

2. VXLAN

As its name indicates, Virtual eXtensible Local Area Network (VXLAN) is designed to provide the same Ethernet Layer 2 network services as VLAN does today, but with greater extensibility and flexibility. Compared to VLAN, VXLAN offers the following benefits:

Flexible placement of multitenant segments throughout the data center:

It provides a solution to extend Layer 2 segments over the underlying shared network infrastructure so that tenant workload can be placed across physical pods in the data center. Higher scalability to address more Layer 2 segments: VLANs use a 12-bit VLAN ID to address Layer 2 segments, which results in limiting scalability of only 4094 VLANs. VXLAN uses a 24-bit segment ID known as the VXLAN Network Identifier (VNID), which enables up to 16 million VXLAN segments to coexist in the same administrative domain.

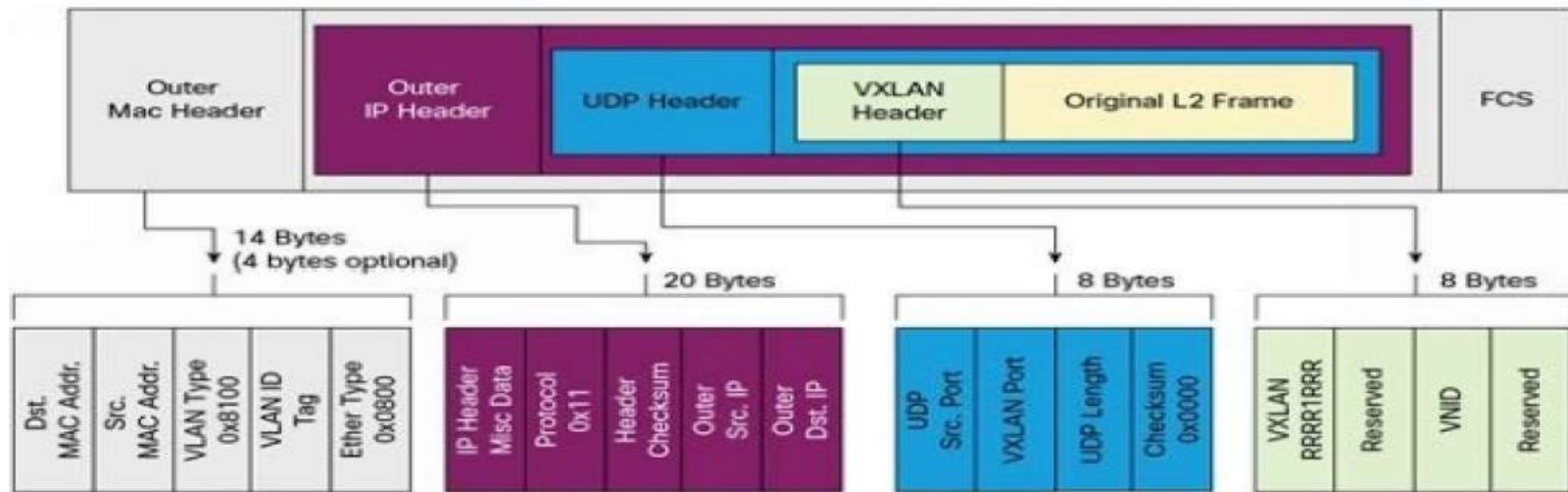
Better utilization of available network paths in the underlying infrastructure:

VLAN uses the Spanning Tree Protocol (STP) for loop prevention, which ends up with not using half of the network links by blocking redundant paths. In contrast, VXLAN packets are transferred through the underlying network based on its Layer 3 header and can take complete advantage of Layer 3 routing, equalcost multipath (ECMP) routing, and link aggregation protocols to use all available paths.

