Arista Universal Cloud Network

Version 2.0

June 2018

About This Document

The Arista Universal Cloud Network Design Guide is based upon common use cases seen from real customers. The UCN Design guide shows a set of solutions, features, and applications that are leveraged to meet the customer's demands.

These designs have been repeatedly tested over the course of time in many different customer and lab installations to ensure their stability and ability to meet the needs of our customers. This design guide provides information concerning Arista Networks technology solutions.

For the most recent design guide information, please see www.arista.com

Table of contents

Table of contents

Table of contents

Table of contents

Table of contents

Arista Universal Cloud Network Overview

Building a modern datacenter to support your critical business applications requires a different approach from that of the legacy enterprise computing network model. The next generation datacenter requires an agile and scalable network infrastructure which supports the ability to rapidly deploy and expand or contract the IT infrastructure needed to run business applications. Next generation enterprise datacenters are highly virtualized and built to provide workload mobility and multi-tenancy. The virtualized environment will benefit from a robust IP fabric where workloads can be placed anywhere in the context of the datacenter. While 1G connectivity is still prevalent in many older environments next generation datacenters are moving to 10/25GbE for compute and storage devices with the ability to quickly scale to 40/50GbE, and even 100GbE without "forklifting" the existing network infrastructure. Arista's Universal Cloud Network Design will highlight and demonstrate differentiators in the areas of:

Available Architecture: Delivering a self-healing architecture of quality across a highly available leaf-spine network with link, path, device and network wide redundancy. Our single binary EOS image works across all platforms and enables more efficient code quality and certification.

Agile Workloads: Legacy networks are typically unaware of micro-services such as workloads, work-streams, or workflows. Arista has the ability to deploy fewer devices with greater port density, thus reducing power, space and DC overhead. This architecture also enables a single IP Fabric for greater efficiency across the entire infrastructure based on open standards giving customers the ultimate choice to choose the best product to meet the business needs. This approach will allow customers to continue to optimize their datacenters and ultimately extend the life cycle.

Analytics: Tracing the workflow information across the different domains enables the ability to quickly pinpoint problems through telemetry tracers that abstract the actionable metadata state for dynamic correlation. Arista can correlate system-wide telemetry coupled with Artificial Intelligence (AI) and Machine Learning (MI) enabled systems to have a positive impact on network availability.

Automation: Arista leverages lessons learned from our cloud customers to bring a new level of business agility to the enterprise. It is through features such as Zero Touch Provisioning and Zero Touch Replacement that Arista is able to provide an unprecedented amount of business agility by allowing the network to be integrated into the overall datacenter solution, not just a networking silo. Automation with CloudVision brings a wealth of provisioning/orchestration to the customer.

AnyCloud API: Being open and programmable with full API support allows for deep integration with existing network management systems and other public cloud providers.

The foundation of this design guide is based on the capabilities and features of Arista's Extensible Operating Systems (EOS), the world's most advanced network operating system. EOS was purpose built for the next-generation datacenter. EOS is a highly modular software design based on a unique multi-process state sharing architecture that separates networking state from the processing itself. This architecture enables fault recovery and incremental software updates on very granular level, without affecting the state of the system. Arista EOS provides an extremely robust, stable and resilient network-operating environment while delivering on the need for openness, software modularity and extensibility. This unique combination offers the opportunity to significantly improve the functionality as part of the evolution of next generation datacenters.

The industry standard EOS command-line interface (CLI) avoids re-training costs. Arista's single binary image for all platforms enables more efficient code certification and time to market for new features and enhancements, as well as bug fixes. This significantly reduces the operational overhead when compared to other operating systems with multiple code release trains. Further, EOS may be extended through third-party software to enable tight integration with any existing in-house network management systems (NMS) and the automation of routine maintenance tasks, including deployment, monitoring, maintenance, and upgrades. This work can be provided through Arista's EOS Consulting Services organization; which can be deployed to provide assistance with integration and customization.

Unique to the industry, Arista has committed to using merchant silicon from multiple suppliers across its portfolio. This market is highly competitive and diversified, ensuring Arista has a range of choices for cutting edge, best-of-breed silicon, and also avoid having to rely on single-source approaches. Prior to Arista Networks, no other networking company has driven its product portfolio towards a path that follows Moore's Law in a way only previously experienced in general purpose computing. With single-chip designs now providing 64x100GbE ports in a 2RU form factor while offering the flexibility to be configured as 128 10GbE or 128 25GbE interfaces, Arista is changing the game and the economics of Ethernet in the datacenter.

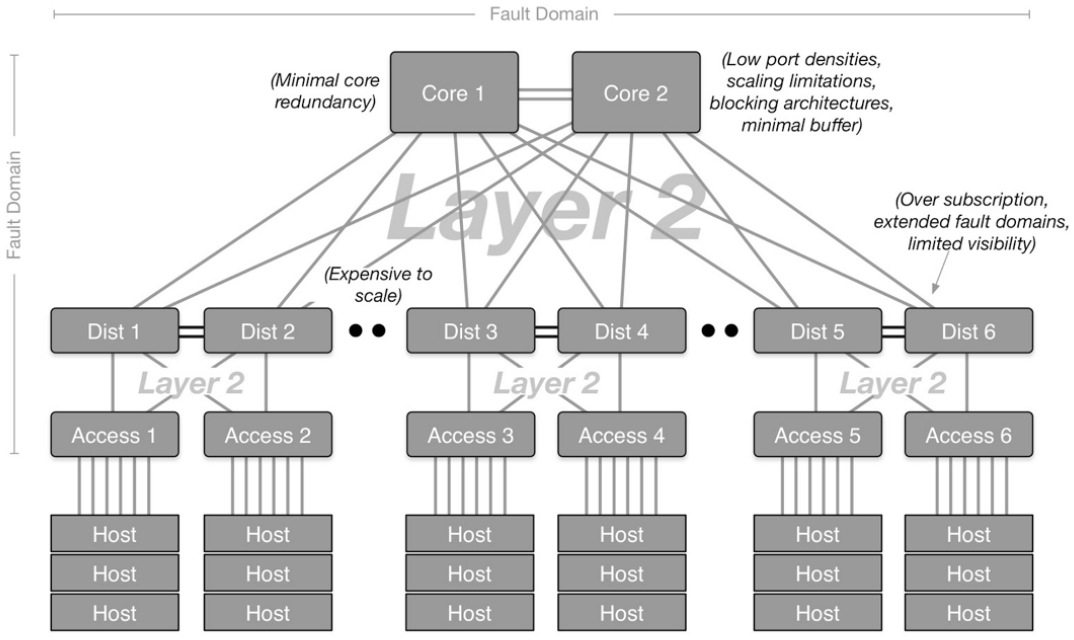
Arista builds best of breed datacenter switches; from ultra low-latency, high density top-of-rack, to the highest performing and feature rich modular chassis switch in the market. This means that customers can trust Arista to deliver market leading platforms, with the highest port density, lowest form factor and most efficient power footprint in the market today and in the future. This guarantees customers can remain at the cutting edge of technology and innovation and achieve the most efficient network.

Arista's commitment to open standards based, highly available, datacenter networks provides customers with the flexibility and resiliency required to simplify the overall infrastructure while minimizing operational costs. Arista's ability to work within the datacenter ecosystem and partner with market leading vendors in SDN, storage, security, virtualization, management and automation, ensures the customer can continue to adopt best of breed solutions without the risk of vendor "lock-in".

By partnering with Arista, customers can expect to benefit from a reduction in operational and capital costs, reduced power requirements as well as a reduced datacenter footprint all while increasing your ability to scale, in an agile way, to meet the business needs.

Traditional Datacenter Network Designs

The traditional approach to building both campus and datacenter networks has been to deploy the well known hierarchical 3-tier topology. The diagram below shows a representation of this architecture.



Traditional Multi-Tiered Models

The general three tier model is comprised of access switches, aggregated across pairs of distribution switches, which in turn uplink to a single pair of core layer devices. This traditional design was the de-facto standard for over 10+ years and was optimized for North-South traffic patterns and non-virtualized workloads. Fast forward to the present day and the demand on the networking infrastructure has changed dramatically. North-South traffic patterns have shifted to East-West due to of the emergence of virtualization and containerized workloads. This shift has made it very apparent that Traditional 3-tier designs were incapable of the increased demand on the network.

The Limitations with LEGACY Designs

Traditional 3-tier designs struggle with today's demands because of the following limitations:

Oversubscription

Oversubscription is expressed as a ratio of contention should all devices on a switch send traffic to a fixed number of uplinks ports at the same time. An example of this could be a 48 Port 25G switch with six 100G uplinks. In this scenario the over subscription ratio would be 1200G:600G or 2:1. Oversubscription can exist in the North-South direction (traffic entering/leaving a datacenter) as well as East-West (traffic between devices in the datacenter). Traditional 3-tier data designs have very large oversubscription ratios, upwards of 20:1, for both North-South and East-West traffic. This was primarily due to outdated networking technologies not keeping pace with Moore's law.

Monitoring

Microbursts caused by traffic patterns such as those seen in TCP in-cast situations (a many-to-one traffic pattern) are missed due to traditional monitoring protocols and methods that rely on polling, like the Simple Network Management Protocol (aka SNMP). Polling intervals, even those that are aggressive, miss sub-second bursts of traffic. This is due to the fact that these very quick, link saturating bursts are too short-lived to affect the averages being reported via SNMP. When left undetected, microbursts can increase latency and ultimately result in packet discards that require applications to retransmit payloads. This inevitably will impact application performance. For today's datacenter, it is extremely important to detect and proactively mitigate issues in the network with tools that can provide better granularity.

Latency

Traditional 3-tier designs were not built with latency sensitive applications in mind. This was primarily because North-South traffic patterns prevailed in the datacenter and applications were built to tolerate high end to end latency. Today, East-West traffic patterns prevail and are driven by disaggregated application workloads. Extra switch hops and queuing from heavily oversubscribed traditional designs introduce unwanted latency which is directly hinders to application performance.

Workload Mobility

Virtual Machine (VM) mobility and legacy enterprise applications regularly require cluster members to reside on the same Layer 2 segment. Providing this type of reachability with traditional 3-tier designs was accomplished by proprietary "lock-in" technologies or separately managed Layer 2 networks. These "lock-in" technologies and bad design practices scale poorly and commonly provide little to no visibility when isolating issues in the network. Datacenters today need the ability to provide the needed reachability without the worrying about increasing the blast radius of the Spanning-Tree Protocol influence and scale past 4096 Layer-2 virtual segments.

Scalability

Traditional 3-tier architectures relied on adding additional tiers when new growth would push the oversubscription ratios past acceptable levels. These additional tiers add complexity, cost and latency to the network. With networking silicon now following Moore's law, oversubscription ratios can be maintained without the need to "scale-up". This promotes "scale-out" architectures that can maintain low oversubscription ratios and support the largest datacenter operators scale. Traditional network designs are no longer practical or sustainable for the demands of today's datacenter.

Management

Managing and operating small to medium datacenter networks can be laborious and cumbersome. Traditional networking operating systems weren't built for automation and software driven architectures. They've historically lacked consistent and reliable APIs, forcing operators to resort to screen-scraping automation that is difficult to support and maintain. Datacenters today need operating systems that present reliable APIs across every platform. Whether the deployment use case call for virtual, traditional hardware or cloud based solutions, the same programmatic interfaces should be used to manage the network. A networking operating system that is open and extensible empowers datacenter operators to streamline their operations and meet demand quickly without the need to increase staffing.

Automation

Practically all processes in legacy datacenter networks required manual intervention. Staging, testing and validating changes in the network was prone to human error and hard to replicate because of the scale of the network. Today's datacenter operators need the ability to automate their workflows into a continuous integration pipeline. This requires an operating system that supports the same tool sets in any scenario. For instance, spinning up a virtual replica of the production network as part of a change control workflow gives the operator the ability to confidently stage, test and validate all network changes in a safe environment before rolling into production. Supporting next generation workflows is just one part of automation, next generation operating systems must also have friendly APIs that promote a 3rd party integrations and allow for a robust partner ecosystem.

Arista Universal Cloud Network (UCN) – The Next Generation Datacenter Network

Due to new and evolving datacenter technologies, including changes in application development, it is clear to see that traffic patterns within the datacenter and out of the datacenter have changed with them over the years. Of these changes, the most notable are general Server Virtualization, Application Containterization,, Multi-Cloud Computing, Web 2.0, Big Data and High Performance Computing (HPC), and more recently AI and machine learning. To optimize and increase the performance of these new technologies, a distributed scale-out, deep-buffered IP fabric has been proven to provide consistent performance that scales to support extreme 'East-West' traffic patterns. To guarantee uninterrupted application performance, an ideal fabric should be non-blocking and provide a scalable network infrastructure to all hosts and applications.

Guiding Principles for Arista UCN designs

Simplicity

A key goal of any good design should be simplicity. A simpler modern datacenter network design can be automated easier, reducing the potential for human error and minimizing outages. Common practices such as broadly spanned VLANs, dense VRF implementations, and proprietary Layer-2 multi-pathing technologies result in overly complicated network designs. To achieve the goal of simplicity, Arista advocates collapsing the datacenter network into two tiers: a Spine Tier(s) and a Leaf Tier. Spine to Leaf connectivity can be via Layer 3 or Layer 2 and should be implemented utilizing open, standards based protocols such as OSPF and BGP for a layer 3 implementation or LACP for a layer 2 implementation.

Open

Arista is a strong believer of customer's ability to select vendors freely to address technical and business outcomes. Customers should not be forced or 'locked into' a particular vendor's products or proprietary protocols. In an open, standards based design, the chosen topology and equipment should be fully interchangeable allowing any leaf or spine switch to be replaced with any other vendor's products without significant redesigns or adverse impact to business.

Investment Protection

Building a future proof datacenter network is comprised of many different components such as system lifecycle and upgradeability, system scale, and a commitment to continuous innovation for both hardware and software elements of the system.

A modular system should have a 7-10 year operating lifecycle and support multiple generations of interoperable line cards within a common chassis. This approach ensures that a customer of that system can continue to use it while transitioning from 10Gb to 25Gb, 40Gb, 50Gb, 100Gb and 400Gb Ethernet at useable densities that continue to meet the scaling needs of the network, applications and the business. Another key element of investment protection is versatility of the hardware, software, and network design to seamlessly support any workload, may it reside on-prem or in the cloud.

Scale

A system is typically measured against power and cooling capabilities of the facility the system operates within; the compute and storage densities have been doubling every few years. Additionally, new facilities are often built adjacent to existing ones to facilitate customer network interconnectivity and the ability to leverage common Internet peering points. A highly scalable network should be designed to support hitless upgrades and be able to double its capacity to support compute and storage connections. Scaling out beyond a simple doubling should be possible without adding incremental 'tiering' to the network, ensuring an optimum and linear cost model for scaling up host densities. A modern Universal Cloud Network must also have the ability to scale into multiple clouds via seamless instantiation of virtual network elements.

Automation

As IT evolves and Devops practices mature, the "if can you deploy it, you can automate it" has become the new normal. Functions like augmenting capacity, software upgrades, change control, or troubleshooting issues such as network congestion should be automated through a simple software-to-machine interface. While most network operators teams are comfortable with the CLI, customers should focus on identifying time-consuming and repetitive workflows that can be automated. Automation will reduce the operating burden allowing time to focus on high value activities such as trouble ticket resolution and striving to improve uptime.

Automation is an important element to reducing the overall operational expense of the network infrastructure. Within web-scale and cloud-titan organizations, Arista is witnessing deployments that are highly automated, where little or zero human interaction is necessary to build and configure network components. Another common trend is CLI access to network equipment is being limited to troubleshooting. All changes are programmatically implemented with change management tools and scripts. While automation doesn't happen overnight, it is important to develop a strategy which embraces automated network change control and orchestration approaches. Automation also entails that as network paradigm shifts from places in the network to places in the cloud, the Universal Cloud Network must support the ability to be deployed in any Cloud utilizing any API.

To aid in this automation journey Arista provides an API (JSON over HTTP/HTTPS) for automation solutions with similar commands to those used in the CLI. This leverages the network engineers' knowledge and paves a path to a future automated state. Please see Appendix A for more information on automation capabilities.

Advanced Telemetry

Networks have often been 'black boxes' where traffic flow was expected to ingress and egress correctly. However, most vendors do not provide the instrumentation necessary to accurately and quickly diagnose network problems, collect historical data, or provide visibility into the traffic to verify accuracy of the flow and physical connectivity. A network must have the ability to recognize when it is experiencing periods of congestion and be able to send a copy of congestion inducing traffic to a monitoring tool. When an interface goes 'down', monitoring tools should provide affected workloads immediately with cross-system event correlation and real-time analytics to diagnose or resolve problem. A miscabling or misconfiguration of the network should be detected and generate an alert and or alarm, not just a re-convergence. Any interface's traffic on the network should be able to be tapped and sent back to a monitoring tool in seconds with only software provisioning, without requiring new cable runs or equipment moves. The network needs to be truly designed to support real-world operations. Please see Appendix B for more information on Telemetry capabilities.

Highly Available, Reliable, and Resilient Network Topology

A highly resilient, reliable, and available network topology is built by providing a modern resilient control plane for the topology that delivers rapid convergence with the most demanding scale. The control plane is a function of the underlying operating system running on the hardware. A modern network operating system should deliver the ability to transparently restart processes, easily extend capabilities to customer's needs, and provide a base platform that is widely adopted and curated with a large technical community.

The control plane must also allow security vulnerabilities to be patched without impacting availability and performance. Patching allows operators to stay on the same version of supported code to maximize the investment in code qualification. This ensures that any potential vulnerabilities are mitigated in a timely manner and reduces the risk of potential outages, enhancing network availability and minimizing adverse business impact.

In addition to a scalable and modular control plane, the network operating system must interact with the underlying hardware in a reliable and highly resilient manner to achieve a blended 99.999% or five 9's availability.

The last design consideration in a resilient, reliable, and available network topology is the redundancy and coordination of the different system elements in the topology. This can include redundant routing paths to forward around failed systems as well as the ability to connect two independent systems virtually together to provide an active/active topology, without merging the two systems into a single point of failure.

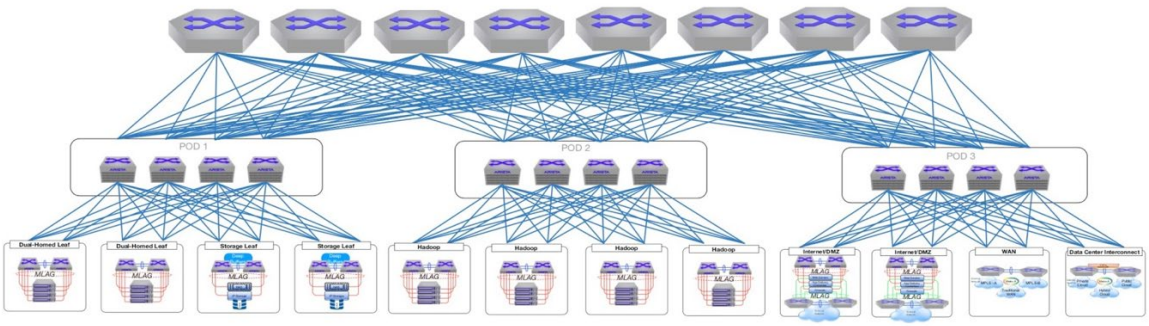
Arista is compelling the industry to re-evaluate the network paradigm of stationary places in the network to encompassing and transitioning to places in the cloud. The 5As principles highlighted above entail that (1) network must be highly available, (2) seamlessly support any workload, (3) have deep analytics, (4) the network OS and hardware should allow for high-degree of automation and (5) the network must be versatile enough to be deployed in any cloud, utilizing any API.

Features of the Arista Datacenter Network Design

Universal Spine

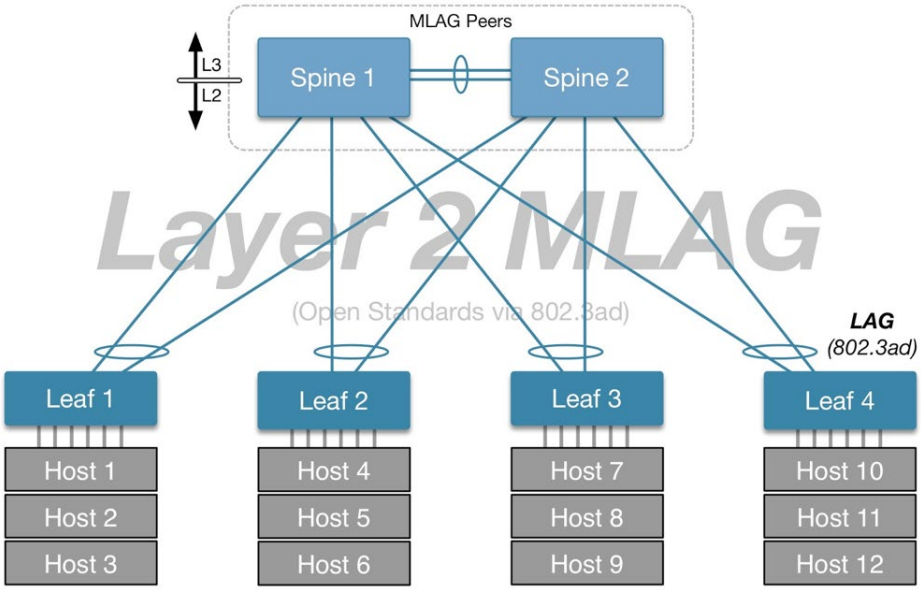
Enterprise resources commonly span multiple datacenters or PODs within a datacenter, including the Public Cloud. The drive to deliver resources in a fast, affordable and reliable manner also drives the need for a flexible, cost-effective, scale-out design at the datacenter core - the Universal Spine. The Universal Spine is non-blocking, supports large scale ECMP, uses standards based protocols and enables internet scale routing with best in class convergence.

The Universal Spine enables architects to design the network around the spine and collapse legacy networking layers into the Universal Spine. Depending on scaling requirements and overall network capacity, the Universal Spine can be leveraged as described in the next section as a 2-tier "Spine-Leaf" approach or as a 3-tier "Spine of Spines" as depicted below. In all of these approaches, the underlying concept of the Universal Spine remains the same - bringing the ideal solution for transforming the datacenter core into a consistent, reliable and scalable design.



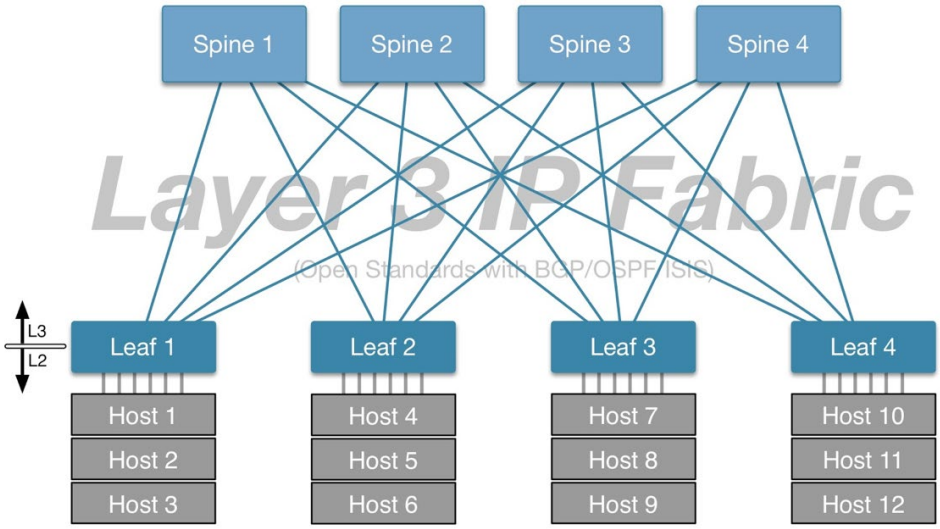
Simplified 2-tier Design

The first design depicted below represents a simple 2-tier network architecture based on a layer 2 implementation. The topology consists of a pair of redundant spine switches running in a MLAG configuration. The spines connect to Top of Rack Leaf switches which terminate connections for hosts, storage, etc. All default gateway and routing functions are served from both of the Spine MLAG peers.



Layer 2 Leaf Spine - with Multi-chassis LAG

The diagram below depicts another 2-tier design, this time with with four spines, using industry standard routing protocols between the leaf and spine. This design leverages Equal Cost Multi Path (ECMP) routing to distribute traffic evenly amongst the spine switches. In this design, leaf switches serve as the default gateways for hosts. One major benefit of using routing here, is eliminating the requirement for running non-scalable L2 protocols like Spanning-tree across the network. It is worth noting that there is an option to use anycast gateways with this design - the same subnets can be used from different Leaf nodes, we will cover this further in the VXLAN section.



Layer 3 Leaf Spine - with 4-way ECMP

Many networking vendors default to recommending complex 3+ tier architectures for datacenter networks. It is important to consider that each additional "specialized" network tier introduces complexity, drives sub-optimal bridging & routing paths and requires unnecessary switch-to-switch interfaces. Multi-tier designs also inherently create multiple points of oversubscription. The recommendation for n-tier datacenter networks is usually based on several sub-optimal factors:

- Proprietary bridging/routing "protocols"
- Insufficient ECMP scale-out capability in proprietary ASICs
- Insufficient table sizes forcing additional layers to remain in the data-path
- Insufficient port density/performance

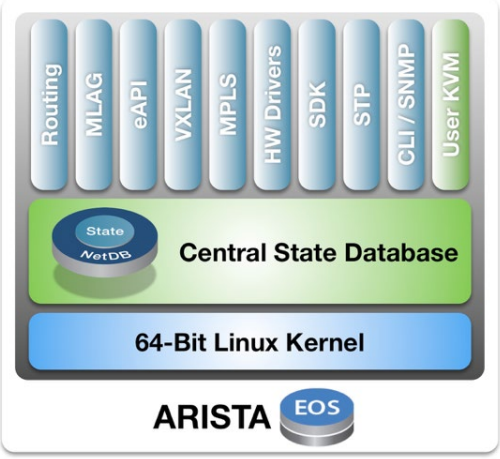
Highly Available, Reliable and Resilient Network Topology

Arista's Extensible Operating System, or EOS, is the most advanced network operating system available. EOS was created with modern-day software and operating system architectures, transparently restartable processes, open platform development, all using an unmodified Linux kernel. At the heart of EOS - System Database (SysDB) drastically simplifies, and thus improves the reliability of, inter-process communication. These building blocks ultimately provide for extreme reliability in how the hardware interacts with the software.

Arista EOS: Use of Native Linux Kernel

Arista leverages an unmodified Linux Kernel to provide resiliency within the network. Building a network operating system on top of an unmodified Linux Kernel has many benefits including:

- Thousands of contributors and millions of users (Fedora has 40+ million users)
- The ability to run countless software packages to provide increased breadth of features
- Quickly resolved O/S, Security and Bug fixes
- Freely available application repositories and source code



Arista EOS: State-Based SysDB Foundation

Separating functional control of the system into multiple processes is great for resiliency and fault isolation, but requires a mechanism for coordinating actions within the system. Arista pioneered this capability through an innovation called SysDB. SysDB is an in-memory database running in user space containing the state of the system. Like traditional databases, SysDB does not contain any application logic and is only responsible for maintaining and propagating the state of the system. However, rather than being optimized for transactions, SysDB is designed for synchronizing state among processes, also called 'agents', by notifying interested processes or agents when state changes.

Arista EOS: SysDB – Publish-Subscribe Programmable Model

A common, recurring, architectural flaw in traditional network operating systems has been how interprocess communication is handled. Complex interdependencies between different processes on a switch require constant refactoring as bugs are fixed/features are added - this burdensome recurring task leads to general instability for other network operating systems. (e.g. memory leaks, regression bugs, etc)

SysDB addresses this problem by operating using its event driven publish/subscribe model. If an attribute or object changes, SysDB will send an event notification to agents that are following an object/attribute, which will then update their local copies. Similarly when the agent writes to the mount point, it only changes its local copy and the write returns immediately. This state-sharing process is key in making EOS extensible, and ultimately the most reliable network operating system in the world.

Hardware resiliency and density impact on MTBF

A highly resilient software model is crucial to providing the customer with the ultimate reliability in their datacenter networks. Although Arista EOS is available in virtual and containerised forms, having an equally innovative hardware model provides customers with real value with rapidly increasing switch densities, high resiliency and reliability through quality system designs and leveraging merchant silicon.

To date, Arista is leading the industry in price, density, and power per port for datacenter networking. Power consumption is one of the leading indicators of the number of components necessary to achieve designed throughput within a device. The more packet processors needed, the more power and subsequent cooling are necessary to utilize these components. Arista has been shipping the world's leading 10/40/100G platform since April of 2013 with the 7508E This platform delivered 1152x10GbE, 288x40GbE, and 96x100GbE with the ability to scale even further.

