Layer 3 Leaf & Spine Design and Deployment Guide

The intention of this guide is to provide a systematic and well thought out series of steps to assist the reader with the design and deployment of a Layer 3 Leaf and Spine (L3LS) topology. The example deployment is based on a design which meets a set of predefined requirements as listed in the System Requirements section of this guide. The intended audiences for this guide are network engineers and administrators as well as well as cloud architects. A good working knowledge of networking principles and specifically the Border Gateway Protocol (BGP) is assumed.

The guide is broken down into four main sections as noted below, each section begins with a high level overview and gets progressively more detailed.

- System Requirements
- Design Overview
- Detailed Design
- Configuration Examples

Table of Contents

Contents

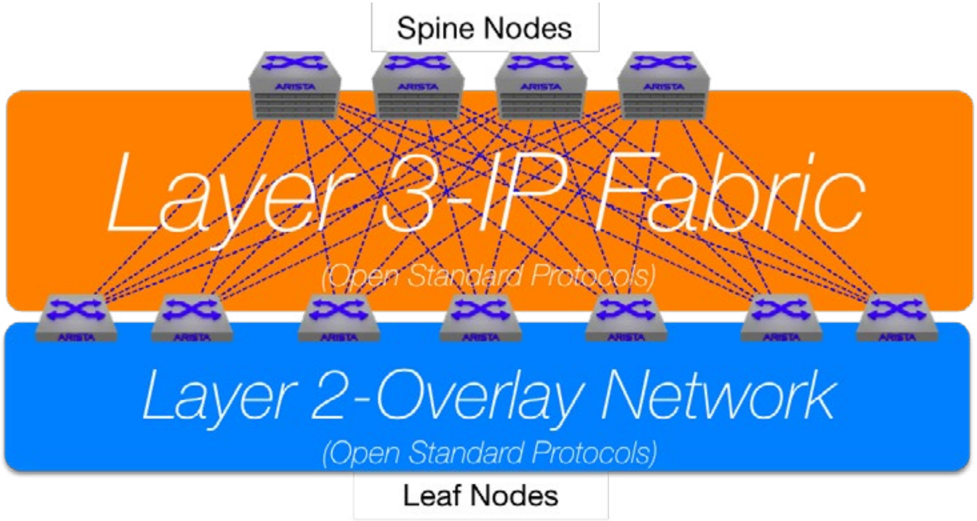
The Drivers for Layer 3 Leaf + Spine Topologies

One of the biggest challenges in today's data centers is the changing traffic patterns and volume of traffic. With the mass adoption of server virtualization and new application architectures it became readily apparent that the legacy three-tier data center model was not the optimal design to support heavy East-West traffic flows.

The traditional 3-tier model, comprised of access, aggregation/distribution and a pair of core switches was designed to be fault tolerant with less emphasis on the proper use of bandwidth. The 3-tier model may still be adequate in a LAN campus setting where traffic patterns remain largely unchanged however the demands of next generation data centers required a new approach.

Arista's L3LS leverages a method of interconnecting switches known as a Clos network. The Clos network was originally envisioned by Charles Clos in 1952 and was based on circuit switching technologies. Arista has pioneered the modern day Clos network by leveraging Ethernet, IP protocols and high density switching silicon.

Clos designs are non-blocking in nature and provide predictable performance and scaling characteristics making them ideal for modern datacenter designs. Interconnections between switches in an L3LS are Layer 3 point-to- point links eliminating the need for traditional Layer 2 loop prevention technologies such spanning-tree or proprietary fabric technologies. Arista's Clos design is comprised of Spine and Leaf switches as depicted below.



Layer 3 Leaf & Spine (L3LS) with L2 Overlay

The role of the spine switches in the Clos topology is to serve as a central backbone or "spine" in which all leafs interconnect. The spine is expected to robust and non-blocking. Each leaf switch has a Layer 3 (point-to-point) link to each spine; thus the name Layer 3 Leaf and Spine (L3LS).

The real value in a Layer 3 Leaf and Spine topology allows customers and service providers alike to design, build and deploy highly scalable, stable and resilient data center networks. By leveraging standard IP routing protocols such as BGP or OSPF large flat Layer 2 failure domains can be eliminated. Adopting an IP based fabric eliminates the possibility of vendor lock-in and allows customers the choice to swap networking components based on their needs, regardless of manufacturer.

Much like the mass adoption of server virtualization technologies customers now have similar choices when it comes to networking. Network virtualization, through the use of overlay technologies like VXLAN, provides the means to extend Layer 2 domains over the routed L3LS. Network overlays support the need for Layer 2 adjacencies in the datacenter and can run seamlessly over an L3LS.

In summary, using standards based IP routing protocols to build redundant and resilient L3LS networks coupled with modern overlay networking technologies provides the best of both worlds and is the foundation for Arista's Universal Cloud Network (UCN) Architecture.

System Requirements

The table below details a list of typical requirements seen in real data center specifications. The goal of this guide is to ensure all of these requirements are met as well as demonstrate the necessary system configurations to deploy them. The requirements include aspects of network and server requirements.

Table 1: System Requirements	
Spine Redundancy	There is requirement to have greater than two spine/core switches to share the load.
Spine Resiliency	The requirement is to have the ability to remove a spine switch from service or suffer a spine failure and have it return to service with little or no impact on application traffic flow.
Scalability	The network must be able to seamlessly scale to support future bandwidth requirements.
Non-Blocking	Design must support the ability to implement leaf-spine subscription ratios based on specific application requirements.
Congestion Avoidance	Design must have the capability to absorb peak traffic loads without losing packets. Network must have mechanisms to ensure traffic can be prioritized and queued if necessary to eliminate the potential of packet loss.
Active/Active Server Connectivity	Each server to have active/active connections, one to a primary leaf switch and one to secondary leaf switch.
Open Standards	The network must support open standards based protocols, no vendor proprietary protocols or features will be used.
Edge Connectivity	Network design to include connectivity into the LAN / WAN environment.
Network Address Translation	Native server IP's must be hidden from the outside network i.e.) Network Address Translation (NAT) but be supported at the network edge.
Traffic Engineering	Mechanisms to ensure traffic can be prioritized and or dropped based on policies.

Arista Universal Cloud Network (UCN) Architecture

The system requirements outlined in the previous section can be met with Arista's Universal Cloud Network (UCN) Architecture. The diagram in Figure 1 depicts the components that make up this architecture. For a comprehensive overview of Arista's UCN architecture you can download the Arista Universal Cloud Network Design Guide at http://www.arista.com/assets/data/pdf/DesignGuides/Arista-Universal-Cloud-Network-Design.pdf

The focus of this guide is the Layer 3 Leaf and Spine (L3LS) and specifically the design, components and configuration details required to build, test and deploy it.

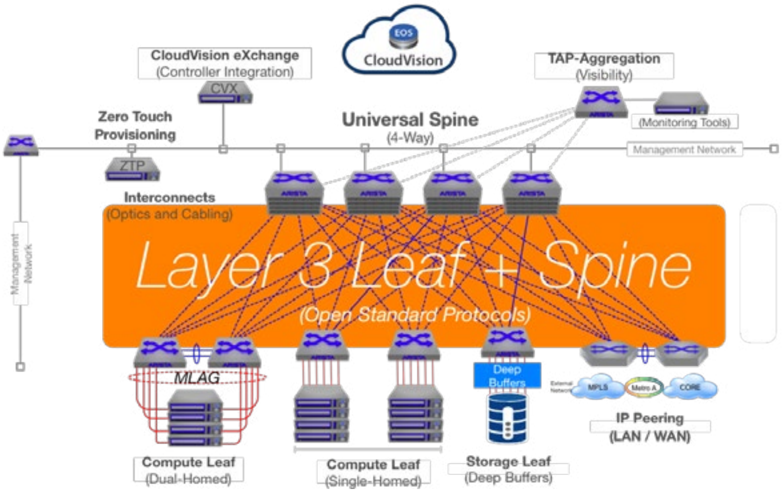


Figure 1: Arista's Universal Cloud Network (UCN) Architecture

Design Overview

The Layer 3 Leaf and Spine (L3LS) topology is the foundation of Arista's Universal Cloud Network Architecture. At a high level the Layer 3 Leaf Spine (L3LS) is simply a collapsed 2-Tier topology comprised of spine and leaf switches. This simple design, when coupled with the advancements in chip technology, a modern operating system and an Open standards approach, provides significant performance and operational improvements.

One of the main advantages of the L3LS design is that the spine can be easily scaled up or down to meet the requirements of small and mid-sized enterprises or up to meet the needs of the largest cloud providers. Spine widths can be from 2 to 128 nodes and supports up to 128-Way ECMP.

By adopting a merchant silicon approach to switch design architects are now able to design networks that have predictable traffic patterns, low latency, minimal oversubscription and the flexibility to scale without changing the overall architecture. Legacy designs often incorporated more than two tiers to overcome density and oversubscription limitations.

Leaf and spine switches are interconnected with routed point-to-point links and each leaf has at least one connection to each spine. Spine switches no longer have a direct dependency on one another, another benefit to the design. BGP, OSPF or ISIS can be used as the routing protocol and Equal Cost Multi Path (ECMP) is used to distribute traffic evenly amongst the individual spine switches, this load sharing behavior is inherent to the design.

A Universal Spine

Arista believes the data center network spine should be universal in nature. What this means is that by using standard protocols and design methods coupled with robust hardware components the data center spine can be leveraged throughout the campus. The concept behind Arista's Universal Spine is:

- Build a spine that meets the needs of all Data Centers
- Reliable and simple to operate
- Interoperate with any and all vendors equipment, leaf switches, firewalls, Application Delivery Controllers etc.

A four-node spine is depicted in Figure 2 and will be used as the working example for this guide.

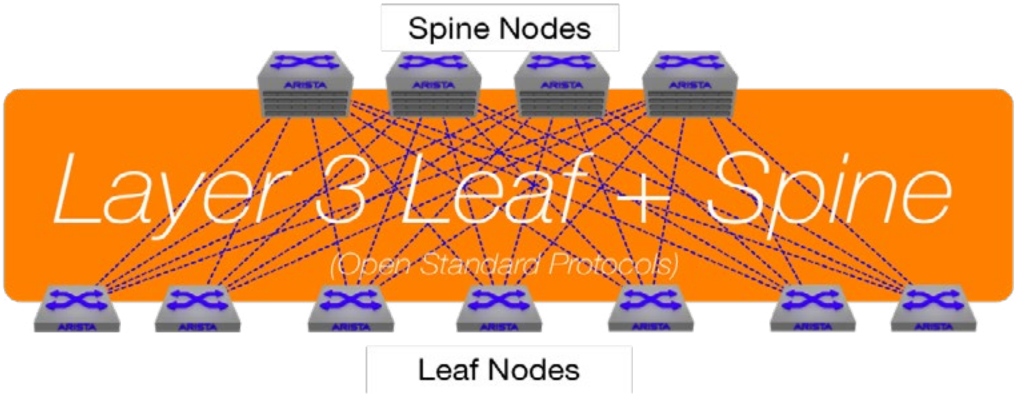


Figure 2: Arista Layer 3 Leaf & Spine (L3LS)

Leaf Options

Leaf switches ensure compute, storage and other workloads get the necessary connectivity and bandwidth dictated by the applications they serve. In the past the inter-connections between the leaf and spine switches have been heavily oversubscribed, careful consideration needs to be taken to ensure the appropriate amount of bandwidth is provisioned. With greater server densities, both virtual and physical, port densities and table sizes are another consideration that need to be taken into account when selecting platforms.

From a leaf design perspective there are three main configurations that need to be considered, designs that support Single-Homed workloads, Dual-Homed workloads as well as a workload such as IP Storage that may benefit from a deep buffer solution such as a Storage Leaf.

