Overlay Network Visibility

April 2019

www.cubro.com

CUBRO
NETWORK VISIBILITY
SIMPLE | SCALABLE | SUSTAINABLE

Index

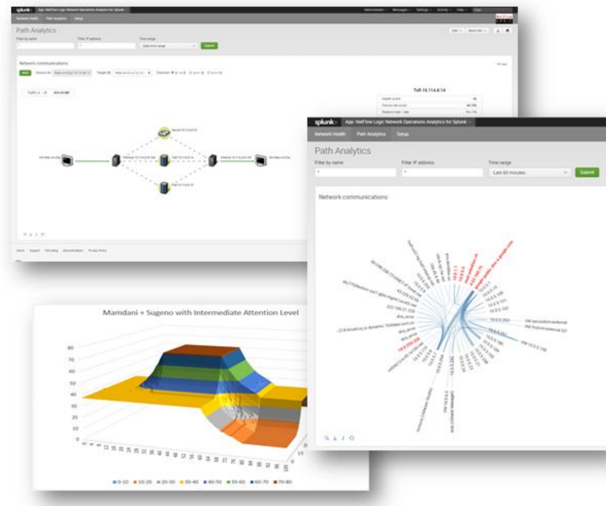
Explanation of overlay networks

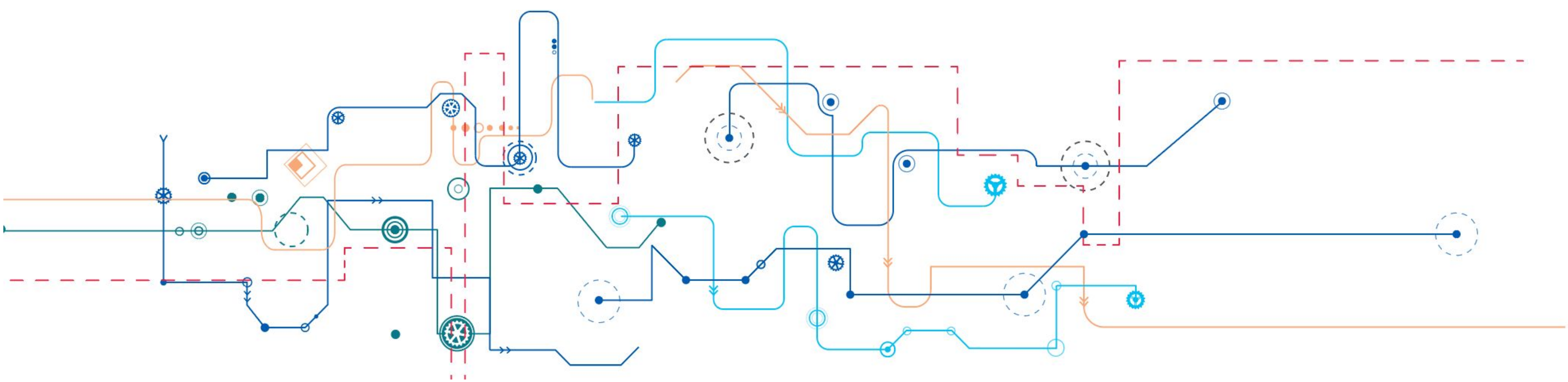
Why and how to do monitoring

Issues with monitoring overlay networks

Cubro’s solution

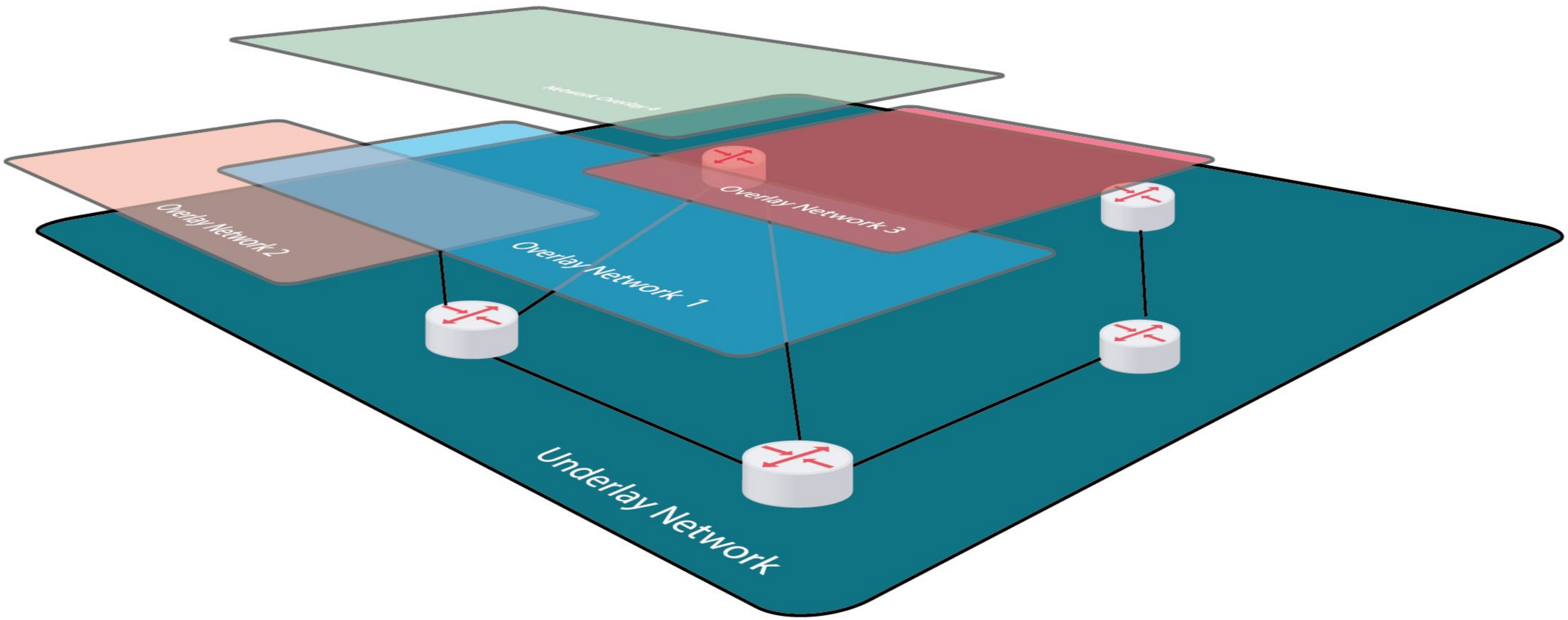
Other typically offered solutions



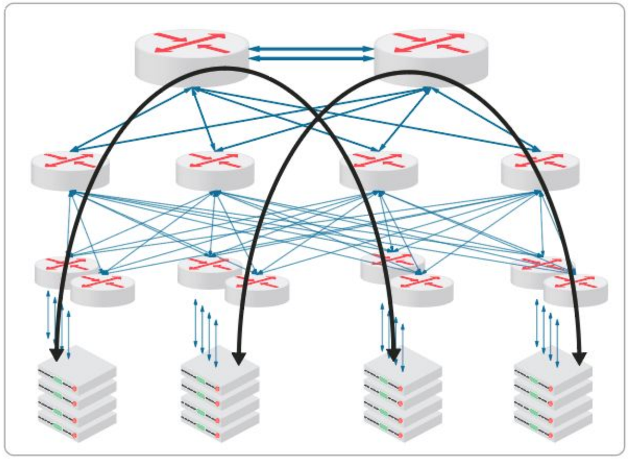


Overlay Networks

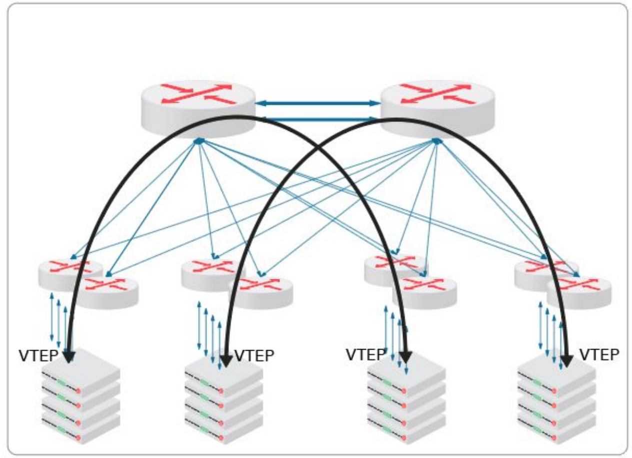
Overlay networks today are a standard in any data center, the only differentiator today versus in the past is that we are talking about hundreds or thousands of overlay networks per data center, with up to multiple thousands of endpoints. Also, overlay networks are more dynamic than in the past such that manual configuration of visibility tools is no longer possible.



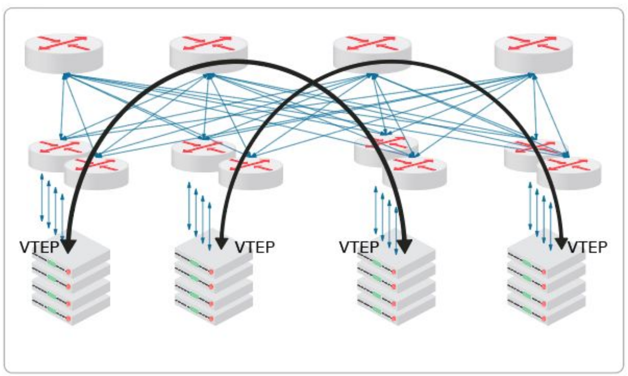
Evolution of Cloud Data Center Networking



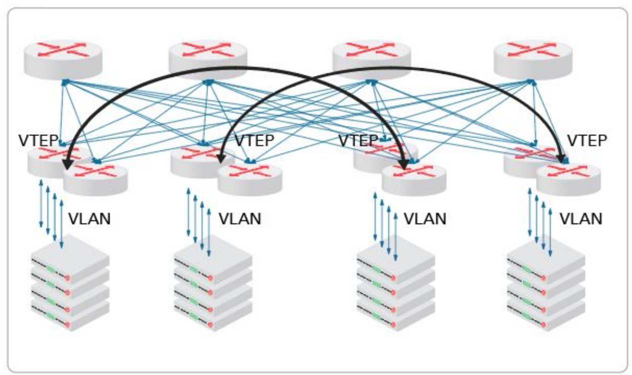
3-Tier Overlay Network



2-Tier Scale-Up Overlay Network



2-Tier Scale-Out Overlay Network



EVPN - Ethernet Virtual Private Network

Ethernet Virtual Private Network (EVPN)[4]: The latest cloud network architecture, proposed by incumbent network mainstream players, offloads the (VXLAN) Virtual Tunnel Endpoint (VTEP) function from compute servers to leaf switches and reduces resource consumption in order to achieve a higher link throughput in compute servers.

□3-Tier Overlay Network: Same 3-tier architecture (core, aggregation and access tiers) as campus networks to provide interconnections for (VXLAN[3]) overlay traffic between compute servers.

□2-Tier Scale-Up Overlay Network: Flatten the network architecture from 3 tiers down to 2 tiers, i.e. tall spine and leaf tiers, to reduce the number of network hops (and therefore, an end to end latency) for overlay traffic between compute servers.

