Принципы построения сетевой инфраструктуры современных ЦОД

Dmitry Zavelskiy
dzavelskiy@juniper.net

Алматы, Казахстан

Апрель 2023

Juniper Networks | Driven by Experience

Disclaimer

This statement of product direction sets forth Juniper Networks' current intention and is subject to change at any time without notice. No purchases are contingent upon Juniper Networks delivering any feature or functionality depicted on this statement.

Agenda

- Modern DC Fabric
- DCI
- Hardware evolution
- Host Routing Bridging
- Automation
- EVPN/VXLAN innovations
- Connected Security

Foundational components of a Cloud-ready DC

Operational economics with service agility

Connected Security 	• Consistent security across clouds • Tight network integration, Micro-segmentation
SDN Overlay 	• Cloud-native SDN for hybrid cloud environments • Multi-cluster management and observability
Automated Datacenter Operations 	• Complete Datacenter lifecycle management (Day 0, 1 & 2) • Consistent, Assured & Reliable
Scalable IP Fabric 	• High-performance 100/200/400G capable underlay • Flexible, standards-compliant inter-connect options

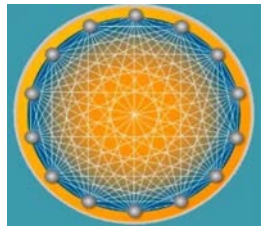
MODERN DC FABRIC

JUNIPER NETWORKS

Data center architecture evolution

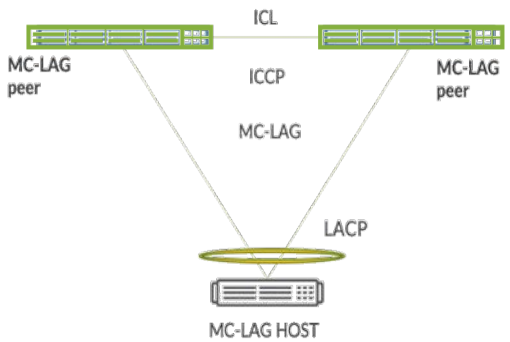
GOAL: APSTRA + EVPN - Solution that enables the experience of a public cloud

End of Life



QFabric

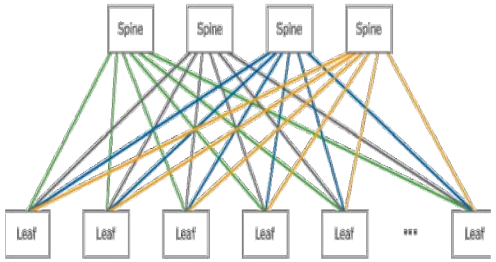
- Proprietary
- Single Point of Failure – High Blast Radius
- Limited scale
- Lower convergence time compared to ESI-LAG
- Limited to dual homing



MC-LAG

Why: Node redundancy

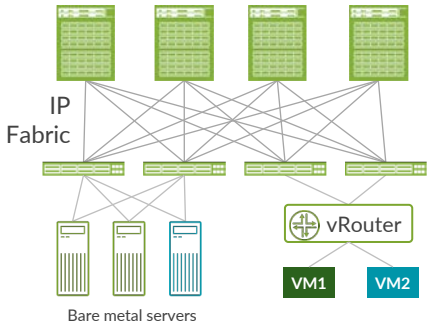
- Proprietary
- Single Point of Failure – High Blast Radius
- Limited scale
- Upgrade/ maintenance - complex



VCF

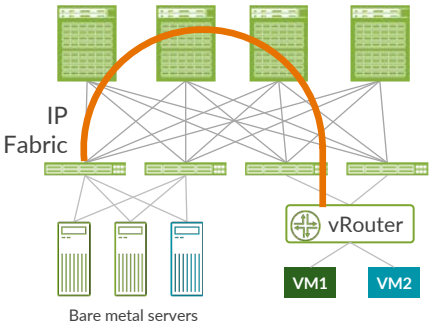
Why: ease of operation

- Open Standards
- No Single Point of Failure
- Internet Grade Scale – scale out architecture
- Disaggregated Architecture
- Fabric always on



IP Fabric

Why: Resilience



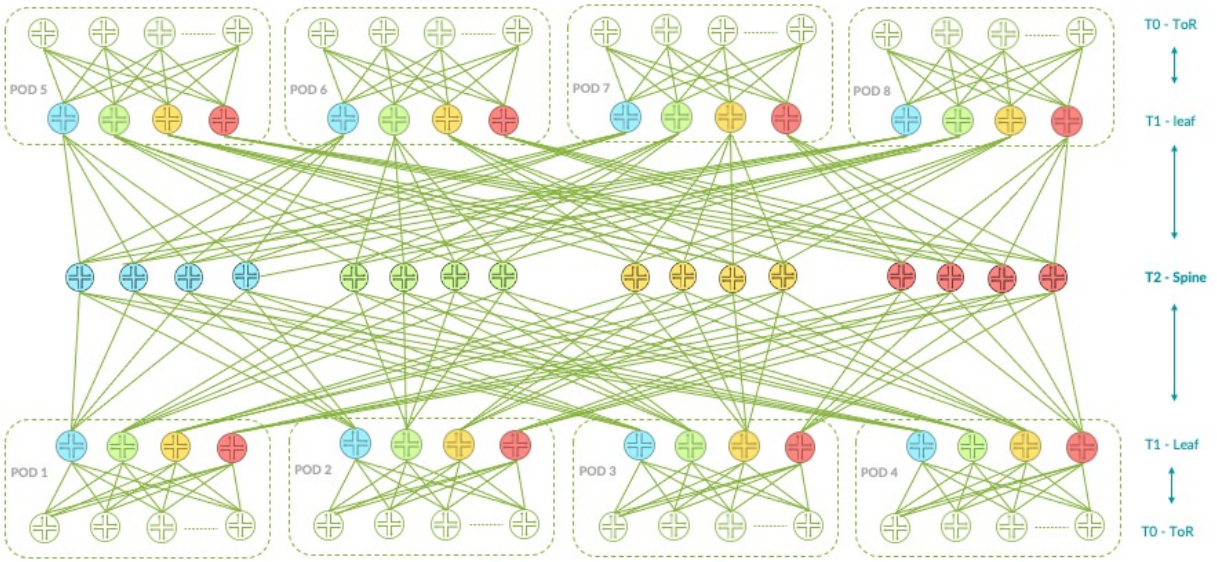
EVPN/VXLAN

Why: Connect L2 domains over L3

Design flexibility of the modern DC fabrics

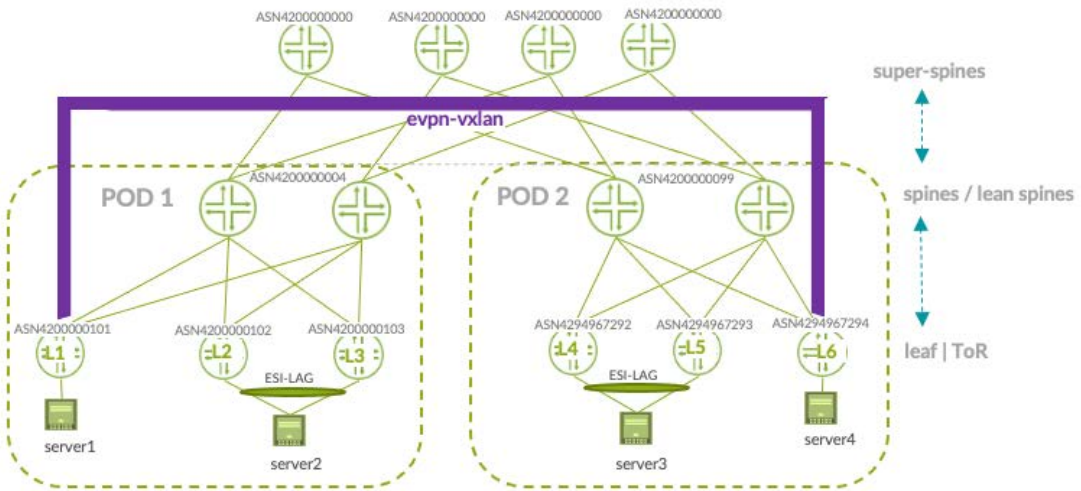
ERB – Edge Routed Bridging
BO – Bridged Overlay
CRB – Centrally Routed Bridging

Native IP Fabric



- Low latency
- Reduced L2 reach
- Network virtualization at the server level
- Server L2 multihoming using MC-LAG (optional) or HBR

EVPN-VXLAN Fabric



- Additional latency to consider
- Flexible L2 reach between servers
- ToR/leaf level network virtualization and services
- Server L2 multihoming using ESI-LAG

DC Architectures & use-cases

IP FABRIC design

Use-cases

#1 IP STORAGE (ROCEV2)

#2 Hyperscalers

#3 Media & Broadcast (Timing)

#4 HBR (Host Based Routing)

10G/25G | 100G | 200G | 400G DC Fabrics

3 – stage IP Clos

5 – stage IP Clos

Collapsed Spine

EVPN-VXLAN design

#5 Traditional Ent/Telco-DC (ESI-LAG)

#6 L2 & L3 DCI (seamless tunnel stitching

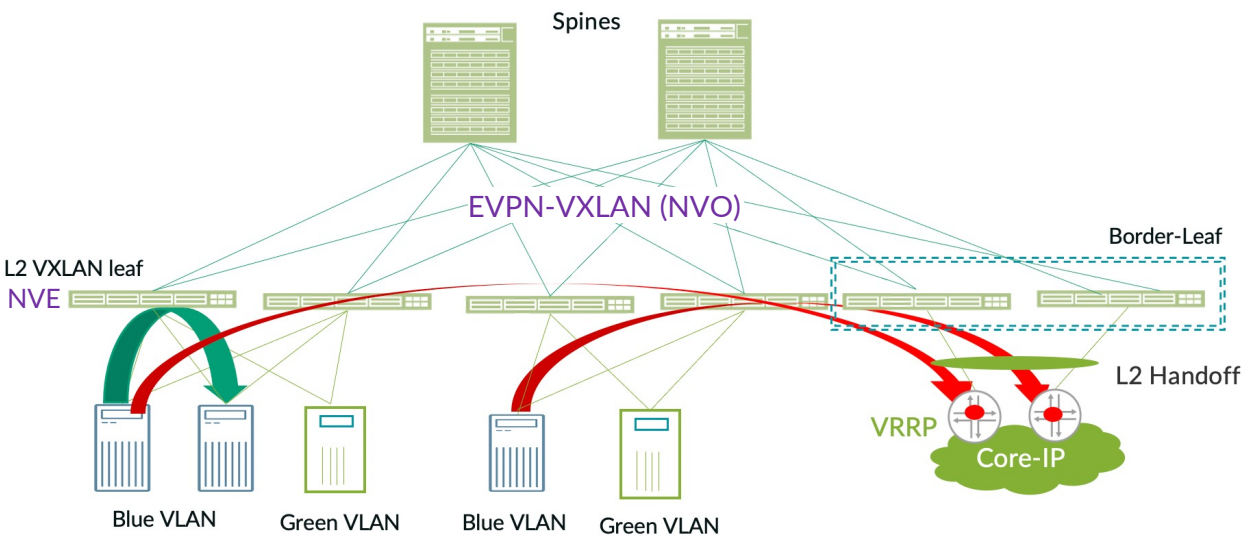
#7 5G/Remote PHY

#8 IX-Fabric Internet Exchange fabric)

Use-cases

EVPN-VXLAN architectures within the Data center

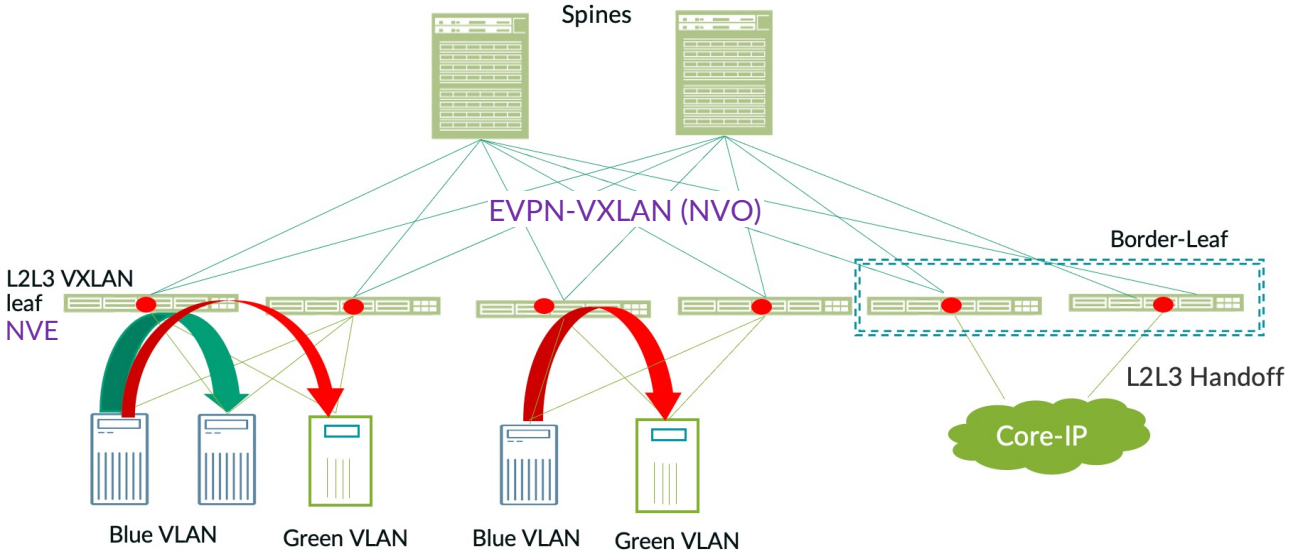
Bridged Overlay - BO



Why?

