BROCADE

# “IP Fabric” Reference Architecture

## Technical Deep Dive

Joerg Ammon jammon@brocade.com

Principal Systems Engineer Service Provider

# Feng Shui of Data Center Design

1. Follow KISS Principle “Keep It Simple…”
2. Minimal features
3. Minimal configuration
4. Configuration repeatability
5. Automate everything

# Reference Architecture

[Docs] [txt|pdf|xml|html] [Tracker] [WG] [Email] [Diff1] [Diff2] [Nits]

Versions: (draft-lapukhov-bgp-routing-large-dc)
00 01

Routing Area Working Group
Internet-Draft
Intended status: Informational
Expires: August 16, 2015

P. Lapukhov
Facebook
A. Premji
Arista Networks
J. Mitchell, Ed.
Microsoft Corporation
February 12, 2015

**Use of BGP for routing in large-scale data centers**
**draft-ietf-rtgwg-bgp-routing-large-dc-01**

Abstract

Some network operators build and operate data centers that support over one hundred thousand servers. In this document, such data centers are referred to as "large-scale" to differentiate them from smaller infrastructures. Environments of this scale have a unique set of network requirements with an emphasis on operational simplicity and network stability. This document summarizes operational experience in designing and operating large-scale data centers using BGP as the only routing protocol. The intent is to report on a proven and stable routing design that could be leveraged by others in the industry.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 16, 2015.

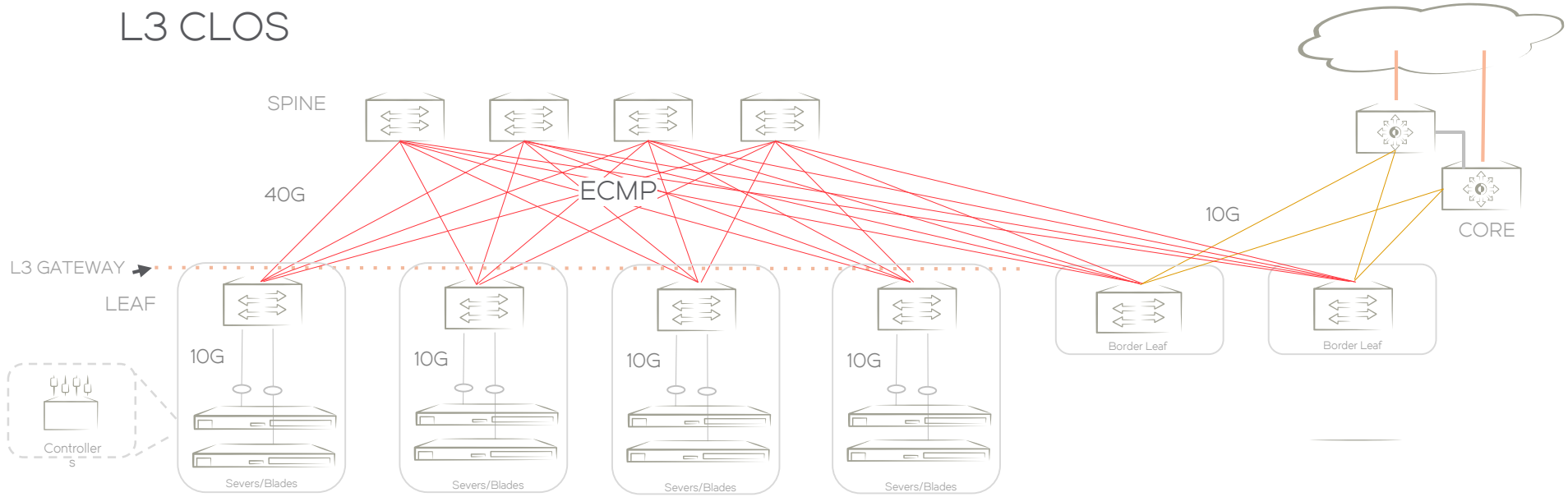
Source:draft-ietf-rtgwg-bgp-routing-large-dc-01

*After experimentation and extensive testing, **Microsoft** chose to use an end-to-end routed network infrastructure with External BGP (**EBGP**) [RFC4271] as the only routing protocol for some of its DC deployments*

Authors
P. Lapukhov@ **Facebook**
A. Premji@ **Arista Networks**
J. Mitchell, Ed@ **Microsoft**

