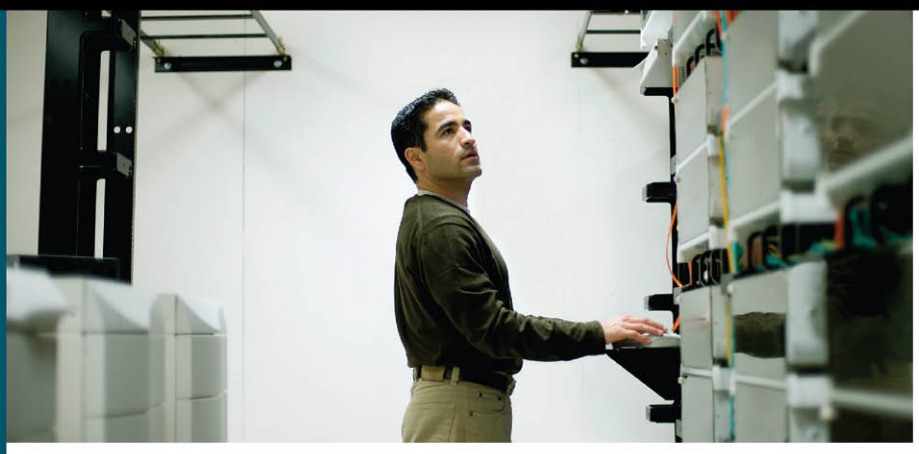

Building Data Centers with VXLAN BGP EVPN

A Cisco NX-OS Perspective

Lukas Krattiger, CCIE No. 21921
Shyam Kapadia
David Jansen, CCIE No. 5952

ciscopress.com

Chapter 4

The Underlay

In this chapter, the following topics will be covered:

- The underlay associated with the BGP EVPN VXLAN fabric
- IP address allocation options and MTU considerations for the underlay
- Underlay unicast and multicast routing options

Network virtualization overlays, including Virtual Extensible LAN (VXLAN), require a network over which the overlay-encapsulated traffic can be transported. Several considerations must be made in this regard. With VXLAN being a MAC in IP/UDP overlay, the transport network needs to carry IP traffic from VXLAN Tunnel Endpoints (VTEPs) in an optimal manner. Recall that every overlay adds additional headers on top of the original packet/frame. The transport network, termed the *underlay*, needs to account for the additional bytes incurred by the overlay headers. In the case of VXLAN, typically, the transport network needs to be provisioned for an additional 50 bytes in the maximum transmission unit (MTU). Likewise, the transport network must match the resiliency and convergence requirements of the overlay as well. It is important to consider scalability, resiliency, convergence, and capacity when considering data center fabric designs. With overlay-based data center fabrics, the importance of the underlay cannot be understated. Typically, the overlay's performance is only as good as the underlay transport that carries it. Even for troubleshooting, debugging, and convergence of the overlay traffic, the underlay is the critical piece.

When designing and building the underlay for the fabric, IP address assignment is an important aspect of this process. The IP address assignment is a prerequisite for any kind of routing protocol that provides reachability between the various devices in the fabric. Multiple options are available in assigning these IP addresses, including, among others, a traditional point-to-point (P2P) method or an address-saving option that uses an

unnumbered IP addressing scheme. Likewise, it is important to be aware of requirements in addressing the interfaces of the VTEP (also called the Network Virtualization Edge [NVE] interface), the loopback interface over which the multiprotocol BGP session is established, and eventually the multicast rendezvous point, when using a multicast based underlay.

The underlay might be required to not only transport unicast traffic but also provide multicast routing in order to handle broadcast, unknown unicast, and multicast (BUM) traffic in the overlay. All these requirements need to be considered when designing and creating the underlay network for a BGP EVPN VXLAN–based data center fabric.

Underlay Considerations

When building the data center fabric, some upfront considerations regarding the underlay take priority. The first involves defining the anticipated topology to be used for the underlay. Today's data centers have requirements for large amounts of north–south as well as east–west traffic patterns, so the topology must be able to accommodate these different kinds of traffic streams and communication profiles.

North–south data traffic, as illustrated in Figure 4-1, is the traditional communication profile for moving data in and out of the data center. In essence, this profile is utilized if a user in the campus or on the Internet wishes to access data from within a data center. Common examples of north–south traffic are traffic associated with classic web browsing or traffic related to email activities. In each of these cases, an end user has an application on a PC to access data hosted within a data center. This type of north–south traffic is important to efficiently move data between data centers and toward the end user.