- No IP gateway migration steps
- Centralized tenant IP management

Edge Routed Bridging - ERB



Why?

- Reduced blast radius and high tenant segmentation
- Distributed tenant IP management

Optimal DC Interconnect

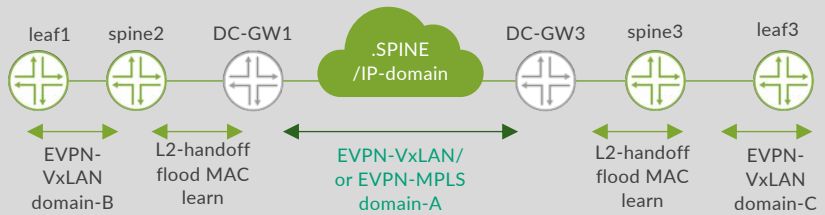
JUNIPER NETWORKS

Optimal DCI and multipod design

- Better DR overlay options for DCI and multipod DC

Option 1

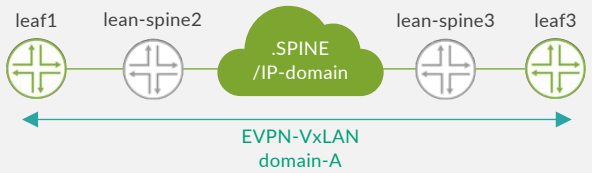
Decoupled Interconnect model using L2 VLAN handoff – L2 DCI



- Harder to manage - more devices/equipment
- Traditional demarcation points
- Flood based MAC learning

Option 2

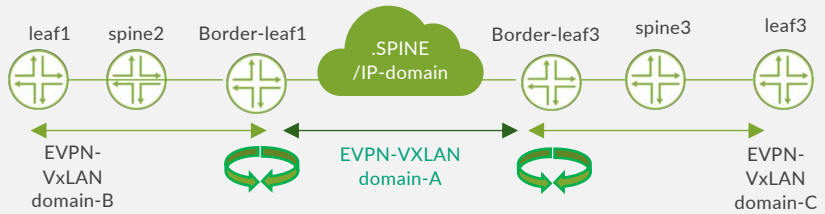
OTT – over the top full-mesh interconnect – L2 DCI



- Intermediate scale and control
- Lack of clear demarcation points
- One EVPN-VXLAN domain between DC sites

Option 3

Seamless EVPN-VxLAN to EVPN-VXLAN seamless stitching - L2 DCI



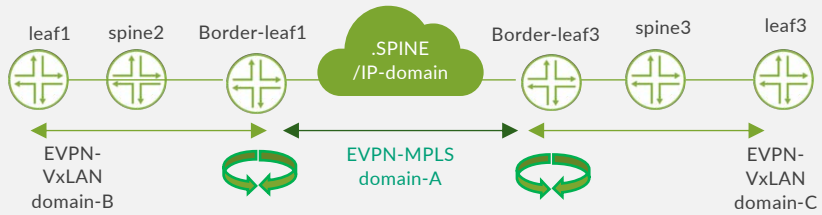
- **Unified EVPN layer 2 solution**
- **Pure overlay with controlled scaling**
- **Higher scale—more DC sites and pods**

Optimal DCI and multipod design

- Better DR overlay options for DCI and multipod DC

Option 4

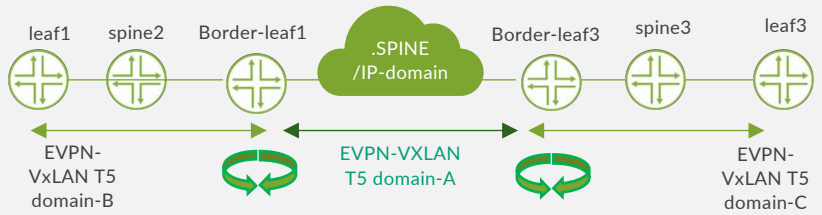
Seamless EVPN-VxLAN to EVPN-MPLS seamless stitching – L2 DCI



- Unified EVPN layer 2 solution
- Pure overlay with controlled scaling
- Higher scale—more DC sites and pods
- MPLS direct connect from border-leaf

Option 5

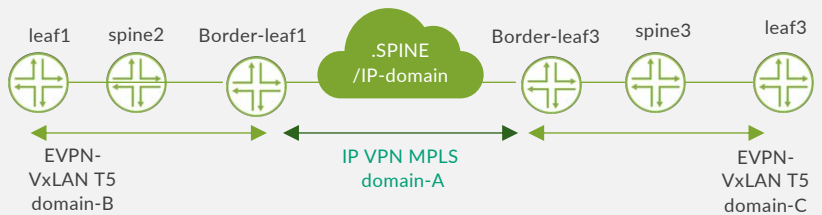
EVPN Type-5 to EVPN Type-5 VxLAN to VxLAN stitching - L3 DCI



- Unified EVPN layer 3 solution
- Pure overlay with controlled scaling
- Higher scale—more DC sites and pods
- VXLAN for IP only DCI

Option 6

EVPN-VXLAN Type-5 to IPVPN MPLS – L3 DCI



- Unified EVPN layer 3 solution
- Pure overlay with controlled scaling
- Higher scale—more DC sites and pods
- MPLS IPVPN using IP only DCI

Why DCI Seamless Interconnect? RFC9014

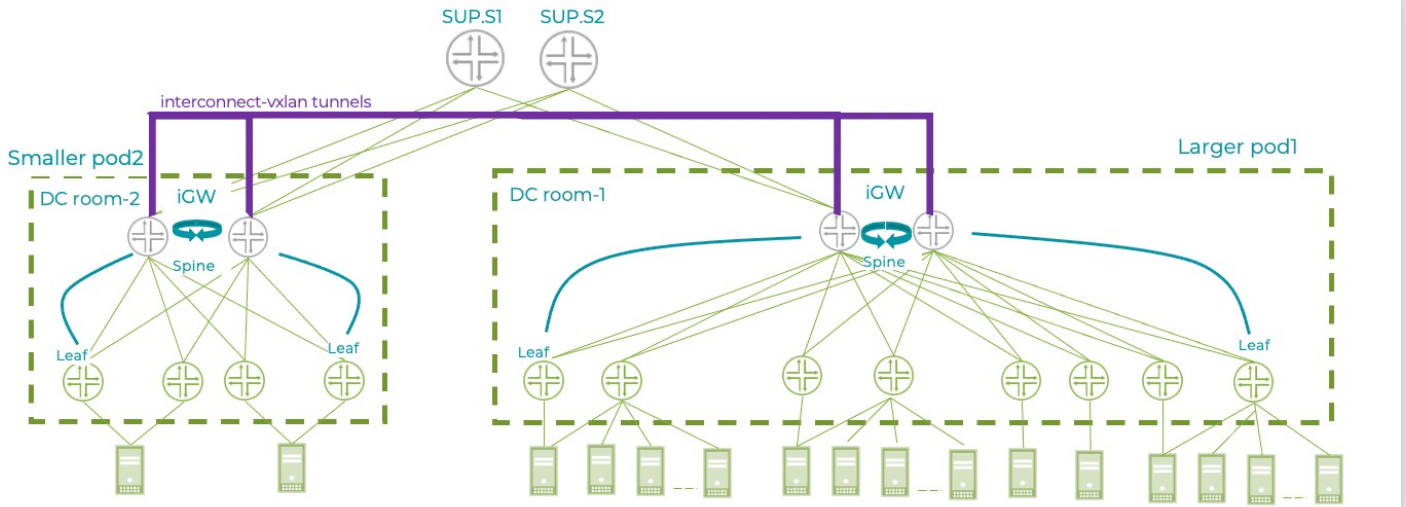
Multi-pod or DCI evpn-vxlan tunnel scale control	The vxlan-to-vxlan stitching allows to reduce and control the number of vxlan tunnels between DC rooms and sites
Easier architecture extensions when workloads are growing	Adding another pod into the existing DC architecture is simpler with no impact on the existing pod leaf operation
More efficient Ethernet flooding	Better flooding between DC rooms or sites when just selected L2& L3 domain are extended at the GW level
Additional virtualization options	Each DC pod can have different set of VNIs for secure tenant isolation, but they can still be selectively interconnected using stitching techniques
Improved operations	The operator/admin can decide which workloads get extended