# IP Fabric: Reference Architecture

## L3 CLOS




**Rack**

1 ToR per Rack – server directly connected

**Leaf**

Uplink port count on the leaf switch determines the max # of spine switches

**Spine**

Spine switch port count determines the max # of leaf switches

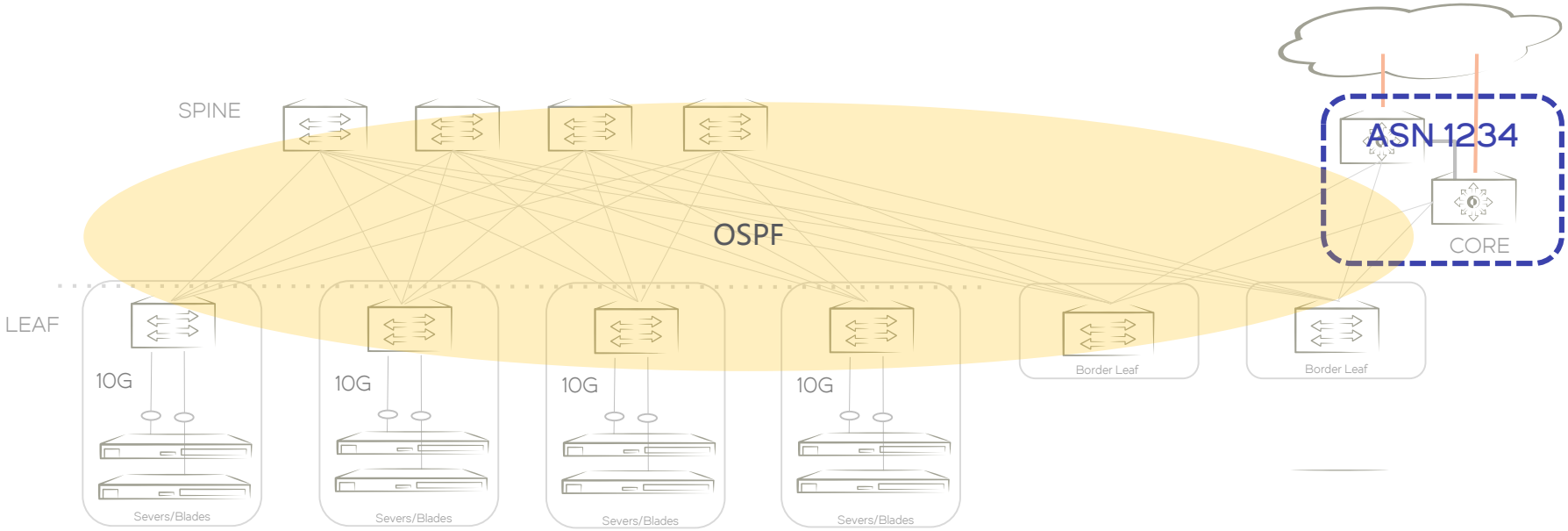
**Edge**

Core or Wan connectivity always through Leaf switch layer

# IP Fabric Underlay

# OSPF Underlay Control Plane

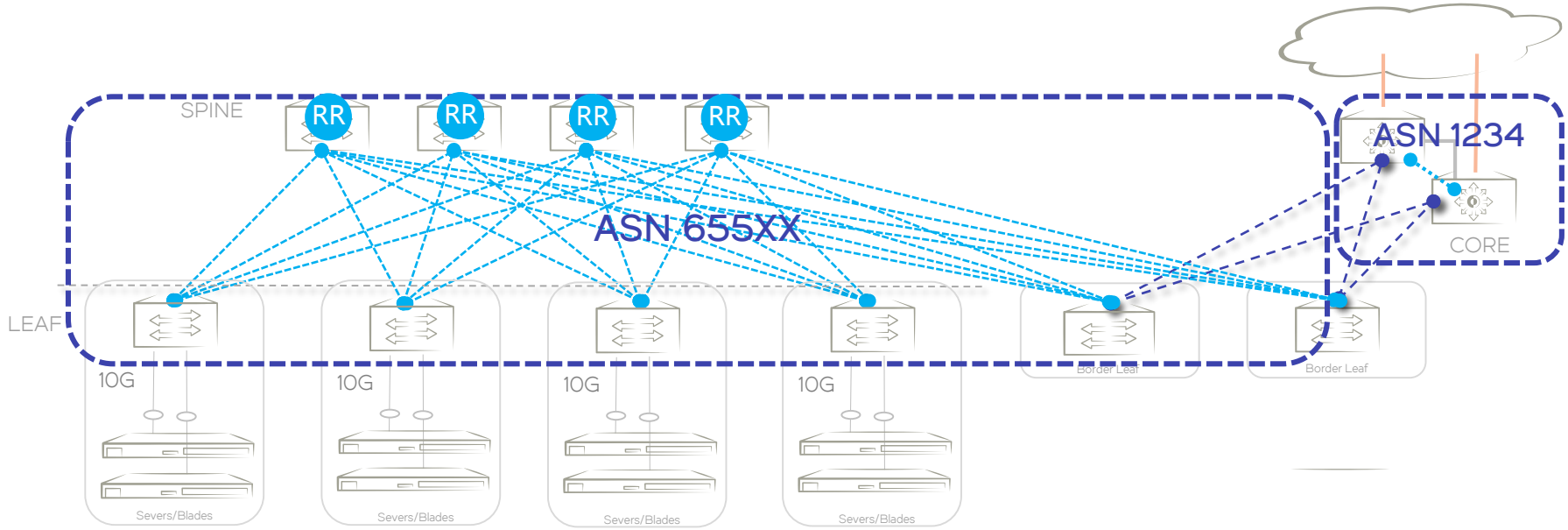
# OSPF Underlay Control Plane




- OSPF across Leaf and Spine
- Single Area or Multiple areas w/ Spine area 0
- Default route into OSPF at Core
- ABR Summary filtering
- Minimal rack level policy control
- Simpler configuration

# iBGP Underlay Control Plane

# iBGP Underlay Control Plane

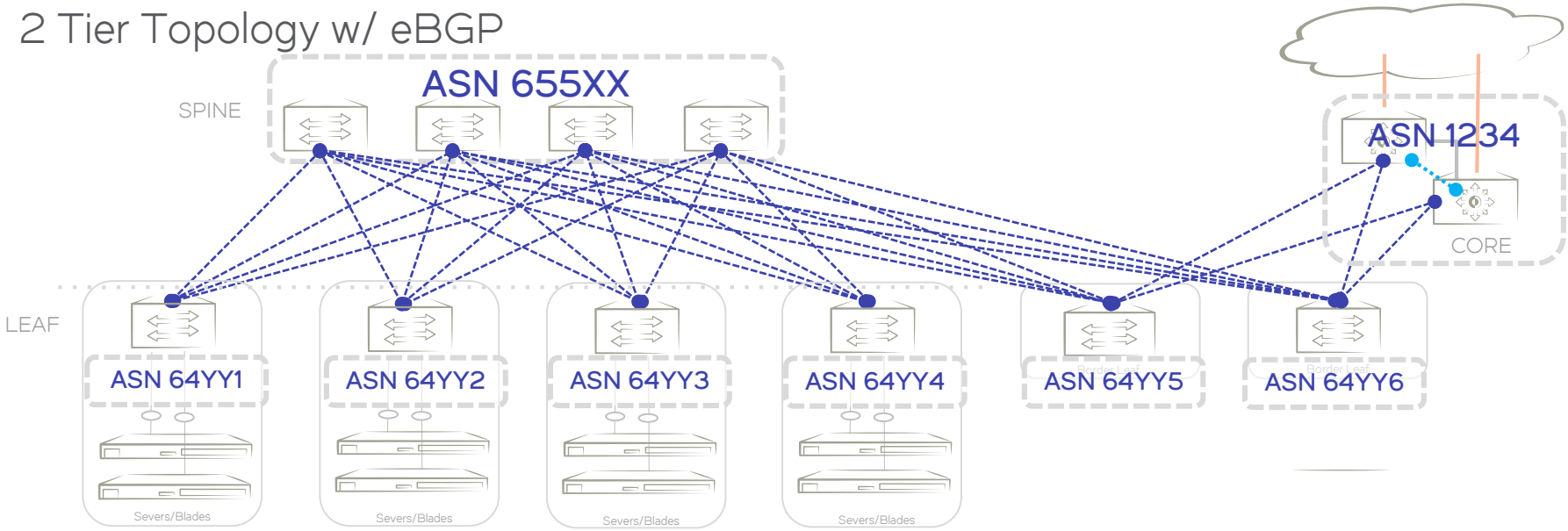



- One ASN across Leaf and Spine
- Each Spine as RR
- No iBGP between Spines
- Default-information originate@ CORE
- BGP **add-path** enabling BGP multi-path updates at RR
- Remove private-as at CORE
- Rack level policies at prefix-level
- RR: bgp next-hop-self

