JUNIPER | Engineering
NETWORKS | Simplicity

EVPN LAG Multihoming in EVPN-VXLAN Cloud Data Center Infrastructures

Juniper Networks, Inc.
1133 Innovation Way
Sunnyvale, California 94089
USA
408-745-2000
www.juniper.net

YEAR 2000 NOTICE

END USER LICENSE AGREEMENT

Table of Contents

About This Guide

Use this guide to learn more about EVPN LAG Multihoming implementation options and features in EVPN-VXLAN Cloud Data Center Infrastructures.

1

CHAPTER

EVPN LAG Multihoming in EVPN-VXLAN Cloud Data Center Infrastructures

About This Guide

The purpose of this document is to provide networking professionals with concepts and tools related to EVPN LAG multihoming in EVPN-VXLAN cloud data center architectures.

The intended audience for this guide includes system integrators, infrastructure professionals, partners, and customers that want to utilize or optimize the use of EVPN LAG multihoming, also known as ESI-LAG multihoming, in their cloud data centers.

Introduction to EVPN LAG Multihoming

Ethernet link aggregation remains a vital technology in modern data center networking. Link aggregation provides the technology to bundle interfaces and to increase bandwidth in a data center by adding new links to bundled interfaces. Link aggregation can significantly increase a network's high availability by providing a redundant path or paths that traffic can utilize in the event of a link failure and has a variety of other applications that remain highly relevant in modern data center networks.

New EVPN technology standards—including RFCs 8365, 7432, and 7348—introduce the concept of link aggregation in EVPNs through the use of Ethernet segments. Ethernet segments in an EVPN collect links into a bundle and assign a number—called the Ethernet segment ID (ESI)—to the bundled links. Links from multiple standalone nodes can be assigned into the same ESI, introducing an important link aggregation feature that introduces node level redundancy to devices in an EVPN-VXLAN network. The bundled links that are numbered with an ESI are often referred to as ESI LAGs or EVPN LAGs. The EVPN LAGs term will be used to refer to these link bundles for the remainder of this document.

Layer 2 Multihoming in EVPN networks is dependent on EVPN LAGs. EVPN LAGs, which provide full active-active link support, are also frequently enabled with LACP to ensure multivendor support for devices accessing the data center. Layer 2 Multihoming with LACP is an especially attractive configuration option when deploying servers because the multihoming is transparent from the server point of view; the server thinks it is connected to a single networking device when it is connected to two or more switches.

shows a fabric topology representing a standard 3 stage spine-leaf underlay architecture. An EVPN-VXLAN overlay offers link and node level redundancy and an active-active Layer 2 Ethernet extension between the leaf devices, without the need to implement a Spanning Tree Protocol (STP) to avoid loops. Different workloads in the data center are integrated into the fabric without any prerequisites at the storage system level. This setup would typically use QFX5110 or QFX5120 switches as leaf devices and QFX10002 or QFX10008 switches as spine devices.

Figure 1: Data Center Example Topology with EVPN LAGs
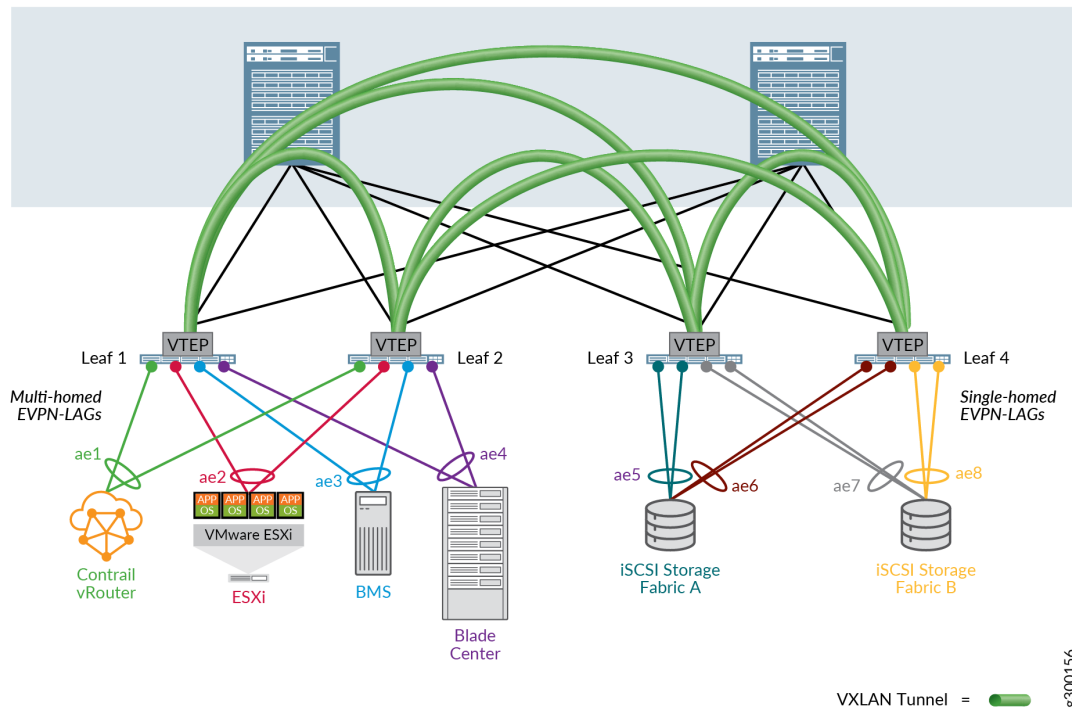
> **NOTE**: Single-homed end systems are still frequently used in modern data center networks. The iSCSi storage array systems in the figure provide an example of single-homed end systems in an EVPN-VXLAN data center.