Extending EVPN-VXLAN overlay

1 Multi-pod L2 DC

Larger DC fabrics with selective L2 VNI stretch

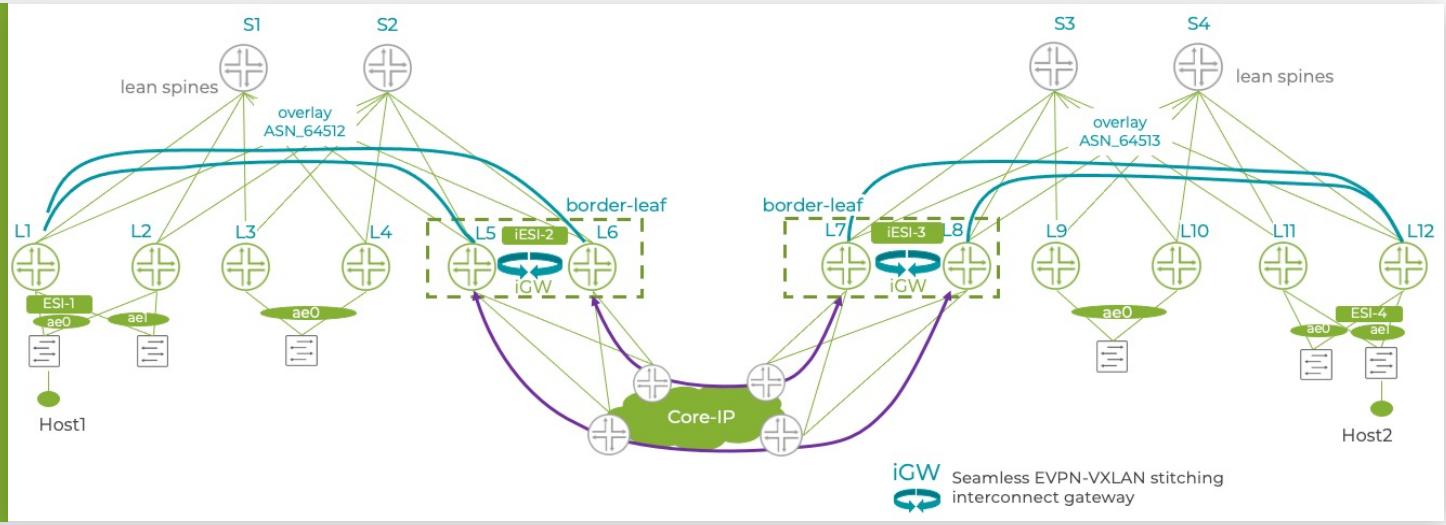
Multi-cloud L2 Data Center



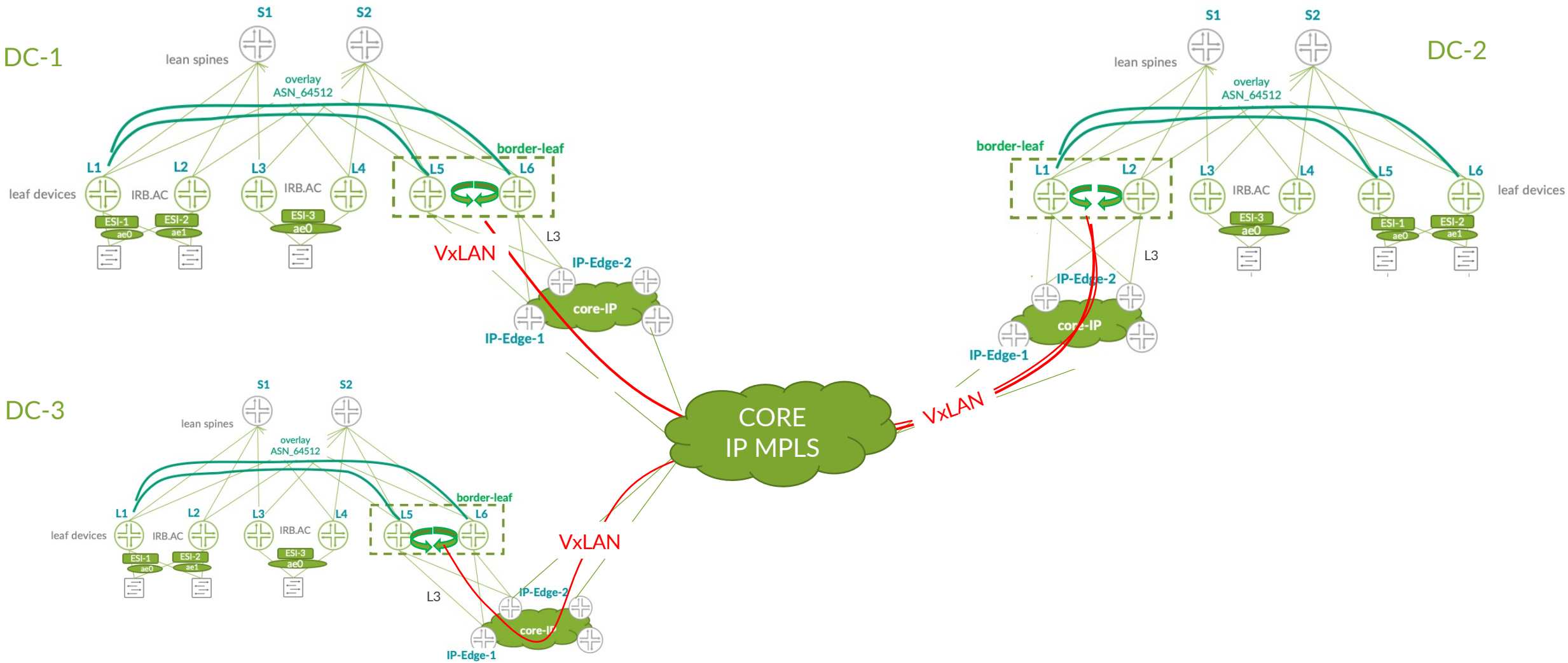
2 L2 DCI

- Multiple DC sites (2 or more)
- Medium/large DC

Multi-cloud L2 DCI

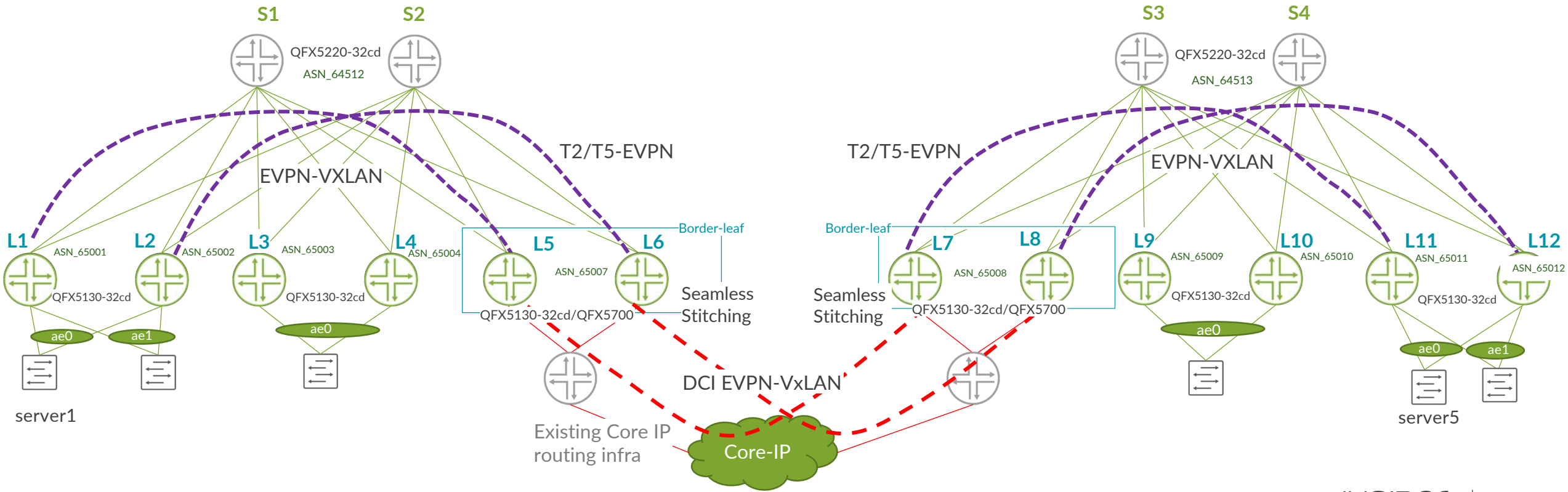


Seamless Multi-Site DCI

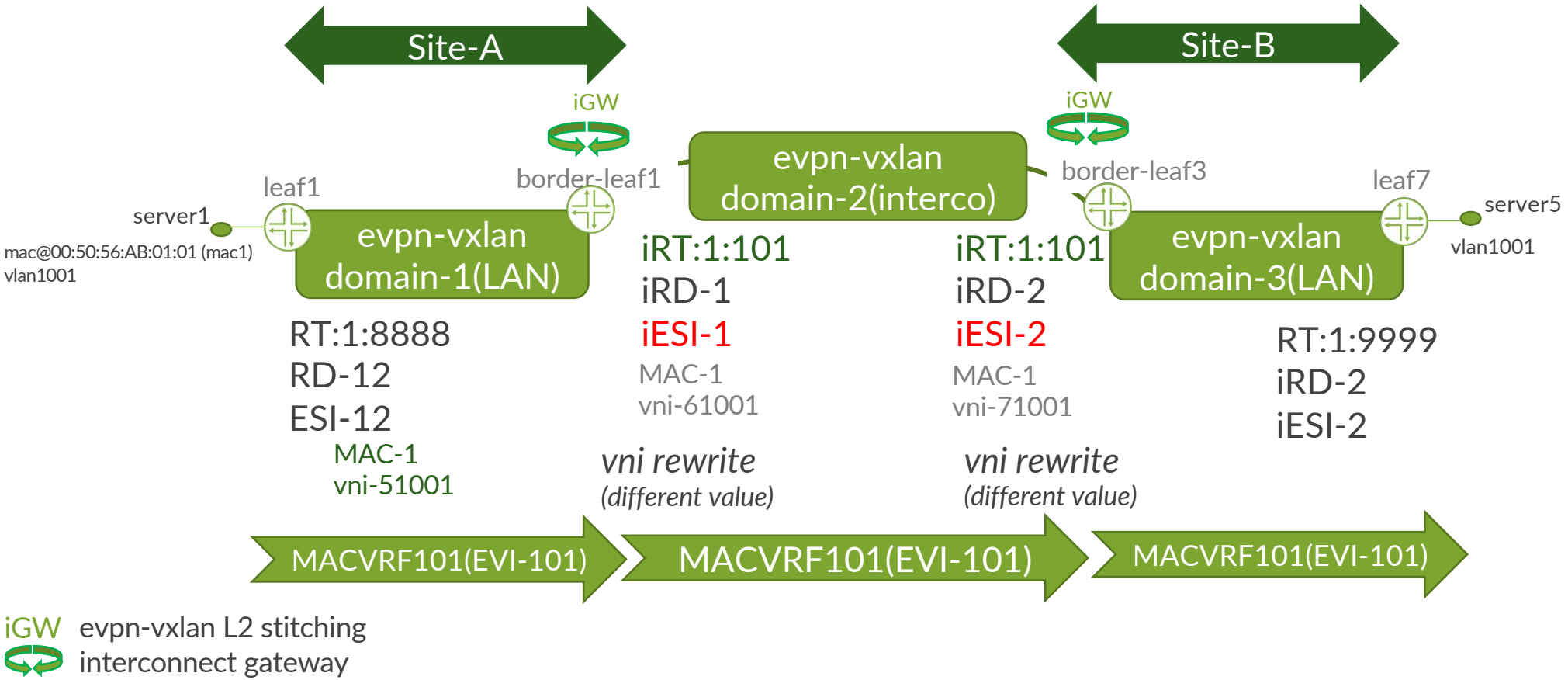


L2/L3 Seamless EVPN-VXLAN stitching

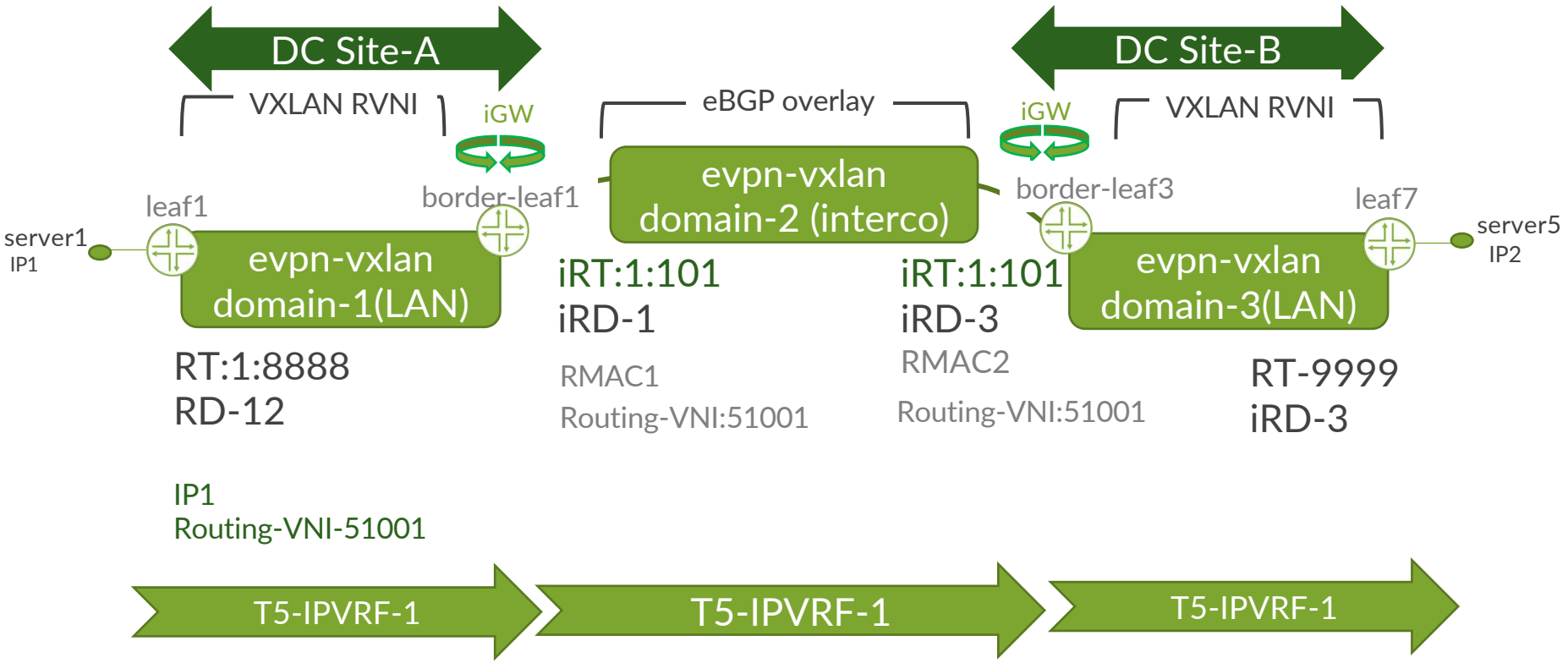
- Manageable demarcation point between LAN Fabric and DCI interconnect domain
- Multi-Site scaling
- L2 and L3 extended between the sites



