BORGcube Blogs

KAMAU WANGUHU'S RANDOM THOUGHTS

Nov 03 2011

VXLAN Primer-Part 1

Networking, vCD, Virtualization Add comments

There has been a lot of chatter in the bloggersphere about the advent of Virtual eXtensible Local Area Network (VXLAN) and all the vendors that contributed to the standard as well as those that are planning on supporting the proposed IETF draft standard. In the next couple of articles I will attempt to describe how VXLAN is supposed to work as well as give you an idea of when you should consider implementing it, and how to implement it in your VMware Infrastructure (VI).

VXLAN Basics:

The basic use case for VXLAN is to connect two or more layer three (L3) networks and make them look like they share the same layer two (L2) domain. This would allow for virtual machines to live in two disparate networks yet still operate as if they were attached to the same L2. See section 3 of the VXLAN IETF draft as it addresses the networking problems that VXLAN is attempting to solve a lot better than I ever could.

To operate a VXLAN needs a couple of components in place:

- Multicast support, IGMP and PIM
- VXLAN Network Identifier (VNI)
- VXLAN Gateway

Categories

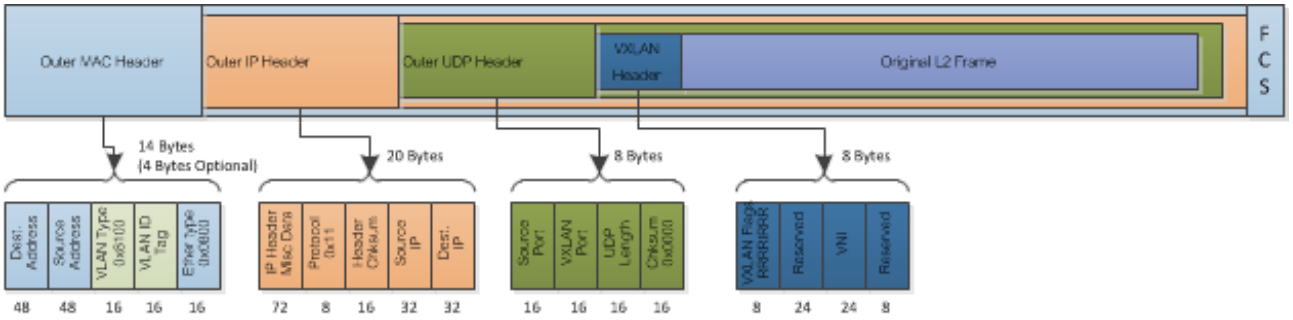
Hardware (1)
Network Virtualization (2)
Networking (15)
Networking (2)
NTP (1)
Servers (3)
Social (4)
Uncategorized (1)
vCD (7)
vCenter (1)
Virtualization (9)
Volunteer (4)
Window$ (1)

Archives

July 2014 (1)
May 2014 (1)
March 2014 (2)
February 2014 (3)
June 2013 (1)
April 2013 (1)

- VXLAN Tunnel End Point (VTEP)
- VXLAN Segment/VXLAN Overlay Network

VXLAN is an L2 overlay over an L3 network. Each overlay network is known as a VXLAN Segment and identified by a unique 24-bit segment ID called a VXLAN Network Identifier (VNI). Only virtual machine on the same VNI are allowed to communicate with each other. Virtual machines are identified uniquely by the combination of their MAC addresses and VNI. As such it is possible to have duplicate MAC addresses in different VXLAN Segments without issue, but not in the same VXLAN Segments.



VXLAN Transport Header Format

Figure 1: VXLAN Packet Header

The original L2 packet that the virtual machines send out is encapsulated in a VXLAN header that includes the VNI associated with the VXLAN Segments that the virtual machine belongs to. The resulting packet is then wrapped in a UDP->IP->Ethernet packet for final delivery on the transport network. Due to this encapsulation you can think of VXLAN as a tunneling scheme with the ESX hosts making up the VXLAN Tunnel End Points (VTEP). The VTEPs are responsible for encapsulating the virtual machine traffic in a VXLAN header as well as stripping it off and presenting the destination virtual machine with the original L2 packet.

