ARISTA

Introduction into VXLAN

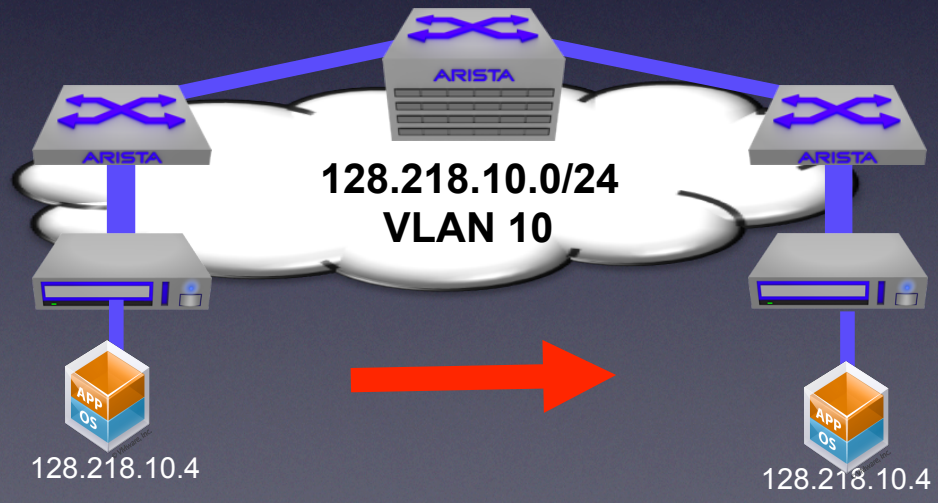
4th DENOG Meeting, Darmstadt
November 15th, 2012