The introduction of the 7500R series expanded the choice of interface speed to support 25GbE and 50GbE in addition to 10/40/100 GbE. Port densities have also increased dramatically with the 36x100GbE linecard. In the 7508R chassis up to 288 ports of 100GbE is possible. The 7500R series also expanded the chassis range by offering the 7512R and 7516R. The 7516R supports up to 576 100GbE ports with under 25W per 100GbE port.

System reliability is not determined by the mean time between any individual device failures, it's how the system resiliently handles the loss of a critical component. Dual supervisors, multiple power supplies, fan and fabric modules ensure resilience. Additionally, fabric module N+1 redundancy provides zero loss of performance even if a fabric is removed. This resilient hardware architecture, coupled with the superior network operating system provided by EOS, makes Arista a highly compelling platform.

Topology Redundancy

Arista believes that a resilient and redundant topology stems from redundant hardware elements in addition to redundant software coordination. The Arista design provides for inherent redundancy in the communication between the leaf and spine by way of equal cost multipath routing protocols such as BGP or OSPF, or in the case of a Multichassis Link Aggregation Group (MLAG) spine, standards based LACP and port channels are used to provide an active/active topology. These protocols allow for the failure of any spine device in the topology. Arista has optimized the ability of the hardware to react to any failures in this scenario and leads the industry in converging around planned and unplanned failures.

Each leaf device can be in a MLAG pair creating an active/active forwarding topology between each host in the environment. Should a leaf switch fail, a sub-second failover can take place to forward traffic toward the other active device.

While topology redundancy is important, maintenance on the topology is becoming equally important. Software upgrades and maintenance require new mechanisms such as Arista's Smart System Upgrade (SSU) functionality, which allows switches to be upgraded while the dataplane is still operating, minimizing traffic impact and can further streamline the process by intelligently routing traffic around the device under maintenance.

Scale-out

Operating from a forecast of a 5x increase in the number of hosts this network will need to support over the next 5 years, the current design supports a scale-out model as follows (assuming edge services pod connectivity stays the same)

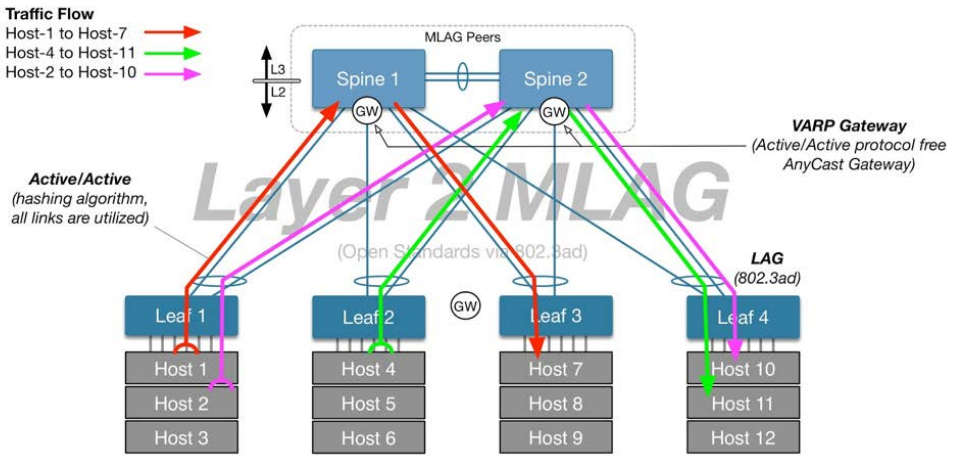
Growth Example (assuming a minimal oversubscription ECMP model)

- 960 Dual Connected Physical Servers/Storage Hosts.
  - » 2x7508R using 80x100GbE per system – scales to 144 Racks
- 4,800 Dual Connected Physical Servers/Storage Hosts.
  - » 4x7508R using 200x100GbE per system – scales to 288 Racks
  - » 4x7512R using 200x100GbE per system – scales to 432 Racks
  - » 4x7516R using 200x100GbE per system - scales to 576 Racks
- 12,000 Dual Connected Physical Servers/Storage Hosts.
  - » 8x7508R using 250x100GbE per system – scales to 576 Racks
  - » 8x7512R using 250x100GbE per system – scales to 864 Racks
  - » 8x7516R using 250x100GbE per system - scales to 1152 Racks

The current model can be scaled 16x with 100% capital equipment retention with a few assumptions (hosts stay connected at 2x10Gb or 2x25Gb, Leaf Switches are connected at 400Gb if equipped with at least 4x100Gb uplinks, etc.).

Layer 2 Active/Active Forwarding

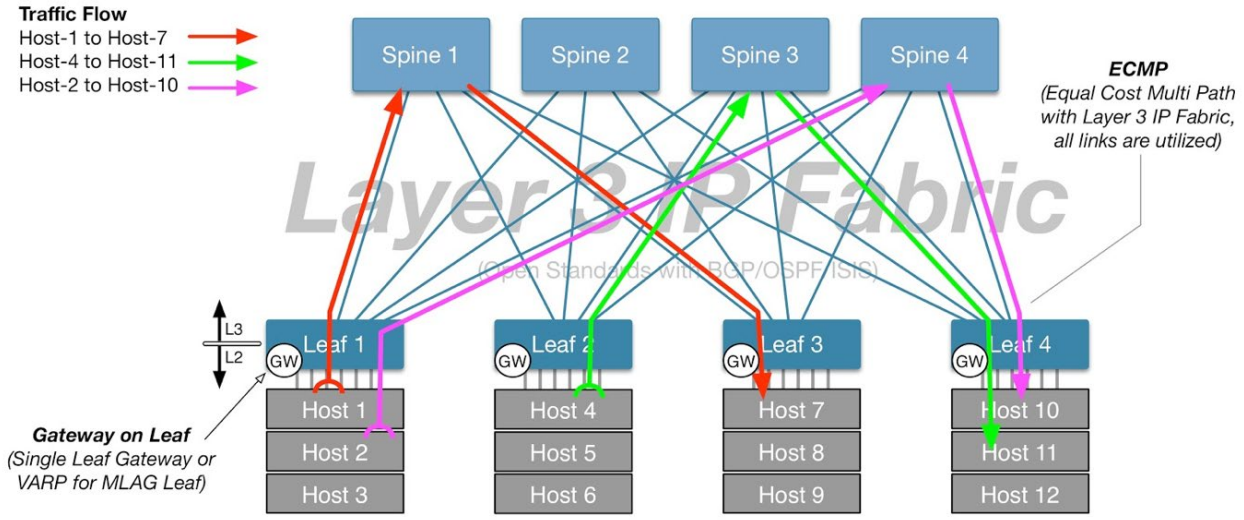
In the Layer 2 Leaf/Spine (L2LS) model Arista leverages the Link Aggregation Control Protocol (LACP) to provide logical connectivity in an active/active topology. LACP allows for complete L2 connectivity between the leaf and spine switches eliminating the need for spanning-tree and a forwarding plane where half of the uplinks are in a blocking state. LACP is widely adopted standard that allows many different vendors to interoperate thus eliminating the need for proprietary forwarding mechanisms. The L2LS design requires that the Spine switches be connected together as MLAG peers as depicted below.



Layer 2 Leaf Spine - with Multi-chassis LAG

Equal Cost Multi-path Routing (ECMP) Forwarding

Utilizing BGP or OSPF as the routing protocol enables the use of equal cost multipath routing (ECMP). With ECMP routing each path is of equal cost and thus will load-balance flows across the IP fabric.



Layer 3 Leaf Spine - with 4-way ECMP

ECMP will utilize all of the available bandwidth in the leaf/spine IP fabric while also providing for resiliency should any of the spine devices or links fail. Load balancing across the spine nodes is flow based to minimize the amount of packet jitter in the session. One potential issue with ECMP or any load-balancing algorithm is the potential for hash-polarization where the load-balancing algorithm result is the same for a number of flows, thus potentially saturating an individual link. Arista addresses this challenge through the use of a configurable hash seed that can be used to rebalance the traffic and minimize hash polarization and link saturation.

Spine Buffering and Congestion Management

In general, buffering and congestion management does not seem to be well understood when it comes to the datacenter network and there is a lot of misleading information when it comes to network buffers. Arista's view on buffering is that any network that may experience congestion will utilize buffers to some degree. When deep buffer systems are required Arista recommends the 7280R or 7500R Series switches. In the 7500R and 7280R Series systems, each port is capable of providing approximately 50ms of packet buffering, which is sufficient to provide a lossless Ethernet fabric for high bandwidth workloads.

While a proper leaf and spine design minimizes oversubscription, shared NAS pools and dedicated IP storage leaf switches present potential bottlenecks in the network and are prone to receiving large amounts of incast traffic that can cause buffers to overflow and ultimately packet loss.

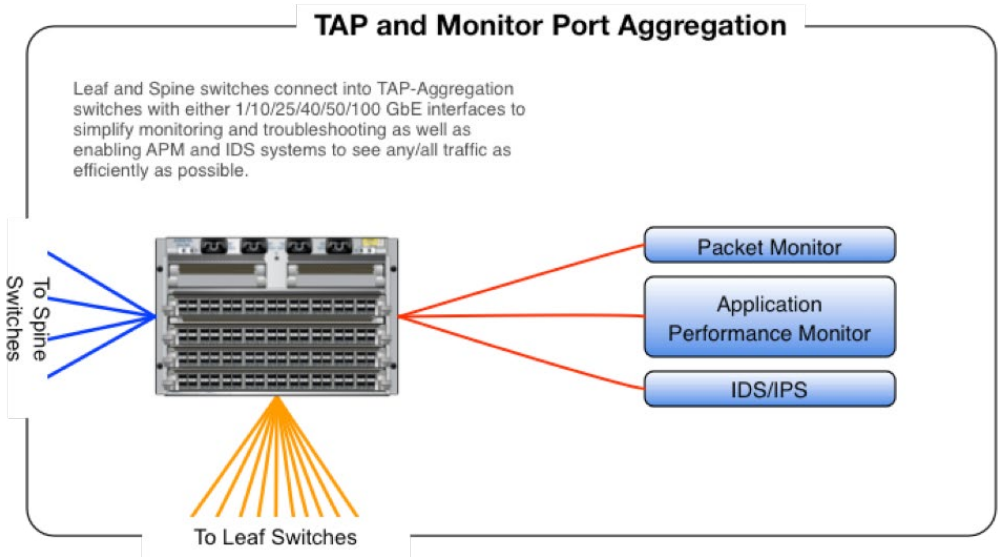
The implementation of a Leaf and Spine architecture with smaller buffers will perform well in a network that does not experience congestion. This fact makes it critically important to understand the performance characteristics required prior to making product decisions. Performing a good baseline analysis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

Single-Image Consistency

All platforms recommended by Arista in this design guide run a single binary image of Arista EOS. With a single image to qualify, across all platforms, the tedious process of testing and bug scrubbing is significantly reduced.

Tap and Monitor Port Aggregation

To support the requirement for network and application visibility Arista supports Tap Aggregation. The Arista 7150S, 7280E/R and 7500E/R Series support Tap Aggregation mode. Monitoring tools such as Application Performance Monitors, IPS/IDS systems and general packet capture devices connect via ports off the Tap Aggregation switches.



The Arista Tap Aggregation Manager (TAM) is a GUI for use by the network operations team used to rapidly provision ports from monitoring sessions on the Leaf/Spine switches back through the Tap Aggregation switch to one or more tools. With this in place any port on the network can be provisioned to send all or a filtered part of its traffic to one or more management/monitoring tools in seconds remotely. Tap Aggregation devices may be managed by Arista CloudVision in the same way regular switches are enabling centralized provisioning and automation for a large fleet of Tap Aggregation devices.

Simplifying Upgrade/Change Control at the Leaf

Multi-chassis Link Aggregation (MLAG) ISSU

MLAG ISSU offers network operators the ability to stagger switch upgrades without the need to schedule a service outage. During the upgrade process, ISSU redirects traffic to a peer switch (in milliseconds) to minimize network impact. EOS is able to synchronize SysDB between versions, allowing the upgraded peer to reestablish MLAG sessions and restore active-active connectivity. The ISSU process can then be repeated on the other switch to complete the, near-hitless upgrade.

Intelligently coordinating the leaf switch upgrade allows migration from version-to-version; adding features and capabilities without causing outages or impactful packet loss. This upgrade coordination can also be managed from within Arista CloudVision through the Change Control functionality.

Accelerated System Upgrade (ASU)

ASU is a rapid-reload feature in which the operating system restarts while maintaining the forwarding state of all packet processors, ensuring that traffic is still forwarding while the control plane is updated. The reliable forwarding capabilities of Arista's EOS operating system allow the downstream network elements to continue to operate during the upgrade process without any knowledge that one of the primary upstream nodes is not available. This allows the network to remain available during the upgrade process, maximizing system uptime and network resource availability.

Utilizing MLAG ISSU and ASU Arista EOS provides the ability to perform system upgrades without needing to schedule service outages. These key features enable network designs that are based on Arista's two-tier Spine/Leaf design.

Programmability and Automation

As more applications rely on the network, the network must be able to provision quickly and roll changes out in a predictable and scalable manner. Automation helps eliminate time-consuming tasks that may be done programmatically such as provisioning VLANs on specific ports. Another benefit to automation is since the actual change is performed programmatically, the risk of an adverse effect is also mitigated by the human error component being removed from the process.

Upgrading network components is a large component of network operations. Running legacy software can make the network vulnerable to security, performance and stability issues that are commonly remedied in newer releases. Automation enables switch upgrades to be a programmatic process of gracefully removing the switch from the network to complete the upgrade procedure and then bringing the switch back into the network after analyzing the switch to ensure that the switch is in the same state prior to the upgrade.

Extensibility

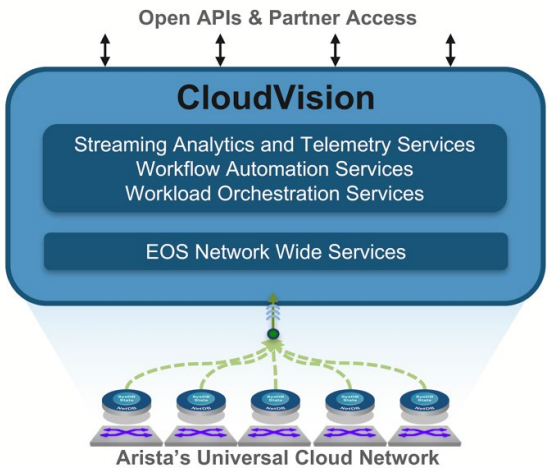
Arista EOS provides the flexibility and extensibility to make end-to-end network automation a reality. This is accomplished by supporting API integration and allowing operators to customize every aspect of the network operating system. Tasks can be streamlined by extending the CLI to incorporate output from a 3rd party API's to give the operations team a holistic view with a single command.

EOS also offers full Linux shell access for root-level administrators, which makes a broad suite of Linux based tools available. This provides operators the flexibility to use the industry standard CLI or launch a bash session and manage the box in the same manner as Linux servers are managed in the rest of the environment. Utilizing a standard Linux kernel and allowing root level access provides a stable and open environment that redefines extensibility in network operations. Root level access may be controlled and limited through standard Role Based Access Control (RBAC).

Automating the Arista UCN with CloudVision

Arista has pioneered the networking industry with its software defined cloud networking approach built on top of Arista EOS with programmable interfaces, publish/subscribe state separation, resilient fault containment and self-healing attributes. CloudVision extends the same architectural approach across the network for state, topology, monitoring and visibility. This enables enterprises to move to a turnkey cloud class automation without needing any significant internal development. CloudVision is a network-wide approach for workload orchestration and workflow automation as a turnkey solution for Cloud Networking.

CloudVision's abstraction of the physical network to this broader, network-wide perspective allows for a more efficient approach for several operational and network telemetry use-cases, including the following highlights:

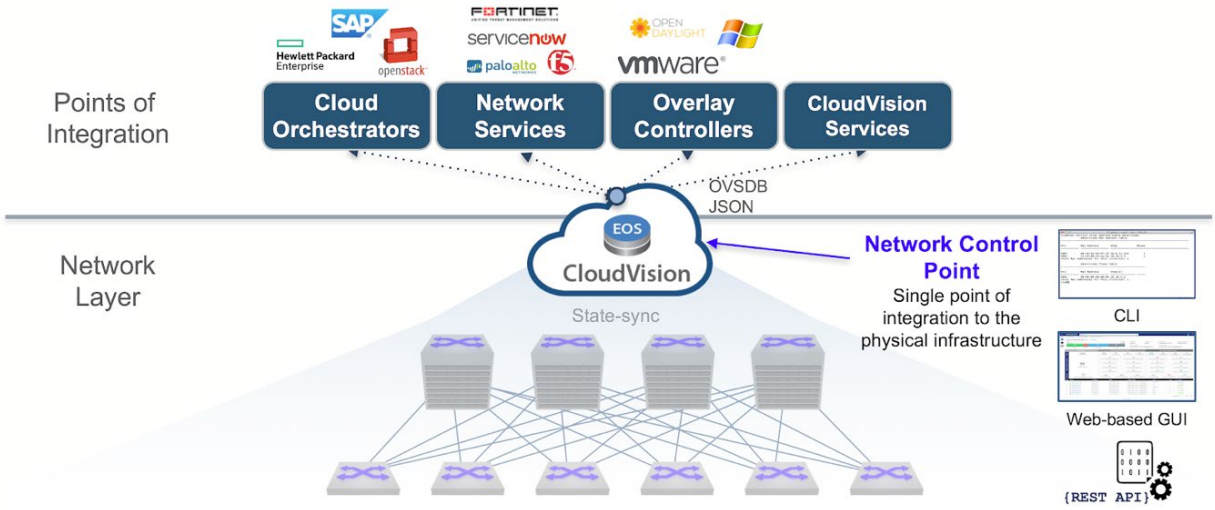


- Centralized representation of distributed network state, allowing for a single point of integration and network-wide visibility and analytics;
- Controller agnostic support for physical and virtual workload orchestration through open APIs such as OVSDB, JSON and Openstack plugins;
- Turn-key automation for initial and ongoing zero touch provisioning, configuration management and network-wide change management, including automated upgrades, network rollback, and network snapshots;
- Compliance Dashboard for Security, Audit and patch management;
- Real-time state streaming for network telemetry and analytics, a modern approach to replace legacy polling per device;

- State Repository, Analytics engines and Telemetry Apps to provide an unprecedented level of granularity in real-time monitoring and historic network state for forensic troubleshooting;
- Cloud Tracer™ for visibility into the availability of network interconnects and services across private, public, and hybrid cloud environments; and
- Macro-Segmentation Service provides automated and seamless service insertion and integration with partner security platforms.

Single Point of Integration

CloudVision's open API infrastructure reduces development costs of orchestration tools. CloudVision communicates with multiple controllers simultaneously to accommodate heterogeneous datacenters enabling a common infrastructure to economically server cloud datacenter workloads. CloudVision delivers this common infrastructure and by enabling a single point of network state with no proprietary lock-in with open partner integration including Dell ASM, HP OneView, F5, Palo Alto Network Panorama, Microsoft (Systems Center and Network Controller), Openstack Neutron ML2 plugin (RedHat, Rackspace, Mirantis, SUSE and VMware), Nuage Networks Virtualized Service Platform, and Infinera Transport SDN.



This section will focus on two components of the CloudVision product suite: CloudVision eXchange (CVX) and CloudVision Portal (CVP). CloudVision Telemetry is covered in detail in the "Visibility and Monitoring of the Arista UCN" section.

CloudVision eXchange (CVX)

CloudVision eXchange (CVX) is built on the same underlying robust architecture as Arista EOS. SysDB is at the heart of EOS providing a centralized state of the Arista switch through which all processes or agents publish and subscribe state information. This same concept is leveraged in CVX to offer the ability to centralize the state of the entire network fabric in an aggregated instance of SysDB.

This centralized state may be utilized to provide workload automation in ways that have not been possible until now in a datacenter network. CVX includes a number of services that may be activated to take advantage of this state.

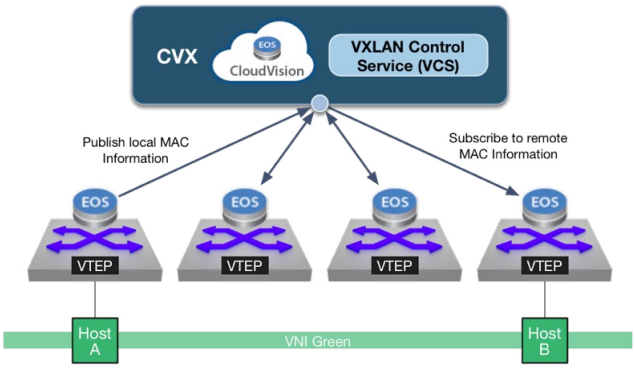
These services include:

- Topology Services - utilizes information gathered from managed switches. This data is aggregated by CVX to build a generic physical view of the network topology.
- VXLAN Control Service (VCS) - centralized control service for MAC learning in a VXLAN environment (VCS is covered in more detail in the L3LS-V Section)

- Macro-Segmentation™ Service (MSS) - integration with network services platforms for automated service insertion
- Bug Alerts service - provides information on known, resolved bugs that are impacting Arista switches. The feature uses learnt switch properties such as EOS release, hardware platform, configuration and operating conditions of all connected switches to determine the list of impacting bugs for each switch in the environment
- Controller Integration services enabling the network wide state knowledge within CVX with third party platforms via API. Technology partners include VMware NSX, OpenStack, Nuage and Microsoft Systems Center. This enables state sharing for technologies like VXLAN to deepen the physical + virtual integration V

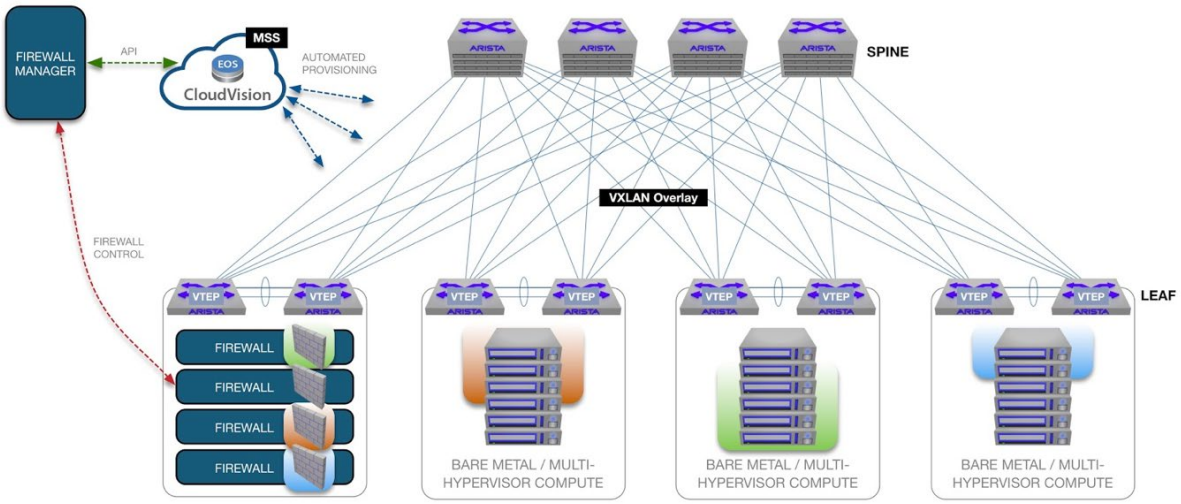
VXLAN Control Service

The VXLAN Control Service (VCS) allows VXLAN tunnel end-points (VTEPs) to publish their state information to the CVX, for example, the MAC addresses local to a VTEP. Conversely, the VTEP subscribes to CVX for the remote MAC-to-VTEP information. This state sharing allows for a VXLAN deployment without the reliance on an additional network protocol such as multicast or BGP to provide the control plane for VXLAN, and simply leverages the connection to CVX which, as described above, is used for more than just MAC learning.



Macro-Segmentation™ Service (MSS)

Macro-Segmentation Service (MSS) provides a software-driven dynamic and scalable network service to insert security devices into the path of traffic with complete flexibility on placement of service devices and workloads. It is specifically aimed at physical-to-physical (P-to-P) and physical to virtual (P-to-V) workloads.



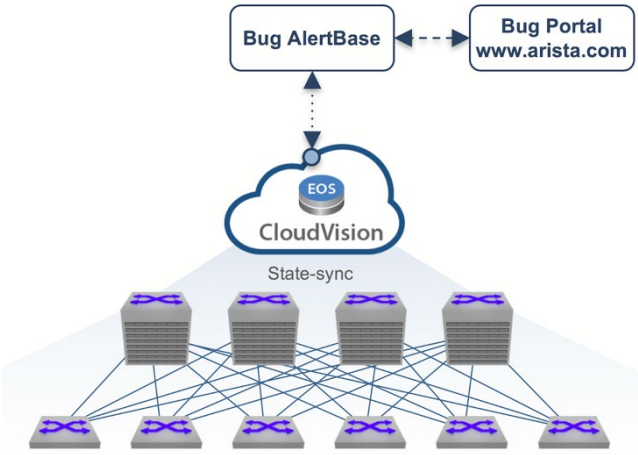
Arista CloudVision provides a single point of integration to the network and ties in automation and orchestration capabilities with a network wide view by aggregating the entire network state. Arista MSS is a service in CloudVision that provides the point of integration between individual firewalls or firewall manager and the Arista network fabric.

When enabled, Arista MSS communicates with the firewall or firewall management platform, using REST APIs and requests security policies of interest. Upon receiving the policy, CloudVision MSS instantiates a new traffic path through the network on the required leaf switches to steer traffic from interesting hosts to the firewall for further inspection.

Bug Alerts Service

Bug Alerts is a service that runs on CVX and provides customers with information on known and resolved bugs that are impacting Arista switches. The feature collects switch properties such as EOS version, hardware platform, configuration and operating conditions of all connected switches. It uses these switch properties and a local database of known bugs (AlertBase) to determine the list of impacting bugs for each switch. This information is then displayed via show commands on CVX.

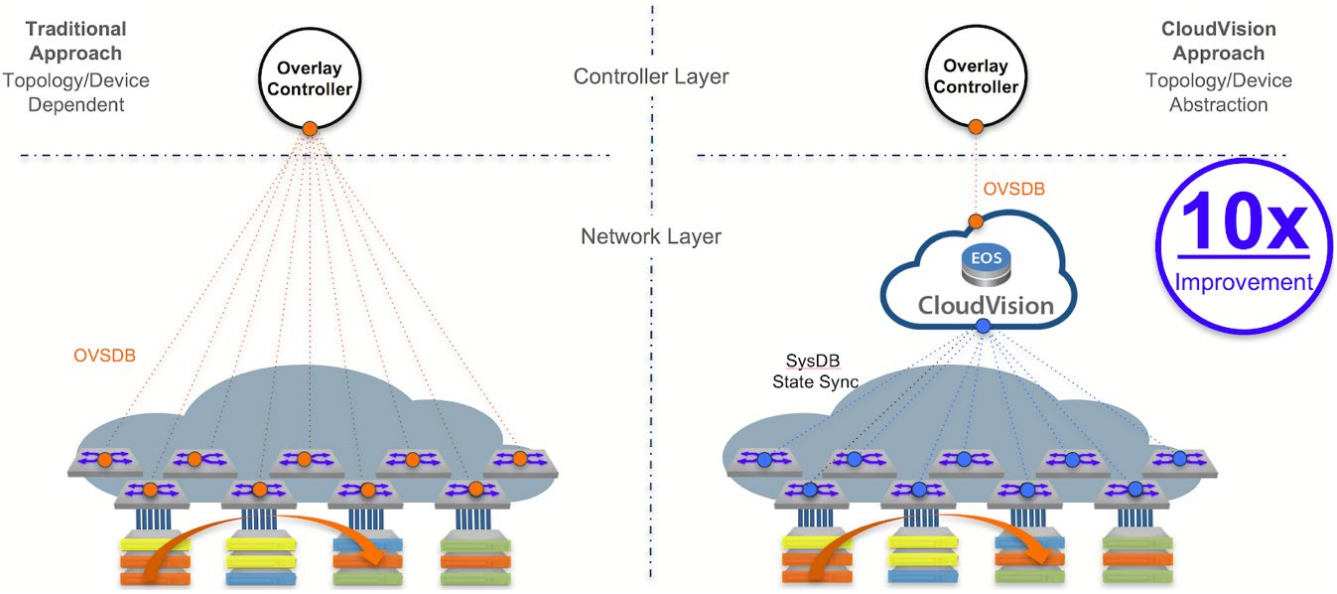
AlertBase – This is the database of known and resolved bugs in Arista switches that is published by Arista Networks as a downloadable file on the software downloads page of www.arista.com. In this page, the file is available under CloudVision->CloudVision eXchange folder. Users registered on the website are required to have access to the customer support section to be able to download the AlertBase.