October 2012 (3)
June 2012 (1)
March 2012 (1)
February 2012 (1)
November 2011 (2)
March 2011 (5)

Meta

Register
Log in
Entries RSS
Comments RSS
WordPress.org

Google Ads

Follow Kamau

The encapsulation is comprised of the following modifications from standard UDP, IP and Ethernet frames:

Ethernet Header:

Destination Address - This is set to the MAC address of the destination VTEP if it is local of to that of the next hop device, usually a router, when the destination VTEP is on a different L3 network.

VLAN - This is optional in a VXLAN implementation and will be designated by an ethertype of 0×8100 and have an associated VLAN ID tag.

Ethertype - This is set to 0×0800 as the payload packet is an IPv4 packet. The initial VXLAN draft does not include an IPv6 implementation, but it is planned for the next draft.

IP Header:

Protocol - Set 0×11 to indicate that the frame contains a UDP packet

Source IP - IP address of originating VTEP

Destination IP - IP address of target VTEP. If this is not known, as in the case of a target virtual machine that the VTEP has not targeted before, a discovery process needs to be done by originating VTEP. This is done in a couple of steps:

a. Destination IP is replaced with the IP multicast group corresponding to the VNI of the originating virtual machine
b. All VTEPs that have subscribed to the IP multicast group receive the frame and decapsulate it learning the mapping of source virtual machine MAC address and host VTEP
c. The host VTEP of the destination virtual machine will then send the virtual machines response to the originating VTEP using its destination IP address as it learned this

from the original multicast frame

d. The Source VTEP adds the new mapping of VTEP to virtual machine MAC address to its tables for future packets

UDP Header:

Source Port - Set by transmitting VTEP

VXLAN Port - IANA assigned VXLAN Port. This has not been assigned yet

UDP Checksum - This should be set to 0×0000. If the checksum is not set to 0×0000 by the source VTEP, then the receiving VTEP should verify the checksum and if not correct, the frame must be dropped and not decapsulated.

VXLAN Header:

VXLAN Flags - Reserved bits set to zero except bit 3, the I bit, which is set to 1 to for a valid VNI

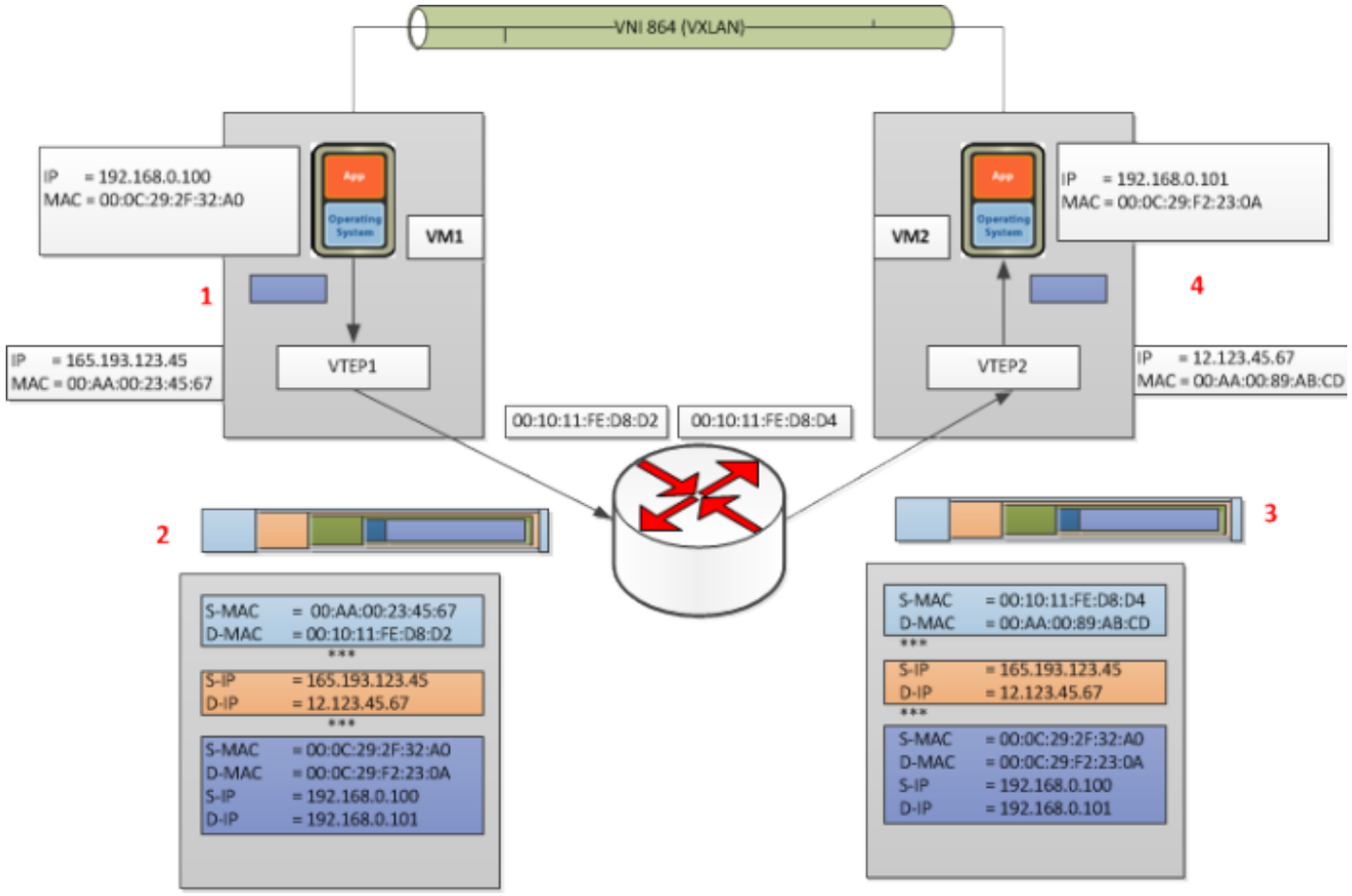
VNI - 24-bit field that is the VXLAN Network Identifier

