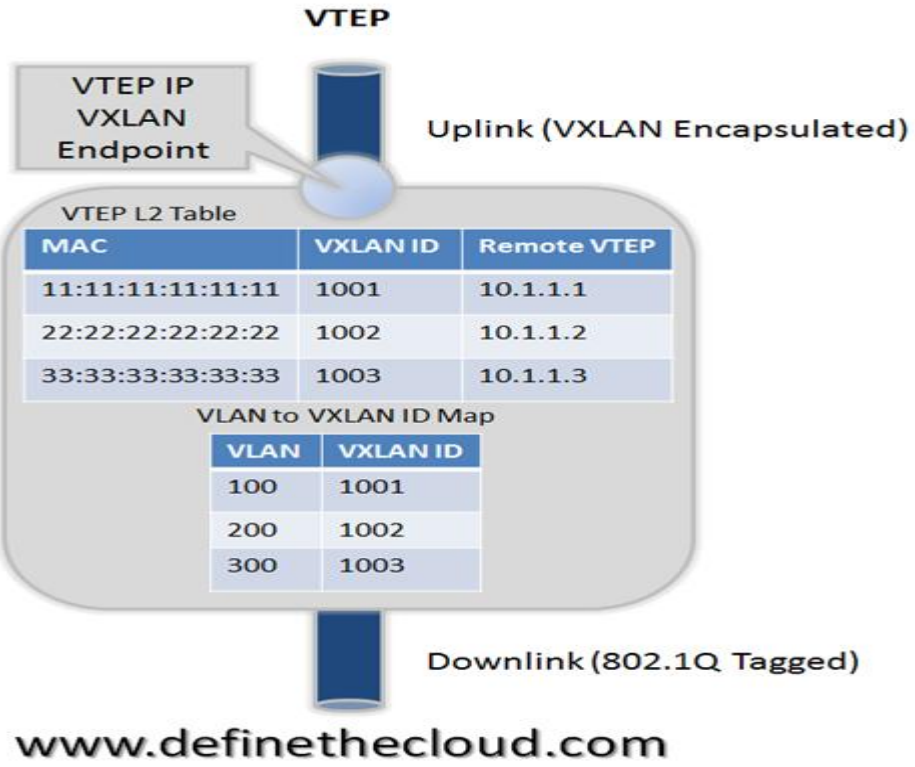
VXLAN DEEP DIVE – PART I

I've been spending my free time digging into network virtualization and network overlays. By far the most popular virtualization technique in the data center is VXLAN. This has as much to do with Cisco and VMware backing the technology as the tech itself. That being said VXLAN is targeted specifically at the data center and is one of many similar solutions such as: NVGRE and STT.) VXLAN's goal is allowing dynamic large scale isolated virtual L2 networks to be created for virtualized and multi-tenant environments. It does this by encapsulating frames in VXLAN packets. The standard for VXLAN is under the scope of the IETF NVO3 working group.

The VXLAN encapsulation method is IP based and provides for a virtual L2 network. With VXLAN the full Ethernet Frame (with the exception of the Frame Check Sequence: FCS) is carried as the payload of a UDP packet. VXLAN utilizes a 24-bit VXLAN header, shown in the diagram, to identify virtual networks. This header provides for up to 16 million virtual L2 networks.

Frame encapsulation is done by an entity known as a VXLAN Tunnel Endpoint (VTEP.) A VTEP has two logical interfaces: an uplink and a downlink. The uplink is responsible for receiving VXLAN frames and acts as a tunnel endpoint with an IP address used for routing VXLAN encapsulated frames. These IP addresses are infrastructure addresses and are separate from the tenant IP addressing for the nodes using the VXLAN fabric. VTEP functionality can be implemented in software such as a virtual switch or in the form a physical switch. VXLAN frames are sent to the IP address assigned to the destination VTEP; this IP is placed in the Outer IP DA. The IP of the VTEP sending the frame resides in the Outer IP SA. Packets received on the uplink are mapped from the VXLAN ID to a VLAN and the Ethernet frame payload is sent as an 802.1Q Ethernet frame on the downlink. During this process the inner MAC SA and VXLAN ID is learned in a local table. Packets received on the downlink are mapped to a VXLAN ID using the VLAN of the frame. A lookup is then performed within the VTEP L2 table using the VXLAN ID and destination MAC; this lookup provides the IP address of the destination VTEP. The frame is then encapsulated and sent out the uplink interface.