See the EVPN Multihoming Implementation section of the EVPN Multihoming Overview document for additional information related to ESIs and ESI numbering. See Multihoming an Ethernet-Connected End System Design and Implementation in the Cloud Data Center Architecture Guide for a step-by-step configuration procedure of an EVPN LAG implementation.

Ethernet Segment Identifiers, ESI Types, and LACP in EVPN LAGs

This section discusses Ethernet Segment Identifiers (ESIs), ESI Types, and LACP in EVPN LAGs.

Ethernet Segment Identifiers and Numbering in EVPN LAGs

An EVPN LAG is identified using an Ethernet segment identifier (ESI). An ESI is a mandatory attribute that is required to enable EVPN LAG server multihoming.

ESI values are encoded as 10-byte integers and are used to identify a multihomed segment. The same ESI value enabled on multiple leaf switch interfaces connected to the same server or BladeCenter form an EVPN LAG. This EVPN LAG supports active-active multihoming towards the connected server.

We recommend using an ESI value that uses the same values for the first 8 bytes with changes only in the 9th and 10th bytes on a per-EVPN LAG basis. The four ESIs used later in this document follow these assignment guidelines and use the following ESI values: 00:03:03:03:03:03:03:03:03:01, 00:03:03:03:03:03:03:03:03:02, 00:03:03:03:03:03:03:03:03:03, and 00:03:03:03:03:03:03:03:03:03:04. Using a structured ESI assignment method simplifies network administration while also ensuring the ESI values can be used across Junos OS releases. The need to be compatible across Junos releases is required in some environments due to changes made to ES import route community handling in Junos OS Release 17.3R3.

See the EVPN Multihoming Implementation section of the EVPN Multihoming Overview document for additional information related to ESIs and ESI numbering.

ESI Types in EVPN LAGs

Junos OS currently supports ESI type 0 (manually hard-coded ESI value), type 1 (auto derived from LACP), and type 5 (IRB-VGA gateway ESI encoded based on autonomous system number) ESI patterns. See the EVPN Multihoming Implementation section of the EVPN Multihoming Overview document for additional information related to ESI types.

EVPN multihoming is managed at the control plane level using EVPN Type 4 routes and the ES-Import Route-Target extended community, which in Junos is an automatically derived 6-byte value taken from the 10-byte ESI value. This extended community is used within the Type-4 ES-Route to signal using EVPN BGP messages when connecting leaf devices to the same multihomed CE device.

LACP in EVPN LAGs

External devices such as servers almost always require an LACP state that matches the ESI-enabled interface on the leaf switches to ensure the server can actively send traffic across all member interfaces in the LAG bundle. Specifically, the same ESI and LACP system identifier should be configured on all links that make up a LAG bundle. To ensure the suggested system ID consistency, its a best-practice to base the system ID on the ESI you assign to each LAG.

For example, if you assign ESI 00:03:03:03:03:03:03:03:03:01 to a LAG, then it is suggested that you derive the corresponding system ID from the last 6 bytes of that ESI, i.e., "03:03:03:03:03:01". The suggested best practice of deriving the system ID from the LAG's ESI is demonstrated in the configurations presented later in this document.

EVPN LAGs in EVPN-VXLAN Reference Architectures

IN THIS SECTION

This section provides an overview of the Juniper EVPN-VXLAN reference architectures and the role of EVPN LAGs in these architectures. It is intended as a resource to help readers understand EVPN LAG capabilities in different contexts.

The standard EVPN-VXLAN architecture consists of a 3-stage spine-leaf architecture. The physical underlay is IP forwarding enabled—all leaf to spine underlay links are typically IPv4 routed—and the logical overlay layer uses MP-BGP with EVPN signaling for control plane-based MAC-IP address learning and to establish VXLAN tunnels between switches.

Juniper Networks has four primary data center architectures:

- Centrally Routed Bridging (CRB)—inter-VNI routing occurs on the spine switches.

- Edge Routed Bridging (ERB)—inter-VNI routing occurs on the leaf switches.

- Bridged Overlay—inter-VLAN and inter-VNI routing occurs outside of the EVPN-VXLAN fabric. Example: Routing occurs at the firewall cluster connected to the EVPN-VXLAN.