2.1. VXLAN Format, Traffic Flow & ECMP

VXLAN encapsulation adds 50 bytes to the original L2 frame by adding 4 different headers:

- VXLAN header
- UDP header
- Outer IP header
- Outer MAC header



Encapsulated VXLAN packets are forwarded between VTEPs (VXLAN Tunnel End Point) based on the native forwarding decisions of the transport network. And because the VXLAN was originally created to be bounded inside the data center where the underlay transport networks are designed and deployed with multiple redundant paths and take advantage of various multipath load-sharing technologies to distribute traffic loads on all available paths. It is desirable to share the load of the VXLAN traffic in the same fashion in the transport network.

A typical VXLAN transport network is an IP-routing network that uses the standard IP ECMP to balance the traffic load among multiple best paths. To avoid out-of-sequence packet forwarding, flow-based ECMP is commonly deployed. An ECMP flow is defined by the source and destination IP addresses and optionally the source and destination TCP or UDP ports in the IP packet header.

Because all the VXLAN packet flows between a pair of VTEPs have the same outer source and destination IP addresses, and all VTEP devices must use one identical destination UDP port that can be either the Internet Assigned Numbers Authority (IANA)-allocated UDP port 4789 (or a customer-configured port), the only variable element in the ECMP flow definition that can differentiate VXLAN flows from the transport network standpoint is **the source UDP port**. A similar situation for Link Aggregation Control Protocol (LACP) hashing occurs if the resolved egress interface based on the routing and ECMP decision is a LACP port channel. The LACP uses the VXLAN outer-packet header for link load-share hashing, which results in the source UDP port being the only element that can uniquely identify a VXLAN flow.

2.2. VXLAN Modes of Operation

From the various industry implementations of VXLAN, we can categorize the VXLAN modes of operation into two main categories:

Control-Plane-Less-VXLAN
Control-Plane-VXLAN

The differences are mainly in the underlay transport network multicast capability, how to deal with BUM (Broadcast, Unknown Unicast & Multicast) traffic as well as the method of discovery and distribution of MAC addresses.

2.3. Control-Plane-Less-VXLAN

For the Control-Plane-less mode of operation of VXLAN, we have two main sub-modes:

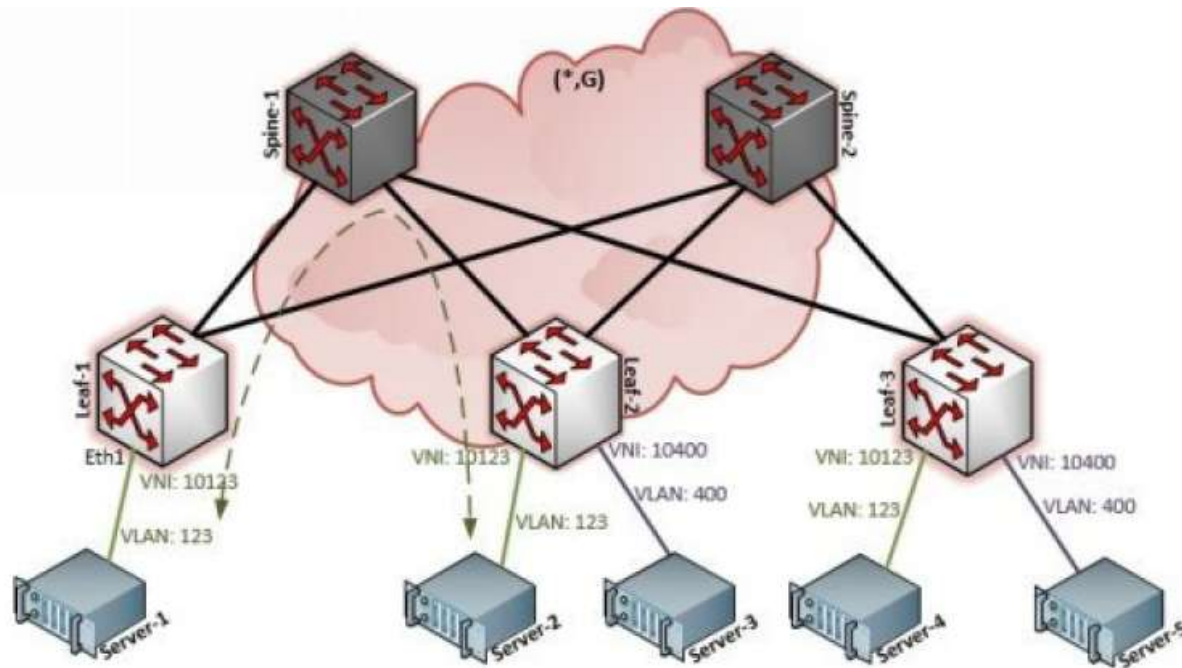
Control-Plane-Less-Multicast-VXLAN
Control-Plane-Less-Unicast-VXLAN