2-Tier Scale-Out Overlay Network: Evolved from 2-tier Scale-Up overlay network by transforming one single (usually much more expansive) tall spine switch system into multiple (more value-efficient) thin spine switch systems in order to provide:

- better redundancy: from 1+1 redundancy in the tall spine architecture to N+1 redundancy in the thin spine architecture
- better scalability: from the small cloud data centers with only a few thin spine switches to the mega cloud data centers with numerous thin spine switches
- ease of logistics: same hardware platform with the pizza-box form factor allows for inter exchange between thin spine switches and leaf switches; therefore, reduce the complexity of the cloud data center asset management and operational logistics

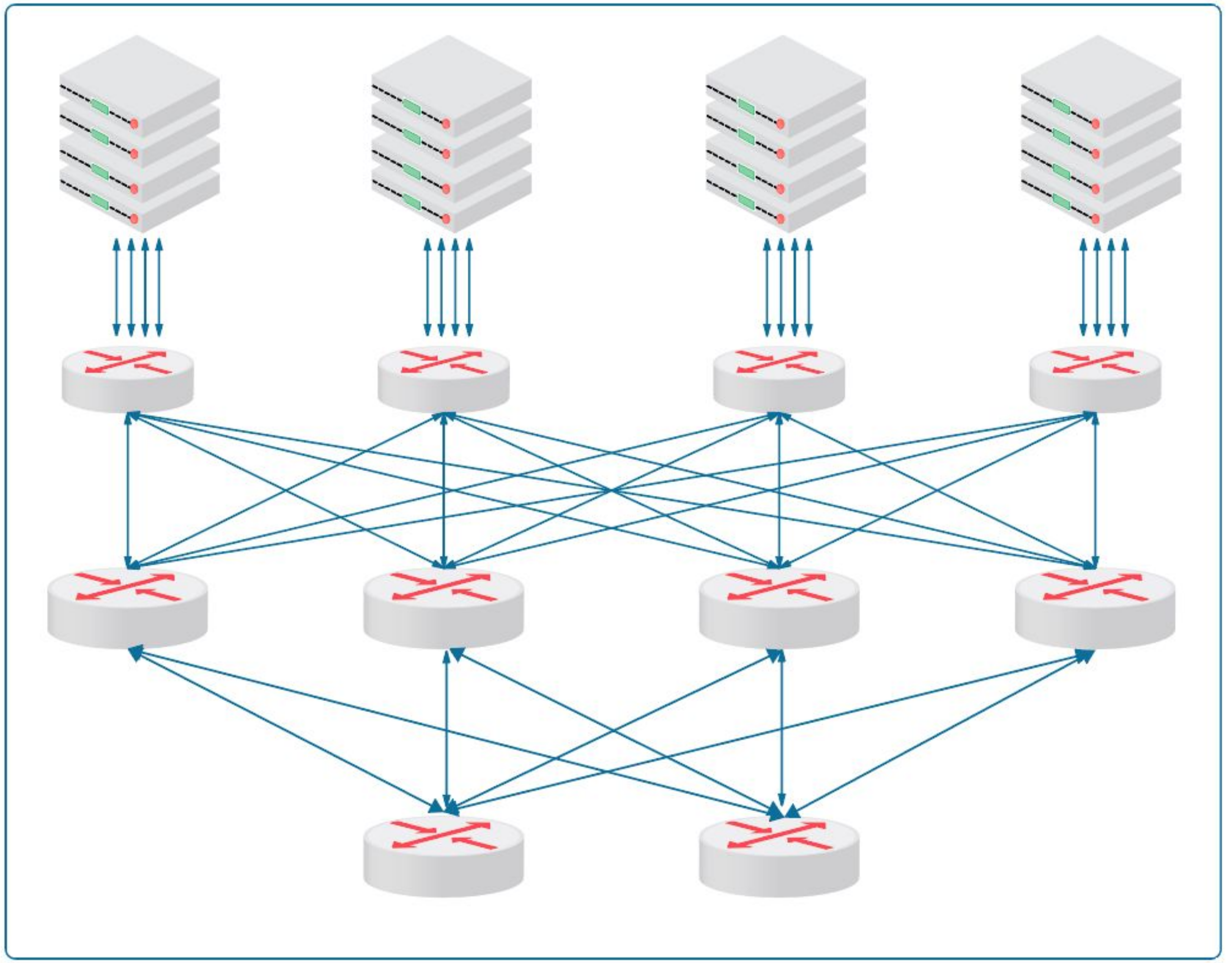
Underlay Data Center Network Design

Server farms

Leaf

Spine

Border Leaf



Underlay Network Design & L2 Overlay

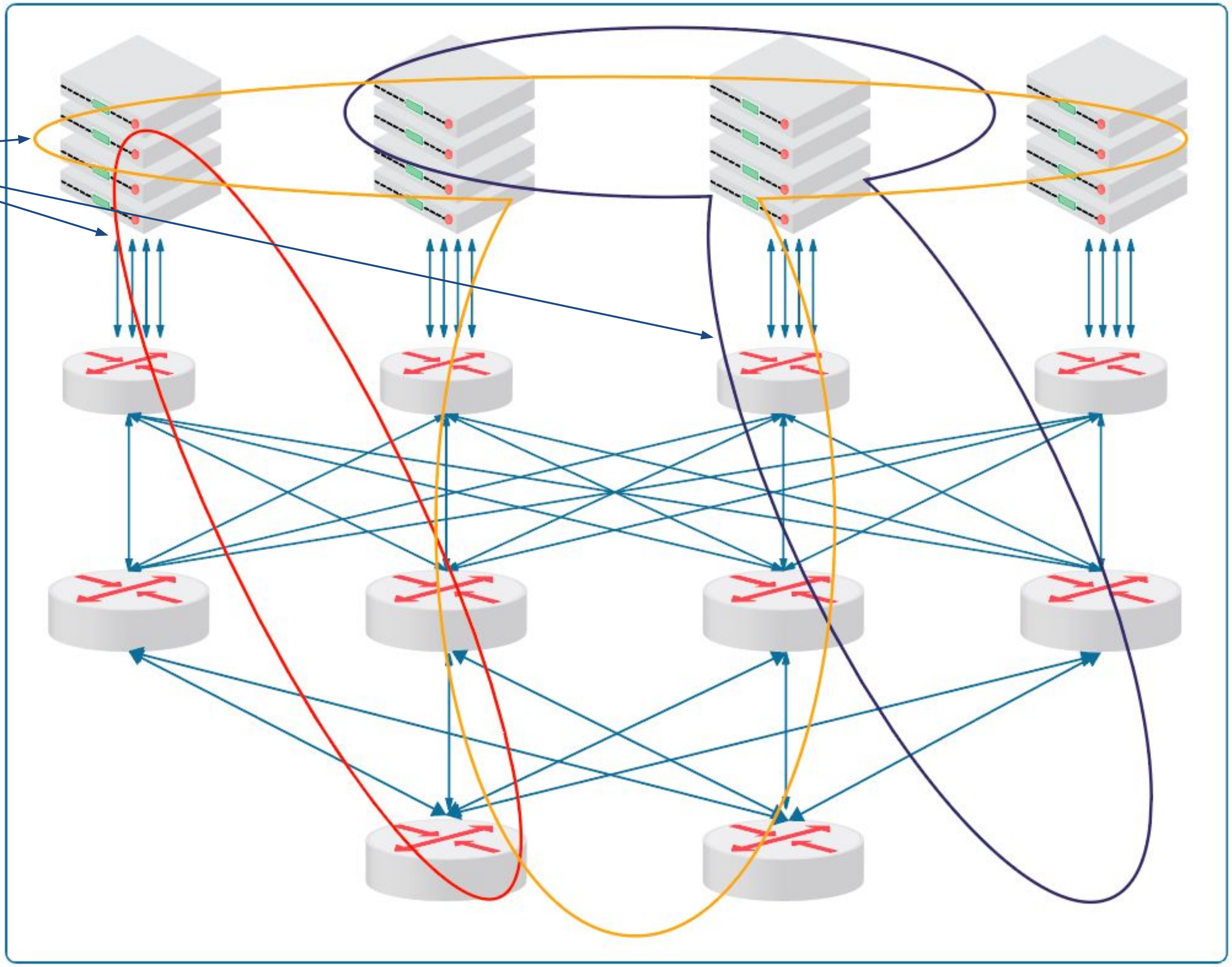
Transparent L2 overlay network

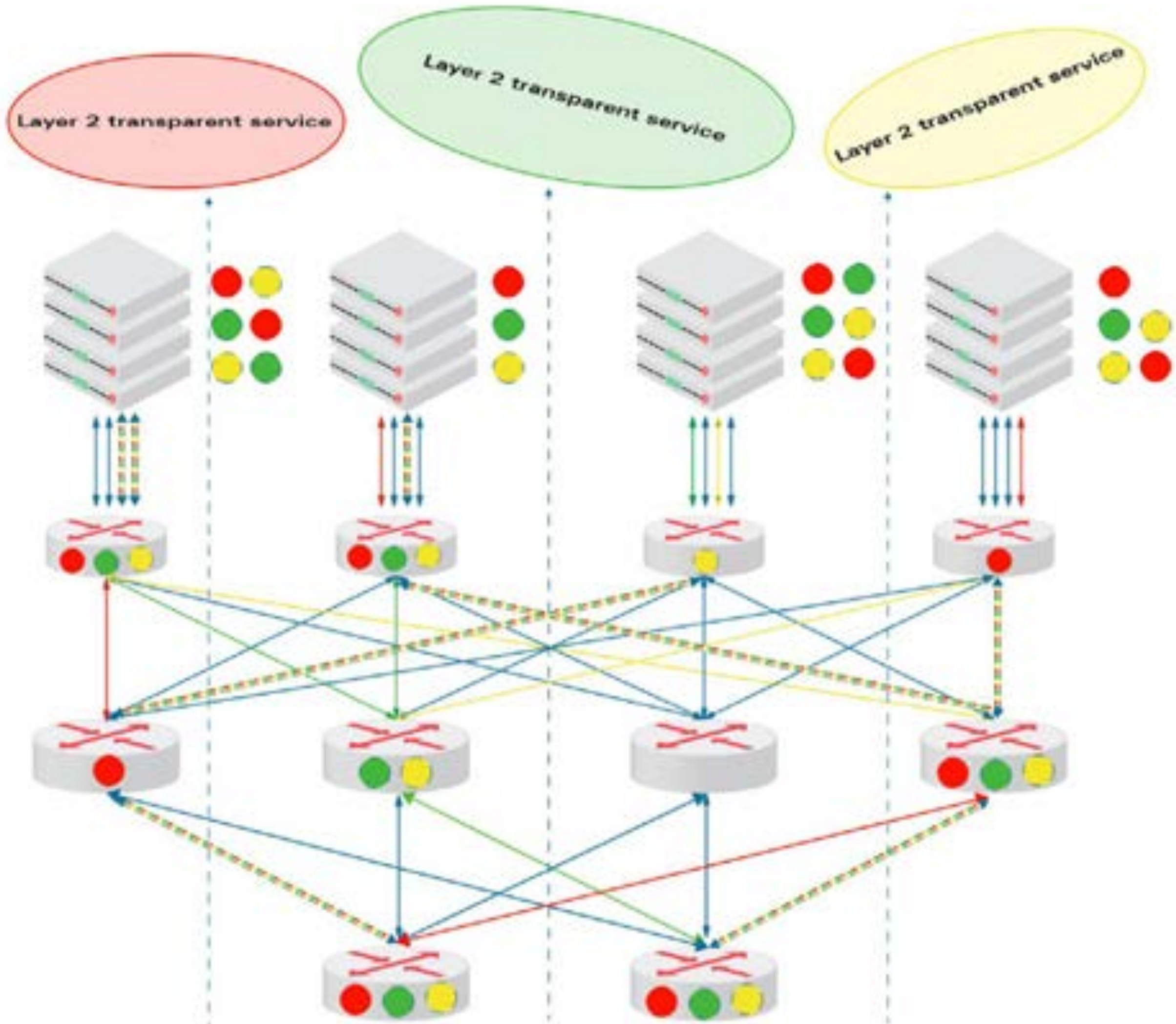
These overlay networks share the same underlay network but for the user it is a fully transparent network.

This is good for the user of the L2 network because they can do whatever they want. For example, use any IP address or any VLAN.

The drawback, however, is the monitoring of the underlay network because you would also see the overlay network.

Additionally, it is complex to determine the different overlay networks.





This diagram shows the general issue.

Each of these services can use the same IP ranges.

This is obviously because the guys who run these services want to make it simple for them.

The underlay network infrastructure handles the separation of these different services.

This is done by tunnels. Today this is typically VXLAN; in the old days it was MPLS or VLAN.

The difference is today it is dynamic.

Visibility Approach on Overlay & Underlay Networks

In principle we have several visibility options

- Visibility of the underlay network.
- Visibility of a specific overlay network.
- Visibility of all overlay networks.
- Visibility of the underlay and overlay at the same time, a “full end to end view”

Network Visibility vs Endpoint Visibility

Network Monitoring and Endpoint Monitoring are often mixed but there is a huge difference!

Network Visibility

- Shows metrics based on network data
- Mostly passive solution
- Agnostic to devices and software
- Low operating cost
- End to end view including the transport path
- Limited application related metrics
- More complex approach in the installation phase (due to being HW)
- Good for troubleshooting

This is Cubro's playground !

Endpoint Visibility

- Show metrics based on logs or active clients
- Typically not passive
- Not agnostic, adoption for each device is usually needed.
- Shows end-to-end performance but not the network path or network parameters. The network parameters are an indirect derivation from the end-to-end parameters.
- Good application-centric metrics
- High and unpredictable operational cost
- Easy to install in the beginning
- Not efficient for troubleshooting

Network Visibility vs Endpoint Visibility

What is more important?

