WHITE PAPER

The Evolution of the Enterprise Access Edge

A Primer on Chassis Models, Stack Technology, and Spine-Leaf Architecture

Table of Contents

Introduction

The enterprise access edge has been dominated by chassis models and stacking technologies for the past 20 years. However, with increased needs for security, stability, and capacity, both the chassis model and the stacking model have limitations that often result in vendor lock-in, difficult expansion options, or network failures. For access network switching, enterprises today require a more flexible, software-defined networking (SDN) technology that can be deployed in an automated fashion while providing less risk of failure.

Overview of the chassis model

For decades the access edge has been dominated by the chassis form factor. But this design was predicated on a common networking closet with all cables running back to a central location. Expansion required the addition of another chassis, which was seen as expensive to add just a few ports.

The chassis models normally had two processing modules that controlled the switch. This created a redundancy model where if a single processing module failed, the chassis would be able to function. However, if the failure resulted in both processing modules failing (or failing in succession) then all the line card modules in the chassis would be off-line.

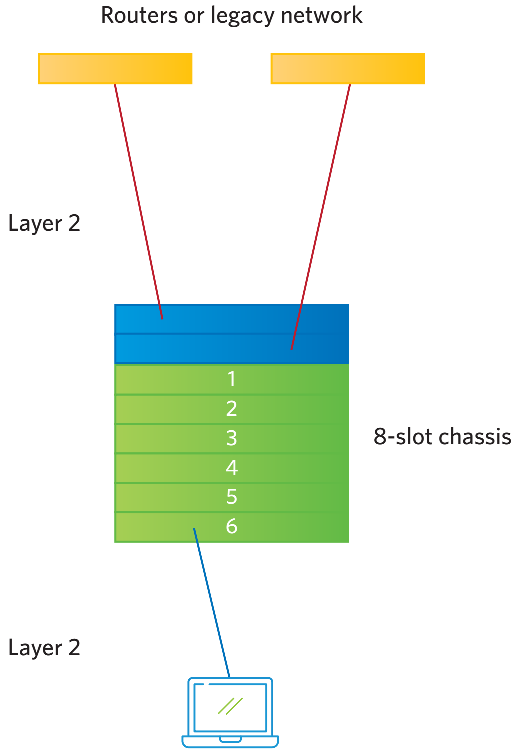

Figure 1 - Access edge chassis model

Consider that a chassis model is just a collection of line cards connected to each other via a shared backplane (see Figure 1). In this example, each line card could represent an individual leaf switch and the fixed backplane could represent a single spine switch to which all the leaf switches (line cards) connect. Even in chassis technology, L2 is typically not utilized in the backplane connections between the line cards. Typically, an optical or electrical layer 1 technology is utilized to interconnect the line cards. This layer 1 connectivity transports the L2 for bridge domains that span multiple line cards. So, by deconstructing the chassis, we have demonstrated that even in a chassis it could be considered as multiple switches interconnecting with each other.

Stacking technologies improve modularity

Stacking technologies were introduced to offset the need to purchase a large chassis when the customer only needed a way to grow by adding blades without the need to purchase a chassis. This allowed for a relatively economical method to expand the access footprint. However, stacking technologies were also proprietary, requiring special cables, major changes in software to run the stacked set of devices as one, weakening the distributed and resilient nature of the solution otherwise. These cables were often kept at rather short distances (often less than 10m) so this still required the centralized cabling design. Given that many switch stacking solutions were utilized to replace older chassis models, this fit the need of the time.

Given that stacking technologies were created to act like mini chassis models, most stacking technologies had the concept of a primary and a secondary stack switch control. Therefore, the same fault problem applies to the stacking technologies. If a problem results in both the primary and secondary stack switch failing, then the entire stack of switches would be off-line. The stacking technologies had another issue in that if the stacking cables failed, there was the possibility of a split stack. This is where both the active and the standby stack control switches are on-line and connected to the network, but they cannot communicate with each other. This could result in all or some of the stack becoming inaccessible to the network.

