Arista 7500 Scale-Out Cloud Network Designs

Arista Networks award-winning Arista 7500 Series was introduced in April 2010 as a revolutionary switching platform, which maximized datacenter performance, efficiency and overall network reliability. It raised the bar for switching performance being five times faster, one-tenth the power draw and one-half the footprint compared to other modular datacenter switches.

Just three years later, the introduction of the Arista 7500E Series modules and fabric delivers a three-fold increase in density and performance with no sacrifices on functionality, table sizes or buffering with industry-leading 1,152 x 10GbE, 288 x 40GbE or 96 x 100GbE in the same quarter-rack 11RU chassis.

This whitepaper details scale-out cloud network designs enabled with the new Arista 7500E second-generation linecard and fabric modules.

Key Points of Arista Designs

All Arista reference designs revolve around these central design goals:

1. *No proprietary protocols or vendor lock-ins. Arista believes in open standards.* Our proven reference designs show that proprietary protocols and vendor lock-ins aren't required to build very large scale-out networks
2. *Fewer Tiers is better than More Tiers.* Designs with fewer tiers (e.g. a 2-tier design rather than 3-tier) decrease cost, complexity, cabling and power/heat. A legacy design that may have required 3 or more tiers to achieve the required port count just a few years ago can be achieved in a 1 or 2-tier design.
3. *No protocol religion.* Arista supports scale-out designs built at layer 2 or layer 3.
4. *Modern infrastructure should be run active/active.* Multi chassis Link Aggregation (MLAG) at layer 2 and Equal Cost Multi-Pathing (ECMP) at layer 3 enables infrastructure to be built as active/active with no ports blocked so that networks can use all the links available between any two devices.
5. *Designs should be agile and allow for flexibility in port speeds.* The inflection point when the majority of servers/compute nodes connect at 1000Mb to 10G is between 2013-2015. This in turn drives the requirement for network uplinks to migrate from 10G to 40G and to 100G. Arista switches and reference designs enable that flexibility.

Figure 1: 7504E and 7508E with up to 1,152 10G ports, 288 40G ports or 96 100G ports

6. *Scale-out designs enable infrastructure to start small and evolve over time.* A two-way ECMP design can grow from 2-way to 4-way, 8-way, 16-way and as far as a 32-way design. An ECMP design can grow over time without significant up-front capital investment.
7. *Large Buffers at the spine (or aggregation) matter.* Modern Operating Systems (OS), Network Interface Cards (NICs) and scale-out storage arrays are increasingly making use of techniques such as TCP Segmentation Offload (TSO) and Generic Segmentation Offload (GSO), collectively termed Large Segment Offload (LSO). These techniques are fundamental to reducing the CPU cycles required when servers send large amounts of data. Large Receive Offload (LRO) is often deployed on the receive side to perform a similar function. A side effect of these techniques is that a server/OS/storage that wishes to transmit a chunk of data will offload it to the NIC, which slices the data into segments and puts them on the wire as back-to-back frames at line-rate. If more than one of these is destined to the same output port then microburst congestion occurs and deep buffers are required to absorb the bursts. The spine (aggregation) layer is at a place in the network where thousands to potentially millions of flows meet and there is consequently a much higher probability of microburst congestion occurring. Lack of deep buffers to absorb the bursts results in packet drops, which in turn results in lower good-put (useful throughput).
8. *Consistent features and OS.* All Arista switches use the same Arista EOS. There is no difference in platform, software trains or OS. It's the same binary image across all switches.
9. *Interoperability.* Arista switches and designs can interoperate with other networking vendors.

Design Choices

Oversubscription

Oversubscription is the ratio of contention should all devices send traffic at the same time. It can be measured in a north/south direction (traffic entering/leaving a datacenter) as well as east/west (traffic between devices in the datacenter). Many legacy datacenter designs have very large oversubscription ratios, upwards of 20:1 for both north/south and east/west, because of the large number of tiers and limited density/ports in the switches but also because of historically much lower traffic levels per server.

Significant increases in the use of multi-core CPUs, server virtualization, flash storage, Big Data and cloud computing have driven the requirement for modern networks to have lower oversubscription. Current modern network designs have oversubscription ratios of 3:1 or less. In a two-tier design this oversubscription is measured as the ratio of downlink ports (to servers/storage) to uplink ports (to spine switches). For a 64-port leaf switch this equates to 48 ports down to 16 ports up. In contrast a 1:1 design with a 64-port leaf switch would have 32 ports down to 32 up.

A good rule-of-thumb in a modern datacenter is to start with an oversubscription ratio of 3:1. Features like Arista Latency Analyzer (LANZ) can identify hotspots of congestion before it results in service degradation (seen as packet drops) allowing for some flexibility in modifying the design ratios if traffic is exceeding available capacity.

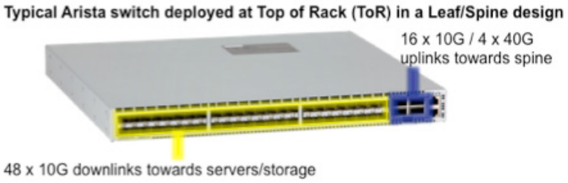

Typical Arista switch deployed at Top of Rack (ToR) in a Leaf/Spine design
16 x 10G / 4 x 40G uplinks towards spine
48 x 10G downlinks towards servers/storage

Figure 2: Leaf switch deployed with 3:1 oversubscription (48x10G down to 4x40G up)

10G Uplinks or 40G Uplinks

For a Leaf/Spine network, the uplinks from Leaf to Spine are typically either 10G or 40G and can migrate over time from a starting point of 10G (N x 10G) to become 40G (or N x 40G). All Arista 10G ToR switches offer this flexibility as 40G ports with QSFP+ can operate as either 1x40G or 4x10G, software configurable. Additionally the AgilePorts feature on some Arista switches allows a group of four 10G SFP+ ports to operate as a 40G port.

Layer 2 or Layer 3

Two-tier networks can be built at either layer 2 (VLAN everywhere) or layer 3 (subnets). Each has their advantages and disadvantages.

Layer 2 designs allow the most flexibility allowing VLANs to span everywhere and MAC addresses to migrate anywhere. The downside is that there is a single common fault domain (potentially quite large), and as scale is limited by the MAC address table size of the smallest switch in the network, troubleshooting can be challenging, L3 scale and convergence time will be determined by the size of the Host Route table on the L3 gateway and the largest non-blocking fan-out network is a spine layer two switches wide utilizing Multi-chassis Link Aggregation (MLAG).

Layer 3 designs provide the fastest convergence times and the largest scale with fan-out with Equal Cost Multi Pathing (ECMP) supporting up to 32 or more active/active spine switches. These designs localize the L2/L3 gateway to the first hop switch allowing for the most flexibility in allowing different classes of switches to be utilized to their maximum capability without any dumbing down (lowest-common-denominator) between switches.

Layer 3 designs do restrict VLANs and MAC address mobility to a single switch or pair of switches and so limit the scope of VM mobility to the reach of a single switch or pair of switches, which is typically to within a rack or several racks at most.

Layer 3 with VXLAN

A VXLAN design complements the Layer 3 designs by enabling a layer 2 overlay across layer 3 via the non-proprietary multi-vendor VXLAN standard. It couples the best of layer 3 designs (scale-out, massive network scale, fast convergence and minimized fault domains) with the flexibility of layer 2 (VLAN and MAC address mobility), alleviating the downsides of both layer 2 and layer 3 designs.

VXLAN capabilities can be enabled in software through hypervisor-resident virtual switches as part of a virtual server infrastructure. This approach extends layer 2 over layer 3 but doesn't address how traffic gets to the correct physical server in the most optimal manner. A software-based approach to deploying VXLAN or other overlays in the network also costs CPU cycles on the server, as a result of the offload capabilities on the NIC being disabled.

Hardware VXLAN Gateway capability on a switch enables the most flexibility, greater scale and traffic optimization. The physical network remains at layer 3 for maximum scale-out, best table/capability utilization and fastest convergence times. Servers continue to provide NIC CPU offload capability and the VXLAN Hardware Gateway provides layer 2 and layer 3 forwarding, alongside the layer 2 overlay over layer 3 forwarding.