Third Party Controller Integration

CloudVision delivers a network services abstraction layer that decouples controllers from the physical datacenter infrastructure. This abstraction insulates controllers from infrastructure OS dependencies removing switch OS and controller version dependencies and thus reducing costs associated with controller certification and network maintenance.

This includes networked applications for workload mobility, network configuration management and control via snapshots, network wide rollback and upgrades, as well as integrated network telemetry. CloudVision's network abstraction dramatically improves controller scaling by up to 10X competitive offerings, with a single touch point to control Arista switches in a cloud data centre. CloudVision brings the benefits of software driven cloud solutions by leveraging Arista's network wide state synchronisation to an Extensible Operating System, EOS, reducing Opex. It delivers customers turnkey solution to cloud-like workload automation and workflow visibility.



Two such examples of integrating the Arista UCN with third party controllers through CloudVision are with Openstack and VMware NSX.

Openstack Integration

Arista CloudVision platform has extensive integration with the OpenStack project, giving customers a powerful network platform on which to run OpenStack deployments. By leveraging the Arista ML2 driver and Layer 3 service plugin, operators can automatically provision tenant networks across the physical infrastructure. This combination provides a high-performance OpenStack networking environment over VLAN and VXLAN based fabrics, along with enhanced visibility into how the virtual tenant networks map onto the physical infrastructure.

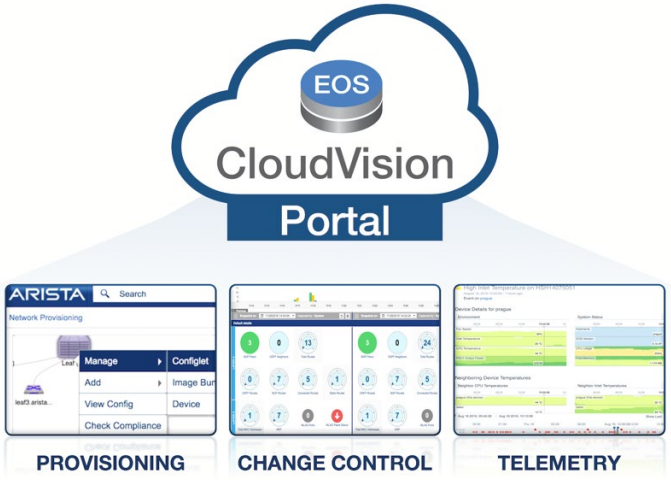
The Arista OpenStack solution provides a number of ways for administrators to orchestrate their Arista switches. The ML2 plugin automates the provisioning of VLANs on Arista switches and can optionally be combined with an Arista VXLAN overlay to provide the functionality across a Layer 3 core. With the Arista layer 3 service plugin, a hardware switch can serve as the routing gateway, even in a VXLAN environment where VXLAN routing is required.

The ability to orchestrate the physical network devices provisioning within the OpenStack solution is a significant achievement for the market. Arista has consistently led the way in providing new and open functionality to the OpenStack community and has an extensive set of features designed to address the needs of scaling an OpenStack cloud.

NSX Integration

Arista CloudVision platform provides network-wide visibility and single point of integration to NSX. Using CloudVision as the integration point allows for changes in the network topology to be abstracted away from NSX. In addition, CloudVision provides software and hardware version independency from joint certification. Since CloudVision runs the same EOS as any other Arista switches, customers need to only certify the CloudVision version with NSX. CloudVision, in turn, provides the aggregate VXLAN state of the physical network to NSX for the most effective physical to virtual synchronization in today's datacenter. This abstraction improves controller scaling, using only one touch point to control all Arista switches in the datacenter. CloudVision also provides redundant hardware L2 Gateways for NSX with the MLAG with VXLAN functionality. MLAG with VXLAN on Arista switches provides non-blocking active-active forwarding and redundancy with hitless failover in an event of switch failure.

In operation, Arista CloudVision will register with the NSX controller and will use the OVSDB protocol to synchronize topology information, MAC to VXLAN Endpoints, and VXLAN ID bindings with NSX. CloudVision will program the Arista switch or switch pairs as the NSX hardware gateway. This hardware gateway integration allows for instantaneous synchronization of state between physical and virtual VXLAN Tunnel Endpoints during any network change or workload modification event. The same mechanism enables provisioning of distributed QoS and security policies at the physical and virtual overlay edge. VMware's NSX Manager front-ends the entire NSX network virtualization platform and acts as the single pane of glass for rolling out your SDDC network. In addition, NSX provides a northbound REST API to further integrate with OpenStack or other in-house orchestration platforms.

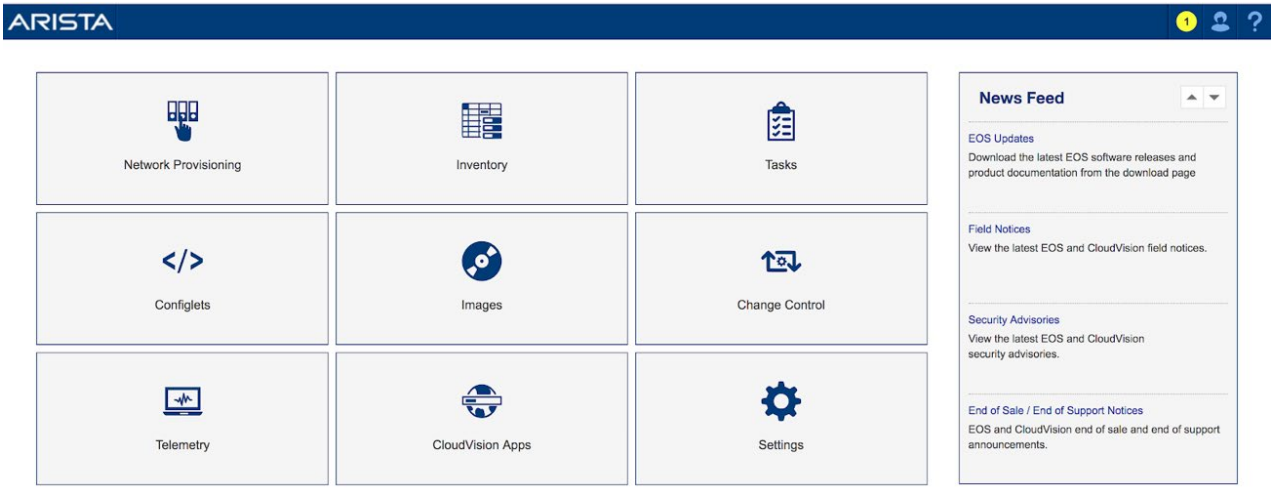


CloudVision Portal (CVP)

CloudVision Portal (CVP) provides administrators a single pane of glass to quickly and efficiently control the network. Tasks such as pulling inventory reports, configuration management, zero touch provisioning and real time state streaming are enabled as turnkey solutions with CVP. CloudVision Portal also offers API connectivity for third party applications or tools to be connected into the CloudVision framework and deployed as part of CVP.

CloudVision Telemetry is an application that exists within CVP and will be detailed in the "Visibility and Monitoring of the Arista UCN" section.

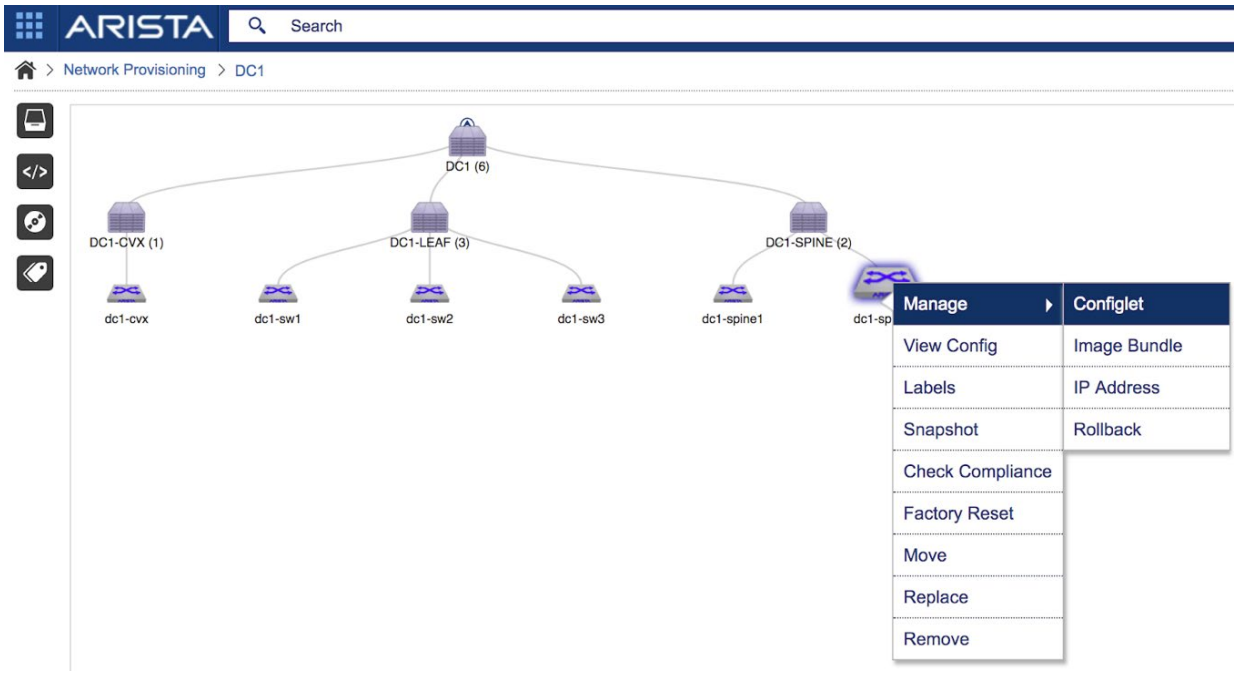
The dashboard provides a simple way to view the various components of CVP. The primary goal of CVP is to provide an easy mechanism for users of the Arista UCN to achieve cloud scale automation without significant development effort.



Network Provisioning

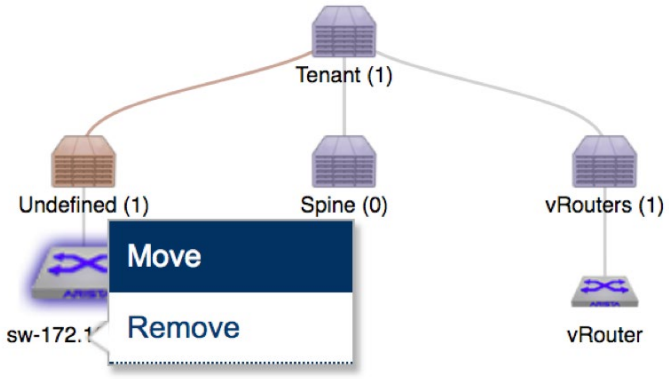
CloudVision Portal enables automated network provisioning for the Arista UCN through features such as: Zero Touch Provisioning (ZTP), Zero Touch Replacement (ZTR), network configlets, programmable configlet builders and task management.

CloudVision presents the user with a logical view of the Arista UCN in which to provision changes to the environment.

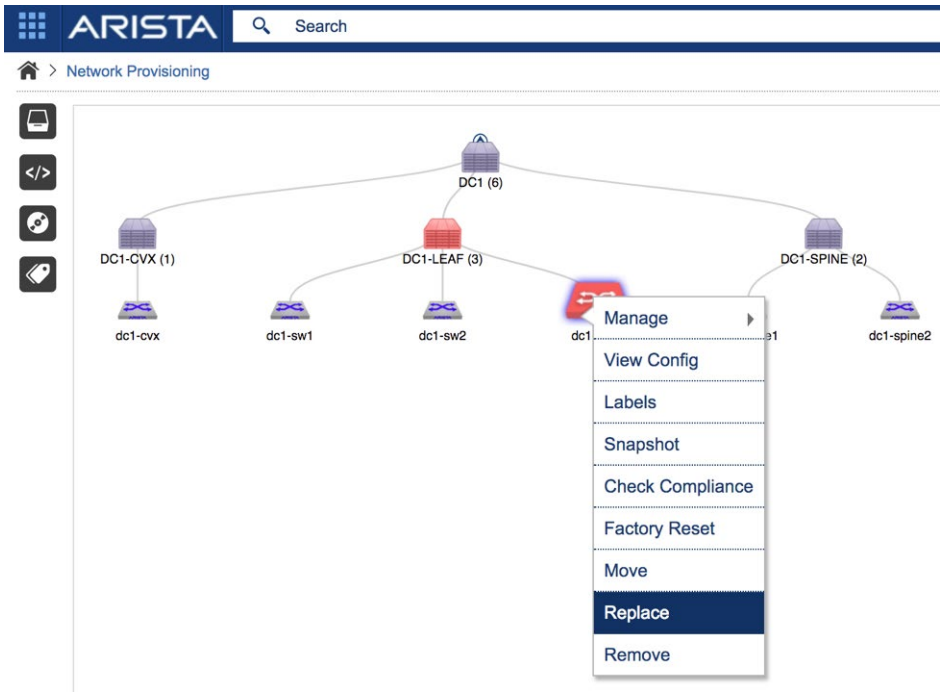


The components of Network Provisioning are detailed as follows:

- Zero Touch Provisioning (ZTP) allows the network operators an automated way to start provisioning switches out of box. This eliminates the need to have console cables on hand when doing initial deployments. Once switches are powered on they will seek out the CVP server using DHCP options provided by the onboard DHCP server or by aid of an 3rd party server. Once in CVP it stays in a ready to deploy state, where EOS upgrades, configurations and extensions can be applied. Furthermore this process can be automated by utilizing scripts & combined with Configlet Builders or the Cloud Builders App can reduce the time to deploy new devices.



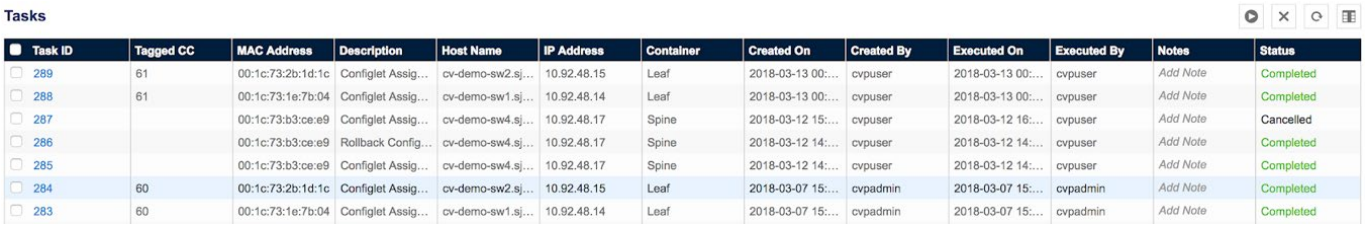
- Zero Touch Replacement (ZTR) allows for quick and efficient RMA replacement of devices. The Zero Touch replacement enables the user to select a replacement switch that has been provisioned using Zero Touch Provisioning. This allows for the quick replacement of devices by inheriting the configuration of the RMA'd device. The restoration would include automated provisioning of the deployed EOS release and any extensions bundled in with the image and device configuration. Compare this to legacy methods of manual upgrade and the last known good configuration of a failed device that may be out of date.



- Network Topology provides a configuration hierarchy view of the network. This allows the network administrator to segment out the running configuration to manageable sections. In large configurations, it may be ideal to apply common segments to a high level container to have all the devices contained within the container to be inherited. This allows for simpler configuration management, as common configurations can be set once and any future changes can be done on one configuration segment call an "configlet" which gets applied to all devices associated with the configuration. The ability to quickly apply hotpatches, extensions and new EOS versions to devices quickly can be done from this screen. This can allow for image consistency and gives you the ability to ensure compliance of devices and configuration from a single screen.

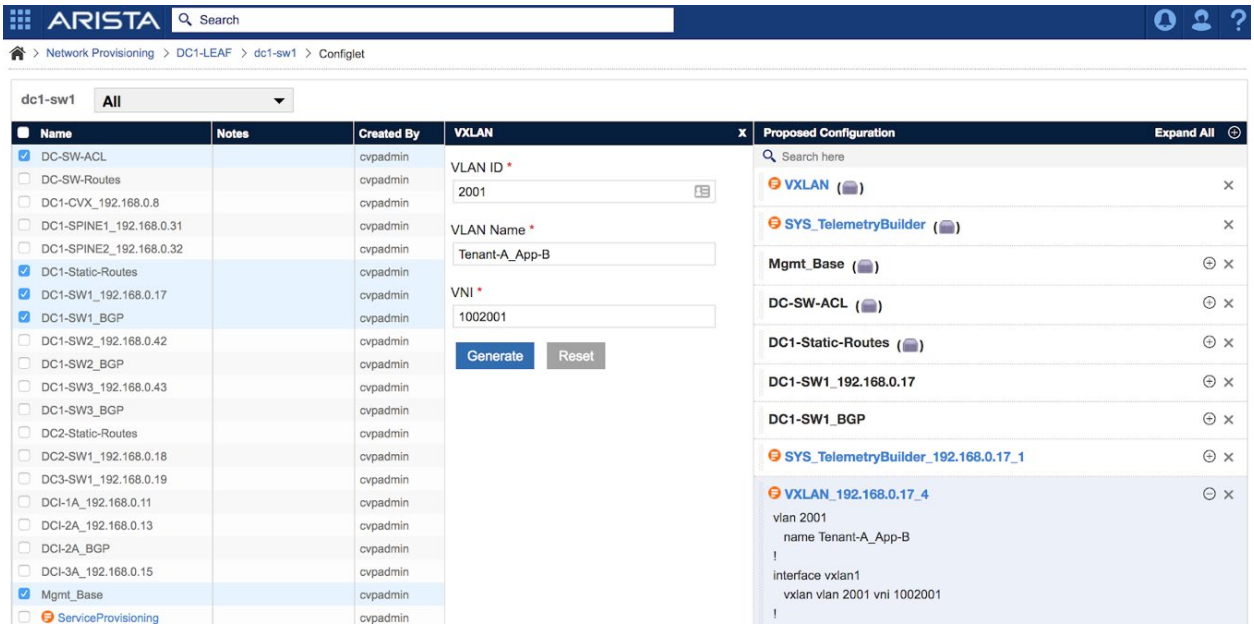
Tasks

- CloudVision Portal Tasks pane shows the granular details of the changes processed and pending within CloudVision Portal along with a detailed log of the changes. Tasks can be executed individually or can be queued to be attached to a change control task to execute with snapshots before and after changes. The executed tasks are granular and can show comparison of the configuration before and after and current state of tasks.



Configlets

- Configlets allow for standard configurations to be pushed out to devices and also allows CloudVision to do sanity checks to ensure the corresponding configuration to be validated with the appropriate devices. Configlet builders are also available with in the configlet section. These builders enhances standard configurations by adding a programmable method to design and apply configurations to devices.



Inventory Management

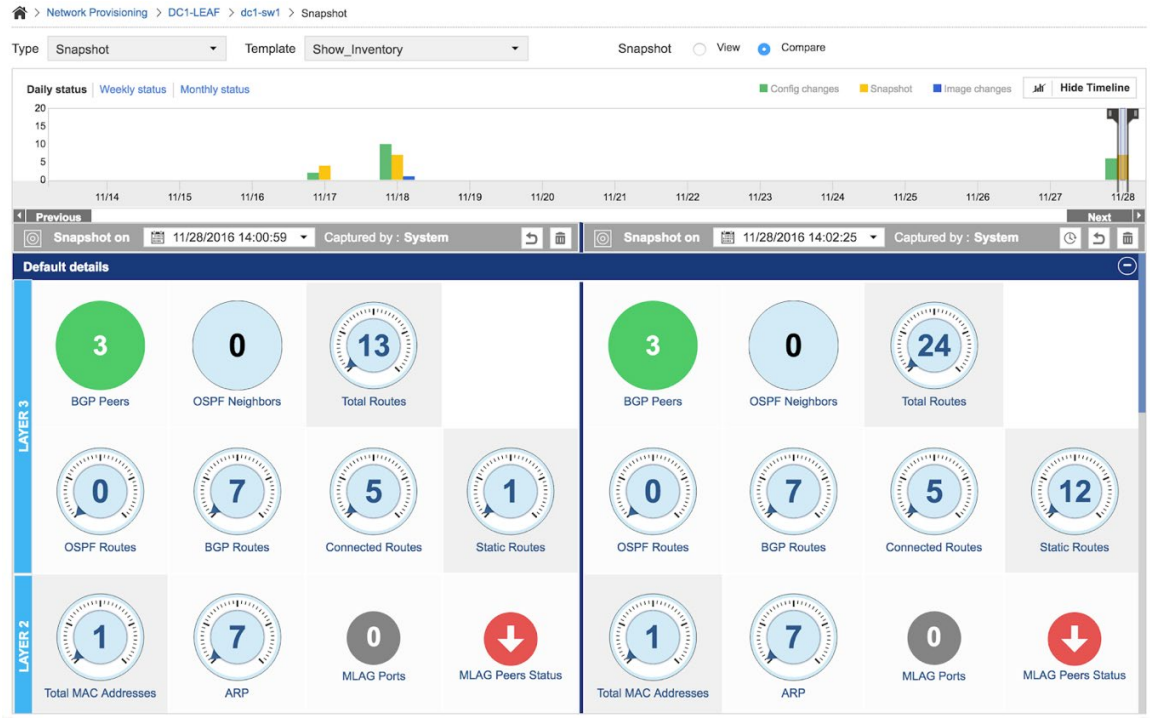
- Inventory provides administrators an exportable view of all devices, both physical or virtual in the network and associated details of the devices including software version, device type and location in the network.

Automated Software Upgrades

- Images pane is a repository for binary software images that can be staged to load on devices. Combined with the ability to combine extensions with EOS images, simplifies code upgrades, new switch deployment and consistent configurations across the infrastructure. Image management can be simplified by creating certified image bundles that can be applied to containers to achieve the same inheritance style application of new EOS images similar configlets. With image bundles any Arista Security hot patches can be added to the image bundles and applied quickly to mitigate vulnerabilities.

Change Control

- Change Control gives the network administrator a mechanism to combine several network changes into a single change. Change controls can consist of multiple change tasks, image upgrades along with the capability scheduling the changes to execute. Change Control based tasks create snapshots before and after executing changes providing a quick comparison method on the verification of changes.



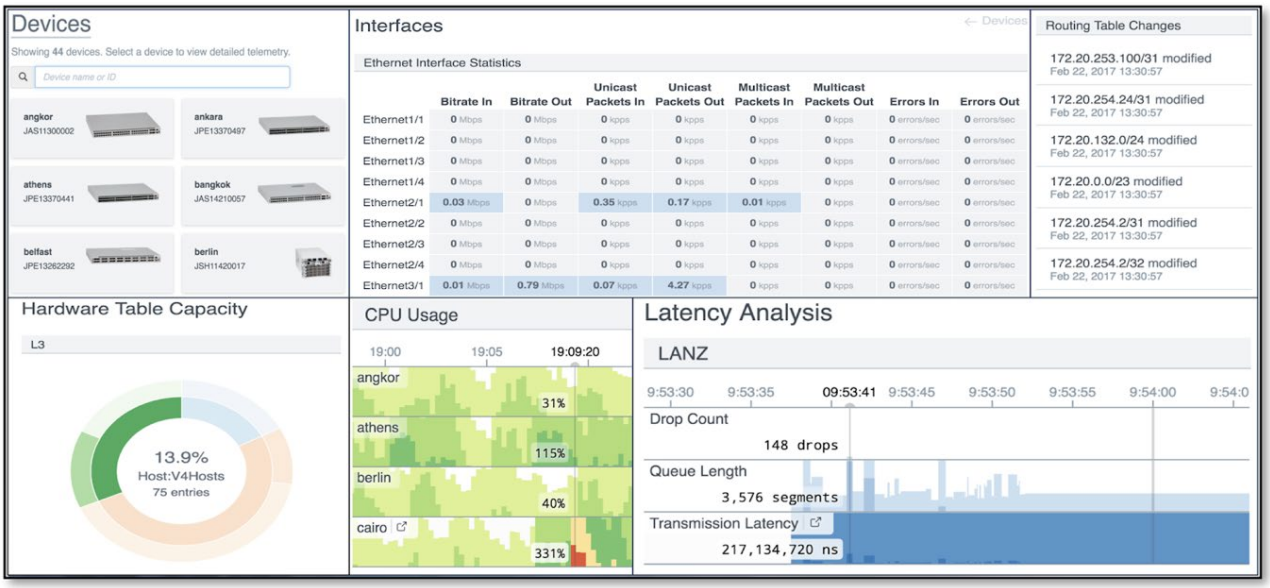
Other CloudVision Apps

- Tap Aggregation shows tap aggregation devices on the network to quickly locate devices and troubleshoot issues.
- CloudVision Apps enable other developers to integrate applications into CloudVision Portal and take advantage of network wide information.

Visibility and Monitoring of the Arista UCN

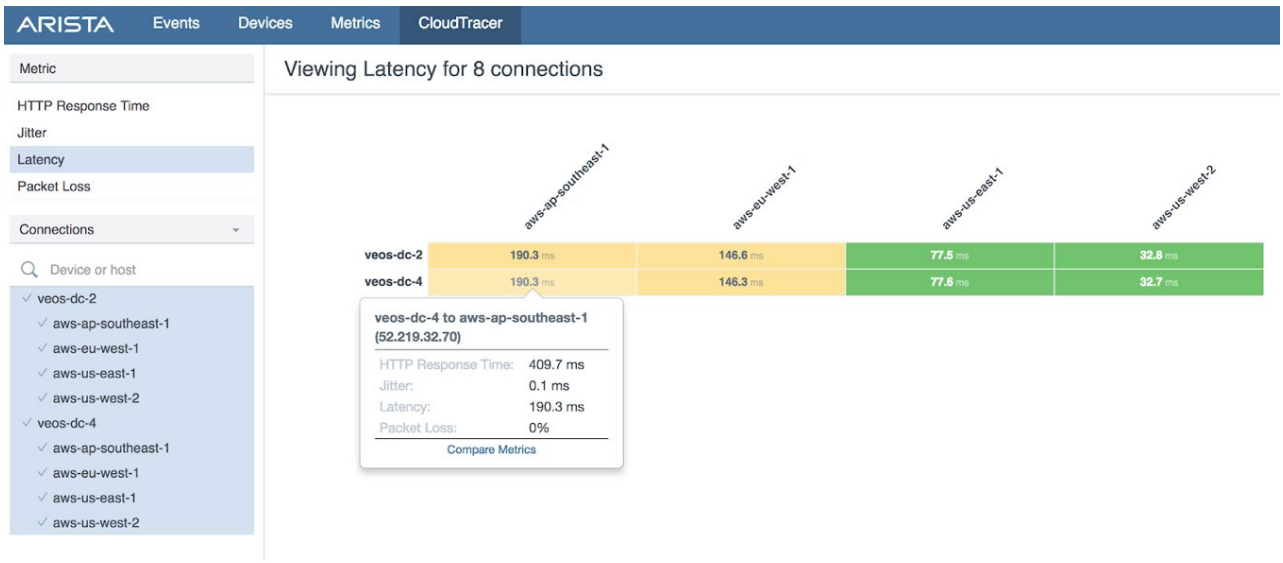
CloudVision Portal Telemetry

CloudVision Portal Telemetry provides a real time view into the state of the network. The telemetry capabilities within EOS leverage a consolidated view of all SysDB instances into a repository called the Network Database or NetDB to provide CloudVision Portal with a data rich repository of network state. This repository of network state provides the base of advanced analytics capabilities to analyze devices in local or remote datacenters or virtual routers in various public cloud services.



Arista Cloud Tracer

Arista Cloud Tracer provides visibility of network paths and services across multiple environments. Regardless of the location of your EOS device, on premises in a hybrid cloud, in a private cloud or out in one of the many public clouds. Arista Cloud Tracer can give you visibility into the latency, jitter, packet loss and http response times of remote applications. With granular visibility captured and stored on CloudVision Portal allows for historical performance analysis.



CloudVision Portal 3rd party integrations

CloudVision API

The foundation in the flexibility of CloudVision Portal and the CloudVision framework is the Northbound API of CloudVision Portal. Everything that CloudVision Portal can do, can also be done via the API. This makes it easy to integrate with third party software as well as in-house developed provisioning and orchestration software.

Some of the 3rd party integrations that exist are:

- Ansible
- Puppet
- ServiceNow

Ansible

In addition to the EOS Ansible modules available for direct provisioning of Arista EOS devices, an Ansible module is available that provides access to CloudVision Portal and create configurations to be provisioned on CloudVision Portal managed devices. The module is highly customisable to provide relevant use cases, since it is written in Python inside the boundaries of the Ansible framework. One such example available when the CloudVision Ansible module is installed, is the ability to configure VLANs on select devices and interfaces to provide a way for server operation teams to simplify their deployment without compromising the rest of the switch configuration maintained by network operation teams.