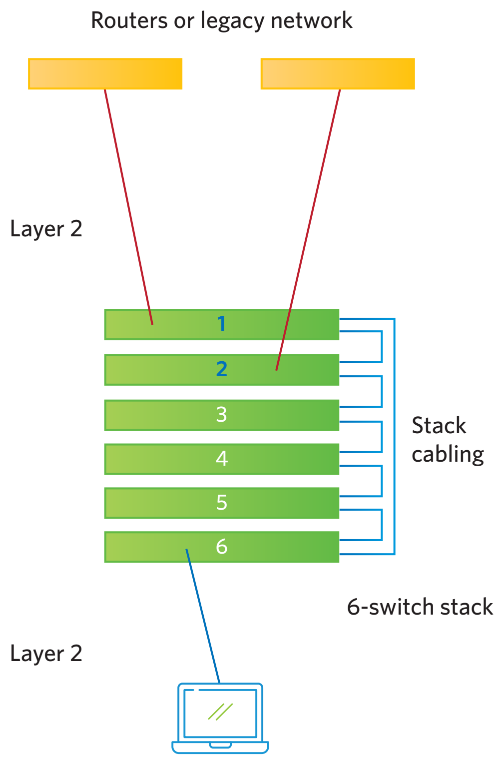

Figure 2 - Access edge stacking technology

As seen in Figure 2, stacking technologies are already multiple switches interconnected. But much like the chassis, these switches are interconnected with physical links. Often, these are proprietary optical links that are just the chassis backplane but in cable form. Most often the stacking cables are limited to two different optical rings (to provide redundancy between the member switches). This is no better than a chassis with redundant backplane paths.

Challenges of the chassis model and switching technology

However, both of the models, chassis and stacking, have major disadvantages when compared to modern software defined networking architectures. Particularly, the fixed backplane that is provided by both the chassis and the stacking technologies. In the chassis model, primarily the backplane was a fixed throughput. If an enterprise needed higher throughput, this meant that the chassis would need to be replaced with another chassis which had a higher backplane throughput. Some vendors developed a non-fixed backplane chassis design. However, in most of these implementations, the matrix of which line cards were supported with which fabric modules created an artificial limitation on the expandability of the chassis' backplane. In the stacking technologies, the backplane of the stack of switches was limited by the stacking ports and the cables needed to interconnect. Most of the stacking solutions did not allow for upgrading of the stacking ports to higher speeds. So if a higher throughput was needed, the entire stack would need to be replaced.

The chassis model often locked enterprises into a proprietary form factor as theses chassis models were often sold only in certain sizes. So, for solutions that needed a total number of ports that were in-between sizes, the enterprise often had to dedicate more cabinet space than required. Stacking technologies often restricted which switch models could be utilized in a given switch stack. Spine-leaf technology solves both issues by allowing the addition of a single switch into the spine-leaf topology. This switch can be of any size.

The SDN spine-leaf architecture

As enterprises look to update and transform their access network, enterprises should consider more modern methods for access network architecture. And in this case, the access network can learn from the data center transformation. Twenty years ago, data centers and cloud technologies were dominated by the chassis model. However, as software-defined networking became more mainstream, the model transformed from the Layer 2 chassis model to the spine-leaf Layer 3 architecture. From an enterprise perspective, this was a direct result of wanting to eliminate the problems that arose from layer 2 failures (Spanning tree failures, Broadcast storms, ARP storms, etc.). However, in the cloud compute model, it was more due to speed reasons which allowed implementation of additional capacity as well as to limit the failure domain of a single network switch failure. With the classic chassis, as many as 768 network ports (server ports) could be offline with a single chassis failure or get impacted by a single bug. However, in the software-defined networking (SDN) model, as few as 48 network ports (server ports) would be off-line.

In the spine-leaf architecture, no single failure of a leaf nor a spine causes the entire spine-leaf model to fail. Also, there is no artificial limitation of the number of spines nor the number of leaf nodes that can be deployed in a spine-leaf technology. The only limitation to the number of spines and number of leaf nodes is the number of ports in each of the switches that can be utilized for the interconnections. This provides for flexibility in access network designs. This also provides for an easier path to adoption since the same technologies that are used in the enterprise data center are now utilized in the access edge.