Frank Laforsch
Systems Engineer, EMEA
f.laforsch@aristanetworks.com

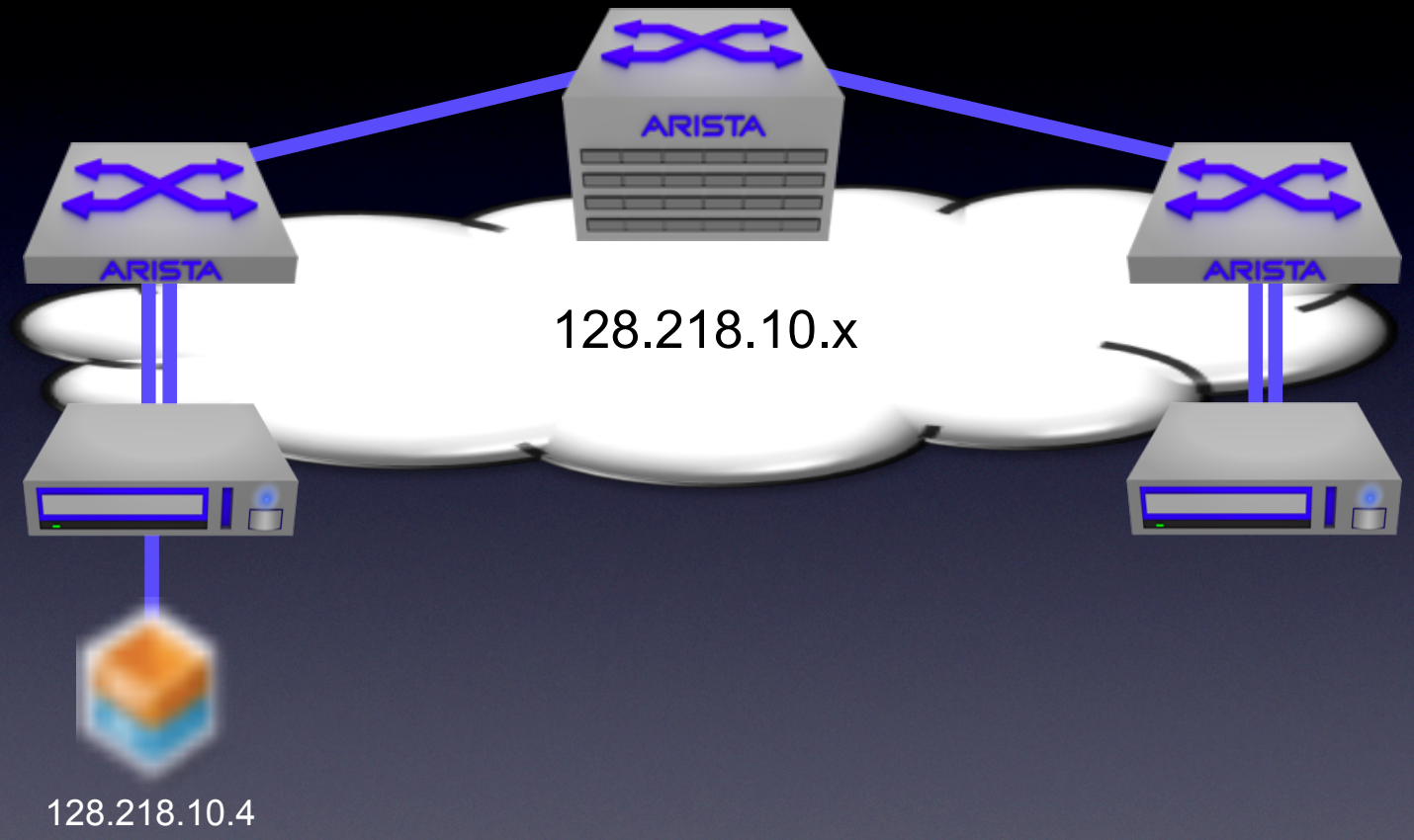

ARISTA
Virtualization Challenges

Virtualization Challenges

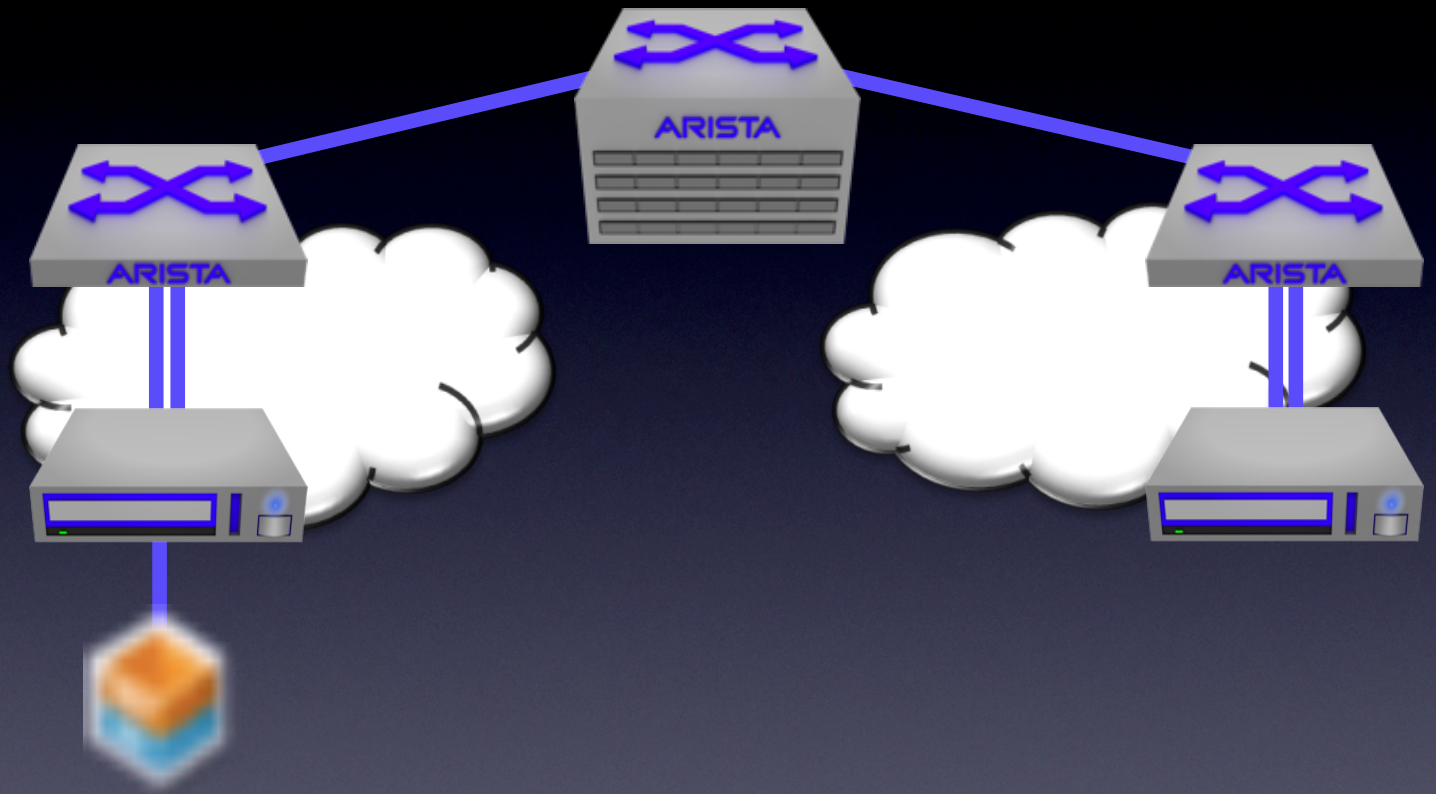
- Virtualization Challenges
  - For stateful vMotion the VM IP address must be preserved after the Vmotion
  - Ensuring zero disruption to any client communicating with the apps residing on the motioned VM
  - To ensure IP address preservation a VM can thus only be motioned to an ESXi host residing in the same subnet/VLAN.