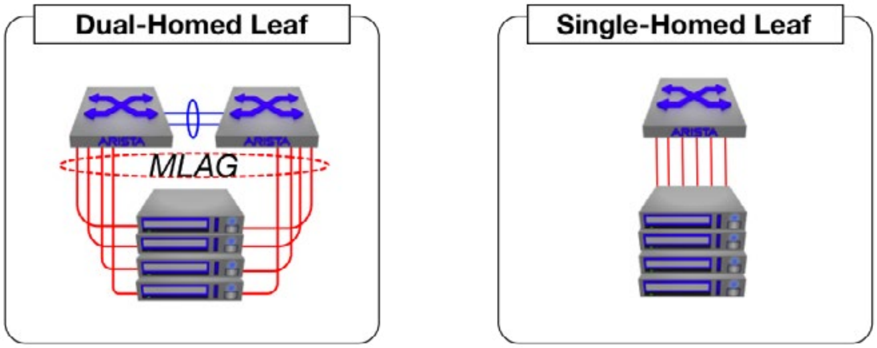
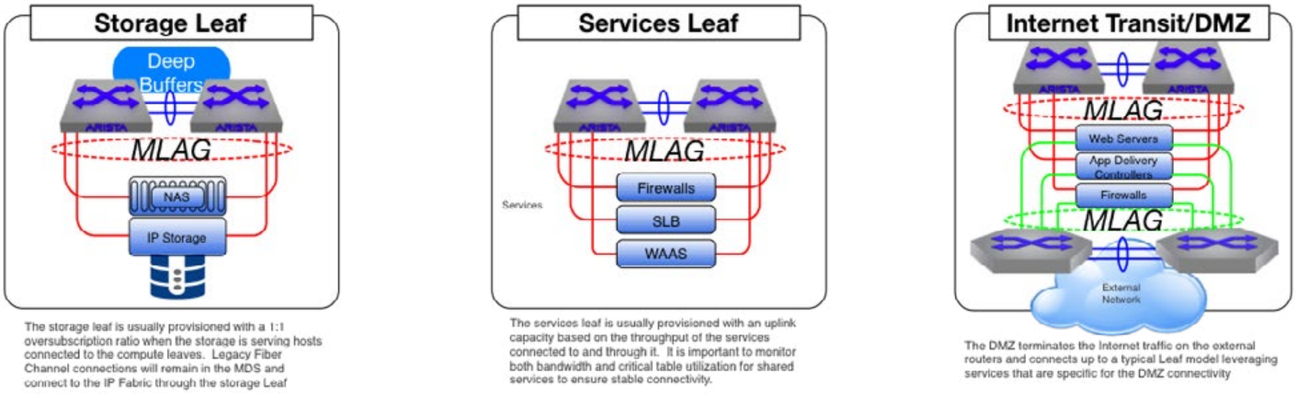


Figure 3: Single-Homed and Dual-Homed Leaf Configurations

Dual-homed systems would leverage a MLAG (Multi-Chassis Link Aggregation) configuration. The MLAG configuration supports the requirement for active/active server connectivity with switch level redundancy at the Top of Rack (ToR).

Other things to consider when making design decisions and platform selections are requirements such as Low Latency, VXLAN Bridging, VXLAN Routing as well as VXLAN Bridging / Routing with MLAG.



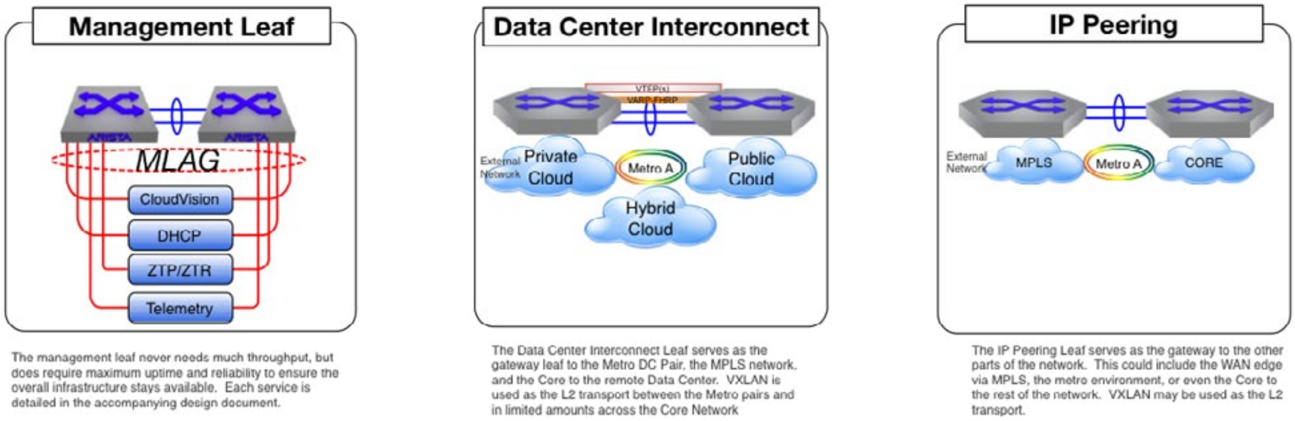
The storage leaf is usually provisioned with a 1:1 oversubscription ratio when the storage is serving hosts connected to the compute leaves. Legacy Fiber Channel connections will remain in the MDS and connect to the IP Fabric through the storage Leaf

The services leaf is usually provisioned with an uplink capacity based on the throughput of the services connected to and through it. It is important to monitor both bandwidth and critical table utilization for shared services to ensure stable connectivity.

The DMZ terminates the Internet traffic on the external routers and connects up to a typical Leaf model leveraging services that are specific for the DMZ connectivity

Figure 4: Storage, Services and Internet DMZ Leaf Configurations

Dedicated Storage Leafs can also be provisioned to ensure IP based storage systems can endure the sustained traffic bursts and heavy incast events. In this design a deep buffer switch would be utilized near the storage system to support the requirements. Deep buffers ensure fairness to all flows during periods of moderate and heavy congestion.



The management leaf never needs much throughput, but does require maximum uptime and reliability to ensure the overall infrastructure stays available. Each service is detailed in the accompanying design document.

The Data Center Interconnect Leaf serves as the gateway leaf to the Metro DC Pair. the MPLS network. and the Core to the remote Data Center. VXLAN is used as the L2 transport between the Metro pairs and in limited amounts across the Core Network

The IP Peering Leaf serves as the gateway to the other parts of the network. This could include the WAN edge via MPLS, the metro environment, or even the Core to the rest of the network. VXLAN may be used as the L2 transport.

Figure 5: Management, Data Center Interconnect and IP Peering Leaf Configurations

Regardless of the leaf configuration, single-homed, dual-home or otherwise, the leaf to spine connectivity is consistent ensuring a common configuration throughout the fabric. Arista has a number of spine and leaf platforms to choose from that meet a broad range of requirements. A more detailed look at the Arista Cloud Network Portfolio can be found at http://www.arista.com/en/products/switches.

Detailed Design

The L3LS has a number of elements that need to be considered during the detailed design. At a high level this work can be broken down into Leaf, Spine, Interconnect and BGP Design.

The diagram below begins to reveal some of the finer points of the Layer 3 Leaf and Spine design. For the purpose of this exercise some assumptions will be made about many of these details however it will be noted and explained as to how and why these design choices could apply to network planning. The stated system requirements will also guide our decision-making. The design includes 100G between leaf and spine, though 40G is also an option depending on requirements.

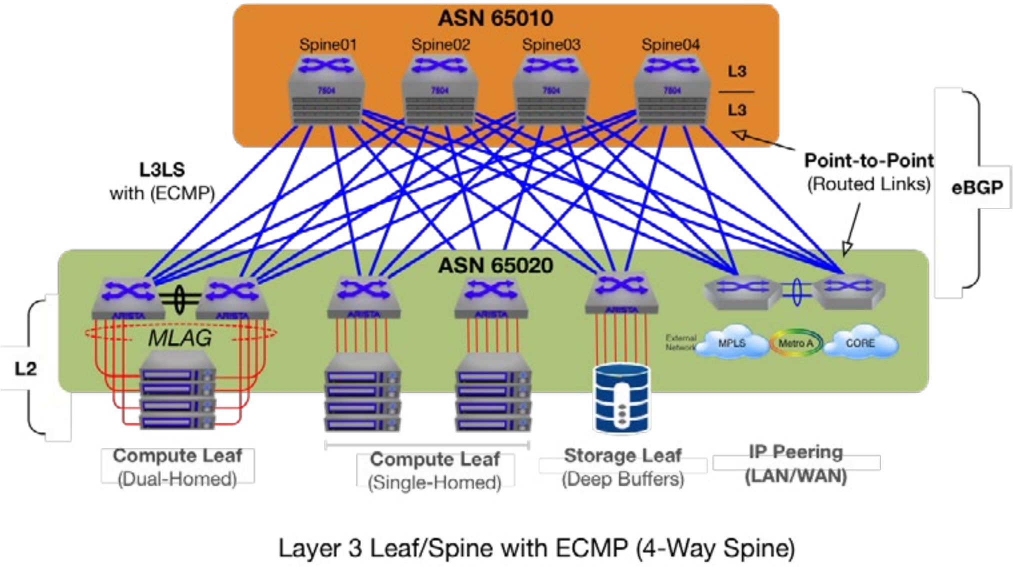


Figure 6: Layer 3 Leaf/Spine with ECMP (4-Way Spine)

Leaf Design Considerations

Application and server requirements dictate leaf design choices and platform selection. The Leaf Design section is broken as follows: Interfaces and Port Densities, Subscription Ratios, Equal Cost Multi-Path routing, Table Sizes, Single and Dual-Home Workloads, Layer 2 Requirements and Transceivers and Cabling.

Interfaces and Port Densities

One of the first things that need to be considered when beginning design work is what type of workload will be supported and how many workloads will need to be scaled to. Getting detailed answers to these questions upfront is paramount to making good design decisions.

There are a number of interface choices for physical server connectivity today and the list is growing. Below is a list of interfaces that require consideration. The list is not a comprehensive of all interface types but focuses on short-range optics commonly seen with the data center.

- Small Form-factor Pluggable (SFP): 1GBASE-T, 1GBASE-SX, 10GBASE-T, 10GBASE-SR
- Quad Small Form-factor Pluggable (QSFP): 40GBASE-SR
- Multi Speed Ports (MXP): 10, 40 & 100G
- QSFP100: 25GBASE-CR, 50GBASE-CR, 100GBASE-SR4

Teams outside of the network-engineering group may drive interface requirements at the end of the day. With traditional DC networks this was less of a concern however new interface types and speeds have changed several things. The first one being parallel lanes and the second being cabling types such as MTP, in both cases it requires a good understanding of optical requirements, specifically 40G but also 25G, 50G and 100G.

There are also situations where retrofitting or upgrading and existing data center network is necessary, which leaves engineers in a situation where they are forced to adopt existing cabling plants.

The quantity of servers within each rack and the anticipated growth rate will need to be documented. This will dictate the switch port density required. Arista has a broad range of platforms in densities from 32x10GBASE-T ports (with 4x40G uplinks) in a 1RU design all the way to 64x100G ports in a 2RU leaf platform, in addition to numerous others.

Transceivers and Cables
As with any network design transceiver and cabling types need to be determined up front. A deep dive on cabling and transceiver selection is beyond the scope of this guide however more information can found at https://www.arista.com/en/products/transceivers-cables.

Connections between leaf and spine may be 40G or 100G. For this design exercise 100G uplinks have already been determined however the cabling/media has not. If runs are short enough Twinax or Direct Attached Cables can be used and are very cost effective. Active Optical Cables (AOC) are another good economical choice, both of these cable types have integrated QSFP100 transceivers. For an existing Multi-Mode (MM) or Single-Mode (SM) fiber plant there are a number of choices. For this design guide a MM plant will be used.

Leaf Uplinks
The type and number of uplinks required at the leaf switch is dictated by the bandwidth and spine redundancy/resiliency needed. With a four-node spine for example a minimum of four 100G uplinks would be utilized.

Traffic Load Balancing
Load balancing traffic over multiple links is the cornerstone for achieving efficiency in a modern data center network. Equal Cost Multi-Path (ECMP) is a routing strategy where next-hop packet forwarding to a single destination can occur over multiple "Best Paths". In a Layer 3 Leaf and Spine topology all leaf switches are directly connected to every spine and as such a network with a four-way spine would provide four paths for a leaf switch to load-balance traffic amongst. To utilize all paths routes to all spines must tie for top place when calculating routing metrics. The maximum-paths feature, covered in more detail below, enables ECMP routing.

Table Sizes
With the adoption of virtualization technologies there has been an explosion of virtual machines on the network. In a somewhat silent manner these virtual machines have increased the MAC addresses count on the network significantly. Packet processors have finite resources that need to be considered during the design. Moving from flat layer 2 networks to a routed L3LS mitigates much of this concern, as MAC addresses are now locally significant and contained at the leaf. It is still important to understand the scaling requirements for MAC address, ARP and route tables in order to make the appropriate platform selections.

Single-Homed Workloads
Certain applications are designed in a fault tolerant manner to support hosts joining and leaving the workload dynamically. Such workloads can be attached to a single leaf and rely on the underlying application for redundancy rather than the network. Single-homed hosts can be connected to the same leaf switches as dual- homed workloads as long as sufficient bandwidth is provisioned on the MLAG peer link.

Dual-Homed Workloads
Another requirement for this design includes network level fault tolerance for server connectivity. Network level fault tolerance is the ability for a workload to survive a single switch failure (at the rack level) without impacting host connectivity and ultimately

application availability. Generally speaking network level fault tolerance assumes active/active server connectivity.

To support these requirements a Dual-Home Leaf Configuration utilizing MLAG will be used. MLAG is standards based and is interoperable with any device that supports the Link Aggregation Control Protocol (LACP) / 802.3ad specification. This configuration supports fault tolerance and active/active load sharing for physical and virtual servers.

IP Peering Leaf

The IP Peering Leaf provides connectivity to resources outside of the Leaf and Spine topology. This may include services such as routers, firewalls, load balancers and other resources. Although the IP Peering Leaf is deployed in a similar manner as other leafs switches, traffic traversing the IP Peering Leaf is typically considered to be North-South traffic rather than East-West. The IP Peering Leaf requires specific consideration as it is often connected to upstream routers at a fraction of speeds of the network spine. For example, a typical leaf-spine interconnect would be running at 100G and an IP Peering Leaf could be connected to the outside world at 1 or 10G. This speed change needs to be understood as higher speed links can easily overwhelm lower speed links, especially during burst events. Routing scale and features also need to be taken into account.