2.3.1. Control-Plane-Less-Multicast-VXLAN

As its name implies; there is NO control or signaling established prior to the VXLAN operation.

This mode is according to the original VXLAN draft as per RFC7348.

This mode requires the underlay transport network to fully support the IP Multicast & every VTEP node to join the proper Multicast domain.



In this mode; the BUM (Broadcast, Unknown Unicast & Multicast) traffic is always carried over Multicast. It's all about the 'Data-Plane' learning (or flow-based learning) that is based on the 'Flood & learn' technique; where the remote VTEPs would know about a MAC address because of the conversational MAC address learning approach:

The destination (receiving) VTEP learns the inner Source MAC of any received VXLAN IP packet (for example a Broadcasted ARP request message carried over Multicast). The source MAC address is then mapped to the Source (Originating) VTEP that originated the VXLAN packet.

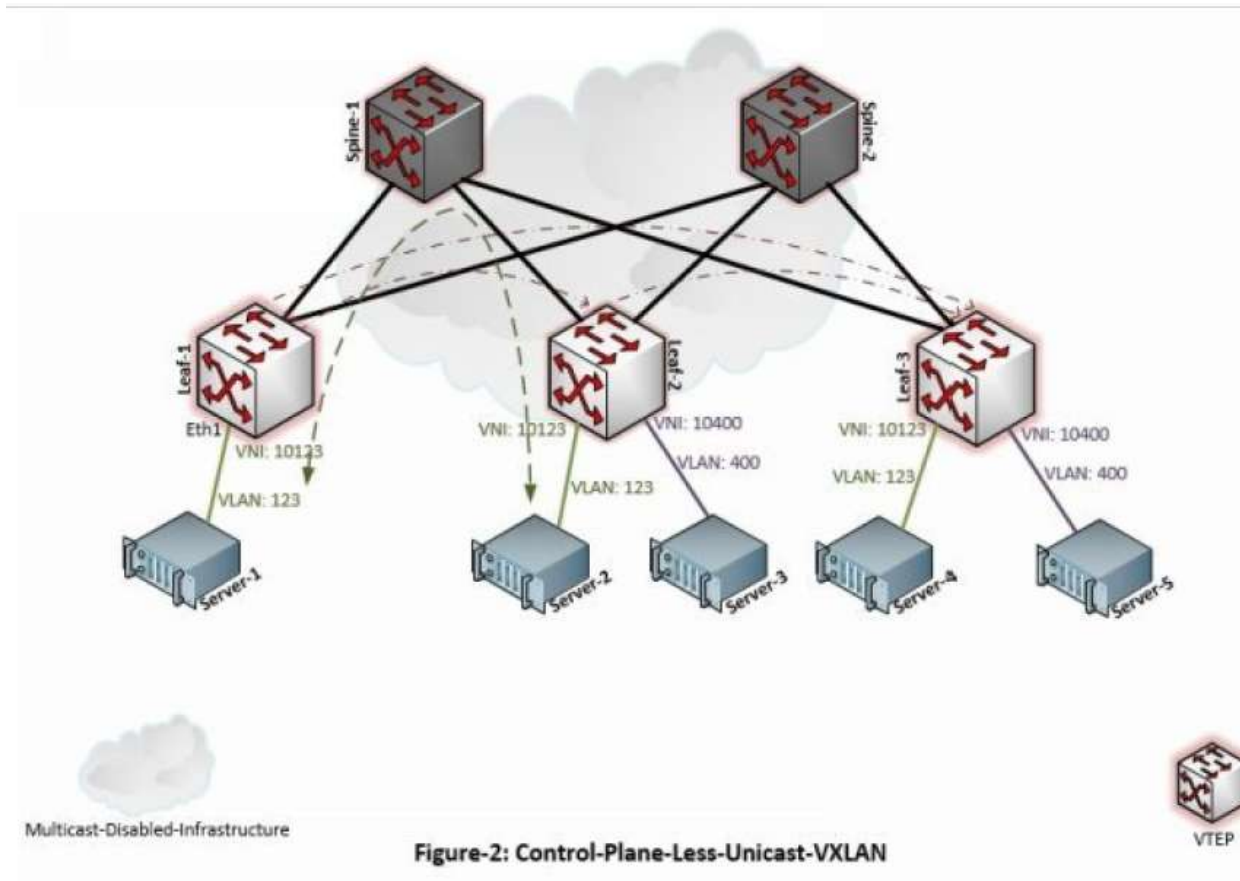
The Originating VTEP will learn the remote MAC address to VTEP mapping once it receives the VXLAN encapsulated 'Unicast' ARP reply message from the receiving VTEP.