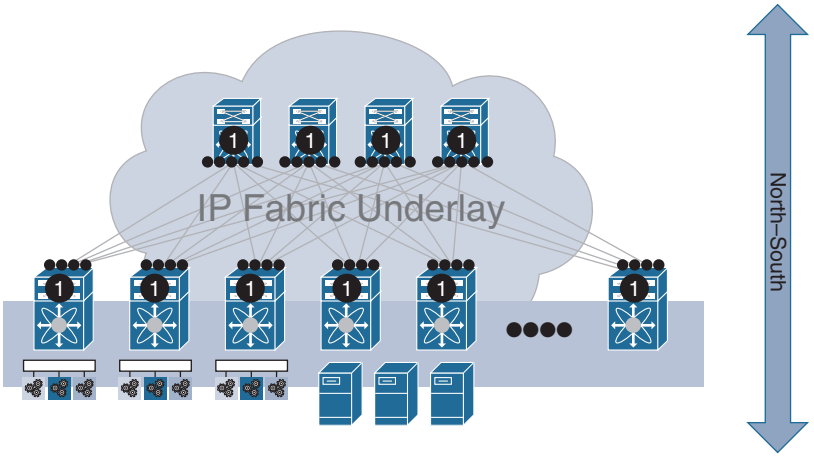

Figure 4-1 *North–South Traffic*

East–west traffic, on the other hand, reflects a slightly different communication profile because this profile describes data communication between servers and/or various applications within the data center (see Figure 4-2). Typically, requests from an end user in a corporate network or in the Internet involve complex preprocessing activities on the underlying data. As a means to demonstrate this need for preprocessing, one example of east–west traffic involves access from a web server (via an app server) to the database. Dynamic rendering of websites and/or business applications often uses a two- or three-tier server architecture, where one tier must communicate with one of the other tiers before delivering the requested data to the end user. When these tiers communicate with one another or need access to data storage or other data residing in the same data center, the term *east–west traffic* is used for this more lateral, or horizontal, traffic profile.

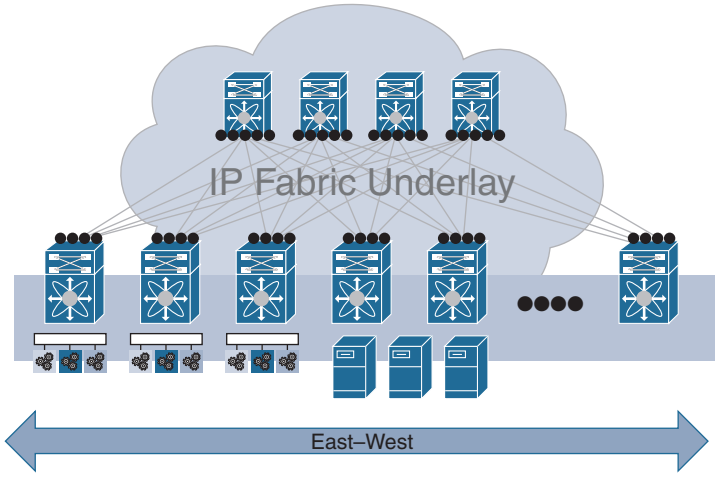

Figure 4-2 *East–West Traffic*

In the early 1950s, when telephone switchboards were manually operated, Charles Clos needed to find a more efficient way to handle call transfers. The switchboards utilized a two-stage network where calls would be transferred by using a single crossbar switch. Unfortunately, frequently calls were blocked because only one path existed, and if another transfer was occupying that path, transfers would fail. Faced with this dilemma, Clos established the best mathematical way for interconnecting two points from an ingress call to an egress one by leveraging multistage fabrics. By having an ingress stage, a middle stage, and an egress stage, calls had multiple path opportunities to reach their transfer destinations. This additional stage created a crossbar matrix of connectivity, and because the resulting networks appeared like woven fibers, the term *crossbar switch* was introduced.

Clos's work and designs, which revolutionized telephone switchboards, also found its way to network switch and data center fabric designs. Most network switches built today are based on Charles Clos's mathematical equations from the 1950s. In a given switch, the front-facing Ethernet ports (first stage) are each interconnected via a network fabric (second stage). As a result of this topology, every front-facing Ethernet port must travel the same distance to every other front-facing Ethernet port (equidistant), thereby ensuring predictable and consistent latency.[1] Figure 4-3 shows the internal architecture of a typical multistage modular network switch.