The IP Peering Leaf can also be leveraged as the peering point with external providers and is capable of carrying Internet scale routing tables. Large routing tables and high performance with low power per port and high density enable the Arista 7500 series and 7280 series to fit seamlessly into the role of 100G IP peering platforms. These platforms also support various tunneling technologies including MPLS, VxLAN, GRE and MPLSoGRE along with programmatic traffic steering options that allowing engineers to optimally route the content.

Default Gateways and First Hop Routers

The Arista Data Center design uses an Anycast default gateway insertion technique known as Virtual Address Resolution Protocol or VARP. On a MLAG pair both switches coordinate the advertisement of an identical MAC and IP address (the VARP address) for the default gateway on each segment. In this model the host can send traffic on a packet-by-packet or flow-by-flow basis to either switch. Each default gateway can receive and service the request making a first hop intelligent routing decision without traversing the peer link. An advantage to using VARP is that there is no control protocol or messaging as utilized in VRRP. This significantly reduces the burden on switch CPUs, as the CPU must process control packets.

Layer-2 Requirements

Another common requirement for many data centers is the need to provide Layer 2 adjacencies between compute nodes located in different racks and connected to separate switches. Many applications, particularly in the enterprise space, require a common/shared subnet between application servers. This is not a new or unique problem for data centers however historically the solution has always been somewhat tough to solve in a low risk cost effective manner. The Virtual eXtensible LAN (VXLAN) protocol is the answer to solving this problem in modern data center design. Deploying an overlay using VXLAN on top of a Layer 3 Leaf and Spine (L3LS) topology ensures L2 requirements can be satisfied while still meeting other specific requirements, such as high availability, that the L3LS design provides.

VXLAN Design & Deployment details are not part of this guide at this time, that being said Arista has a broad range of VXLAN capable switches which can be utilized in a L3LS design.

Link Layer Discovery Protocol

The Link Layer Discovery Protocol (LLDP) is a vendor-neutral link layer protocol used by devices for advertising their identity, capabilities, and neighbors on a local area network.

As data centers become more automated LLDP is being leveraged by many applications as a way to automatically learn about adjacent devices. Adjacent is a key word as LLDP only works between devices that are connected at layer two, i.e. on the same physical segment and in a common VLAN.

At the network level LLDP becomes important when integrating with provisioning systems such as OpenStack and VMware. Through

LLDP, switches are able to learn details about connected devices such physical / virtual machines, hypervisors as well as neighboring switches. As an example of this type of integration, when VM instances are created on compute nodes the Ethernet trunk port between the leaf switch and compute node can be automatically configured to allow the required VLANs, this is enabled by using information learned by LLDP.

The use of LLDP should be considered and reviewed with virtualization and cloud architects during the design activities.

Spine Design Considerations

Characteristics of a Network Spine

As a subset of the initial requirements presented in this guide the network spine requires careful consideration. The network spine is expected to be robust and support all types of workloads at both low and peak loads. A proper spine design should be able to scale as the business grows without the need for forklift upgrades and have a 5+ year lifespan. Last but not least the spine should provide deep visibility into switch performance data while at the same time be easy to update and automate.

Key Spine Attributes

There are also several more specific attributes that require consideration. In larger networks or networks built with chassis based systems the design needs to take into consideration the Internal Switching Fabric itself. Switching fabrics can be broken down into two main categories, Ethernet/Flow-Based and Cell-Based. Much the same as leaf design, queuing and buffering are also considerations as is the tolerance level to accept packet loss in the network. Table sizes, power and density as well as cost are always considerations as well. Using open standards based protocols are also key attributes of any good design.

Internet Switching Fabrics

In chassis based systems (as well as multi-chip systems) there needs to be a way to connect front panel ports from one linecard to ports on other linecards. These connections are made behind the scenes via specific fabric packet processors or other types of internal switching chips. The key take away here is that there is more than one type of "internal switching fabric" and it is important to understand the differences when making spine design decisions.

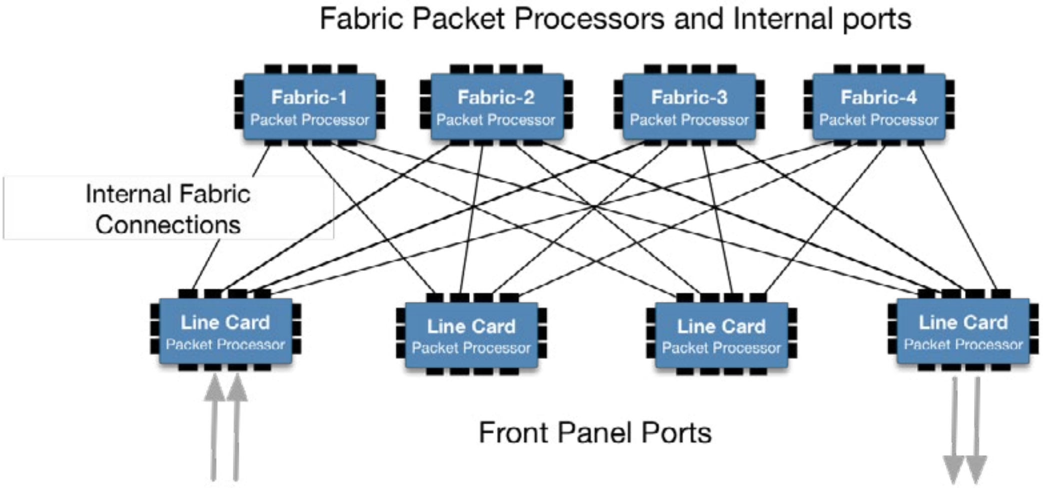

Fabric Packet Processors and Internal ports
Fabric-1 Packet Processor
Fabric-2 Packet Processor
Fabric-3 Packet Processor
Fabric-4 Packet Processor
Internal Fabric Connections
Line Card Packet Processor
Line Card Packet Processor
Line Card Packet Processor
Line Card Packet Processor
Front Panel Ports

Figure 7: Internal Switching Fabric

Ethernet-Based Fabric

The first type of fabric is known as an Ethernet-Based fabric. As the name might suggest an Ethernet-Based fabric is largely bound by the rules of Ethernet. Consider a single chip as a switch connecting to other chips with patch cables all using Ethernet, an internal Clos design

Within an Ethernet-based design there are limits to the efficiency that can be achieved. In general 80-90% efficiency is deemed achievable using bandwidth aware hashing and Dynamic Load Balancing (DLB) techniques on the linecard to fabric connections.

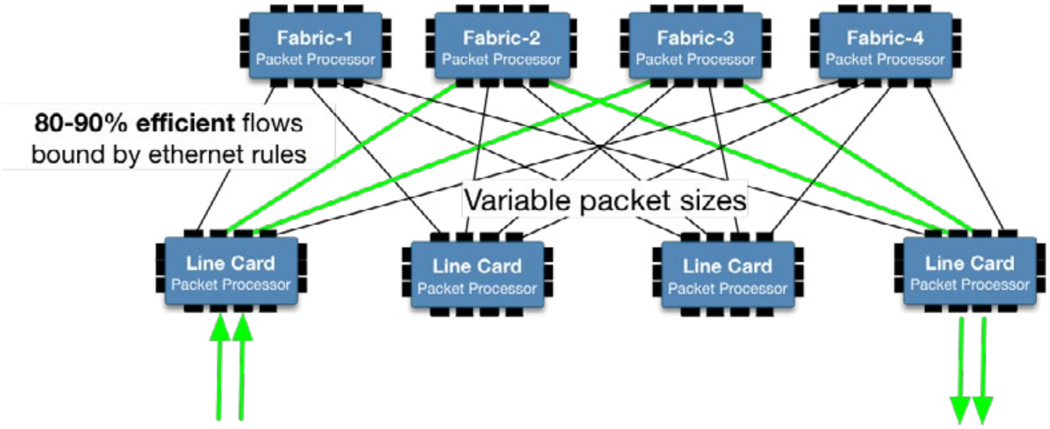


Figure 8: Ethernet-Based Fabric

Cell-Based Fabric - With Virtual Output Queuing (VOQ)

Cell based architectures are quite different as they are not bound by the rules of Ethernet on the switch backplane (the front panel port to fabric connections). A cell-based fabric takes a packet and breaks it apart into evenly sized cells before evenly "spraying" across all fabric modules. This spraying action has a number of positive attributes making for a very efficient internal switching fabric with an even balance of flows to each forwarding engine. Cell- based fabrics are considered to be 100% efficient irrespective of the traffic pattern.

Because the cell-based fabric does not utilize Ethernet it is inherently good at dealing with mixed speeds. A cell- based fabric is not concerned with the front panel connection speeds making mixing and matching 100M, 1G, 10G, 25G, 40G, 50G and 100G of little concern. Adding Advanced Queuing Credit based schedulers with Virtual Output Queues (VOQs) and deep buffers (for congestion handling) to a cell-based platform provides for a lossless based system that deserves consideration.

Cell based systems will give you more predictable performance under moderate to heavy load, the addition of Virtual Output Queue (VOQ) architectures will also help protect against packet loss during congestion. These two capabilities coupled with a deep buffer platform all but guarantee the lossless delivery of packets in a congested network.

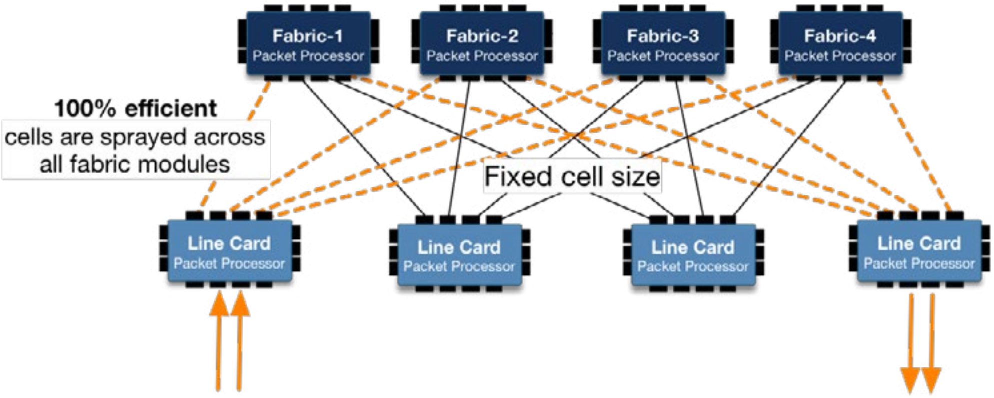


Figure 9: Cell-Based Fabric with VOQ

Choosing a Spine Platform

Like many design choices it comes down to having good data to work with. Some conditions to consider are: low loads, heavy loads, loads during failure conditions (loss of spine, or two) and loads during maintenance windows when a spine may be taken out of service. Ideally having good baselines is a good start, for net new builds this often comes down to modeling and predicated application demands. A proper capacity-planning program is essential to day two operations, ensuring your design can absorb or scale to meet future demands.

Here are some general rules of thumb that can help with the decision-making if you don't have enough data up front.

- In typical networks the first collision is seen at ~30% utilization;
- Excessive collisions begin to happen at around ~70% utilization; and
- Networks begin to collapse above ~70% utilization if something isn't done.

If you can't get a handle on the current and/or future design details, cell based large buffer systems are a catch all that makes the most sense when trying to minimize risk.

Hardware Specifications

The level of redundancy built in the spine as a whole (all switches in the spine) will dictate the level of redundancy required at the platform level. A key consideration is that of supervisor redundancy. In a multi-spine design (three or more switches) the spines ability to lose a node without adversely impacting performance increases. By and large, multi-spine designs are configured to use a single supervisor.

Leaf-Spine Interconnects

All leaf switches are directly connected to all spine switches through fiber optic or copper cabling. In an L3LS topology all of these interconnections are routed links. These routed interconnects can be designed as discrete Point-to-Point links or as Port-Channels. There are pros and cons to each design, discrete point-to-point links are most common configuration and will be the focus of this guide. Leaf-Spine interconnects require careful consideration to ensure uplinks are not over-subscribed. Subscription ratios can be engineered (as discussed in the Leaf Design section of this document) to ensure uplinks are properly sized to prevent congestion and packet loss. Leaf-Spine Interconnects can be 10G, 40G or 100G interfaces.

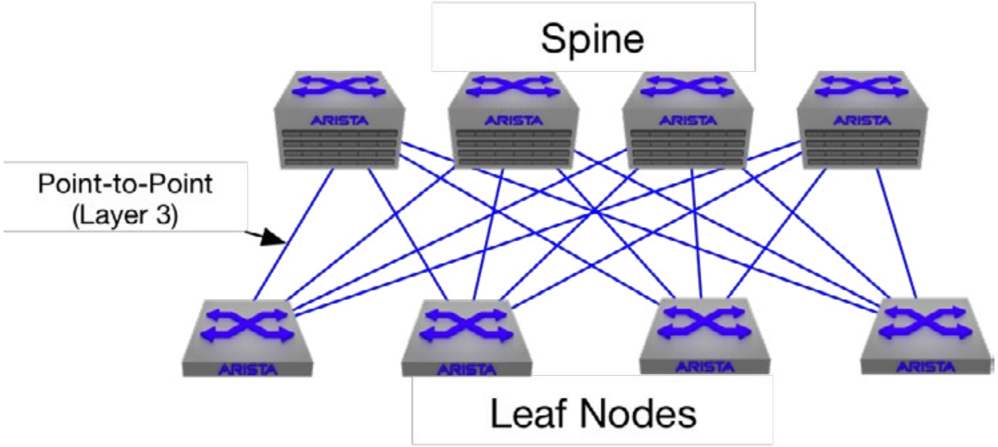


Figure 10: Routed Point-to-Point links

Congestion Management

Subscription Ratios

Subscription and oversubscription ratios are expressed as a ratio of downlink to uplink capacity. An example of this would be a 48 Port 10G switch with four 40G uplinks. In this scenario the oversubscription ratio would be 480G:160G or 3:1. Oversubscription can exist in the North-South direction (traffic entering/leaving a data center) as well as East-West (traffic between devices in the data center).