Reserved - A set of fields, 24 bits and 8 bits, that are reserved and set to zero

Putting it Together:

Tags Used

/24 /56 /64 Active Directory AD Architecture ARIN Broadcast Cloud Cloud Director Core Nap Distributed Switch dvfilter dvs External Network Global IANA IETF IP Block IPv4 IPv6 Isolation ISP Link-Local Loopback MTU Multicast NAT/Routed Networking Network Pool Organization Network RFC 1918 RFC 2462 RFC 3849 RIR Site-local Testing Time Warner ULA Uniform Local vApp Network vCD vCDNI VMware VXLAN



VXLAN: VM to VM communication

Figure 2: VM to VM communication

When VM1 wants to send a packet to VM2, it needs the MAC address of VM2 this is the process that is followed:

- VM1 sends a ARP packet requesting the MAC address associated with 192.168.0.101
- This ARP is encapsulated by VTEP1 into a multicast packet to the multicast group associated with VNI 864

- All VTEPs see the multicast packet and add the association of VTEP1 and VM1 to its VXLAN tables
- VTEP2 receives the multicast packet decapsulates it, and sends the original broadcast on portgroups associated with VNI 864
- VM2 sees the ARP packet and responds with its MAC address
- VTEP2 encapsulates the response as a unicast IP packet and sends it back to VTEP1 using IP routing
- VTEP1 decapsulates the packet and passes it on to VM1

At this point VM1 knows the MAC address of VM2 and can send directed packets to it as shown in in Figure 2: VM to VM communication:

1. VM1 sends the IP packet to VM2 from IP address 192.168.0.100 to 192.168.0.101
2. VTEP1 takes the packet and encapsulates it by adding the following headers:
    - VXLAN header with VNI=864
    - Standard UDP header and sets the UDP checksum to 0×0000, and the destination port being the VXLAN IANA designated port. Cisco N1KV is currently using port ID 8472.
    - Standard IP header with the Destination being VTEP2's IP address and Protocol 0×011 for the UDP packet used for delivery
    - Standard MAC header with the MAC address of the next hop. In this case it is the router Interface with MAC address 00:10:11:FE:D8:D2 which will use IP routing to send it to the destination
3. VTEP2 receives the packet as it has it's MAC address as the destination. The packet is decapsulated and found to be a VXLAN packet due to the UDP destination port. At this point the VTEP will look up the associated portgroups for VNI 864 found in the VXLAN header. It will then verify that the target, VM2 in this case, is allowed to receive frames for VNI 864 due to it's portgroup membership and pass the packet on if the verification passes.
4. VM2 receives the packet and deals with it like any other IP packet.