Traditional access networking relied upon Layer 2 and spanning tree to bridge all the ports together in bridge domains (VLANs) to support the local access connectivity requirements. This enabled many chassis-based switches to belong to the same bridge domain, but as multiple chassis were interconnected, this created Layer 2 loops. The spanning tree protocol is available to prevent the looping of broadcast, unknown unicast, and multicast (BUM) traffic which would result with access network meltdown. However, the spanning tree protocol relies upon every switch in the Layer 2 bridge domain to behave properly. If one of the people does not behave properly, there can be a meltdown in the network. This is very similar to a traffic intersection.

The traffic signals only prevent car crashes if each operator of the vehicles obeys the traffic laws. When an operator disobeys a traffic signal, then a crash happens. This means that Layer 2 bridge domains are subject to meltdown due to software or hardware failures of the switches in the domains. Data centers and cloud compute models solved this by utilizing virtual extensible local area network (VXLAN) technologies. Instead of utilizing Layer 2 protocols to communicate between the switches which share a common bridge domain, Layer 3 protocols are utilized and the layer 2 bridge domain is tunneled between relevant switches. The Layer 3 protocol utilized is the BGP-EVPN. This is an extension of BGP that allows for ethernet segments to be extended across the layer 3 fabric via a virtual private network. The BGP-EVPN model utilizes the split horizon concept to eliminate the looping of the layer 2 BUM traffic. In the Layer 2 model, each switch was not aware of the origination of the BUM traffic, and as a result, switches would just flood BUM traffic that was received. However, in the BGP-EVPN model, the originating switch is included in the advertisement of the BUM traffic. Therefore, when a switch receives the BUM traffic, it will not flood the traffic preventing loops.

Also, the use of a Layer 3 protocol to interconnect the switches eliminated the restrictions on the number of switches that could be interconnected. The only limitation is the number of ports in a switch that can be used to interconnect.

Advantages of the spine-leaf architecture

By replacing a chassis with a spine-leaf architecture, the spine-leaf architecture provides the following advantages:

1. **Distributed architecture** – Not all of the switches need to be placed within a particular location.
2. **Flexible and upgradeable backplane** – The effective throughput is not limited by a single switch or cable. Increases in throughput can be achieved by implementation of new switches with higher throughput.
3. **Isolated fault domains (no single point of failure)** – the use of multiple switches that independently classify and forward traffic creates multiple failure domains.
4. **Highly scalable access architecture** – The architecture is not limited to a single location nor a single switch platform. Multiple different switch types in a multitude of different topologies can be deployed.

5. Flexibility in switch choice – This technology has no limitation on the switch models utilized in the architecture. Any BGP-EVPN VXLAN capable switch can be utilized.
6. Standard protocols – This solution relies on standard BGP-EVPN and VXLAN which provides for interoperation between different vendors which comply to BGP-EVPN and VXLAN standards.
7. No vendor lock-in – Assuming that all switches comply with BGP-EVPN and VXLAN standards, the enterprise can utilize any switch in the topology.

Unlike the chassis model architecture, the SDN-based switches utilized in the spine-leaf architecture do not need to be placed in a centralized location. These individual switches can be placed at any location that is within the cabling limitations (100m for copper ethernet but can be as far as 2000m for fiber ethernet).

Leaf switch interconnectivity and redundancy

In the spine-leaf architecture, all leaf switches connect to one or more spine switches. Two spine switches is the typical minimum for a spine-leaf architecture (to avoid a single point of failure); however, there is no technical reason a single spine or more than two spine switches could not be utilized. But a single spine switch would not be ideal as failure of that single spine switch would cause the access network to fail. In the same vein, a spine-leaf architecture does not require two and only two spine switches. To expand or add additional redundancy, more than two spine switches can be utilized. In fact, BGP-EVPN allows for a minimum of 8 equal cost paths and many vendors support up to 32 equal cost paths. This implies that the limitations on the number of spines in a spine-leaf is around 8 to 32 spine switches. However, this assumes that all leaf switches connect to all the same spine switches. But the spine-leaf architecture does not require that all leaf switches must connect to all the same spine switches.