Hard to say, it depends on the customers needs... “Customers business cases”

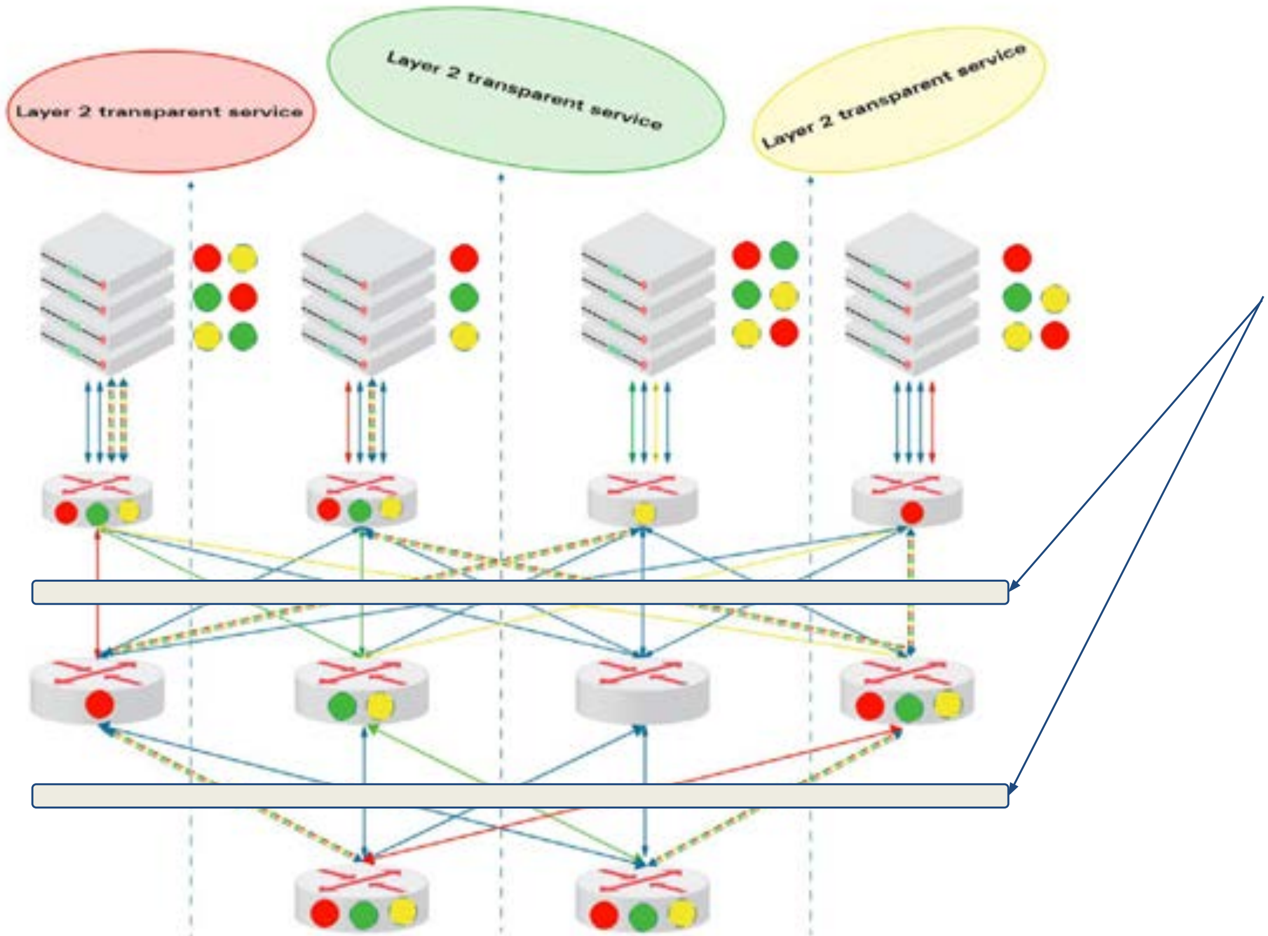
- **Service Providers = 80 % Network 20 % Endpoint**
- **Datacenter Providers = 90 % Network 10 % Endpoint**
- **Large Enterprise = 40 % Network 60 % Endpoint**
- **Enterprises with their own cloud infrastructure = 50 % Network 50 % Endpoint**
- **Enterprises with public cloud infrastructure = 10 % Network 90 % Endpoint**

Bottom line: both solutions can work together to provide total visibility!

Check what you need and decide which solution works best for you!

https://www.egscyber.com/blog/what-would-be-better-out-of-network-monitoring-and-endpoint-monitoring/

Underlay Network Design & L2 Overlay and Monitoring



The issue is seen clearly in this picture if you tap and monitor at these points:

We see two issues:

The same traffic can be seen several times.

All overlay networks are seen at the same time.

“ The L2 networks can run the same IP range and therefore, it is very complex for the classical monitoring to separate the streams because typical monitoring solution works with IP addresses to determine the different paths in the network “

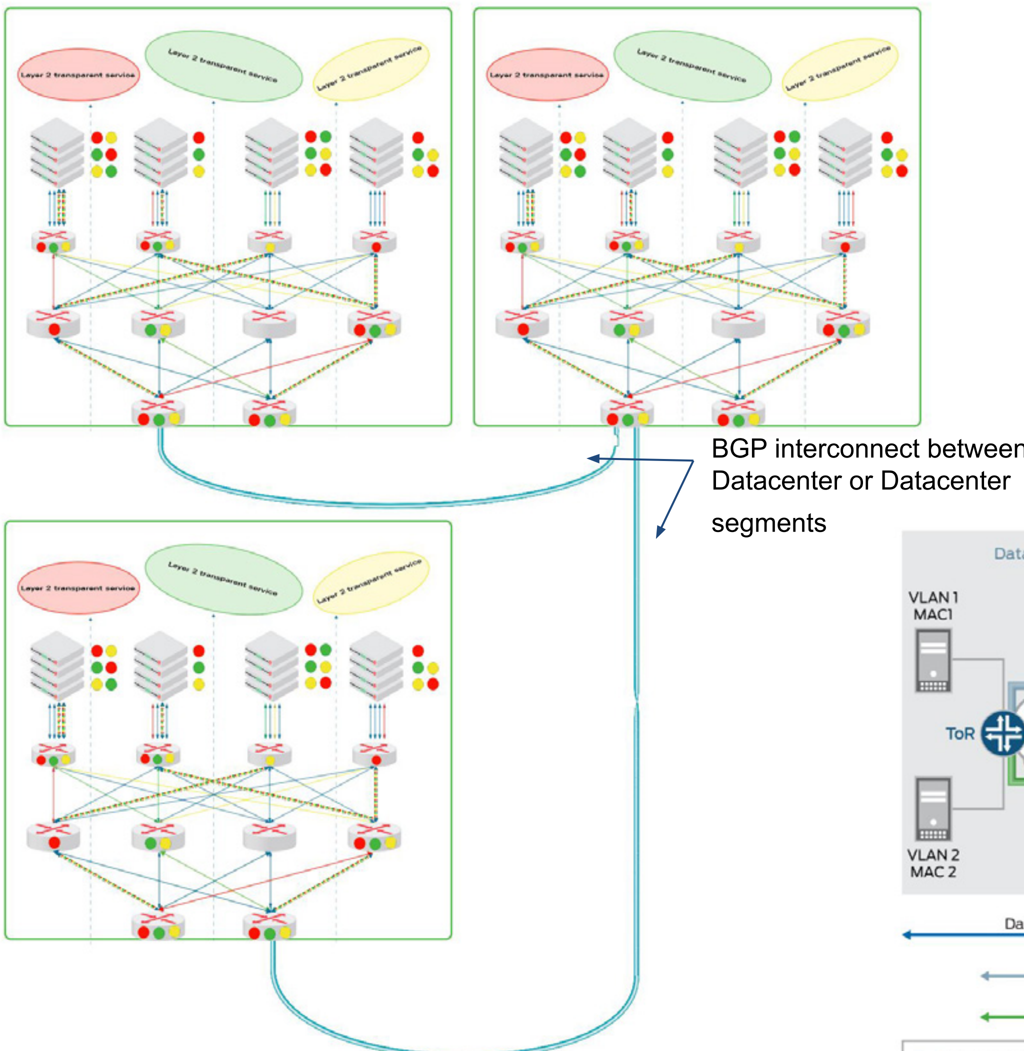
Typical monitoring tools cannot handle tunneled traffic.

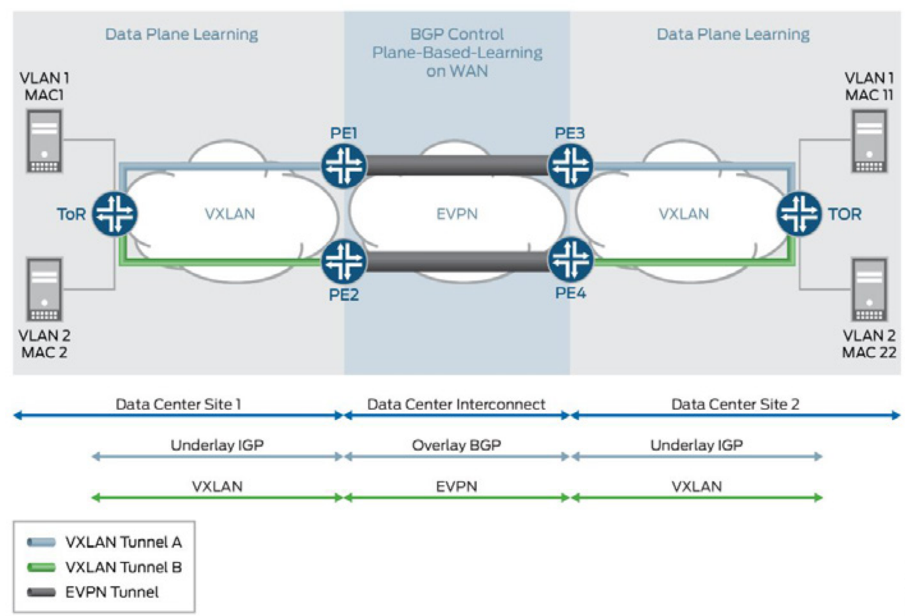
Nearly all monitoring tools are designed to handle traffic only on one port, or on one logical network layer. This tool cannot usually correlate traffic.

Issue 2

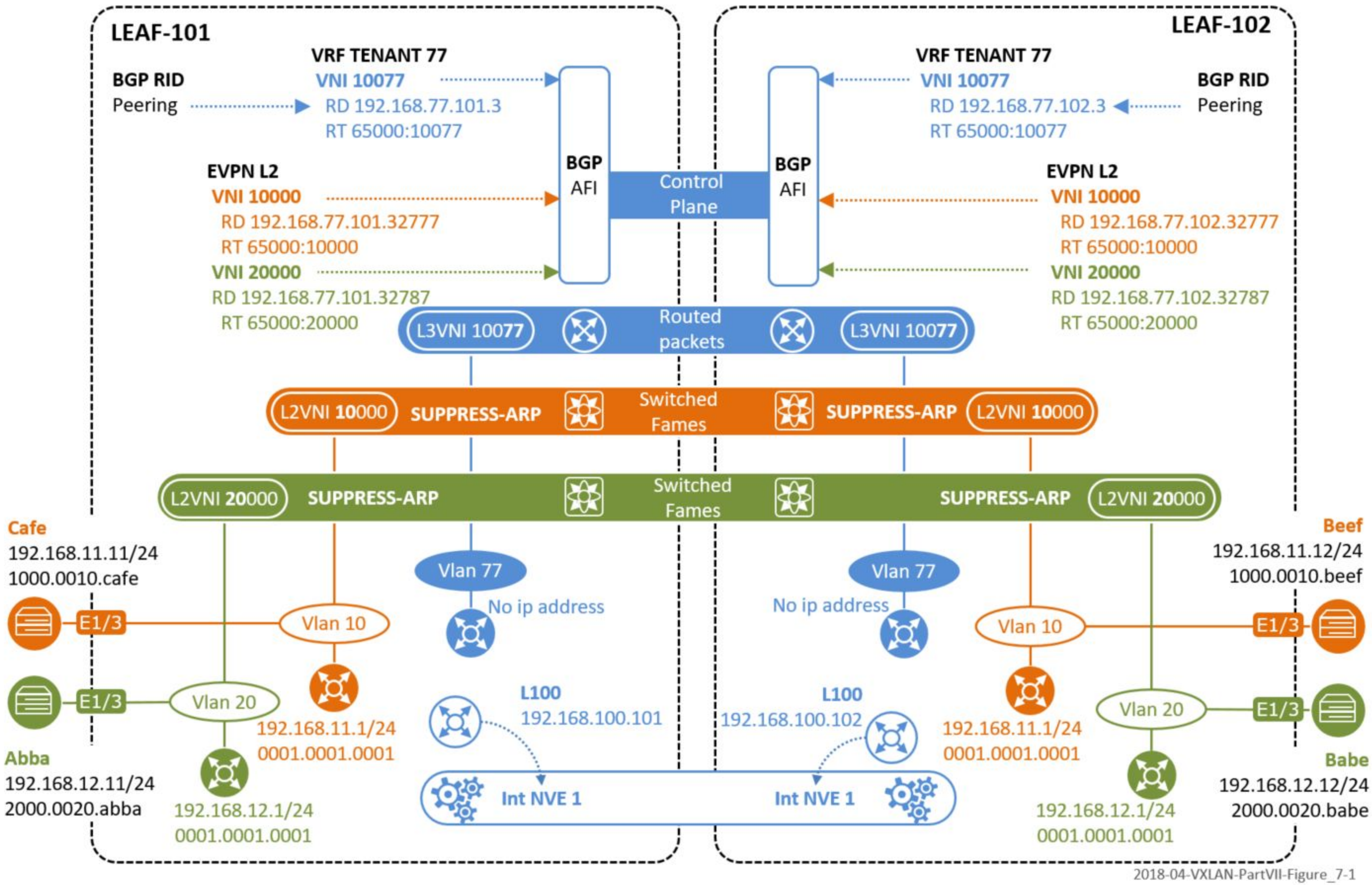
This issue is even more complex.
The overlay network can be distributed over different DCs. These different DCs are typically connected over BGP links.