How does this apply to a data center network or a data center fabric? In applying the Clos fabric concept from the network switch to data center fabric topology, a front-facing Ethernet port is replaced with a top-of-rack (ToR) switch, also known as a *leaf*. The leaf is responsible for providing connectivity to and from servers as well as making forwarding decisions, based on bridging or routing lookups, within the data center fabric. Hundreds of leafs potentially exist, and interconnections among them are needed. A leaf connects to a second stage in the data center fabric, which is composed of a set of spine switches. Depending on the requirements for scale and bandwidth for the fabric, the number of spines can vary between 2 and 32, which is good enough for practical deployments.

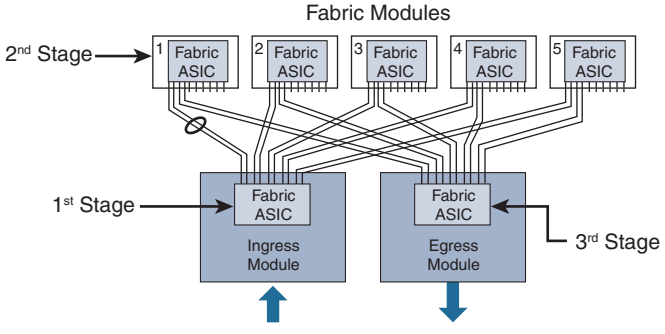

Figure 4-3 *Multistage Network Switch*

Within the data center fabric, a spine is connected to every leaf, thereby providing *N* paths from every leaf to every other leaf, where *N* is the number of spines (see Figure 4-4). The spines in turn may connect to another spine layer, termed the super-spine layer, thereby allowing construction of N-stage spine leaf fabrics. Because the spine itself is a connectivity "backbone" in the overlay-based data center fabric, it has no visibility into the user traffic itself because the overlay encapsulation occurs at the leaf. The spine simply ensures that VTEP-to-VTEP communication is done efficiently between different leafs by using the underlay IP network's equal-cost multipath (ECMP) capability provided by the routing protocol. In comparing this to Clos fabric topology concepts, multiple paths are available via the spines, with each offering the same distance and latency between any two endpoints in the data center fabric.

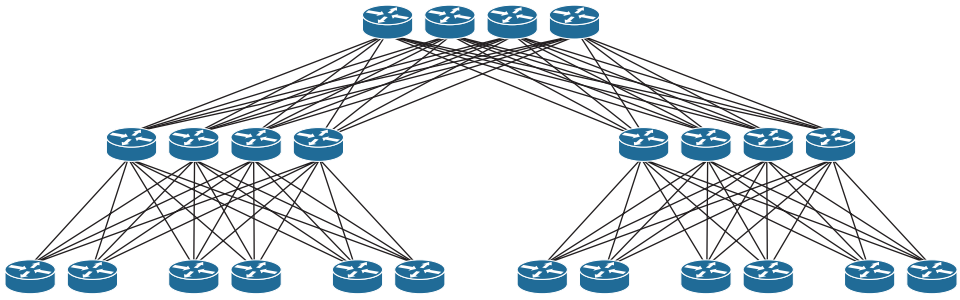

Figure 4-4 *Multistage Clos Network*

MTU Considerations

Recall that VXLAN, as a network virtualization overlay technology, places some additional overhead on the original data frames. This overhead manifests in the form of additional identifier information required for the functioning of VXLAN itself. This helps solve some of the existing deficiencies present in traditional network transport technologies (such as Ethernet). However, due to the overhead, additional considerations need to be taken into account specific to the underlay network design.

As a general rule, fragmentation (splitting of a frame or data packet because it is too large for a transport network) should be avoided. Fragmentation and reassembly put additional burden on the switch resources as well as the server resources, which results in transport inefficiency. Computers, servers, PCs, and other network hardware equipped with Ethernet network interface cards (NICs) have a standard maximum transmission unit (MTU) of 1500 bytes.[2] The total size of the Ethernet frame is 1518 (or 1522, with the additional optional 802.1Q tag), with 6 bytes each for the source and DMAC addresses, 2 bytes for the Ethertype, and 4 bytes for the frame check sequence (FCS). This means that a computer can send a payload of 1500 bytes or less, which includes all header information, from Layer 3 and above. Likewise, a server could potentially send unfragmented data up to an MTU of 1500 bytes as well through its default configuration. However, if VXLAN is employed between the network switches to which the servers are attached, the MTU is reduced to 1450 bytes through its default configuration. This is because VXLAN adds either 50 or 54 bytes of overhead (see Figure 4-5) as part of identifier information.