- Centrally-Routed Bridging Mutual (CRB-M)—architecture where the spine switches are also connecting the existing data center infrastructure with the EVPN LAG. CRB-M architectures are often used during data center migrations.

EVPN LAGs in Centrally Routed Bridging Architectures

In the CRB architecture, we recommend provisioning the EVPN LAGs at the leaf layer and connecting two or more leaf devices to each server or BladeCenter.

Figure 2 on page 7 illustrates EVPN LAG provisioning in a CRB architecture.

Figure 2: EVPN LAGs in a CRB Architecture

ESI	LACP system-id
● ae1 00:03:03:03:03:03:03:03:03:01	00:03:03:03:03:01
● ae2 00:03:03:03:03:03:03:03:03:02	00:03:03:03:03:02
● ae3 00:03:03:03:03:03:03:03:03:03	00:03:03:03:03:03
● ae4 00:03:03:03:03:03:03:03:03:04	00:03:03:03:03:04

IRBs = ● ●

VXLAN Tunnel = ▬

g300152
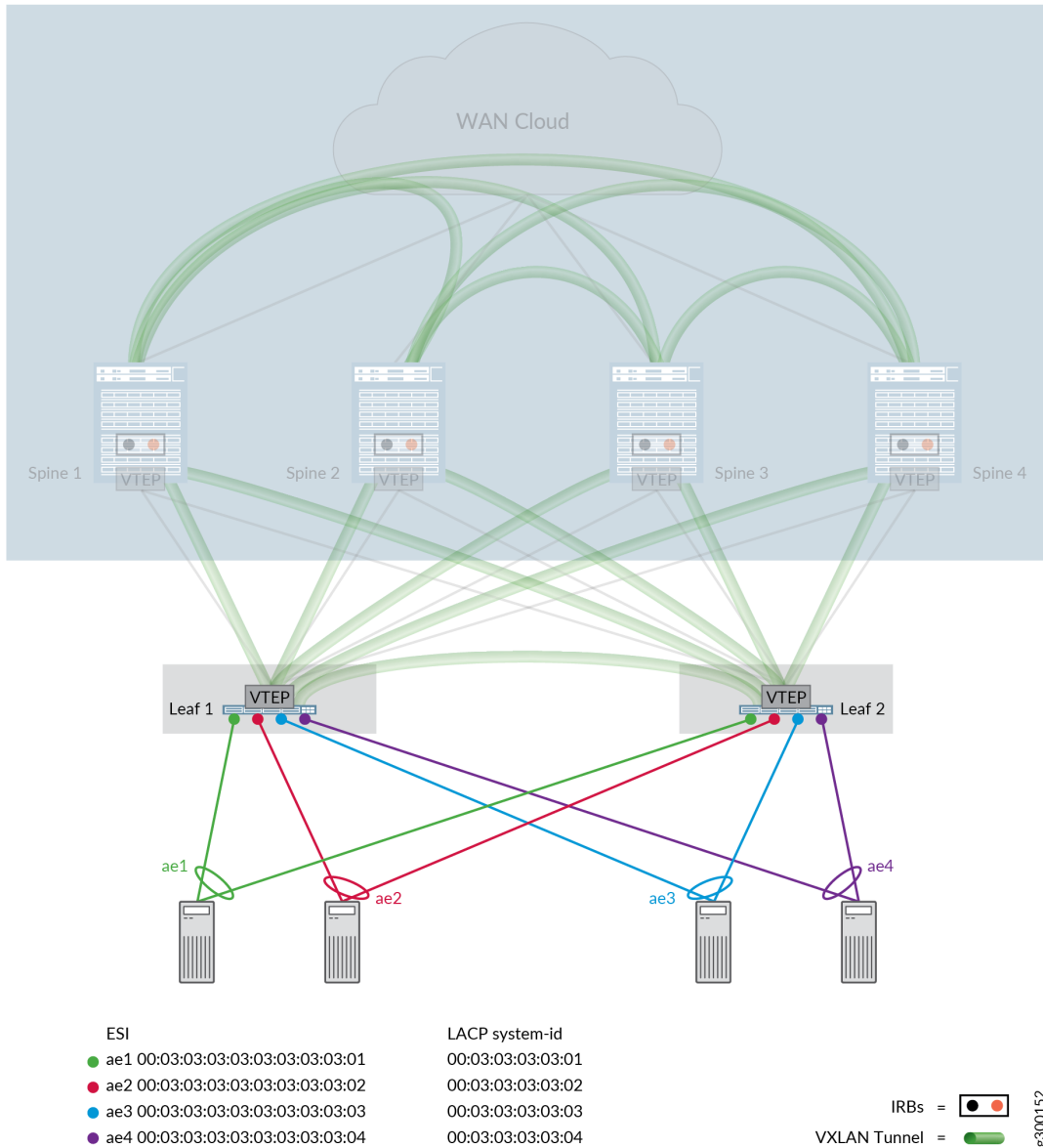
BEST PRACTICE: The same ESI value and LACP system ID should be used when connecting multiple leaf devices to the same server. Unique ESI values and LACP system IDs should be used per EVPN LAG.