As the functionality enabled in Hardware VXLAN Gateways is rapidly evolving, the topologies possible with VXLAN Gateway support continue to expand providing maximum flexibility. We suggest working closely with Arista on these designs.

Table 1: Pros/Cons of Layer 2, Layer 3 and Layer 3 with VXLAN designs

Design Type	Pros	Cons
Layer 2	VLAN everywhere provides most flexibility MAC mobility enables seamless VM mobility	Single (large) fault domain Redundant/HA links blocked due to STP Challenging to extend beyond a pod or datacenter without extending failure domains L3 gateway convergence challenged by speed of control plane (ARPs/second) L3 scale determined by Host route scale @ L3 gateway Scale can be at most 2-way wide (MLAG active/active) Maximum Number of VLANs x Ports on a switch limited by Spanning Tree Logical Port Count Scale Challenging to Troubleshoot
Layer 3	Extends across pods or across datacenters Very large scale-out due to ECMP Very fast convergence/ re-convergence times	VLAN constrained to a single switch MAC mobility only within a single switch

Forwarding Table Sizes

Ethernet switching ASICs use a number of forwarding tables for making forwarding decisions: MAC tables (L2), Host Route tables (L3) and Longest Prefix Match (LPM) for L3 prefix lookups. The maximum size of a network that can be built at L2 or L3 is determined by the size of these tables.

Historically a server or host had just a single MAC address and a single IP address. With server virtualization this has become at least 1 MAC address and 1 IP address per virtual server and more than one address/VM if there are additional virtual NICs (vNICs) defined. Many IT organizations are deploying dual IPv4 / IPv6 stacks (or plan to in the future) and forwarding tables on switches must take into account both IPv4 and IPv6 table requirements.

If a network is built at Layer 2 every switch learns every MAC address in the network, and the switches at the spine provide the forwarding between layer 2 and layer 3 and have to provide the gateway host routes.

If a network is built at Layer 3 then the spine switches only need to use IP forwarding for a subnet (or two) per leaf switch and don't need to know about any host MAC addresses. Leaf switches need to know about the IP host routes and MAC addresses local to them but don't need to know about anything outside the local connections. The only routing prefixes leaf switches require is a single default route towards the spine switches.

Regardless of whether the network is built at layer 2 or layer 3 it's frequently the number of VMs that drives the networking table sizes. A currently modern x86 server is a dual socket with 6 or 8 CPU cores/socket. Typical enterprise workloads allow for 10 VMs/CPU core, such that for a typical server to have 60-80 VMs running is not unusual. It is foreseeable that this number will only get larger in the future.

For a design that is 10 VMs/CPU, quad-core CPUs with 2 sockets/server, 40 physical servers per rack and 20 racks of servers, this would drive the forwarding table requirements of the network as follows:

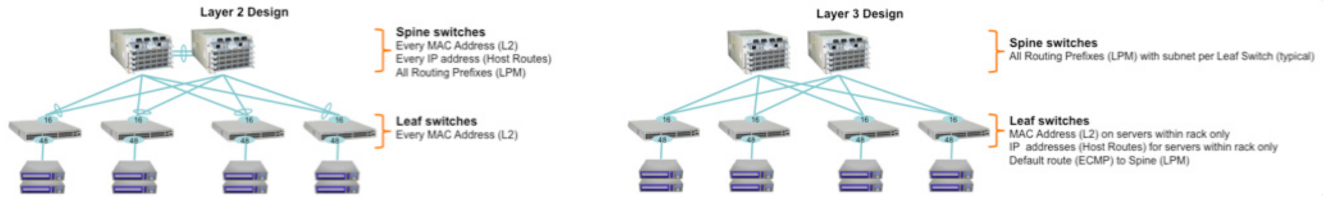

Layer 2 Design
Spine switches
Every MAC Address (L2)
Every IP address (Host Routes)
All Routing Prefixes (LPM)
Leaf switches
Every MAC Address (L2)
16
48
Layer 3 Design
Spine switches
All Routing Prefixes (LPM) with subnet per Leaf Switch (typical)
Leaf switches
MAC Address (L2) on servers within rack only
IP addresses (Host Routes) for servers within rack only
Default route (ECMP) to Spine (LPM)

Figure 3: Layer 2 and Layer 3 Designs contrasted

Regardless of whether the network is built at layer 2 or layer 3 it's frequently the number of VMs that drives the networking table sizes. A currently modern x86 server is a dual socket with 6 or 8 CPU cores/socket. Typical enterprise workloads allow for 10 VMs/CPU core, such that for a typical server to have 60-80 VMs running is not unusual. It is foreseeable that this number will only get larger in the future.

For a design that is 10 VMs/CPU, quad-core CPUs with 2 sockets/server, 40 physical servers per rack and 20 racks of servers, this would drive the forwarding table requirements of the network as follows:

Table 2: Forwarding Table scale characteristics of Layer 2 and Layer 3 Designs

Forwarding Table	Layer 2 Design		Layer 3 Design	
	Spine Switches	Leaf Switches	Spine Switches	Leaf Switches
MAC Address (1 vNIC / VM)	1 MAC Address/VM x 10 VMs/CPU x 4 CPUs/socket x 2 sockets per server = 80 VMs/server x 40 servers/rack = 3,200 MAC addresses/rack x 20 racks = 64K MAC addresses		Minimal (Spine switches operating at L3 so L2 forwarding table not used)	1 MAC Address/VM x 10 VMs/CPU x 4 CPUs/ socket x 2S = 80 VMs/server x 40 servers/rack = 3,200 MAC addresses
IP Route LPM	Small number of IP Prefixes	None (Leaf switch operating at L2 has no L3)	1 subnet per rack x 20 racks = 20 IP Route LPM prefixes	Minimal (Single ECMP route towards Spine switches)
IP Host Route (IPv4 only)	1 IPv4 host route/VM 3200 IPv4 host routes/rack x 20 racks = 64K IP Host routes	None (Leaf switch operating at L2 has no L3)	Minimal (No IP Host Routes in Spine switches)	1 IPv4 host route/VM 3200 IPv4 host routes/rack = 3200 IP Host routes
IP Host Route (IPv4 + IPv6 dual stack)	1 IPv4 and IPv6 Host route/VM 64K IPv4 Host Routes + 64K IPv6 Host Routes	None (Leaf switch operating at L2 has no L3)	Minimal (No IP Host Routes in Spine switches)	1 IPv4 and IPv6 Host route/VM 3200 IPv4 Host Routes + 3200 IPv6 Host Routes

Layer 2 Spanning Tree Logical Port Count Scale

Despite the common concerns with large layer 2 networks (large broadcast domain, single fault domain, difficult to troubleshoot), one limiting factor often overlooked is the control-plane CPU overhead associated with running the Spanning Tree Protocol on the switches. As a protocol, Spanning Tree is relatively unique in that a failure of the protocol results in a 'fail open' state rather than the more modern 'fail closed' state. If there is a protocol failure for some reason, there will be a network loop. This characteristic of spanning tree makes it imperative that the switch control plane is not overwhelmed.

With Rapid Per VLAN Spanning Tree (RPVST), the switch maintains multiple independent instances of spanning tree (for each VLAN), sending/receiving BPDUs on ports at regular intervals and changing the port state on physical ports from Learning/Listening/ Forwarding/Blocking based on those BPDUs. Managing a large number of non-synchronized independent instances presents a scale challenge unless there is careful design of VLAN trunking. As an example, trunking 4K VLANs on a single port results in the state of each VLAN needing to be tracked individually.

Multiple Spanning Tree Protocol (MSTP) is preferable to RPVST as there are less instances of the spanning tree protocol operating and moving physical ports between states can be done in groups. Even with this improvement layer 2 logical port count numbers still need to be managed carefully.

The individual scale characteristics of switches participating in Spanning Tree varies but the key points to factor into a design are:

- The number of STP Logical Ports supported on a given switch (this is also sometimes referred to as the number of VlanPorts).
- The number of instances of Spanning Tree that are supported if RPVST is being used.