Puppet

Similar to Ansible, a Puppet module is available that enables the creation of configuration for the network when deploying compute resources. This configuration is provisioned through CloudVision Portal on the CloudVision Portal managed devices. Alternatively, Puppet may be installed on the Arista EOS devices directly through the standard Puppet Enterprise RPM.

ServiceNow

Many customer deploy ServiceNow for managing change within IT Operations. CloudVision has been integrated with ServiceNow via the REST APIs available within each system to create a seamless workflow for change management. The high level review and approval process remains in the ServiceNow system and through the integration, the creation, scheduling and execution of change tickets and CMDB Configuration Items may be automated via CloudVision Portal. When a task is generated in CloudVision Portal the Service Now integration creates a ServiceNow Change request. The Change request can now be treated as any ServiceNow task including approval procedures and scheduling execution of tasks. When the ServiceNow change request has been approved and scheduled time and date arrives, task execution inside CloudVision Portal is triggered from Service Now. The ServiceNow Change request is then automatically updated and closed when the CloudVision Portal task is completed. In addition, when new network devices are provisioned from CloudVision Portal, the CMDB is populated with these new devices in an automated fashion.

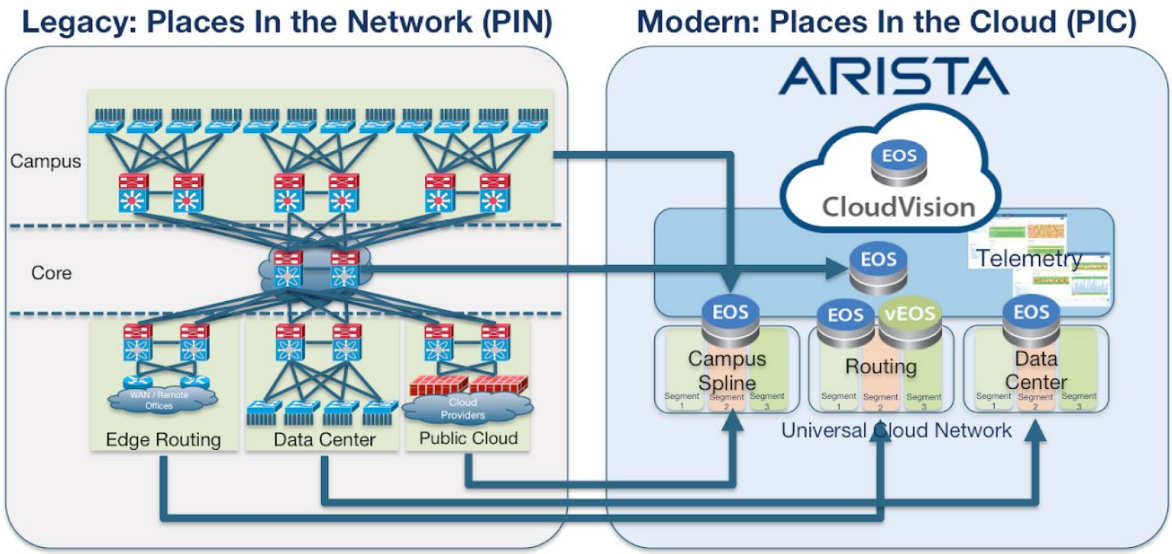
Network Topologies Overview

To simplify the maintenance and monitoring Arista recommends using a consistent architecture across all leaf switches and cabinets. This ensures operations teams can easily support any portion of the network without having to relearn how each section of the topology is constructed.

The main design consideration for each type of Leaf is the oversubscription ratio between the downlink ports and the uplink ports and the overall function of the leaf. Most leaf switches can be grouped into three major types: Datacenter, Routing, and Campus. Please see Appendix C for detailed port configurations for Leaf devices.

As the Universal Cloud Network becomes a reality, it is important to begin migrating the various Places in the Network (PINs) to Places in the Cloud (PICs). This will accelerate the advantages that customers are able to reap from a modern network design.

As introduced in the Universal Spine, the PINs will be subsumed by the Datacenter PIC, Edge Routing PIC, Public Cloud PIC, and the Campus PIC:



Datacenter Leaf Designs

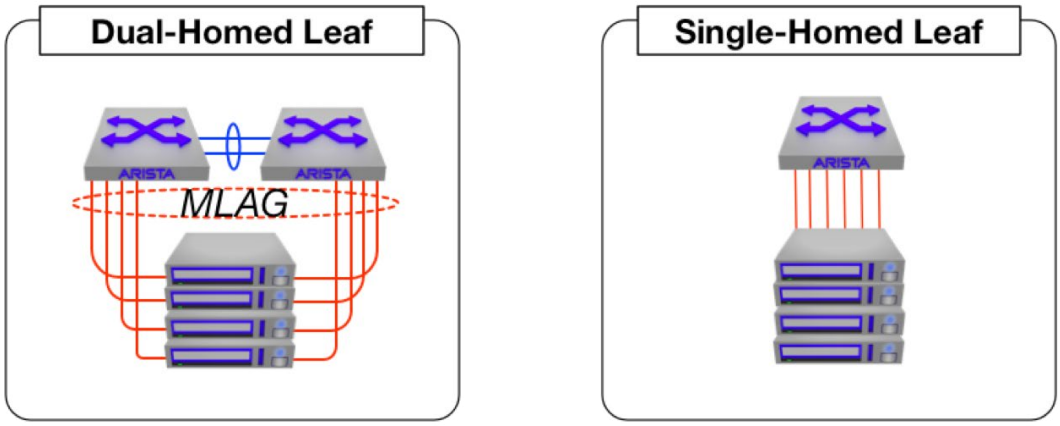
Single / Dual Homed Compute Leaf

In this leaf model each host is single or dual homed to one or two separate leaf switches. The dual-homed model has the leaf switches interconnected. These two switches advertise themselves as a single LACP peer to all connected hosts. The host sees a single LAG group connected to a single LACP peer.

Additionally, in a dual homed model each leaf switch acts as the first-hop router using VARP, a stateless method of providing default gateway redundancy. Unlike HSRP or VRRP, both VARP enabled leaf switches route, allowing full use of available bandwidth.

Compute leaf switches commonly connect at a 3:1 oversubscription ratio or better - this allows up to 48 10/25GbE ports to be used to connect to hosts and telemetry devices and leaves 6x40GbE, 6x100GbE, or 12x100G interfaces for uplinks.

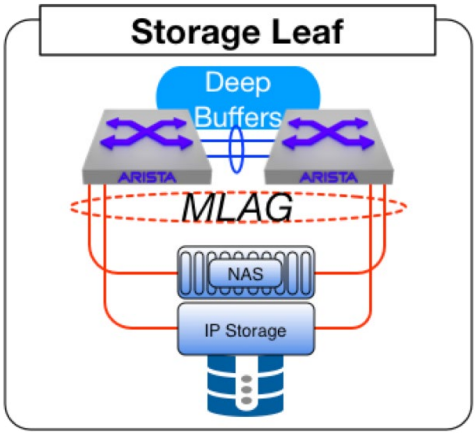
A separate 1GbE Arista switch that aggregates the lights-out management ports can provide the out-of-band management of each host. It can either connect to a completely separate OOB management network, or to the leaf on a separate IPv4 VRF. The OOB Management model will be consistent across all leaf models.



Storage Leaf

The storage strategy will have a key influence on the architecture of the datacenter IP fabric. The technology choice - whether it be block, file, IP, Ethernet - will affect not only the logical topology of the infrastructure (layer 2 or 3), but also the flexibility and operational complexity of the switching platforms that can be deployed.

Storage entails high levels of incast traffic and often interface speed mismatches. It is not uncommon to have the storage infrastructure teams upgrade the storage nodes to 10/25GbE or 40/50/100GbE while some of the devices accessing data are doing so at 1Gb or 10Gb. This will lead to significant packet loss and retransmission reducing overall application performance.



Storage is an often-requested resource requiring the network to handle a significant amount of concurrent, bursty transactions. Because of this, oversubscription ratios in the storage network are often 1:1 or 2:1 and rarely higher. The devices used for storage traffic should have larger buffers and also may have advanced monitoring features like latency analysis. Any problem on the storage network will affect hundreds or thousands of hosts and applications, so traditional reactive network management is not sufficient.

Although Arista does not see FCoE as the way forward for storage in the datacenter, there will be a requirement at small scale to bridge into the FC infrastructure. The Arista platforms provide support for IEEE Datacenter Bridging to allow the transport of lossless Ethernet (IEEE 802.1Qbb Priority Flow Control) traffic. Its use however needs to be carefully considered as it has the potential to impact all traffic types, not just the specific ones it is deployed to assist. It also introduces a significant level of administrative complexity.

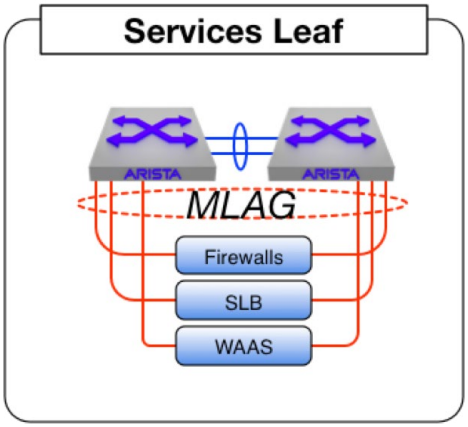
With the adoption of 10/25/40/50/100 GbE, IP-based block (iSCSI) and file (NFS) storage present a better-cost per-performance for the datacenter. These are established technologies, which can take advantage of the multi-pathing capabilities of the leaf-spine architecture to provide linearly scalable performance. Application performance can also be guaranteed using standard QoS frameworks to ensure the storage traffic is correctly prioritized and serviced as it traverses the leaf and spine nodes.

The IP approach to storage is also "Server centric" due to the fact that it can be transported over a standard IP infrastructure, including traversal of layer 3 boundaries. While providing a greater level of flexibility with regards to the deployment of storage nodes, iSCSI and NFS don't present any additional overhead for the network operation and management as they utilize tried and trusted TCP and UDP protocols. It is this simplicity and flexibility that iSCSI and NFS offers that makes them a better fit.

While IP storage can be delivered over any L2/L3 infrastructure. To achieve optimal performance, the latency should be as low as possible between storage writes and the acknowledgement from the storage target. This allows the host to process data and applications more efficiently.

A second and sometimes overlooked problem with IP storage is the bursty nature of the traffic flows. These traffic bursts can consequently result in traffic loss within the network infrastructure due to buffer exhaustion. In a leaf-spine architecture this is often observed on the spine nodes, as traffic bursts converge from multiple leaf nodes. It can also be experienced at the leaf node when a large number of storage devices are connected to a pair of leaf switches. Within the proposed design careful consideration can be given for both the leaf and spine nodes. The 7500R and 7280R series are well suited for this operation due to their large buffers enabling up to 50ms of buffering per port. These features allow the spine and/or the leaf to absorb the microbursts and avoid packet loss. This reduces retransmissions thus improving efficiency and overall performance of the storage.

Services Leaf



Advanced services such as security, application front ends, VPN appliances, wireless controllers and WAN optimization usually go into one of three locations in a datacenter design: a dedicated services leaf or series of leaf switches distributed amongst the compute as virtual appliances, or deployed on or attached to the network edge.

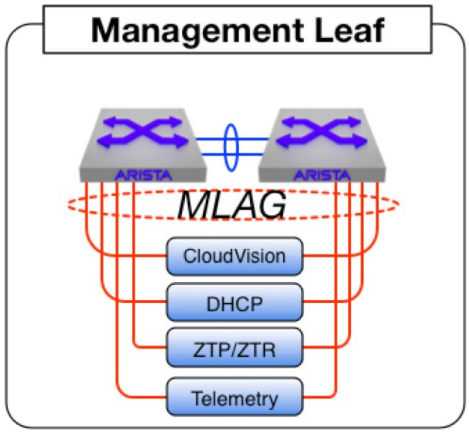
Deploying dedicated leaf switches for the Network Services centralizes all of the services and also simplifies service chaining between the compute and the storage. When deploying this architecture, Arista recommends provisioning uplink capacity inline with the maximum throughput of the aggregate number of attached services to avoid any oversubscription and congestion.

Virtual appliances are an increasingly popular trend as the provisioning of a service is very simple and the services can be repositioned along with its supported workloads. Many appliance vendors are bringing virtual appliances to the market for this and many more reasons. While the performance is generally not as high as the hardware-accelerated appliances, it is possible to distribute these across many servers to deliver the performance required.

Management Leaf

Arista recommends the use of the compute leaf architecture for providing management services to the rest of the network. This includes shared network services such as DNS, DHCP, ZTP/ZTR, Log Analysis, any SDN controllers and orchestration/automation systems.

These services are generally not high throughput so the management leaf often is designed with higher-than-normal oversubscription while maintaining the maximum levels of resilience and reliability.



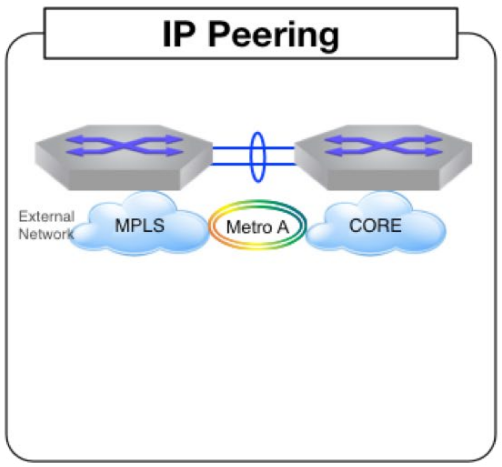
Routing Leaf Designs

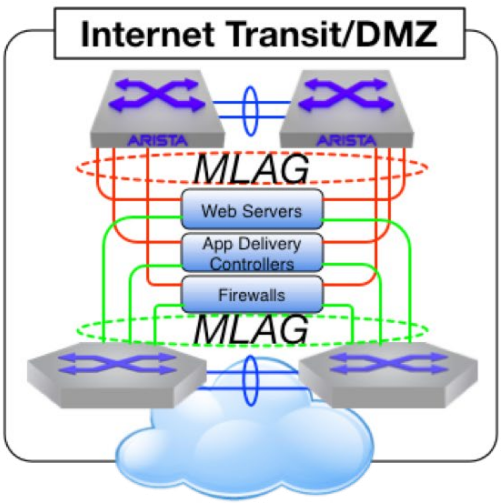
Internet Transit/DMZ and IP Peering Leaf

The Internet Transit/DMZ leaf is the primary external ingress and egress point from the Internet into the datacenter. Peering with external transit providers and exchanging of Internet scale route tables can also be performed within either of these leaf designs. The Internet/DMZ leaf will often house services specific to Internet filtering, load balancing, optimization, and security prevention. It is designed to uplink directly to the spine once the traffic has been inspected and scrubbed to maximize throughput between the Internet business services and the internal resources.

A secondary use of the Internet Transit/DMZ leaf would be to leverage it for extending external services to the internal offerings. This would be done using VXLAN as the encapsulated transport from the DMZ into the specific location inside of the network. Services such as wireless mobility or specific service offerings could take advantage of this Internet firewall bypass functionality without compromising the internal security of the enterprise.

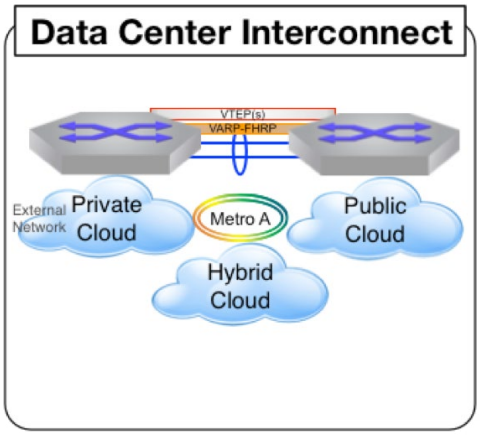
The Internet IP Peering leaf is the peering point with external providers and is capable of carrying Internet scale routing tables. Large routing tables in excess of 1-2M routes and high performance with low power per port and high density enable the Arista 7500R series and 7280R series to fit seamlessly 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.





Datacenter Interconnect

The Datacenter Interconnect leaf provides the ability to interconnect one or more datacenters, but can also be considered as the launch point for private, hybrid or public cloud solutions as a basis for a customer's multi-cloud solution.

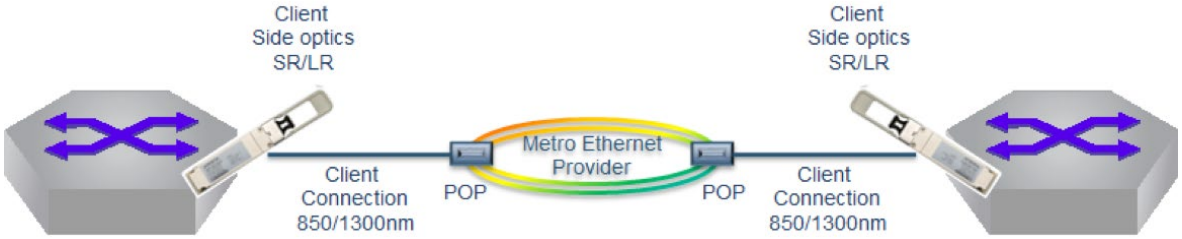


Physical Data Interconnectivity

From a L1 standpoint, Arista provides a number of solutions for providing direct access between datacenters, from using client optics connecting into a managed Metro/WAN service provider, to privately owned DWDM solutions.

DCI Managed Solution

For a managed solution, customers lease Metro/WAN ethernet solutions from service providers between their datacenters, the appropriate client side optics are then used to match the service providers connection connecting them directly into the DCI switches. Providing a point to point connection between sites allowing it to be configured as a L2 or L3 interlink.



This solution scales as you need to grow your bandwidth, but is limited to a specific speed per connection, i.e. each connection will be provisioned at a set speed and be become inflexible as bandwidth requirements increase.

Managed Wavelength Solution

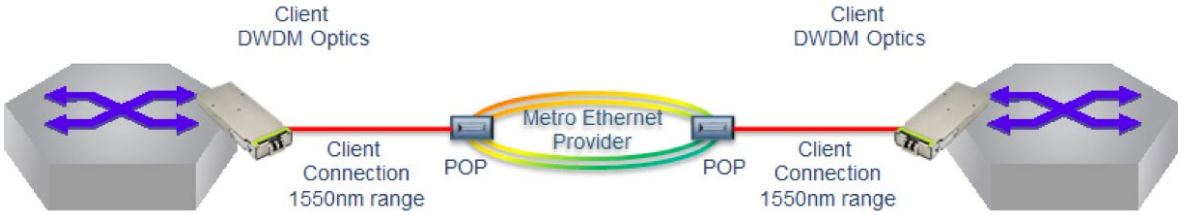
Similar with the managed connection solutions, clients can lease a wavelength or Lambda from their service provider. The service provider handles these connections as "Alien Wavelengths" within their DWDM network and as such only switch light along a managed path between the customers datacenters. The advantage is that customers can provision the connection at the speed of their choosing. Arista provides two optics options for this solution:

- 10G SFP DWDM (SFP-10G-DZ-T)

  These optics will work in any 10G SFP+ cable switch, are tunable to any lambda in the Full C-Band 50 GHz ITU Grid and drive up to 80km.

- 200G CFP2 DWDM (CFPX-200G-DWDM)

  Thes optics work with the 7500R DWDM 8 port line card and offer the flexibility of 100G, 150G and 200G speeds and will operate over distances in excess of 5,000Km at 100G. Each optic is tunable and will work on the 100GHz, 50GHz and 37.5GHz ITU Grids. Further the line card offers MACSec encryption at line rate at the speed selected.

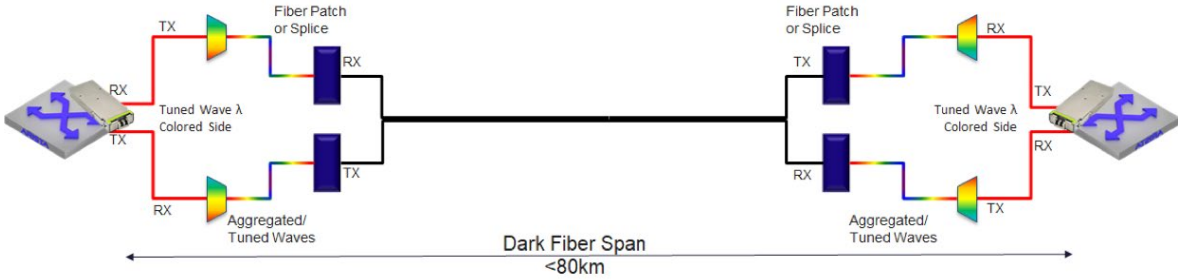


Private DWDM Solution

Where the scale of bandwidth and complexity of a multi-site DCI solution make a managed solution uneconomical, customers may want to implement their own private DWDM network. Dark fibers should be provisioned by a service provider between locations. Customers should then decide on creating either and Active or Passive DWDM solution.

Passive DWDM

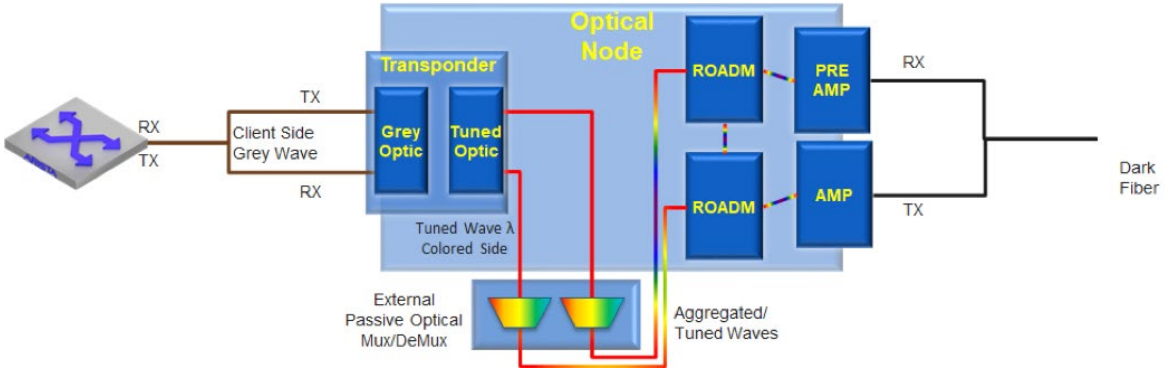
As the name implies uses passive DWDM components rather than an integrated system, care should be taken when selecting this option as a number of factors will dictate the maximum distance between DCI nodes, including quality of fiber, number of connections or splices, each of which will cause attenuation and degradation of the signal.



Arista provides three optics that will work with this solution, the SFP-10G-DZ-T and CFPX-200G-DWDM which have been discussed in the Managed Wavelength Solution. Another option is to use the 100G QSFP DWDM (QSFP-100G-DZ2-xx). It is important to understand that this solution requires a higher level of engineering.

Active DWDM

Active DWDM solutions provide a more robust solution in a multi-site optical topology, allowing for wavelengths to pass between two endpoints directly or indirectly via another location using ROADM (Reconfigurable Optical Add Drop Multiplexers). Traditionally DWDM nodes have been used to convert client side (grey) optical signals to lambda's which are then MUX'ed together through a passive mux and then sent over a dark fiber connection via a ROADM.



Arista's solution produces the lambdas directly from the source switch, negating the need for client side optics, expensive transceiver line cards with both client side and channelized optics to convert the signal to a lambda. These optics are tuned to a specific lambda and connected directly into the same passive Mux, and on to the dark fiber via the ROADM.

Customers just need to create Network Circuits between DWDM nodes and introduce the Arista Lambda signal as an Alien wavelength. The image belows shows a two Arista switches connected together with redundancy connections, each connection travels through the DWDM network via different paths ensuring diversity.

Arista provides two optics that will work with this solution, the SFP-10G-DZ-T and CFPX-200G-DWDM which have been discussed in the Managed Wavelength Solution.

Campus Spline Design

Campus Spline Requirements

The total number of devices that are in the campus network will continue to grow with the addition of IoT devices. It will become increasingly important to provide services for these devices similar to that of other connected hosts. To help manage these devices a new cloud paradigm must be employed which includes cloud scale, cloud reliability, and cloud automation.

Campus Cloud Scale

The Campus Spline has two primary designs that can be deployed depending upon the closet leaf connectivity and customer needs. As with every Arista Spine switch, the most common scale limitation is the number of front-panel ports for the device. For smaller customer campus networks, a single Campus Spline pair of 1RU switches will allow scale out of 30+ closet leaf switches connected at 100G. For the larger customer needs it may be necessary to utilize a chassis based ECMP Campus Spline to maximize the total number of closet leaf devices that can be connected back. Implementing campus leaf-spine designs can deliver better price performance than legacy three tier, access-aggregation-core, architectures. Removing mid-tier IDFs reduces equipment count thus saving costs while also increasing reliability. Speed will also be a large determining factor and customers should use 10/25/40/50/100G connections between the closet uplink ports and the Campus Spline. Lastly because the Campus Spline is built on the same datacenter principles, the use of modern technologies such as VXLAN to provide L2 adjacency can also be considered.

Cloud Reliability

Many host devices will have a single connection to the closet leaf switch. The recommended design is to have multiple uplinks from each closet or access device to the Campus Spline. Newer generation, active-active, load sharing technologies, can improve spine to leaf bandwidth utilization while also ensuring reliability. Similarly, active-active collapsed spine platforms should rely less on brittle, active-passive control plane architectures and instead provide hitless maintenance and upgrade features that avoid network degradation let alone failure.

Cloud Automation

Being able to manage the Campus Spline as another set of switches connected to the Universal Spine has many advantages including:

- Standardized Software Releases
- Industry standard CLI
- Uniform Telemetry through CloudVision
- Management and Change Control with existing CloudVision Systems
- Automated patching and upgrades

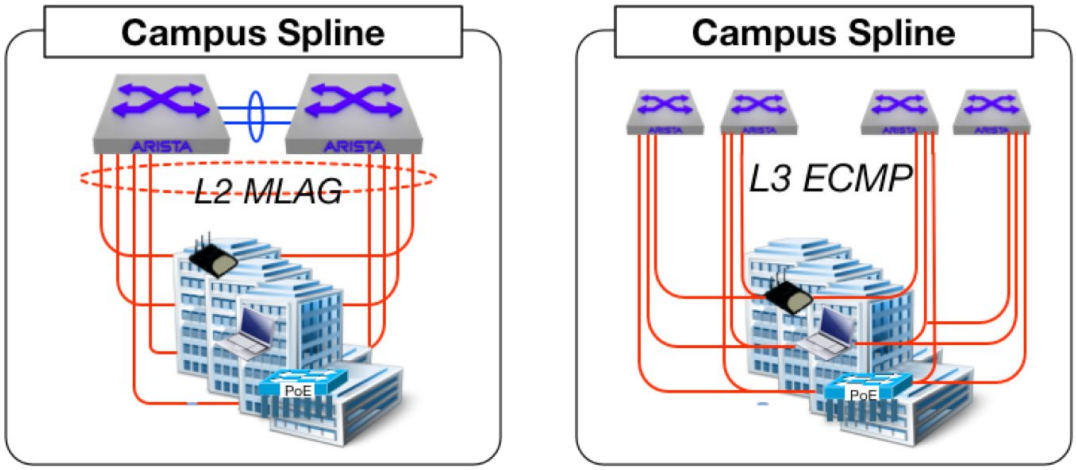
For those customers that want a single pane of glass to manage their entire datacenter and campus, Arista CloudVision can provide exactly that. Given that CloudVision and EOS work seamlessly together today, the addition of a new capability in the switches with the introduction of the Campus Spline means that day one CloudVision will be able to wrap the Campus Spline in a set of rich automation features such as change control, orchestration, telemetry and even code upgrades.

Design Considerations

As depicted below, there are two unique types of Campus Spline designs. The first of the two employs the traditional L2 MLAG model that will then connect up to the various devices and switches throughout the campus. This model provides for simple L2 services across all the leaf devices as well as an Active/Active topology.

The second design is for the customers that have a more robust campus need AND have the appropriate amounts of cabling that is backhauled to the Campus Spline. In this design each leaf switch would be L3 ECMP connected back to all 4 Campus Spline devices. L2 Adjacency services would be provided through a mechanism such as VXLAN. This design also allows you to put any of the Campus Spline devices in maintenance mode and remove it from the overall load balancing, where intrusive operations can be performed such as disruptive network configuration or code upgrades.

In both of these designs, each Campus Spline switch would be connected back to the Arista Universal Spine within the datacenter. It is recommended that these devices be connected at 40/50/100G in order to have adequate bandwidth to all the campus leaf switches.



Arista UCN Scale-Out Designs

An accepted principle of network designs is that a given design should not be based on the short-term requirements but instead the long-term requirement of how large a network or network pod may grow over time. Network designs should be based on the maximum number of usable ports that are required and the desired oversubscription ratio for traffic between devices attached to those ports over the life of the network.