# eBGP Underlay Control Plane
## (Draft Lapukhov)

# eBGP Underlay Control Plane

## 2 Tier Topology w/ eBGP




- Public ASN at Core/Edge
- One Private - ASN per Rack
- One Private - ASN for SPINE
- eBGP session between Leaf-Spine
- IP Address Optimization using /31, /127
- _OR_ IP unnumbered

- Allows BGP policies per TOR
- IPv4/IPv6
- Remove private AS at CORE
- Default-information originate @CORE
- 4 Byte ASN for Scale

# Layer 2 Overlay

# L2 Overlay Choice

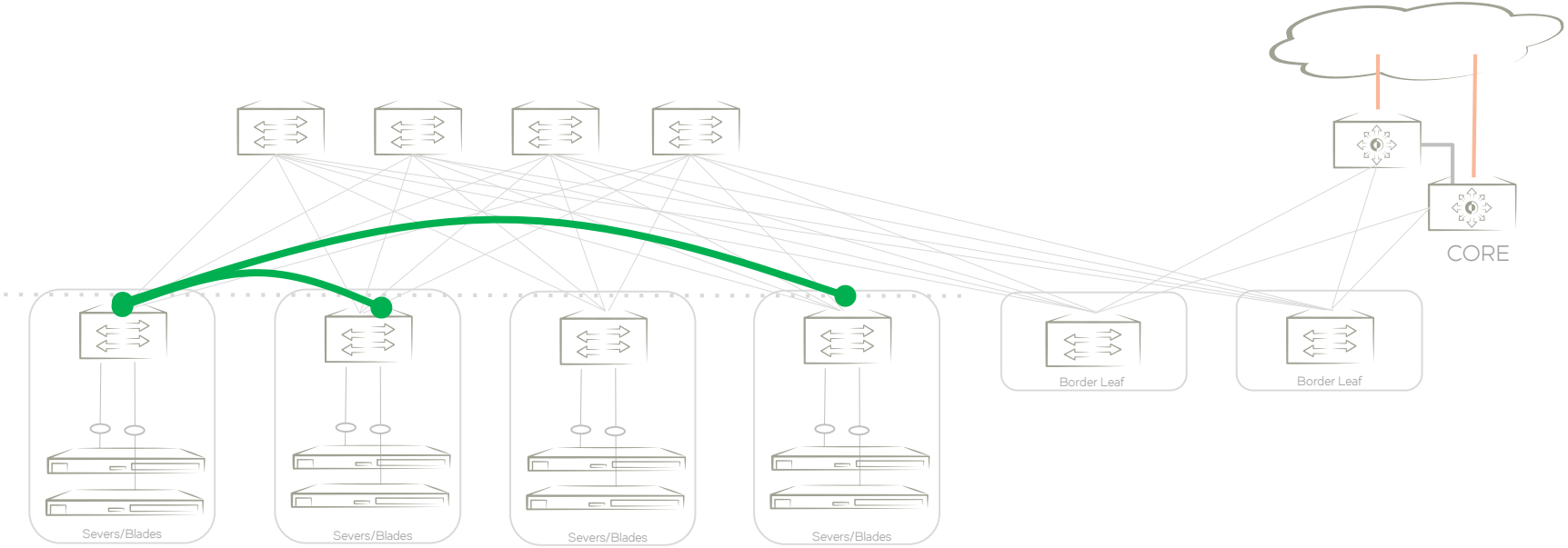
- IP Fabric L2-Extension
- Partner-Specific Network Virtualization
  - NSX
- Standards Based Network Virtualization
  - BGP-EVPN

# Solutions – IP Fabric with L2 Extension

## Powered by Brocade IP Fabric

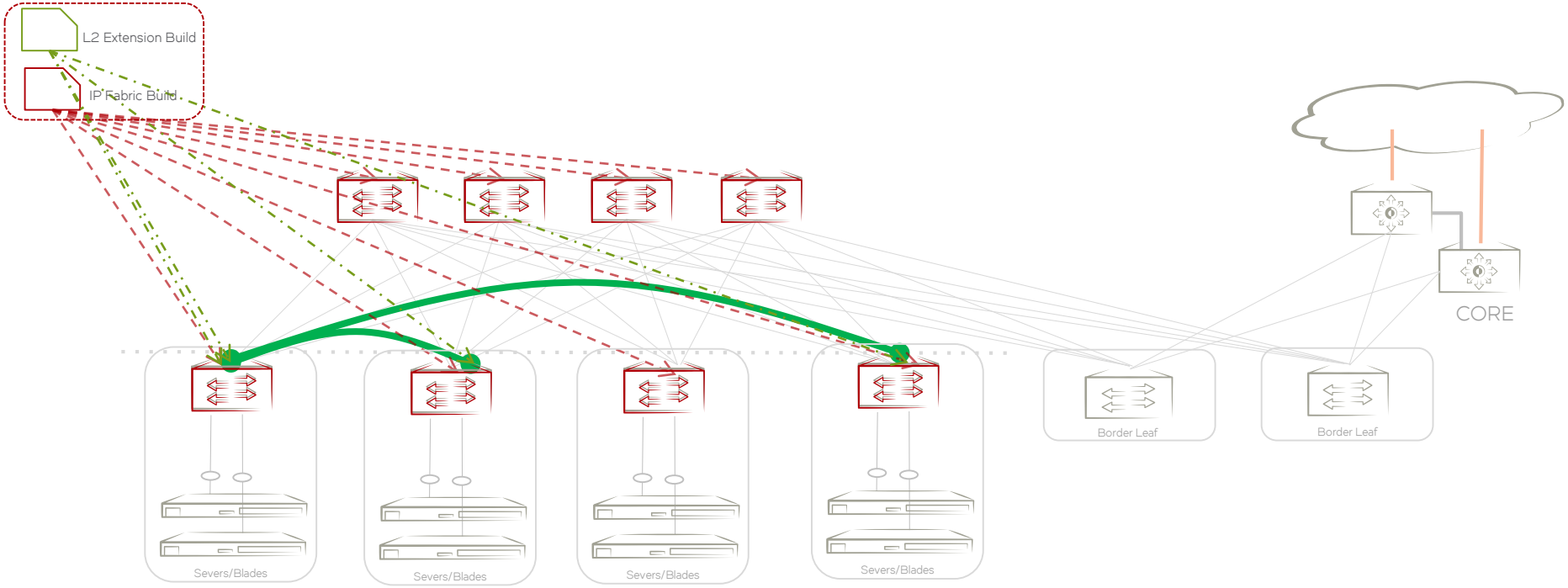
- IP Fabric – L2 Extension

  L2 extension between racks using VxLAN.