Less Tiers Compared to More Tiers

A design with more tiers offers higher scalability compared to a design with less tiers. However it trades this off against both higher capital expense (capex) and operational expense (opex). More tiers means more devices, which is more devices to manage as well as more ports used between switches for the fan-out interconnects between switches.

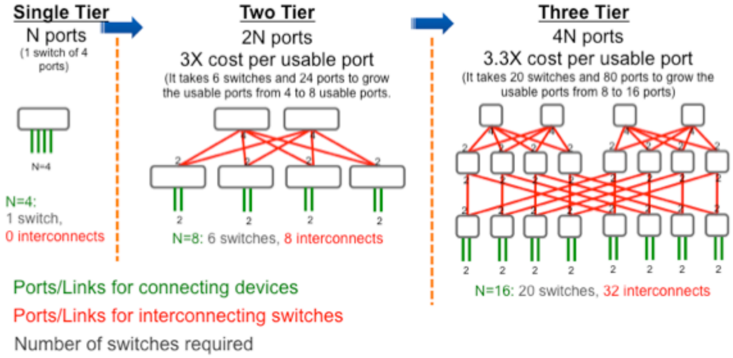

Single Tier
N ports
(1 switch of 4 ports)
N=4
N=4:
1 switch,
0 interconnects
Two Tier
2N ports
3X cost per usable port
(It takes 6 switches and 24 ports to grow the usable ports from 4 to 8 usable ports.
2
N=8: 6 switches, 8 interconnects
Three Tier
4N ports
3.3X cost per usable port
(It takes 20 switches and 80 ports to grow the usable ports from 8 to 16 ports)
2
N=16: 20 switches, 32 interconnects
Ports/Links for connecting devices
Ports/Links for interconnecting switches
Number of switches required

Figure 3: Layer 2 and Layer 3 Designs contrasted

Using a 4-port switch as an example for simplicity and using a non-oversubscribed Clos network topology, if a network required 4 ports, the requirements could be met using a single switch. (This is a simplistic example but demonstrates the principle).

If the port requirements double from 4 to 8 usable ports, and the building block is a 4-port switch, the network would grow from a single tier to two-tiers and the number of switches required would increase from 1 switch to 6 switches to maintain the non-oversubscribed network. For a 2x increase of the usable ports, there is a 3-fold increase in cost per usable port (in reality the cost goes up even more than 3x as there is also the cost of the interconnect cables or transceivers/fiber.)

If the port count requirements doubles again from 8 to 16, a third tier is required, increasing the number of switches from 6 to 20, or an additional 3.3 times increase in devices/cost for just a doubling in capacity. Compared to a single tier design, this 3-tier design now offers 4x more usable ports (16 compared to 4) but does so at over a 20x increase in cost compared to our original single switch design.

Capital expense (capex) costs go up with increased scale. However, capex costs can be dramatically reduced if a network can be built using fewer tiers as less cost is sunk into the interconnects between tiers. Operational expense (opex) costs also decrease dramatically with fewer devices to manage, power and cool, etc. All network designs should be looked at from the perspective of the cost per usable port (those ports used for servers/storage) over the lifetime of the network. Cost per usable port is calculated as:

$$\frac{\text{(cost of switches + power + optics + fiber)}}{\text{(total nodes x oversubscription)}}$$

Arista Scale-Out Designs

Starting Point

An accepted principle of network designs is that a given design should not be based on the short-term requirements but instead the longer-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 longer-term.

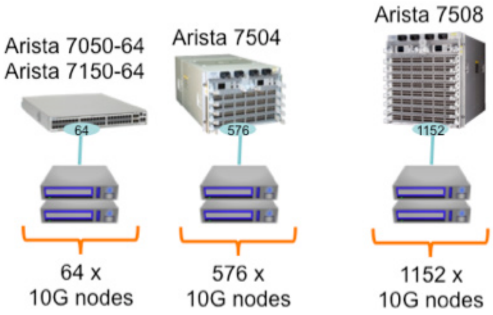


Figure 5: Maximum scale of a single tier

If the longer-term requirements for number of ports can be fulfilled in a single switch (or pair of switches in a HA design), then there's no reason why a single tier design couldn't be used. Single tier designs will always offer the lowest capex and opex as there are a minimal number of devices and no ports used for interconnecting tiers of switches.

For designs that don't fit a single tier then a two-tier design is the next logical step. A two-tier design has spine switches at the top tier and leaf switches at the bottom tier.

In a two-tier spine/leaf 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.

Arista Scale-Out Designs

Scale out designs start with one pair of spine switches and some quantity of leaf switches. A two-tier leaf/spine 3:1 oversubscribed network design for 48 servers per leaf switch (Arista 7050S-64) has 48x10G connections to devices and 16x10G uplinks to the spine (48x10G : 16x10G = 48:16 = 3:1 oversubscribed). A pair of Arista 7500 modular switches (7504 or 7508) is the spine. Each leaf switch attaches to each spine switch with 8x10G links. With a single DCS-7500E-36Q linecard (36x40G / 144x10G) in each spine switch, the initial network expands to enable connectivity for 18 leaf switches (864 x 10G attached devices @ 3:1 oversubscription end-to-end) as shown in Figure 6.

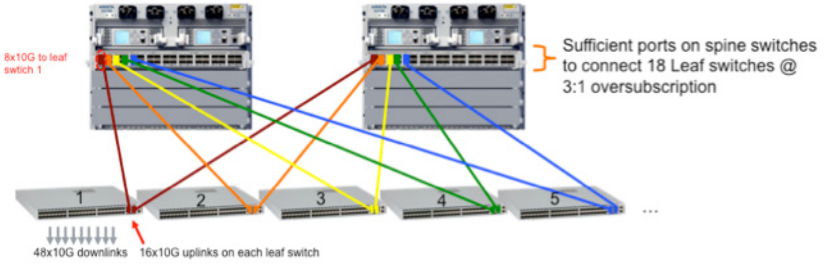

8x10G to leaf swtich 1
Sufficient ports on spine switches to connect 18 Leaf switches @ 3:1 oversubscription
1
2
3
4
5
...
48x10G downlinks
16x10G uplinks on each leaf switch

Figure 6: Starting point of a scale-out design: one pair of switches each with a single linecard

As more leaf switches are added and the ports on the first linecard of the spine switches are used, a second linecard is added to each chassis and half of the links are moved to the second linecard. The design can grow from 18 leaf switches to 36 leaf switches (1,728 x 10G attached devices @ 3:1 oversubscription end-to-end as shown in Figure 7.

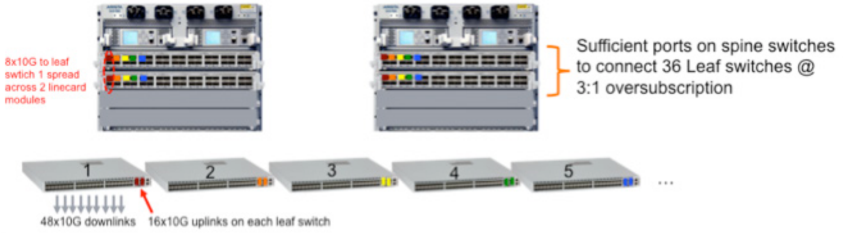

8x10G to leaf swtich 1 spread across 2 linecard modules
Sufficient ports on spine switches to connect 36 Leaf switches @ 3:1 oversubscription
1
2
3
4
5
...
48x10G downlinks
16x10G uplinks on each leaf switch

Figure 7: First expansion of spine in a scale-out design: second linecard module

This process repeats a number of times over. If the uplinks between the leaf and spine are at 10G then each uplink can be distributed across 4 ports on 4 linecards in each switch.

The final scale numbers of this design is a function of the port scale/density of the spine switches, the desired oversubscription ratio and the number of spine switches. Provided there are two spine switches the design can be built at layer 2 or layer 3. Final scale for two Arista 7504 spine switches is 72 leaf switches or 3,456 x 10G @ 3:1 oversubscription end-to-end. If the design used a pair of Arista 7508 switches then it is double that, i.e., 144 leaf switches for 6,912 x 10G @ 3:1 oversubscription end-to-end as shown in Figure 8.