In this case a BGP correlation is needed to produce useful results.





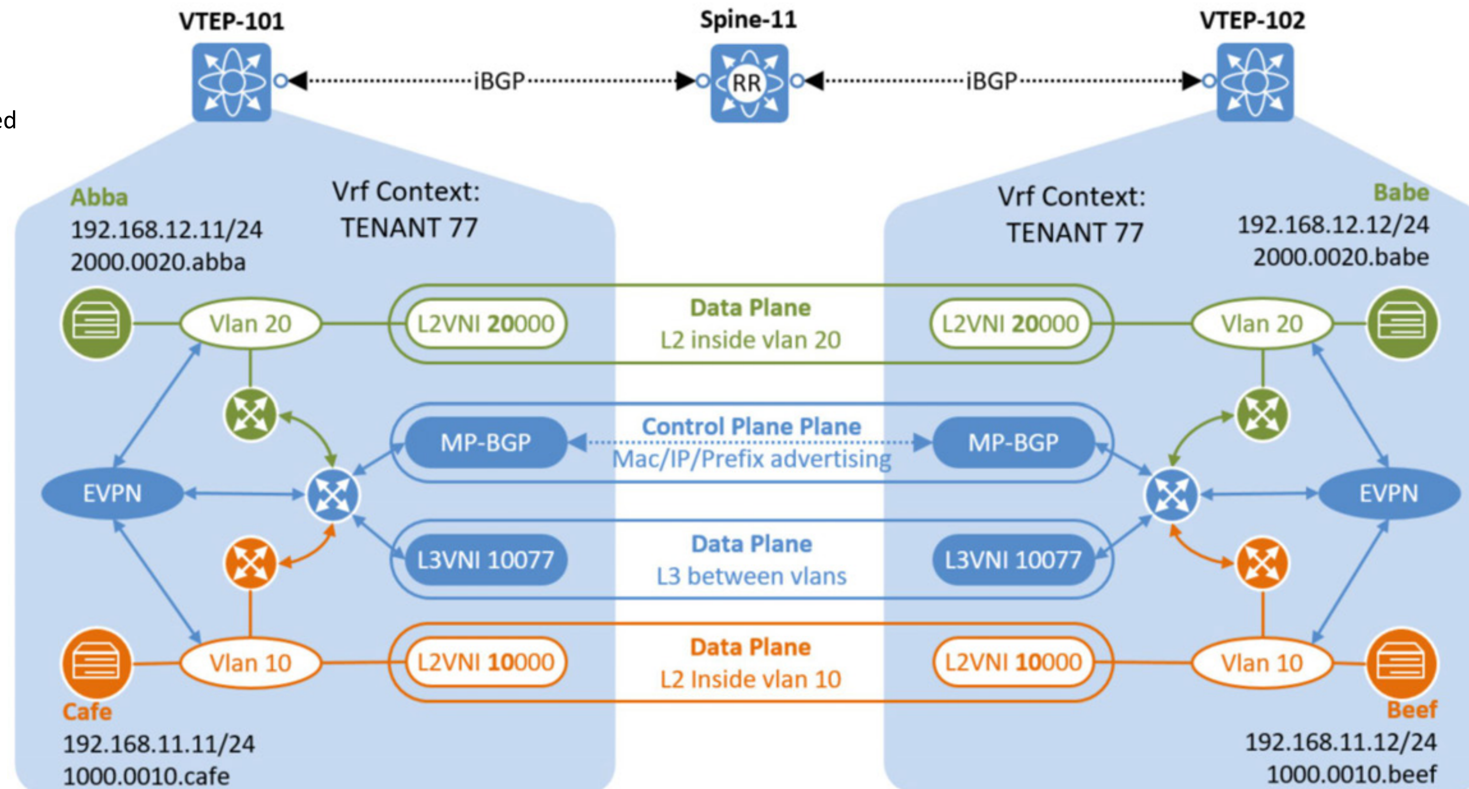
VXLAN Concept



VXLAN Concept

green and orange is L2 switched
blue is L3 routed

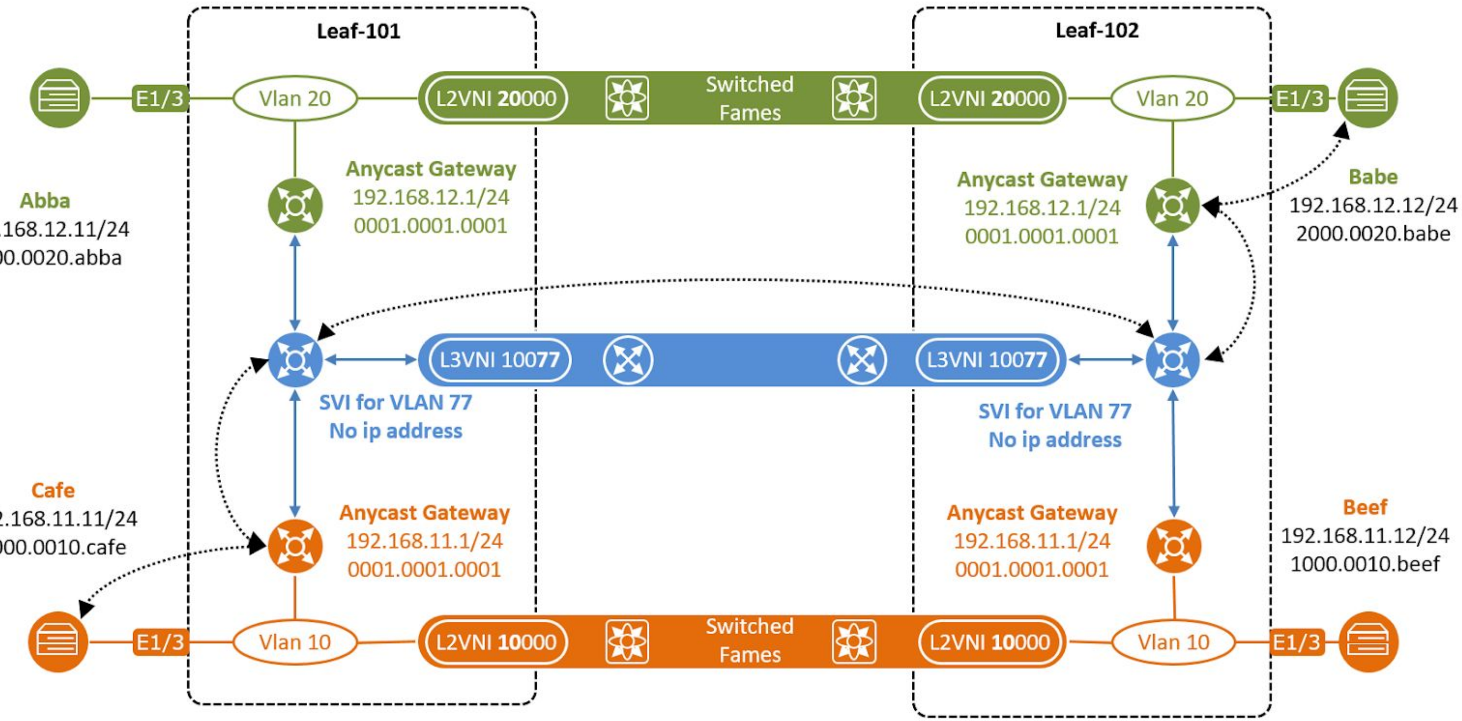
see previous slide



If you not want to monitor all traffic from Cafe, filtering on only one VXLAN is not enough, in this example you see L2 switched traffic is in VNI 10000 but routed traffic is in VNI 10077.

The challenge now is to know which VXLAN belongs together because when you have multiples routing endpoints you would have multiple VXLAN ID’s. The connection is BGP !

VXLAN Concept



2018-04-VXLAN-PartVII-Figure_7-11

only on VXLAN ID 10077 you would see the routed packet with

VLAN 77

on ID 10000 or ID 20000 you would see only the packet to the GW

with VLAN 10 and VLAN 20

If you now remove the VXLAN you would see this packet
3 times with different header MAC and VLAN

such traffic cannot removed by a deduplication function
because only the content is the same but the headers are different *

```
> Frame 368: 164 bytes on wire (1312 bits), 164 bytes captured (1312 bits)
> Ethernet II, Src: 5e:00:00:00:00:07 (5e:00:00:00:00:07), Dst: 5e:00:00:02:00:07 (5e:00:00:02:00:07)
> Internet Protocol Version 4, Src: 192.168.100.101, Dst: 192.168.100.102
> User Datagram Protocol, Src Port: 60963, Dst Port: 4789
v Virtual eXtensible Local Area Network
  > Flags: 0x0800, VXLAN Network ID (VNI)
    Group Policy ID: 0
    VXLAN Network Identifier (VNI): 10077
    Reserved: 0
> Ethernet II, Src: 5e:00:00:00:00:07 (5e:00:00:00:00:07), Dst: 5e:00:00:01:00:07 (5e:00:00:01:00:07)
> Internet Protocol Version 4, Src: 192.168.11.11, Dst: 192.168.12.12
> Internet Control Message Protocol
```

VXLAN Concept timing

Basic connectivity test

We are going to test basic connectivity between the hosts with ping.

Ping from Café to Beef (L2VNI service over VXLAN fabric)

Figure 7: ping Café to Beef

```
Cafe#ping 192.168.11.11
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 192.168.11.11, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/1/2 ms
```

Ping from Café to Abba (Local routing)

Figure 8: ping Café to Abba

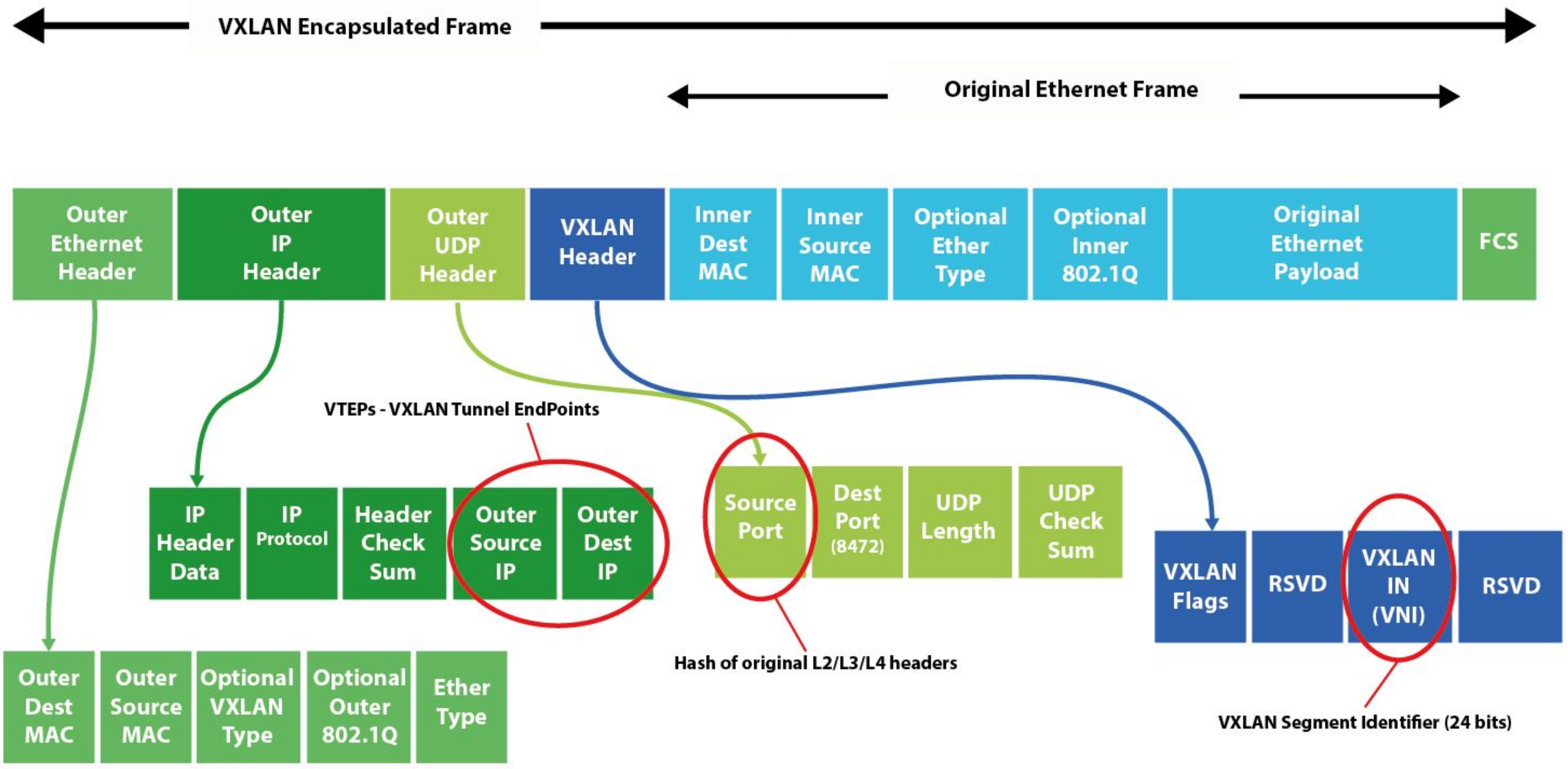
```
Cafe#ping 192.168.12.11
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 192.168.12.11, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 2/8/13 ms
```