# Open Automation Suite in action

## IP Fabric – L2 Extension

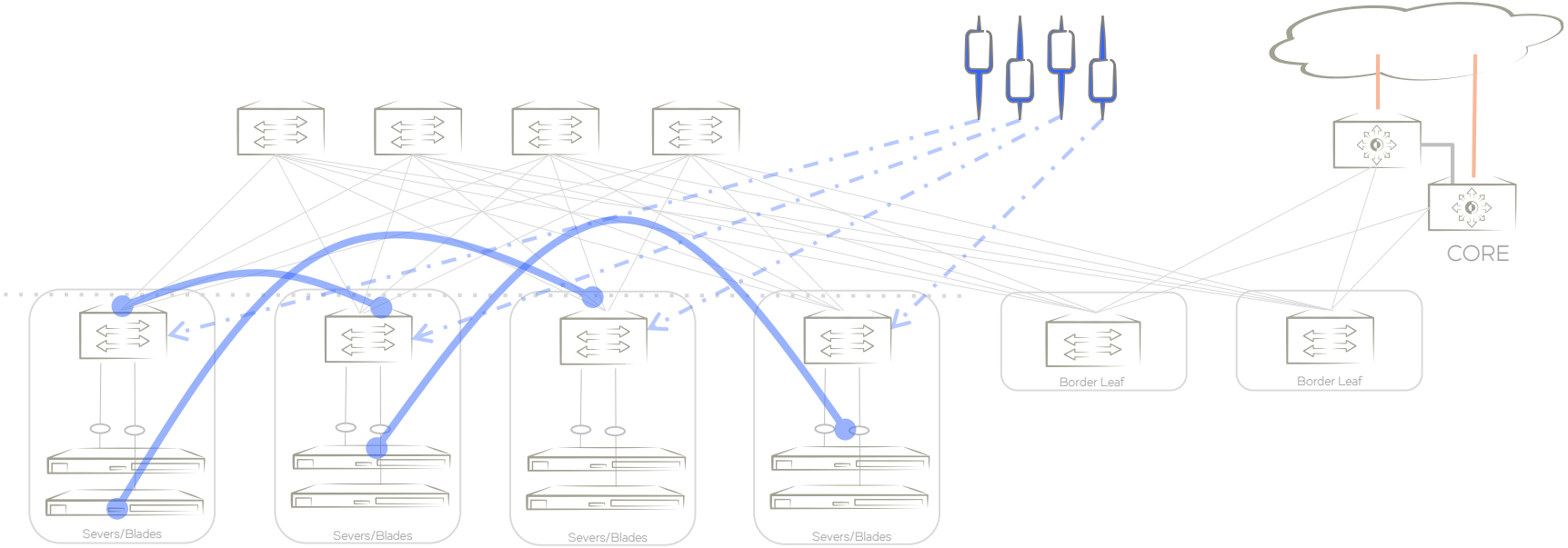


# Solutions – IP Fabric with Controller Based Overlays

## Powered by Brocade IP Fabric

- IP Fabric – Underlay for Controller Based Overlays

  As underlay for controller based overlays



# EVPN Control Plane

## Highlights

- Based on Industry standard (Multi-vendor Interoperability)
  - RFC 7432 https://tools.ietf.org/html/rfc7432 (EVPN - MPLS)
  - RFC 7209 https://tools.ietf.org/html/rfc7209 (Requirements for EVPNs)
  - https://tools.ietf.org/html/draft-ietf-bess-evpn-overlay-01 (EVPN - VXLAN)
- Minimizes BUM Network Flooding through MAC/IP Host/subnet routes and ARP suppression on VTEP's
- Optimal East-West and north-south traffic leveraging underlay ECMP
- Workload mobility with distributed anycast functionality with no changes in workload L3 Configuration
- VTEP peer discovery and Authentication
- Supporting Intra Subnet VxLAN Bridging and L3 Inter Subnet VxLAN Routing
- Multi-tenancy using VxLAN in Data Center using BGP-VRF

© 2015 BROCADE COMMUNICATIONS SYSTEMS, INC. INTERNAL USE ONLY

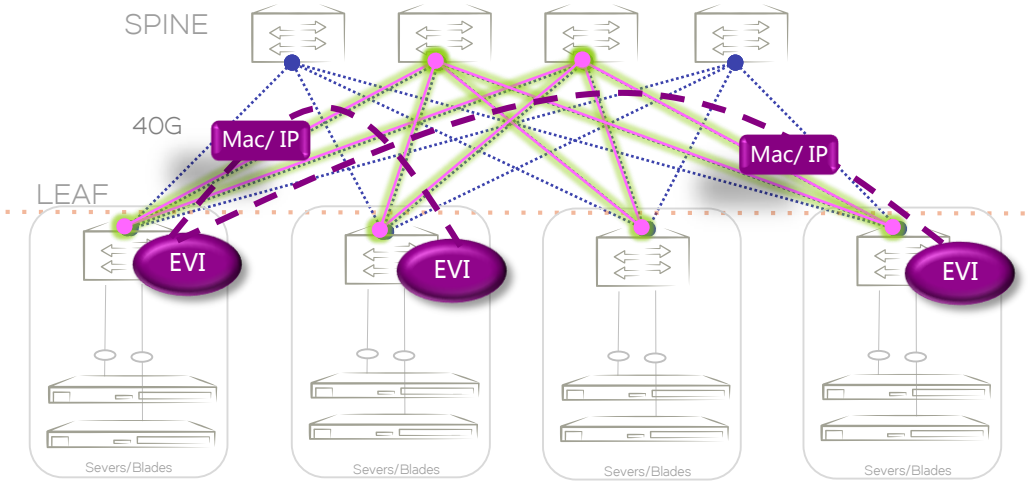
# BGP EVPN Constructs

- EVPN Instance a.k.a. MAC VRF
- VLANs assigned to an EVPN Instance
- RD, Import and Export Route-Target communities configured for the EVPN instance
- Ethernet Segments and Ethernet Segment Identifier
- Route Types
  - ES and ES-AD Routes
  - MAC Routes
  - MAC-IP Routes
  - Inclusive Multicast Routes
  - Prefix Routes
- Spine switches do not need EVPN instance configuration
- Spine switches propagate EVPN routes without “importing” them.

