VXLAN

Extending Networking to Fit the Cloud

KAMAU WANGŨHŨ

Kamau Wangũhgũ is a Consulting Architect at VMware and a member of the Global Technical Service, Center of Excellence group. Kamau's focus at VMware is in Cloud Architecture and virtual networking futures. Kamau is also a VMware Certified Design Expert (VCDX #003) and was involved in a number of VCDX defenses at the beginning of the program. Kamau has also presented and been involved in all VMworlds, some Technical Solution eXchanges, and Partner Exchanges around the world. Previous to working for VMware, Kamau performed infrastructure architecture work for high transaction systems. Follow Kamau on his blog at http://www.borgcube.com/blogs, or on twitter at http://twitter.com/Borgcube. Kamau@BORGcube.com

When assembling a cloud infrastructure, you want the flexibility of locating your physical resources anywhere in your datacenter. Networks in a datacenter, on the other hand, may constrain virtual machine (VM) location due to the need for shared Layer 2 connectivity among correspondent virtual machines. VXLAN is an overlay network that overcomes this constraint, allowing you to locate your VMs anywhere in the datacenter without the network being the limiting factor. In this article, I will explain what VXLAN is and how the technology is implemented.

Scalable Networks

One of the biggest infrastructure obstacles faced in building out cloud environments is providing sufficient numbers of isolated networks for tenants. In addition, there is a need to decouple the scaling of these networks from the constraints imposed by the physical network topology and datacenter network architecture. These problems are in no way constrained to cloud-based networking, but are exacerbated by the need for large numbers of on-demand and scalable networks consumed in cloud environments. A key requirement for cloud networking is the need to place VMs anywhere in the datacenter for scaling purposes. This needs to be done while maintaining Layer 2 adjacency between these VMs in order for them to use the same IP address space. However, this requirement implies that the physical networking infrastructure between these communicating VMs should be a Layer 2 (broadcast) network.

Enter VXLAN

VXLAN, in a nutshell, is an overlay Layer 2 over Layer 3 technology that provides physical infrastructure-independent networking to VMs. The motivation behind VXLAN overlay networks was to address the scale and isolation needs encountered in cloud-based infrastructures without imposing any requirements on the physical network infrastructure. VXLAN addresses the flexibility and location-independence requirements of the virtual infrastructure without requiring any networking hardware changes.

In this article, I will concentrate on virtual infrastructure and cloud implementations, as that is where most of the benefits will come into play today. Figure 1 is a key to the icons used in this article.

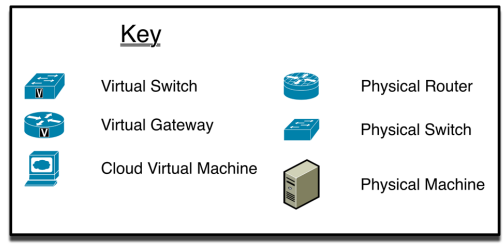

Figure 1: Key to icons used in this article

What Is VXLAN

As an overlay Layer 3 network, VXLAN provides physical infrastructure-independent networking to VMs. Each overlay network is known as a VXLAN Segment and is identified with a 24-bit VXLAN Network Identifier (VNI). This makes it possible to overlay over 16 million networks on a single VXLAN fabric. This is illustrated in Figure 2 as a networking fabric overlaid on virtual switches across multiple server clusters.

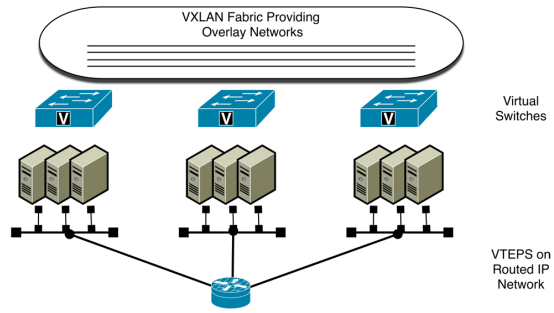

Figure 2: VXLAN fabric overlaid on multiple virtual switches associated with clusters of servers on different Layer 3 networks

A VXLAN fabric is a homogeneous namespace maintained by a single entity on which overlay networks are instantiated. Connectivity between different VXLAN fabrics will not be discussed in this article.