EVPN LAGs in Edge Routed Bridging Architectures

Figure 3 on page 9 illustrates the use of EVPN LAGs within an Edge Routed Bridging (ERB) architecture. The recommended EVPN LAG provisioning in an ERB architecture is similar to the CRB architecture. The major difference between the architectures is that the IP first hop gateway capability is moved to the leaf level using IRB interfaces with anycast addressing.

The ERB architecture offers ARP suppression capability complemented by the advertisement of the most specific host /32 Type-5 EVPN routes from leaf devices toward the spine devices. This technology combination efficiently reduces data center traffic flooding and creates a topology that is often utilized to support Virtual Machine Traffic Optimization (VMTO) capabilities.

Figure 3: EVPN LAGs in ERB Architecture
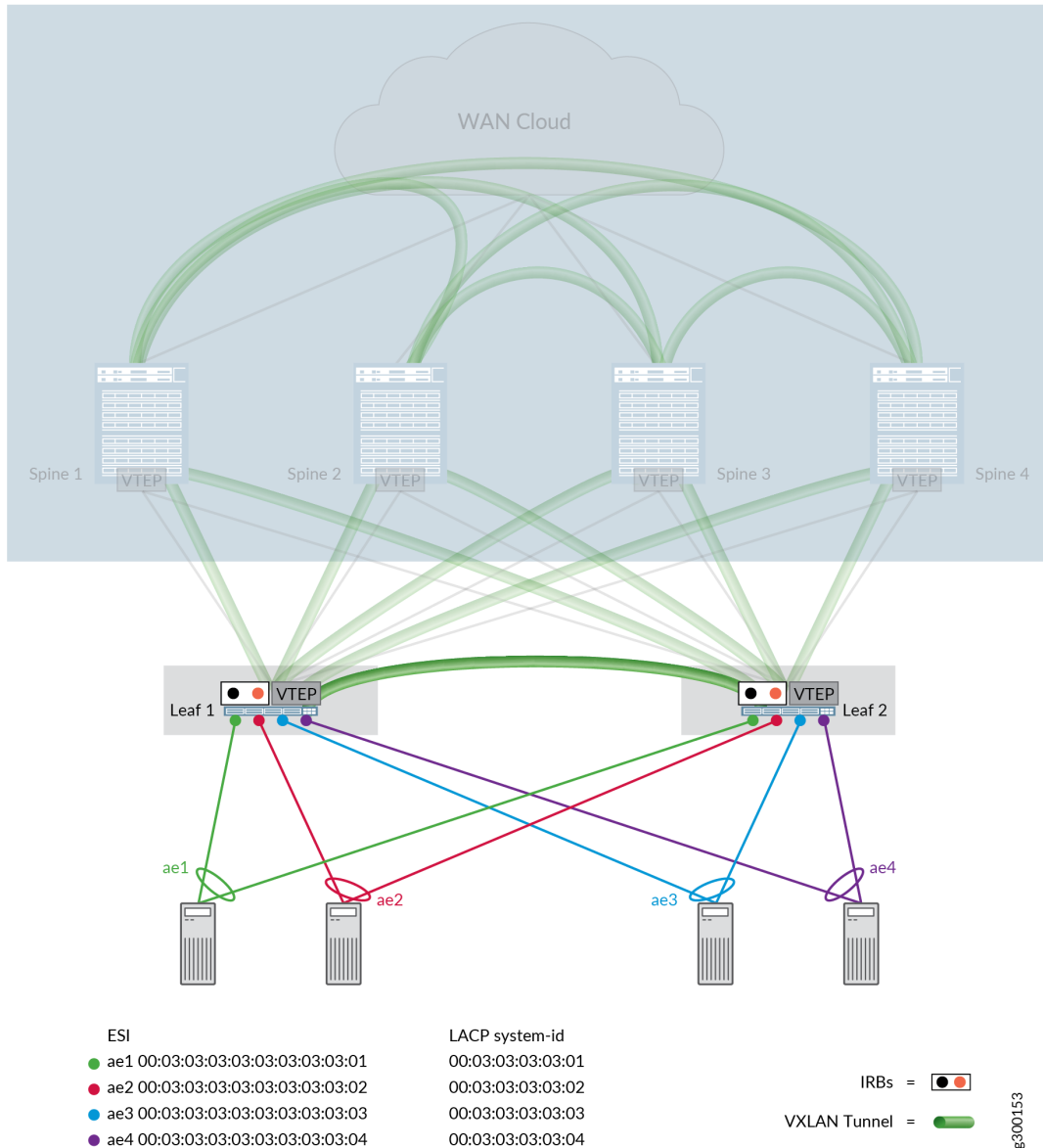
EVPN LAGs in Bridged Overlay Architectures

In a bridged overlay architecture, VLANs are extended between leaf devices across VXLAN tunnels. EVPN LAGs are used in a bridged overlay to provide multihoming for servers and to connect to first hop gateways outside of the EVPN-VXLAN fabric, which are typically SRX Series Services Gateways or MX Series routers. The bridged overlay architecture helps conserve bandwidth on the gateway devices and

increases the bandwidth and resiliency of servers and BladeCenters by delivering active-active forwarding in the same broadcast domain.