Ping from Café to Babe (L3VNI service over VXLAN fabric)

Figure 9: ping Café to Babe

```
Cafe#ping 192.168.12.12
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 192.168.12.12, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 20/23/29 ms
```

Frame Structure in the Underlay Network



Issue 1

Outer Ethernet Header	Outer IP Header	Outer UDP Header	VXLAN Header 10020	Inner Dest MAC	Inner Source MAC	Inner IP Dest 10.10.10.10	Inner IP Source 10.10.10.1	Original Ethernet Payload	FCS

Outer Ethernet Header	Outer IP Header	Outer UDP Header	VXLAN Header 10010	Inner Dest MAC	Inner Source MAC	Inner IP Dest 10.10.10.10	Inner IP Source 10.10.10.1	Original Ethernet Payload	FCS

This is the common issue; same IP range but different overlay.

Normally the standard monitoring devices do not see the outer header.

Therefore, the result of the inner IP measurement is often wrong!

The overlay information is usually lost.

You get a result but it is incorrect!

VXLAN Monitoring scenarios with spanports

1:) Inter leaf traffic with same subnet and EVPN termination on hypervisor

2:) Inter leaf traffic with same subnet and EVPN termination switch port

3:) Inter leaf traffic with same subnet and EVPN termination one side switch port other side hypervisor

4:) Intra subnet traffic over differents leafs and EVPN termination on hypervisor (inter L2)

5:) Intra subnet traffic over differents leafs and EVPN termination switch port (inter L2)

6:) Intra subnet traffic over differents leafs and EVPN termination one side switch port other side hypervisor (inter L2)

7:) Inter leaf traffic with different subnet and EVPN termination on hypervisor (inter L3)

8:) Inter leaf traffic with different subnet and EVPN termination switch port (inter L3)

9:) Inter leaf traffic with different subnet and EVPN termination one side switch port other side hypervisor (inter L3)

10:) Different subnet traffic over differents leafs and EVPN termination on hypervisor

11:) Different subnet traffic over differents leafs and EVPN termination switch port

12:) Different subnet traffic over differents leafs and EVPN termination one side switch port other side hypervisor

VXLAN Monitoring scenarios with spanports

1:) Inter leaf traffic with same subnet and EVPN termination on hypervisor

only egress monitoring with span port

no duplicates, but VXLAN and LAN tags on the traffic

VXLAN Monitoring scenarios with spanports

2:) Inter leaf traffic with same subnet and EVPN termination switch port

only egress monitoring with span port

no duplicates, LAN tags on the traffic

VXLAN Monitoring scenarios with spanports

3:) Inter leaf traffic with same subnet and EVPN termination one side switch port other side hypervisor

only egress monitoring with span port

no duplicates,

on the switch port termination no VXLAN but VLAN

on the hypervisor termination VXLAN and VLAN

In this case the request has no VXLAN and the answer has a VXLAN or vice versa

VXLAN Monitoring scenarios with spanports

4:) Intra subnet traffic over differents leafs and EVPN termination on hypervisor

only egress monitoring with span port

duplicates,

packet 1:) on the spine egress VXLAN and VLAN

packet 2:) on the leaf egress VXLAN and VLAN (both the same)

remove duplicates ??

deduplication does not support VXLAN so VXLAN must be removed first
very expensive approach could not be done via filters.

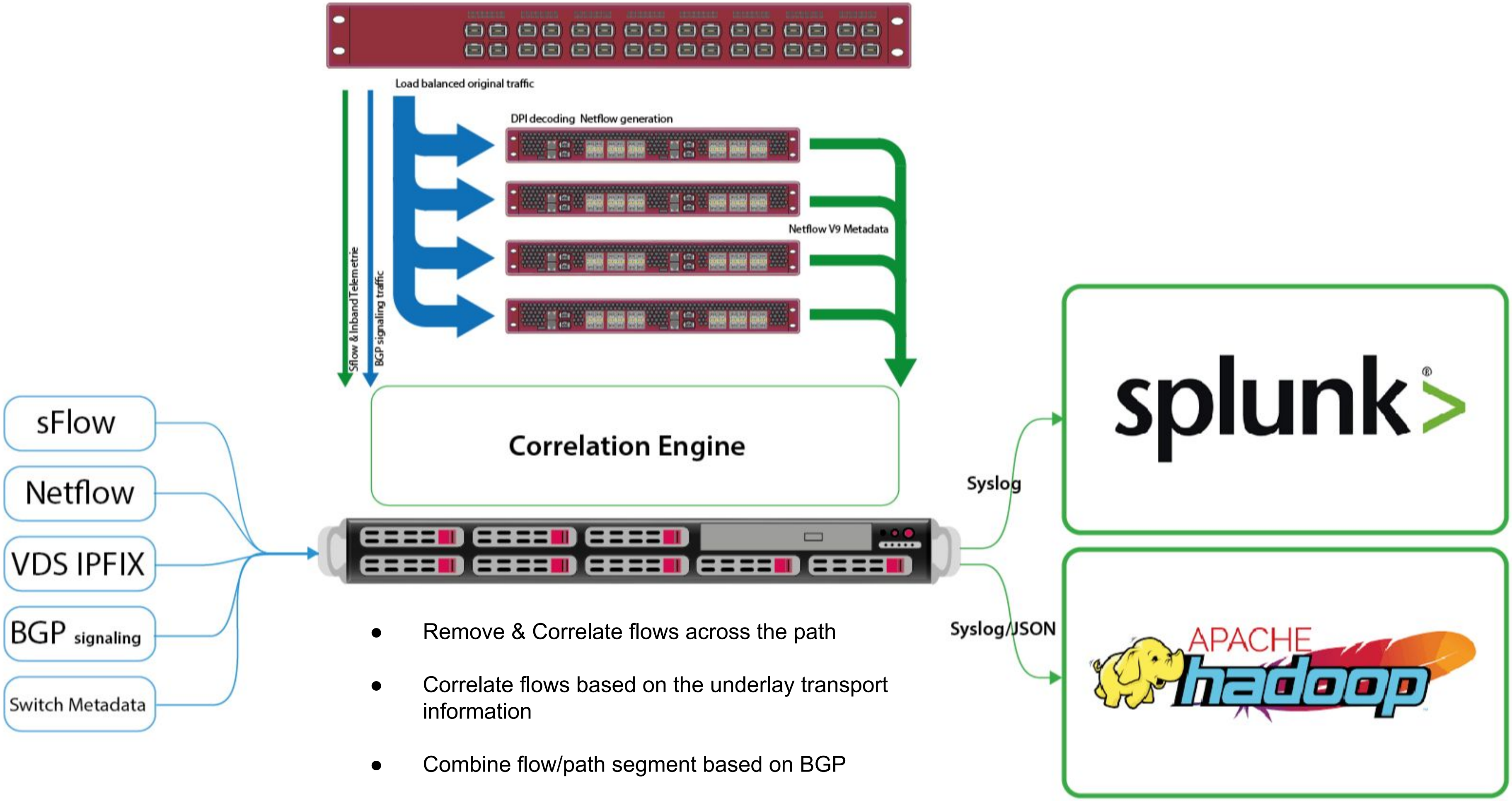
The problem is it could not be done based on links because the duplicates are on different links
so first we must remove the VXLAN than aggregate all to one big pipe, but this big pipe overloads
the deduplication CPU (100 Gbit +) so LB is need to feed different deduplication CPU's very complex !

(It is also not clear is the Network Layer untouched ? "IP ID field")

5:) Intra subnet traffic over differents leafs and EVPN termination switch port (inter L2)
Same as 4 but no VXLAN

6:) Intra subnet traffic over differents leafs and EVPN termination one side switch port other side hypervisor (inter L2)
Same as 4 but one packet with VXLAN + VLAN and one with only VLAN

Cubro Solution Design 1



- Remove & Correlate flows across the path
- Correlate flows based on the underlay transport information
- Combine flow/path segment based on BGP
- Enrich Data from Switch inband Telemetry
- Enrich Data with switch Table Information

Cubro Solution Design 1

Network Path

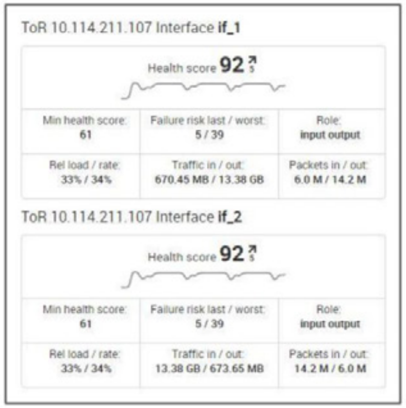
Now that applications and hosts impacted by physical network outages are identified, an SDDC administrator can select end nodes to view where VM to VM traffic is encapsulated, and can see specifically which physical network devices the traffic went through.

Now SDDC administrators can pinpoint which of the network devices in the path are a cause of application performance problems.

The following image shows network device interfaces involved in VM to VM communication.

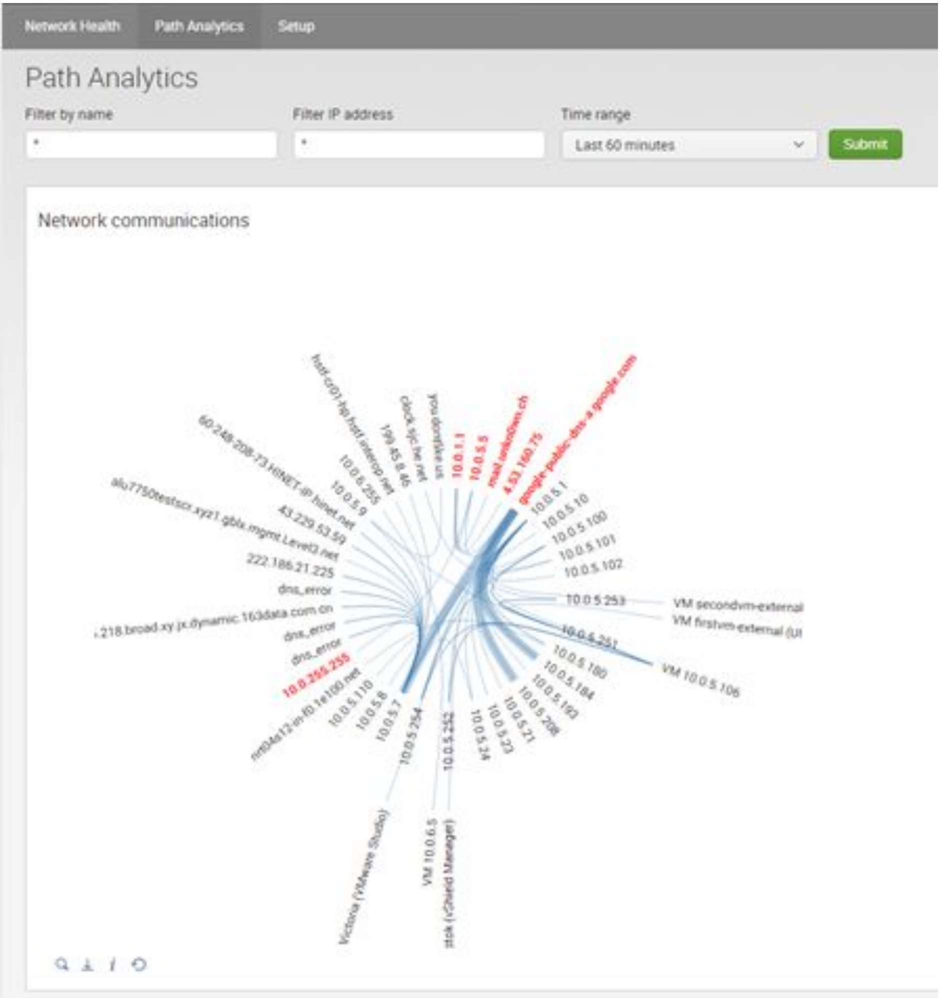
For interfaces relaying a traced communication the following information is presented:

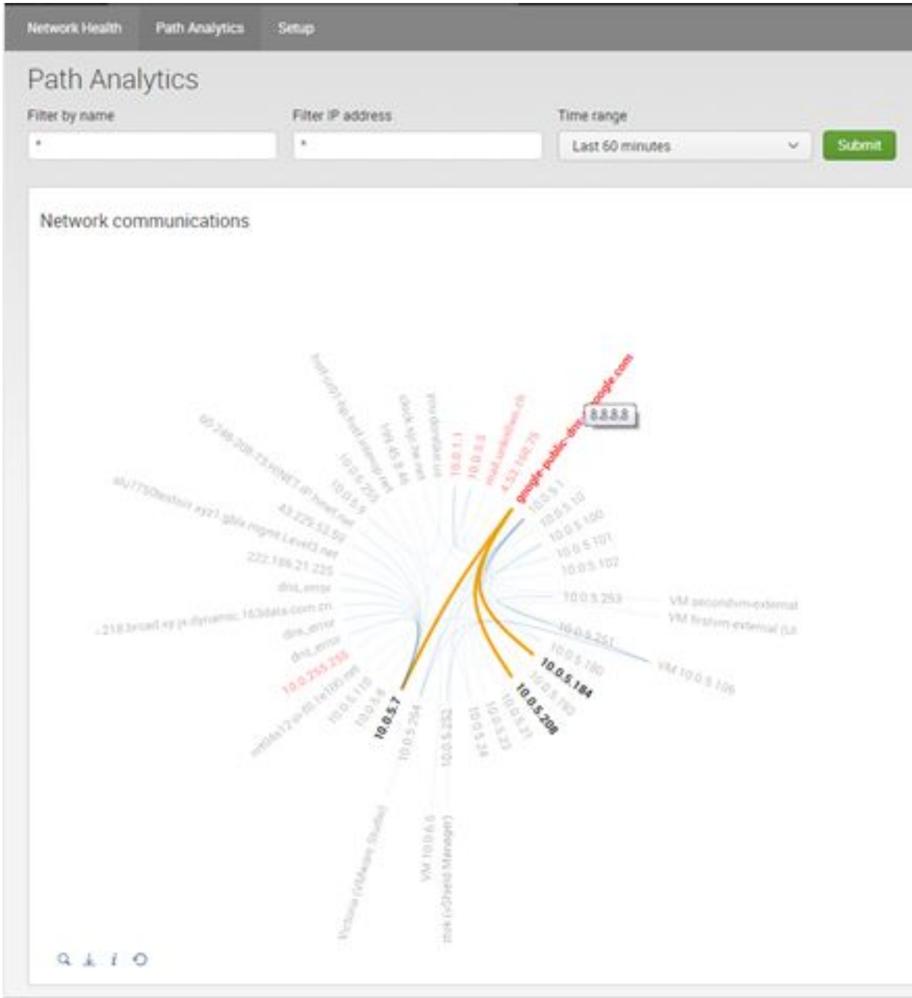
- Relative traffic load on this interface as a percent of its nominal capacity
- Relative packet rate on this interface as a percent of a maximal packet rate sustainable at a current average packet size
- A total number of bytes passed in each direction through this interface over a selected time interval
- A total number of packets passed in each direction through this interface over a selected time interval



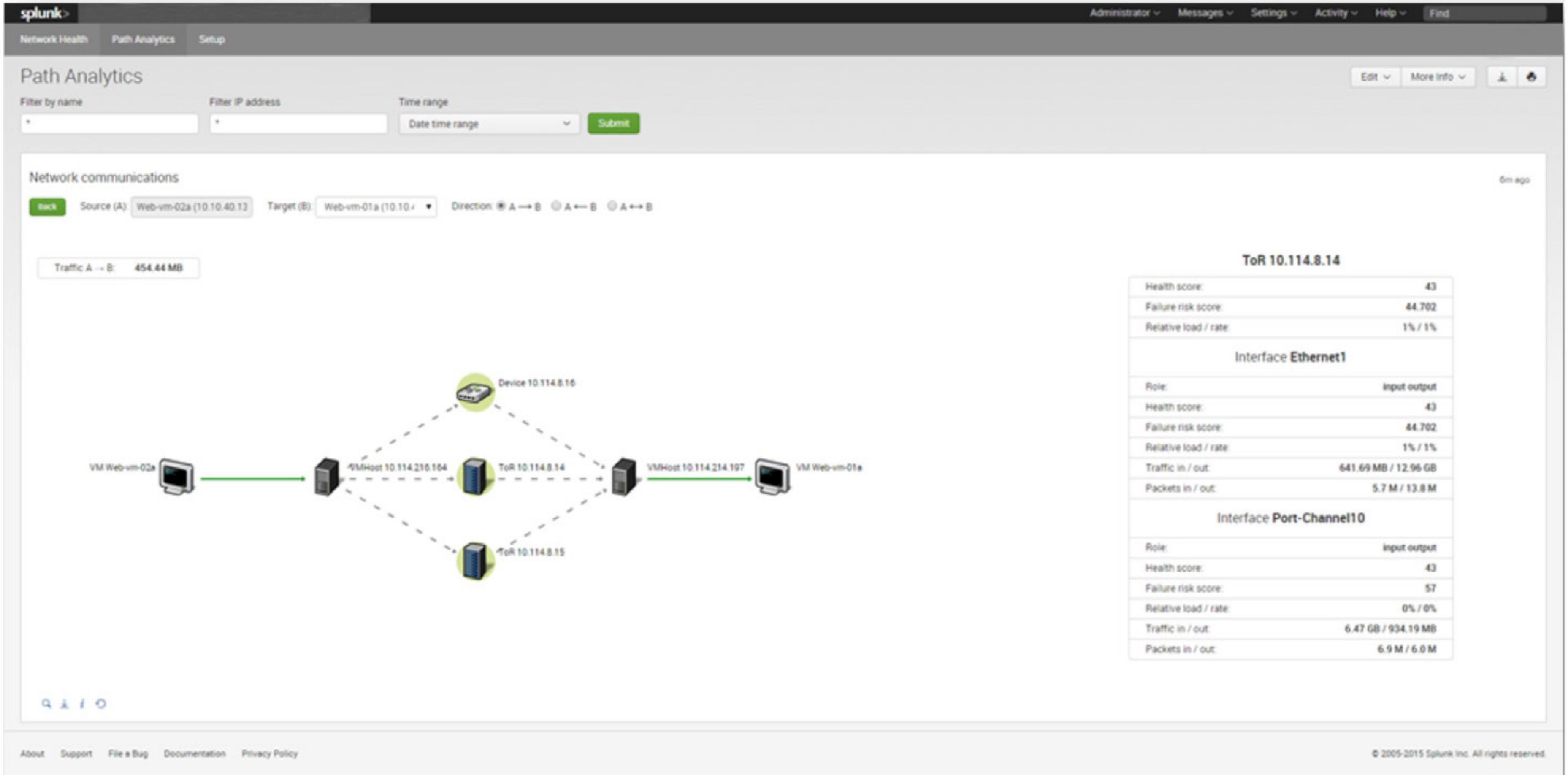
The Path information is available not only for VM to VM (East-West traffic) within the data center, but also for VM to gateways (North-South traffic). This capability is useful in identifying network congestion and abnormal activity such as data exfiltration.

Conversations in Virtual Overlay Networks

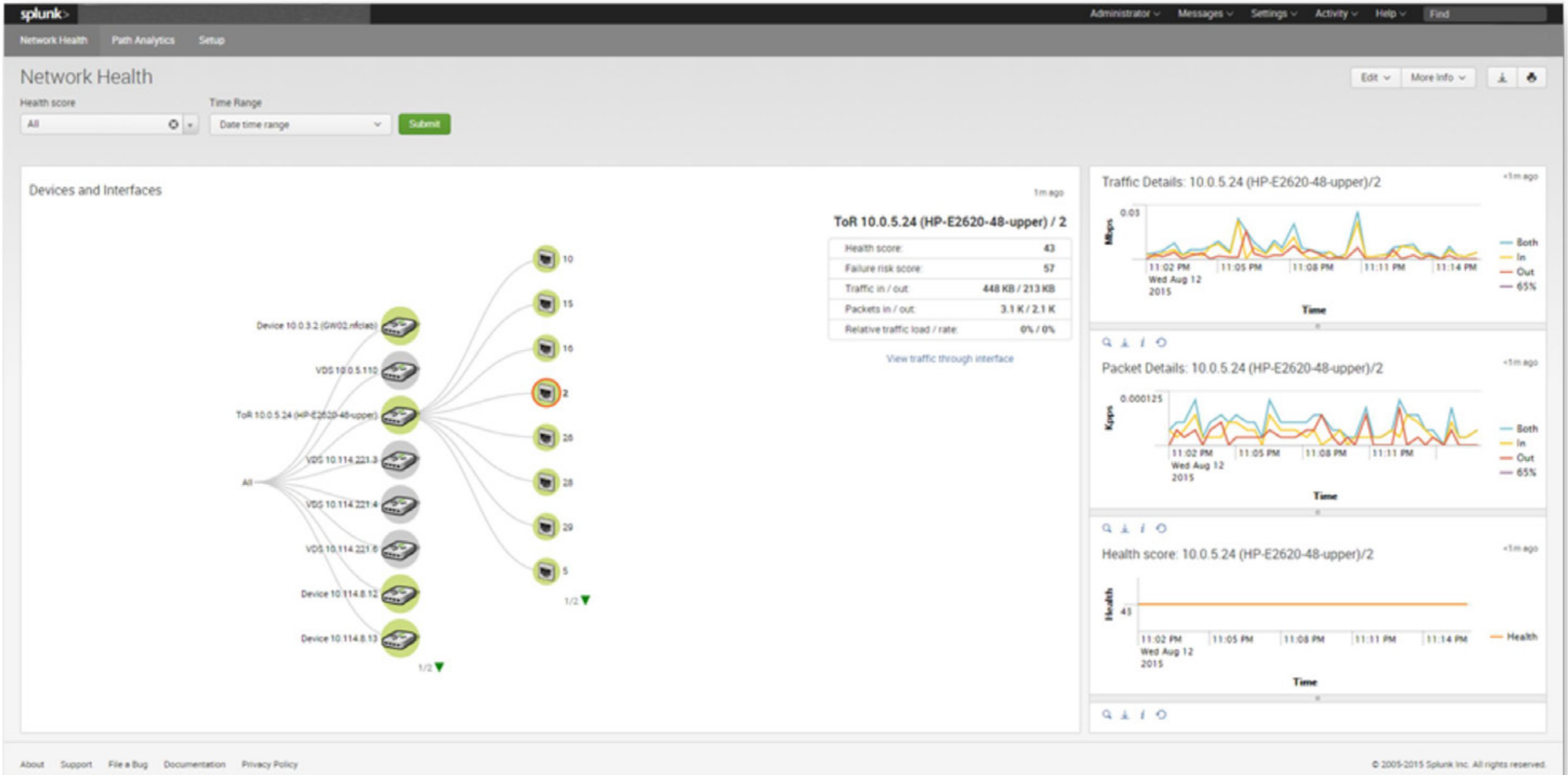




Physical Path Rendering for VM Conversations



Network Health



Other features

- Top Tunnels – show top tunnels by traffic
- Top Flows – show top flows within a tunnel
- Distributed Firewall (DFW) - coming
- Distributed Logical Router (DLR) - coming