All subsequent traffic to a known MAC address will be Unicast IP encapsulated VXLAN.

2.3.2. Control-Plane-Less-Unicast-VXLAN

Exactly like the Control-Plane-Less-Multicast-VXLAN; there is NO control or signaling established prior to the VXLAN operation, instead; a list of all available & participating VTEPs are configured on each VTEP per supported VXLAN.

In this mode; the underlay transport network doesn't need to support IP Multicast.



For the BUM traffic; instead of being 'Multicast' over the underlay transport network as in the previous mode of operation, the 'head-end replication' is used here where the originating VTEP has to replicate the VXLAN packet & sends a copy to every other VTEP participating in this same VXLAN. The list of VTEPs must be configured (changed & updated) manually on each & every VTEP in the VXLAN domain.

The 'Data-Plane' learning technique is also here in this mode of operation.

2.4. Control-Plane-VXLAN

In this mode; there is no need for the IP Multicast in the underlay transport network; for any traffic that would require to be sent to all VTEPs (like Broadcast or Multicast); the head-end replication will be used instead (as in the previous mode)

Dealing with the Unknown Unicast traffic is what really differentiates this mode of operation from the previous modes. In this mode, a 'Control-Plane' does exist to distribute the MAC-to-VTEP mapping entries between the different VTEPs, hence no need for any data-plane learning technique (like flood & learn).

This control plane piece could be a **Controller** (like VMware NSX, Midokura, Nuage, Openstack...), a signaling protocol (like MP-BGP in the **EVPN**-based VXLAN)

2.4.1. Controller-based-VXLAN

In the Controller-based VXLAN service, Data-Plane (flow-based) learning is optional or even not needed; the controller synchronizes all the MAC addresses as soon as the different switches learn them locally from their local ports.