Figure 4 on page 10 illustrates EVPN LAGs in a sample bridged overlay architecture.

Figure 4: EVPN LAGs in Bridged Overlay Architecture
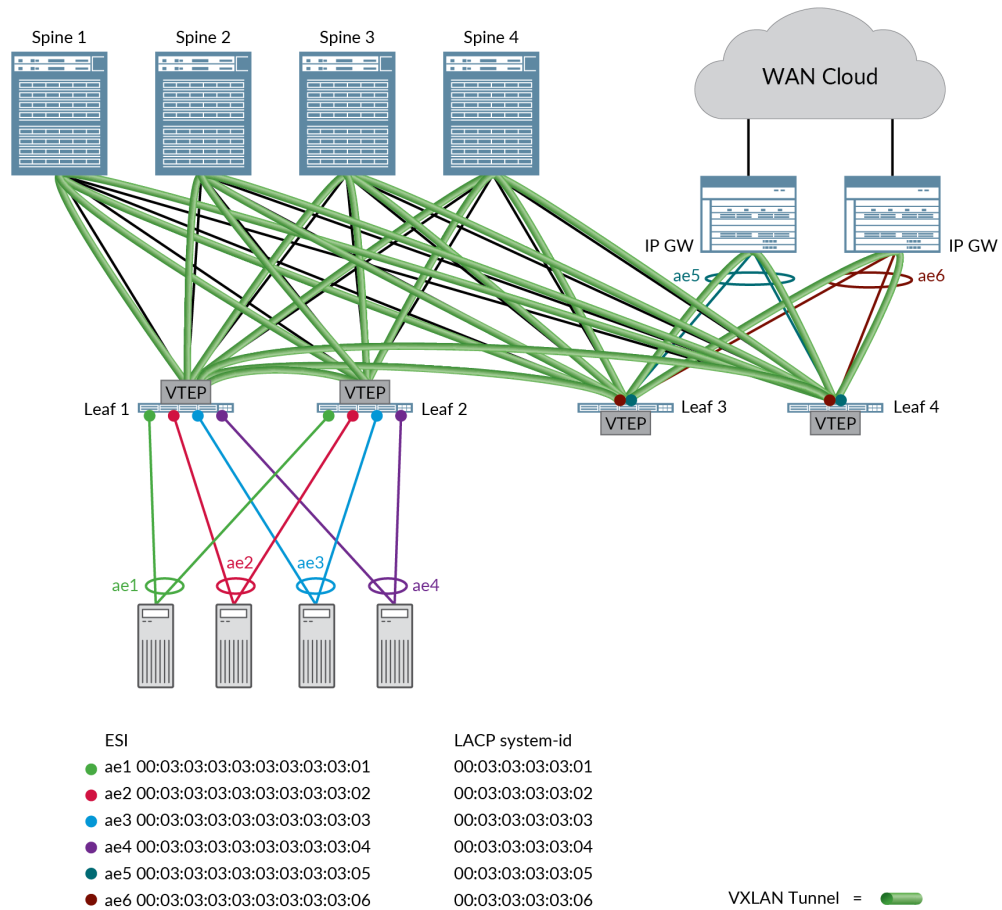
EVPN LAGs in Centrally Routed Bridging Migration Architectures

EVPN LAGs might be introduced between spine and leaf devices during a migration to one of the aforementioned EVPN-VXLAN reference architectures. This EVPN LAG is needed in some migration scenarios to integrate the existing legacy ToR-based infrastructure to the EVPN-VXLAN architecture.

Figure 5 on page 11 shows a Virtual Chassis and an MC-LAG architecture connected to spine devices using an EVPN LAG. The EVPN LAG provisioning is done from the spine devices during the migration of these topologies into an EVPN-VXLAN reference architecture.