For this design servers/workloads are attaching to the leaf at 1/10/25/50G. This is important because the bandwidth each server demands is aggregated when it comes to upstream connectivity i.e. the bandwidth consumed on the uplinks. Even though it's possible to have a wirespeed switch it does not mean servers will not encounter congestion. Server virtualization further compounds this problem as virtual machines (workloads) can pop up anywhere at anytime with no need for physical cabling. To ensure servers, both physical and virtual, do not suffer packet loss due to congestion, subscription ratios need to be taken into consideration. Below are some general rules of thumb when it comes to subscription ratios.

- Subscription-ratio of 1:1 for racks hosting IP Storage
- Subscription-ratio of 3:1 for General computing racks
- Subscription-ratio of 6:1 for racks containing Network services
- Subscription-ratio of 1:1 for IP Peering (match external BW with Spine BW)

When calculating subscription ratios it is important to know a little more about the quantity of servers (both virtual and physical) connecting to leaf switches as well as the expected bandwidth requirements. With 48 1/10G ports, there is a maximum of 480G of data coming into (ingress) and out (egress) of the switch. Using a subscription ratio of 3:1 can determine the uplink capacity required, in this example 480G / 3 = 160G. A switch with four 40G uplinks can meet this requirement (4x40G = 160G).

With the introduction of 100G uplinks the oversubscription level can be reduced in the 1/10G general computing deployment to 1.2:1. In this example, 48 1/10G host facing ports and at least 4 100G uplinks towards the spine, the oversubscription is 1.2:1 (480G / 1.2 = 400G).

Some common examples of host facing port to uplink ratios are outlined in the table below with options for 1/10G, 25G and 50G.

Host Ports	Uplinks	Over-Subscription	Example Hardware Configuration
48 1/10G	4 40G	3:1	48 1/10G (SFP+ or Copper) + 6 40G QSFP ports
48 1/10G	4 100G	1.2:1	48 1/10G (SFP+ or Copper) + 6 100G QSFP100 ports
48 1/10G	6 100G	1:1	48 1/10G (SFP+ or Copper) + 6 100G QSFP100 ports
96 25G [48 50G]	8 100G	3:1	96 25G [48 50G] (24 QSFP100) + 8 100G QSFP100 ports

Buffering

In general, congestion management and buffering does not seem to be well understood when it comes to data center networks. In reality, any network can experience congestion, and when it does buffers are utilized in an effort to avoid dropping packets. While a well thought out leaf and spine design minimizes oversubscription, services such as dedicated IP storage systems are prone to receiving large amounts of incast traffic. These types of traffic patterns have the potential to create bottlenecks.

Incast or TCP incast is a many to one communication pattern that is most commonly seen in environments that have adopted IP based storage as well as High Performance Computing applications as such as Hadoop. Incast can occur in different scenarios but a simple example is one where many hosts request data from a single server simultaneously. Imagine a server connected at 10G trying to serve 40G of data from 1000 users, this is a many to one relationship. Sustained traffic flows that exceed the capacity of a single link can cause network buffers to overflow causing the switch to drop packets.

For a detailed explanation on the benefits of buffers in the data center see the following white paper titled Why Big Data Needs Big Buffer Switches http://www.arista.com/assets/data/pdf/Whitepapers/BigDataBigBuffers-WP.pdf

When deep buffer systems are required Arista recommends our 7280 or 7500 series switches. In the 7500 and 7280 series systems, each port is capable of providing up to 50ms of packet buffering.

The implementation of a Leaf and Spine architecture with smaller buffers will perform well in a network that does not experience congestion. This makes it critically important to understand the performance characteristics required prior to making product decisions. Performing a good baseline analysis of traffic flows, particularly during periods of high utilization, will ensure your design meets your requirements.

Data Center Bridging and Priority Flow Control

Data Center Bridging (DCB) and Priority Flow Control (PFC) are two additional protocols used to assist with the lossless delivery of Ethernet.

DCB has two important features: Data Center Bridging Exchange (DCBX) and Priority Flow Control (PFC). EOS uses the Link Layer Discovery Protocol (LLDP) and the Data Center Bridging Capability Exchange (DCBX) protocol to help automate the configuration of Data Center Bridging (DCB) parameters, including the Priority-Based Flow Control (PFC) standard, which enables end-to-end flow-control.

As an example, these features enable a switch to recognize when it is connected to an iSCSI device and automatically configure the switch link parameters (such as PFC) to provide optimal support for that device. DCBX can be used to prioritize the handling of iSCSI traffic to help ensure that packets are not dropped or delayed.

PFC enables switches to implement flow-control measures for multiple classes of traffic. Switches and edge devices slow down traffic that causes congestion and allow other traffic on the same port to pass without restriction. Arista switches can drop less important traffic and tell other switches to pause specific traffic classes so that critical data is not dropped. This Quality of Service (QoS) capability eases congestion by ensuring that critical I/O (in the storage example) is not disrupted or corrupted and that other non-storage traffic that is tolerant of loss may be dropped.

BGP Design

In a short time the Border Gateway Protocol has become the routing protocol of choice for large data center and cloud scale networks. During this time much has been learned about the design and operations of these networks and the benefits are now being realized by organizations of all sizes. Below are a number of the key benefits BGP has over other routing protocols.

- Extensive Multi-Vendor interoperability
- Native Traffic Engineering (TE) capabilities
- Minimized information flooding, when compared to linkstate protocols
- Reliance on TCP rather than adjacency forming
- Reduced complexity and simplified troubleshooting
- Mature and proven stability at scale

EBGP vs. IBGP

Arista recommends using the exterior Border Gateway Protocol (eBGP) in a Layer 3 Leaf and Spine design. There are number of reasons for this choice but one of the more compelling reasons is simplicity, particularly when configuring load sharing (via ECMP) which is one of the main design goals of the L3LS. Using eBGP ensures all routes/paths are utilized with the least amount of complexity and fewest steps to configure.

In contrast an iBGP design would require additional considerations to achieve this behavior. An iBGP design would require all switches in the L3LS to peer with every other device in the network. The use of route reflectors could reduce this burden however the default route reflection behavior introduces additional complexity. By default, route reflectors will only reflect the best prefix, which hampers the ability to perform load sharing. To work around the default route reflector behavior the BGP 'AddPath' feature could be used. AddPath supports the advertisement of multiple paths for the same prefix through a single peering session without the new paths replacing previously learned ones; this is extra work for little return.

Although an iBGP implementation is technically feasible using eBGP allows for a simpler less complex design that is easier to troubleshoot. There is however a corner case where iBGP is still used in this design, this scenario is covered in the BGP Leaf Design section.

Autonomous System Number - Design Options

BGP supports several designs when assigning Autonomous System Numbers (ASN) in a L3LS topology. For this design a Common (shared) ASN will be assigned to the Spine nodes and another ASN to the Leaf nodes. For completeness a second option will be reviewed, Common Spine ASN and discrete Leaf ASNs however this guide will not cover the configuration details.

To start, a quick review of ASN numbering. IANA has reserved, for Private Use, a contiguous block of 1023 Autonomous System numbers from the 16-bit Autonomous System Numbers registry, 64512 to 65534. IANA has also reserved for Private Use a contiguous block of 94,967,295 Autonomous System numbers from the 32-bit (4 Byte) ASN registry, namely 4200000000 - 4294967294. These reservations have been documented in the IANA "Autonomous System (AS) Numbers" registry [IANA.AS].

This guide will use the private AS numbers between 64512 through 65535.

Common Spine ASN - Common Leaf ASN

As the name implies the Common Spine ASN - Common Leaf ASN design uses a single ASN for all spine nodes and another ASN for all leaf nodes. This design has a number of advantages including better use of available 16 bit ASNs and the ability to have a more standard leaf switch configuration. For larger customers preserving 16 bit ASNs reduces the requirement to move to 32 bit ASNs, which all vendors may not support yet.

Having a common leaf configuration simplifies deployment, operations and automation tasks.

When using a common ASN for all leaf switches the rules of BGP need to be reviewed. As a rule an eBGP neighbor, in this case a leaf switch will ignore an update from a neighbor if its local ASN (the leafs ASN) appears in the path (the AS_PATH). In other words the update would be denied on ingress and not be considered for installation in the Routing Information Base (RIB), this is a loop prevention mechanism and is normally desirable. In an isolated environment such as a data center the network topology is well understood, the allowas-in feature was developed for conditions like this.

The **allowas-in** feature permits the switch to "allow" the advertisement of prefixes that contain duplicate autonomous system numbers. This command programs the switch to ignore its ASN in the AS path of routes and allows the entry(s) to be installed into the RIB. The allow-as in features allows us to have a common ASN for all leaf switches. Loop prevention is still at work in the spine due to the spine switches residing in a common AS; allowas-in is not configured on the spine switches.

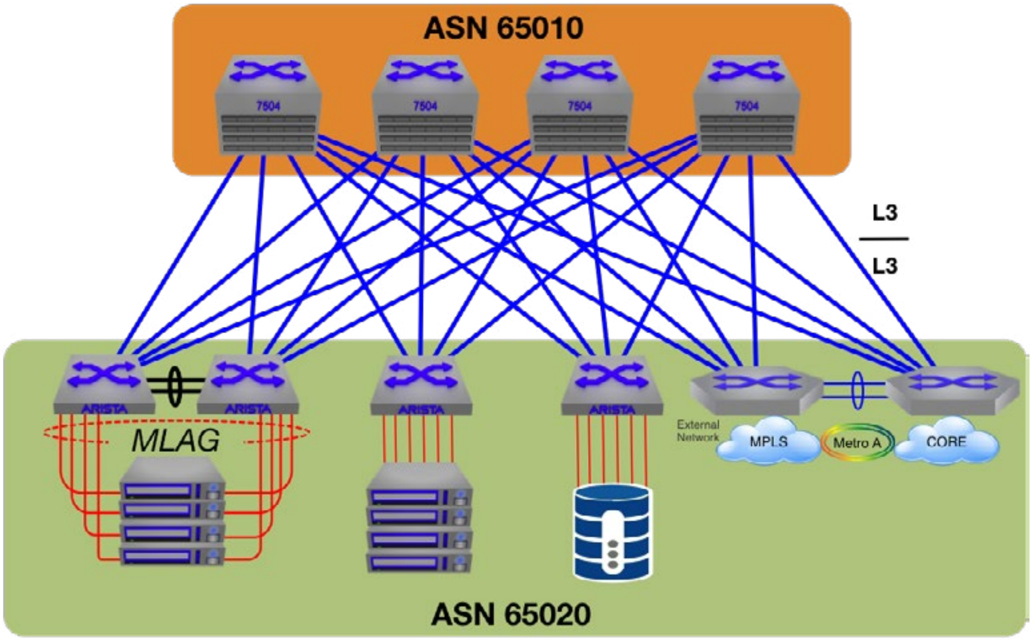


Figure 11: Common Spine - Common Leaf ASN

Common Spine ASN - Discrete Leaf ASN

The Common Spine ASN - Discrete Leaf ASN design uses a single ASN for all spine nodes and discrete ASNs for each leaf nodes.

This design has a number of advantages such as;

- Each rack can now be identified by its ASN
- Traceroute and bgp commands will show discrete AS making troubleshooting easier
- Uses inherent BGP loop prevention mechanisms and therefore does not require the allowas-in feature
- Unique AS numbers help troubleshooting and don't require flexing the EBGP path selection algorithm with allowas-in

Large Scale (where 1023 ASNs is not enough)
One thing to keep in mind if choosing to deploy discrete ASNs at the leaf is that there are a fixed number of ASN available, that being 1023 (64512 to 65534). This is likely not a concern for most customers but it is important to know. Arista EOS does support 4-byte ASNs to avoid this scaling limitation of 2-byte ASNs. Another difference to consider is that having discrete ASN means less of a common configuration.