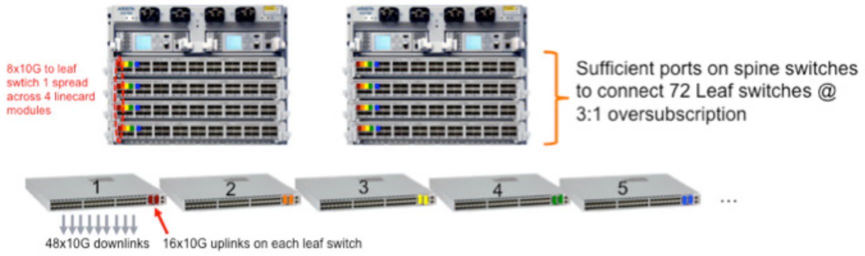

8x10G to leaf
swtich 1 spread
across 4 linecard
modules
Sufficient ports on spine switches
to connect 72 Leaf switches @
3:1 oversubscription
1
2
3
4
5
...
48x10G downlinks
16x10G uplinks on each leaf switch

Figure 8: Final expansion of spine in a scale-out design: add a fourth linecard module to each Arista 7504

Spine/Leaf Design 1G Nodes Using 2 Spine Arista 7500 Series

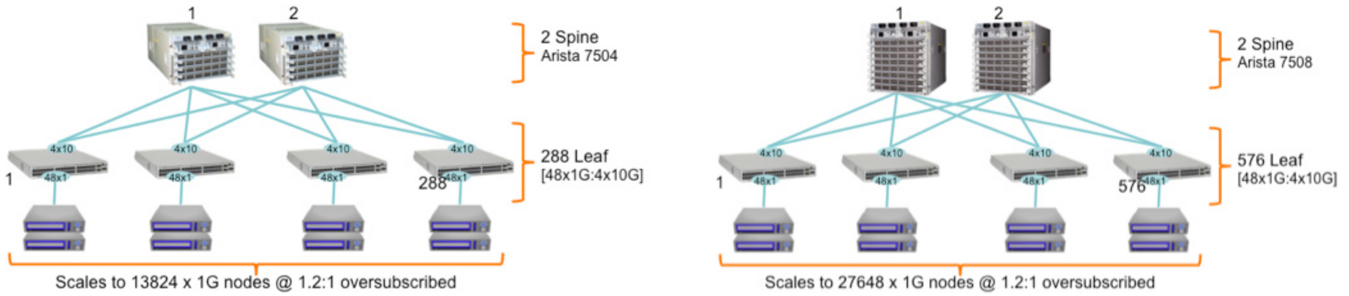


Figure 9: Spine/Leaf network design for 1G attached nodes using Arista 7504/7508 spine switches (maximum scale with 2 switches) with uplinks at 10G

Spine/Leaf Design 10G Nodes @ 3:1 Oversubscription Using 2 Spine Arista 7500 Series

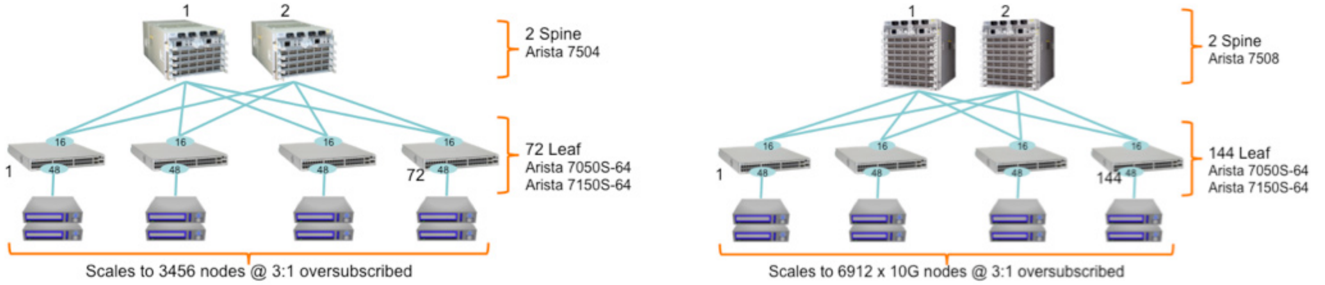


Figure 10: Spine/Leaf network design for 10G attached nodes @ 3:1 oversubscription using Arista 7504/7508 spine switches (maximum scale with 2 switches) with uplinks at 10G

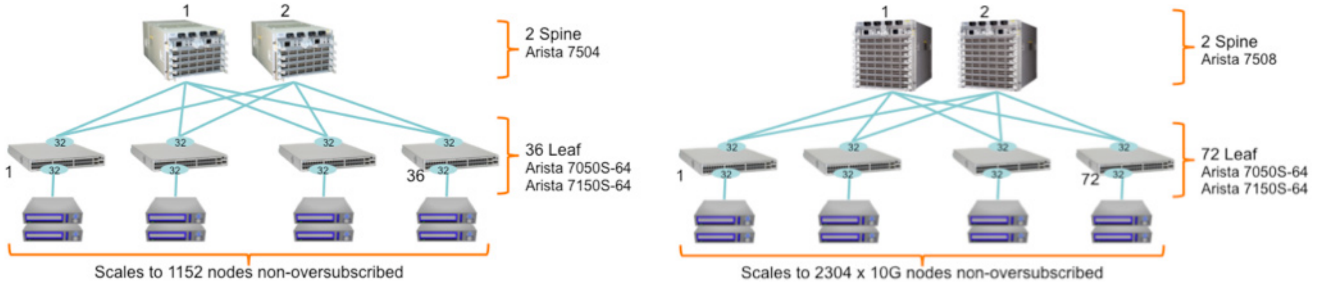


Figure 11: Spine/Leaf network design for 10G attached nodes non-oversubscribed using Arista 7504/7508 spine switches (maximum scale with 2 switches) with uplinks at 10G

These topologies can all be built at layer 2 or layer 3. If the designs are layer 2, MLAG provides an L2 network that runs active/active with no blocked links, which requires an MLAG peer-link between the spine switches.

It may also be desirable to use MLAG on the leaf switches to connect servers/storage in an active/active manner. In this case, a pair of leaf switches would be an MLAG pair and would have an MLAG peer-link between them. The MLAG peer-link can be a relatively small number of physical links (at least 2) as MLAG prioritizes network traffic so that it remains local to a switch for dual-attached devices.

Large-Scale Designs with 10G Uplinks

The designs can also scale out using layer 3 with up to as many as 32 spine switches in an ECMP layout allowing for a very large fan out of leaf switches. Just as a spine switch has linecard modules added to it over time as the network grows, the same approach can be used for spine switches too. A network may evolve from 2 spine switches to 4, 8, 16 and eventually as many as 32 spine switches. All paths between spine and leaf run active/active utilizing standard routing protocols like BGP and OSPF and up to 32-way ECMP is used to run all paths in active/active mode. The following diagrams demonstrate how a network can evolve from 4 spine switches to 8 and 16 in a 3:1 oversubscribed 10G design (see Figure 12.)