The VXLAN fabric uses the networking provided by the virtual switches to interconnect the VMs and the VXLAN tunnel endpoints (VTEP). VXLAN traffic is tunneled by VTEPs through the physical Layer 3 network infrastructure to other VTEPs. The VTEP is the component of the VXLAN networking stack that encapsulates and decapsulates the VXLAN tunnel traffic that is overlaid on the physical network as illustrated in Figure 3. The VTEP is usually implemented as part of the hypervisor in order to reduce latency. Each VTEP has knowledge of all the virtual machines that it handles and gleans information about virtual machines handled by other VTEPs through data plane-based learning. Here, an unknown destination MAC frame is transmitted over the multicast group associated with that VXLAN segment. When it arrives at the individual VTEPs, they learn the association between the VTEP IP address and the VNI + VM MAC address.

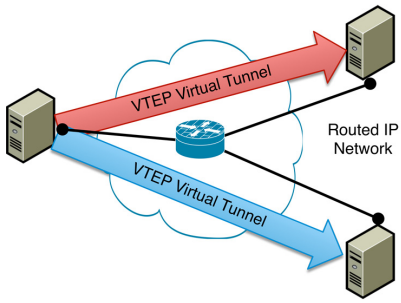

Figure 3: Hosts participating in a VXLAN exchange virtual machine traffic over logical VTEP tunnels that are established through a routed IP network.

The VTEP creates point-to-point tunnels between itself and other VTEPs for the transport of VXLAN encapsulated traffic. The encapsulated traffic can be transported over dedicated transport VLANs or it can share existing VLANs if traffic volumes allow. A VTEP does not provide access to the physical network for the virtual machines it fronts. Access to networks outside the VXLAN segments is provided through the use of a gateway. The simplest use case for a gateway is as a Layer 2 bridge between VXLAN and VLAN environments as shown in Figure 4. It is also possible for routers, or Layer 3 switches, to be VXLAN aware so that they can forward traffic at Layer 3.

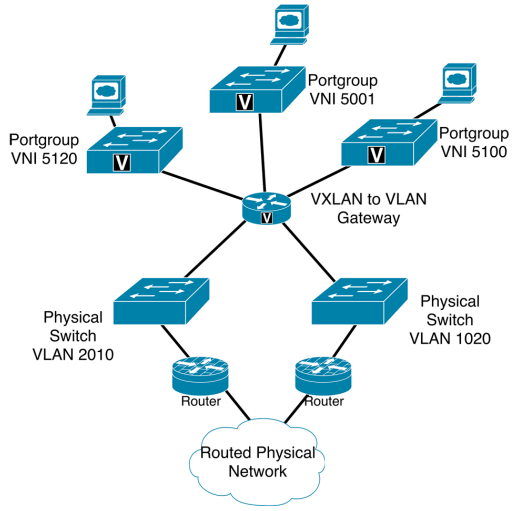

Figure 4: A VXLAN to VLAN gateway is needed to allow virtual machines on a VXLAN-backed network to communicate with nodes outside its Layer 2 network.

When a virtual machine is communicating with other virtual machines on its Layer 2 network, it is not aware in any way of VXLAN. The original Layer 2 frame is encapsulated in an outer UDP packet along with a VXLAN header as illustrated in Figure 5. The VXLAN header contains the VNI (added by the VTEP) that identifies which isolation network the packet belongs to. This UDP frame is then transported on an IP-based network like any normal IP traffic.

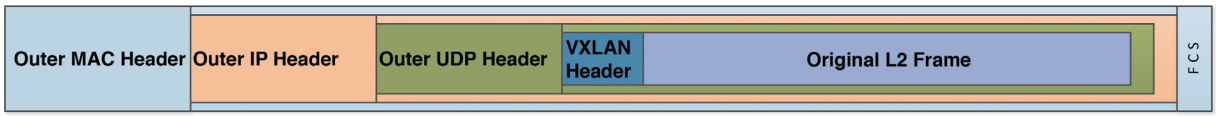

Figure 5: Virtual machines Layer 2 packet is encapsulated in a UDP packet after a VXLAN header is added to the original packet. This allows the packet to be tunneled in an IP Network.

If the destination virtual machine happens to be on the same virtual switch as the source (on the same physical host), then no VXLAN encapsulation happens, and the original Layer 2 frame is passed to the destination.

The outer IP packet has the source address of the source VTEP (VTEPS) and the destination IP of the destination VTEP (VTEPD) fronting the destination MAC address contained in the original Layer 2 frame. VTEPD is discovered when a node sends out an ARP looking for the MAC address corresponding to an IP address. If VTEPS does not know the IP address of the VTEPD, the following process is followed to discover this information:

1. VNI of the VXLAN segment is matched to its associated multicast address.
2. ARP request is encapsulated in a multicast packet whose address corresponds to the multicast address associated with this VNI.
3. Multicast packet is received by all VTEPs that have subscribed to this multicast address, because they are configured with virtual machines in the VNI.
4. All VTEPs glean from this multicast packet the IP address of VTEPS and the source virtual machine's MAC address, which are added to local lookup tables.
5. All VTEPs forward the ARP request to their port group that is associated with the VNI in the multicast packet.
6. The destination virtual machine responds to the ARP request as normal.
7. VTEPD encapsulates the response and sends out a unicast packet back to VTEPS with the ARP response.
8. VTEPS decapsulates and forwards the packet on to the associated port group.
9. In the process of decapsulating the packet, VTEPS gleans the destination node's MAC address and VTEPD IP address from the packet and adds them to its local lookup tables for future unicast communication.