EVPN-VXLAN Type-2 stitching – EVI - RT/RD/ESI/i-ESI unicast MAC

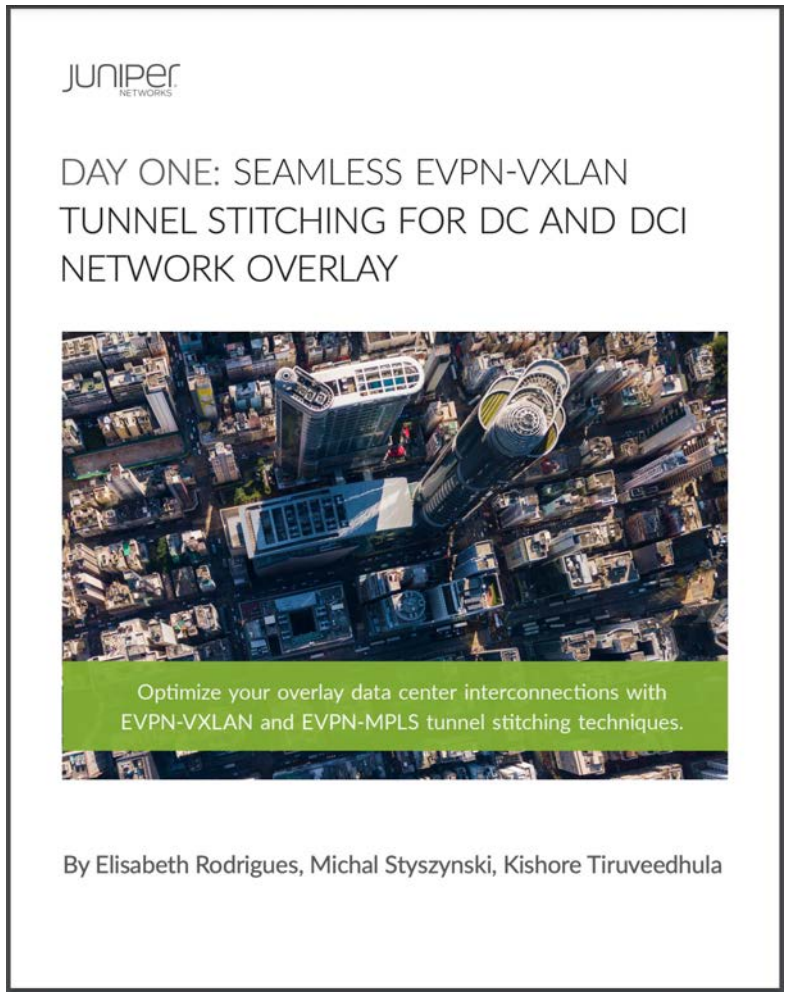


EVPN-VXLAN Type-5 stitching - RT/RD/RMAC rewrite



iGW evpn-vxlan L3 stitching interconnect gateway

New Day One Book



DAY ONE: SEAMLESS EVPN-VXLAN TUNNEL STITCHING FOR DC AND DCI NETWORK OVERLAY

https://www.juniper.net/documentation/en_US/day-one-books/DayOne-Green-Seamless_EVPN.pdf

HARDWARE EVOLUTION

Scalable IP Fabric

100/400G

Silicon

DCI

Telemetry

JUNIPER NETWORKS

QFX5K	ACX	PTX	MX
			
Shallow buffer fixed Merchant Leaf/spine	Deep buffer fixed Merchant Leaf/spine/DCI	Deep buffer fixed/modular Juniper Spine/DCI	Programmable fixed/modular Juniper DCI/DC edge

Comprehensive Datacenter portfolio

Industry-leading EVPN-VXLAN

Standards-based Inter-operability

Data Center Switching Portfolio

Rich portfolio of leaf, spine and DCI

Suffix	Typical speeds supported
T	1/10G RJ-45
S	1/10G SFP/SFP+
Y	1/10/25G SFP/SFP+
Q	40G QSFP+
L	10/25/50G SFP-56
C	40/100G QSFP+/QSFP28
CD	40/100/400G QSFP56-DD
xM	MACsec enabled (sometimes suffix not used)

(More options with breakout cables)

QFX5110-48S QFX5110-32Q	QFX5120-48Y QFX5120-48YM QFX5120-48T QFX5120-32C	QFX5130-32CD QFX5700	QFX5200-32C	QFX5210-64C	QFX5220-32CD QFX5220-128C	ACX7100-48L ACX7100-32C	PTX10001-36MR	PTX10004 PTX10008
Trident 2+	**Trident 3**	**Trident 4**	**Tomahawk**	**Tomahawk 2**	**Tomahawk 3**	**Jericho**	**Juniper Custom Silicon**	
48x10GbE and 4x40/100GbE 32x40GbE 20x40GbE and 4x100GbE	48x10/25GbE and 8x40/100GbE 48x1/10GbE and 6x40/100GbE 32x40/100GbE	32x40/100/400GbE and 2x10GbE 32x400GbE 128x40/100GbE 64x40/100GbE and 16x400GbE	32x40/100GbE	64x100GbE and 2x10GbE	32x400GbE 128x100GbE	48x25/50GbE and 6x400GbE 32x100GbE and 4x400GbE	Multi-rate 24x400GbE and 12x100GbE	576/1152 x100GbE 144/288 x400GbE
Junos OS/ with EVPN-VXLAN, IP Fabric								

QFX5K - UPDATE

JUNIPER NETWORKS

QFX5K family positioning

QFX51xx family

Chip – Trident Series

- Balance feature set
- Optimized for features
  - L2 over L3 with EVPN-VXLAN
  - Overlay multicast replication
- Server connectivity with native SFP optics (mostly)

QFX52xx family

Chip – Tomahawk Series

- Optimized for speed
  - Multi-stage IP fabric
- Lower latency
- Dynamic load balancing
- Server connectivity with QSFP optic break-out (mostly)

QFX5120 Series - Trident3 10G/25G/100G (1 TBPS - 3.2 TBPS)

QFX5120-48Y

48x 1G/10G/25G SFP + 8 X100 QSFP

QFX5120-32C

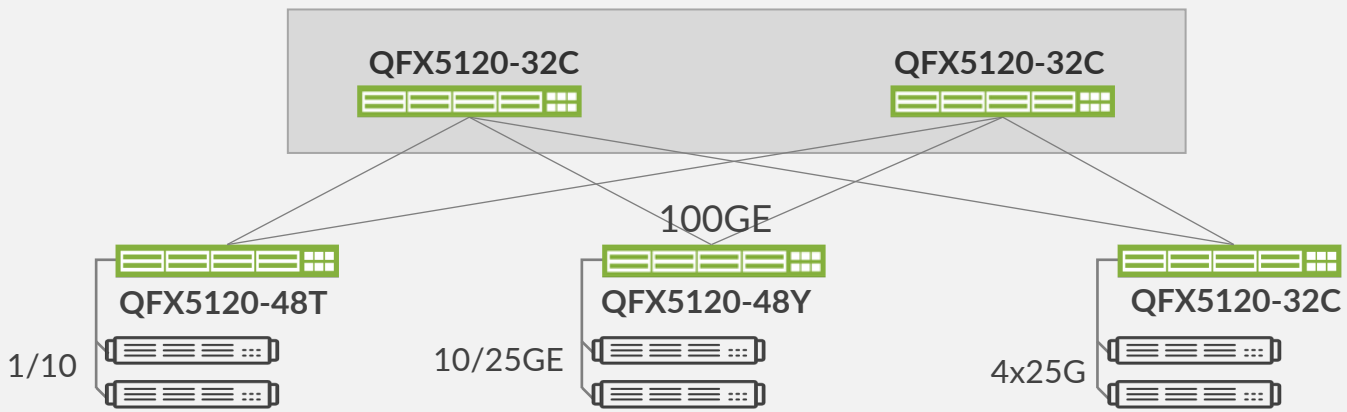
32x 100G QSFP28 + 2x 10G SFP

QFX5120-48T

48x 1G/10G Copper + 6x 100G QSFP
PTP timing

QFX5120-48YM

48x 1G/10G/25G SFP + 8 X100 QSFP
PTP timing - MACSec on all ports



Features	5110	5120
VFI	8K	16K
L3 Host (UFT)	128K	168K
LPM (UFT)	128K	350K
Next Hops	48K	64K
VxLAN L3	Yes	Yes
Buffer	16MB	32MB

QFX5130/QFX5700 - Trident4 based platforms

QFX5130-32CD

Spine, Server Leaf, Border Leaf

- 1 RU
- TD4 based (12.8Tbps)
- 32x400G
- 4x100G (break out)

QFX5700

Spine, Border Leaf

- 5U single PFE chassis
- TD4 based (12.8Tbs)
- 8 linecards
  - 4x400G
  - 16x100G (MACsec capable)
  - 20x10G/25G (MACsec capable)
- Up to 32x400G or 128X100G

Features

Features
Over 1M LPM table (IPv4) 620K IPv6 LPM table
JUNOS -EVOLVED
Higher buffer (132 MB)
1M LPM table (IPv4)
24K filters/pipe
384K Mac table
PTP
Higher buffer (132 MB)
Inband Telemetry (IFA)

QFX5200 Series - 25G/100G (3.2Tbps - 6.4 Tbps)

QFX5200-32C

32x100G

Tomahawk (3.2Tbps)
16MB buffer
Leaf and spine in native IP fabric
Lean spine in EVPN-VxLAN fabric

QFX5210-64C

64x100G
2 RU
Tomahawk2 (6.4Tbps)
42MB buffer
Lean spine in EVPN-VXLAN
ERB/BO design

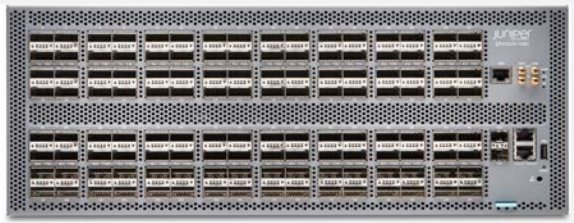
Features	Tomahawk	Tomahawk2
IPv4	160K	320K
IPv6	84K	160K
ECMP (groups/members)	2K/16K	4K/32K
ECMP	64-way	64-way
MPLS labels pushed	3	8
MPLS entries	16K	32K
VxLAN Layer2	Yes	Yes
VxLAN Layer3	No	No

QFX5220 Series - 50G/100G/400G
(TH3-12.8 TBPS)

QFX5220-32CD

32x400G/128x100G/64x200G
Leaf and spine in native IP fabric
Lean spine in EVPN-VxLAN fabric

QFX5220-128C



128x100G
Lean spine in EVPN-VXLAN
ERB/BO design

Features	Tomahawk	Tomahawk2	Tomahawk3
IPv4	160K	320K	> 400K
IPv6 (64b)	84K	160K	> 300K
ECMP (groups/members)	2K/16K	4K/32K	4K/64K
MPLS labels pushed	3	8	8
MPLS entries	16K	32K	16K