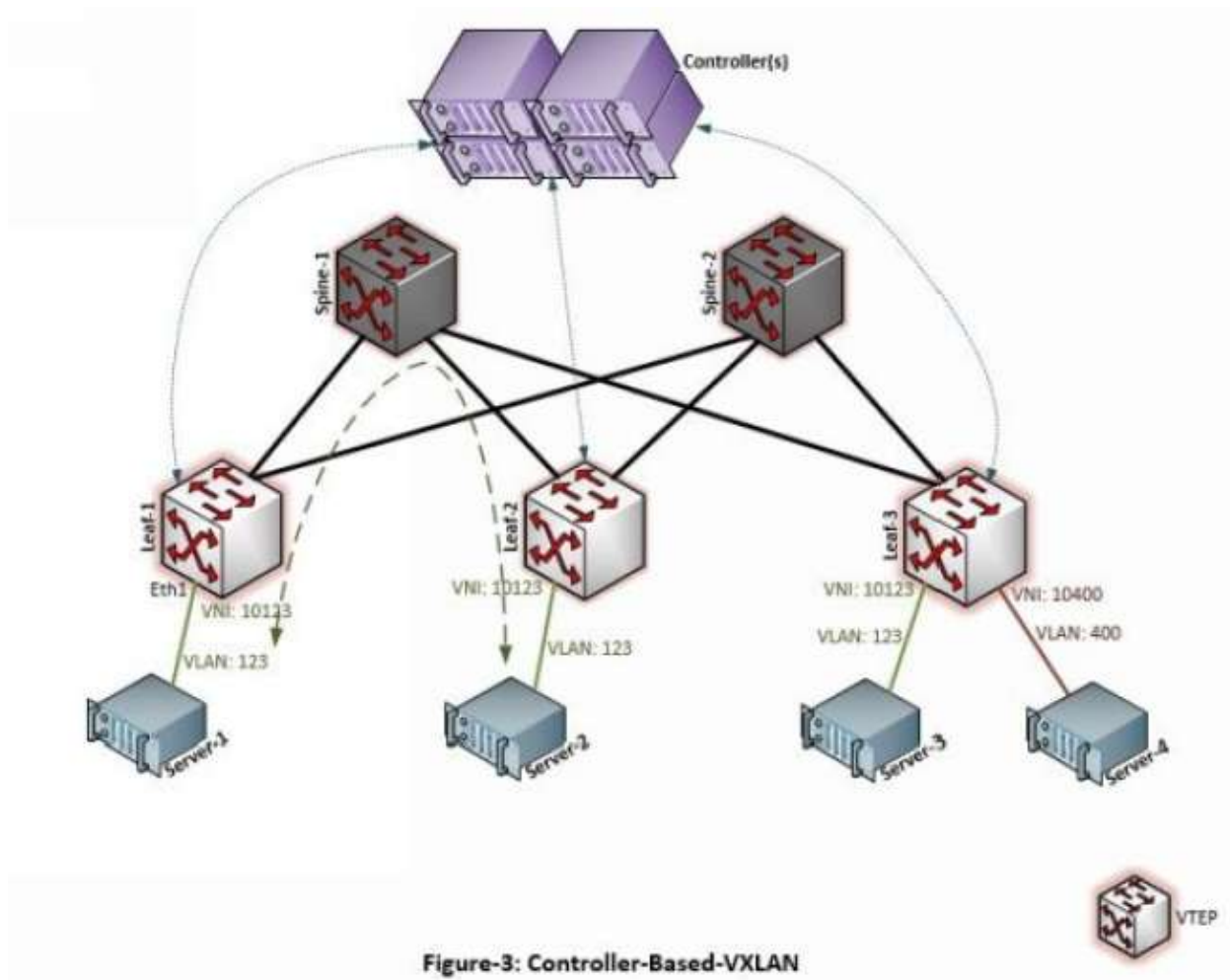


Figure-3: Controller-Based-VXLAN

For example in Figure-3, Leaf-1 learns the MAC address of Server-1 from its local port Eth1. This information is automatically and immediately synchronized with the controller that in turn pushes that info to Leaf-2, Leaf-3 & any other VTEP in the same VXLAN domain.

This VXLAN operation depends on the distribution of all learnt MAC addresses from the different VTEPs via the controller that 'pushes' to all VTEPs a complete (and always updated) list of MAC-to-VTEP mapping entries.

Because of that, for this mode; there would be no Unknown Unicast as the list of all communicating MACs is always present and updated on each VTEP, but in case of an unknown MAC (maybe for a destination outside the local VXLAN domain) & depending on the configuration; the local VTEP can direct it via the default entry towards the VXLAN gateway. For the other Broadcast & Multicast traffic; the head-end replication is always the solution.