The return path for packet from VM2 to VM1 would follow the same IP route through the router on the way back.

Share this:

Print | Twitter 37 | Reddit | StumbleUpon | Facebook | Email | LinkedIn 16 | Google

Posted by Kamau Wanguhu on 2011/11/03 Tagged with: Multicast, Networking, vCD, VLAN, VMware, VXLAN

29 Responses to “VXLAN Primer-Part 1”

1. **Kamau Wanguhu’s VXLAN Primer – Part 1 | NSX Insight** says:
2014/07/28 at 07:30

[…] still sticking your toes in the water with VXLAN and want to take a deep dive Kamu Wanguhu’s VXLAN Primer – Part 1 is an excellent way to get […]

2. **VXLAN Primer-Part 1 (Clarifications) - BORGcube Blogs** says:
2014/05/05 at 17:57

[…] on feedback I have received so far on Part 1, I need to make some clarifications to what I have covered specifically with how the protocol […]

3. **Ying Wang** says:
2014/04/09 at 04:52

Nice article!
But i can’t understand the “Destination IP” section. The process described below, in my opinion, is to find the MAC addresss, not “Destination IP”. Am i right?

Reply

Kamau Wanguhu says:
2014/04/10 at 20:30

You are right in a way.

The process is for discovering the IP address of the VTEP hosting a VM. One reason you need the IP address of the VTEP could be because you are trying to discover the MAC address of a virtual machine. Another reason could be because you know the MAC address of the VM, but you have no idea on which host it is on. This last one should not happen unless the VMs ARP cache lifetime is longer than the VTEP destination maps.

In short anytime an ESXi host does not know where to send traffic for a VM, it's VTEP will need to discover the IP address of the destination VTEP which is the process that I walked you through.

Reply

4. **Ashwani** says:
2014/04/08 at 06:17

Great explanation…keep it up

Reply

5. **Moses** says:
2014/03/21 at 16:07

Great primer, met Martin Casado discussed NSX, interesting stuff. Keep up the good work fellow countryman

Reply

6. **chirag** says:
2014/03/17 at 01:14

Nice explanation

Reply

7. **VXLAN | squarey's blog** says:
2013/12/18 at 09:24

[...] VxLAN - Extending VLAN into the Cloud VXLAN basics and use cases (when / when not to use it) VXLAN Primer-Part 1 VXLAN Primer-Part 2: Let's Get Physical Typical VXLAN Use Case Digging Deeper into [...]

8. **Tytus** says:
2013/11/07 at 07:19

Clearly explained. Thank you!

Reply

9. **Devi Prasad Ivaturi** says:
2013/09/10 at 02:26

Small correction, IMO. The DMAC, in the payload has to be all F's as you were explaining about ARP resoultion. Agree?

Reply

Kamau Wanguhu says:
2013/09/10 at 09:54

Not sure which DMAC you are referring to. If it is an ARP request from the VM then the DMAC will be all F's as it is a broadcast at the Ethernet level. For the outer packet, from the VTEP, the

DMAC is always the MAC of the next hop.

Reply

10. **Some VMware vCloud Director related news | Timo Sugliani** says:
2013/08/12 at 15:51

[...] VXLAN Primer Part 1 [...]

11. **EIad Al-Aqqad** says:
2013/02/07 at 18:22

Thanks for the great post.

Reply

12. **VXLAN & Multicast « CloudAssassin – Cloud Architecture Simplified** says:
2013/01/22 at 11:26

[...] Multicast and how to enable this in vCloud Director, if you dont this has been blogged about here:http://www.borgcube.com/blogs/2011/11/vxlan-primer-part-1/ and [...]

13. **Internets of Interest for 18th January 2013 — EtherealMind** says:
2013/01/18 at 21:34

[...] VXLAN Primer-Part 1 - BORGcube Blogs - Good details on VXLAN. Lots of hard work in here. Bookmarked for later reference. [...]