A two-tier leaf-spine design has spine switches at the top tier and leaf switches at the bottom tier. In a two-tier leaf-spine design, every leaf switch attaches to every spine switch. The design can be built at either layer 2 or layer 3, however layer 3 designs scale higher as there can be more than 2 spine switches and MAC entries and host routes are localized to a given leaf switch or leaf switch pair.

When a collapsed leaf/spine model is not desirable Arista recommends one of three leaf/spine topologies. As explained in the following sections each design will meet certain scaling and functional requirements.

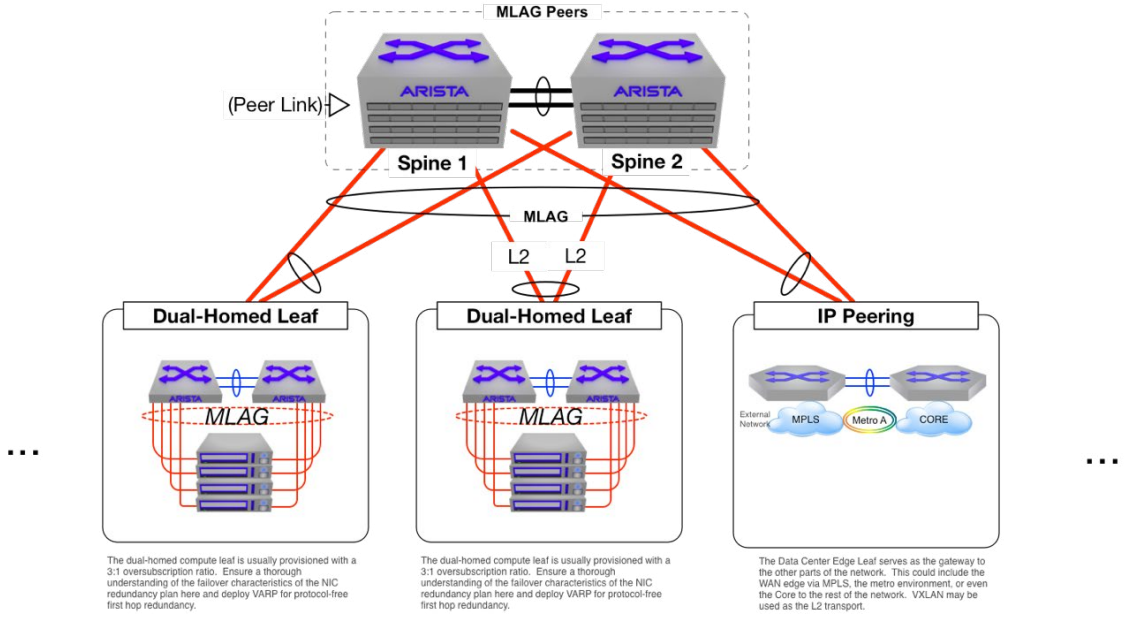
- Layer-2 Leaf-Spine (L2LS) Network
- Layer-3 Leaf-Spine (L3LS) Network
- Layer-3 Leaf-Spine w/VXLAN (L3LS-V) Network
- Arista AnyCloud

Layer-2 Leaf-Spine Network Design

Topology Overview

The Layer 2 Leaf Spine (L2LS) network design is a two-tier architecture comprising of two spine switches and one or more Top-of-Rack (ToR) leaf switches. Leaf switches provide connectivity to storage, compute, service, and datacenter edge network elements. Leaf switches may be deployed by themselves or in a pairs for high availability. Spine switches aggregate and provide a fast backbone for the leaf switches. Any two switches in the Arista family of devices can provide a MLAG pair – not just a chassis based product.

Layer 2 Leaf/Spine with Multi-chassis LAG (MLAG)



All-active multipathing is achieved at layer two via Multi-Chassis Link Aggregation (MLAG). MLAG is a standards-based approach that effectively presents two physical switches as a single logical device to a connected host by leveraging Link Aggregation Control Protocol (LACP). This eliminates blocked links via Spanning Tree Protocol (STP) while still providing high availability and increased bandwidth.

Key Benefits

Layer 2 designs provide the simplest design to meet the majority of the enterprise customer's requirements. These designs allow Layer 2 between racks and localize the L3 gateway at the spine switches. Hosts in a Layer 2 MLAG design utilize a default gateway configured on each spine switch with an Anycast Default Gateway or vARP (Virtual Arp) address that provides an active-active default gateway function. VRRP is also supported as a FHRP protocol.

Key benefits provided by a L2LS design are ease of workload mobility, ease of segmentation provided by 802.1Q VLANs, and in some cases it can be simpler to provision and manage. In a MLAG deployment, the spanning-tree protocol remains in place as a fallback protection in the unlikely event of a MLAG failure or misconfiguration. The simplicity and resiliency makes a L2LS a solid choice for a small to medium size deployment.

The main benefits of L2 MLAG designs include:

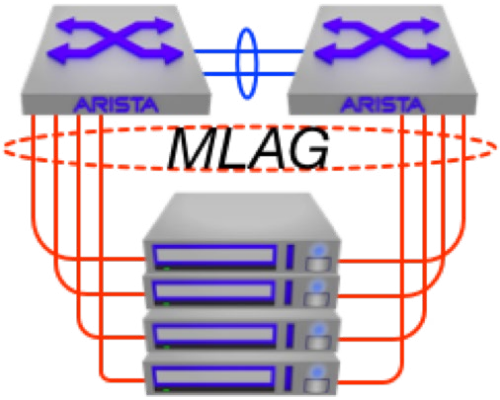
- Ease of Workload Mobility
- Segmentation proved by 802.1Q VLANs
- Simple design to provision and manage
- All links are active with the minimization of the Spanning-Tree Protocol

A L2LS design can theoretically scale to nearly 200,000 10/25Gb nodes, however particular consideration must be given to the amount of broadcast traffic, mac table sizing, and control plane limits necessary.

System Components

The main components of the L2LS design are:

Multi-chassis Link Aggregation Groups (MLAG)

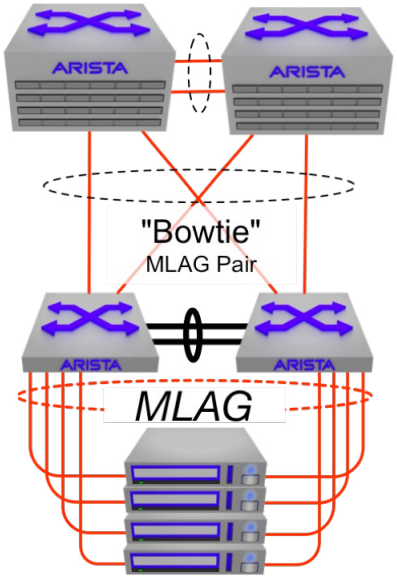


MLAG eliminates the shortcomings of STP and provides an all-active topology. Two switches, spine or leafs, are configured as MLAG peers (a MLAG domain) and synchronize over a MLAG peer link which runs between peers. The peer link is commonly a port channel to support resiliency. In a chassis based system it is common practice to distribute the peer links across line cards. The two MLAG peers advertise the same LACP device ID to a downstream-connected device. When connecting a MLAG to a standards compliant LACP (IEEE 802.3ad) device the upstream switches see the MLAG peers as a single device and forms the port channel. Spanning Tree Protocol considers this aggregated connection a single logical link and does not place any of the individual uplinks into a blocking state.

Spanning Tree Protocol is still functional in a MLAG configuration but its functionality is reduced to a fallback/safety mechanism in case of a MLAG failure or misconfiguration.

"Bowtie MLAG"

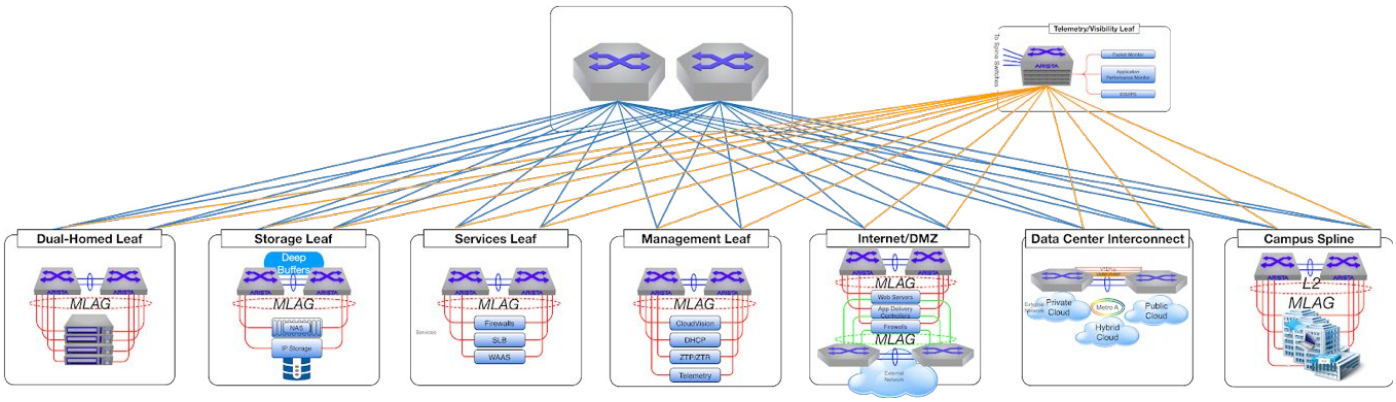
The bowtie MLAG is simply a large MLAG between two pairs of MLAG peers. All links are forwarding. This is a large improvement over the complexity of attempting to load balance VLANs across alternating root bridges – typically in an odd/even fashion.



Virtual Address Resolution Protocol (VARP)

As described above, VARP (L3 Anycast Gateway) provides first hop redundancy by sharing a virtual IP and MAC address. Both leaf switches in a MLAG pair respond to ARP requests for a virtual IP address. Virtual Router Redundancy Protocol (VRRP) can also be used in the L2LS design as an alternative to VARP if desired.

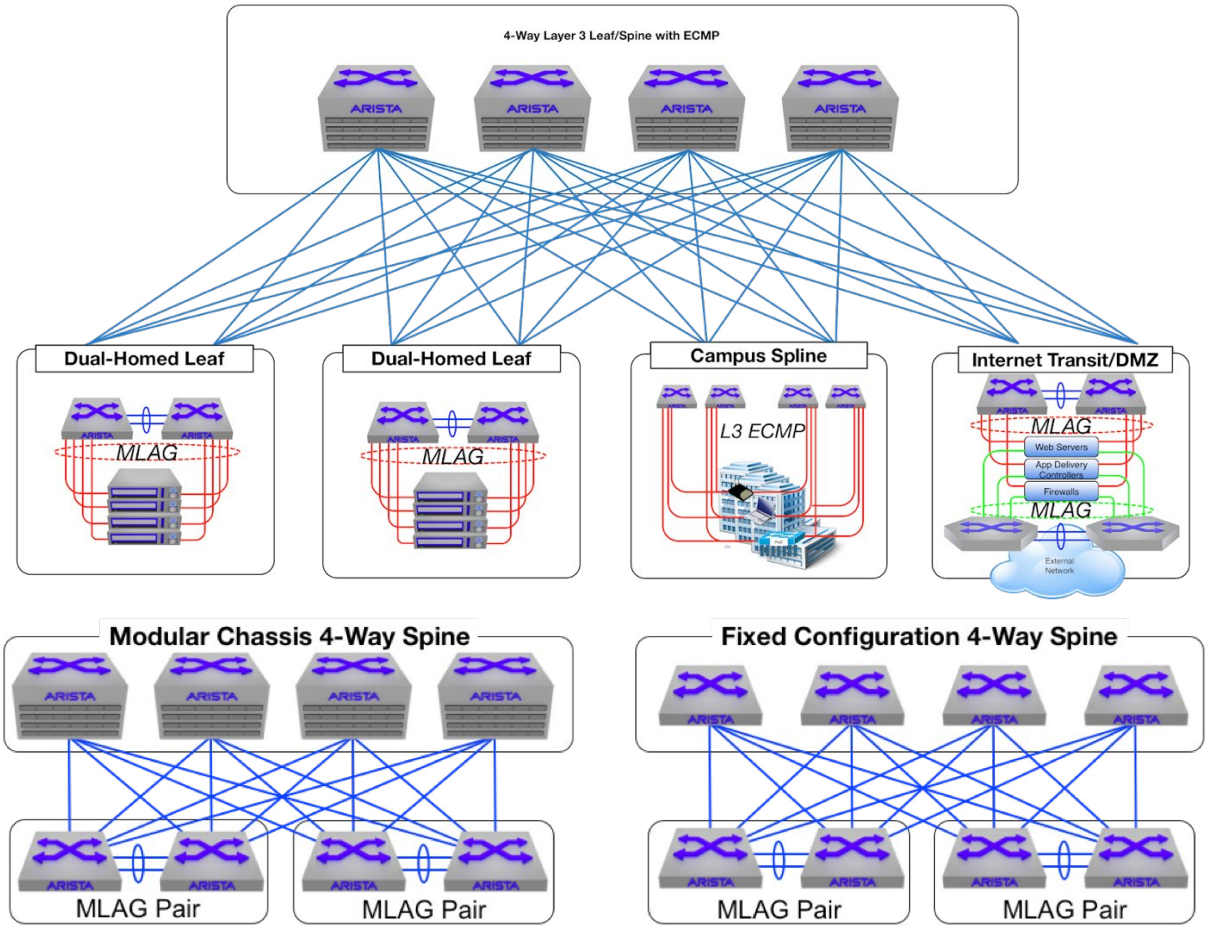
Complete Design



Layer-3 Leaf-Spine Network Design

Topology Overview

The Layer 3 Leaf Spine (L3LS) network design is a two-tier architecture comprising of 2-128 spine switches and one or more Top-of-Rack (ToR) leaf switches. Leaf switches provide connectivity to storage, compute, service, and datacenter edge network elements. Leaf switches may be deployed by themselves or in a pairs for high availability. Spine switches aggregate and provide a fast backbone for the leaf switches. Any two switches in the Arista family of devices can work in an L3 ECMP topology – not just a chassis based products. For example, use cases where deep buffers aren't mandatory at the spine layer, Arista 7300X Series chassis or the Arista 7260X would suffice depending on how far the spine layer needs to scale. Below is a topological overview of designs with Modular and Fixed switches:



All-active multipathing is achieved at layer 3 via Equal Cost Multipath Routing (ECMP). ECMP leverages a standards-based approach that effectively provides a large scale IP fabric leveraging standards based routing protocols to deliver spine/leaf connectivity.

Key Benefits

To support the maximum scalability, reliability and simplicity Arista recommends a two-tier Leaf/Spine architecture. For increased scalability and stability a Layer 3 Leaf/Spine (L3LS) design, leveraging Equal Cost Multipath (ECMP) has proven itself to be the topology of choice. For small to mid-size deployments OSPF works well as a routing protocol however for larger implementations BGP is commonly recommended for its ability to scale and overall ease of management.

Layer 3 designs provide the fastest convergence times and the largest scale utilizing fan-out ECMP supporting up to 128 active spine switches. These designs localize the L2/L3 gateway to the first hop switch and routing takes place at the top of rack. Hosts in a Layer 3 ECMP design utilize a default gateway configured on the first hop switch. The default gateway can exist on a single-homed switch or if MLAG is utilized (to support redundant connections to hosts) then an Anycast Default Gateway or VARP (Virtual ARP) address may be used to provide an active/active default gateway function, VRRP is also supported.

The main benefits of L3 ECMP designs include:

- Spine redundancy and capacity
- Ability to grow/scale as capacity is needed (from 2 to 64 spines supported)
- Collapsing of fault/broadcast domains (due to Layer 3 topologies)
- Deterministic failover and simpler troubleshooting
- Readily available operational expertise as well as a variety of traffic engineering capabilities

The countless benefits of Layer 3 designs are tough to dispute however there is limitation that needs consideration. Layer 3 designs restrict VLAN and MAC address mobility to that of a single switch or pair of switches. In doing so this can limit the scope of VM mobility to a single switch or pair of switches. In order leverage the many benefits of a layer 3 design one has to consider network virtualization technologies and the Arista L3LS-V Design. As an example, stateful vMotion of a workload from one subnet to another is not possible without extending the layer 2 domain between racks, a problem network virtualization technologies solve. Another common consideration is the interaction between the networking team and the server/virtualization teams regarding IP addressing and workload provisioning. In some circumstances a large/flat network with one large subnet is deemed easier to provision because any workload and IP address can be deployed anywhere, a challenge that network virtualization concepts also address.

Differentiators

Such designs can also be achieved using standards based non-proprietary protocols such as OSPF and BGP and can scale up to 128 ways. This design gives us the ability to eliminate Layer 2 for exceptional scalability and fault tolerance.

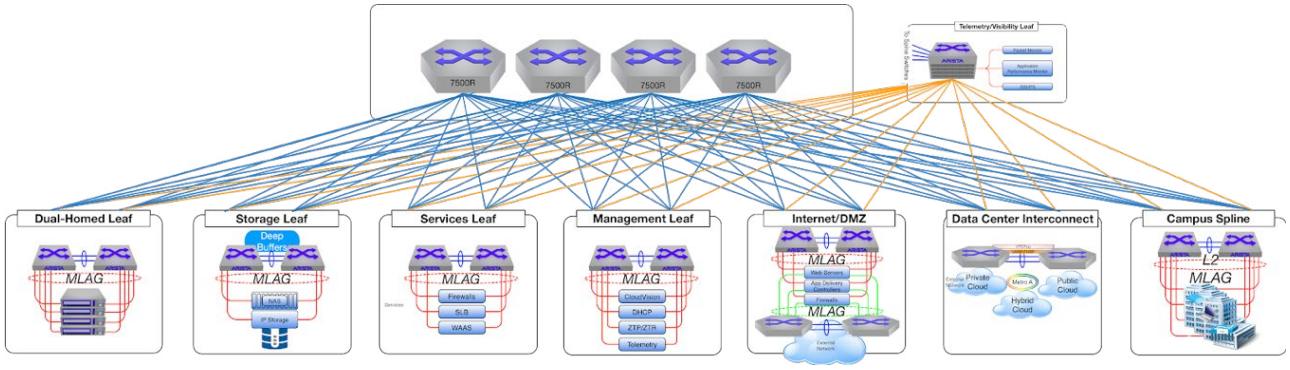
Routing Protocol

Arista recommends using BGP as the routing protocol for the ECMP architecture. This is primarily based on the rapid convergence properties of BGP and its ability to scale out without requiring a significant increase in routing protocol 'chatter'. It is also a deterministically peered routing protocol meaning that a simple cabling error is not likely to result in an adverse traffic pattern or malformed topology. For the smaller rack solution, OSPF could be introduced as an alternative, but quickly reaches the scale limits as the size of the design increases.

The recommendation to use BGP is also based on the principle of keeping the network Open, by Open we believe customers should have the choice to mix and integrate multiple vendor's products. Arista does not use proprietary mechanisms to link network layers or minimize a customer's ability to choose the best products to support their business requirements.

The BGP design for the architecture would configure the spine devices with a single BGP AS and each leaf node(s) would be assigned its own unique autonomous system number. This design leverages the inherent advantages of BGP such as deterministic performance and minimal flooding during a network failure. BGP also supports robust route-policies that can be leveraged to influence traffic paths during normal operations as well as during maintenance windows. Manipulating route policies to "drain" traffic from selected leafs/spines switches during a scheduled maintenance window is a luxury we have not realized before.

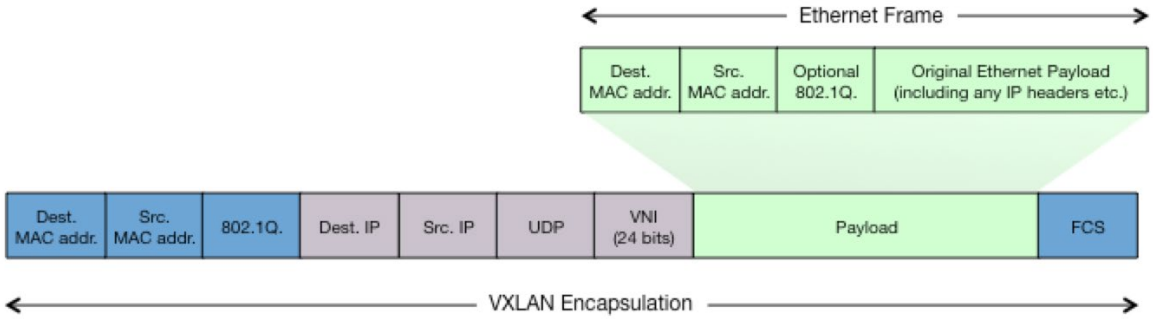
Complete Design



Layer-3 Leaf-Spine with VXLAN Network Design

The Layer 3 Leaf Spine architecture is highly scalable and provides a consistent ECMP design for workloads. It is however limited in that a Layer 2 domain is contained within a leaf node or MLAG leaf pair. In order to enable Layer 2 extension across the Leaf Spine, while at the same time maintaining the benefits of the Layer 3 Leaf Spine architecture, a network virtualization technology is required.

The network virtualization technology leveraged in the Arista Networks Layer 3 Leaf Spine with VXLAN (L3LS-V) solution is Virtual eXtensible Local Area Network (VXLAN). VXLAN is an open IETF specification designed to standardize an overlay encapsulation protocol, capable of relaying Layer 2 traffic over IP networks. VXLAN has wide industry support and was authored by Arista, VMware and others in the industry.



The main use case for VXLAN has been for network virtualization in the datacenter, allowing the creation of logical Layer 2 domains on top of an underlying IP network. Individual Layer 2 domains are identified using a 24-bit Virtual Network Identifier (VNI), allowing for up to 16 million independent domains to be specified. Layer 2 frames are encapsulated in IP UDP datagrams and are relayed transparently over the IP network. It is this inherent ability to relay unmodified Layer 2 traffic transparently over any IP network makes VXLAN an ideal technology for datacenter interconnection.

Within the VXLAN architecture, Virtual Tunnel End Points (VTEP) performs the encapsulation and decapsulation of layer 2 traffic. Each VTEP is identified by an IP address, which is assigned to a Virtual Tunnel Interface (VTI). The VTEP receives standard layer 2 Ethernet frames, selects the correct VNI and forms an IP UDP packet for transmission to one or more destination VTEPs. The source IP address is that of the sending VTI, the destination IP address is that to of the receiving VTI.

The VNI is typically determined based on the IEEE 802.1Q VLAN tag of the frame received. The destination VTEP (or VTEPs in the case of multicast or broadcast traffic) is selected based a destination to VTEP map, very similar in function as a normal MAC bridging table, except MAC addresses are associated with IP addresses rather than switch interfaces.

This network virtualization capability allows us to create logical, overlay virtual networks that are decoupled from the underlying physical network.

Key Benefits

One of the key benefits of an overlay networks is the ability to stretch Layer 2 boundaries over layer 3 or routed infrastructures. This capability delivers an optimal level of flexibility and mobility so that compute nodes can now be dynamically placed anywhere in the datacenter, removing the traditional layer 3 boundaries of the physical infrastructure. There are still many applications that require servers to share a common subnet between racks, PODs and even datacenters. Network Virtualization based on the open VXLAN standard addresses this problem.

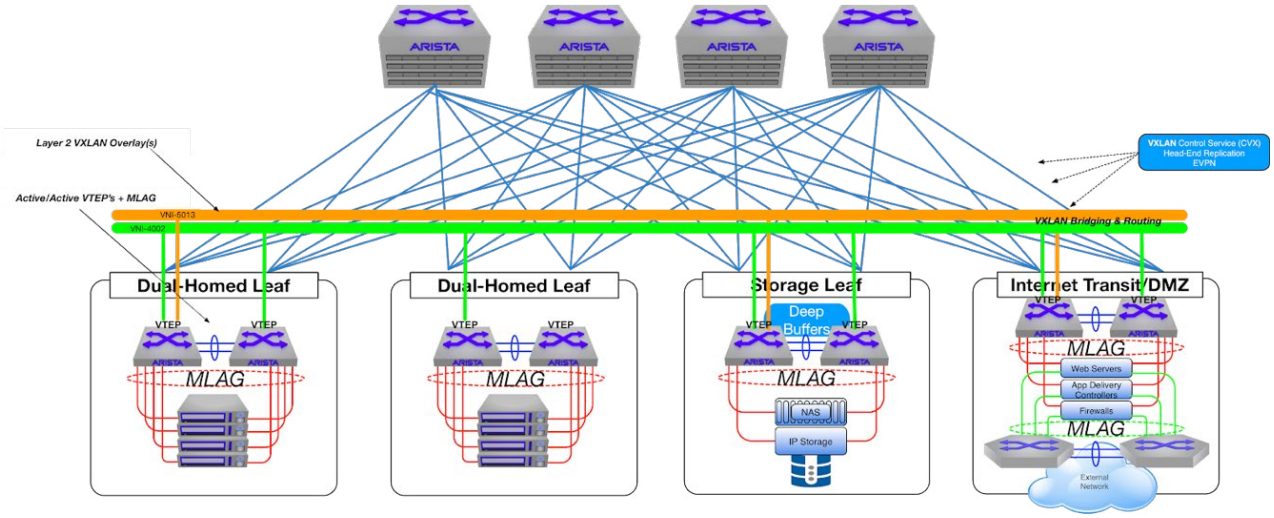
Network Virtualization also enables other teams and or automation tools to make changes to the overlay network without modifying the existing underlay/physical network. This capability can reduce the time to deploy new applications as tasks such as VM deployment and network configuration can be performed simultaneously by either a single individual or via orchestration/ automation tools.

VXLAN also allows network engineers to build the L3 Datacenter Networks they have always dreamed of. This allows engineers to collapse large Layer 2 topologies, which have plagued us for years, with Layer 3 IP based fabrics. Creating L3 "IP Fabrics" reduces the size of failure domains and becomes fundamentally more stable due to this design. Arista believes any size of datacenter can benefit from a L3 topology however larger datacenters are the most vulnerable and have the most to gain by adopting VXLAN Overlay technologies with L3 underlays.

Topology Overview

As mentioned above, Network Virtualization gives us the ability to create separate virtual networks that are decoupled and abstracted from their physical topologies. Decoupling has advantages and disadvantages. As a comparison, when virtualizing servers the hypervisor is still tightly coupled to the underlying physical hardware. The hypervisor is well aware of the state and availability of the underlying resources (Memory, CPU, I/O and Storage). This tight coupling is clearly necessary to ensure that Virtual Machines get the proper allocation of the underlying server resources.

In contrast, Network Virtualization deployments have different levels of coupling depending on the overlay model chosen. By their nature, networks have relied on their underlying protocols to ensure applications and servers received the resources they demanded. In a virtualized network protocols remain the main mechanism for data plane forwarding, however control plane learning can come in several flavors depending on the model that is chosen. The characteristics of each type of overlay are noted below.



Network Based Overlay

- As a best practice Virtual Tunnel End Points (VTEPs) reside on physical switches at the Leaf
- Tight coupling between physical and virtual network
- Data plane learning is integrated into the physical hardware/software
- Hardware accelerated VXLAN encap/decap up to the line-rate speed of the switch
- Support for all workload types, Baremetal or Virtual Machines, IP Storage, Firewalls, Load balancers/Application Delivery Controllers etc
- Native VXLAN gateway support

Hypervisor Based Overlay

- VTEPs reside at the Hypervisor / vSwitch level
- No coupling between physical and virtual networks (other than underlying protocols)
- Tight integration with Virtualization management suites
- Support for Virtual Machine workloads only (unless VXLAN Gateways are included)
- Software based VXLAN encap/decap
- VXLAN Gateway support is an add-on

Hybrid Overlay

- VTEPs reside on physical switches and hypervisors
- Integration between hardware and software control planes supports tighter coupling
- Support for all workload types
- With integration, management tools have a comprehensive view into all components of the overlay network

Due to the decoupling of the overlay network, a robust, reliable and normalized underlay network becomes critically important. Arista recommends the Layer-3 Leaf-Spine (L3LS) Topology, utilizing VXLAN capable switches at the Leaf and Spine of network. This topology is flexible and robust enough to support any overlay model.

VXLAN Control plane in the Arista UCN

The VXLAN standard specifies a learning process for overlay traffic that follows traditional Ethernet semantics - i.e. learning is achieved by the flooding of traffic with an unknown destination until a valid destination can be learned.

To facilitate a similar mechanism, the initial VXLAN standard leveraged an IP multicast based control plane. In this approach VTEP members of a VNI join an associated IP multicast group and any unknown, broadcast or multicast traffic is forwarded to each of the VTEP members of the group. This approach places a requirement for IP multicast support on the L3 Leaf-Spine IP fabric.

Multicast deployments, however, do not cater to all customer or orchestration layer application requirements; Arista therefore provides support for an additional control plane model, which is termed Head End Replication (HER).