2.4.2. EVPN-VXLAN

In the EVPN-VXLAN; each VTEP is now a PE (Provider Edge) node & will learn the local MAC addresses associated to its VXLANs from its local ports as usual. Using MP-BGP EVPN address family; these entries will be propagated between the different PEs (ideally through a set of MP-BGP RR 'Route Reflectors')

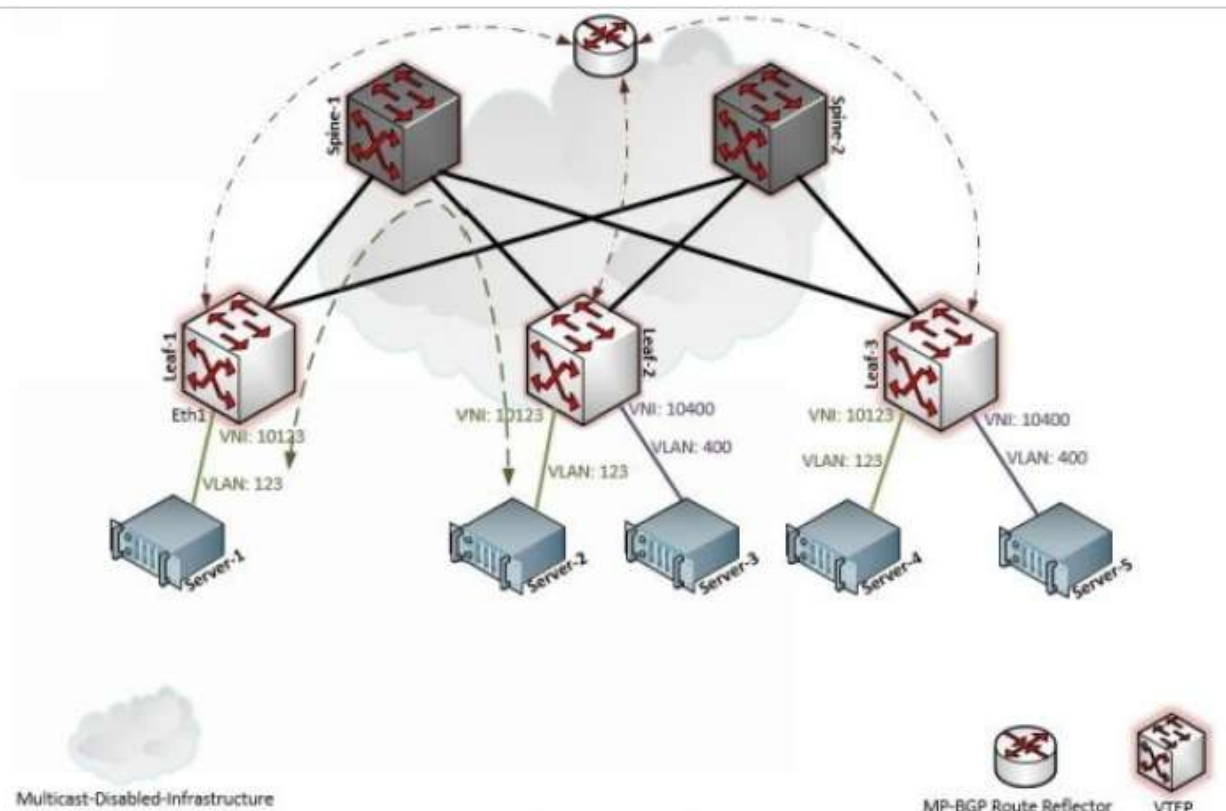


Figure-4: EVPN-VXLAN

As in the Controller-based VXLAN; there would be no Unknown Unicast as the list of all communicating MACs is on each VTEP, but in case of an unknown MAC; the local VTEP can direct the traffic via the default entry towards the VXLAN gateway. Again for the other Broadcast & Multicast traffic; the head-end replication is always the solution.