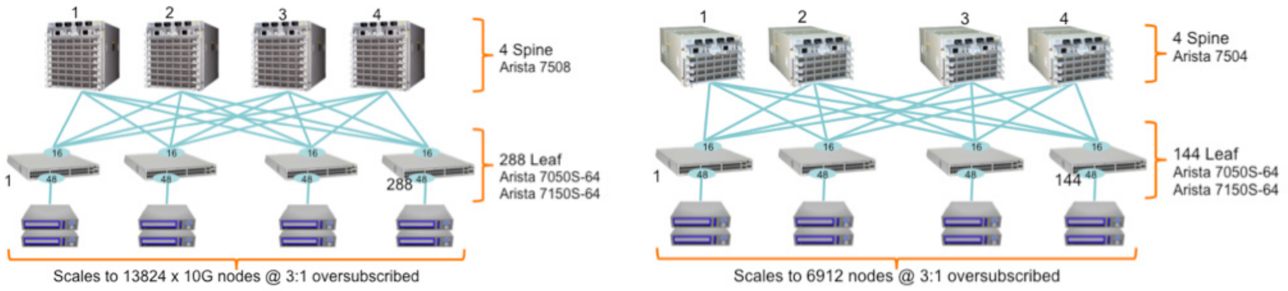


Figure 12: Arista 7504/7508 Spine 4-way ECMP to Arista 64-port 10G Leaf switches @ 3:1 Oversubscription

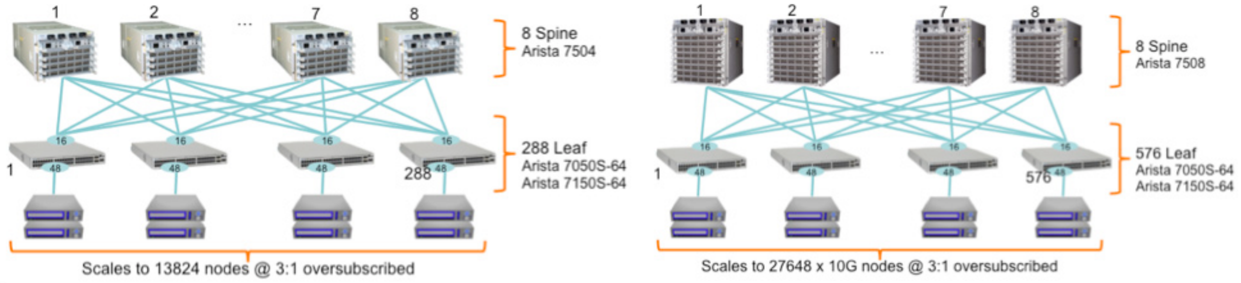


Figure 13: Arista 7504/7508 Spine 8-way ECMP to Arista 64-port 10G Leaf switches @ 3:1 Oversubscription

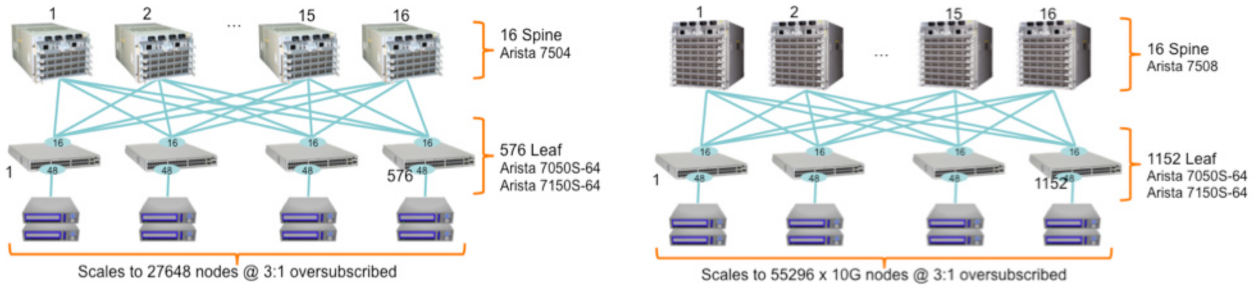


Figure 14: Arista 7504/7508 Spine 16-way ECMP to Arista 64-port 10G Leaf switches @ 3:1 Oversubscription

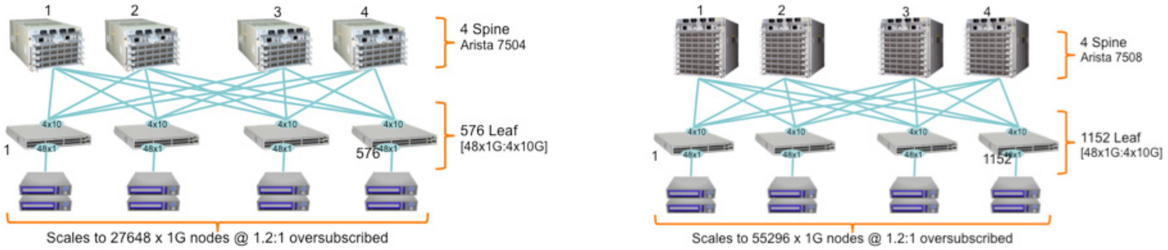


Figure 15: Arista 7504/7508 Spine 4-way ECMP to Arista 48x10G Leaf switches @ 1.2:1 Oversubscription

The same design principles can be applied to build a 10G network that is non-oversubscribed. The network size can evolve over time (pay as you grow) with a relatively modest up-front capex investment:

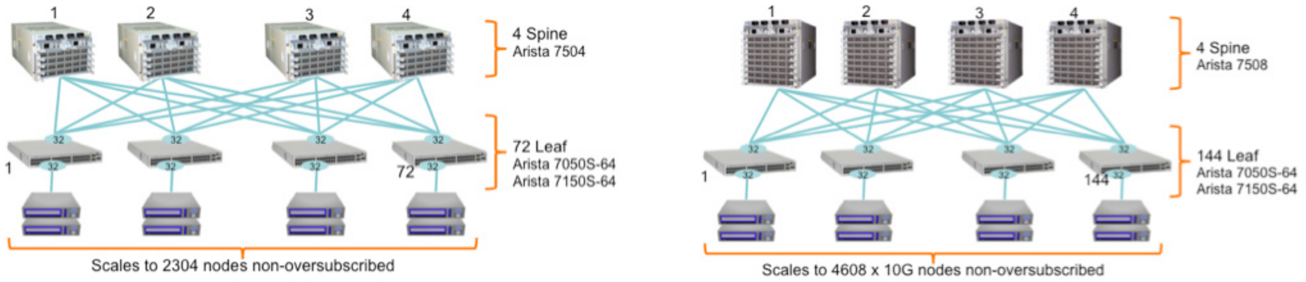


Figure 16: Arista 7504/7508 Spine 4-way ECMP to Arista 64-port 10G Leaf switches non-oversubscribed

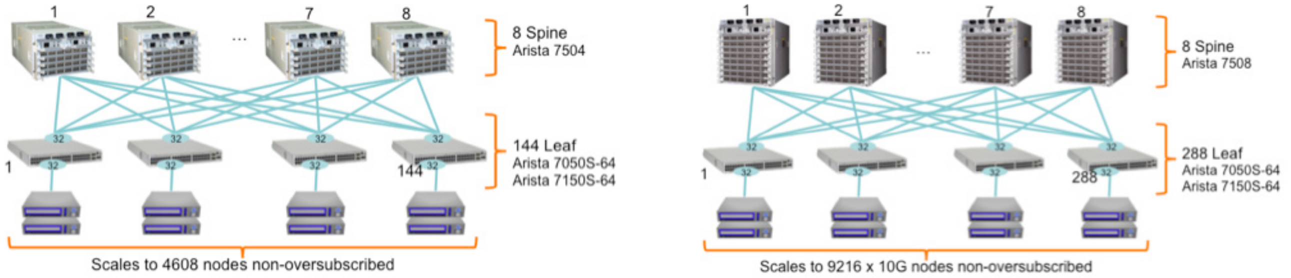


Figure 17: Arista 7504/7508 Spine 8-way ECMP to Arista 64-port 10G Leaf switches non-oversubscribed

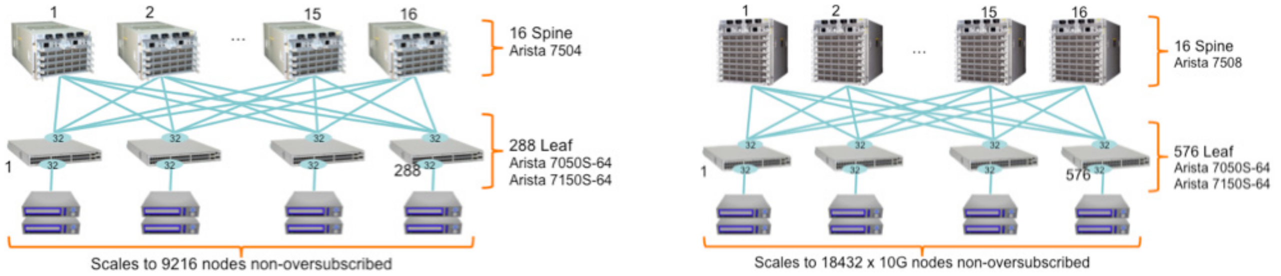


Figure 18: Arista 7504/7508 Spine 16-way ECMP to Arista 64-port 10G Leaf switches non-oversubscribed