Consider that there are six spine switches within the local access network and there are 44 leaf switches. If my spine switches are 32 port switches, and my leaf switches are 48 access port switches with four uplink ports, and each spine utilized two ports to connect to the routed access out of the site, then my access network would have each leaf switch connected to certain spine switches in a pattern that would keep it contiguous. Consider leaf switches L1 through L44 and spine switches S1 to S6. Then leaf L1 would connect to spine switches S1, S2, S3, and S4. Leaf switch L2 would connect to spines S2, S3, S4, and S5. Leaf switch 3 would connect to spines S3, S4, S5, and S6. Leaf switch L4 would connect to S4, S5, S6, and S1 (repeating back to the original spine and creating a contiguous fabric) as depicted in Figure 3. This pattern would continue until all leaf switches were connected and would result in four ports not utilized and allowing for one more leaf switch expansion. But this design would allow for multiple failures without a single switch from being isolated from the enterprise network. Any spine failure would result in only two ports out of the site being down, and only 29 or 30 inter-switch connections being down. Any failure of a leaf switch would result in only the loss of 48 access ports.

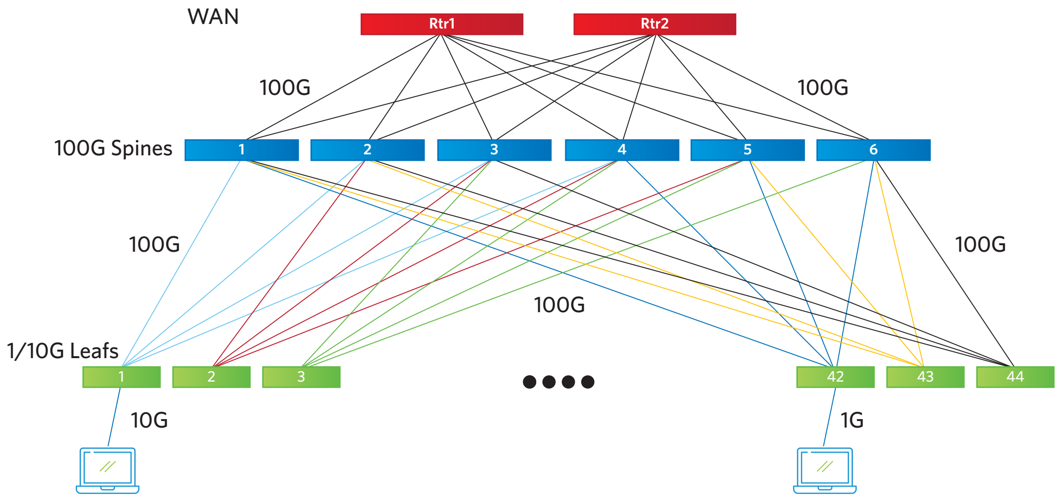

Figure 3 - Spine-leaf architecture

Leaf switch uplink requirements

Unlike the flexible chassis fabric, the spine-leaf architecture only requires that the uplinks of the leaf switches have a compatible connection on the spine switch. So, if some of the leaf switches utilize 10G as the uplink (consider a 48 port 1G switch with 4x10G uplinks), the spines only need to be able to terminate a 10G connection. But if in the same spine leaf architecture, some leaf switches connected at 100G, the spine switches would only need to support the 100G connections. Consider a 48 port 10G with 6x100G uplink spine. If each spine utilized one 100G link to connect north out of the access network, then this architecture could support 48 leaf switches (48x1G, 4x10G) and 5 leaf switches (48x10G, 6x100G) with just 4 spine switches as depicted in Figure 4. If added throughput was needed, additional spine switches could be added to provide additional bandwidth north out of the access site, or to provide additional bandwidth between all the leaf switches.