The Head End Replication approach allows the configuration of a remote VTEP flood list for each VNI. Any Broadcast, Unknown unicast or Multicast traffic (BUM) traffic for the VNI, would be encapsulated by the VTEP and individually unicast to each of the remote VTEPs within the configured flood list. The functionality therefore removes any requirement for IP multicast support in the IP fabric for forwarding VXLAN BUM traffic. The programming of the flood-list can be achieved via a control plane such as BGP EVPN or CloudVision Exchange (CVX), it could be a manual process, or it could be automated via the Arista eAPI by a third party virtualization platform or script.

The MAC address-learning function in this model can be achieved using the normal flood-and-learn process or to reduce the level of flooding locally learnt MAC addresses can be dynamically distributed to all remote VTEPs within the specific VNI to reduce the level of flooding locally. The dynamic distribution of the MAC addresses is achieved by enabling a VXLAN control plane service. Arista EOS supports VXLAN control plane via two services: the Cloudvision Exchange (CVX) based VXLAN Control Service (VCS); or BGP EVPN.

CVX VXLAN Control Services (VCS) provides configurable, automated, data plane BUM services for VTEPs in the datacenter. VCS complements orchestration platforms by providing mechanisms that ensure reachability of all devices running on dynamically configured virtual networks. VCS serves many different deployments, providing an alternative to multicast based BUM learning. Administrators save maintenance and bring-up costs because VCS is easy to use and doesn't depend on complex IP multicast services. CloudVision VCS also coexists with alternative BUM forwarding services. The network can now support multiple hypervisor architectures simultaneously allowing more workloads to run in the same datacenter. CloudVision provides user commands and API primitives that help associate physical to virtual network topology. Administrators can use CloudVision to troubleshoot virtualized networks or can visualize virtual topologies using custom tools or commercial applications that leverage CloudVision's programmable APIs. Enhanced physical-to-virtual visualization improves monitoring and reduces troubleshooting time thus reducing costs incurred from network errors or outages.

BGP EVPN defines a control plane to compliment the VXLAN data plane. EVPN is not the only control plane available for VXLAN. There are alternative control planes such as SDN controllers and overlay orchestrators. However these alternative solutions do not run on the EVPN NVE (Network Virtualization End-Point) or VTEPs. They are independent software applications running on external servers. This architecture separates the control plane from the physical data plane. EVPN by contrast is a protocol that runs directly on the EVPN NVE/VTEP. With EVPN the VXLAN control plane runs directly on the underlay.

EVPN

Technical Overview

EVPN is a standards-based BGP control plane to advertise MAC addresses, MAC and IP bindings and IP Prefixes. The standard was first defined in RFC 7432 for an MPLS data plane. That work has since been extended in the BESS (BGP Enabled ServiceS) working group, with additional drafts published by the group defining the operation in the context of Network Virtualization Overlay (NVO) for VXLAN, NVGRE and MPLS over GRE data planes (https://tools.ietf.org/html/draft-ietf-bess-evpn-overlay-07)

In EVPN, a standards-based control-plane, multiprotocol BGP (MP-BGP), is used to discover remote VTEPs and advertise MAC address and MAC/IP bindings in the VXLAN overlay. As a standards-based approach, the discovery and advertisement of the EVPN service models can inter-operate amongst multiple vendors.

Terminology

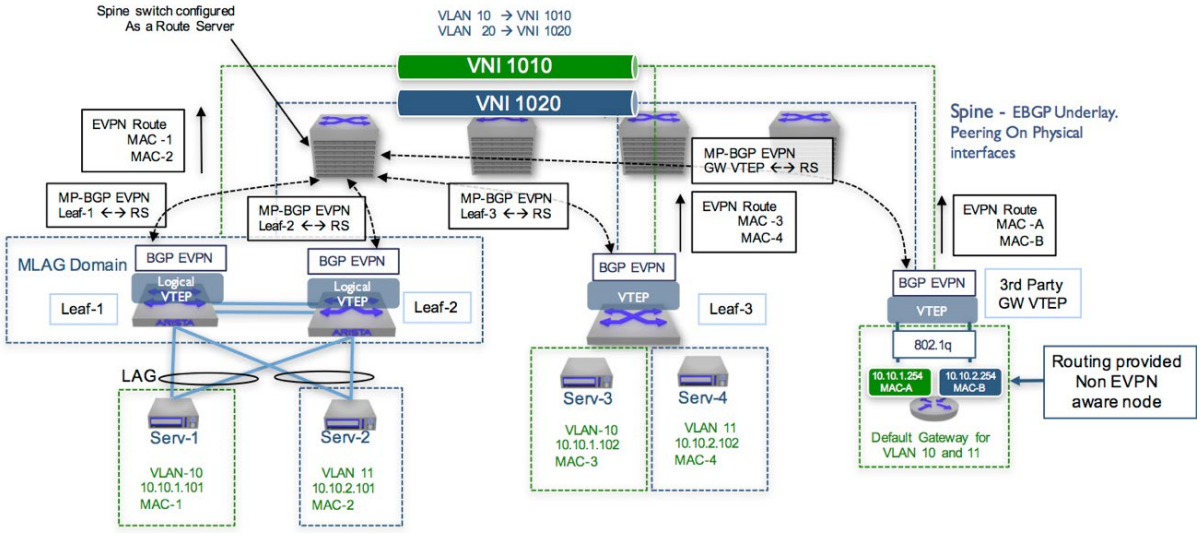
EVPN was designed to be independent of the tunnel encapsulation mechanism used in the underlay. For example EVPN supports both VXLAN and MPLS as the tunnel encapsulation mechanism. To remain agnostic the EVPN standard defines several new terms to describe components in an EVPN domain. The list below summarizes the terms used to describe the different components in an EVPN domain.

- Virtual Network Identifier (VNI) - A numerical identifier
- Network Virtualization Overlay (NVO) - An overlay domain, this can be a one or more VNIs
- Network Virtualization End-Point (NVE) - An EVPN capable network device. Also known as a Virtual Tunnel Endpoint (VTEP)
- EVPN Instance (EVI) - A logical container that represents an EVPN instance on an NVE/VTEP
- Tenant - the term tenant refers to the non-VXLAN side of the EVPN domain such as an interface, VLAN or set of VLANs that are being stretched across the EVPN domain.

EVPN Implementations

L2 Virtual Private Network (VXLAN Bridging)

Stretching a L2 domain across a L3 network is known as VXLAN bridging. In VXLAN bridging, an NVE/VTEP learns MAC addresses from its local tenant interfaces using standard source MAC learning. When an NVE learns a MAC address on the tenant network the NVE advertises the MAC address to the remote NVEs using EVPN Type 2 Route update messages via MP-BGP. When a NVE receives a frame with a destination of a remote VTEP, the NVE adds the necessary VXLAN header and routes the packet to the appropriate VTEP. When a VXLAN frame is received by an NVE, the NVE decapulates the VXLAN packet and bridges the frame to the necessary VLAN.



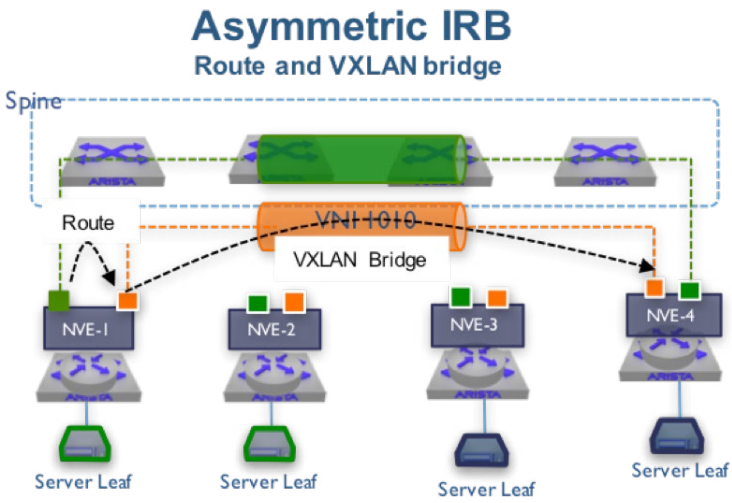
Integrated Routing and Bridging (IRB)

There are several ways to add a routing layer to an EVPN domain. One solution would be to add a L3 leaf to the spine. In this solution inter-VLAN and external traffic would be sent to the L3 leaf where it would be routed as needed. Although this is an option it may not be the most optimal solution, especially if the data is being routed within the rack. The most scalable solution is to implement L3 routing on each NVE at the leaf layer.

Routing EVPN traffic at the leaf layer is known as Integrated Routing and Bridging. IRB works by assigning an IP anycast address to each VLAN on every NVE participating in a VNI. The anycast IP address is used as the default gateway for the hosts in the L2 segment. Defining an anycast address allows the same IP address to be defined on every NVE in the domain. This means every host in the EVPN domain can use the same default gateway regardless of where the host resides in the EVPN domain. Using an anycast address also eliminates the need to run a first hop redundancy protocol. IRB is also supported in a multi-tenant deployment.

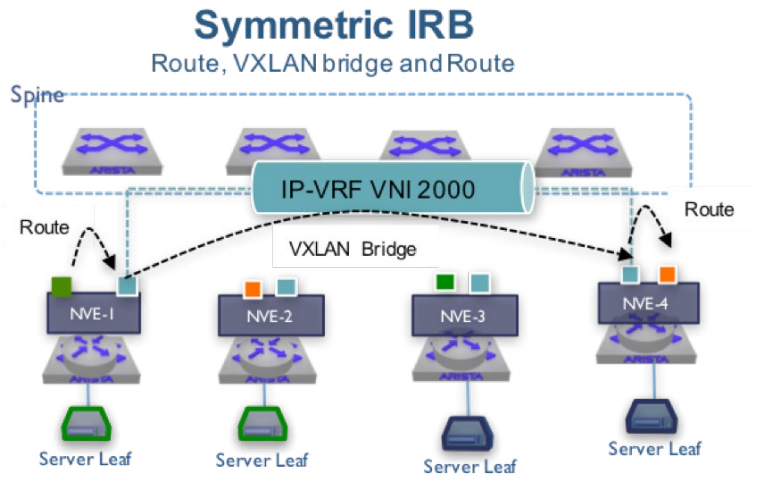
Asymmetric IRB

Asymmetric IRB is the default IRB mode in EVPN. Inter-subnet routing functionality is performed by the ingress VTEP, with the packet after the routing action being VXLAN bridged to the destination VTEP. The egress VTEP then needs to remove the VXLAN header and forward the packet onto the local layer 2 domain based on the VNI to VLAN mapping. In the return path, the routing functionality is reversed with the destination VTEP now performing the ingress routing and VXLAN bridging operation, hence the term asymmetric IRB.



Symmetric IRB

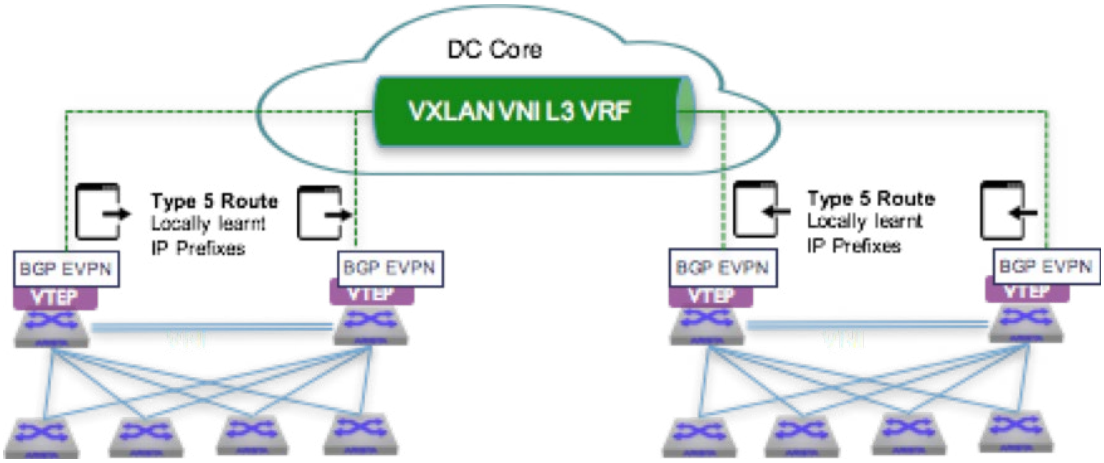
Symmetric IRB addresses the scale issues associated with Asymmetric IRB by configuring networks that are directly attached. Connectivity to non-local subnets on a remote VTEP is achieved through an intermediate IP-VRF. In this model, the ingress VTEP routes the traffic between the local subnet and the IP-VRF, which both VTEPs are a member of. The egress VTEP then routes the frame from the IP-VRF to the destination subnet. The forwarding model results in both VTEPs performing a routing function, hence the term symmetrical IRB.



L3 VPN

With L2 VPN, there is a defined bridging and routing hierarchy. Tenant interfaces are connected to MAC-VRFs. MAC-VRFs are then connected to an IP-VRF through an IRB interface. This allows hosts on the tenant network to use EVPN bridging to reach remote hosts in the same VLAN or use IRB to route packets to hosts in other VLANs. L3 VPNs simply eliminate the MAC-VRF and connect the tenant interface directly to the IP-VRF. This eliminates support for EVPN bridging, but allows the IP-VRF to establish L3 connections to the tenant network.

L3 VPNs support dynamic routing protocols between the IP-VRF and a router running in the tenant network. This allows the tenant router to advertise routes to the EVPN domain and the EVPN domain to advertise routes to the tenant router. In order to exchange routes across the EVPN domain the IP-VRF must be stretched across the EVPN domain. Stretching an IP-VRF simply means EVPN type 5 route updates are used to advertise L3 subnets learned from the tenant router to remote NVE/VTEPs.



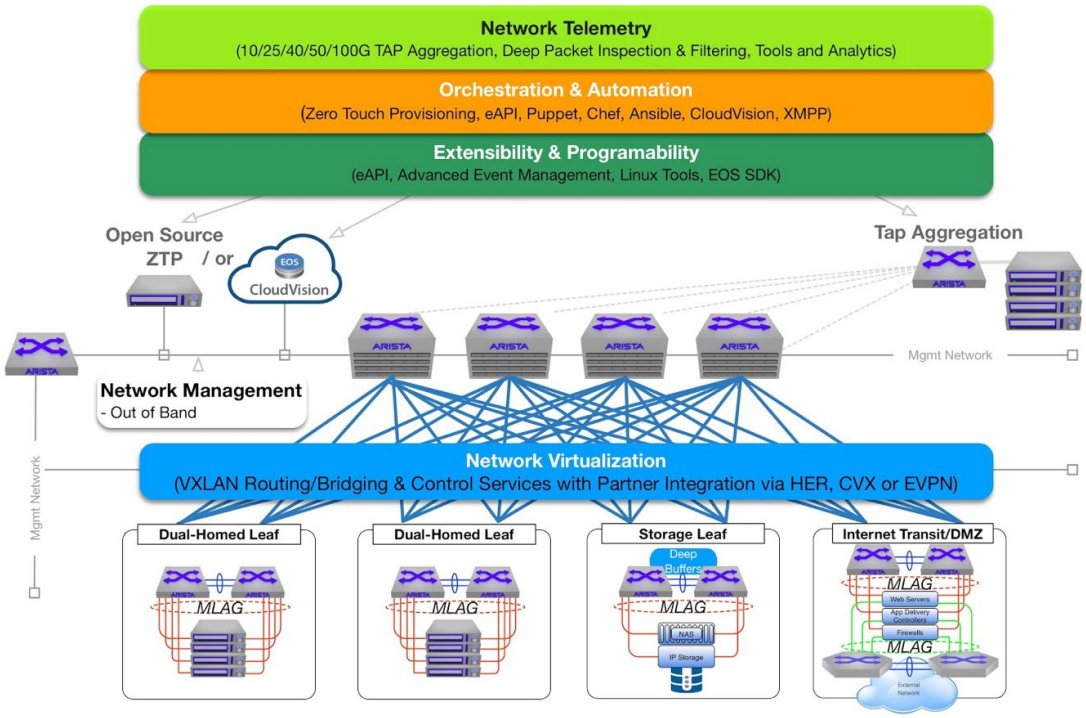
Arista Differentiators

Network virtualization is one of the primary building blocks for multi-tenant cloud hosting and large-scale virtual machine farms. Network virtualization technology compliments server hypervisors by enabling seamless workload mobility regardless of underlying network addressing and protocol choices. Arista is capable of bringing any combination of physical servers, storage, load balancers, firewalls and the network monitors into any virtual network segment. The provisioning is achieved seamlessly in software natively or via central controllers. These capabilities are based on hardware-accelerated Virtual Tunnel End-points (VTEPs) and the mapping of these between physical and virtual networking technologies.

Network virtualization overcomes the limitations of traditional VLANs through the use of overlay technologies such as VXLAN and NVGRE, and includes programmable interfaces for control by cloud providers, virtualization platforms, or provisioning systems.

Datacenter server strategy continues to shift towards a virtualized network spearheaded by virtualization at the server layer. Such a shift requires scaling MAC address limitations on the physical network infrastructure as well as the current bridging domain limitation of 4094 VLANs. Further fuelling this drive is the need to provide mobility within and across the datacenters. This also allows for optimal utilization of all hardware resources within the DC while simplifying the provisioning of on-demand resources. Therefore this solution allows for limiting the scope of failure domains, creating a simplified abstraction that identifies a customer/tenant/application for the application of network policy, enables greater scale by reducing MAC/IP host table expansion in the network and is built off of open standards with no proprietary vendor lock-in.

The goal of Network Virtualization as an overlay network is the decoupling of the physical topology from the logical topology, to allow connectivity between compute (virtual or physical) and network services (virtual or physical) regardless of where they may reside within the datacenter. This approach delivers an optimal level of flexibility and mobility, so that compute nodes can now be dynamically placed anywhere in the datacenter, removing the traditional layer 3 boundaries of the physical infrastructure.



This also allows for optimal utilization of all hardware resources within the DC while simplifying the provisioning of on-demand resources, limiting the scope of failure domains, creating a simplified abstraction that identifies a customer/tenant/application for the application of network policy, enables greater scale by reducing MAC/IP host table expansion in the network and is built off of open standards with no proprietary lock-in.

Any Cloud Network Design

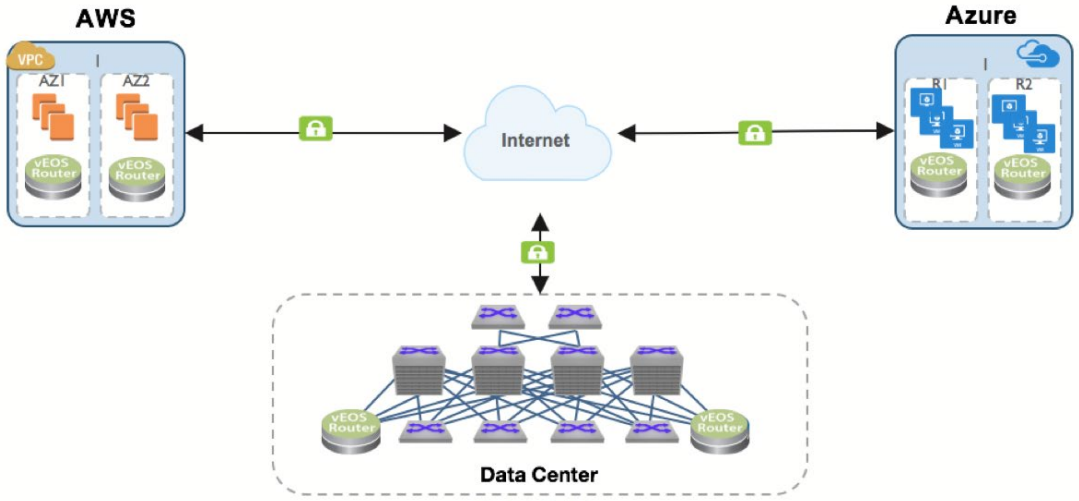
Solution Description and Components

The Arista Hybrid Cloud Solution consists of two basic components: vEOS and CloudVision. Arista vEOS is the same high-performance reliable software as Arista's EOS with added features for Hybrid Cloud. These include IPSEC and GRE tunneling, a high-performance packet forwarding engine and Cloud Tracer (synthetic telemetry probe). CloudVision provides its same reliable provisioning and monitoring capabilities with added support for Cloud Tracer Telemetry.

The Arista Hybrid Cloud Solution allows a consistent management/monitoring, configuration and control plane. This allows consistent interfaces and transport across multiple public cloud vendors as well as deployment as an NFV platform.

Topology overview

Arista vEOS, in the cloud, would serve as the default gateway in each unique availability zone for each virtual network (VNET, VPC, etc) or on the hypervisor in an NFV topology. IPSEC and/or GRE tunneling would link regions within Cloud Providers, as well as other Cloud Providers, and a customer's on-premise based network.



Key Benefits

Business requirements and strategic planning mandate today the use of public cloud services. As with self owned infrastructure the move towards vendor independence and avoidance of technology lock-ins are recommended when looking for cloud services. The target solution needs to be an easy, clear and seamless expansion, not a new technology or a break in established business processes. It is recommended to target a hybrid cloud approach always with two or more target environments in mind. This provides the benefit of competition regarding pricing and technology.

The first two key **strategy benefits** of the Arista Hybrid Cloud Solution are:

- Natural growth into the cloud by using the same operating system (vEOS-Router) as well as the same tooling (e.g. CloudVision) for deployment, operation and visibility. Visibility includes the same market leading streaming services to create unmatched analytics and telemetry services.
- By using one consistent operating system across public cloud offerings (e.g. Amazon Web Services (AWS), or Microsoft (AZURE), more to be added over time) low OPEX and low operation RISK can be achieved.

The **operational benefits** of one operating systems are:

- Limited training needs for the architecture, implementation and operation teams.
- Reuse of existing configuration and operation templates ("configlet" in the case of CloudVision).
- Consistent audibility and verification of on-premise and cloud configurations and compliance
- Minimal effort required to validate new software trains due to one image across all domains

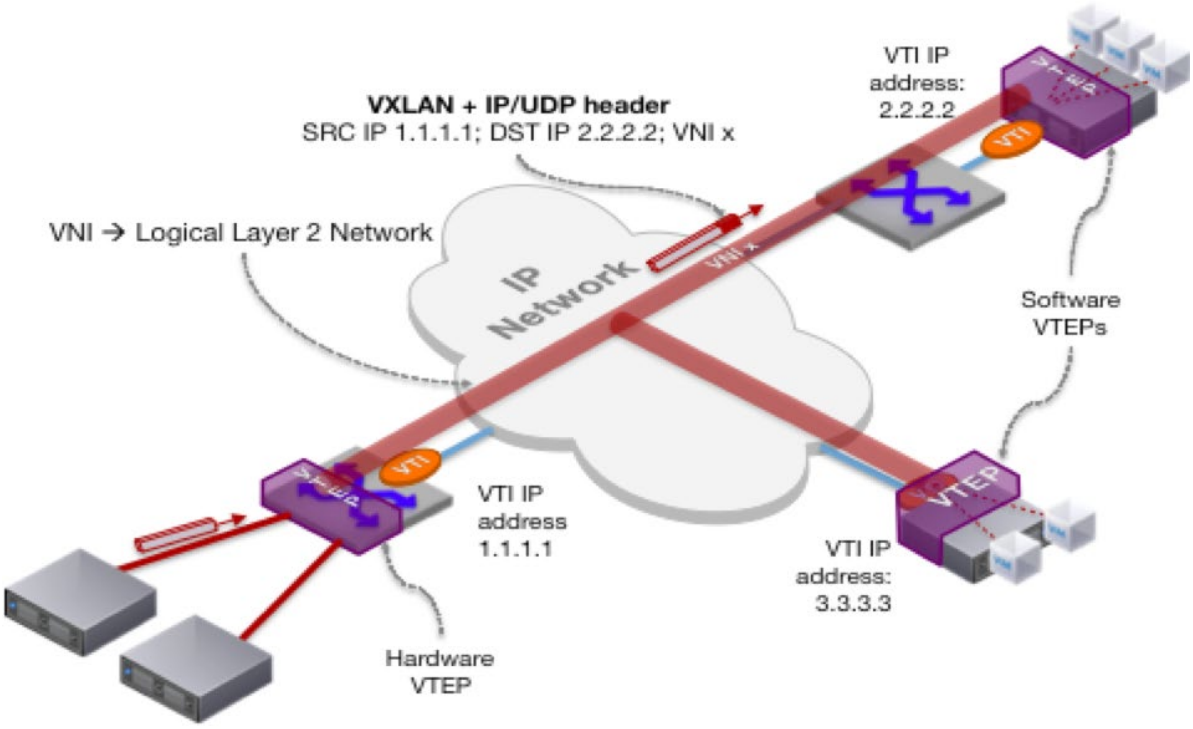
The consumption is flexible by providing either an own license or to use a bundled offering of the cloud provider. To accommodate the individual needs regarding performance and encryption, the vEOS-Router offers licensing based on discrete performance steps and a separate option to utilize US-export controlled encryption.

The **architecture benefits** often become visible once a certain size or complexity is reached. By using the Arista hybrid-cloud approach, scaling and simple connectivity can be achieved by utilizing a design comparable to the on-premise leaf-spine architecture. The use of a transit environment (e.g. Transit VPC if using AWS) is resembling the spines in a leaf-spine network. This allows for a clean and consistent architecture for inter-connecting a large number of public cloud building blocks (e.g. tenant or organizational based VPCs if using AWS).

Design Topology Decision Matrix				
Criteria	**Layer 2 Leaf/ Spine w/MLAG**	**Layer 3 Leaf/ Spine**	**Layer 3 Leaf/ Spine with VXLAN**	**Any Cloud**
Layer 2 Between Hosts	X		X	
Layer 3 Between Hosts		X	X	
Routing Protocol Expertise		X	X	
<5000 Hosts (Mac addresses)	X	X	X	
>5000 Hosts (Mac addresses)		X	X	
Minimal Loss on SW Upgrade	X			
Lossless SW Upgrade		X	X	
L2 Overlay			X	
Public Cloud Integration/Routing				X

Arista Networks DCI with VXLAN Solution

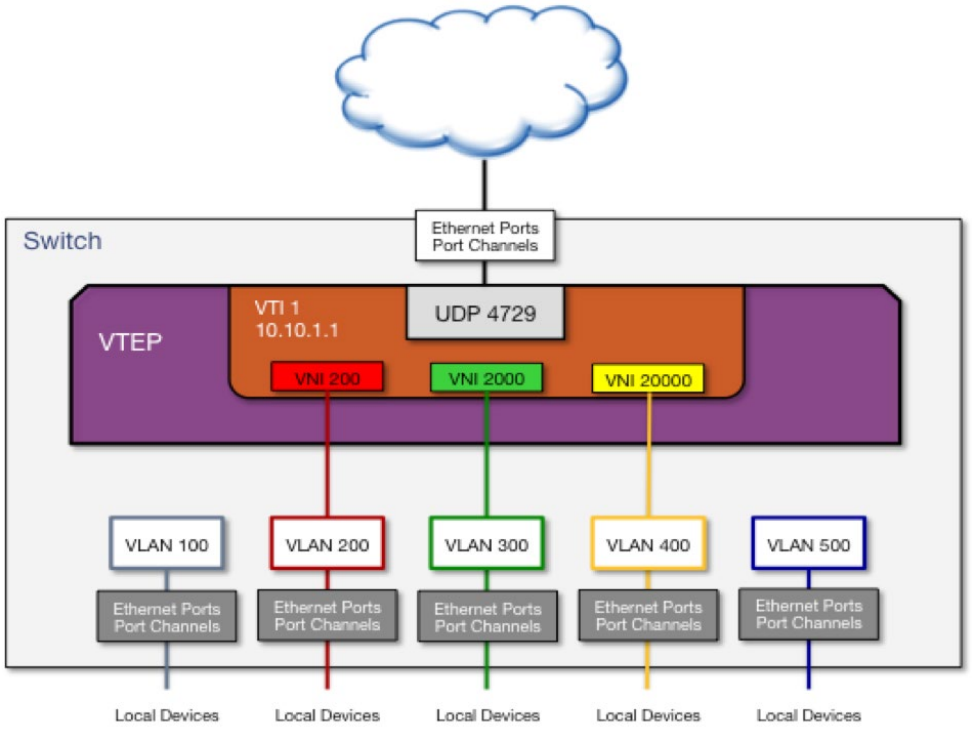
Arista DCI with VXLAN Solution Components



The foundation for Arista Networks solution for Datacenter Interconnection is the hardware VXLAN VTEP gateway function coupled with Multi-Chassis Link Aggregation (MLAG). In addition, Arista's Virtual ARP (VARP) mechanism can be used to ensure redundant first hop router interfaces are localized within the datacenter, in order to avoid traffic destined to exit the DC from consuming the DCI interconnect links.