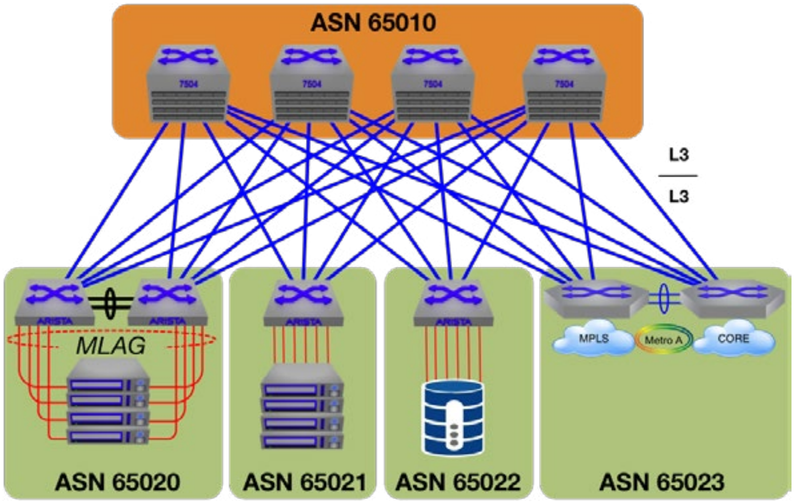


Figure 12: Common Spine Common Leaf ASN

ECMP and BGP Multipath

Leveraging Equal Cost Multi-Path (ECMP) routing is the cornerstone for achieving efficiency in a modern data center network. To utilize all available paths, leaf and spine switches must first process routing updates with common metrics. Under normal circumstances there can only be one winner. By using the maximum-paths feature routes that have "tied" for top place will be added to the Forwarding Information Base (FIB).

The maximum-paths command controls the maximum number of parallel eBGP routes that the switch supports. The default maximum is one route. The command provides an ECMP (equal cost multiple paths) parameter that controls the number of equal-cost paths that the switch stores in the routing table for each route.

Equal Cost Multi Path (ECMP) will also be used to ensure an even distribution of traffic flows throughout the fabric.

In a four-way spine design the maximum paths would be four and the ecmp choices would also be four. Below is an example of the command.

```
!
switch01(config-router-bgp)# maximum-paths 4 ecmp 4
!
```

Maximum Routes

The neighbor maximum-routes command configures the maximum number of BGP routes the switch will accept from a specified neighbor. The switch disables peering with the neighbor when this number is exceeded.

```
!
switch01(config-router-bgp) # neighbor 10.10.1.10 maximum-routes 12000
!
```

BGP and Route Convergence

Convergence Times

One of the most commonly asked questions when running BGP in the data center is, "How fast can the network converge during a link or peer failure?"

The need for fast failure detection in a L3LS design is critical to ensure failures are detected as quickly as possible. EOS's default behavior is to immediately tear down an eBGP neighbor adjacency when a link goes down. In turn, link-down events instantly trigger BGP changes. This ability to rapidly detect and act on unreachable peers results in sub-second (~100ms) removal of eBGP peers and subsequent withdrawal of the associated routes.

In a modern data center it is necessary to have the ability to rapidly detect and remove peers as a catch all for a number of common conditions that can exist in any topology. Scenarios such as asymmetric routing where sending traffic flows are taking different paths than returning traffic flows. A condition such as this can result in lengthy re-convergence in specific scenarios if not managed correctly. Having the switch immediately tear down these connections allows for predictable network behavior and optimal network convergence times.

Bidirectional Forwarding Detection

Bidirectional Forwarding Detection (BFD) is a simple mechanism that detects if a connection between adjacent systems is up, allowing it to quickly detect failure of any element in the connection. It does not operate independently, but as an aide to routing protocols. The routing protocols are responsible for neighbor detection, and create BFD sessions with neighbors by requesting failure monitoring from BFD. Once a BFD session is established with a neighbor, BFD exchanges control packets to verify connectivity and inform the requesting protocol of failure if a specified number of successive packets are not received. The requesting protocol is then responsible for responding to the loss of connectivity.

Routing protocols using BFD for failure detection continue to operate normally when BFD is enabled, including the exchange of hello packets. The basic behavior of BFD is defined in RFC 5880.

In networks without data link signaling, connection failures are usually detected by the hello mechanisms of routing protocols. Detection can take over a second and reducing detection time by increasing the rate at which hello packets are exchanged can create an excessive burden on the participating CPUs. BFD is a low-overhead, protocol-independent mechanism, which adjacent systems can use for faster detection of faults in the path between them.

As a rule Arista does not recommend BFD on directly connected point-to-point links where data signaling exists. BFD is strictly a failure-detection mechanism and does not discover neighbors or reroute traffic. In this application EOS's default behavior performs optimally, this is in part due to Link Fault Signaling (LFS), which is inherent in the 10/40/100G Ethernet standard.

Link Fault Signaling (LFS)

The rapid detection of link failures is in part due to the Link Fault Signaling (LFS) mechanism in the 10/40/100G standard. LFS has provided a whole new meaning to connectivity state and provides a low-level health checker that performs 'local-fault' and 'remote-fault' signaling. LFS signaling behaves much like BFD but at a lower layer, because of this BFD is an added complexity with no benefit.

The Link Fault Signaling mechanism lives below the MAC layer in a sublayer called the Reconciliation Sublayer or RS. Its job is to monitor the status between a local RS and a remote RS and perform link status notification; LFS relies on the Sub layers at the PHY level to detect the faults that render a link unreliable.

At the end of the day a RS reports the fault status of a link, the fault status could be a Local Fault (LF) or a Remote Fault (RF). Upper layer protocols can take action on this status.

- Local Fault indicates a fault detected on the receive data path
- Remote Fault indicates a fault on the transmit path

Tuning BGP

This section of the guide is reserved for specific parameters that can be configured to tune the L3LS. It is important to understand your particular deployment to determine if tuning will be beneficial in your environment.

MLAG (Leaf) Integration with L3LS

To support the requirement of dual-homed active/active server connectivity MLAG will be used at leaf layer using a Dual-Homed Compute Leaf configuration. A standard MLAG configuration is used with a couple of additions to support eBGP peering to the spine as well as iBGP peering between the MLAG peers. For completeness the standard MLAG configuration is covered in the configuration examples section of this guide.

When integrating a MLAG Leaf configuration into a Layer 3 Leaf Spine, iBGP peering is recommend between the MLAG peers. The reason for the peering is due to specific failure conditions that the design must take into consideration. In normal operation paths learned via eBGP (leaf to spine uplinks) will always be preferred over paths learned via iBGP (given other tie-breaker criteria are equal) this is desirable.

Failure Scenarios

During specific failure conditions routes learned via iBGP will come into effect, consider the following failure scenario (illustrated in Figure 13 below):

Node1 is communicating with Node2. Leaf-1 and Leaf-2 are MLAG peers and are configured to run VARP to provide an active/active redundant first hop gateway for Node1, but they are not BGP peers. On Leaf-1, both of its uplinks have failed. While both Spine-1 and Spine-2 would certainly notice this failure and reconverge, some things will not be affected, such as:

- MLAG running between Leaf-1 and Leaf-2 would not notice any change, and continue functioning as normal, which in turn means that the port-channel between Node1 and Leaf-1/Leaf-2 would remain up and function as normal, VARP would also continue to function as normal.
- This is important because traffic leaving Node1 that gets hashed to Leaf-2, which would be fine, but any traffic hashed to Leaf-1 would effectively be black holed.

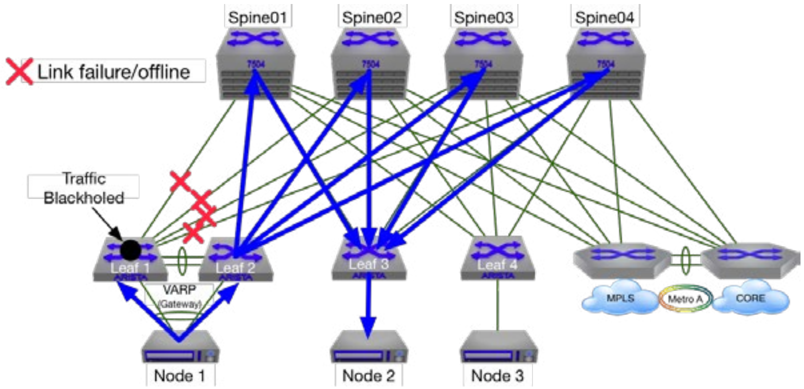


Figure 13: MLAG Integration with L3LS (without iBGP peering)

Peering Leaf-1 and Leaf-2 alleviates this problem. In the same failure scenario (Figure 14 below) with an iBGP peering between Leaf-1 and Leaf-2, any traffic hashed to Leaf-1 would follow the remaining route pointing to Leaf-2 and then be ECMP-routed to the spine. During normal operation the path through the peering leaf switch is less desirable (a longer path) than those directly connected to the spine alleviating any suboptimal path through Leaf-2.

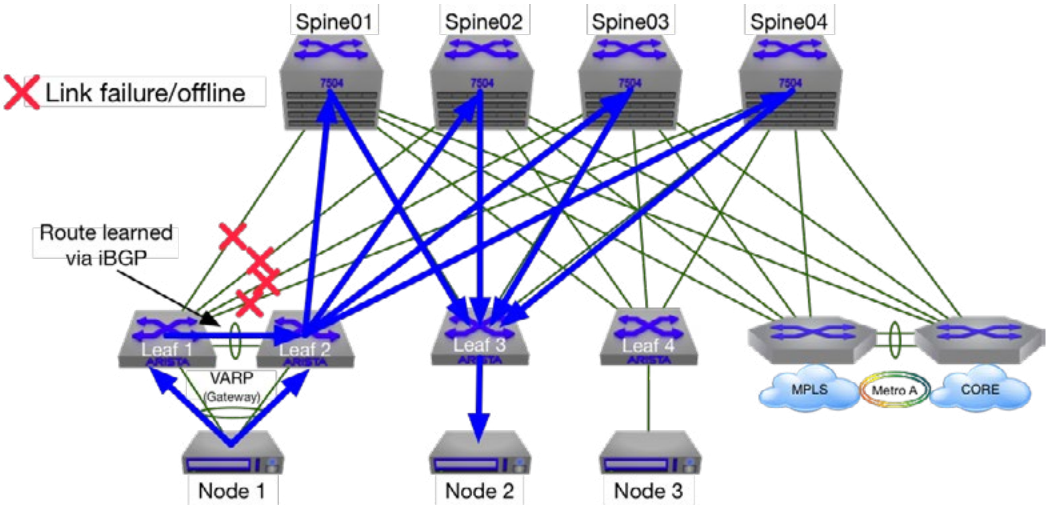


Figure 14: MLAG Integration with L3LS (with iBGP Peering)

IP Addressing in the L3LS

Like any network design IP addressing and logical networks need to be allocated and assigned. It is nearly impossible to give definitive answers for the network design however the examples in this guide will help. As of writing the examples only take into account IPv4 addressing.

Addressing can be broken down into the following sections: Base Prefix, Point-to-Point Networks, Loopback and Management Networks, Server / Workload addressing and Network Address Translation (NAT) addresses.

Base Prefixes

For an optimal addressing plan prefixes within the L3LS should strive to have a common base prefix or set of prefixes. As an example a point-to-point link would have a specific prefix for each localized network i.e. 10.10.0.0/30. In a L3LS with one hundred point-to-point links all prefixes could be summarized by a base prefix of 10.10.0.0/24.

Point-to-Point Networks

Depending on the size of the data center a L3LS may have hundreds of point-to-point links. Each leaf has a point- to-point network between itself and each spine. When carving up address space for these links be sure to strike the right balance between address conservation and leaving room for the unknown. Using a /31 mask will work as will a /30, your personal circumstance will dictate your decision.

Server / Workload Addressing

Server and other workloads need addresses too. When deploying a native L3LS (i.e. No Overlay network) each Leaf hosts a number of networks/prefixes for the local workloads. As a rule a VLAN will be associated with each prefix as well as a local gateway. Each leaf may have upwards of ten VLAN/prefixes and often more. When creating your address plan try to keep a base prefix in mind.

Loopback and Management Interfaces

All switches with the L3LS require a least one loopback address, loopbacks use a /32 mask. Loopbacks are not dependent on link state and therefore are always up making them excellent for troubleshooting. Management Interfaces also require IP addresses as well.

Network Address Translation (NAT)

To hide internal data center addressing or to host applications on publicly routable IP addresses NAT may be required. NAT can be performed in a number of places such as Firewalls, Load-balances and/or Border Routers. For Internet facing services public addresses are required, internal facing services may utilize private addressing.

Regardless of where NAT occurs it is important to allocate address space for this and ensure the device performing the address translation can handle the workload.

Sample addressing schemes are provided in the configuration section of this guide.

Edge Connectivity

To connect the IP Peering Leaf to the Edge Routers, a peering relationship must be established. Like the L3LS, BGP provides the best functionality and convergence for this application. Whether to use iBGP or eBGP truly depends on the specific requirements and capabilities of the edge devices. This guide will focus on using eBGP with the IP Peering Leaf. External BGP (eBGP) is primarily used to connect different autonomous systems and as such lends itself well to this design scenario whereby an IP Peering Leaf, representing the data centers autonomous system needs to connect to the external or transit providers, which may represent any number of different autonomous systems.

Depending on the use case and specific design requirements, attention must also be given to how public prefixes will be shared with the Edge Routers and how access to the Internet will be routed. If this is a requirement, the IP Peering Leaf must be prepared to accept and redistribute a default route to the spine. Conversely, any private autonomous system numbers (ASN) must be removed before advertising to upstream providers and transit peers, this can be done in a number of ways and is beyond the scope of this document.