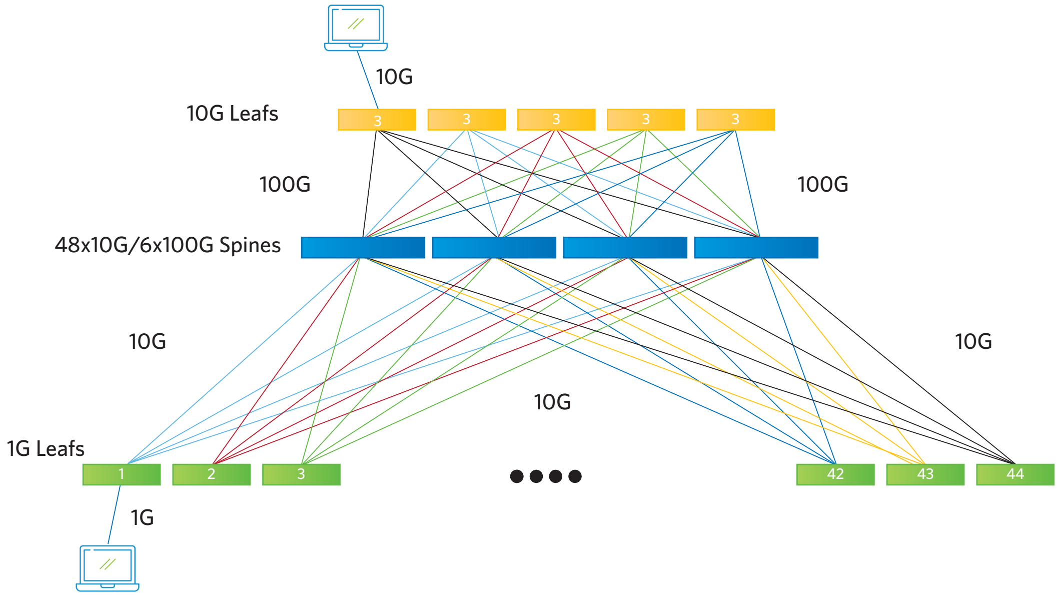

Figure 4 - Spine-leaf mixed switch topology

Multi-tiered spine-leaf architecture

A typical two-tier architecture is all that is needed to satisfy very large access networks at a site; however, there is nothing in the spine-leaf architecture that prohibits a multi-tiered architecture to further expand the size and bandwidth capacity of the local access. Consider our previous scenario of 48 1G access switches and five 10G access switches. If the number of 10G access switches increased, then there would be a need for more 100G connections. Without purchasing new 1G access switches (with 100G uplinks vs 10G uplinks), an additional tier of 100G could be created. The 100G spine (Tier 1) would consist of 32 port 100G switches. This would allow for each switch to have 4 100G links north out of the access location and allow the other 28 ports to be utilized to connect to either 10G access leaf switches or the second spine tier. So, the second spine tier would be four 48x10G/6x100G switches. These second-tier spine switches would utilize four 100G links out of each of the first-tier spine switches. This would leave 24 100G links in the first-tier spine switches to connect 24 10G access switches. The second-tier spine switches could connect 48 1G access switches as seen in Figure 5.

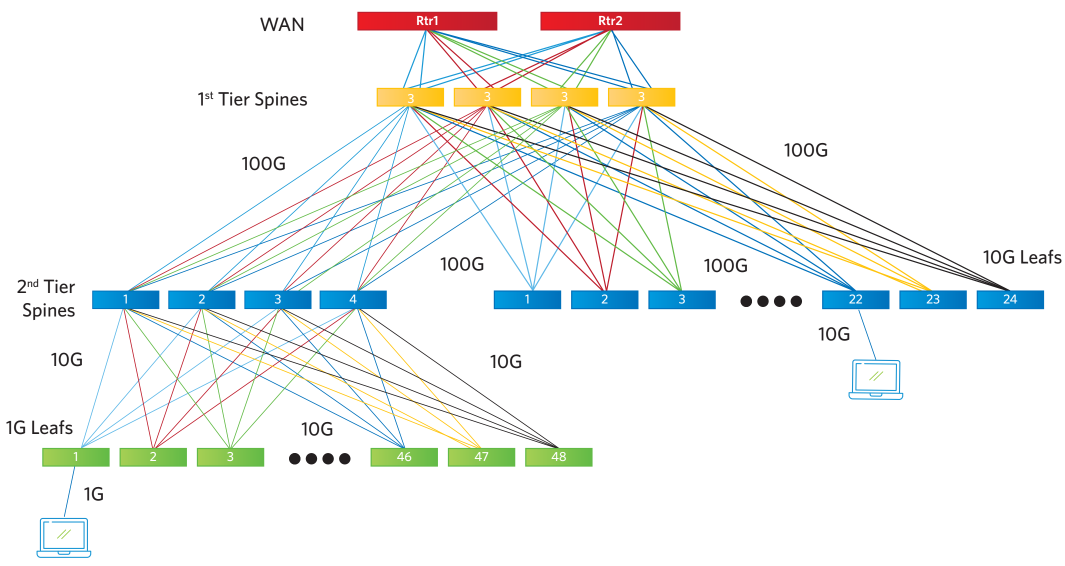

Figure 5 - Two-tiered spine-leaf architecture

Since standard routing protocols are utilized in the spine-leaf architecture, the number of tiers is not limited. So this is a very extensible and flexible architecture. It can fit a local access network of any size.