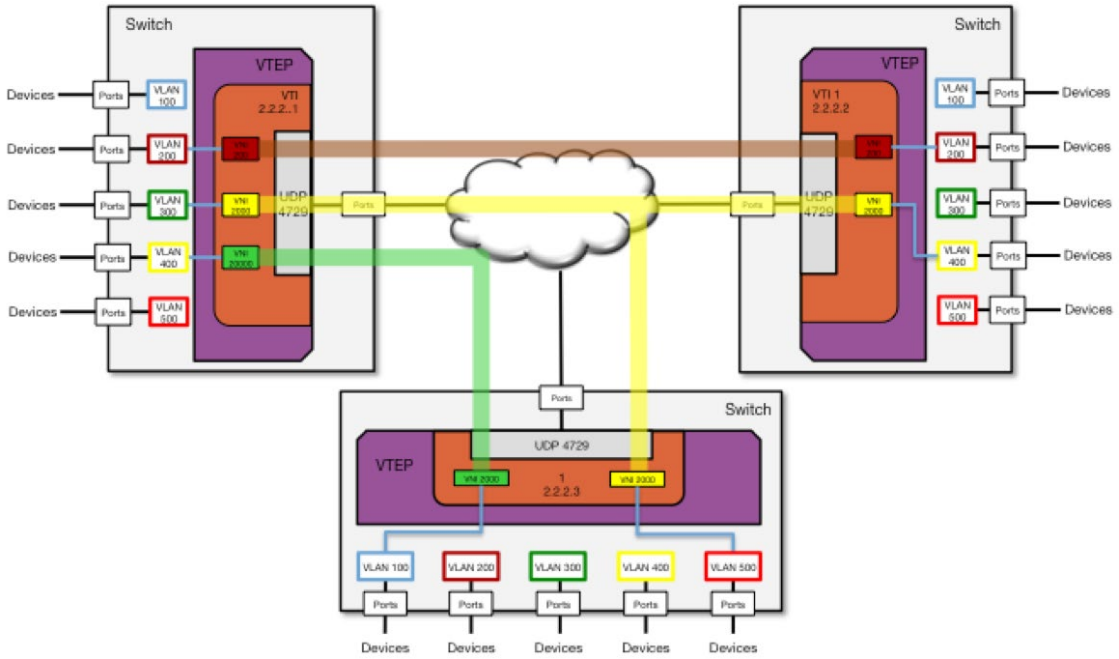
Arista Networks Hardware VXLAN Hardware VTEP Gateway

Arista was the first networking supplier to provide a hardware based VXLAN bridging, when it introduced the capability on the Arista 7150S with EOS 4.12. This version supported a single VTEP per switch, delivering the ability to transparently bridge Layer 2 over IP networks. The IP address for each VTEP is specified using a loopback interface which is in turn mapped to the VTI, referred to as the vxlan 1 interface. The default IP port for VXLAN traffic is UDP 4729, allow this can be changed if required.



Arista VXLAN Hardware VTEP Architecture

Each VTEP performs local VLAN to VXLAN VNI mapping, enabling VLAN translation to be performed as part of the tunneling process if required. The Arista VXLAN Hardware Gateway is capable of emulating layer 2 LAN networks, so both point to point and point to multi-point logical layer 2 topologies are supported with full support for the delivery of broadcast, multicast and unknown unicast traffic. Each VTEP filters spanning tree BPDUs ensuring that each DC is an independent spanning tree domain, helping isolate potential faults.

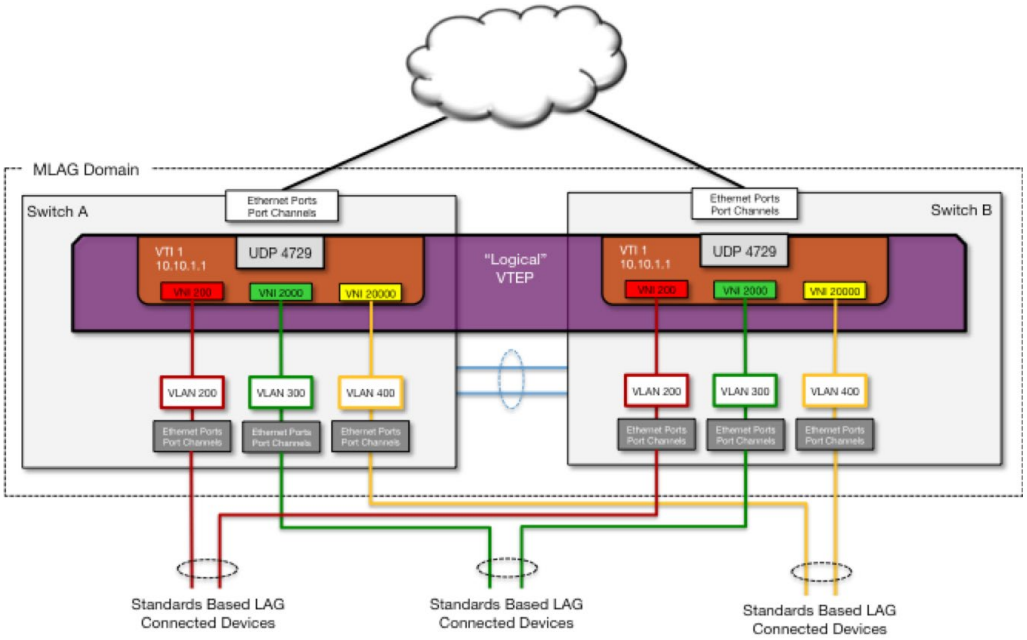


Arista VXLAN Point-to-Point & Multipoint VNIs with VLAN Translation

The initial version of Arista's VXLAN gateway handled the flooding of any broadcast, multicast and unknown unicast traffic using a single, configurable multicast group. With EOS 4.13, Arista introduced the option to use Head End Replication allow the conversion of traffic requiring flooding to be converted to individual unicast packets, thus eliminating the need to implement multicast on the layer 3 transport. As described in the L3LS-V section, Arista EOS now supports BGP EVPN as a control plane for VXLAN networks.

Arista VXLAN+MLAG

With the release of EOS 4.14, Arista introduced support allowing the VXLAN bridging function to span a pair of switches interconnected with MLAG. This allows for the implementation of a VTEP that operates on two separate switches simultaneously, effectively creating a "logical" VTEP, thus doubling performance as well as providing an active-active, fully redundant highly available system.



Arista Networks DCI with MPLS solution

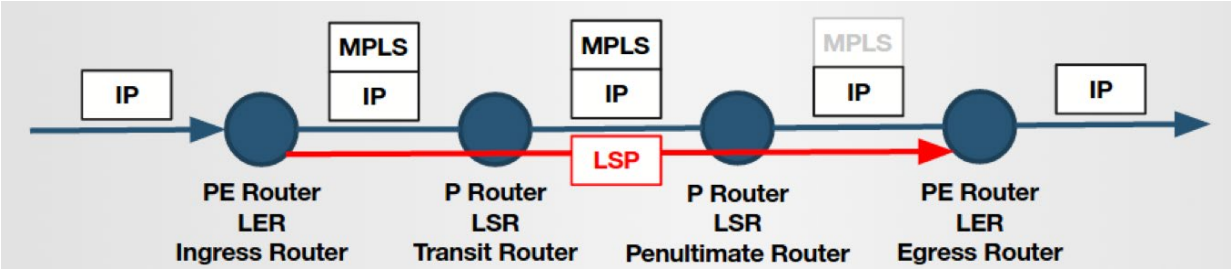
Arista MPLS

Arista currently offers MPLS support for both L2VPN (Pseudowire), L3VPN (VPNV4 BGP), and EVPN (L2/L3) services. Both Label Distribution Protocol (LDP) and Segment Routing (SR) are supported for label distribution and labels are used to forward the traffic - LDP allows customers to build simple MPLS services; where as SR offers traffic engineering and automation options. The Arista MPLS portfolio supports most of the common use cases that include DCI, SP WAN services, Mobile backhaul, and Traffic-engineering controller integration.

MPLS is supported on both fixed format and chassis based switches. Arista offers the highest density switches (form factor), lowest per port power, and the MPLS support that makes Arista the best option for customers who are struggling with power and space constraint in DC space.

Please see: https://www.arista.com/en/support/product-documentation/supported-features for specific supported features by platform.

Arista EOS supports LDP, static and even the latest in MPLS Segment routing. MPLS may be used as an overlay encapsulation that can be run within the DCI or within the datacenter. MPLS has three main actions it takes with a label: Push, Pop or Swap.



As traffic will enter an MPLS environment a label is pushed on the packet. As the label is label swapped throughout the environment a P node with simply swap labels to its peers and finally the PE at the corresponding end will Pop the label off. Starting in 4.20 release of EOS, Arista supports all components of this functionality. This means that an Arista switch can function as both a PE or P node. To provide switches within the DCI rather than routers providers the ability to have high performance switching speed and bandwidth with the possibility with using all MPLS features as in L3VPN, L2VPN, EVPN and Pseudowires.

Segment Routing for the enterprise

Segment routing is a new method of label swapping. Segment routing is a very unique way of swapping labels throughout a MPLS environment because everything is handled through the routing protocol. There is no need to have a secondary protocol on top of IS-IS or OSPF to distribute MPLS labels. In the Arista segment routing implementation IS-IS handles segment routing within the routing protocol updates. So within a IS-IS TLV packet it will also carry label information for all the routers within a IS-IS area.

IS-IS SR supports four different ways of defining segments, namely, Node-SIDs, Prefix-SIDs, Proxy-Node SIDs and Adjacency Segments (SIDs). Both IPV4 and IPV6 prefixes are supported in IS-IS SR.

- Node-SID - Node segment IDs are typically tied to a switches loopback that represents the node.
- Prefix-SID - This is any prefix represented in segment routing. Point to points, loopbacks etc.
- Proxy-Node SIDS - Proxy-Node-SIDs are for backwards compatibility within a mix SR and LDP environment.
- Adjacency Segments - Adjacency Segments are the segments between different switches. So any point to point link between any PE/P or P/P nodes will be given a SR label.

Segment routing allows for traffic engineering or manipulation of traffic. RSVP-TE is generally the means of traffic engineering with in a MPLS DCI environment. Segment Routing is much simpler as Segment routing does not need RSVP for bandwidth reservation SR simply will stack adjacency, node, or prefix SIDs on top of each other in a ordered fashion so traffic will egress the way the user would prefer. Typical use cases would be application or VRF steering across different links due to bandwidth or latency sensitive applications.

EVPN with Segment routing use cases

As described in the L3LS-V section, EVPN is a standard technology (RFC 7432). EVPN implementation uses BGP for control plane learning and traditional flood learning is not required for MAC learning. Arista supports commonly used data plane for transport including VXLAN and MPLS. EVPN technology primary use case includes flexible L2/L3 VPN services for DCI and WAN connectivity over the IP and MPLS core networks, and VXLAN overlay deployment in datacenter. There are number of benefits that EVPN offers and some are listed below:

- Standard Technology (IETF Focus)
- Supports both L3 and L2 VPN services
- Control Plane Learning via BGP for L2 MAC Addresses
- A/A Multi Homing With Per Flow Load Balancing (both L2 & L3 traffic)
- MAC Routes Advertisement Similar To IP As In L3VPN Service
- IRB (Distributed Gateway/ Anycast Gateway)
- Various Encapsulation Support (VXLAN, MPLS)
- IP Prefixes Advertisement (L3VPN Services)
- Proxy ARP Resolves The Arp Requests Locally (Broadcast Storm Suppression)
- Minimize Broadcast And Unknown Unicast Traffic
- Large Scale Multi-tenancy (VXLAN Scale in DC fabric Environment)
- Bum Traffic Replication (DF)
- L3 Like Fast Convergence
- VPN Traffic Can Use MPLS For Traffic Engineering

EVPN services can leverage SR to steer application traffic on a controlled path if there is an SLA defined for those applications. SR simpler example could be a case where a low latency application should always flow across the lowest latency path available across the infrastructure. Some of the common use cases are listed below:

- For WAN & DCI connectivity, an application SLA can be defined which forces the application traffic to follow certain paths. In this case there could be multiple paths available to reach the destination. The SLA defined can be based on path latency, packet drop, jitter, BW availability, protected circuit, path coloring(never use or always use that link), and device capability.
- For DC Fabric, SR routing can be used to contain certain applications like storage to use defined uplinks. This will allow to create a virtual data plane domains within the datacenter fabric.
- Controller based traffic steering can be use in complying to application business logic. The intelligence in this case is derived from application and infrastructure analytics. The controller can create SR tunnels carrying application traffic in compliance to business logic. This use case eventually will mature into a self healing closed loop system.

There could be other use cases that can leverage the SR technology to transport EVPN traffic. EVPN and Segment Routing are expected to play a critical role in building intelligent SDN networks.

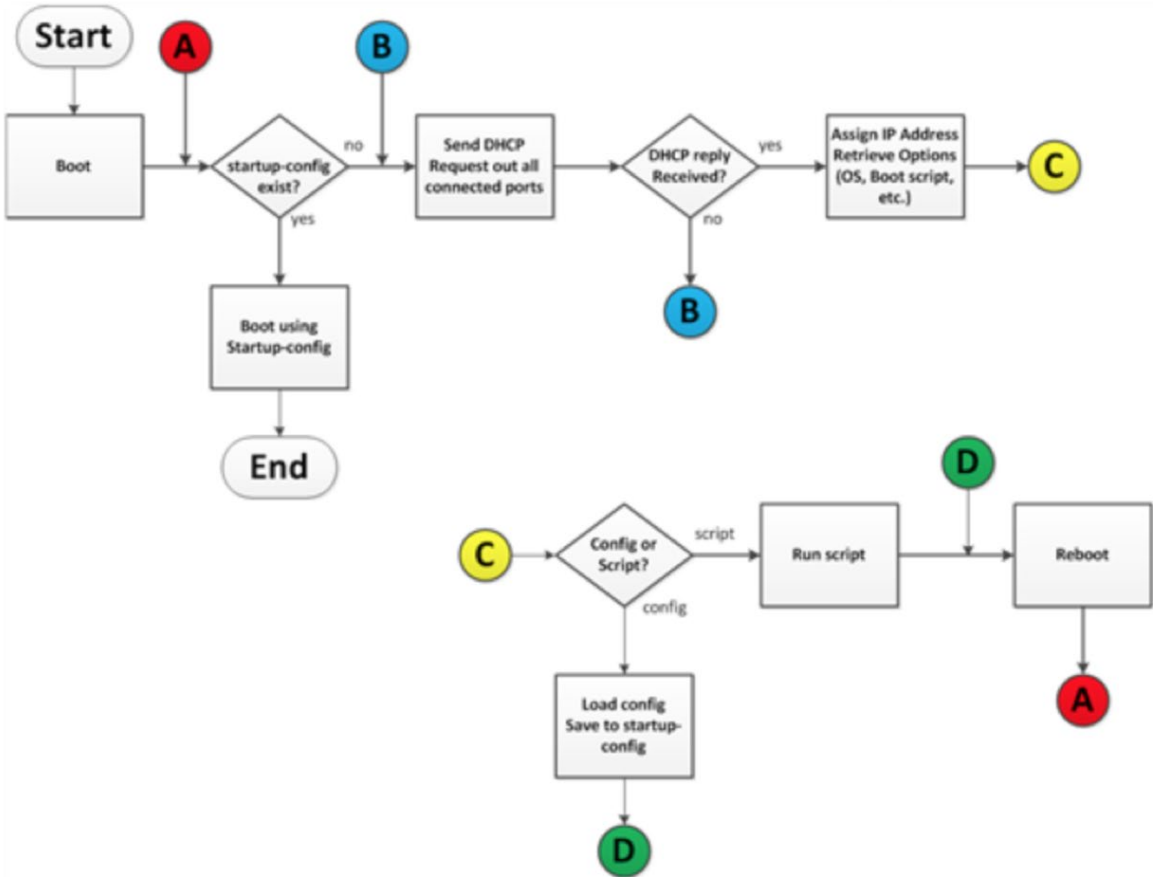
Appendix A – Network Automation

Arista EOS operating system is the most open and programmable network operating systems in the industry today. Much of this has come from an early understanding of what is required to build a world-class datacenter. A core tenant of Arista's EOS is reducing operational costs or OPEX. One of the key components of the Arista automation story focuses on bare metal provisioning (ZTP). Arista also includes support for OpenConfig and automation tools such as Chef, Puppet and Ansible.

A note about automation initiatives based on industry experience: investing the time and effort up front to properly build and test an automation solution will truly set the customer on a path for success. Layering in automation to an already provisioned network can be a challenge. There are many benefits to automating new deployments. By making automation a requirement from the start, implementation of the solution becomes ingrained in the network operators, managers and engineers. This type of ground-up deployment has the most chance to break the boundaries of what has been traditionally possible with legacy network equipment. Arista also has an EOS+ service offering to assist with this automation.

Zero Touch Provisioning (ZTP)

Key to realizing any automation strategy is a focus on bare metal provisioning. Arista's Zero Touch Provisioning (ZTP) leverages a standards-based (DHCP, TFTP, and HTTP) approach to provide the flexibility normally associated with server provisioning in the network layer. Arista also includes support for automation tools such as Chef, Puppet and Ansible further ensuring an existing compute automation strategy can be extended into the network layer. ZTP combined with CloudVision support, network administrators can now rapidly provision an entire IP based datacenter network.



ZTP is enabled by default on all new Arista products. When installed and powered on, the switch will send DHCP requests out all connected interfaces. When received by the DHCP server, DHCP options sent by the Arista switch will identify the device and send other information such as (System MAC, Hardware Version, Serial number etc.) to the DHCP server. This information can then be used to automate the system configuration based on unique system attributes. After receiving DHCP option 67, the switch requests the defined boot script with instructions on what to do..

With the onset of CloudVision, we have further simplified switch deployment for our customers. Once switches are powered-on they will seek out the CVP server using DHCP options provided by the onboard DHCP server or by aid of a third party server. Once in CVP the new device stays in a ready to deploy state, where EOS upgrades, configurations and extensions can be applied. Furthermore, this process can be automated by utilizing scripts and combined with Configlet Builders, or the Cloud Builders App can reduce the time to deploy new devices.

While a simple feature, ZTP lowers the engineering cost to deploy and replace network hardware by removing the need for on-site technicians at bring-up time. By automating this process an engineer can boot up an entire new site in minutes instead of hours; along with the surety that configurations and methodology would be consistent. Once devices are live, the further possibilities created by participation in a wider orchestration system are limitless.

Custom Scripting (Beyond ZTP)

ZTP in its basic form is very useful for deploying and maintaining infrastructure but some want to take it further. The basic operation of ZTP can be further enhanced to provide advanced automation for example using LLDP to determine the location and role of the switch in the network. Once the role and location are determined from pre-defined criteria the device can auto insert into the network. This scripting can then download extensions, as well as perform a variety of other useful tasks. This capability is exactly what CloudVision does out of the box.

Arista is committed to helping the customer take full advantage of all the automation capabilities of our platforms. Our Systems Engineering and Advanced Services teams working in conjunction with your automation teams can design an automation solution that meets your requirements. We also offer a software engineering service called EOS Consulting Services that can help to install a custom solution to fit the requirements in a turnkey manner.

At Arista we take great pride in contributing to the open nature of our EOS network operating system. We have a powerful open development community (http://eos.arista.com) where users can find tech tips, automation scripts and share experiences on Arista devices. The community approach allows us to share information and create an ecosystem of user driven open tools. Many of these tools and tech tips, have a common theme of automation, proactive notification and true differentiation.

Arista API (eAPI)

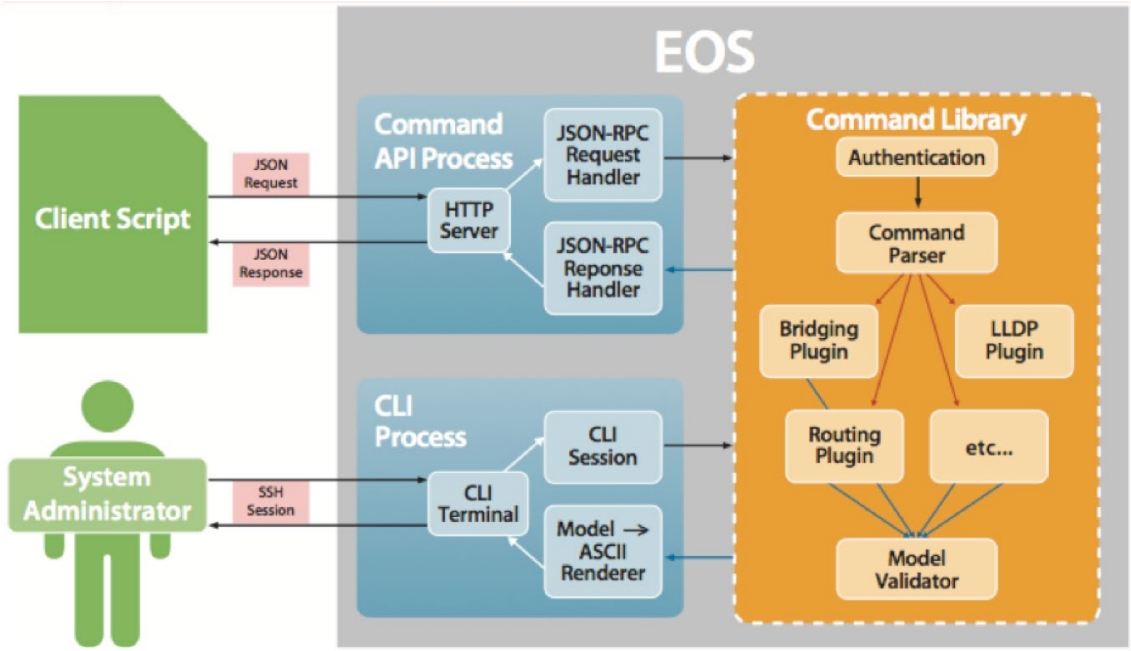
Arista has developed an open API for issuing commands to the switch or group of switches based on JSON RPC over HTTP/HTTPS and optionally with SSL certificate support. EOS API (eAPI) is a comprehensive tool for network managers. It allows applications or controllers to gain visibility into system state and customers to integrate applications and other network components with Arista's switching platforms. eAPI provides a mechanism to run legacy applications seamlessly against newer versions of EOS.

Arista EOS offers multiple programmable interfaces for applications. Arista has the capability of using various programmable interfaces such as Linux, Go, Python, Ruby, as well as Arista EOS extensible API (eAPI) using JSON.

eAPI is programming language and data model agnostic – allowing the customer to choose the most familiar programming language for third party integration.

Some of the advantages of eAPI's JSON structure:

- Less verbose/more efficient
- More careful in use of data types
- Deeply interoperable with any programming language
- Ability to represent most general data structures: records, lists and trees
- Significantly simpler syntax – more efficient parsing
- Can represent any non-recurrent data structure as is – more flexible
- Continually updated in line with EOS's CLI ensuring consistent support across all products



Arista JSON eAPI model

Only a scalable API such as eAPI can support true network agility. eAPI is complete and portable. It extends across various layers of the network from Leaf to Spine. eAPI's command library provides a shared mechanism to support CLI, SNMP and programming APIs under a common scheme, which maps consistently across all EOS based products. eAPI provides a seamless integration point for controller or applications to read and modify the state of the switch within the network – thus fully bringing the software-defined networking concept to life.

(To view, enable the API and visit "http://<your-switch's-ip-address>/ explorer.html" in a web browser).

OpenConfig

OpenConfig provides open data models for network management. The goal of OpenConfig is to normalise configuration and monitoring data across platforms with common data and device interactions. The two primary components of the OpenConfig are YANG based data models for configuration and operational state; and an open transport over which to stream this state or instantiate configuration.

Arista EOS supports writing, reading and streaming various OpenConfig configuration and state models over gNMI (gRPC Network Management Interface), RESTCONF, and NETCONF transports.

Automation Tools (Chef, Puppet, CFengine, Ansible and Saltstack)

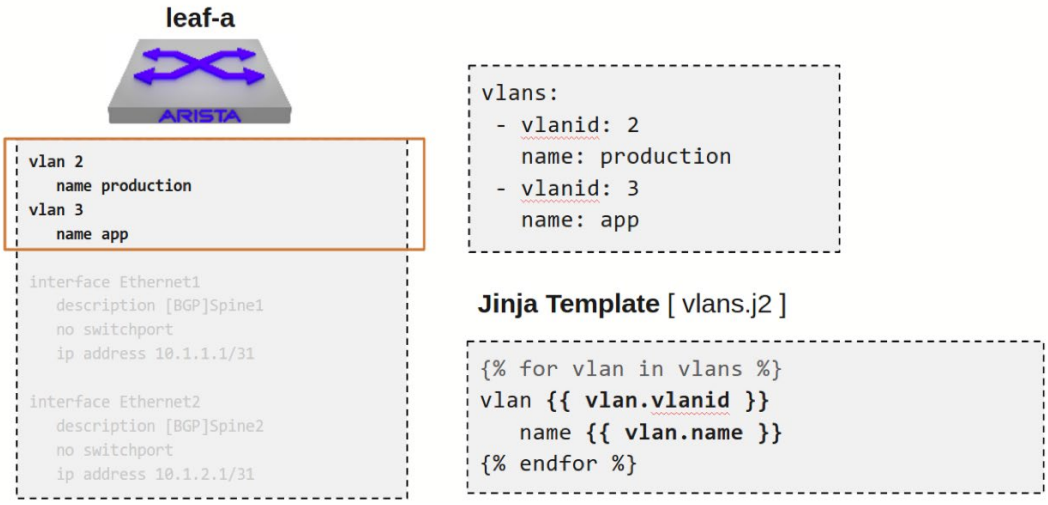
Arista EOS supports native integration with a variety of network orchestration and configuration management tools. The same tools used by system administrators to automate provisioning across large numbers of servers can also be used to orchestrate Arista's switch portfolio. Arista's open and extensible operating system provides unparalleled support for off-the-shelf Linux tools. Reference configuration as well as code examples are published on our EOS Central website as well as Github for general deployment and use.

Ansible integration

Ansible is a great configuration management tool which supports thousands of different modules including Arista EOS network devices. Ansible does not require a agent which means a switch can be zero touch provisioned then using Ansible can be used to either send ad-hoc commands or configure the switch throughout the process of added new configuration. Ansible can be used within either the eAPI which is preferred or through SSH which leverages with Paramiko library. Currently in EOS there is support for Ansible, AWX and Ansible tower.

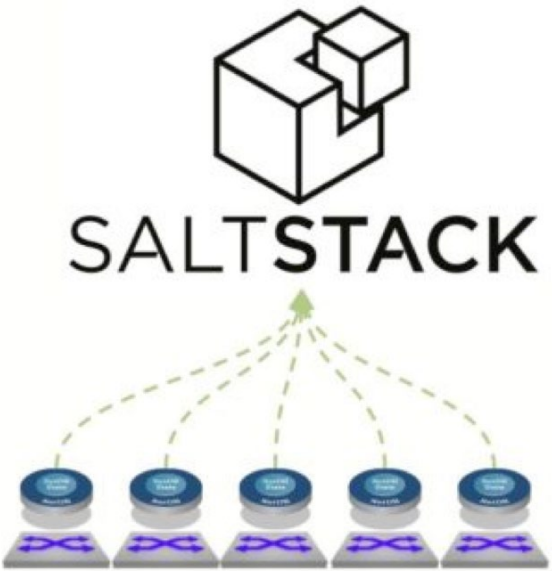
Since Ansible 2.1 Arista EOS supports Jinja2 templating language for any configuration changes. A Jinja2 templating language is simply a way to render texts using python objects either strings, lists, dictionaries or tuples.

A common area where Jinja2 templating language can be used within Ansible is when it comes to templating snippets of config. For example, adding a VLAN or interface config to a jinja2 template is very simple.



Within the diagram VLAN 2 and VLAN 3 need to be added to the switch. Within Jinja they are added with simple Jinja for-loop that will loop through VLANs then add a the first string under vlan.vlan.vlanid then vlan.name. This will result in the CLI configuration of VLAN 2 and VLAN 3 as shown above.

Saltstack integration



Salt is an event driven infrastructure management tool that can configure, manage and react to events as they come through the network on its high speed ZeroMQ messaging bus that runs between the master and the minions or in our case EOS switches. Salt does require an agent for Arista EOS which can be installed on the switch due to Arista EOS uses a Fedora core linux or Salt can be used agentless in a concept called a proxy minion.

Salt can be used to configure devices from multiple template languages such as Jinja or YAML. Salt can target devices and run ad-hoc commands to multiple switches. Although, the most powerful feature of Salt is the events. All events are logged on the high speed, ZeroMQ messaging bus. Therefore, Salt can react to any event such a network switch being added, a new host coming up, a new BGP peering session etc. Salt also has many different integrations or commonly known as formulas to third parties like Slack integration.

