of spines and makes them more expensive
- Need some magic if you don't want EVPN in FIB
- Limited scale
- Harder to upgrade spines

Route Reflectors: VM

- Broader choice of options
- You need to host it «somewhere»
- More difficult than it seems to be
  - Must be reachable in the underlay
  - Must not depend on the overlay (including admin network, IPMI, storage etc)
  - Physical connection type (spines are normally 40G/100G)
  - Must ensure that RR is never used as transit node
  - VM has no direct link (virtual switch)

Route Reflectors: How We Do It



Route Reflectors: How We Do It

- VM on a dedicated KVM server connected to edge-leafs
  - Decision driven by the need of 10G link
  - In future to spines (40/100G)
- Local strorage, local boot, admin access over underlay
- Two L3 p2p links towards uplink switches:
  - VLANs on the hypervisor's virtual switch
  - eBGP to announce RR's loopback (like one more Clos stage)
- BFD supporting these eBGP sessions
  - This is the only place we use BFD
  - No BFD elsewhere, including iBGP between VTEPs and reflectors

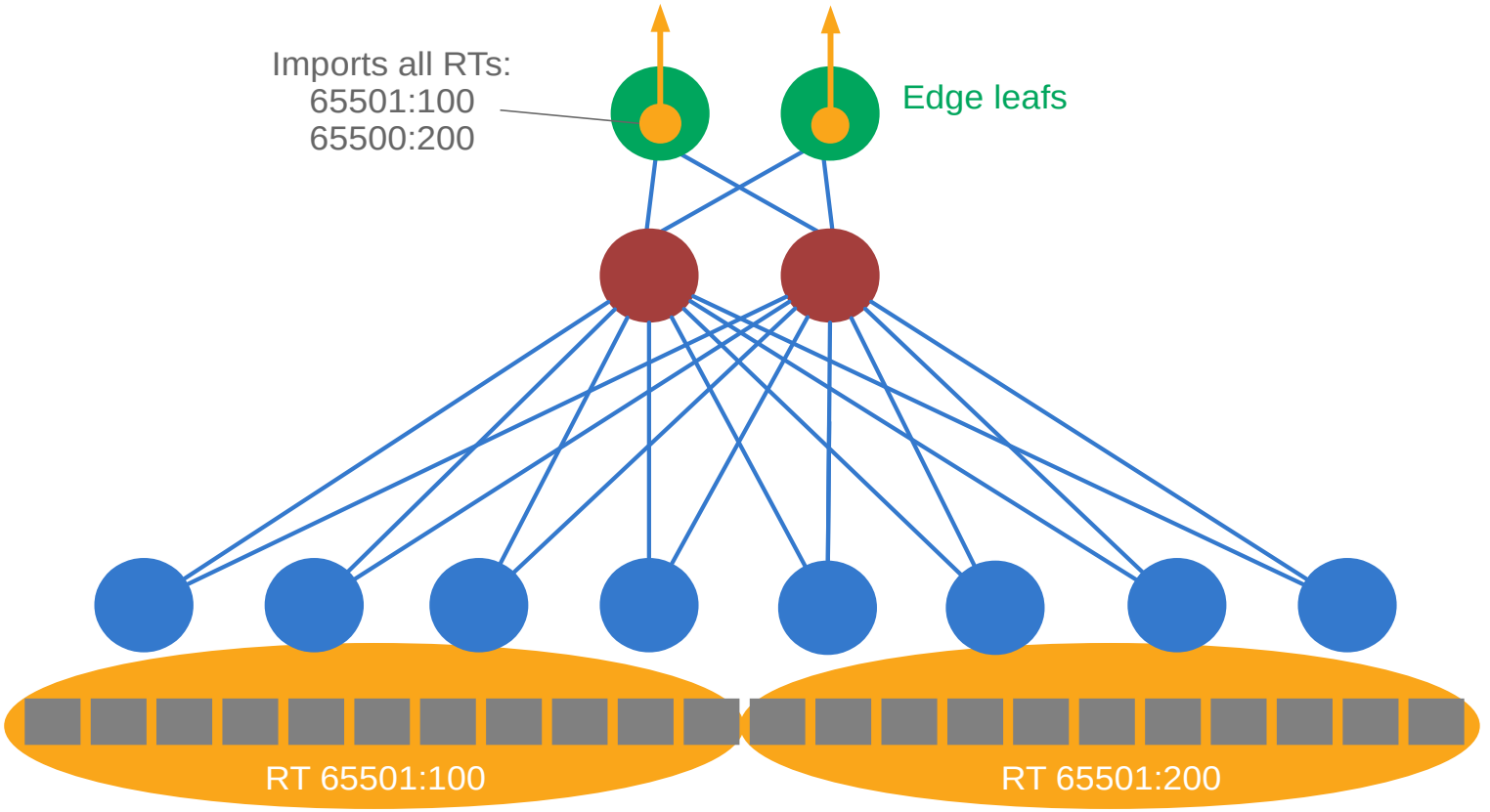
Overlay Scaling

- In some VRFs the number of routes depends on the number of hypervisors
- In others it depends on the number of VMs
  - And it's a challenge
  - We have a lot of VMs and can't have all the routes on all leaf VTEPs

Overlay Scaling: Sharding

- EVPN, like other BGP VPNs, uses route target concept
- Well-known RT-based techniques exist to limit or extend the scope of BGP VPN routes:
  - Hub and Spoke VPN
  - Extranet
  - See, for example, «MPLS Enabled Applications» book for more details
- It's simple to split some VRFs into multiple shardes and then import all routes on a hub device (edge leaf)

VRF Sharding



Multi-Homing Options: Pure L3

Pure L3 with BGP

- Like one more clos stage
- Requires BGP on the host
- As traffic can't be labeled (there is no MPLS and BGP-LU), you need a session per VRF
  - Won't scale