Highlights

- Industry leading 100G/400 port density
- 64MB buffer
- Low latency
- JUNOS-EVO
- Target customers: large-scale cloud data centers
- PTP BC
- Data-packet timestamping
- L2 multicast: IGMP snooping/PIM snooping
- 100G/400G FR/DR

ACX 7100 - UPDATE

JUNIPER NETWORKS

Deep buffer to support cloud DC applications

Need for deep buffers

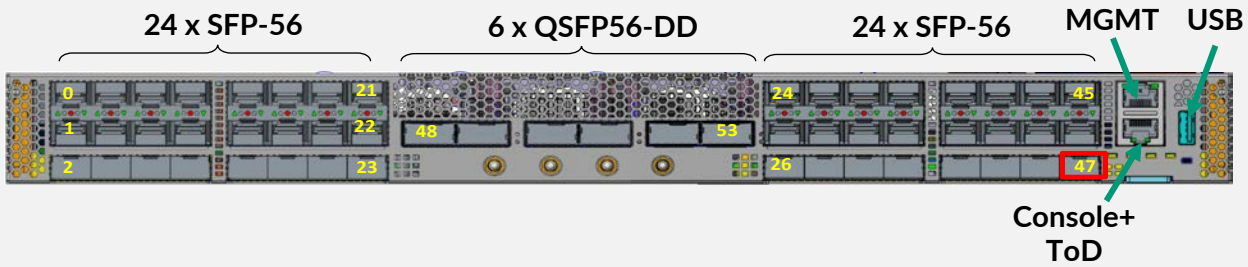
- Handle traffic bursts and minimize traffic loss
- Deep buffers work hand in hand with QoS functionality and help keep Service Level Agreements
- Keep interface links fully utilized with minimal trade offs
- Reduce or eliminate choke points in network
- Enable smooth performance of data center and Big Data applications that involve many connections with varying latency and throughput patterns

ACX7100-48L

ACX7100-32C

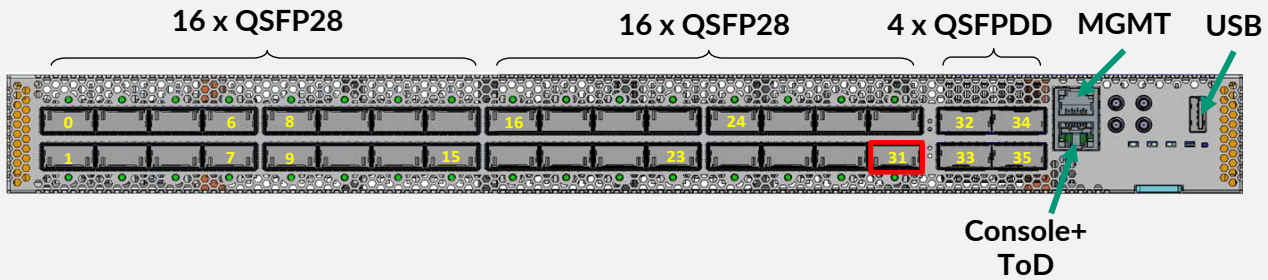
ACX7100 Series - deep buffer fixed spine and leaf

ACX7100-48L



Deep Buffer
TOR/Leaf/Border leaf
48 x 10G/25G/50G
100G/400G uplinks
8GB buffer

ACX7100-32C



Deep Buffer
Spine/Leaf/Border leaf
100G optimized with 400G uplinks (Leaf) OR
36x100G (spine)
8GB buffer

PTX

JUNIPER NETWORKS

PTX10K - Juniper 400G silicon for spine, DCI, peering, core

8 - slot

PTX10008

36x400G LC 144x100G
32x100G + 4X400G LC*

4 - slot

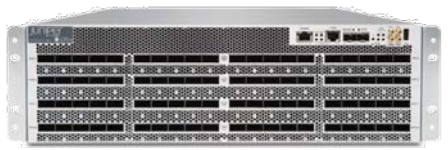
PTX10004

36x400G LC 144x100G
32x100G + 4X400G LC

1 RU

PTX10001-36MR

4x400G + 32x100G
24x400G +12x100G



3 RU

PTX10003-160C

32x400G + 32x100G
160x100G

3 RU

PTX10003-80C

16x400G + 16x100G
80x100G

MACsec on chip

Deep buffer

High scale

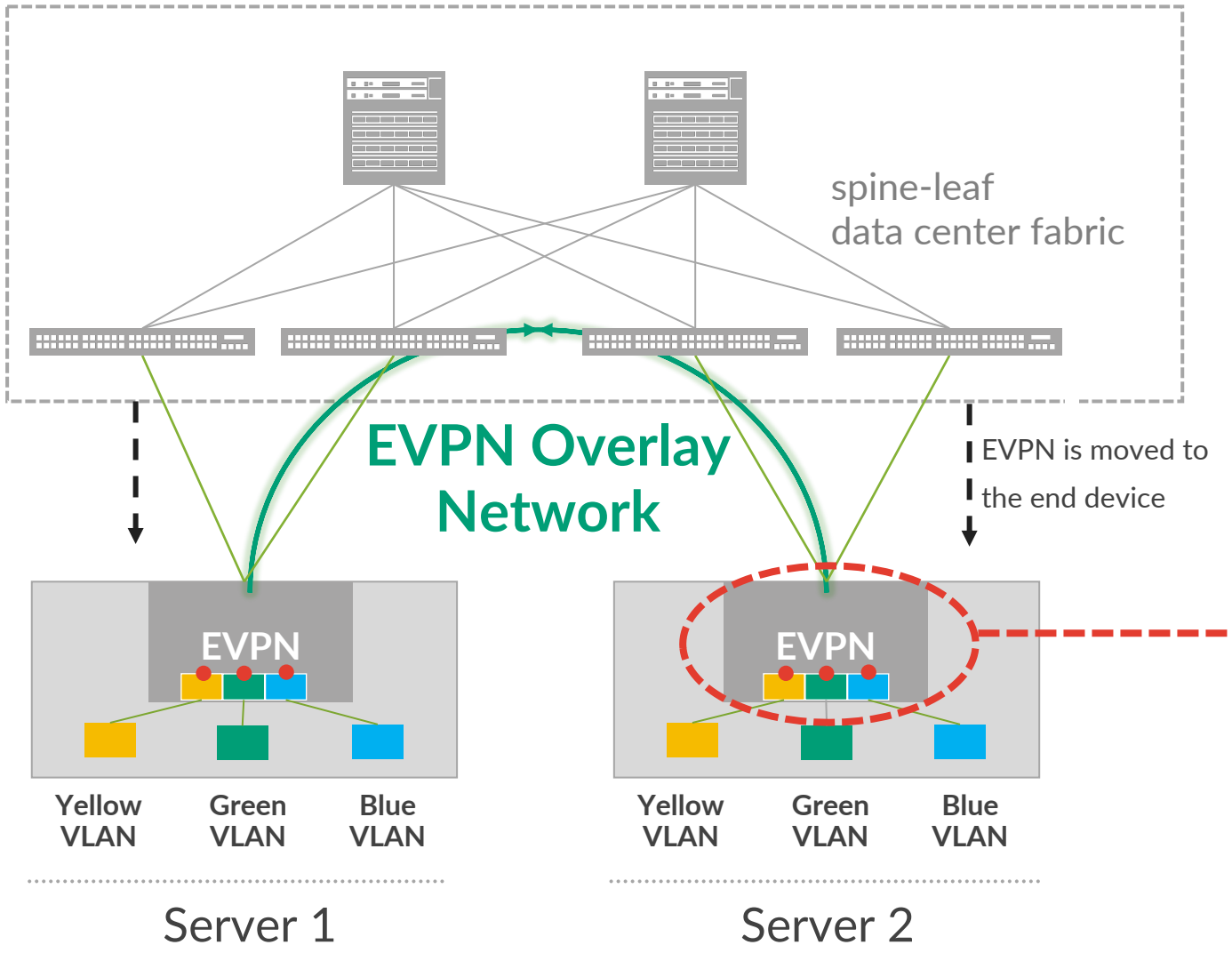
Converged portfolio for spine, DCI, peering and WAN backbone

NATIVE IP WITH HOST ROUTED BRIDGING

JUNIPER NETWORKS

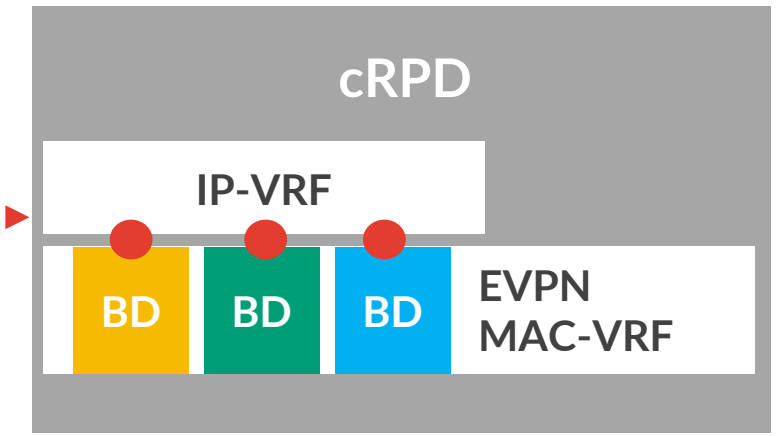
What is EVPN Host Routed Bridging (HRB)?

cRPD - Deployable on demand as a microservice with Kubernetes integration (CNI)



Eliminating L2 Looping	Load balance
No need for multihoming for link failure protection between "PE" and "CE"	Up to 128 ECMP in the underlay
No need for DF election No need for ES split horizon or local bias	A scale out data center with K8s automatic failure recovery
Increase network reliability	Auto life cycle management

HRB



● Integrated Routing and Bridging (IRB)

BD: Bridge Domain

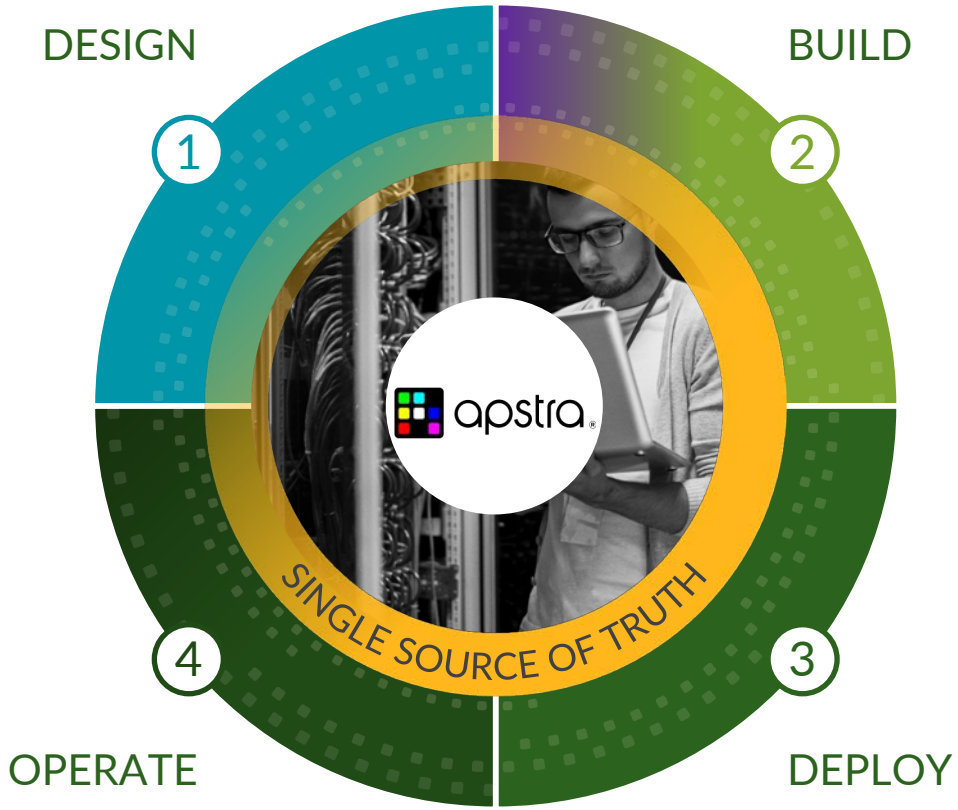
apstra®

JUNIPER NETWORKS

Why should you care about Data Center automation?