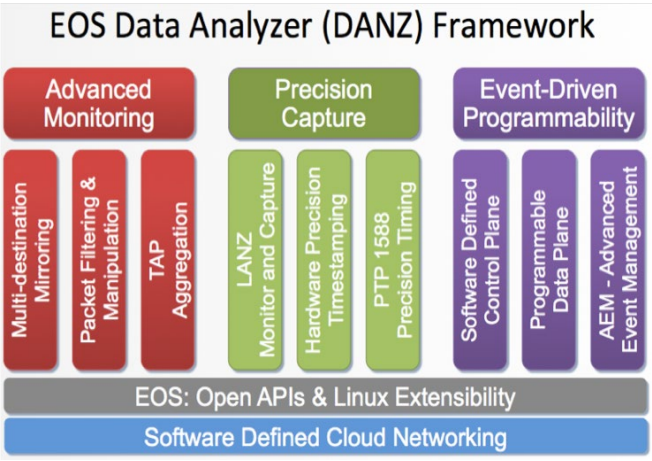
APPENDIX B - Network Telemetry and Visibility

Support for Statistical Flow Monitoring (sFlow)

Arista EOS supports sFlow collection services. Managers can configure the switch to report statistics and forward sampled traffic flows to an sFlow monitor/collector. Arista EOS supports up to two sFlow collectors. This tool is preferred when managing multi-terabit switched datacenter networks.

Arista Data ANalyZer (DANZ)

Arista Data ANalyZer (DANZ) is an integrated feature set of Arista's Extensible Operating System and is currently available on Arista Networks 7150, 7280, and 7500 series switches. DANZ provides a solution to monitoring and visibility challenges at 10/40Gbps and higher with unmatched price-performance and precision.



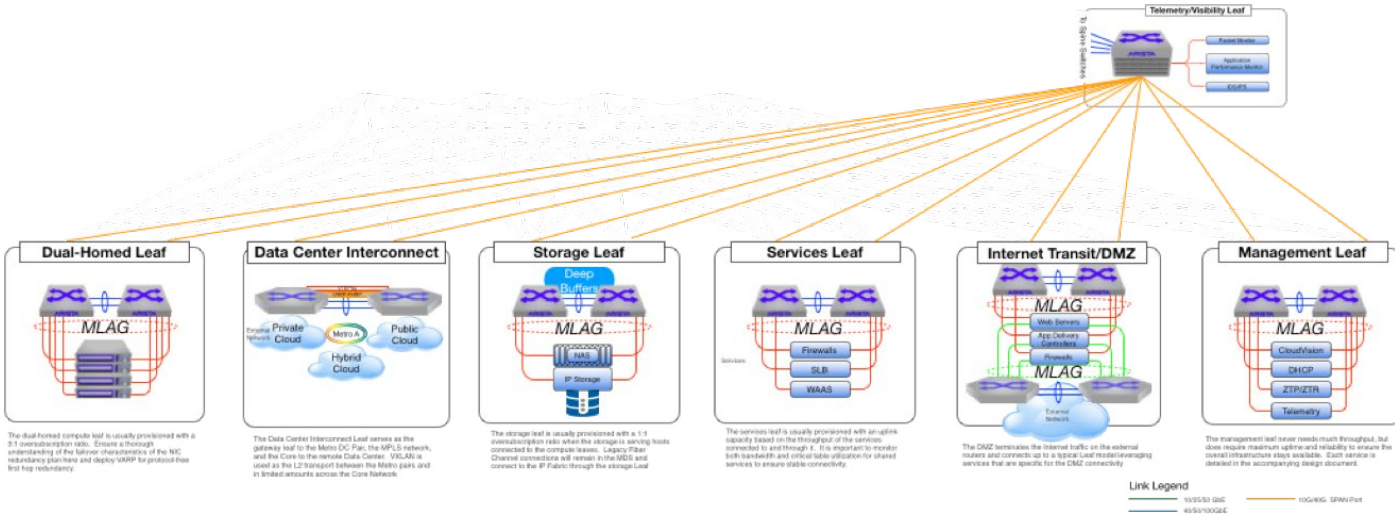
Arista has focused on delivering a new class of high precision instrumentation; bringing management tools up to speed and aligned with the visibility needs of high performance environments for 10GbE speeds and beyond.

EOS Data Analyzer (DANZ) tool-set provides a broad range of features for fine and coarse-grained troubleshooting in three main categories:

- Advanced Mirroring
- Precision Capture
- Event-Driven Programmability

Advanced Mirroring and TAP Aggregation

The advanced mirroring functionality of the Arista platform is a suite of enhancements to common port mirroring functionality that enables users to leverage capabilities normally only found in dedicated aggregation devices directly in the datacenter switch – significantly reducing the complexity, cost and footprint of precision monitoring.



Latency Analyzer (LANZ)

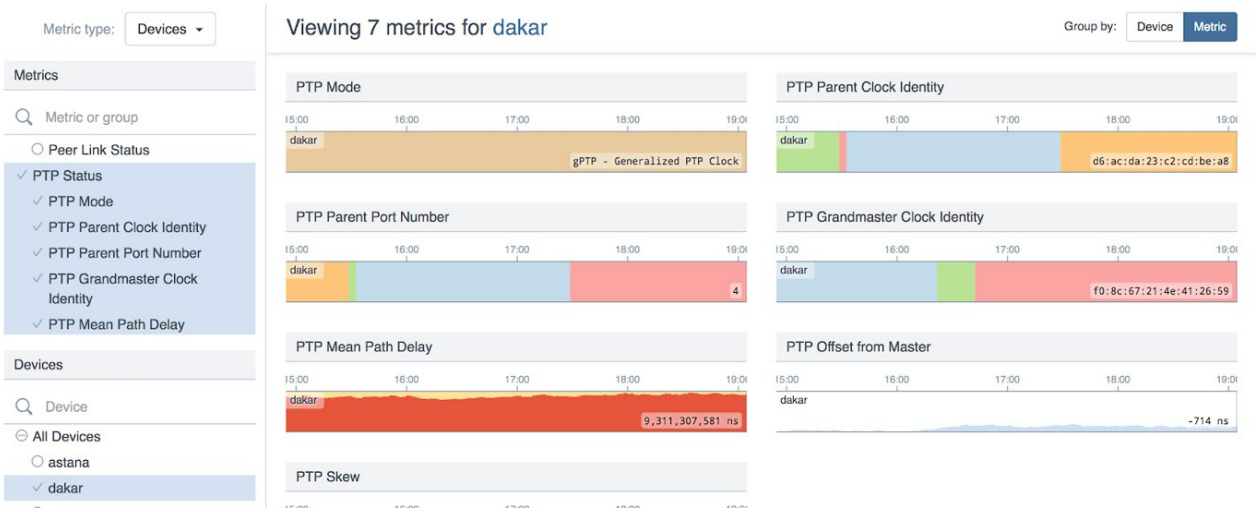
LANZ is a unique tool for microburst detection and monitoring interface contention. It allows an administrator to interpret events typically occurring at nanosecond and microsecond rates in the network.

LANZ allows the user to set thresholds for triggered alerts in hardware. LANZ works at the byte level, capturing, rather than sampling data. It can be configured to alert on the smallest levels of interface contention, being able to instrument to the granularity of just one packet queued behind another.

LANZ visibility is now available through CloudVision Telemetry as well:

Hardware Timing (precision oscillator and PTP services)

The 7150 series leverages the on board high-precision oscillator coupled with integrated hardware time stamping to apply accurate timestamps to every packet at wire speed without incurring additional latency. The integration of hardware time stamping provides new possibilities for accurate measurement of both one-way and round-trip traffic latency with basic tools or software (such as a network sniffer). Hardware based Precision Time Protocol (IEEE-1588) for sub-microsecond accuracy is also supported across a wide selection of Arista hardware platforms. The precision oscillator and hardware time stamping capabilities of the enables accuracy measures to tens of nanoseconds (±10ns).



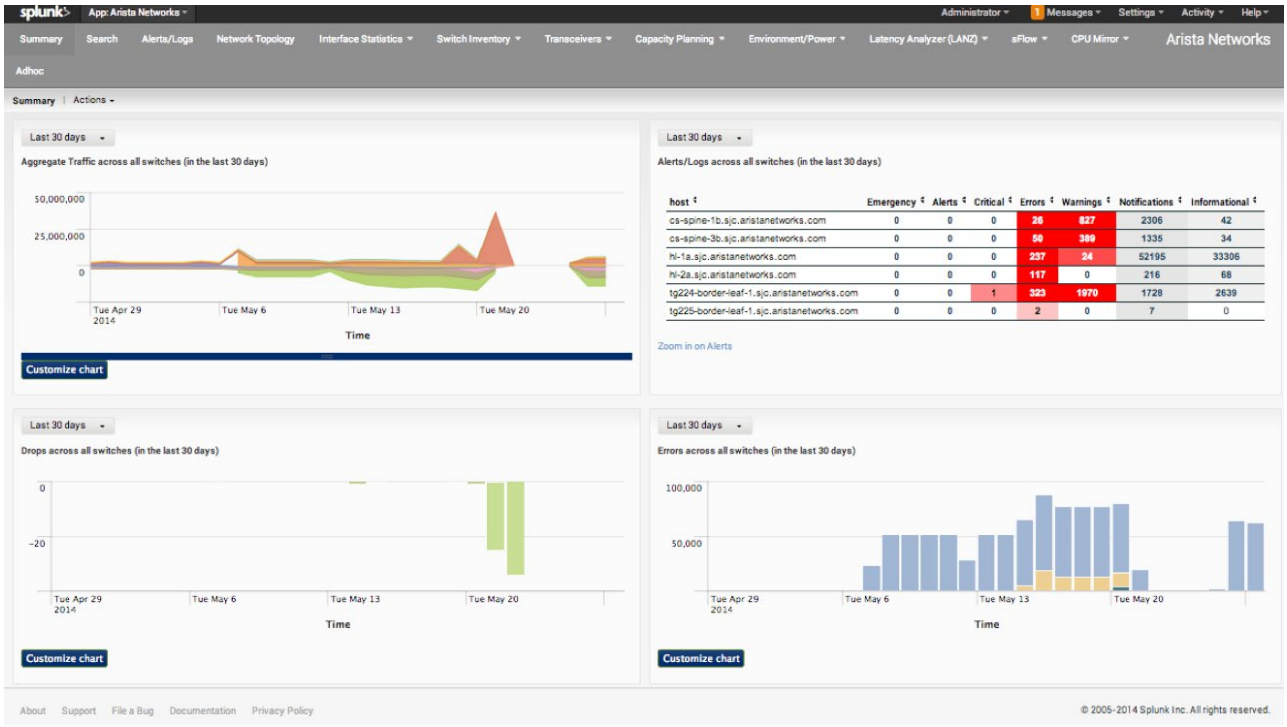
Advanced Event Management (AEM)

AEM is a suite of tools that allows EOS platforms to easily schedule events and react to network changes in customized ways. AEM can execute customized scripts or packages to control or reconfigure the device, extract data, automate operational processes, and even customize the behavior of the switch on the fly. It enables the network operator to build a rich set of action/reaction events to proactively manipulate what happens on the switch without human interaction, lowering operational overhead and improving reaction time to issues.

With AEM, any EOS platform can log key state indicators in a SQL database. This is done in real-time, providing operators a comprehensive view of what's happening in the network at any time index. AEM is implemented on All EOS platforms allowing administrators to analyze and correlate network events from multiple points in the network.

Splunk Integration

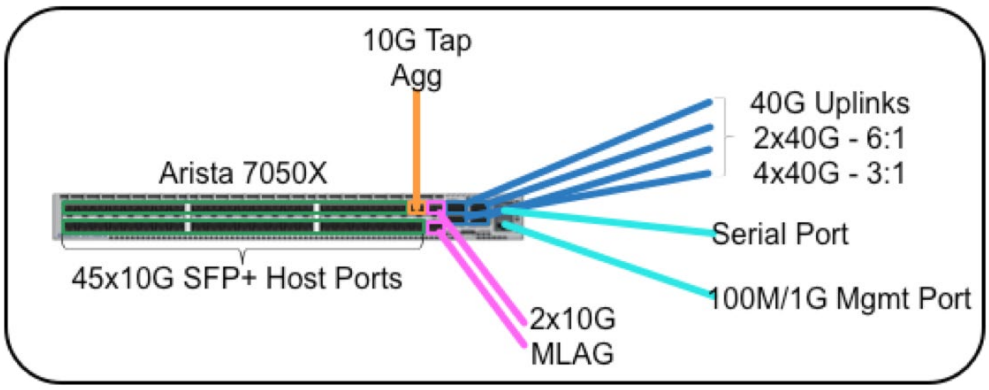
Arista EOS has the unique functionality to provide direct event reporting to a Splunk data collector. This native functionality allows operators to add networking statistics to single-pane monitoring systems that previously only included application data. This enables the NOC to correlate network errors in the network with events in real time. The Splunk integration also delivers the ability to track inventory, environment, and interface statistics that can be used to further analyze and troubleshoot issues.



APPENDIX C – LEAF SWITCH CONFIGURATION

Compute, Services and Management Leaf switches

For a 64 port TOR configuration, Arista recommends the Arista 7050X for 10G to the hosts and 40G up to the spine and the 7160, 7050X3 or 7060X2 switches for 10/25G to hosts and 100G up to the spine.. The 7050X switch supports 48 ports of 10Gb via SFP+ or 10G-BaseT connectors as well as 6x40Gb uplinks via the QSFP connector. The 7160 and 7060X3 support 48 ports of 10/25Gb and 6x100Gb uplinks via QSFP and the 7050X3 with 12x100Gb uplinks. As deployed the port allocation would be:



45x 10GbE ports – Host facing

1x 100MbE/1GbE port – IPMI connection to OOB management network

1x 10GbE port – Tap Aggregation connection to TapAgg network

2x 10GbE ports – MLAG Peer Links

2 or 4x 40GbE ports – Uplinks to Spine Switches

The 7050X-64 supports wire-speed forwarding, local L3 routing, hardware VXLAN virtual tunnel end-point (VTEP) support, and has buffers designed to support 3:1 downlink to uplink ratios leveraging PAUSE framing for flow control during periods of congestion.

For more dense rack solutions, Arista recommends the Arista 7050SX2-128 or 7050TX2-128. This switch supports 96 ports of 10Gb via SFP+ or 10G-BaseT connector as well as 8x40Gb uplinks via the QSFP connector. As deployed the port allocation would be:

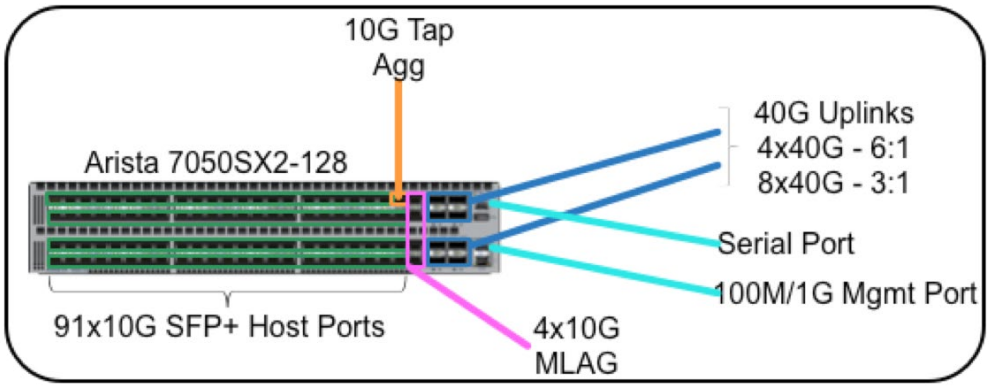
91x 10GbE ports – server facing

1x 100MbE/1GbE port – IPMI connection to OOB management network

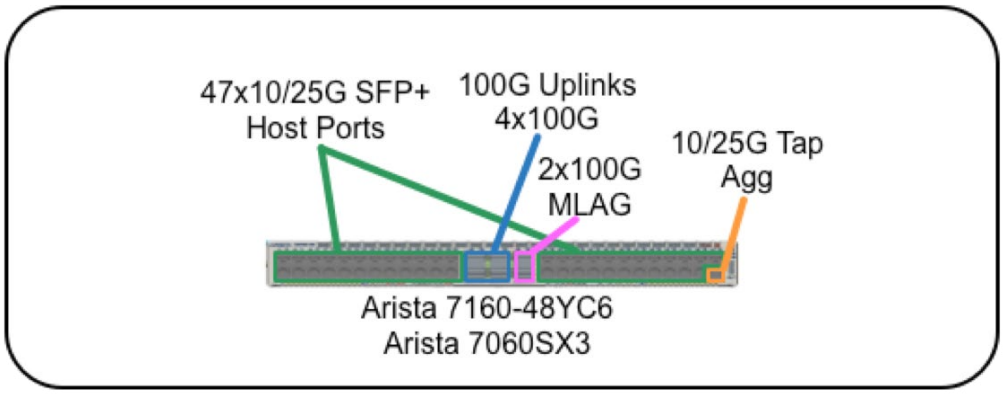
1x 10GbE port – Tap Aggregation connection to TapAgg network

4x 10GbE ports – MLAG Peer Links

8x 40GbE ports – Uplinks to Spine Switches



The 7050SX2-128 supports wire-speed forwarding, local L3 routing, hardware VXLAN virtual tunnel end-point (VTEP) support, and has buffers designed to support 3:1 downlink to uplink ratios leveraging PAUSE framing for flow control during periods of congestion.



47x 10/25GbE ports – Host facing

1x 100MbE/1GbE port – IPMI connection to OOB management network

1x 10/25GbE port – Tap Aggregation connection to TapAgg network

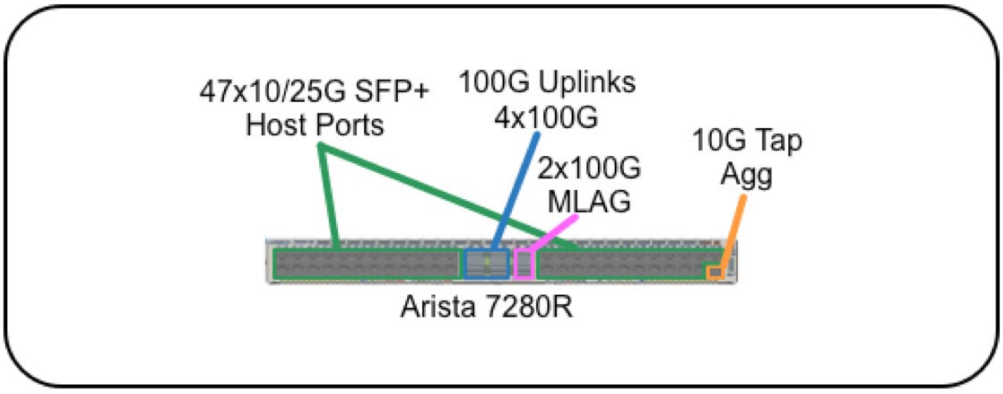
2x 100GbE ports – MLAG Peer Links

2 or 4x 100GbE ports – Uplinks to Spine Switches

The 7050X-64 supports wire-speed forwarding, local L3 routing, hardware VXLAN virtual tunnel end-point (VTEP) support, and has buffers designed to support 3:1 downlink to uplink ratios leveraging PAUSE framing for flow control during periods of congestion.

Internet Transit/DMZ and Storage Leaf Switches

Arista recommends the Arista 7280R series for Internet Transit/DMZ and storage leaf deployments. This switch family is built upon the same architecture as the 7500 series except it is in a 1RU form factor. The 7280R series provides deep buffers, similar to the 7500R Spine switches to ensure reliable delivery of data. As deployed a typical port allocation would be:



47x 10/25GbE ports – provider/server facing

1x 1GbE/1GbE – IPMI connection to OOB management network (on rear of switch)

1x 10/25GbE port – Tap Aggregation connection to TapAgg network

2x 100GbE ports – MLAG Peer Links

4x 100GbE ports – Uplinks to Spine Switches

The 7280R series supports wire-speed forwarding, Internet scale routing tables, wire-speed L3 routing, hardware VXLAN virtual tunnel end-point (VTEP) support, and has buffers designed to support 1:1 downlink to uplink ratios. The 7280R series is optimally suited for storage networks. The stability of EOS is a key factor in successfully deploying an IP storage network. Aside from stability, the 7280R provides features key for IP storage deployments such as deep buffers, VoQ based priority flow control for lossless connectivity without the head of line blocking caused by collateral pause in a non-VoQ based switch, and ECN marking for seamless integration with host based congestion management. The 7280R series also has the flexibility of offering multiple connectivity options including 10/25/40/50/100 GbE to provide flexibility with host connectivity while allowing engineers to select the optimum form factor for deployment.

Datacenter Interconnect

Arista recommends the Arista 7280R for the datacenter interconnect deployment. This switch supports 48 ports of 10/25GbE via SFP+ connector as well as 6x100Gb uplinks via the QSFP connector. In addition to the deep buffer architecture, the 7280 series provides line rate VXLAN for DCI transport. In a DCI type application where speed mis-matches are common, leveraging deep buffers can make a tremendous impact on DCI performance. As deployed the port allocation would be:

46 x 10/25GbE ports – host/router facing

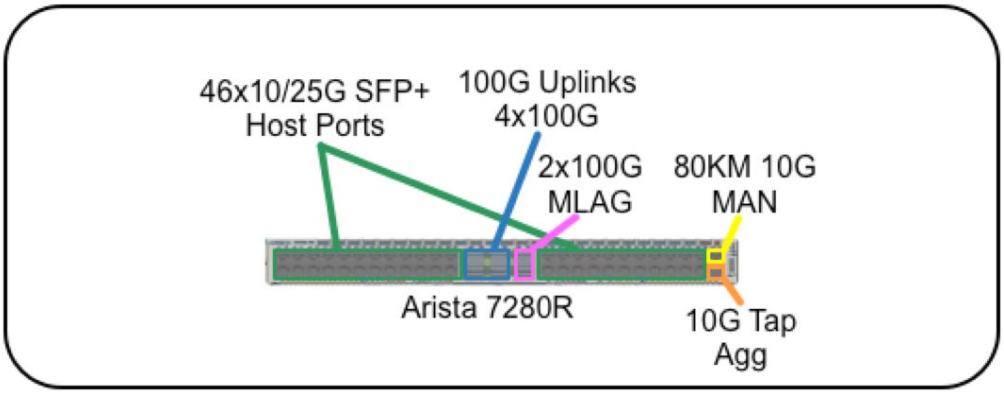
1 x 1 GbE/1GbE – IPMI connection to OOB management network

1 x10/25GbE – Long Haul MAN Connection

1 x 10/25GbE port – Tap Aggregation connection to TapAgg network

2 x 100GbE ports – MLAG Peer Links

4 x 40/100GbE ports – Uplinks to Spine Switches



Campus Spline

Arista recommends the Arista 7050X3 and 7300X3 switches for the Campus Spline deployments. The 7050X3 switch supports 32x100Gb ports via the QSFP+ connector. The 7050X3 also supports a series of security features in addition to VXLAN encapsulation for a complete solution in the Campus Spine. As deployed the port allocation would be:

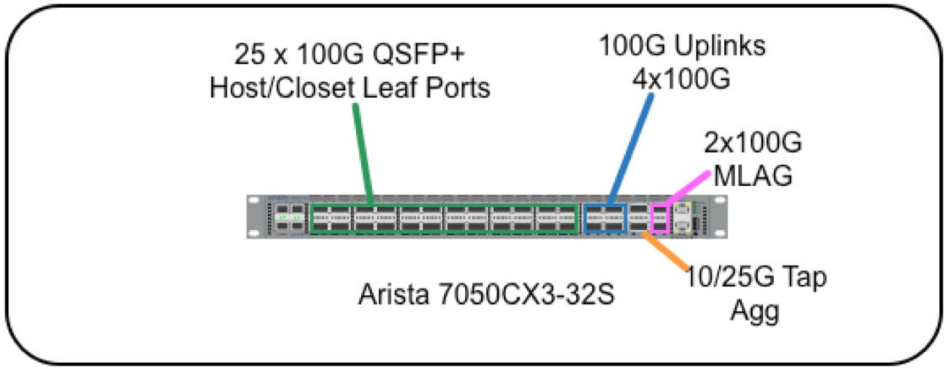
25 x 100GbE ports – Closet Leaf or Campus host connections (may be broken in 4x10/25GbE as well)

1 x 1 GbE/1GbE – IPMI connection to OOB management network

1 x 100GbE port – Tap Aggregation connection to TapAgg network

2 x 100GbE ports – MLAG Peer Links (not required with an ECMP design)

4 x 100GbE ports – Uplinks to Spine Switches



The Campus Spline could also be a chassis based systems. The Arista 7300X3 is the recommended chassis for the Campus Spline deployment. The 7300X3 utilized 32x100Gb ports via the QSFP+ connectors. It also has the same rich set of features and security capabilities that the 7050X3 has - as they are built using the same packet processor. As deployed the 7300X3 would be:

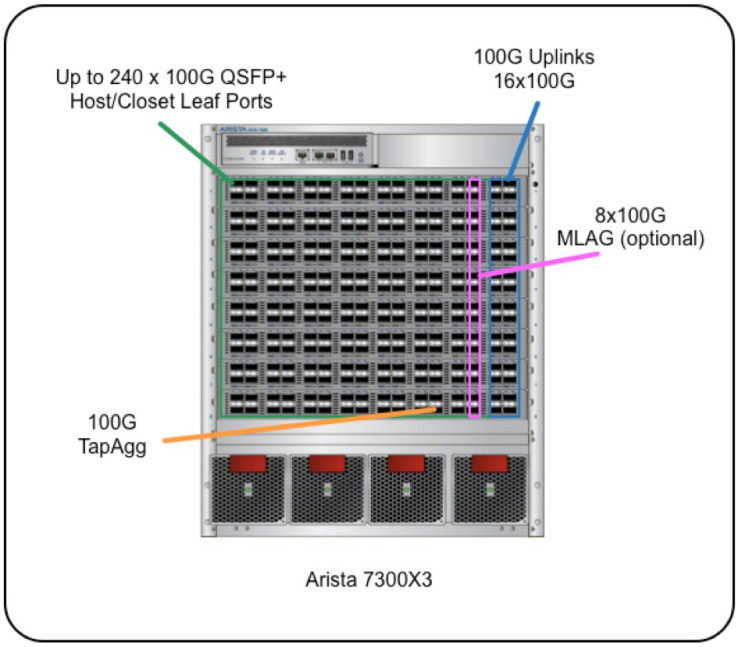
Up to 240 x 100GbE ports – Closet Leaf or Campus host connections (may be broken in 4x10/25GbE)

1 x 1 GbE/1GbE – IPMI connection to OOB management network

1 x 100GbE port – Tap Aggregation connection to TapAgg network

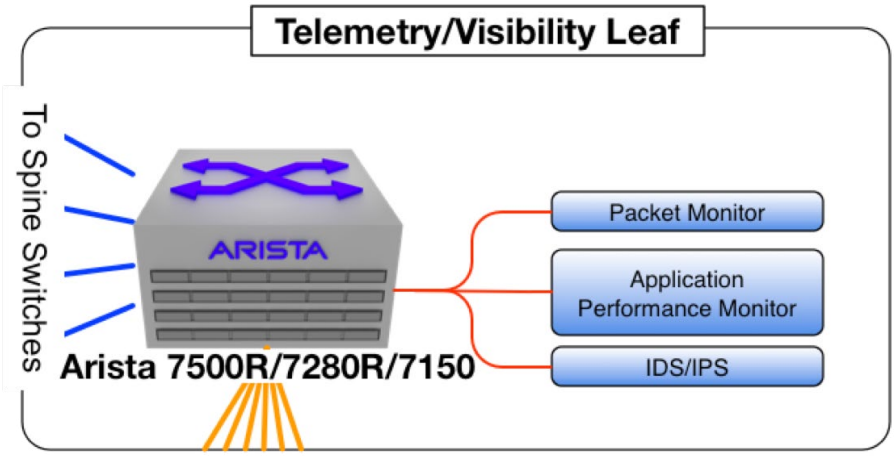
8 x 100GbE ports – MLAG Peer Links (not required with an ECMP design)

16 x 100GbE ports – Uplinks to Spine Switches



Tap Aggregation/Visibility

Arista recommends the Arista 7150, 7280, and 7500 series for the tap aggregation deployment. The 7150S switch could be combined with a 7500 series to provide extreme scale tap aggregation across every device in the datacenter. The 7150S-64 switch supports 48 ports of 10Gb via SFP+ connector as well as 4x40Gb uplinks via the QSFP connector. The 7280R supports 48 ports of 10/25G via SFP+ connector as well as 6x100Gb uplinks via the QSFP connector. The 7500R supports up to 576 x 100Gb via the QSFP connector.



LOM/IPMI 1GbE Switch

Arista recommends the Arista 7010T for the out-of-band management switch for the network to provide single-image consistency across the entire network.

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