Figure 5: EVPN LAGs in CRB Migration Architectures
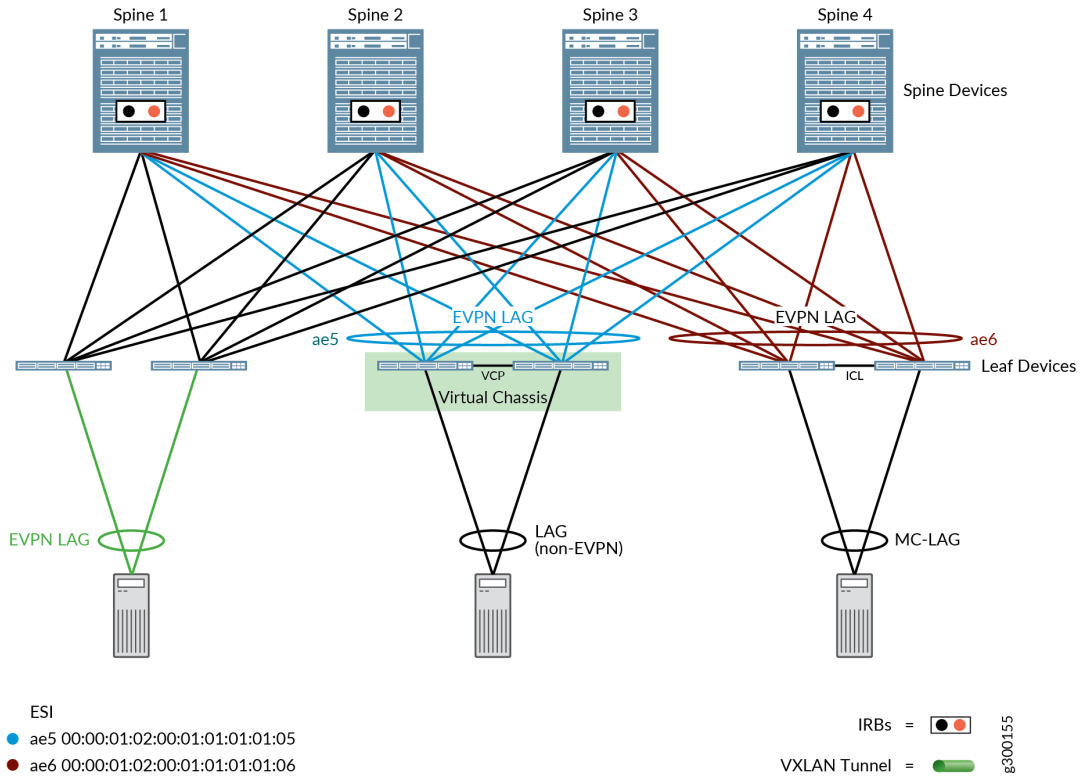
The CRB Migration architecture is often used when migrating an MC-LAG or Virtual Chassis-based data center in phases. In this architecture, the EVPN LAG capability is introduced at the spine level and only one overlay iBGP session is running between the two spine switches. The top-of-rack switches connected to the spine devices are legacy switches configured as Virtual Chassis or MC-LAG clusters with no EVPN iBGP peerings to the spine switches.

This architecture helps when deploying EVPN-VXLAN technologies in stages into an existing data center. The first step is building an EVPN LAG-capable spine layer, and then sequentially migrating to an EVPN control plan where MAC addresses for the new leaf switches are learned from the spine layer switches. The new leaf switches, therefore, can benefit from the advanced EVPN features, such as ARP suppression, IGMP suppression, and optimized multicast, supported by the new switches.

The default EVPN core isolation behavior should be disabled in CRB Migration architectures. The default EVPN core-isolation behavior disables local EVPN LAG members if the network loses the last iBGP-EVPN signaled peer. Because this peering between the two spine devices will be lost during the migration, the default behavior-which can be changed by entering the `no-core-isolation` option in the `edit protocols evpn` hierarchy—must be changed to prevent core isolation events.

EVPN Features in EVPNs using EVPN LAGs

This section introduces some commonly-used features in EVPNs that are using EVPN LAGs.

EVPN Aliasing

EVPN aliasing is the ability of a remote device to load balance Layer 2 unicast traffic across an Ethernet segment towards an endpoint device. See the Aliasing section of the EVPN Multihoming Overview topic for additional information on aliasing.

Servers and BladeCenter switches connected to an EVPN LAG with multiple links may send ARP requests to one of the leaf devices. The leaf devices respond by advertising the MAC address to the rest of the EVPN topology using the overlay iBGP peerings. The other leaf devices in the EVPN network use the default EVPN aliasing (EVPN Route type 1 per EVI) functionality to learn the EVPN Type 2 MAC routes from other leaf devices. All of the leaf devices are connected to the same ESI, however, for forwarding purposes.

EVPN LAG Multihomed Proxy Advertisement

A leaf device in a network running EVPN responds to a server's ARP request with a MAC and IP EVPN Type 2 route.

Starting in Junos OS Release 18.4, multiple leaf devices that have links in the same EVPN LAG have the capability of also advertising MAC and IP EVPN Type 2 routes to the server that sent the ARP request. This capability is possible using the proxy M-bit in EVPN Type 2 routes. This capability is especially beneficial in failure scenarios. If a switch is multihomed to two leaf devices and the link to one of those

devices fails, a Type 2 message can be sent toward the server that initially sent the ARP request and traffic can be sent across the network over the active links.

High Availability in EVPN LAGs

EVPN LAGs include many built-in high availability capabilities.

EVPN LAG link level fast convergence is delivered using the EVPN IBGP route type I massive withdrawal message. Node level fast convergence is handled using the BFD for IBGP in the overlay network. Node-level Layer 2 EVPN LAG redundancy is available in an EVPN-VXLAN fabric using built-in EVPN loop prevention mechanisms like split horizon and designated forwarding. Spanning tree protocol (STP) does not need to run in an EVPN-VXLAN fabric. See Designated Forwarding Election and Split Horizon in the EVPN Multihoming Overview topic for additional information on these features.