- Configuration consistency
- Managing fabric as a single entity
- Abstracts away the complexity of Clos fabric deployments
- Integration with northbound orchestrator
- Closed-loop automation
- Monitoring, analytics and service assurance
- Emphasis on Customer Experience
- Lower TCO

Automated DC Operations using APSTRA



Day 0 — **Design & Build**

"Networks must be agile, not fragile"

Day 1 — **Deploy**

"Automation By Default, Manual By Exception"

Day 2+ — **Operate**

"Visualize and Troubleshoot The Network"

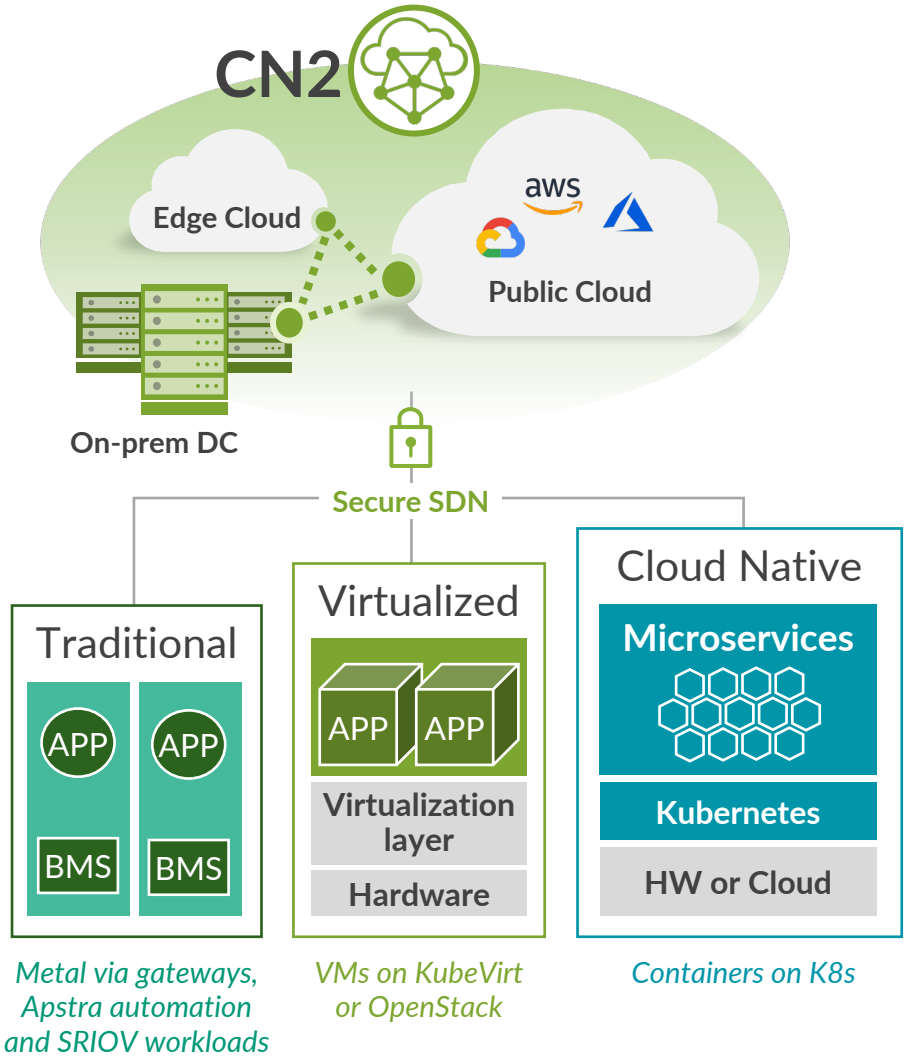
- Intent-based
- Multi-vendor
- Root Cause Identification
- Closed-loop automation

Contrail (CN2) differentiation

JUNIPER NETWORKS

CN2: Cloud-Native Contrail Networking

Values: Investment protection, flexibility, agility, and improved economics



K8s-Native SDN

A seamless experience built into Kubernetes itself

Cloud-Native Networking

Hybrid/Multicloud consistency for better operational economics

Hybrid SDN for K8s and OpenStack

Infrastructure investment protection and evolvable infrastructure

NetOps-Driven Automation

Simple, repeatable CI/CD pipeline test assurance at cloud scale

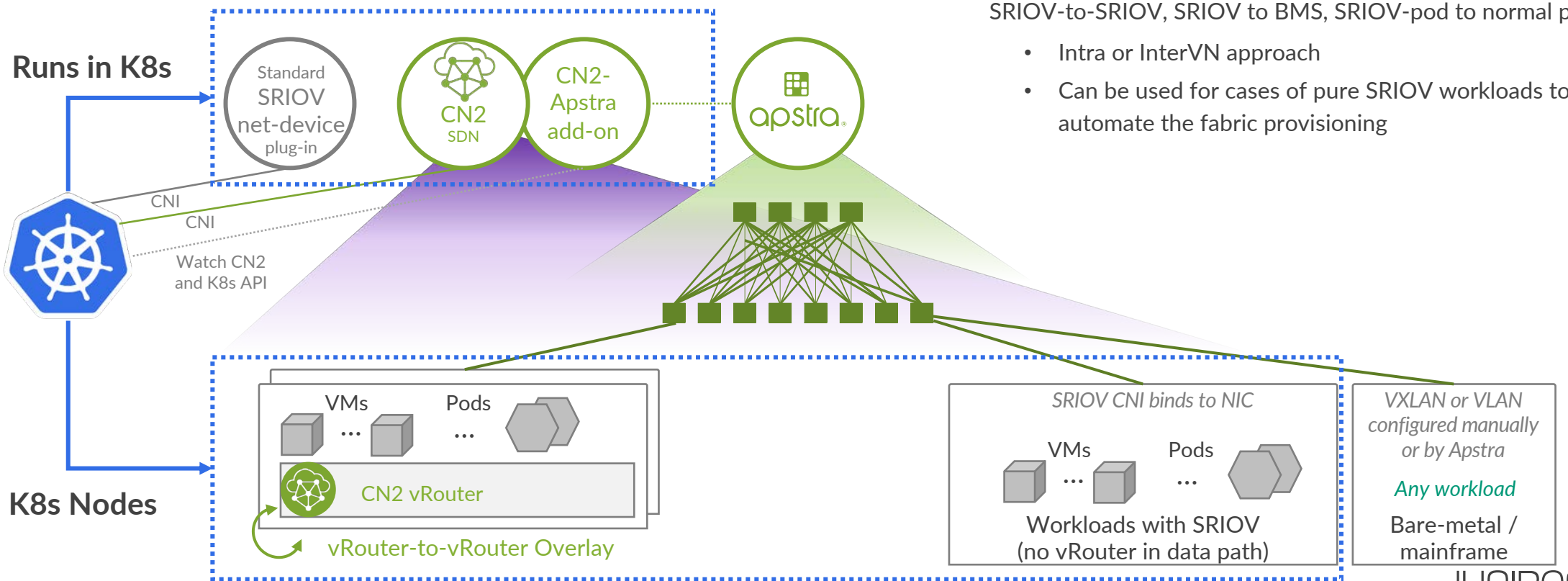
One-to-Many Operational Economics

Centralized multi-cluster networking and monitoring for scalable ops

CN2 and Apstra Integration: Use Cases

New Automated Connectivity for:

1. **Enterprise use cases**
   Connect normal pods to metal/VM or non-K8s containers
   - Intra VN (extend the subnet)
   - Inter VN (different subnets)
2. **Telco Cloud and ML-learning use cases with SRIOV**
   SRIOV-to-SRIOV, SRIOV to BMS, SRIOV-pod to normal pod
   - Intra or InterVN approach
   - Can be used for cases of pure SRIOV workloads to automate the fabric provisioning



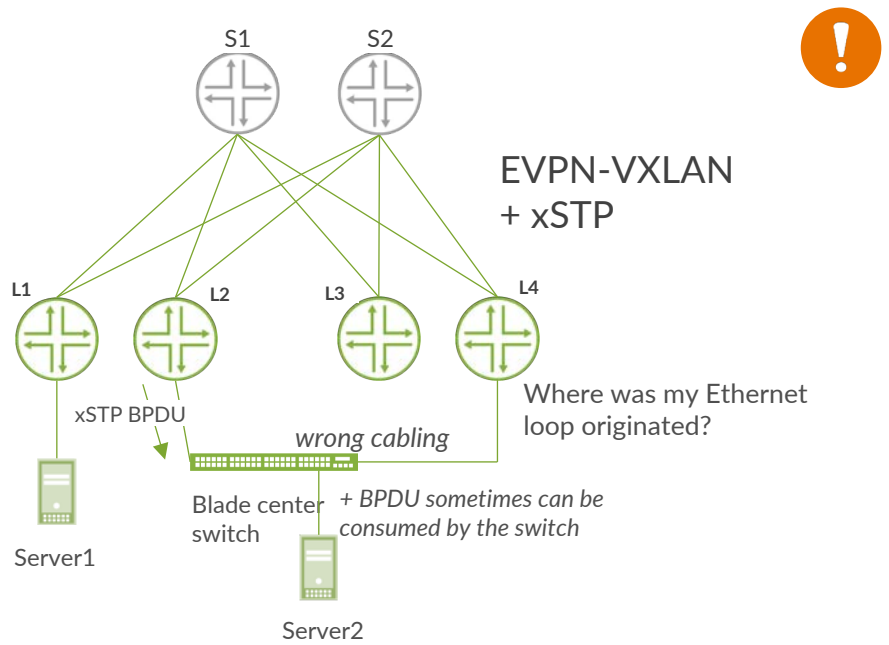
EVPN/VXLAN Innovations

JUNIPER NETWORKS

Infrastructure stability through enhanced loop detection

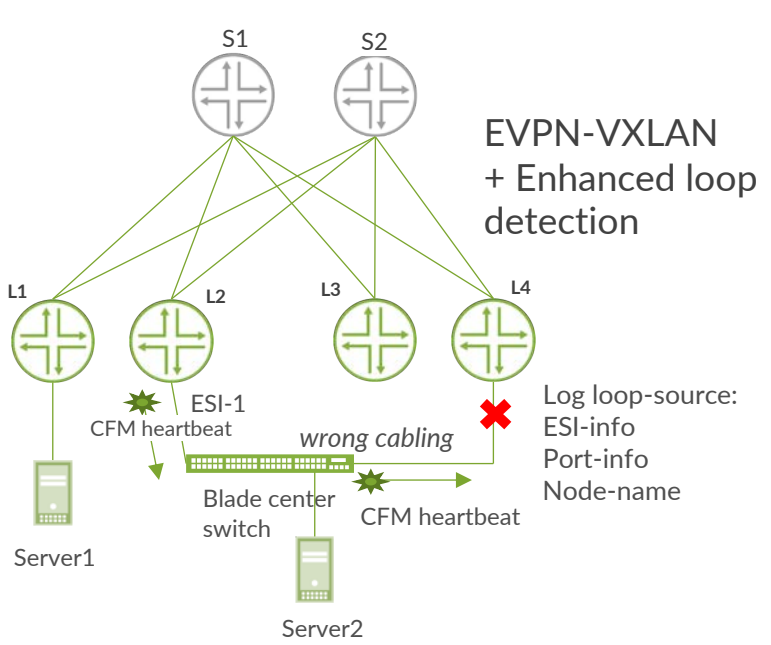
Other DC Fabric Solutions

EVPN-VXLAN yes, but… still spanning-tree for loop-detection



Juniper DC Solution

No more spanning-tree on EVPN-VXLAN nodes!