In addition to providing a general entry and exit point to the traditional network, the IP Peering Leaf may also provide an entry point to the overlay network. When an overlay network is deployed, traffic is typically encapsulated prior to transiting the spine and remains encapsulated until it reaches the destination. For overlay traffic that requires connectivity external to the data center, the traffic must be forwarded to the IP Peering Leaf which can then decapsulate the traffic and forward it to the appropriate destination, whether that be a firewall or host external to the data center.

Overlay Networks

As mentioned in the introduction to this guide the vast majority of data centers still have Layer 2 dependencies. The modern data center solution to supporting these Layer 2 dependencies is network virtualization. Network virtualization technologies, or overlay networks as they are often called, can come in several forms. The Virtual eXtensible Local Area Network (VXLAN) protocol has become the standard for building overlay networks as it seamlessly supports Layer 2 adjacencies over all routed networks, the L3LS included.

To create an overlay network via the physical network, networking equipment must support VXLAN bridging and routing functionality at the hardware level. Hardware based overlays are able to encapsulate and decapsulate traffic destined for the overlay (VXLAN) networks at wirespeed.

Detailed design and configuration examples for VXLAN and other overlay solutions are outside of the scope of this design guide. That being said, all overlay models rely heavily on the design of the Layer 3 underlay.

Configuration Examples

This section of the guide is intended to provide configuration examples for all aspects of the Layer 3 Leaf and Spine deployment. It includes sample configurations as well as steps to verify the configuration where appropriate. The following configuration examples will be covered;

- Base Configuration
- Loopbacks and Management Interfaces

- Point-to-Point Links
- BGP Configuration & Base Routing
- Route Advertising
- MLAG Configuration with L3LS (for Dual-Homed Servers)

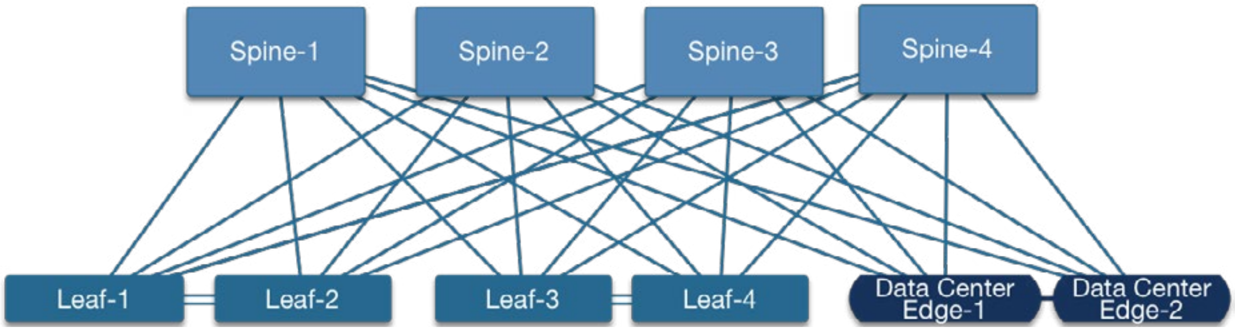


Figure 15: Topology for Configuration Examples

For a more thorough explanation of the specific commands found in this guide please see the Arista System Configuration Guide for the version of EOS you are running. System Configuration Guides can be found on the Arista Support Site at http://www.arista.com/en/support/product-documentation.

Note: Configuration guidance for spine to IP Peering Leaf connectivity is common with the general leaf configuration examples in this guide. IP Peering Leaf configuration can be quite variable depending on the external providers in use and is therefore outside of the scope of this guide. Please refer to the design section of this guide for an over view of Edge Connectivity design considerations.

Base Configuration (All Switches)

Below is a base configuration; hostname, DNS server addresses, Domain name, as well as NTP server information addresses are required before proceeding.

```
! Spine-1 Example
!
hostname Spine-1
ip name-server $DNSHostAddress
ip name-server $DNSHostAddress
ip domain-name $CompanyDomainName
!
ntp source Management1/1
ntp server $NTPHostAddress1 prefer
ntp server $NTPHostAddress2
!
username admin role network-admin secret 5 $Password
!
clock timezone GMT
!
```

```
Notes:

The variables below represent
company specific information.

$DNSHostAddress
$DNSHostAddress
$CompanyDomainName
$NTPHostAddress1
$NTPHostAddress2

•
```

Loopbacks and Management Interfaces

Use Table 1 below as a reference to configure the management and loopback interfaces for all spine and leaf switches. The configurations for Spine-1 are shown below. Note that a VRF is used for management.

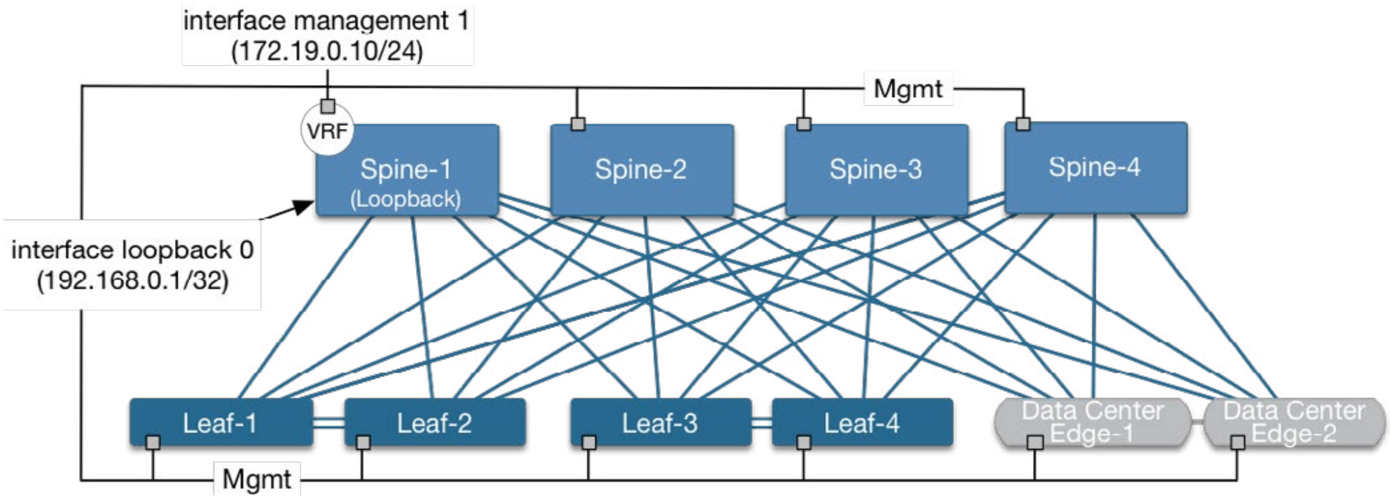


Figure 16: Loopback and Management Interfaces

```
! Spine-1 Example
!
vrf definition MGMT
     rd 0:65010
!
interface Management1
     vrf forwarding MGMT
     ip address 172.19.0.10/24
!
interface Loopback0
     description Router-ID
     ip address 192.168.0.1/32
!
```

```
Notes:

This example is for Spine-01,
see table x for the details
for other spines.
```

Table 1: Loopback and Management IP Addressing

Node	Interface	IP/Mask	Interface	IP/Mask
Spine-1	Management 1	172.19.0.10/24	Loopback 0	192.168.0.1/32
Spine-2	Management 1	172.19.0.11/24	Loopback 0	192.168.0.2/32
Spine-3	Management 1	172.19.0.12/24	Loopback 0	192.168.0.3/32
Spine-4	Management 1	172.19.0.13/24	Loopback 0	192.168.0.4/32
Leaf-1	Management 1	172.19.0.14/24	Loopback 0	192.168.0.5/32
Leaf-2	Management 1	172.19.0.15/24	Loopback 0	192.168.0.6/32
Leaf-3	Management 1	172.19.0.16/24	Loopback 0	192.168.0.7/32
Leaf-4	Management 1	172.19.0.17/24	Loopback 0	192.168.0.8/32

Point-to-Point Links

This section of the configuration will assign the appropriate addresses to the spine and leaf switches so they can reach one another via IP. Each leaf switch will utilize four 100Gbps interfaces. Each of the four leaf uplinks will connect to a separate spine and each spine will have a connection to each leaf switch. The 100G interfaces are configured as routed (no switchport) point-to-point interfaces with logging enabled for changes in link status. The ARP timeout is also set to 900 seconds to ensure stale dynamic ARP entries are removed from tables.

To identify each point-to-point link a Link # can be used to keep track of the interfaces and addresses, the Link # is not part of the configuration, just a way to keep track of the links, see Table 2 for details. Use this table as a reference when configuring the links.

Note:

- In the examples the interface speed is forced to 100G full, and the default MTU is 9214.
- Depending on your hardware platform interface names may be different.

Table 2: Addressing for Spine and Leaf Point-to-Point Links

Link #	Node	Interface	IP / Mask	Node	Interface	IP / Mask
1	Spine-1	Ethernet 3/1/1	10.10.0.1/30	Leaf-1	Ethernet 49	10.10.0.2/30
2	-	Ethernet 3/2/1	10.10.0.5/30	Leaf-2	Ethernet 49	10.10.0.6/30
3	-	Ethernet 3/3/1	10.10.0.9/30	Leaf-3	Ethernet 49	10.10.0.10/30
4	-	Ethernet 3/4/1	10.10.0.13/30	Leaf-4	Ethernet 49	10.10.0.14/30
5	Spine-2	Ethernet 3/1/1	10.10.0.17/30	Leaf-1	Ethernet 50	10.10.0.18/30
6	-	Ethernet 3/2/1	10.10.0.21/30	Leaf-2	Ethernet 50	10.10.0.22/30
7	-	Ethernet 3/3/1	10.10.0.25/30	Leaf-3	Ethernet 50	10.10.0.26/30
8	-	Ethernet 3/4/1	10.10.0.29/30	Leaf-4	Ethernet 50	10.10.0.30/30
9	Spine-3	Ethernet 3/1/1	10.10.0.33/30	Leaf-1	Ethernet 51	10.10.0.34/30
10	-	Ethernet 3/2/1	10.10.0.37/30	Leaf-2	Ethernet 51	10.10.0.38/30
11	-	Ethernet 3/3/1	10.10.0.41/30	Leaf-3	Ethernet 51	10.10.0.42/30
12	-	Ethernet 3/4/1	10.10.0.45/30	Leaf-4	Ethernet 51	10.10.0.46/30
13	Spine-4	Ethernet 3/1/1	10.10.0.49/30	Leaf-1	Ethernet 52	10.10.0.50/30
14	-	Ethernet 3/2/1	10.10.0.53/30	Leaf-2	Ethernet 52	10.10.0.54/30
15	-	Ethernet 3/3/1	10.10.0.57/30	Leaf-3	Ethernet 52	10.10.0.58/30
16	-	Ethernet 3/4/1	10.10.0.61/30	Leaf-4	Ethernet 52	10.10.0.62/30

Spine Switches

Use this example to configure the point-to-point links on the spine and leaf switches. The configuration for Spine-1 is shown in the example below.

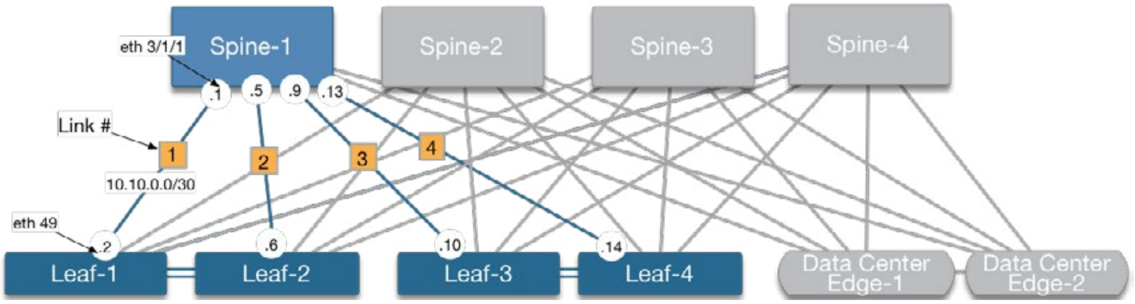


Figure 17: Configuring Point-to-Point Links

```
! Spine-1 Point-to-Point Link Configuration Example
!
interface Ethernet3/1/1
   description – P2P Link to Leaf-1
   speed forced 100gfull
   logging event link-status
   no switchport
   ip address 10.10.0.1/30
   arp timeout 900
   no shutdown
!
interface Ethernet3/2/1
   description – P2P Link to Leaf-2
   speed forced 100gfull
   logging event link-status
   no switchport
   ip address 10.10.0.5/30
   arp timeout 900
   no shutdown
!
interface Ethernet3/3/1
   description – P2P Link to Leaf-3
   speed forced 100gfull
   logging event link-status
   no switchport
   ip address 10.10.0.9/30
   arp timeout 900
   no shutdown
!
interface Ethernet3/4/1
   description – P2P Link to Leaf-4
   speed forced 100gfull
   logging event link-status
   no switchport
   ip address 10.10.0.13/30
   arp timeout 900
   no shutdown
!
```

Notes:

Configuration details are shown for Spine-1.