# eBGP Underlay + BGP Overlay(EVPN)

## Controller-less VxLAN Overlay

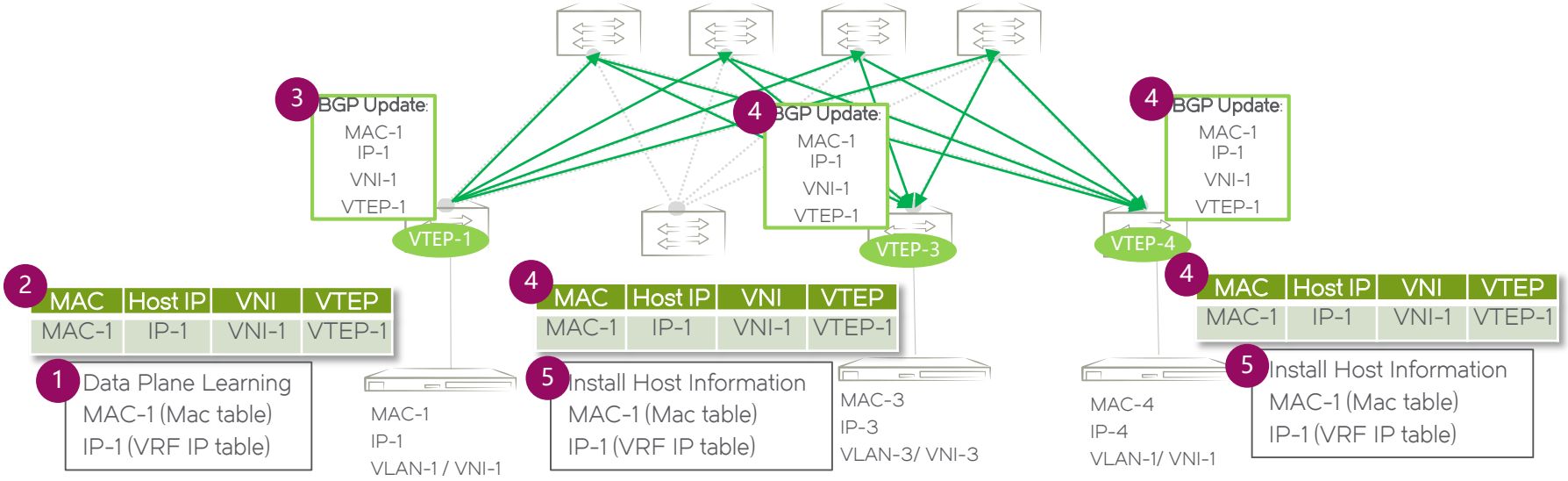
- Auto VTEP Discovery
- L2 VNI & Multi-tenancy
  (Intra Subnet w/ VxLAN)
- **L3 VNI and Multi-tenancy**
  **(Inter** Subnet Routing w/ VxLAN)
- Single Pass in ASIC:
  VxLAN [De] Encapsulation + Routing
- Supports IPv4 & IPv6




eBGP Underlay eBGP Overlay

# EVPN: Host NLRI Learning and Distribution

## Control Plane Signaling.



# Use Case: Intra-Subnet Bridging

## L2 Bridging

VxLAN Encap Packet Egress

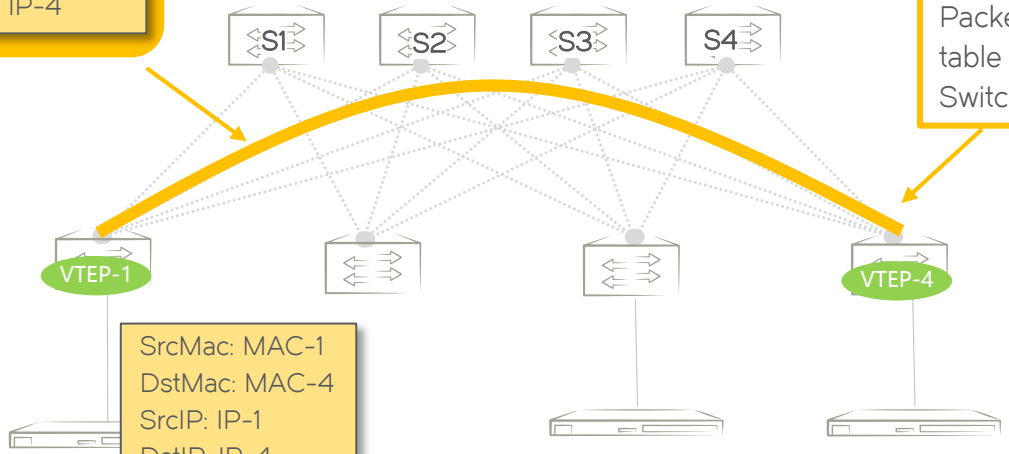
SrcMac VTEP-1
DstMac S1-4
SrcIP **VTEP-1**
DstIP **VTEP-4**
VNI **VNI-1**

SrcMac: MAC-1
DstMac: MAC-4
SrcIP: IP-1
DstIP: IP-4

Ingress VTEP
De-encapsulates the VxLAN Packet, Checks the MAC table for DstMac: **MAC-4** and Switches

Mac Table::
MAC-1 VLAN-1 Port 1/0
**MAC-4 VNI-1 VTEP 4**

Route Table;
**VTEP-4** S1, S2, S3, S4




SrcMac: MAC-1
DstMac: MAC-4
SrcIP: IP-1
DstIP: IP-4

MAC-1
IP-1
VLAN-1 / VNI-1

MAC-4
IP-4
VLAN-1/ VNI-1

Mac Table:
**MAC-1 VNI-1 VTEP 1**
MAC-4 VLAN-1 Port 1/0

Route Table:
**VTEP-1** S1, S2, S3, S4

# Use Case: L2 Bridging & Mult-Tenancy

VxLAN Encap Packet Egress

SrcMac VTEP-1
DstMac S1-4
SrcIP **VTEP-1**
DstIP **VTEP-4**
VNI **VNI-1**

SrcMac: MAC-1
DstMac: MAC-4
SrcIP: IP-1
DstIP: IP-4

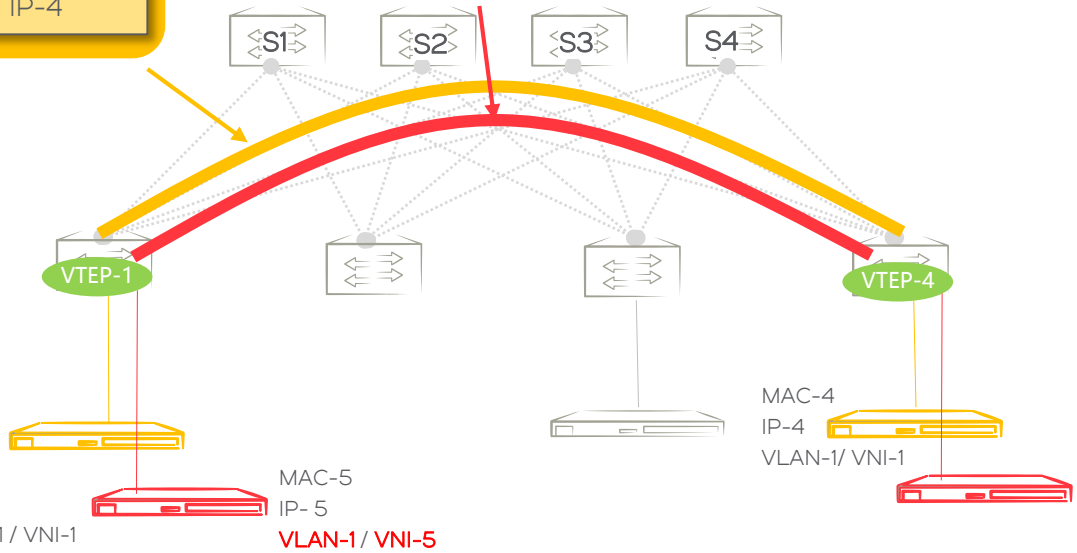
SrcMac VTEP-1
DstMac S1-4
SrcIP **VTEP-1**
DstIP **VTEP-4**
VNI **VNI-5**