The Layer 2 network is extended over the Layer 3 fabric utilizing VXLAN and BGP-EVPN. This is achieved by associating a VXLAN tunnel endpoint (VTEP) with each switch and associating bridge domains, which need to be extended across the l3 fabric, with the VTEP. BGP-EVPN advertises the IP address associated with the VTEP to all the other switches within the fabric. In this manner, VXLAN tunnels for the extended bridge domains can be established to encapsulate the L2 traffic in that bridge domain, and send it across the VXLAN tunnel to the destination switch.

Switch/connector issues

One major issue with interconnecting multiple Layer 2 switches with Layer 1 connectors is that Layer 1 cannot protect the access network from a Layer 2 failure. In the OSI model, higher layers in the stack can solve lower-layer issues by placing mechanisms that limit the propagation of lower-layer problems. This is why an SD-WAN, with its application aware routing, can solve issues with problems in WAN/ISP connectivity and route applications over L3 networks that are not experiencing problems. There are vendors that attempt to solve the Layer 2 network issues with Layer 1 solutions (no need for routing as every switch/device is one hop from all other devices); however, in these solutions, the solution still relies upon the spanning tree protocol to stop the broadcast loop. And, as discussed, this is prone to hardware or software failure, which might result in the access network being completely isolated.

What to look for in a modern SDN access solution

A modern on-premises zero-trust LAN solution consists of leaf switches that provide L2 access to hosts and are interconnected to other switches via Layer 3 links. The switches run the OSPF protocol to establish the underlay fabric. BGP-EVPN is run as an overlay protocol to establish the VXLAN tunnels for the L2 extension between switches. Not all of the switches in the design should need to be controlled by the on-premises zero-trust access provider. Since OSPF is utilized to create the underlay, the upstream switches (spines or aggregation switches) only need to be able to run OSPF for the VXLAN overlay to work. This solution utilizes standard BGP-EVPN so it can integrate with any switches that utilize BGP-EVPN and VXLAN to extend the L2 bridge domains.

The ideal solution provides for an automated workflow to create a spine-leaf architecture that would replace either a chassis or a switch stack. This workflow automates the configuration of the switches which includes creation of the VTEPs (VXLAN Tunnel Endpoint), the association of the bridge domains with the VTEP, the creation of the L3 underlay and the BGP-EVPN configuration.

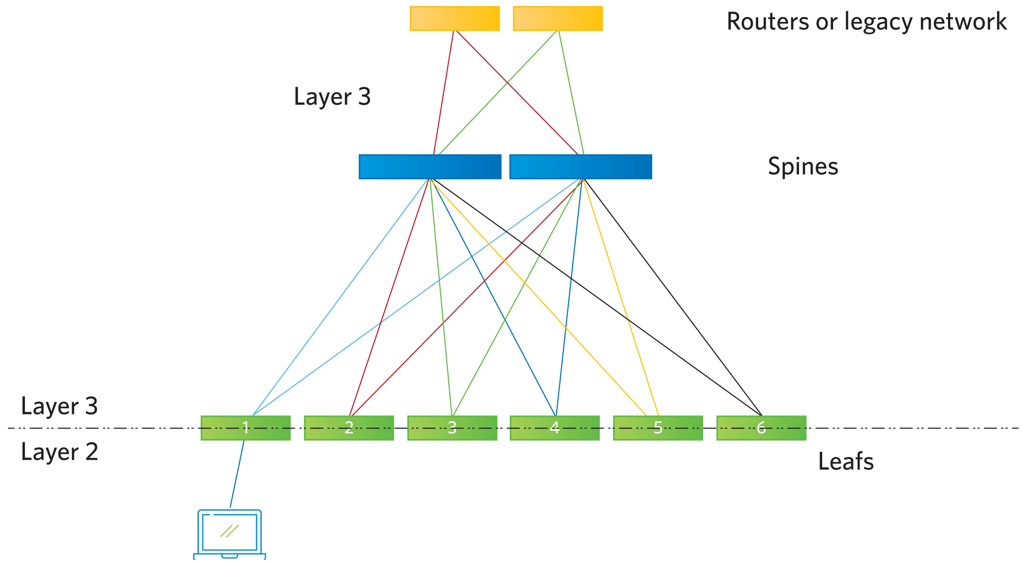

Figure 6 - Spine-leaf (L2-L3 boundary at access edge)