Configuration is for the Point-to-point interfaces that connect Spine-1 to leaf switches 1,2,3 & 4.

Leaf Switches

Use this example to configure the point-to-point links on the spine and leaf switches. The configuration for Leaf-1 is show in the example below.

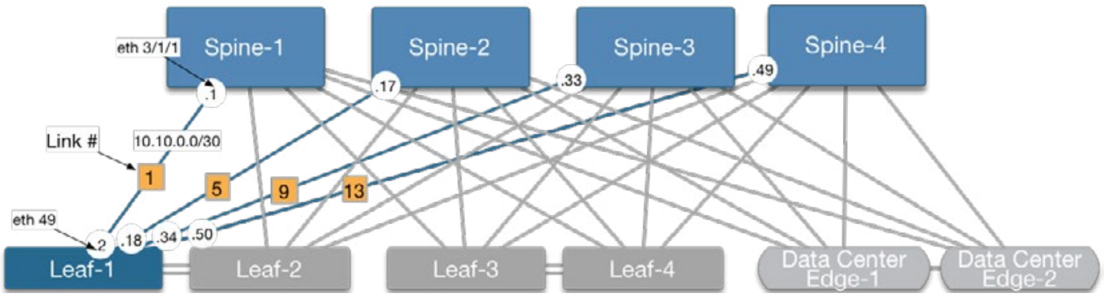


Figure 18: Configuring point-to-point links on leaf switches

```
! Leaf-1 Example
!
interface Ethernet49/1
  description - P2P Link to SPINE switch-1
  speed forced 100gfull
  mtu 9214
  logging event link-status
  no switchport
  ip address 10.10.0.2/30
  arp timeout 900
  no shutdown
!
interface Ethernet50/1
  description - P2P Link to SPINE switch-2
  speed forced 100gfull
  mtu 9214
  logging event link-status
  no switchport
  ip address 10.10.0.18/30
  arp timeout 900
  no shutdown
!
interface Ethernet51/1
  description - P2P Link to SPINE switch-3
  speed forced 100gfull
  mtu 9214
  logging event link-status
  no switchport
  ip address 10.10.0.34/30
  arp timeout 900
  no shutdown
!
interface Ethernet52/1
  description - P2P Link to SPINE switch-4
  speed forced 100gfull
  mtu 9214
  logging event link-status
  no switchport
  ip address 10.10.0.50/30
  arp timeout 900
  no shutdown
```

Notes:

This example uses the configuration specifics for Leaf-1.

Base Routing & BGP Configuration

The diagram below depicts the Autonomous System Numbering scheme that will be used in the configuration examples. Note that this example uses the Common Spine – Common Leaf AS numbering scheme.

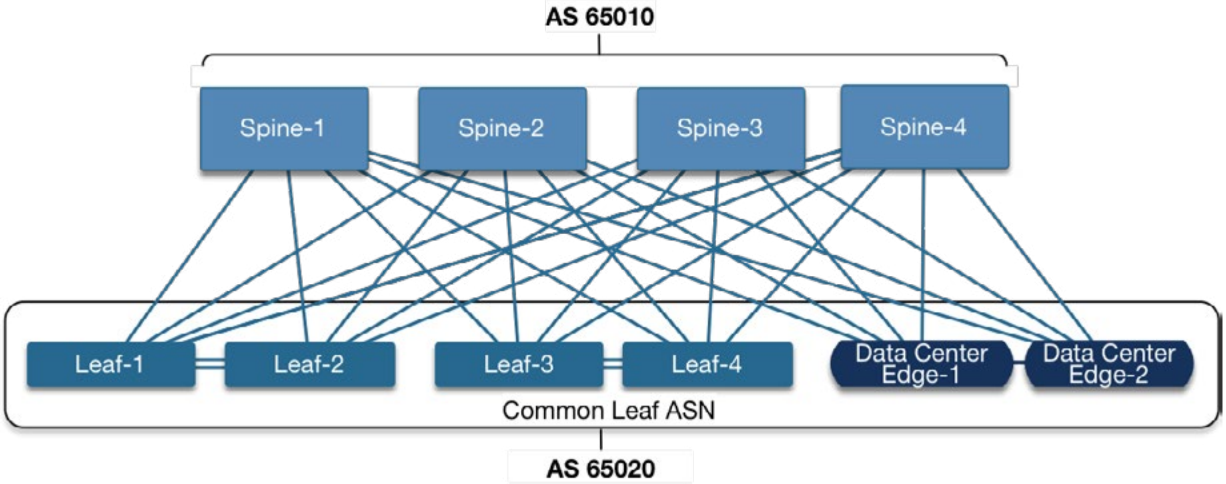


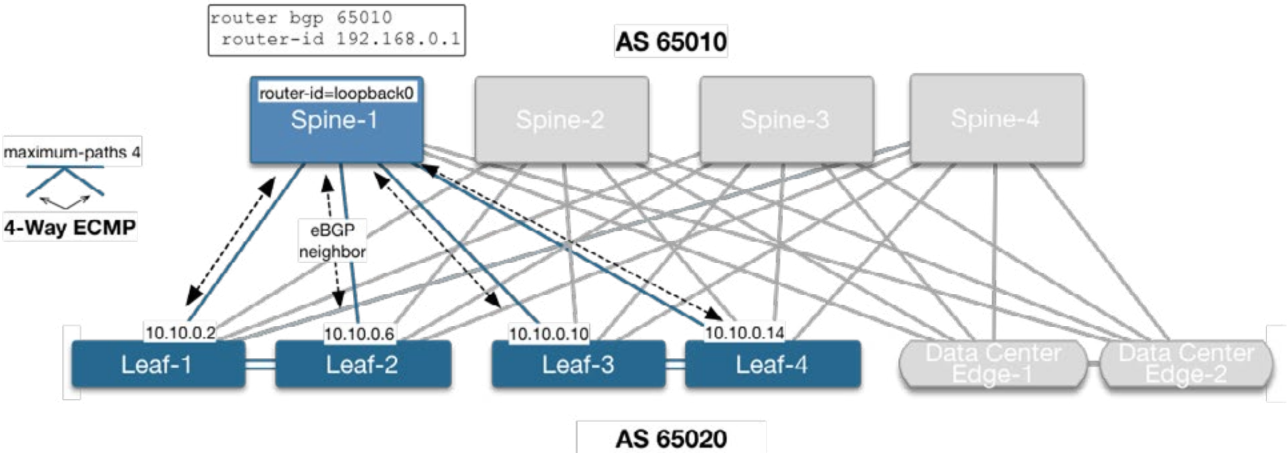
Figure 19: ASN Number Scheme

Spine Configuration (BGP)

To configure Border Gateway Protocol (BGP) IP Routing must first be enabled on the device. For the spine configurations the default BGP distance is altered to give preference to external BGP routes. Leaf neighbors are also defined and utilize a peer-group to simplify configuration.

Perform the following configuration on each spine switch. Note that all spine switches share a common ASN in this example, see diagram for details. The loopback 0 address will be used as the router-id.

The example below uses static BGP peer groups. A static BGP peer group is a collection of BGP neighbors, which can be configured as a group. Once a static peer group is created, the group name can be used as a parameter in neighbor configuration commands, and the configuration will be applied to all members of the group. Neighbors added to the group will inherit any settings already created for the group. Static peer group members may also be configured individually, and the settings of an individual neighbor in the peer group override group settings for that neighbor. When the default form of a BGP configuration command is entered for a member of a static peer group, the peer inherits that configuration from the peer group.



```
router bgp 65010
 router-id 192.168.0.1
```

Figure 20: BGP Spine Configuration

```
! Spine-1 Example
!
ip routing
no ip routing vrf MGMT
!
router bgp 65010
   router-id 192.168.0.1
   bgp log-neighbor-changes
   distance bgp 20 200 200
   maximum-paths 4 ecmp 64
   neighbor EBGP-TO-LEAF-PEER peer-group
   neighbor EBGP-TO-LEAF-PEER remote-as 65020
   neighbor EBGP-TO-LEAF-PEER maximum-routes 12000
   neighbor 10.10.0.2 peer-group EBGP-TO-LEAF-PEER
   neighbor 10.10.0.6 peer-group EBGP-TO-LEAF-PEER
   neighbor 10.10.0.10 peer-group EBGP-TO-LEAF-PEER
   neighbor 10.10.0.14 peer-group EBGP-TO-LEAF-PEER
   network 192.168.0.1/32
!
```

```
Notes:

Spine ASN = 65010
Leaf ASN = 65020

Leaf-1 = 10.10.0.2
Leaf-2 = 10.10.0.6
Leaf-3 = 10.10.0.10
Leaf-4 = 10.10.0.14
```

Below is a simplified configuration that does not use peer-groups.

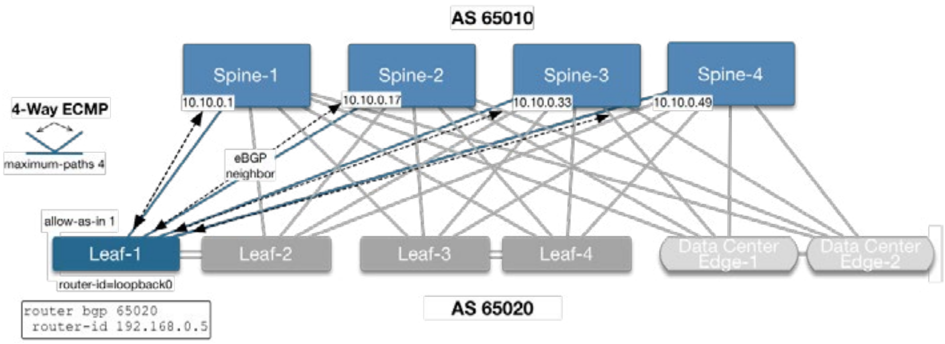
```
! Spine-1 Simple Example
!
router bgp 65010
   router-id 192.168.0.1
   bgp log-neighbor-changes
   distance bgp 20 200 200
   maximum-paths 4 ecmp 64
   maximum-routes 12000
   neighbor 10.10.0.2 remote-as 65020
   neighbor 10.10.0.6 remote-as 65020
   neighbor 10.10.0.10 remote-as 65020
   neighbor 10.10.0.14 remote-as 65020
   network 192.168.0.1/32
!
```

```
Notes:

Note this is just another example;
there is a single neighbor
statement for each leaf switch.
```

Leaf Configuration (BGP)

The leaf switch configuration is very similar to the spine BGP configuration. Routing must be enabled on the device and distance is again altered to make sure the switch prefers external BGP routes. A single peer-group is utilized to peer with the spine with a standard configuration. The allowas-in feature is also configured.



```
router bgp 65020
 router-id 192.168.0.5
```

Figure 21: BGP Leaf Configuration

```
! Leaf-1 Example
!
ip routing
no ip routing vrf MGMT
!
router bgp 65020
   router-id 192.168.0.5
   bgp log-neighbor-changes
   distance bgp 20 200 200
   maximum-paths 4 ecmp 4
   neighbor EBGP-TO-SPINE-PEER peer-group
   neighbor EBGP-TO-SPINE-PEER remote-as 65010
   neighbor EBGP-TO-SPINE-PEER allowas-in 1
   neighbor EBGP-TO-SPINE-PEER maximum-routes 12000
   neighbor 10.10.0.1 peer-group EBGP-TO-SPINE-PEER
   neighbor 10.10.0.17 peer-group EBGP-TO-SPINE-PEER
   neighbor 10.10.0.33 peer-group EBGP-TO-SPINE-PEER
   neighbor 10.10.0.49 peer-group EBGP-TO-SPINE-PEER
   network 192.168.0.5/32
   redistribute connected
!
```

```
Notes:

Spine ASN = 65010
Leaf ASN = 65020

Spine-1=10.10.0.1  -Lnk 1
Spine-2=10.10.0.17 -Lnk 5
Spine-3=10.10.0.33 -Lnk 9
Spine-4=10.10.0.49 -Lnk 13

Loopback = 192.168.0.5
```

Route Advertising (BGP)

To ensure that only the proper routes are advertised from leaf switches, a route map must be applied to the BGP peer. The route map references the prefix-list which contains the routes that are intended to be advertised to the spine. Inside the BGP configuration, the route map is applied to the neighbors using the peer-group in the "out" direction. Connected routes are redistributed to the spine switches to advertise local subnets and interfaces.