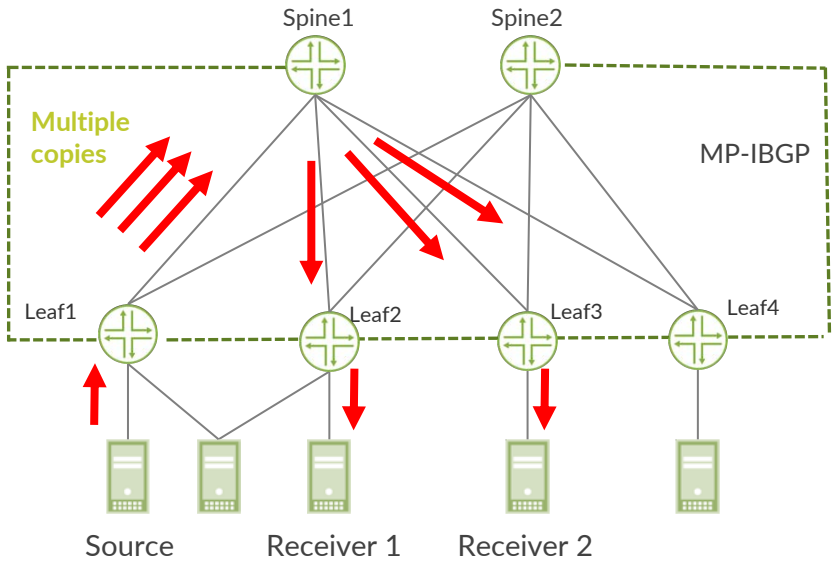


Full overlay solution

- Lower TCO assisted replication at the selected spines

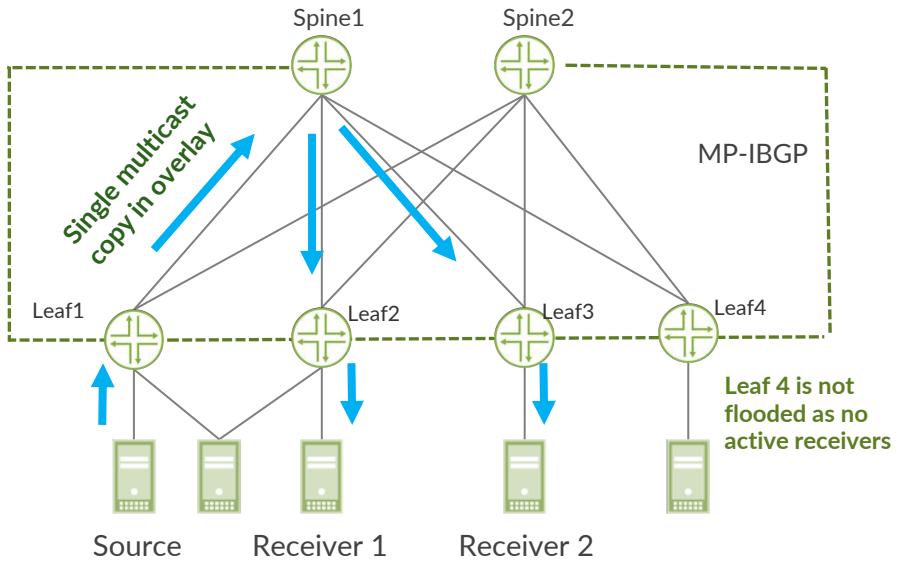
Other DC solutions

Multicast uses EVPN overlay



Juniper DC solutions

Multicast w/ EVPN overlay + assisted replication + selective



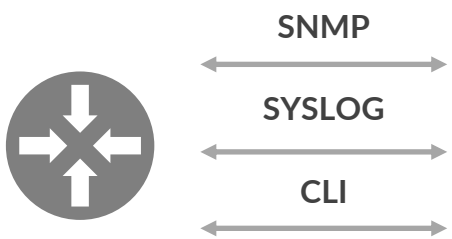
Experience enabled with telemetry

Circa ~2017 | Today | 2022+

TD3 – 2H 2021; TD4/TH3 – 2022

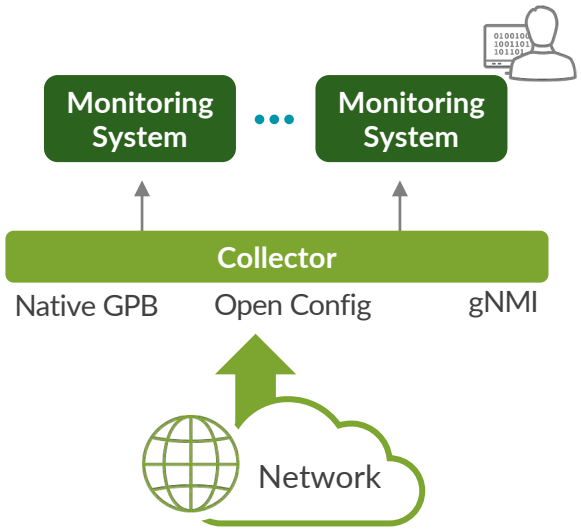
LEGACY MONITORING

Resource intensive,
hard to automate



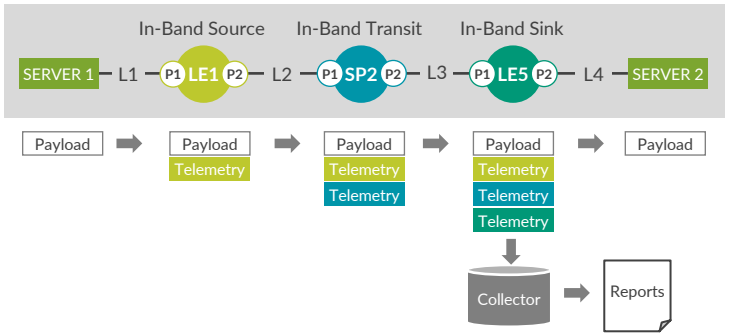
STREAMING TELEMETRY

Real time, data model driven

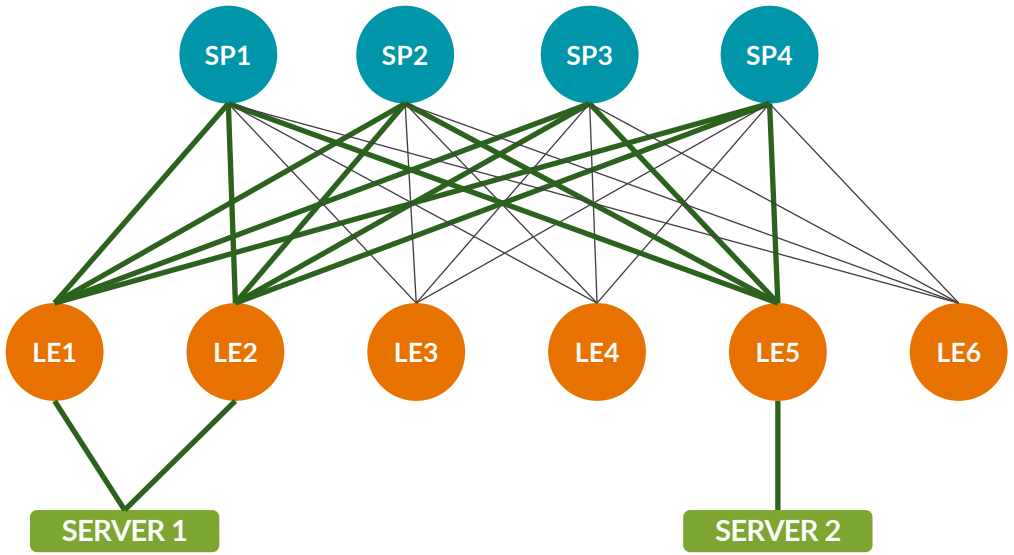


IN-BAND TELEMETRY (IFA)

Latency, congestion



What is the need for Inband Telemetry?

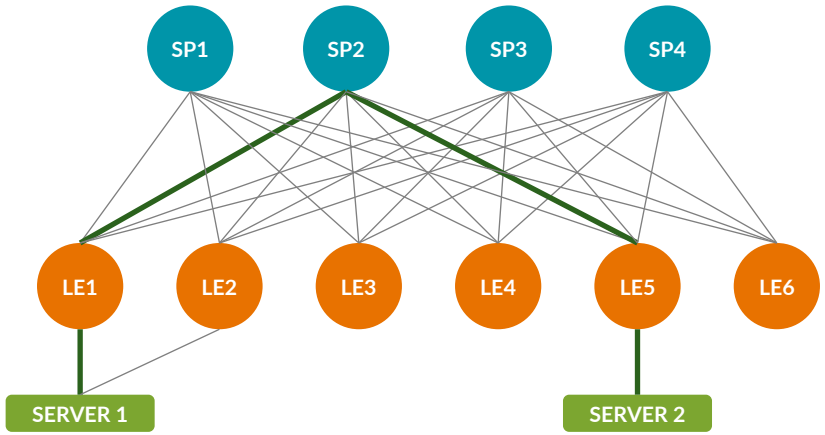


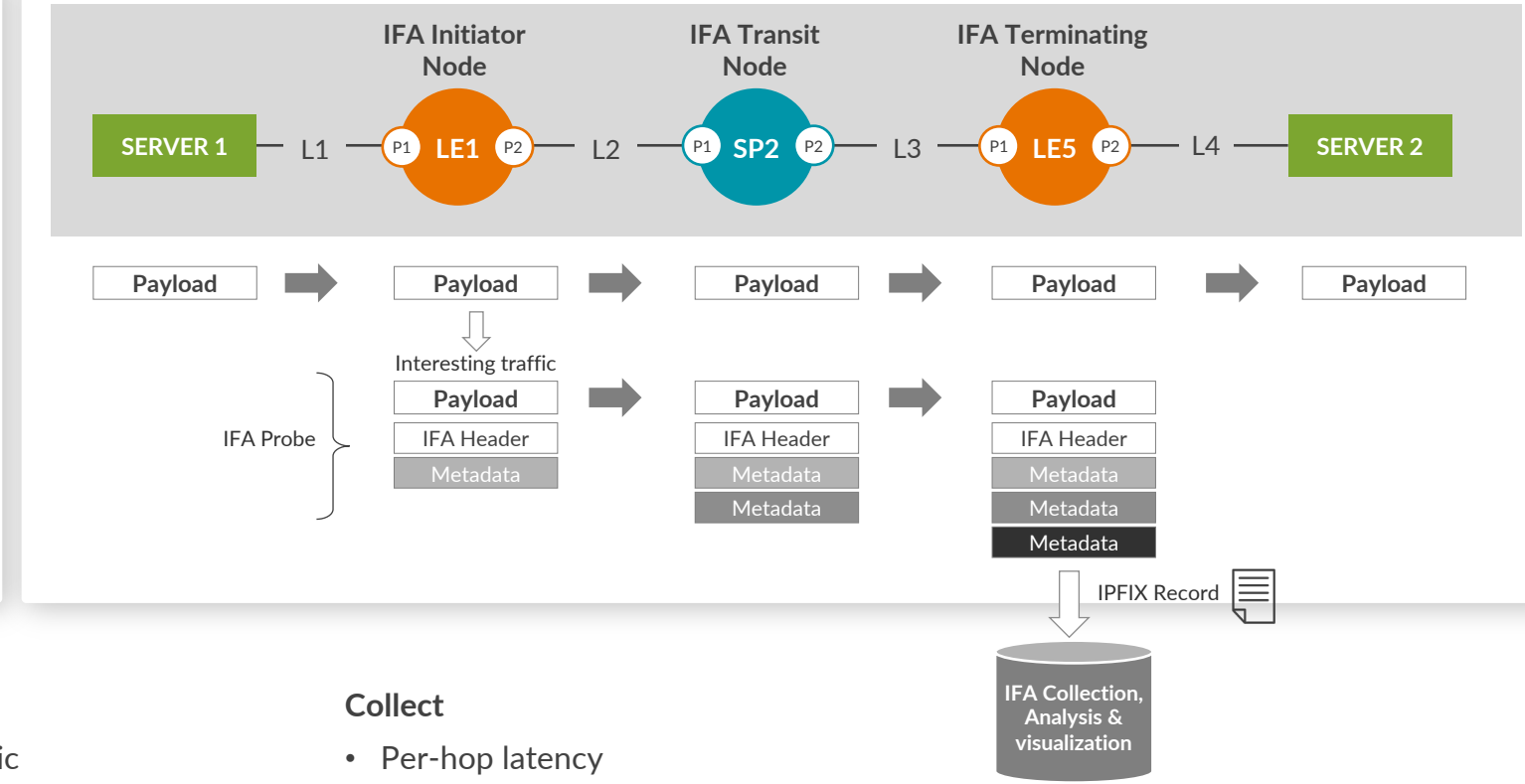
Overlay fabrics introduce additional layers for troubleshooting