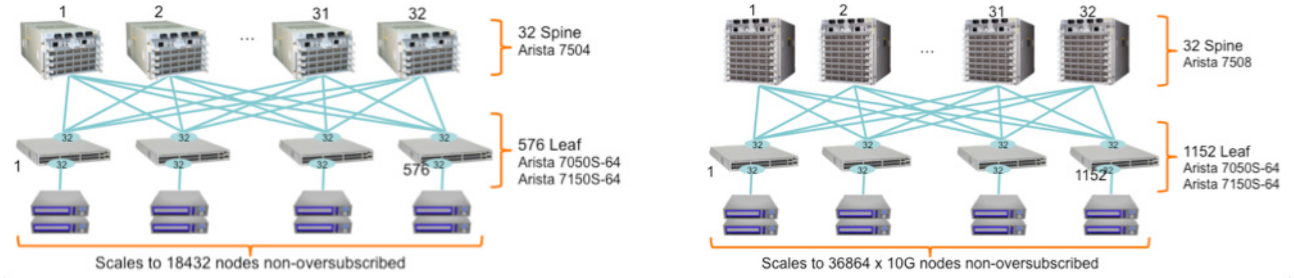


Figure 19: Arista 7504/7508 Spine 32-way ECMP to Arista 64-port 10G Leaf switches non-oversubscribed

Large-Scale Designs with 40G Uplinks

The same simple design principles can be used to build networks with 40G uplinks between spine/leaf instead of 10G uplinks. On Arista switches any 40G QSFP+ port can be configured as either 1x40G or 4x10G and using optics breakout to individual 10G links. Many designs can very easily evolve from 10G uplinks to 40G uplinks or support a combination. On Arista switch platforms that support AgilePorts (e.g. Arista 7150S), 4 SFP+ interfaces can be configured into a 40G port allowing further flexibility in selecting uplink speeds.

The following diagrams show the maximum scale using 40G uplinks from leaf to spine in a layer 3 ECMP design for 3:1 oversubscribed 10G nodes:

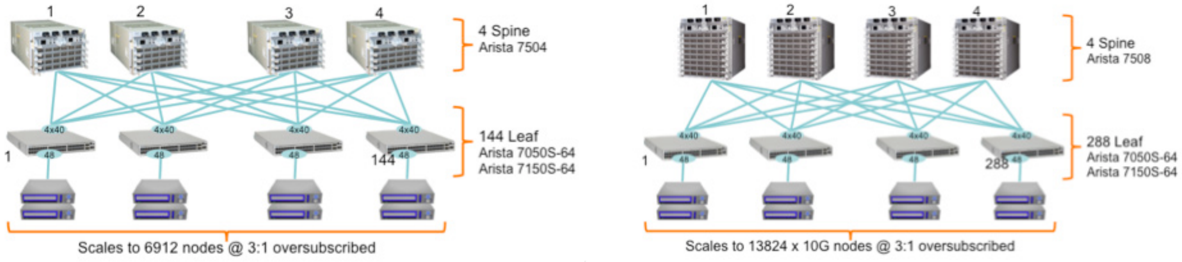


Figure 20: Arista 7504/7508 Spine 4-way ECMP to Arista 48x10G + 4x40G Leaf switches @ 3:1 Oversubscription

Optics, Cabling and Transceiver Choices

There are a variety of transceiver, optics and cabling choices available:

SFP+/SFP is the most common transceiver for 10G/1G with support for a wide range of distances:

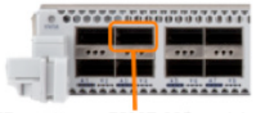
SFP+ port on an 7500E-48S/72S module

Figure 21: SFP+/SFP Port

Table 3: SFP+/SFP Transceiver Options

Type	Speed	Reach	Media	Notes
10GBASE-CR	10G	0.5m, 1m, 1.5m, 2m, 2.5m, 3m, 5m, 7m	Direct Attach (DAC) CX1 Twinax	Since cable is pre-terminated it comes with transceivers at both ends fused to a copper cable
10G-AOC	10G	3m to 30m	Active Optical Cable (AOC)	Since cable is pre-terminated it comes with transceivers both ends fused to a fiber cable
10GBASE-SRL	10G	100m (OM3) 150m (OM4)	50µ MMF	Optically interoperable with 10GBASE-SR up to 100m
10GBASE-SR	10G	100m (OM3) 150m (OM4)	50µ MMF	Optically interoperable with 10GBASE-SRL up to 100m
10GBASE-LRL	10G	1km	9µ SMF	Optically interoperable with 10GBASE-LR up to 1km
10GBASE-LR	10G	10km	9µ SMF	Optically interoperable with 10GBASE-LRL up to 1km
10GBASE-ER	10G	40km	9µ SMF	
10GBASE-ZR	10G	80km	9µ SMF	
10GBASE-DWDM	10G	40km/80km	9µ SMF	40+ wavelengths available
1000BASE-T	100M/1G	100m	Cat5e	100M support available on some switches/ transceivers
1000BASE-SX	1G	550m	50µ MMF	
1000BASE-LX	1G	10km	9µ SMF	

QSFP+ transceivers are used for 40G connectivity. These also allow breaking out a single physical port as either 1x40G or 4x10G:



QSFP+ port on an 7500E-36Q module

Figure 22: QSFP+ Port

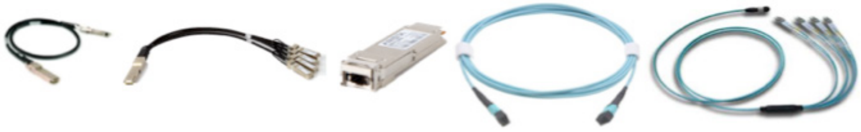


Table 4: QSFP+ Transceiver Options				
Type	Speed	Reach	Media	Notes
40GBASE-CR4	40G	0.5m, 1m, 2m, 3m, 5m, 7m	Direct Attach (DAC)	Since cable is pre-terminated it comes with transceivers both ends fused to a copper cable
40GBASE-CR4 to 4x10GBASE-CR	40G x 10G	0.5m, 1m, 2m, 3m, 5m	Direct Attach (DAC)	Since cable is pre-terminated it comes with QSFP+ one end and 4xSFP+ the other fused to a copper cable
40G-AOC	40G	3m to 100m	Active Optical Cable (AOC)	Since cable is pre-terminated it comes with transceivers both ends fused to a fiber cable
40G UNIV	40G	150m (OM3) 150m (OM4)	50µ MMF	Duplex MMF
40G UNIV	40G	500m	9µ SMF	Duplex SMF Optically interoperable with 40GBASE-LRL4 and 40GBASE-LR4 up to 500m
40GBASE-SR4	40G	150m (OM3) 150m (OM4)	50µ MMF	Can operate as 1x40G (40GBASE-SR4) or 4x10G (compatible with 10GBASE-SR/SRL)
40GBASE-XSR4	40G	300m (OM3) 450m (OM4)	50µ MMF	Can operate as 1x40G (40GBASE-XSR4) or 4x10G (compatible with 10GBASE-SR/SRL) Optically interoperable with 40GBASE-SR4 up to 150m
40GBASE-LR4	40G	10km	9µ SMF	Optically interoperable with 40GBASE-LRL4 up to 1km
40G-LRL4	40G	1km	9µ SMF	Optically interoperable with 40GBASE-LRL4 up to 1km
40G-PLR4	40G	10km	9µ SMF	Can operate as 1x40G (40GBASE-PLR4) or 4x10G (compatible with 10GBASE-LR/LRL) Optically interoperable with 40GBASE-PLRL4 up to 1km
40G-PLRL4	40G	1km	9µ SMF	Can operate as 1x40G (40GBASE-PLRL4) or 4x10G (compatible with 10GBASE-LR/LRL) Optically interoperable with 40GBASE-PLR4 up to 1km