The core isolation feature quickly brings down the local EVPN-LAG when all IBGP EVPN overlay peerings are lost. This feature diverts traffic to other paths when a link is down. See When to Use the Core Isolation Feature for additional information on core isolation.

EVPN LAG Interfaces Advanced Capabilities

EVPN LAG multihoming has a series of advanced feature capabilities when compared to other LAG multihoming technologies such as Virtual Chassis or Multichassis link aggregation (MC-LAG). The advanced features include Proxy ARP and ARP suppression, IGMP proxy, MAC mobility, MAC history, and duplicate MAC detection.

ARP suppression in an ERB architecture helps reduce Ethernet broadcast traffic flooding across the Layer 2 fabric, thereby freeing up server resources across the fabric. ARP requests, therefore, are not flooded to other leaf devices when the ARP table is already populated with the address from other signaling events, most notably EVPN route sharing. See EVPN Proxy ARP and ARP Suppression, and NDP and NDP Suppression for additional information on Proxy ARP and ARP suppression.

IGMP proxy helps reduce IGMP membership report flooding in a data center by translating the IGMP messages into EVPN type 6 routes to send to all of the leaf devices in the data center. See Overview of IGMP Snooping in an EVPN-VXLAN Environment for additional information on IGMP Proxy. MAC mobility history allows EVPN LAGs to track how often a MAC address moves between ESIs. MAC mobility history allows network administrators to create security actions based on MAC address movements while also simplifying MAC address-related administration in an EVPN. See Overview of MAC Mobility for additional information on this feature.

EVPN LAG Pros and Cons

The following table summarizes the pros and cons of EVPN-VXLAN and EVPN multihomed data centers.

Table 1: EVPN LAG Pros and Cons

Pros	Cons
Flexibility. EVPN-VXLAN is an industry standard based approach to data center fabric creation that appears to be gaining traction and wide adoption across the networking industry.	More devices to manage than in case of a single aggregated platform like Virtual Chassis.
Full support for active-active multihomed access links.	Perceived as complex.
Integrates with Contrail provisioning.	Use with FCoE/FIP is not supported.
Functional parity with Virtual Chassis for network layer functions in next generation access (NGA) use cases.	
Simplified overall topology. EVPN, in particular when configured at the leaf layer, reduces the total number of technologies in a data center fabric.	
Simplified configuration. The same configuration can be shared per given EVPN LAG between different leaf devices and there is no need to provision back to back links between the leaf devices.	
Flexible configuration. EVPN-VXLAN is supported using enterprise-style and service-provider style (ELS and non-ELS) configurations.	

Table 1: EVPN LAG Pros and Cons *(Continued)*

Pros	Cons
Improved services delivery. EVPN-VXLAN introduces the concept of VLAN normalization at the services dedicated border-leafs level and the overlapped vlan-id (unique VNI) capability at the per POD level (pair of leafs).	

EVPN LAG Configuration Best Practices

IN THIS SECTION

Junos OS on QFX Series switches support Enterprise style configuration and Service Provider style configuration.

Both configuration styles can be used to configure EVPN LAGs. We recommend using the Enterprise configuration style because it supports more data center features like storm control profiles and BPDU blocking without having to enable RSTP on the leaf devices.

This section covers EVPN LAG configuration using both configuration styles.

EVPN LAG Configuration of a CRB Architecture Using the Enterprise Configuration Style

The Enterprise configuration style is the recommended way to enable EVPN LAG functionality in most data center architectures. Enterprise style provisioning is generally simpler than Service Provider style provisioning and generally is more compatible with other Layer 2 features.