Top Tunnels, Top Flows

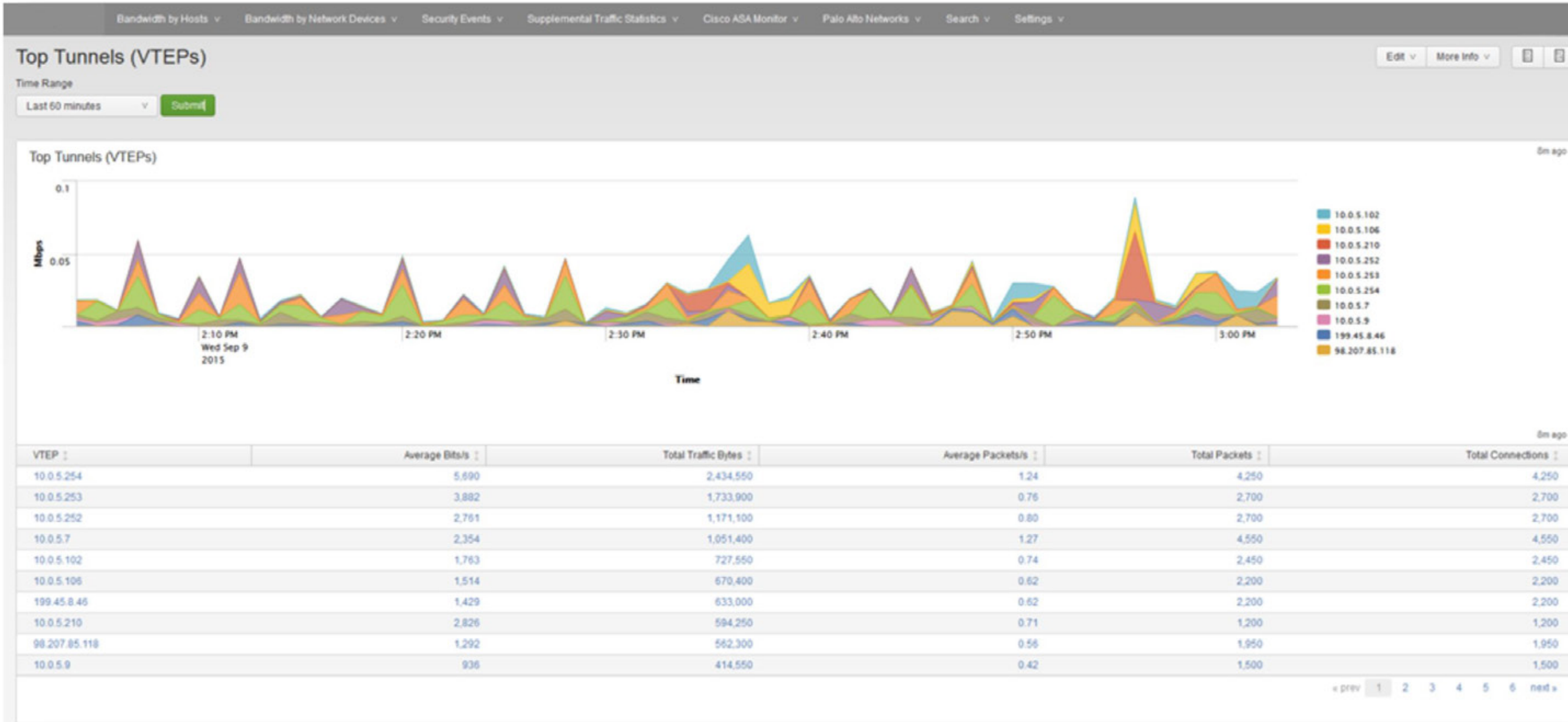
- Top Tunnels (VTEPs) by Traffic; Select Time Interval; Show:

VTEP, Average Bits/s, Total Traffic Bytes, Average Packets/s, Total Packets, Total Connections

- Drill down by selecting VTEP from the list above; Show:

VXLAN_ID, Source VM IP, Source VM Name, Source VTEP, Destination VM IP, Destination VM Name, Destination VTEP, Average Bits/s, Total Traffic Bytes, Average Packets/s, Total Packets, Total Connections

Top Tunnels (VTEPs)



VTEP	Average Bits/s	Total Traffic Bytes	Average Packets/s	Total Packets	Total Connections
10.0.5.254	5,690	2,434,550	1.24	4,250	4,250
10.0.5.253	3,882	1,733,900	0.76	2,700	2,700
10.0.5.252	2,761	1,171,100	0.80	2,700	2,700
10.0.5.7	2,354	1,051,400	1.27	4,550	4,550
10.0.5.102	1,763	727,550	0.74	2,450	2,450
10.0.5.106	1,514	670,400	0.62	2,200	2,200
199.45.8.46	1,429	633,000	0.62	2,200	2,200
10.0.5.210	2,826	594,250	0.71	1,200	1,200
98.207.85.118	1,292	562,300	0.56	1,950	1,950
10.0.5.9	936	414,550	0.42	1,500	1,500

Drill down VTEP

Top Flows for VTEP: 10.0.5.252

5m ago

5m ago

VXLAN_ID	Source VM IP	Source VM Name	Source VTEP	Destination VM IP	Destination VM Name	Destination VTEP	Average Bits/s	Total Traffic Bytes	Average Packets/s	Total Packets	Total Connections
5001	10.0.5.9	vm_10_0_5_9	10.0.5.252	10.0.5.14	vm_10_0_5_14	10.0.5.253	28,858	223,650	3.23	200	200
5001	10.0.5.9	vm_10_0_5_9	10.0.5.252	10.0.5.15	vm_10_0_5_15	10.0.5.253	1,116,800	139,600	100.00	100	100
5001	10.0.5.9	vm_10_0_5_9	10.0.5.252	10.0.5.16	vm_10_0_5_16	10.0.5.253	4,953	94,100	0.99	150	150
5002	10.0.5.10	vm_10_0_5_10	10.0.5.252	10.0.5.17	vm_10_0_5_17	10.0.5.253	9,303	72,100	1.61	100	100
5002	10.0.5.10	vm_10_0_5_10	10.0.5.252	10.0.5.18	vm_10_0_5_18	10.0.5.253	558,400	69,800	50.00	50	50
5002	10.0.5.10	vm_10_0_5_10	10.0.5.252	10.0.5.19	vm_10_0_5_19	10.0.5.253	138	21,200	0.16	200	200
5002	10.0.5.10	vm_10_0_5_10	10.0.5.252	10.0.5.20	vm_10_0_5_20	10.0.5.253	156,400	19,550	150.00	150	150
5002	10.0.5.10	vm_10_0_5_10	10.0.5.252	10.0.5.21	vm_10_0_5_21	10.0.5.253	118,800	14,850	50.00	50	50
5002	10.0.5.10	vm_10_0_5_10	10.0.5.252	10.0.5.22	vm_10_0_5_22	10.0.5.253	83,600	10,450	50.00	50	50
5002	10.0.5.10	vm_10_0_5_10	10.0.5.252	10.0.5.23	vm_10_0_5_23	10.0.5.253	43,600	5,450	50.00	50	50

« prev 1 2 3 4 next »

Top VMs

- Top VMs by Traffic; Select Time Interval; Show:

VM_IP, VM_Name, Bytes_IN, Bytes_OUT, Packets_IN, Packets_OUT, Total Connections

- Drill down by selecting **Bytes_IN** from the list above; Show (selected VM is dest_VM):

VXLAN_ID, src_vm, src_VTEP, dest_vm, dest_VTEP, Avg Bits/s, Bytes, Avg Packets/s, Packets, Connections

- Drill down by selecting **Bytes_OUT** from the list above; Show (selected VM is src_VM):

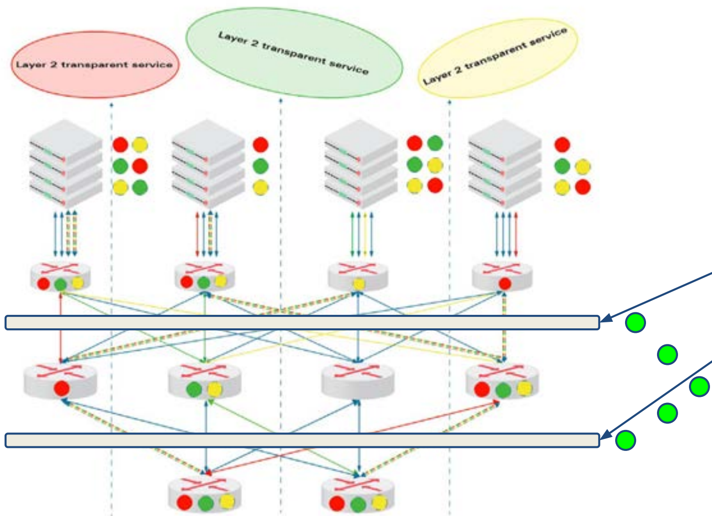
VXLAN_ID, src_vm, src_VTEP, dest_vm, dest_VTEP, Avg Bits/s, Bytes, Avg Packets/s, Packets, Connections

Cubro Solution Design 2

The other possible option is dynamic VXLAN filtering. This solution is needed for packet based solutions, like Wireshark and for instance mobil monitoring systems

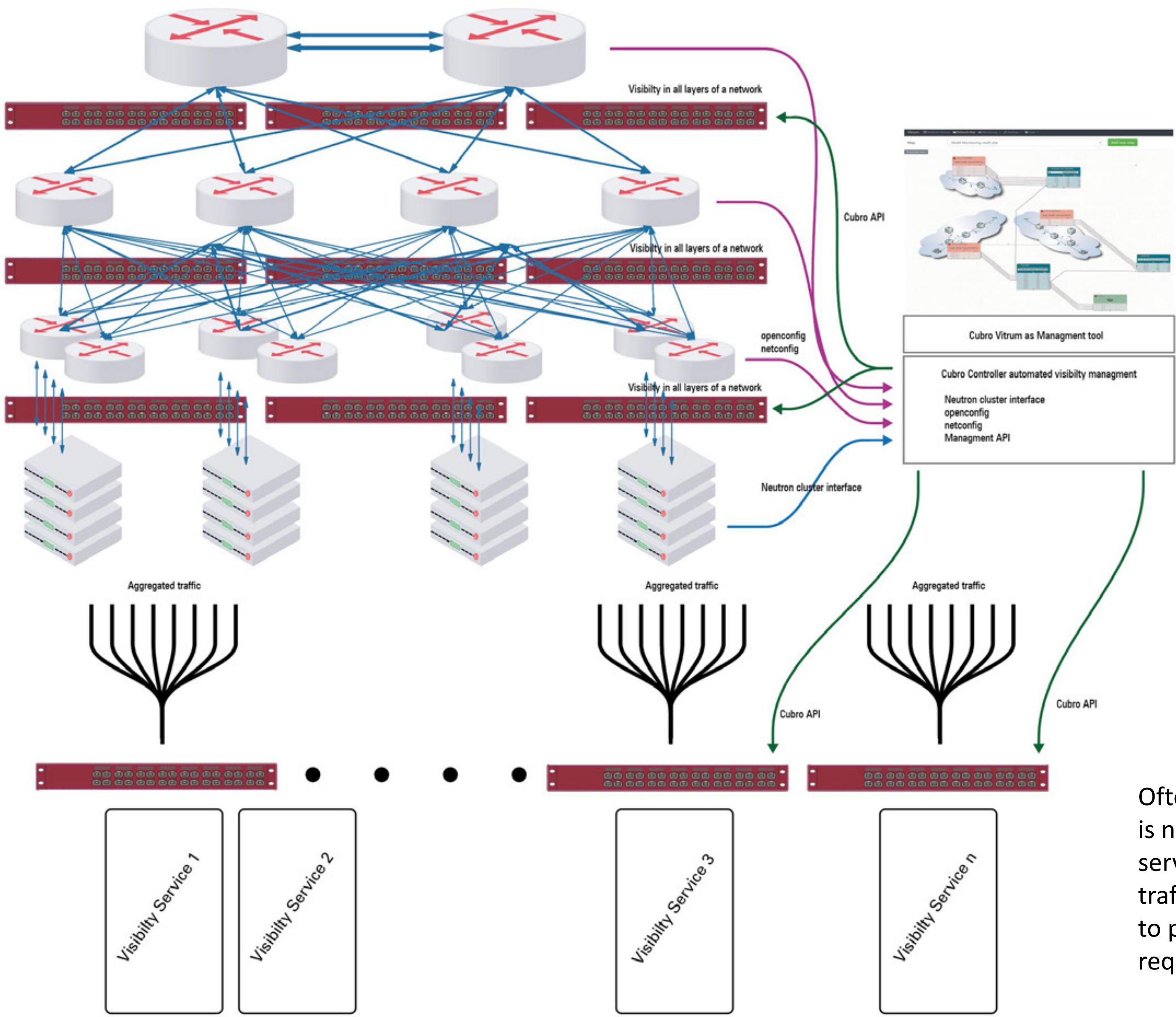
Old equipment can be repurposed because dynamic VXLAN filtering would assure that only the traffic from the relevant overlay is filtered out and sent to legacy monitoring tools.

The challenge is that only a few NPBs are capable of VXLAN filtering. The second issue is this must be done dynamically. For that reason some signaling protocols must be decoded by the packet broker or a external appliance. - > Cubro Cloud Switch!