SrcMac: MAC-5
DstMac: MAC-6
SrcIP: IP-5
DstIP: IP-6

Overlapping VLAN-IDs possible

Mac Table::
MAC-1 VLAN-1 Port 1/0
MAC-4 VNI-1 VTEP 4
MAC-5 VLAN-1 Port 9/0
MAC- 6 VNI-5 VTEP4

Route Table;
VTEP-4 S1, S2, S3, S4

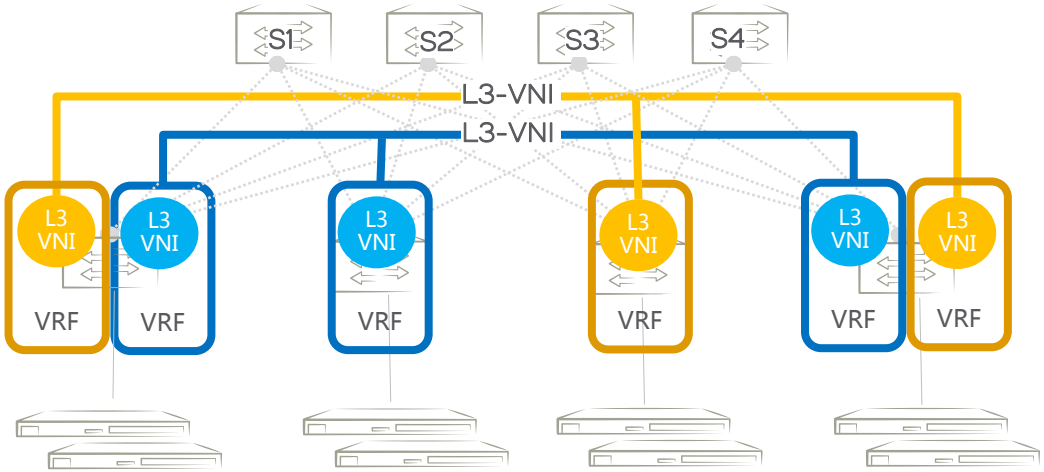



Mac Table:
MAC-1 VNI-1 VTEP 1
MAC-4 VLAN-1 Port 1/0
MAC- 6 VLAN-1 Port 6/0
MAC-5 VNI-5 VTEP-1

Route Table:
VTEP-1 S1, S2, S3, S4

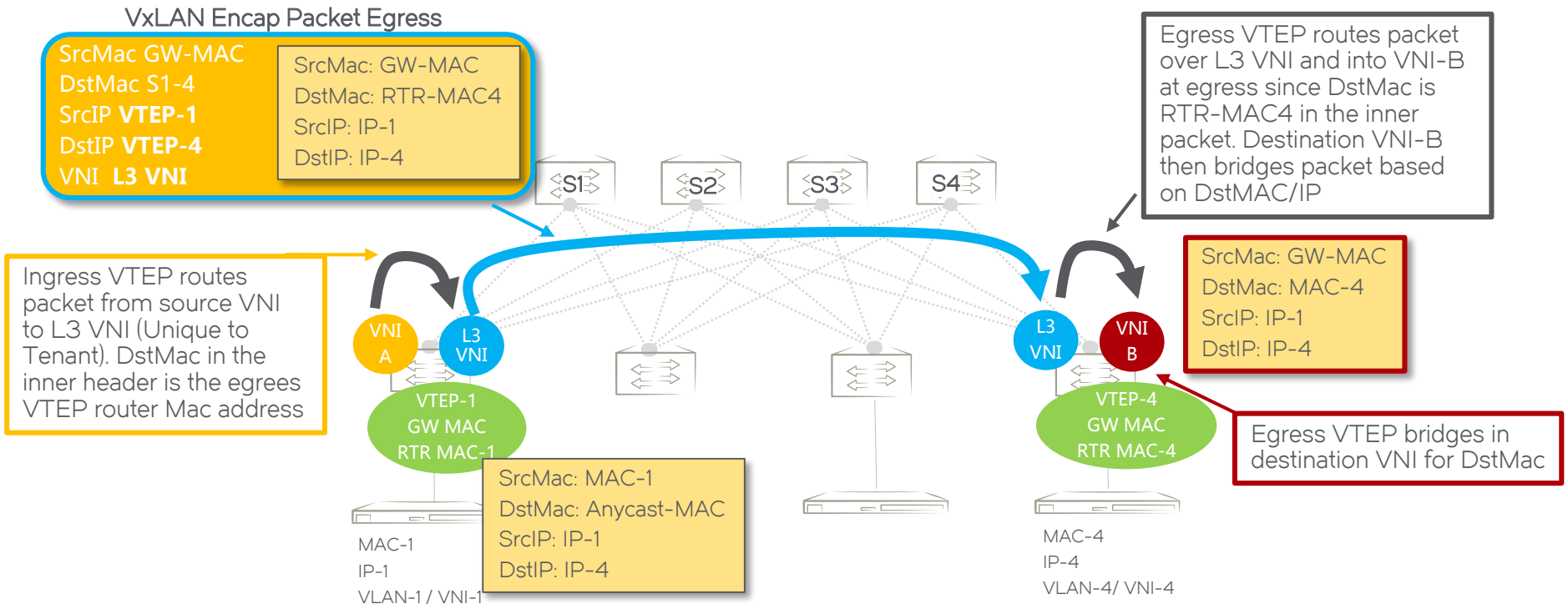
# L3 Multi-Tenancy w/ VxLAN

- VxLAN Based L3 Multi-Tenancy
- VRF + L3 VNI
- Standards based Interop
- Single Protocol Instance for multiple VRF.
- No MPLS complexity
- RT/RD Import Export Policies supported
- Scale TBD Tenants/TOR