Multi-Homing Options: ESI

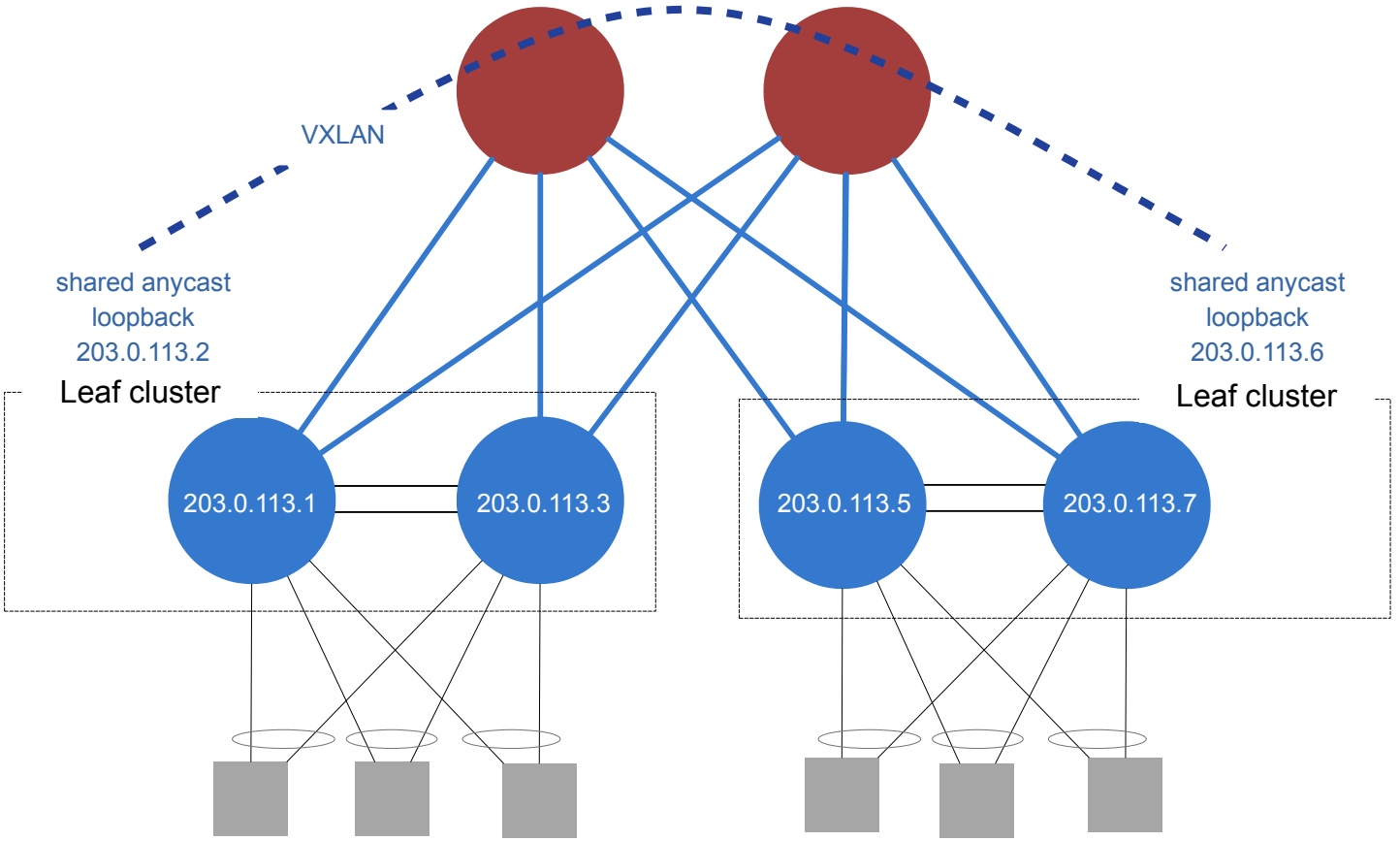
EVPN ESI with MC-LAG light

- MC-LAG light is just the same LACP ID, simple and efficient, no need for ISL link
- ESI is a special EVPN multihoming mechanism
- Flexibe: allows more than two node multihoming, mixed single- and multih-oming, cross pair multihoming etc
- ECMP is done by ingress VTEP between underlay routes
- Vendor support is somewhat poor: not all vendors, not all product lines

Multi-Homing Options: MC-LAG + Anycast VTEP

- Requires a real MC-LAG
  - Vendor specific, often painful
  - Most implementations require an ISL-link
  - Some implementations are not flexible in terms of mixing multi- and single-homed ports
- Two leafs share the same «secondary» loopback IP to announce EVPN routes
  - Decreases the number or overlay routes by 2
- ECMP is performed by spines between underlay routes towards the anycast VTEP
  - Based on VXLAN UDP source port, set by the ingress VTEP as a flow hash key
- Strangely, the most convinient option these days

Anycast VTEP



A Note on Software VTEP

- It's more and more common to terminate overlay directly on the hypervisors
  - E. g. https://ripe77.ripe.net/archives/video/2001/
- It's a matter of good implementation and performance
  - For a commercial cloud platform the CPU ressources is what you'd prefer to customers that spend for infrastructure needs
- We are also starting to do it but in a hypervisor-to-hypervisor fashion, so not yet really integrated with the fabric

Note on Management

- Often vendors recommend OOB management
- Some people love it but...
- Put a dedicated OOB management switch into each racks is a little bit too expensive and cumbersome
  - What about OOB switches to manage hundreeds of OOB switches?
  - Most vendors do crazy things for OOB management (copper only, 10/100 only, can't add default route, can't add to VRF, etc)
  - These things are different from vendor to vendor

Note on Management

- We use underlay for management
  - Routine management, monitoring etc
  - Bootstrap in a remote DC
    - One of the reasons (but not the only one) why Internet is in a VRF
    - Requires attention to default interface speed (40G or 100G)
    - Foreign SFP activation with a hidden command is not an option

Questions

Thank you

Scaleway