Embedded 100G Optics are used on a range of Arista fixed and modular systems including the *Arista 7500E-72S-LC* and *7500E-12CM-LC* linecard modules to provide industry-standard *100GBASE-SR10, 40GBASE-SR4* and *10GBASE-SR* ports without requiring any transceivers. *This provides the most cost effective and highest density 10/40/100G connectivity in the industry.* Each port mates with a standard MPO/MTP cable (12 fiber pairs on the one cable/connector) and provides incredible configuration flexibility enabling one port to operate as any of:

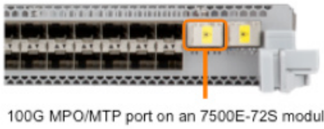


Figure 23: 100G MPO/MTP connector on an Arista Multi-speed Port (MXP

- 1 x 100GBASE-SR10
- 3 x 40GBASE-SR4
- 2 x 40GBASE-SR4 + 4 x 10GBASE-SR
- 1 x 40GBASE-SR4 + 8 x 10GBASE-SR
- 12 x 10GBASE-SR

These ports can be used with OM3/OM4 MMF supporting distances of 100m (OM3) and 150m (OM4). MPO to 12xLC patch cables break out to 12 x LC connectors for connectivity into SFP+.

Arista AgilePorts on some switches can use a group of 4 or 10 SFP+ ports to create an industry-standard 40GBASE-SR4 or 100GBASE-SR10. This provides additional flexibility in how networks can grow and evolve from 10G to 40G and 100G while providing increased flexibility in terms of distances supported.

100GbE CFP2 and QSFP100 transceivers are used for plug and play 100G connectivity on a variety of fixed and modular systems including the *Arista 7500E-6C2-LC (CPF2)* and *7500E-12CQ-LC (QSFP100)* linecard modules to provide support for a wide range of industry-standard 100G optics and cables. The CFP2 hot pluggable transceiver for the 7500E Series is approximately half the size of the first generation CFP optics and is optimized for density of 100G whilst enabling support for long haul options of up to a distance of 40km based on a power budget of up to 12W per transceiver. The QSFP format of the 100G optics is termed QSFP100 to differentiate it from the 40GbE QSFP+ design but shares a common mechanical design. This allows QSFP 100G ports to support both 40GbE and 100GbE optics for dual speed capability. In addition the QSFP100 form factor allows for higher density interfaces due the size and lower power budget. A variety of 100GbE optics in CFP2 and QSFP100 form factors are available including cables, short reach and long reach options.

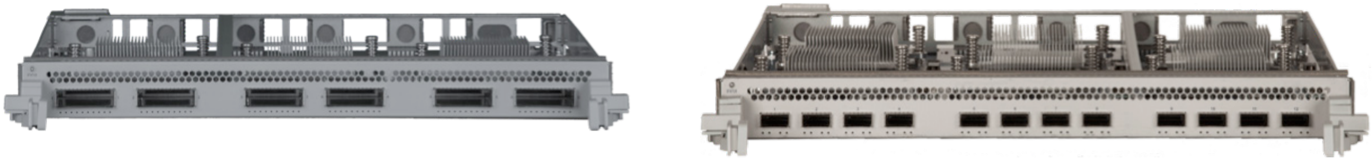


Figure 24: Arista 7504/7508E Series CFP2 and QSFP100 linecards

Arista EOS Foundation Features That Enable These Designs

Arista's scale-out cloud network designs are underpinned on a number of foundation features of Arista's award-winning Extensible Operating System:

Multi Chassis Link Aggregation (MLAG)

MLAG enables devices to be attached to a pair of Arista switches (an MLAG pair) with all links running active/active. MLAG eliminates bottlenecks, provides resiliency and enables layer 2 links to operate active/active without wasting 50% of the bandwidth as is the case with STP blocked links. L3 Anycast Gateway (Virtual ARP / VARP) with MLAG enables the L3 gateway to operate in active/active mode without the overhead of protocols like HSRP or VRRP.

To a neighboring device, MLAG behaves the same as standard link aggregation (LAG) and can run either with Link Aggregation Control Protocol (LACP) (formerly IEEE 802.3ad, more recently IEEE 802.1AX-2008) or in a static 'mode on' configuration.

The MLAG pair of switches synchronize forwarding state between them such that the failure of one node doesn't result in any disruption or outage as there are no protocols to go from standby to active, or new state to learn as the devices are operating in active/active mode.

Zero Touch Provisioning (ZTP)

ZTP enables switches to be physically deployed without any configuration. With ZTP, a switch loads its image and configuration from a centralized location within the network. This simplifies deployment, enabling network engineering resources to be used for more productive tasks by avoiding wasting valuable time on repetitive tasks such as provisioning switches or requiring network engineers to walk around with serial console cables.

An extension to ZTP, Zero Touch Replacement (ZTR) enables switches to be physically replaced, with the replacement switch picking up the same image and configuration as the switch it replaced. Switch identity and configuration aren't tied to switch MAC address but instead are tied to location in the network where the device is attached (based on LLDP information from neighboring devices). While a hardware failure and RMA is not likely to be a common event, ZTR means that in this situation the time-to-restoration is reduced to the time it takes for a new switch to arrive and be physically cabled, and is not dependent on a network engineer being available to provide device configuration, physically in front of the switch with a serial console cable.

VM Tracer

As virtualized datacenters have grown in size, the physical and virtual networks that support them have also grown in size and complexity. Virtual machines connect through virtual switches and then to the physical infrastructure, adding a layer of abstraction and complexity. Server side tools have emerged to help VMware administrators manage virtual machines and networks, however equivalent tools to help the network administrator resolve conflicts between physical and virtual networks have not surfaced.

Arista VM Tracer provides this bridge by automatically discovering which physical servers are virtualized (by talking to VMware vCenter APIs), what VLANs they are meant to be in (based on policies in vCenter) and then automatically apply physical switch port configurations in real time with vMotion events. This results in automated port configuration and VLAN database membership and the dynamic adding/removing VLANs from trunk ports.

VM Tracer also provides the network engineer with detailed visibility into the VM and physical server on a physical switch port while enabling flexibility and automation between server and network teams.

VXLAN

VXLAN is a multi-vendor industry-supported network virtualization technology that enables much larger networks to be built at layer 2 without the inherent scale issues that underpin large layer 2 networks. It uses a VLAN-like encapsulation technique to encapsulate layer 2 Ethernet frames within IP packets at layer 3 and as such is categorized as an 'overlay' network.

VXLAN provides solutions to a number of underlying issues with layer 2 network scale, namely:

- Enables large layer 2 networks without increasing the fault domain
- Scales beyond 4K VLANs
- Enables layer 2 connectivity across multiple physical locations or pods
- Potential ability to localize flooding (unknown destination) and broadcast traffic to a single site
- Enables large layer 2 networks to be built without every device having to see every other MAC address

From a virtual machine perspective, VXLAN enables VMs to be deployed on any server in any location, regardless of the IP subnet or VLAN that the physical server resides in.

VXLAN is an industry-standard method of supporting layer 2 overlays across layer 3. As multiple vendors support VXLAN there are subsequently a variety of ways VXLAN can be deployed: as a software feature on hypervisor-resident virtual switches, on firewall and load-balancing appliances and on VXLAN hardware gateways built into L3 switches. Arista's approach to VXLAN is to support hardware-accelerated VXLAN gateway functionality across a range of switches, progressively enabled through CY2013.

LANZ

Arista Latency Analyzer (LANZ) enables tracking of network congestion in real time before congestion causes performance issues. Today's systems often detect congestion when someone complains, "The network seems slow." The network team gets a trouble ticket, and upon inspection can see packet loss on critical interfaces. The best solution historically available to the network team has been to mirror the problematic port to a packet capture device and hope the congestion problem repeats itself.