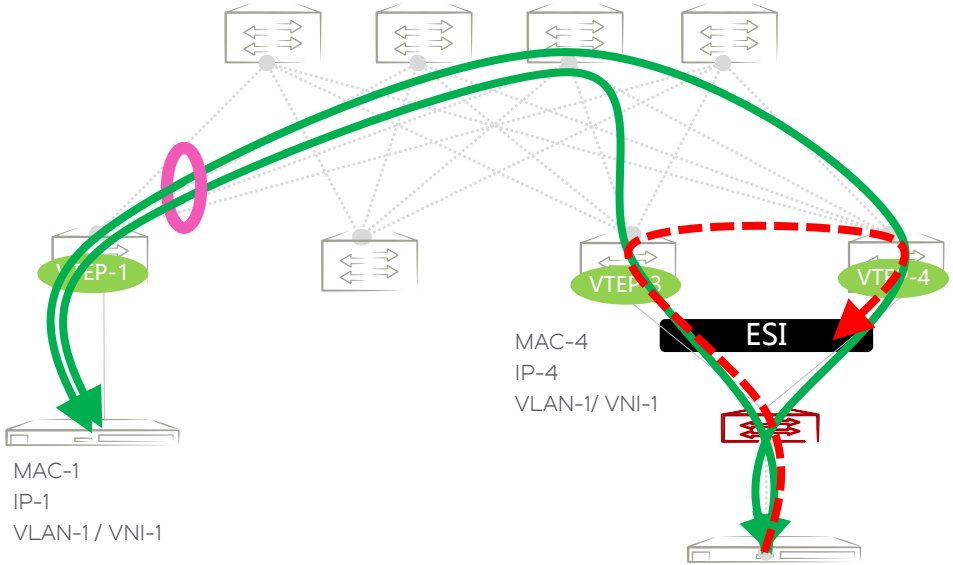
# Use Case: Inter-Subnet Routing

## Symmetric Integrated Routing and Bridging



# Redundant connections into EVPN - simplified

## Multi-homing / Ethernet Segment Identifier - ESI




- Configure / auto derive an ESI
- Autodiscover PEs attached to ES
- Elect designated forwarder (per VLAN)
- Split-horizon (impacts data plane)
  - MPLS: ESI-label
  - VXLAN: Track source IP-addresses and filter
- Aliasing: Associate remote MAC with all PEs announcing the same ES

# ARP/ND Suppression

- ARP/ND Suppression enabled on a per VLAN.
- BGP EVPN derived ARP/ND Entries installed in the ARP/ND Suppression Cache.
- Leaf switches intercept ARP/ND requests from servers
- Generate ARP/ND response if destination IP found in ARP/ND suppression cache
- Flood over VxLAN tunnels only if not found

© 2015 BROCADE COMMUNICATIONS SYSTEMS, INC. INTERNAL USE ONLY

# Conversational ARP/ND

- More hosts in IP Fabric than the leaf switch can hold in HW
- BGP EVPN Control plane capable of learning lot more MAC-IP (ARP/ND) routes than HW capacity
- Leaf switch should install only those host routes in HW for which there are active flows
- If Host Route is available and HW space is available, host route is programmed in HW
- Age out ARP/ND host entries from HW on inactivity