Employing VXLAN incurs an overhead consisting of a 14-byte outer MAC header, a 20-byte outer IP header, an 8-byte UDP header, and an 8-byte VXLAN header. The 50 or 54 bytes of overhead introduced by VXLAN must be added to the MTU considerations of Ethernet itself. The optional additional 4 bytes that might push the overhead from 50 bytes to 54 bytes would come from preserving the IEEE 802.1Q tag in the original Ethernet frame. In most instances, the 802.1Q tag is mapped to the VNI and stripped off before the VXLAN encapsulation is added to the original frame. However, certain use cases related to Q-in-Q[3] or Q-in-VNI[4] exist (for example, nested hypervisor, Layer 2 tunneling, etc.), that would require the original IEEE 802.1Q tag to be preserved. In any case, these additional bytes of the VXLAN overhead need to be considered as part of the underlay design.

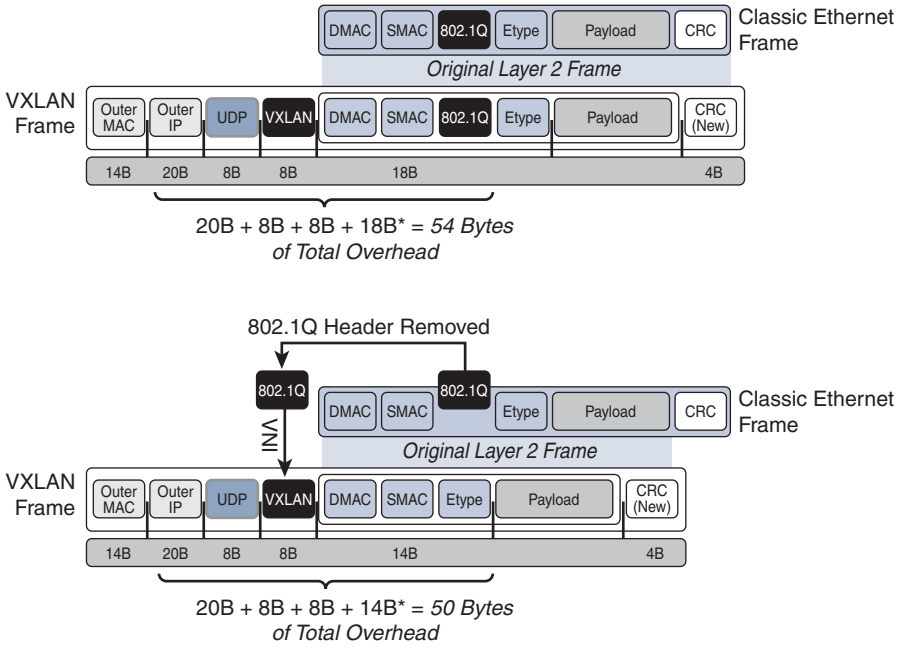

Figure 4-5 *VXLAN Frame Format, Including Byte Count (50/54 Byte Overhead for Comparison)*

In data center networks, efficiency is essential, and therefore, the MTU on the server side is often increased to accommodate jumbo frames (data units that are larger than the traditional 1500 bytes typical of Ethernet frames). Jumbo frames coming from the server side can typically be up to a maximum of 9000 bytes, which most NICs and/or virtual switches offer. With a need to now support MTUs of 9000 bytes from the server side when using jumbo frames, and with the additional bytes from the VXLAN encapsulation, a value around 9050 or 9054 bytes needs to be allowed in order to avoid fragmentation.

Fortunately, most of the network switches from Cisco and other vendors provide a maximum MTU value of 9216 bytes, though some have to reduce this value by 20 to 30 bytes to accommodate a network switch internal service header. Therefore, a jumbo frame MTU of 9000 bytes from the server side can be accommodated by network switches without introducing fragmentation. This is not uncommon because the use of large transmission units is more efficient, and in a high-speed network, using them reduces the number of round trips required over the network. Because MTU configuration is performed at many locales, and because this can potentially affect the overlay during subsequent reconfigurations, an initial consideration regarding the MTU size of the network is critical. The jumbo frame MTU needs to be configured consistently across all points in the network, including the host-facing server interfaces. As an additional safeguard, certain network protocols check to assess whether a neighboring device is using the same MTU