Identifying physical path for traffic flow can be challenging with IP ECMP and/or LAG

Other troubleshooting tools like sFlow, overlay ping/traceroute, streaming telemetry lack granularity and latency information

How does IFA 2.0 Inband Telemetry work and its benefits





Benefits

- Inline probes help monitor live traffic
- Traverse same path as original packet
- Get exact pathtrace (despite Leaf-Spine ECMP or Server-Leaf LAG)
- Monitor performance issues like latency and congestion

Collect

- Per-hop latency
- Per-hop ingress/egress port numbers
- Congestion indication
- Queue id
- Egress port speed
- RX timestamp

Connected Security (fabric + security)

JUNIPER NETWORKS

JUNIPER THREAT LABS

#1 In Every Test For The Past Four Years

Independently Validated Security Efficacy

NetSecOPEN
2022
99.8%
effectiveness against exploits

Outperforming "leading" vendors
- Fortinet
- Palo Alto Networks
- Zscaler
- Checkpoint
- Cisco

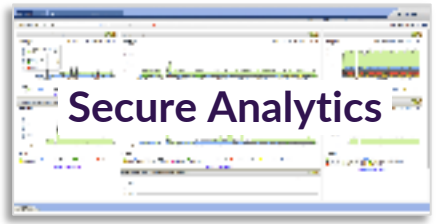
Cyber Ratings — AAA Rating
2021 Next-Gen Firewall
99.5%
against client and server-side exploits

Cyber Ratings — AAA Rating
2022 Cloud Network Firewall
100% Exploit block rate
0 False positives

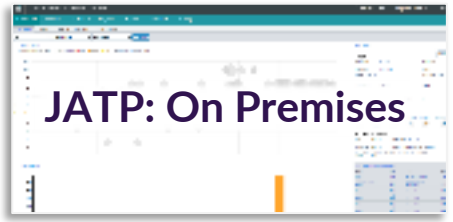
ICSA Labs
Advanced Threat Defense
Q3 2022
Q2 2022
Q1 2022
Q4 2021
Q3 2021
Q2 2021
Q1 2021
Q4 2020
100% malware block rate
<0.5% false positives

Security portfolio

Management, Automation



SIEM



Advanced Threat Prevention

IPS/IDS
Anti Malware
Application Security
URL Filtering
IoT Security

Next-Gen Security Services

+ Cluster capabilities
Active/Standby, Active/Active

4 Gb/s (2 vCPU)
40 Gb/s (16 vCPU)

vSRX

cSRX*

1U
5Gb/s

SRX1500

1U
20Gb/s

SRX4100

1U
40Gb/s

SRX4200

1U
80Gb/s

SRX4600

Advanced Security Acceleration (SPC3)

5U
320 Gb/s

SRX5400

8U
960 Gb/s

SRX5600

16U
1.4 Tb/s

SRX5800

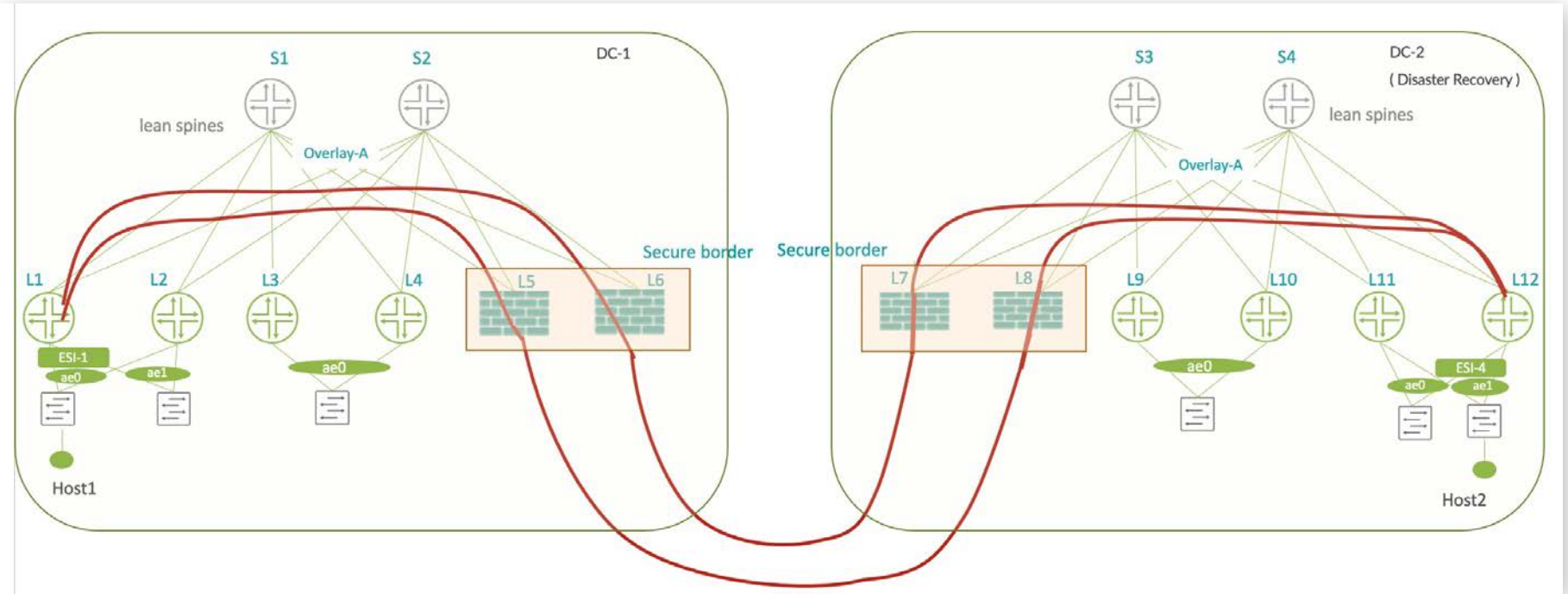
Campus | Private Cloud/Multicloud | Large Data Center/Service Provider

Routing/SD-WAN | IPsec/VPN | High Availability | SSL/TLS Proxy

SRX in underlay - tunnel inspection

Problem Statement: inspect tunnel content (inner packet) anywhere in the network where there is no opportunity to terminate the tunnel first.

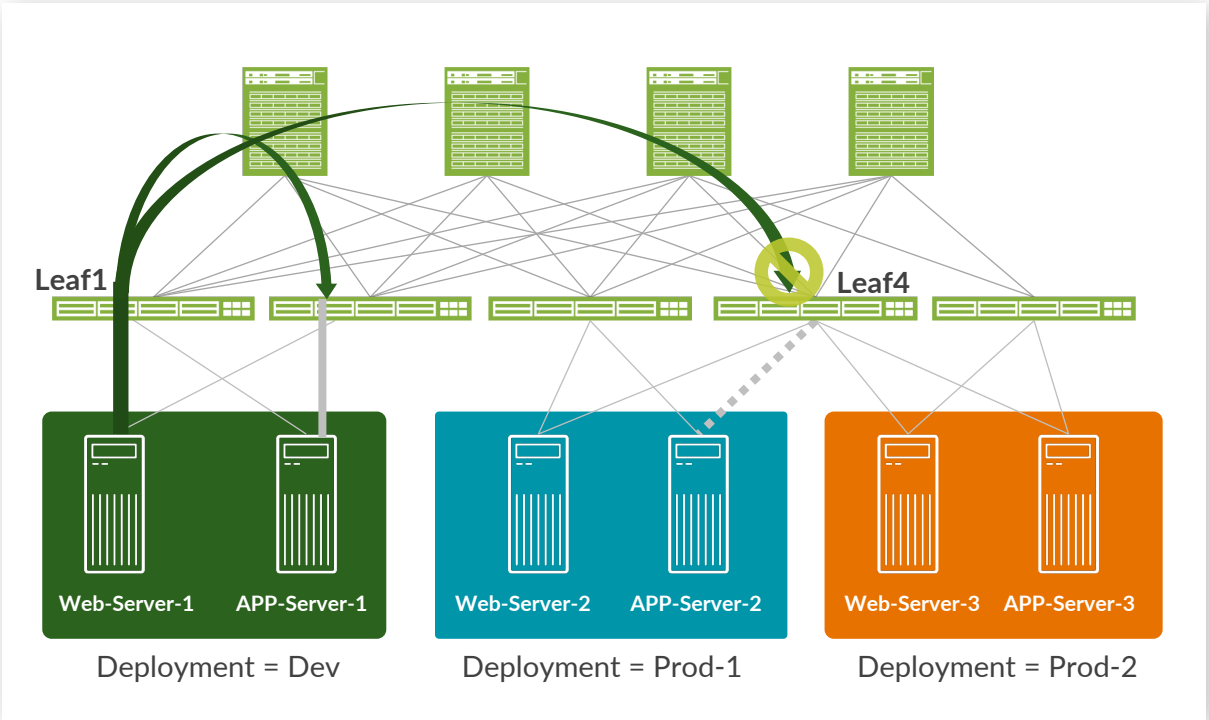
Solution: as long as SRX is in the path of the tunnel, it can inspect the VXLAN tunnel content for Layer 4/Layer 7 inspections – UTM (Unified Threat Management), IDP (Intrusion and Detection Policy), IPS (Intrusion Prevention System) IDS (Intrusion Detection System) and more.



- **Supports IPV6, IPV4 traffic**
- **Per VNI inspection**
- **Junos OS Release 21.1R1**

Group based policy

Goals
- Solve Scalability
- Manageability



Egress enforcement at leaf 4

Takeaway: Support larger number of workloads

Classification

Tier	Tag	Deployment
WebServer-1	10	Dev
AppServer-1	10	Dev
WebServer-2	20	Prod
AppServer-2	20	Prod
WebServer-3	30	Prod
AppServer-3	30	Prod

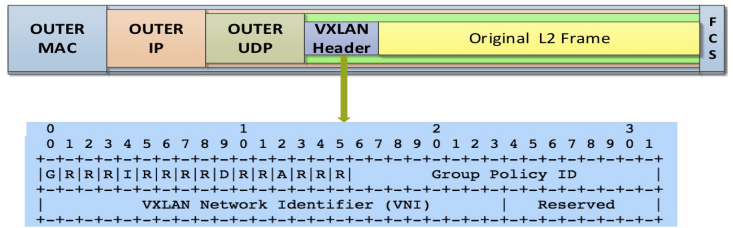
```
>> set firewall family ethernet-switching filter f1 term t1 from match { smac <Web-Server1-1, App-Server-1> }
>> set firewall family ethernet-switching filter f1 term t1 then gbp-src-tag 10
>> set firewall family ethernet-switching filter f1 term t1 from match { smac <Web-Server-2, App-Server-2> }
>> set firewall family ethernet-switching filter f1 term t1 then gbp-src-tag 20
>> set firewall family ethernet-switching filter f1 term t1 from match { smac <Web-Server-3, App-Server-3> }
>> set firewall family ethernet-switching filter f1 term t1 then gbp-src-tag 30
```

Firewall rules (TCAM space)

Src. GroupTag	Dst.Group.Tag	Policy
10	10	ALLOW
10	20	DENY

```
set firewall family ethernet-switching filter f1 term t1 from {gbp-src-tag 10}
set firewall family ethernet-switching filter f1 term t1 from {gbp-dst-tag 10}
set firewall family ethernet-switching filter f1 term t1 from  port 80
set firewall family ethernet-switching filter f1 term t1 then accept
set firewall family ethernet-switching filter f1 term t2 then deny
```

Each one of the dest Leaf will have these 5 lines of config



VXLAN header format with 16-bits of Group Policy ID (Tag)

СПАСИБО!

ask-se@juniper.net

dzavelskiy@juniper.net

JUNIPER NETWORKS | Driven by Experience™

Q&A

JUNIPER NETWORKS