14. **Word Of Caution About Overextending The Use Of VXLAN « vArchitect Musings** says:
2013/01/04 at 04:25

[...] explanation of VXLAN, I encourage everyone to read Kamau Wanguhu's extremely informative primer and

Scott Lowe's blog post on "Examining VXLAN." The benefits of VXLAN for data [...]

15. **VXLAN: The Right Data Center Interconnect Technology For vCloud Connector? « vArchitect Musings** says:
2013/01/03 at 04:48

[...] explanation of VXLAN, I encourage everyone to read Kamau Wanguhu's extremely informative primer and Scott Lowe's blog post on "Examining VXLAN." The benefits of VXLAN for data [...]

16. **Mike Laverick** says:
2012/12/13 at 14:49

Great - post! Although its my understanding that VXLAN is not yet supported across two sites… yet…

Reply

Kamau Wanguhu says:
2012/12/13 at 16:03

Mike, That is correct for the cases where the sites are managed by different vCenter/vShield Manager pairs.

If you want the VXLAN backed networks to be available in a multisite configuration the sites have to be managed b the same vCenter/vShield Manager pair. This will allow for the VXLAN name space to be the same. Note that for a multi site configuration that you are still bound to the support restrictions of vCenter such as latency and WAN bandwidth.

The other question you always have to ask is what it is you are trying to achieve by spanning your L2 across sites. There are a lot of application implications to having L2 adjacency with high latency, take high transaction systems as an example…

Reply

17. **Capcorne** says:
2012/12/12 at 12:20

Hi,

What if the two or more L3 networks are in two different DC, this will work as the transport layer is not aware of the communication between VM and then VXLAN could achieve the same thing than OTV or I' m missing something ?

Reply

> **Kamau Wanguhu** says:
> 2012/12/12 at 17:14
>
> 
>
> @7f8096629719bcec75ce114cdf3f44c7:disqus that is correct and is one of the use cases for VXLAN. In vCD 5.1, for this to work, the two data centers would need to be managed by the same vCenter as the name space is maintained by the vShield Manager.
>
> Reply

18. **Capcorne** says:
2012/12/11 at 17:55

Very clear explanation and schema! thanks for sharing.

Reply

19. **Gary** says:
2012/11/06 at 19:03

Great explanation! Thanks

Reply

20. **Carl P** says:
2012/11/05 at 12:16

Very cool stuff!

Reply

21. **Tim Oudin** says:
2012/11/02 at 19:20

Thanks Kamau!

Reply

22. **VXLAN basics and use cases (when / when not to use it)** says:
2012/11/02 at 15:11

[...] some excellent articles about how this works, and I suggest you read that if you are interested. (VXLAN Primer Part 1, VXLAN Primer Part 2) On top of that I would also highly recommend Massimo's Use Case [...]

23. **VXLAN and Layer 3 Connectivity - blog.scottlowe.org - The weblog of an IT pro specializing in virtualization, storage, and servers** says:
2011/12/02 at 17:47

[...] Digging Deeper into VXLAN, Part 1 VXLAN Deep Dive, Part 2: Looking at the Options Digging Deeper in VXLAN, Pt 3, More FAQs The Care and Feeding of VXLAN VXLAN Part Deux VXLAN Conclusion Google+ discussions on VXLAN VXLAN Primer - Part 1, BORGcube Blogs [...]

24. **vitocorleone** says:
2011/11/09 at 18:04

This is an excellent primer on a critically important new cloudy technology.

Reply

25. **VXLAN Primer – flyingpenguin** says:
2011/11/08 at 01:58

[...] If you are new here, you might want to subscribe to the RSS feed for updates on this topic.A nice summary and introduction to VXLAN is on BORGcube. Note the responsible system for encapsulation since that is always a focus of [...]

Leave a Reply

Connect with:

Powered by OneAll Social Login

Name (required)

E-mail (required)

URI

Your Comment

You may use these HTML tags and attributes: <a href="" title=""> <abbr title=""> <acronym title=""> <b> <blockquote cite=""> <cite> <code> <del datetime=""> <i> <q cite=""> <strike>

Submit Comment

Notify me of follow-up comments by email.

Notify me of new posts by email.

vCD Network Isolation-vCDNI VXLAN Primer-Part 1 (Clarifications)