Although not mandatory, using a route-map provides a level of protection in the network. Without a route-map random networks could be created at the leaf, which would automatically be added to the data center's route table. This configuration should be done at the leaf switch. A statement for each network requiring routing needs to be added to the prefix-list.

```
! Leaf-1 Example
!
route-map ROUTE-MAP-OUT permit 10
   match ip address prefix-list PREFIX-LIST-OUT
!
ip prefix-list PREFIX-LIST-OUT seq 10 permit 192.168.0.5/32
ip prefix-list PREFIX-LIST-OUT seq 20 permit 10.10.0.0/16
ip prefix-list PREFIX-LIST-OUT seq 30 permit 172.20.0.0/24
ip prefix-list PREFIX-LIST-OUT seq 40 permit 172.30.0.0/24
!
router bgp 65020
   redistribute connected
   network 192.168.0.5/32
   neighbor EBGP-TO-SPINE-PEER route-map ROUTE-MAP-OUT out
!
end
```

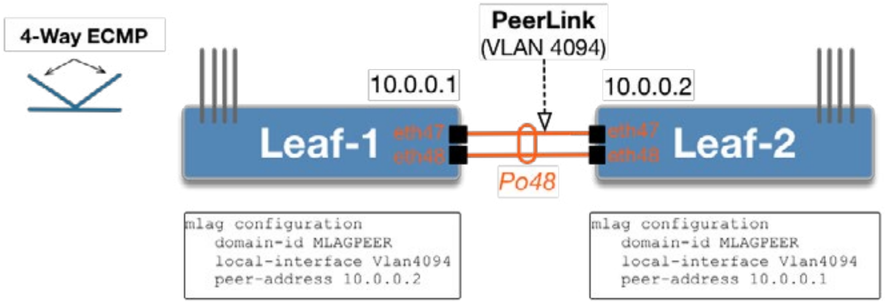
```
Notes:

Loopback = 192.168.0.5
VLAN 200 = 172.20.0.0/24
VLAN 300 = 172.30.0.0/24
```

MLAG Configuration - Dual-Homed Compute Leaf

The Multi-Chassis Link Aggregation Group configuration must be identical on both MLAG peers. The MLAG peer VLAN must be created and added to the MLAGPEER trunk group. The MLAG peers also must have IP reachability with each other over the peer link.

To ensure forwarding between the peers on the peer link, spanning-tree must also be disabled. Once the port channel is created for the peer link and configured as a trunk port, additional VLANs may be added if necessary to transit the peer link. The MLAG configuration consists of four main pieces of information: the MLAG domain must be unique for each MLAG pair, the local interface for IP reachability (VLAN 4094 in this example) and the physical peer link (Port Channel 48 in this example).



```
mlag configuration
   domain-id MLAGPEER
   local-interface Vlan4094
   peer-address 10.0.0.2
```

```
mlag configuration
   domain-id MLAGPEER
   local-interface Vlan4094
   peer-address 10.0.0.1
```

Figure 22: MLAG Configuration for Dual-Homed Compute

```
! MLAG Configuration for Leaf-1
!
interface Port-Channel48
   description – L2 Trunk Link to Leaf-2
   logging event link-status
   switchport trunk allowed vlan 100,200,4094
   switchport mode trunk
   switchport trunk group MLAGPEER
   qos trust dscp
   no spanning-tree portfast auto
   spanning-tree portfast network
   no shutdown
!
interface Ethernet47
   description -L2 Trunk Link to Leaf-2
   mtu 9214
   no logging event link-status
   no snmp trap link-status
   channel-group 1 mode active
   lacp port-priority 16000
   no shutdown
!
interface Ethernet48
   description -L2 Trunk Link to Leaf-2
   mtu 9214
   no logging event link-status
   switchport mode trunk
   switchport trunk allowed vlans 100,200,4094
   no snmp trap link-status
   channel-group 1 mode active
   lacp port-priority 16000
   storm-control broadcast level 1
   no shutdown
!
```

```
! MLAG Configuration for Leaf-2
!
interface Port-Channel48
   description – L2 Trunk Link to Leaf-1
   logging event link-status
   switchport trunk allowed vlan 100,200,4094
   switchport mode trunk
   switchport trunk group MLAGPEER
   qos trust dscp
   no spanning-tree portfast auto
   spanning-tree portfast network
   no shutdown
!
interface Ethernet47
   description -L2 Trunk Link to LEAF switch-1
   mtu 9214
   no logging event link-status
   no snmp trap link-status
   channel-group 1 mode active
   lacp port-priority 32000
   no shutdown
!
interface Ethernet48
   description -L2 Trunk Link to LEAF switch-1
   mtu 9214
   no logging event link-status
   switchport mode trunk
   switchport trunk allowed vlans 100,200,4094
   no snmp trap link-status
   channel-group 1 mode active
   lacp port-priority 32000
   storm-control broadcast level 1
   no shutdown
!
```

```
interface Vlan4094
   description InterSwitch_MLAG_PeerLink
   mtu 9214
   ip address 10.0.0.1/30
   arp timeout 900
   no shutdown
!
mac address-table aging-time 1200
!
ip virtual-router mac-address 00:1c:73:00:00:99
!
mlag configuration
   domain-id MLAGPEER
   local-interface Vlan4094
   peer-address 10.0.0.2
   peer-link Port-Channel1
   reload-delay 60
!
```

```
interface Vlan4094
   description InterSwitch_MLAG_PeerLink
   mtu 9214
   ip address 10.0.0.2/30
   arp timeout 900
   no shutdown
!
mac address-table aging-time 1200
!
ip virtual-router mac-address 00:1c:73:00:00:99
!
mlag configuration
   domain-id MLAGPEER
   local-interface Vlan4094
   peer-address 10.0.0.1
   peer-link Port-Channel1
   reload-delay 60
!
```

IBGP Configuration for the MLAG Configuration

In general a standard MLAG configuration is used with a couple of additions to support eBGP peering to the spine as well as iBGP peering between the MLAG peers. When integrating an MLAG Leaf configuration into a Layer 3 Leaf Spine, iBGP peering is recommend between the MLAG peers. The reason for the peering is due to specific failure conditions that the design must take into consideration. In normal operation paths learned via eBGP (leaf to spine uplinks) will always be preferred over paths learned via iBGP (given other tie-breaker criteria are equal) this is desirable.

The following configuration will enable iBGP between the MLAG peers. Notice the ASN is that of the peer switch and not the Spine. The neighbor next-hop-self command configures the switch to list its address as the next hop in routes that it advertises to the specified BGP-speaking neighbor or neighbors in the specified peer group. This is used in networks where BGP neighbors do not directly access all other neighbors on the same subnet.

```
! iBGP Configuration for MLAG Leaf-1
!
router bgp 65020
   neighbor 10.0.0.2 remote-as 65020
   neighbor 10.0.0.2 next-hop-self
   neighbor 10.0.0.2 allowas-in 1
   neighbor 10.0.0.2 maximum-routes 12000
!
```

```
! iBGP Configuration for MLAG Leaf-2
!
router bgp 65020
   neighbor 10.0.0.1 remote-as 65020
   neighbor 10.0.0.1 next-hop-self
   neighbor 10.0.0.1 allowas-in 1
   neighbor 10.0.0.1 maximum-routes 12000
!
```

Server Connectivity (VLANs and Gateways)

General configuration examples for connecting servers into the leaf switches are shown below. A more general-purpose single-homed compute configuration as well as a dual-homed (active/active) compute configuration is also shown. Note that server level configuration always needs to be reviewed as well, particularly with dual-homed active/active configurations.

Single-Homed Leaf Configuration

An example of VLAN and Gateway configuration for a single-homed compute leaf is shown below. Notice ports are configured as access ports and assigned a VLAN. A Switched Virtual Interface (SVI) is created for each VLAN, which acts as the default gateway for the host/workload. LLDP is also enabled on the switch to support learning of neighbor information.

```
! Leaf-1 Single-Homed Server VLAN Configuration
!
lldp run
!
vlan 100
   name SERVER-VLAN-172.20.0.0/24
!
vlan 200
   name SERVER-VLAN-172.30.0.0/24
!
interface Vlan100
   description SVI-FOR-VLAN-100
   mtu 9214
   ip address 172.20.0.1/24
   arp timeout 900
   no shutdown
!
interface Vlan200
   description SVI-FOR-VLAN-200
   ip address 172.30.0.1/24
   arp timeout 900
   no shutdown
!
interface Ethernet 21
   switchport access vlan 100
   no snmp trap link-status
   storm-control broadcast level 1
   spanning-tree portfast
   spanning-tree bpduguard enable
   no shutdown
!
interface Ethernet 22
   switchport access vlan 200
   no snmp trap link-status
   storm-control broadcast level 1
   spanning-tree portfast
   spanning-tree bpduguard enable
   no shutdown
!
```

Dual-Homed Leaf Configuration

Below is the VLAN and port configuration for Leaf-1 and Leaf-2. Note that Leaf-1 and Leaf-2 are MLAG peers and will share a common configuration. Default gateway configuration uses the system VARP address. Both access and trunk port configurations are show. LLDP configuration is also shown in this example.

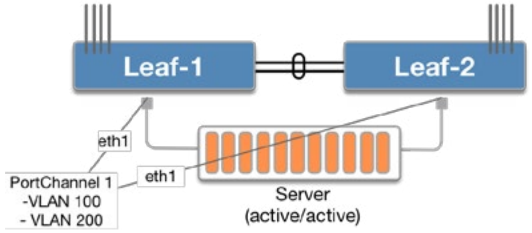


Figure 23: Dual-Homed Leaf Configuration for Server Connectivity

```
! Leaf-1 Server VLAN Configuration
!
lldp run
!
vlan 100
   name SERVER-VLAN-172.20.0.0/24
!
vlan 200
   name SERVER-VLAN-172.30.0.0/24
!
interface Vlan100
   description SVI-FOR-VLAN-100
   mtu 9214
   ip address 172.20.0.2/24
   ip virtual-router address 172.20.0.1/24
   arp timeout 900
   no shutdown
!
interface Vlan200
   description SVI-FOR-VLAN-200
   ip address 172.30.0.2/24
   ip virtual-router address 172.30.0.1/24
   arp timeout 900
   no shutdown
!
interface Port-Channel1
   switchport access vlan 100
   no snmp trap link-status
   port-channel lacp fallback static
   mlag 1
   storm-control broadcast level 1
   spanning-tree portfast
   spanning-tree bpduguard enable
   no shutdown
!
interface Ethernet 1
   description MLAG-To-Server
   channel-group 1 mode active
   no shutdown
!
interface Port-Channel2
   switchport access vlan 200
   no snmp trap link-status
   port-channel lacp fallback static
   mlag 2
   storm-control broadcast level 1
   spanning-tree portfast
   spanning-tree bpduguard enable
   no shutdown
!
interface Ethernet 2
   description MLAG-To-Server
   channel-group 2 mode active
   no shutdown
!
```

```
! Leaf-2 Server VLAN Configuration
!
lldp run
!
vlan 100
   name SERVER-VLAN-172.20.0.0/24
!
vlan 200
   name SERVER-VLAN-172.30.0.0/24
!
interface Vlan100
   description SVI-FOR-VLAN-100
   mtu 9214
   ip address 172.20.0.3/24
   ip virtual-router address 172.20.0.1/24
   arp timeout 900
   no shutdown
!
interface Vlan200
   description SVI-FOR-VLAN-200
   ip address 172.30.0.3/24
   ip virtual-router address 172.30.0.1/24
   arp timeout 900
   no shutdown
!
interface Port-Channel1
   switchport access vlan 100
   no snmp trap link-status
   port-channel lacp fallback static
   mlag 1
   storm-control broadcast level 1
   spanning-tree portfast
   spanning-tree bpduguard enable
   no shutdown
!
interface Ethernet1
   description MLAG-To-Server
   channel-group 1 mode active
   no shutdown
!
interface Port-Channel2
   switchport access vlan 200
   no snmp trap link-status
   port-channel lacp fallback static
   mlag 2
   storm-control broadcast level 1
   spanning-tree portfast
   spanning-tree bpduguard enable
   no shutdown
!
interface Ethernet 2
   description MLAG-To-Server
   channel-group 2 mode active
   no shutdown
!
```

List of Acronyms

AS - Autonomous System

ASN - Autonomous System Number

BFD - Bidirectional Forwarding Detection BGP - Border Gateway Protocol

eBGP - External Border Gateway Protocol ECMP - Equal Cost Multi-Path

EGP - Exterior Gateway Protocol

iBGP - Internal Border Gateway Protocol IETF - Internet Engineering Task Force

L2 - Layer 2

L3 - Layer 3

MLAG - Multi-chassis Link Aggregation RFC - Request for Comment

RIB - Routing Information Base

VARP - Virtual Address Resolution Protocol VXLAN - Virtual eXtensible LAN

References

"A Border Gateway Protocol 4 (BGP-4)", https://tools.ietf.org/pdf/rfc4271.pdf

"Use of BGP for routing in large-scale data centers", https://tools.ietf.org/pdf/draft-ietf-rtgwg-bgp-routing-large-dc-07.pdf

Santa Clara—Corporate Headquarters
5453 Great America Parkway,
Santa Clara, CA 95054

Phone: +1-408-547-5500
Fax: +1-408-538-8920
Email: info@arista.com

Ireland—International Headquarters
3130 Atlantic Avenue
Westpark Business Campus
Shannon, Co. Clare
Ireland

Vancouver—R&D Office
9200 Glenlyon Pkwy, Unit 300
Burnaby, British Columbia
Canada V5J 5J8

San Francisco—R&D and Sales Office
1390 Market Street, Suite 800
San Francisco, CA 94102

India—R&D Office
Global Tech Park, Tower A & B, 11th Floor
Marathahalli Outer Ring Road
Devarabeesanahalli Village, Varthur Hobli
Bangalore, India 560103

Singapore—APAC Administrative Office
9 Temasek Boulevard
#29-01, Suntec Tower Two
Singapore 038989

Nashua—R&D Office
10 Tara Boulevard
Nashua, NH 03062