The following configuration provides a sample EVPN LAG configuration done using the Enterprise configuration style on the leaf devices in a Centrally Routed Bridging architecture.

```
user@leaf_node1>show configuration interfaces ae0
esi {
     00:33:33:33:33:33:33:33:33:01;   a unique but same value at leafs connected to given server
     all-active;
}
aggregated-ether-options {
    lacp {
        active;
        system-id 00:33:33:33:33:01;   a unique but same value at leafs connected to given server
    }
}
unit 0 {
    family ethernet-switching {
        interface-mode trunk;
        vlan {
            members [ 100-101 ];   the explicit vlan-ids enabled with vxlan - can't be mixed
with regular vlans on same ESI-LAG
        }
    }
}

user@leaf_node1> show configuration vlans
vlan100 {
    vlan-id 100;   value provisioned at the ESI-LAG interface in enterprise-mode
    vxlan {
        vni 50100;
    }
}
vlan101 {
    vlan-id 101;
    vxlan {
        vni 50101;
    }
}

user@leaf_node1> show configuration interfaces et-0/0/50
description esi-lag-member-link;
ether-options {
    802.3ad ae0;
```

```
}


user@leaf_node1> show configuration switch-options
vtep-source-interface lo0.0;      lo0.0 must be in the global routing table
route-distinguisher 1.1.1.21:1;   RD must be a unique value per leaf
vrf-import MY-FABRIC-IMPORT;
vrf-target target:1:5555;


root@dc-tme-qfx5110-1> show configuration protocols evpn
vni-options {
    vni 10001 {
        vrf-target target:1:100;
    }
    vni 10002 {
        vrf-target target:1:101;
    }
}
encapsulation vxlan;
multicast-mode ingress-replication;
default-gateway do-not-advertise;
extended-vni-list [ 50100 50101 ];



user@leaf_node1> show configuration policy-options policy-statement MY-FABRIC-IMPORT
term term1 {
    from community MY-FAB-COMMUNITY;
    then accept;
}
term term2 {
    from community COM-VNI-50100;
    then accept;
}
term term3 {
    from community COM-VNI-50101;
    then accept;
}
then reject;

user@leaf_node1> show configuration policy-options community MY-FAB-COMMUNITY
members target:1:5555;
```

```
user@leaf_node1> show configuration policy-options
policy-statement LB {
    term term1 {
        from protocol evpn;
        then {
            load-balance per-packet;
        }
    }
    term term2 {
        then {
            load-balance per-packet;
        }
    }
}
policy-statement MY-FABRIC-IMPORT {
    term term1 {
        from community MY-FAB-COMMUNITY;
        then accept;
    }
    term term2 {
        from community ESI-SPINE;
        then accept;
    }
    term term3 {
        from community COM-VNI-50100;
        then accept;
    }
    term term4 {
        from community COM-VNI-50101;
        then accept;
    }
    then reject;
}
policy-statement MY_VTEPS {
    term term1 {
        from {
            route-filter 1.1.1.0/24 prefix-length-range /32-/32;
        }
        then accept;
    }
    then reject;
}
community COM-VNI-50100 members target:1:100;
```

```
community COM-VNI-50101 members target:1:101;
community MY-FAB-COMMUNITY members target:1:5555;
community SPINE-ESI members target:1:8888;

user@leaf_node1> show configuration protocols bgp
log-updown;
group overlay {
    type internal;
    local-address 1.1.1.21;
    family evpn {
        signaling;
    }
    vpn-apply-export;
    local-as 64512;
    bfd-liveness-detection {
        minimum-interval 300;
        multiplier 3;
    }
    multipath;
    neighbor 1.1.1.11;
    neighbor 1.1.1.12;
}
group underlay {
    type external;
    export MY_VTEPS;
    multipath multiple-as;
    neighbor 10.10.19.1 {
        peer-as 65511;
    }
    neighbor 10.10.21.1 {
        peer-as 65512;
    }
}

user@leaf_node1> show configuration interfaces et-0/0/48
description spine0_connected;
mtu 9216;
unit 0 {
    family inet {
        address 10.10.19.2/24;
    }
}
```

```
user@leaf_node1> show configuration interfaces et-0/0/49
description spine1_connected;
mtu 9216;
unit 0 {
    family inet {
        address 10.10.21.2/24;
    }
}
```

EVPN LAG Configuration of a CRB Architecture Using the Service Provider Configuration Style

The QFX5110 switches also support the Service Provider style of configuration. When configuring EVPN LAGs in the Service Provider configuration style, multiple units per given EVPN LAG are assigned. These multiple units provide an opportunity to enable more selective filter and rate limiting per unit interface, but these assignments must be done on a per unit basis and are therefore burdensome to configure and maintain. This granularity is generally not needed in data center architectures using spine and leaf topologies, so we do not recommend using Service Provider style configuration when enabling EVPN LAGs in most data center environments.

The following configuration provides a sample EVPN LAG configuration done using the Service Provider configuration style. This configuration is supported when QFX5110 or QFX5120 switches are operating in the leaf device role in an architecture where the IRBs are not on the leaf devices, which are the Centrally Routed Bridging (CRB) and Bridged Overlay (BO) architectures. The Enterprise configuration style must be used to enable EVPN LAGs when a QFX5110 or QFX5120 switch is used as the leaf device in an Edge Routed Bridging (ERB) architecture.

```
user@leaf_node1> show configuration interfaces ae11
flexible-vlan-tagging;
encapsulation extended-vlan-bridge;
esi {
    00:33:33:33:33:33:33:33:33:01;
    all-active;
}
aggregated-ether-options {
    lacp {
        active;
        periodic fast;
        system-id 00:33:33:33:33:01;
```

```
    }
}
unit 100 {
    vlan-id 100;
}
unit 101 {
    vlan-id 101;
}


<- Corresponding VxLAN - > VLAN Mapping

user@leaf_node1> show configuration vlans vlan100
interface ae11.100;
vxlan {
    vni 50100;
}
{master:0}

user@leaf_node1> show configuration vlans vlan101
interface ae11.101;
vxlan {
    vni 50101;
}
```

RELATED DOCUMENTATION

Cloud Data Center Architecture Guide

EVPN Feature Guide

Multihoming an Ethernet-Connected End System Design and Implementation

EVPN Multihoming Overview