This solution can be instituted in a greenfield environment where all new switches are BGP-EVPN and VXLAN extended (see Figure 6), or it can be in a brownfield environment where the chassis replacement is L2 towards the hosts and L2 towards the rest of the access network, but BGP-EVPN and VXLAN between the replacement set of switches (see Figure 7).

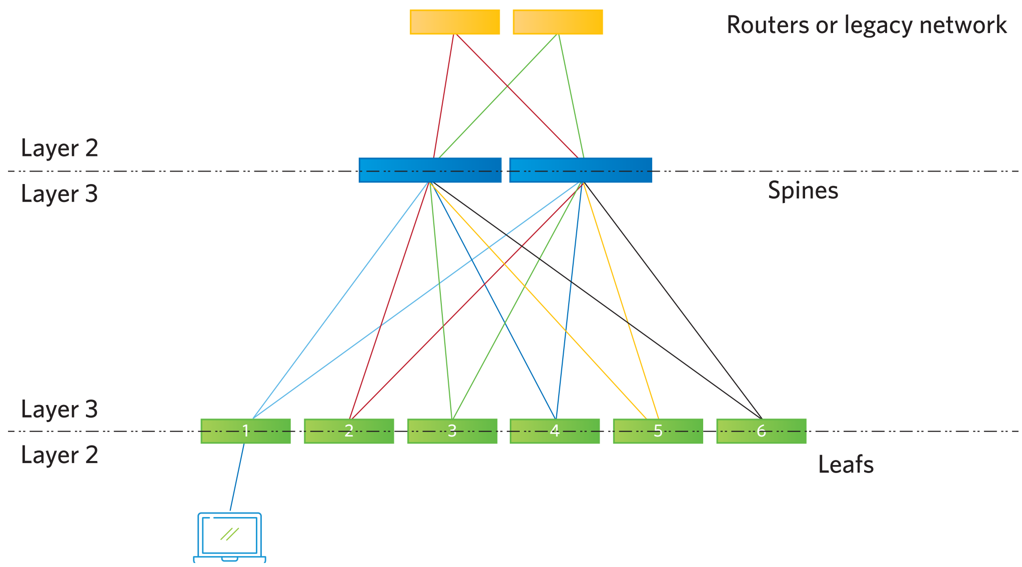

Figure 7 - Spine-leaf (L2-L3 boundary at router)

Even if a vendor's solution is a spine-leaf architecture, the solution should not necessarily be yet another IP routing-based solution. Ideally the vendor has taken the concepts of SD-WAN and applied them to the local access network, with policy-based routing to determine the best path across the VXLAN fabric for a given application. Just as in SD-WAN, the enterprise should be in a position to define policies to determine which paths are acceptable for a given application. This solves the Layer 2 forwarding issue where only a single path is selected for forwarding in a given bridge domain. Now all available paths can be utilized by a given application.

About Versa Networks

Versa Networks, the leader in single-vendor Unified SASE platforms, delivers AI/ML-powered SSE and SD-WAN solutions. The platform provides networking and security with true multitenancy, and sophisticated analytics via the cloud, on-premises, or as a blended combination of both to meet SASE requirements for small to extremely large enterprises and service providers.

Thousands of customers globally with hundreds of thousands of sites and millions of users trust Versa with their mission critical networks and security. Versa Networks is privately held and funded by Sequoia Capital, Mayfield, Artis Ventures, Verizon Ventures, Comcast Ventures, BlackRock Inc., Liberty Global Ventures, Princeville Capital, RPS Ventures and Triangle Peak Partners. For more information, visit https://www.versa-networks.com or follow Versa Networks on X (Twitter) @versanetworks.

Versa Networks, Inc, 2550 Great America Way, Suite 350, Santa Clara, CA 95054
+1 408.385.7660 | info@versa-networks.com | www.versa-networks.com