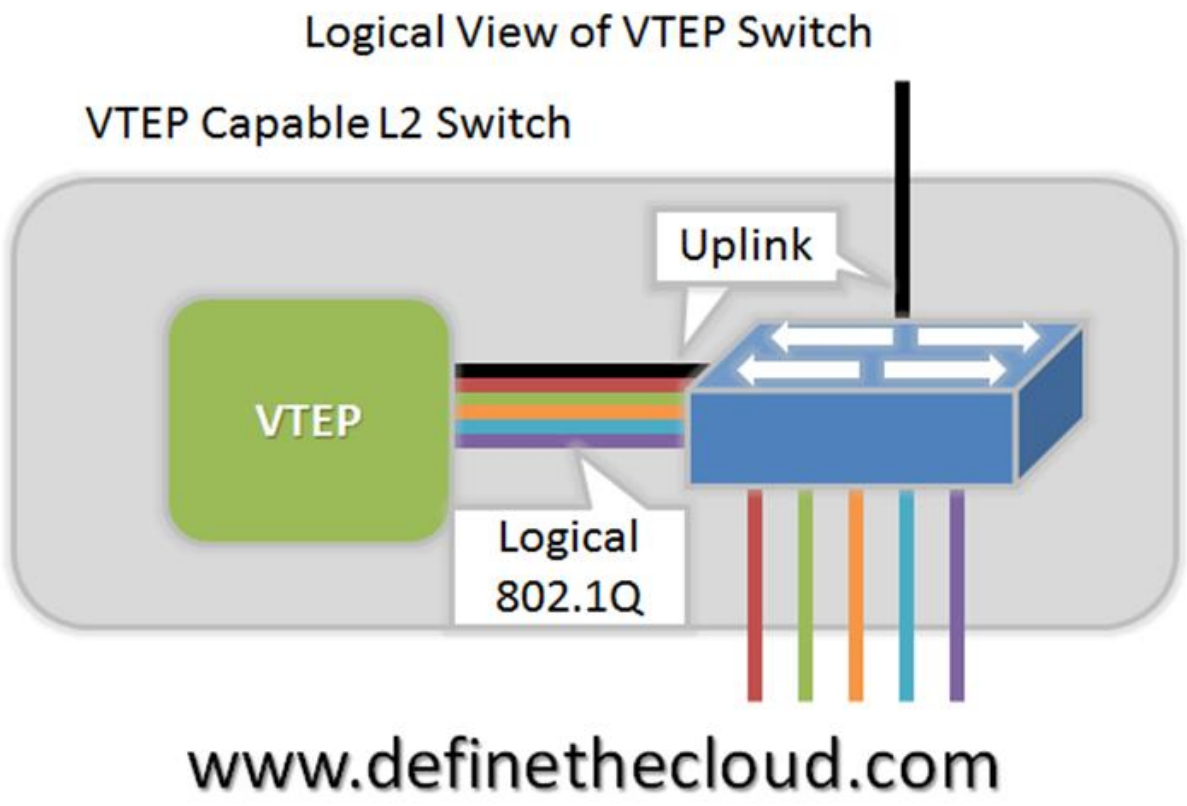


Using the diagram above for reference a frame entering the downlink on VLAN 100 with a destination MAC of 11:11:11:11:11:11 will be encapsulated in a VXLAN packet with an outer destination address of 10.1.1.1. The outer source address will be the IP of this VTEP (not shown) and the VXLAN ID will be 1001. In a traditional L2 switch a behavior known as flood and learn is used for unknown destinations (i.e. a MAC not stored in the MAC table. This means that if there is a miss when looking up the MAC the frame is flooded out all ports except the one on which it was received. When a response is sent the MAC is then learned and written to the table. The next frame for the same MAC will not incur a miss

because the table will reflect the port it exists on. VXLAN preserves this behavior over an IP network using IP multicast groups.

Each VXLAN ID has an assigned IP multicast group to use for traffic flooding (the same multicast group can be shared across VXLAN IDs.) When a frame is received on the downlink bound for an unknown destination it is encapsulated using the IP of the assigned multicast group as the Outer DA; it's then sent out the uplink. Any VTEP with nodes on that VXLAN ID will have joined the multicast group and therefore receive the frame. This maintains the traditional Ethernet flood and learn behavior.

VTEPs are designed to be implemented as a logical device on an L2 switch. The L2 switch connects to the VTEP via a logical 802.1Q VLAN trunk. This trunk contains an VXLAN infrastructure VLAN in addition to the production VLANs. The infrastructure VLAN is used to carry VXLAN encapsulated traffic to the VXLAN fabric. The only member interfaces of this VLAN will be VTEP's logical connection to the bridge itself and the uplink to the VXLAN fabric. This interface is the 'uplink' described above, while the logical 802.1Q trunk is the downlink.



Source: http://www.definethecloud.net/vxlan-deep-dive/