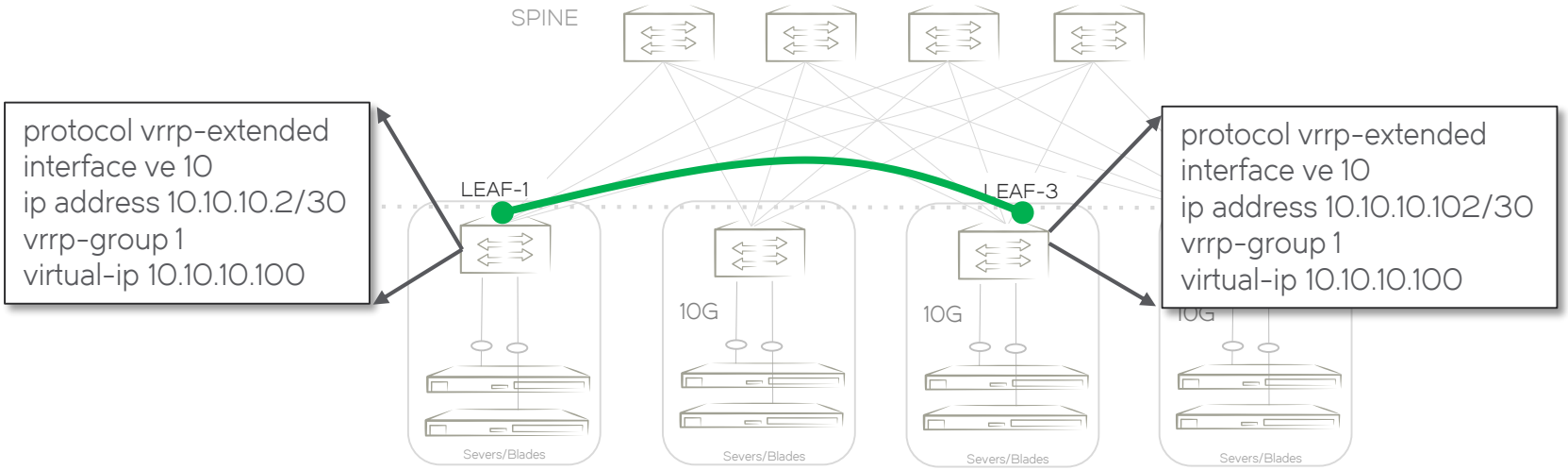
© 2015 BROCADE COMMUNICATIONS SYSTEMS, INC. INTERNAL USE ONLY

# Layer 3 Gateway

© 2015 BROCADE COMMUNICATIONS SYSTEMS, INC. INTERNAL USE ONLY

# L3 Gateway

## VRRP-E

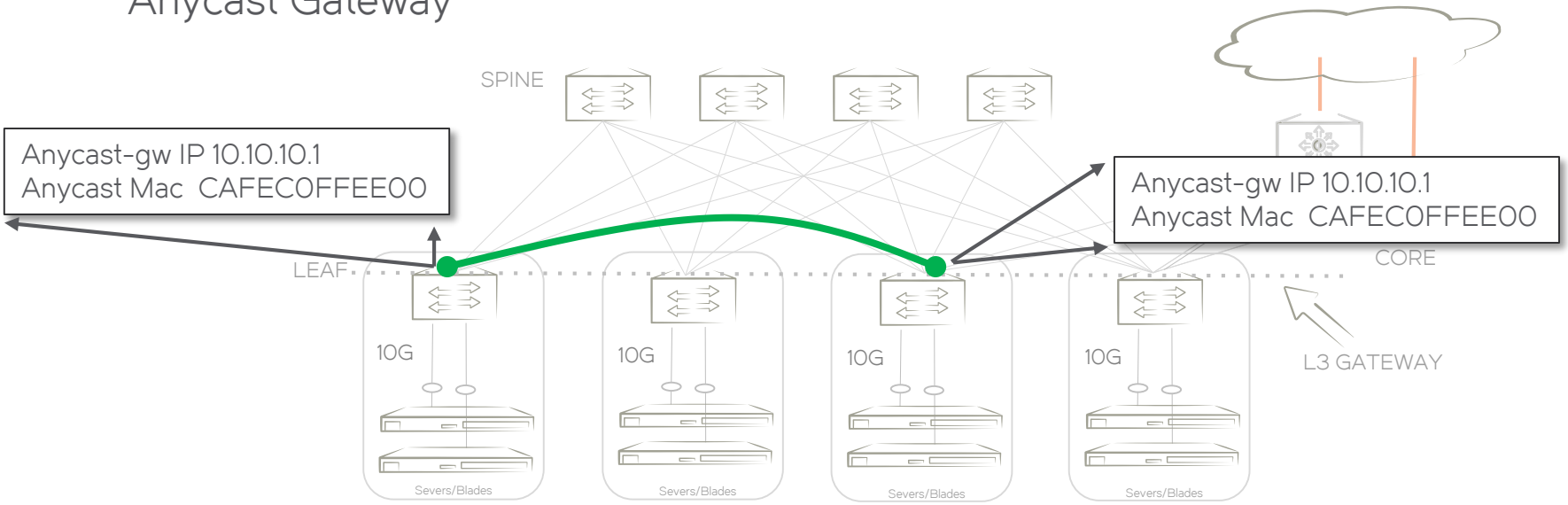



- Same IP Subnet across Leafs
- Same VRRP-E Virtual IP on Leaf switches across VxLAN Tunnel
- VRRP-E Short Path forwarding at Leaf.

© 2015 BROCADE COMMUNICATIONS SYSTEMS, INC. INTERNAL USE ONLY

# L3 Gateway Redundancy

## Anycast Gateway

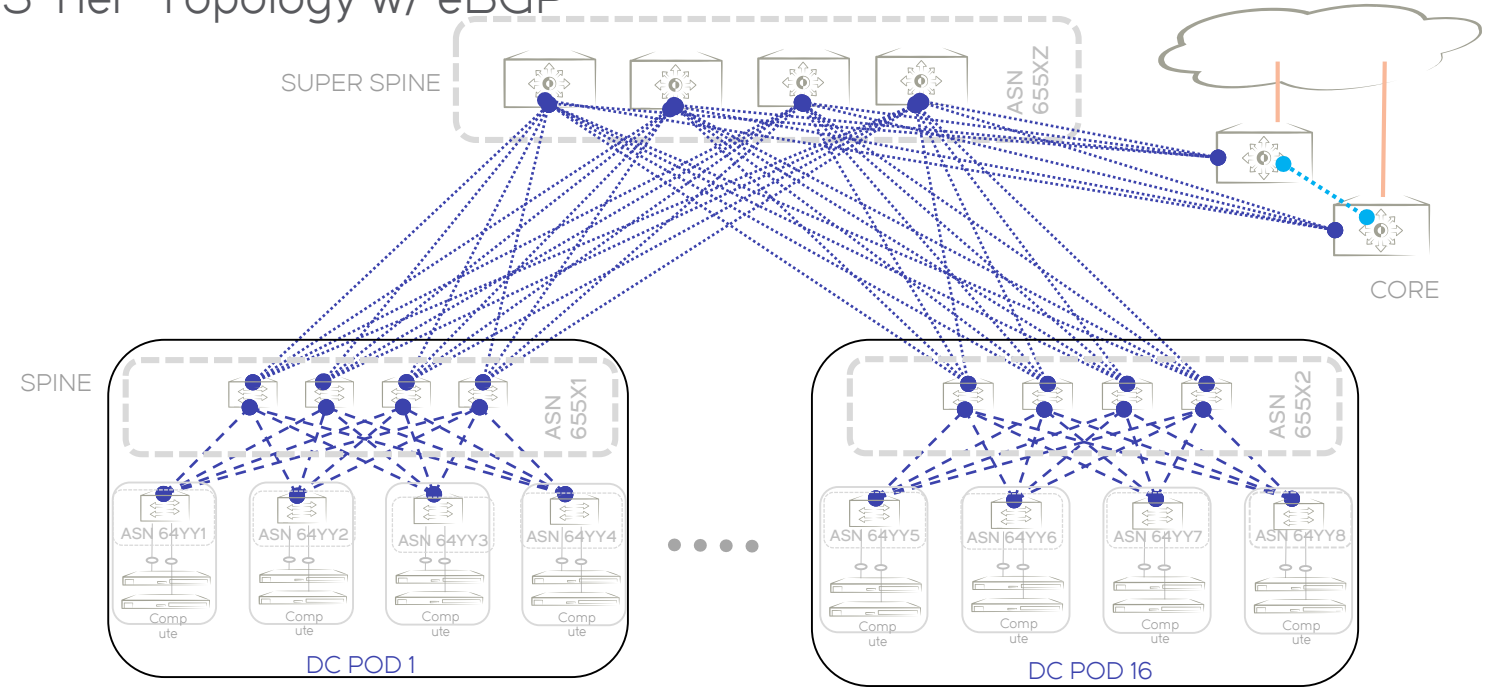



- Anycast Gateway IP is configured on the leaf switches (just as Virtual IP)
- Anycast Gateway MAC address configured to be the same on all leaf switches`

© 2015 BROCADE COMMUNICATIONS SYSTEMS, INC. INTERNAL USE ONLY

# Underlay Control Plane

## 3 Tier Topology w/ eBGP




### Spine

eBGP Peering across Spine and Super Spine

ECMP Multipath

### Super Spine

CORE directly connected to Super Spine

Default-information originate@CORE to Super Spine

Reference: Facebook Data Center Architecture

# Summary

- Many DC architectures exist – there is no „one size fits all“
- IP fabrics are one robust alternative
- IP fabrics definitely need automation for provisioning and operations
- EVPNs are one way to provide L2-overlay networks

© 2015 BROCADE COMMUNICATIONS SYSTEMS, INC. INTERNAL USE ONLY

BROCADE

# Thank you

© 2015 BROCADE COMMUNICATIONS SYSTEMS, INC. INTERNAL USE ONLY