All unknown destination packets, broadcasts, and multicasts are treated in a similar manner, by encapsulating them in multicast packets. This makes it possible for VXLAN-backed networks to support any type of protocol that rides on top of Ethernet, even non IP-based protocols.

VM to VM traffic is encapsulated and sent as unicast traffic between the VTEPs. Only unknown, broadcast and multicast traffic is encapsulated and sent out as multicast traffic.

IGMP snooping [1] needs to be enabled on the physical switches. This enables the physical switches to treat multicast traffic, not like broadcasts, but with a little more intelligence. Physical switches with IGMP snooping enabled will filter out ports that have not subscribed to multicast traffic, thus reducing the amount of data the attached nodes need to process. For IGMP snooping to work, a router, or a Layer 3 switch, needs to be configured as an IGMP queryer. The router will send regular queries about multicast subscription on the network, spurring responses from all the VTEPs about the multicast addresses they are subscribed to. The physical switch snoops these responses and uses the information to maintain its multicast subscription tables (Figure 6).

For VXLAN to work in an environment where VTEPs are not all in the same Layer 2 network (i.e., there are routers used for forwarding between the VTEPs which are on different Layer 3 networks), multicast routing needs to be enabled across the IP network. This is usually done through a protocol like PIM (Protocol-Independent Multicast).

What Does VXLAN Buy Me?

VXLAN creates a network abstraction layer over available physical networks. With a VXLAN fabric in a datacenter, it is possible to overlay multiple VXLAN-backed Layer 2 networks all over the datacenter, providing Layer 2 adjacency to VMs hosted in the datacenter. This makes it possible to create on-demand networks on top of this fabric, allowing unconstrained virtual machine placement within the datacenter and, at the same time, affording unencumbered virtual machine mobility.

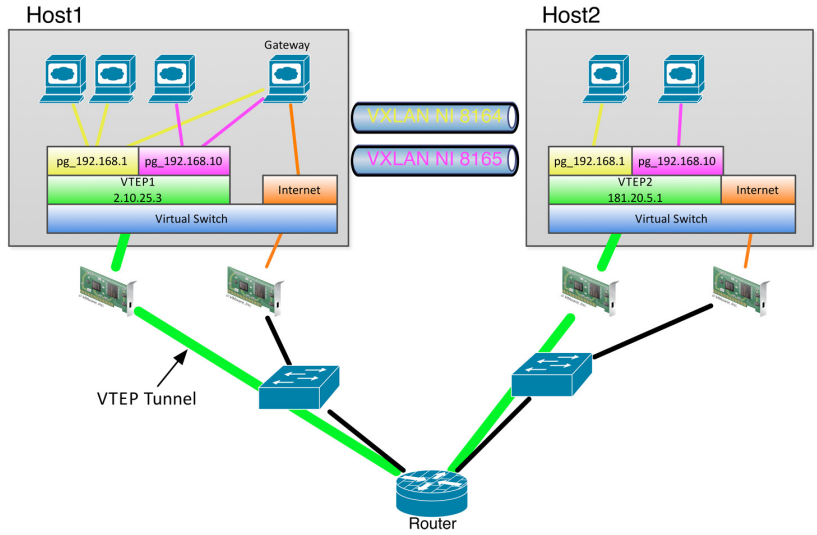

Figure 6: Physical VXLAN topology showing virtual machine connectivity and tunneled VXLAN traffic together with regular traffic from a VXLAN to VLAN gateway

There is no requirement in the VXLAN specification (IETF draft [2]) as defined, for the nodes on a VXLAN backed Layer 2 network to be virtual. There is a requirement though on the physical hosts, or Layer 2 physical switches, for a VXLAN stack in order for physical nodes to attach to a VXLAN fabric.

Acknowledgments

Many thanks go to T. Sridhar for his edits and mentorship.

References

[1] IGMPv3 and IGMP Ssnooping switches: http://tools.ietf.org/html/draft-ietf-idmr-snoop-00.

[2] A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks: http://tools.ietf.org/html/draft-mahalingam-dutt-dcops-vxlan-01.