Now, with LANZ's proactive congestion detection and alerting capability both human administrators and integrated applications can:

- Pre-empt network conditions that induce latency or packet loss
- Adapt application behavior based on prevailing conditions
- Isolate potential bottlenecks early, enabling pro-active capacity planning
- Maintain forensic data for post-process correlation and back testing

Arista EOS: A Platform For Stability and Flexibility

The Arista Extensible Operating System, or EOS, is the most advanced network operating system available. It combines modern-day software and O/S architectures, transparently restartable processes, open platform development, an un-modified Linux kernel, and a stateful publish/subscribe database model.

At the core of EOS is the System Data Base, or SysDB for short. SysDB is machine generated software code based on the object models necessary for state storage for every process in EOS. All inter-process communication in EOS is implemented as writes to SysDB objects. These writes propagate to subscribed agents, triggering events in those agents. As an example, when a user-level ASIC driver detects link failure on a port it writes this to SysDB, then the LED driver receives an update from SysDB and it reads the state of the port and adjusts the LED status accordingly. This centralized database approach to passing state throughout the system and the automated way the SysDB code is generated reduces risk and error, improving software feature velocity and provides flexibility for customers who can use the same APIs to receive notifications from SysDB or customize and extend switch features.

Arista's software engineering methodology also benefits our customers in terms of quality and consistency:

- Complete fault isolation in the user space and through SysDB effectively convert catastrophic events to non-events. The system self-heals from more common scenarios such as memory leaks. Every process is separate, with no IPC or shared memory fate-sharing, endian-independent, and multi-threaded where applicable.
- No manual software testing. All automated tests run 24x7 and with the operating system running in emulators and on hardware Arista scales protocol and unit testing cost effectively.
- Keep a single system binary across all platforms. This improves the testing depth on each platform, improves time-to-market, and keeps feature and bug resolution compatibility across all platforms.

EOS provides a development framework that enables the core concept of Extensibility. An open foundation, and best-in-class software development models deliver feature velocity, improved uptime, easier maintenance, and a choice in tools and options.

Arista EOS Extensibility

Arista EOS provides full Linux shell access for root-level administrators, and makes a broad suite of Linux based tools available to our customers. In the spirit of 'openness' the full SysDB programming model and API set are visible and available via the standard bash shell. SysDB is not a "walled garden" API, where a limited subset of what Arista uses is made available. All programming interfaces that Arista software developers use between address spaces within EOS are available to third party developers and Arista customers.

Some examples of how people customize and make use of Arista EOS extensibility include:

- Want to back up all log files every night to a specific NFS or CIFS share? Just mount the storage straight from the switch and use rsync or rsnapshot to copy configuration files
- Want to store interface statistics or LANZ streaming data on the switch in a round-robin database? Run MRTG right on the switch.
- Like the Internet2 PerfSonar performance management apps? Just run them locally.
- Want to run Nessus to security scan a server when it boots? Create an event-handler triggered on a port coming up.
- Using Chef, Puppet, CFEngine or Sprinkle to automate your server environment? Use any or all of these to automate configuration and monitoring of Arista switches too.
- Want to PXE boot servers straight from the switch? Just run a DHCP and TFTP server right on the switch.

If you're not comfortable running code on the same Linux instance as what EOS operates on we allow guest OSs to run on the switch via KVM built in. You can allocate resources (CPU, RAM, vNICs) to Guest OSs and we ship switches with additional flash storage via enterprise-grade SSDs.

Other Software Defined Cloud Networking (SDCN) Technologies

In addition to the EOS foundation technologies outlined, Arista Software Defined Cloud Networking (SDCN) incorporates various other technologies that enable scale-out automated network designs. Some of these other technologies include:

- Advanced Event Monitoring (AEM)
- Automated Monitoring/Management
- Arista CloudVision
- Arista eAPIs
- OpenFlow, OpenStack, Open Virtual Switch integration and others

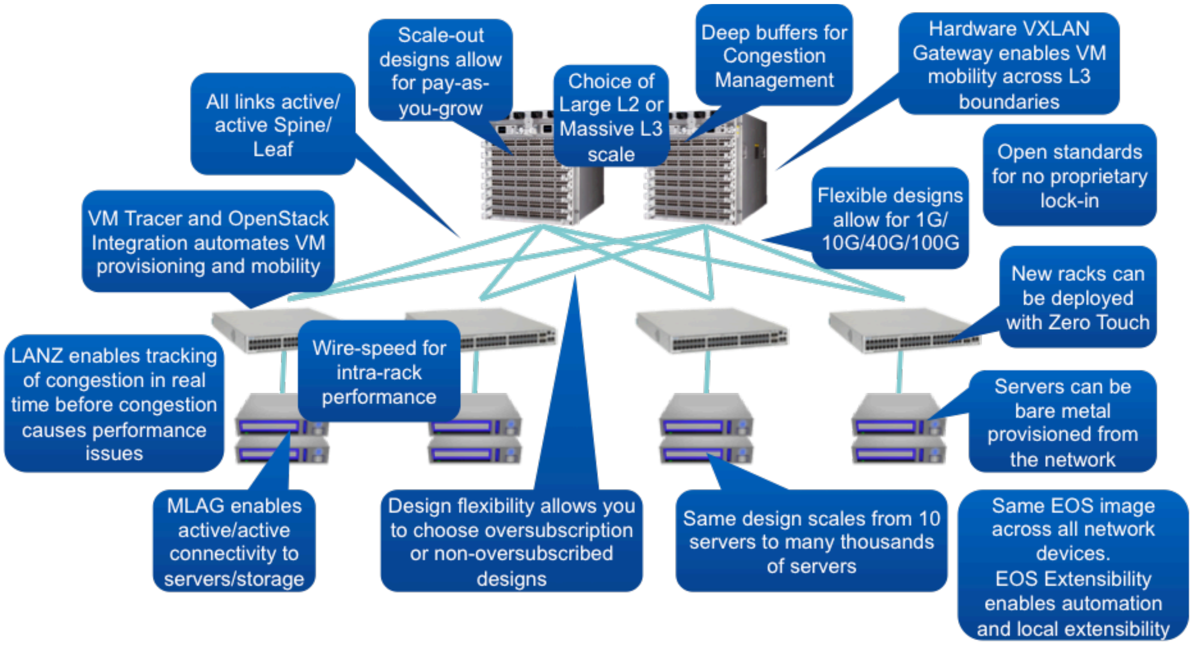

Scale-out designs allow for pay-as-you-grow
Deep buffers for Congestion Management
Hardware VXLAN Gateway enables VM mobility across L3 boundaries
All links active/ active Spine/ Leaf
Choice of Large L2 or Massive L3 scale
Open standards for no proprietary lock-in
VM Tracer and OpenStack Integration automates VM provisioning and mobility
Flexible designs allow for 1G/ 10G/40G/100G
New racks can be deployed with Zero Touch
LANZ enables tracking of congestion in real time before congestion causes performance issues
Wire-speed for intra-rack performance
Servers can be bare metal provisioned from the network
MLAG enables active/active connectivity to servers/storage
Design flexibility allows you to choose oversubscription or non-oversubscribed designs
Same design scales from 10 servers to many thousands of servers
Same EOS image across all network devices.
EOS Extensibility enables automation and local extensibility

Figure 25: Arista EOS foundation features and cloud network scalability

Conclusion

Arista's scale-out cloud network designs take the principles that have made cloud computing compelling (automation, self-service provisioning, linear scaling of both performance and economics) and combine them with the principles of Software Defined Networking (network virtualization, custom programmability, simplified architectures, and more realistic price points).

This combination creates a best-in-class software foundation for maximizing the value of the network to both the enterprise and service provider datacenter: a new architecture for the most mission-critical location within the IT infrastructure that simplifies management and provisioning, speeds up service delivery, lowers costs and creates opportunities for competitive differentiation, while putting control and visibility back in the hands of the network and systems administrators.

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

Copyright © 2016 Arista Networks, Inc. All rights reserved. CloudVision, and EOS are registered trademarks and Arista Networks is a trademark of Arista Networks, Inc. All other company names are trademarks of their respective holders. Information in this document is subject to change without notice. Certain features may not yet be available. Arista Networks, Inc. assumes no responsibility for any errors that may appear in this document. Sep 2016 02-0019-01