Dynamic VXLAN filtering see next slide

Cubro Solution Design 2 the multi service approach



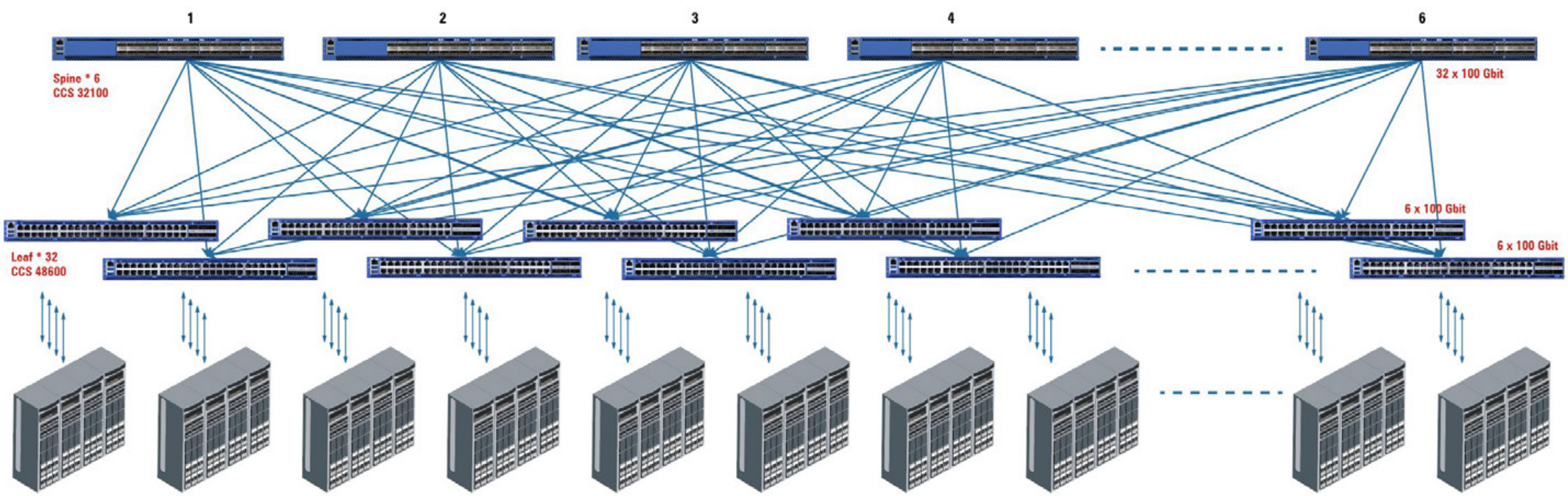
The center of this solution is the Cubro Controller which interacts with openstack and the other network elements, and controls the cubro visibility nodes. The cubro visibility nodes are not packet pushers they have an advanced API to
cope with the visibility needs for today's networks!

Often the same or overlapped traffic is needed for different visibility services. In this case “aggregated” traffic is send to several CSV units to provide the final preparation to the requested service.

Cubro Solution Design 3

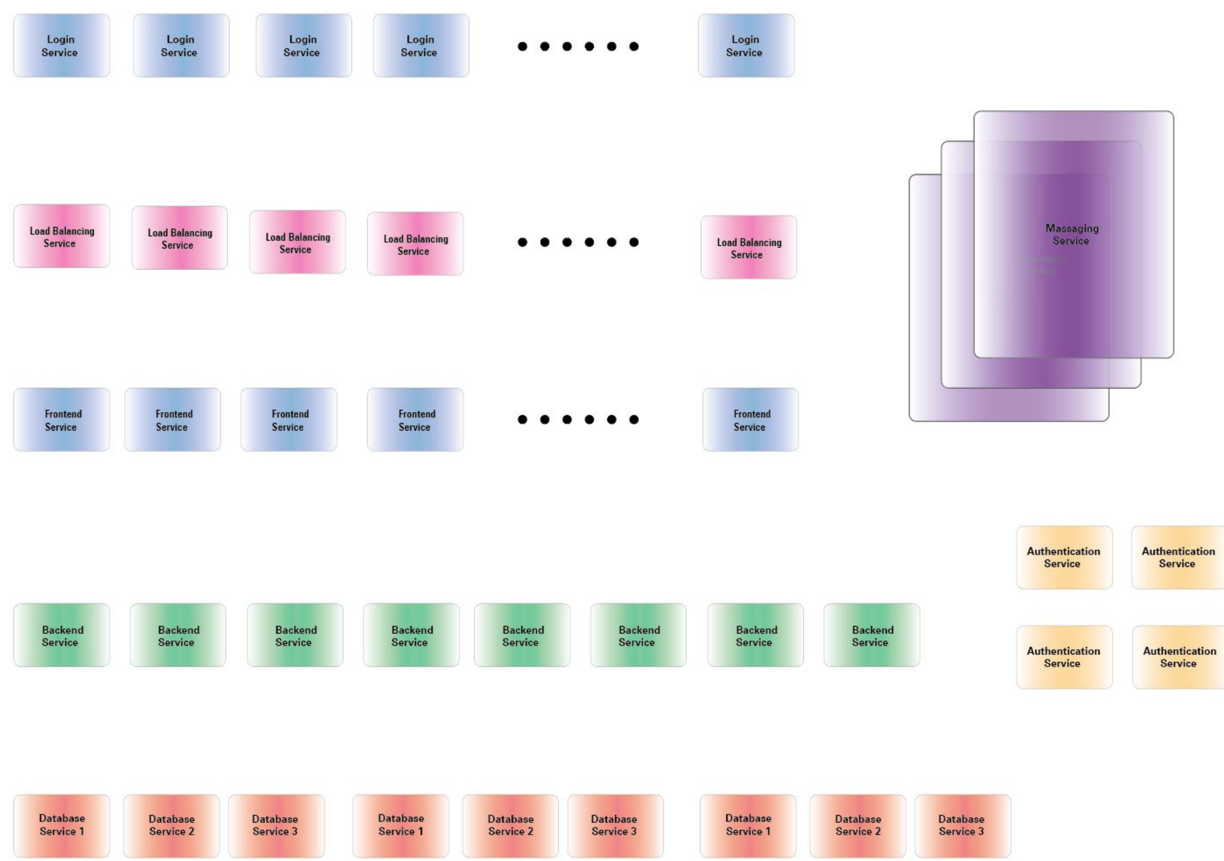
The most advanced solution would be to use the Cubro Cloud Switch because the CCS combines an advanced switching fabric with a visibility fabric.

In the cloud the visibility must be a part of the network infrastructure !



The Cloud is Breathing

Application Breathing



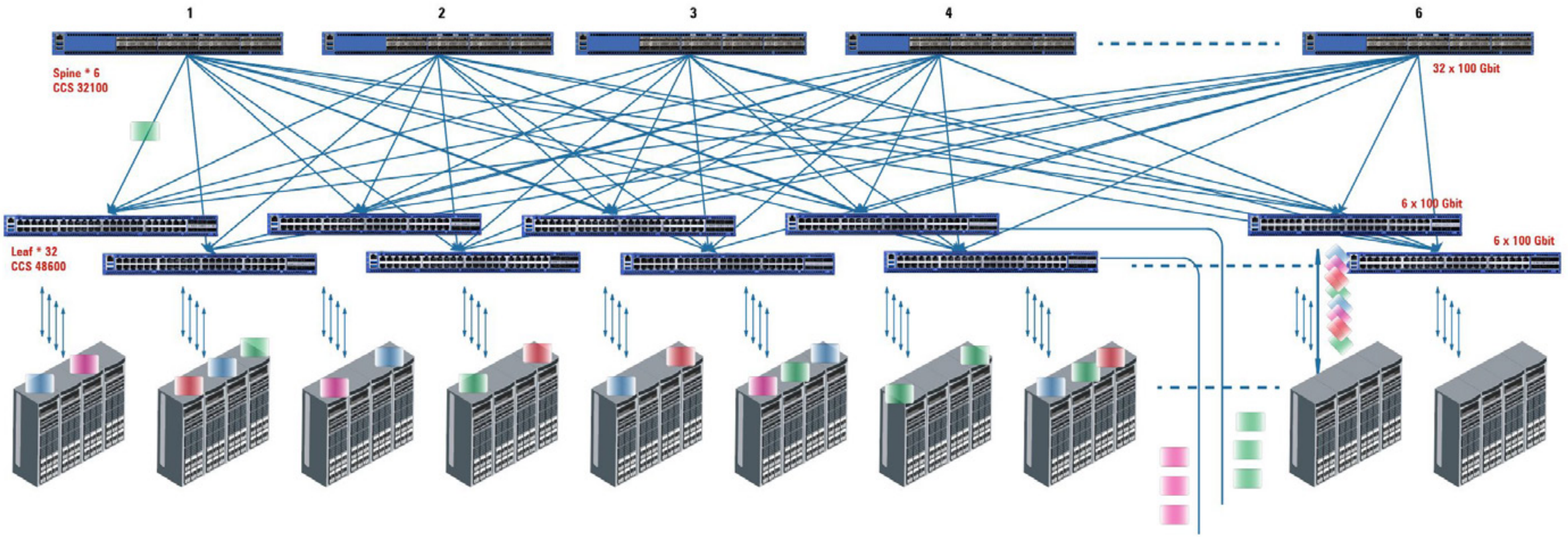
There is one big advantage that the application can dynamically grow when it needs more resources.

But the same “application” can also run in different data centers, spread out over the world.

An “application” can easily breathe under load to several 1000 micro services.

Cubro Solution Design 3

The visibility solution had to follow the Cloud breathing in realtime and this is only possible, if the visibility solution is a part of the network infrastructure.



EX - EXA - CVN - CCS

EX = classical hi end NPB with L4 functions

- aggregation
- filtering
- load balancing
- .....

EXA = classical hi end NPB with L7 functions

- aggregation
- filtering up to L7 “keyword search”
- time stamping
- GTP load balancing
- VXLAN metadata filtering
- ....

static approach / manual configuration

CVN = Cubro Visibility Node “self organizing”

This is not a NPB any more, because it interacts with the Cubro Visibility Controller, and supports dynamic packet handling approaches for modern overlay networks.

- dynamic visibility service steering
- dynamic load balancing
- dynamic packet modification
- ......

CCS = Cubro Cloud Switch “part of the network”

The CCS is currently the end in this evolution from L4 NPB to a active network device with visibility functions included.

- inband dynamic visibility service steering
- cloud centric visibility
- application breathing support
- ....

dynamic approach / self organizing visibility

CUBRO NETWORK VISIBILITY
SIMPLE | SCALABLE | SUSTAINABLE

EX - EXA - CVN - CCS

EX

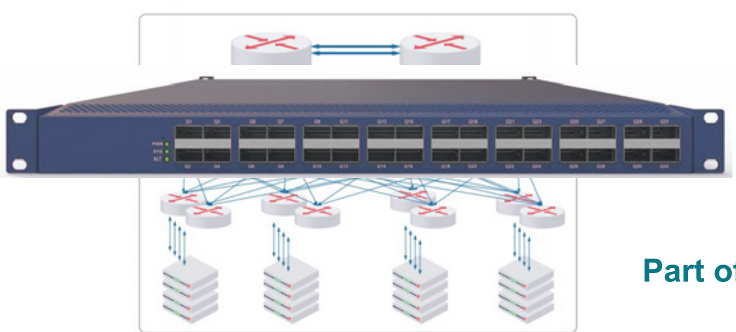
EXA

CVN

+

Cubro Controller

CCS



Part of the network infrastructure

E
V
O
L
U
T
I
O
N

CUBRO NETWORK VISIBILITY

SIMPLE | SCALABLE | SUSTAINABLE

EMEA

Cubro Network Visibility
Ghegastraße 1030 Vienna,
Austria

Tel.: +43 1 29826660
Fax: +43 1 2982666399

Email: support@cubro.com

APAC

Cubro Asia Pacific

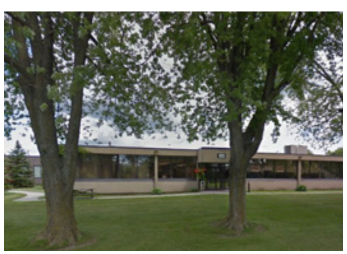
8, Ubi Road 2 #04-12 Zervex
Singapore 408538

Tel.: +65-97255386

Email: jl@cubro.com

Thank you

US & North America



Cubro US & North America
225 Peachtree Street NE
Suite 1100, Atlanta, GA,
30303, USA

Email: americas@cubro.com

Japan

Cubro Japan

8-11-10-3F, Nishi-Shinjuku,
Shinjuku,
Tokyo, 160-0023 Japan

Email: japan@cubro.com