Traditional Stateful vMotion



Non-Stateful vMotion Across L3 Subnets

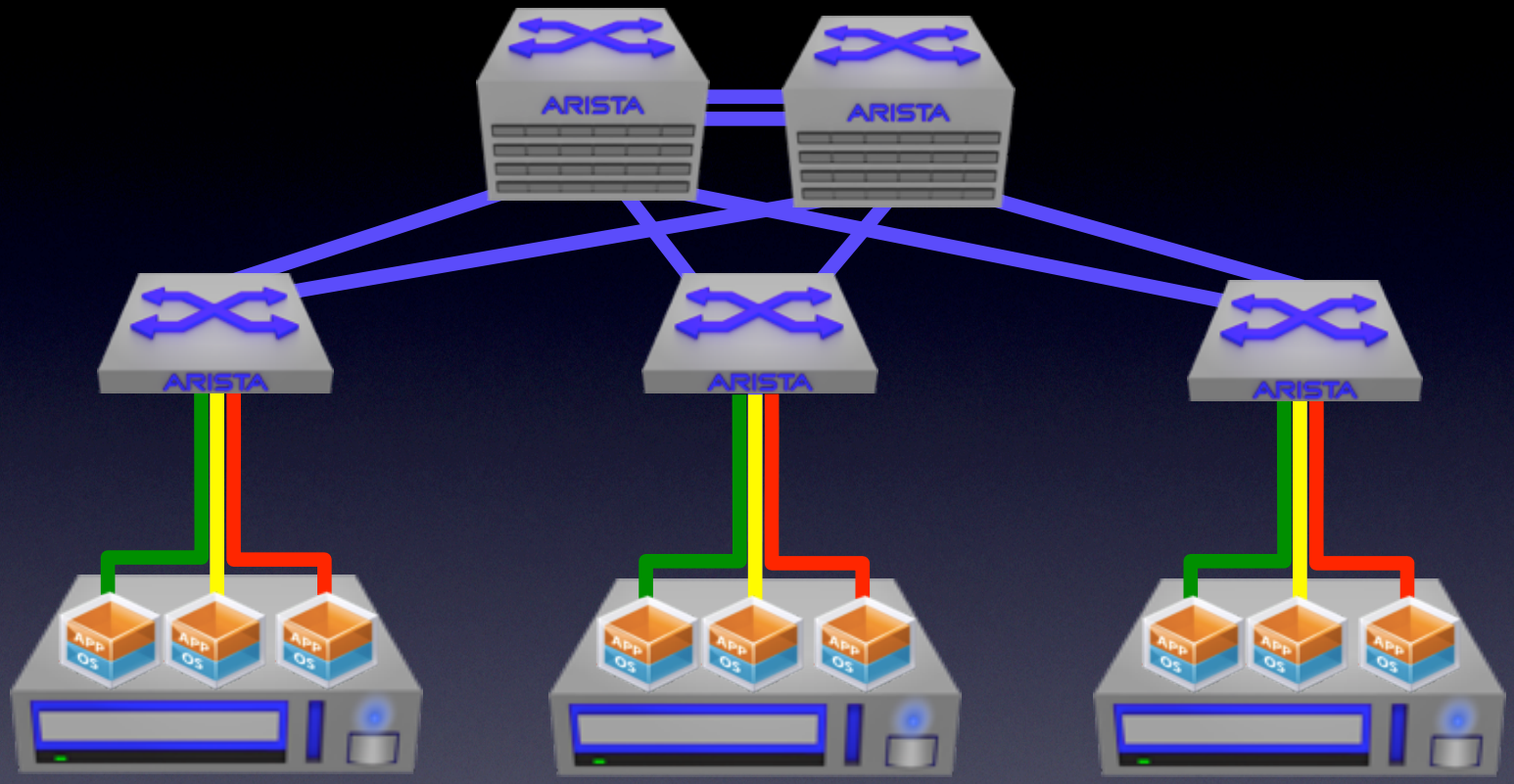


128.218.10.4

128.218.11.4

- Breaks TCP Sockets
- NFS/CIFS/iSCSI Mounts Go Away
- Reachability?

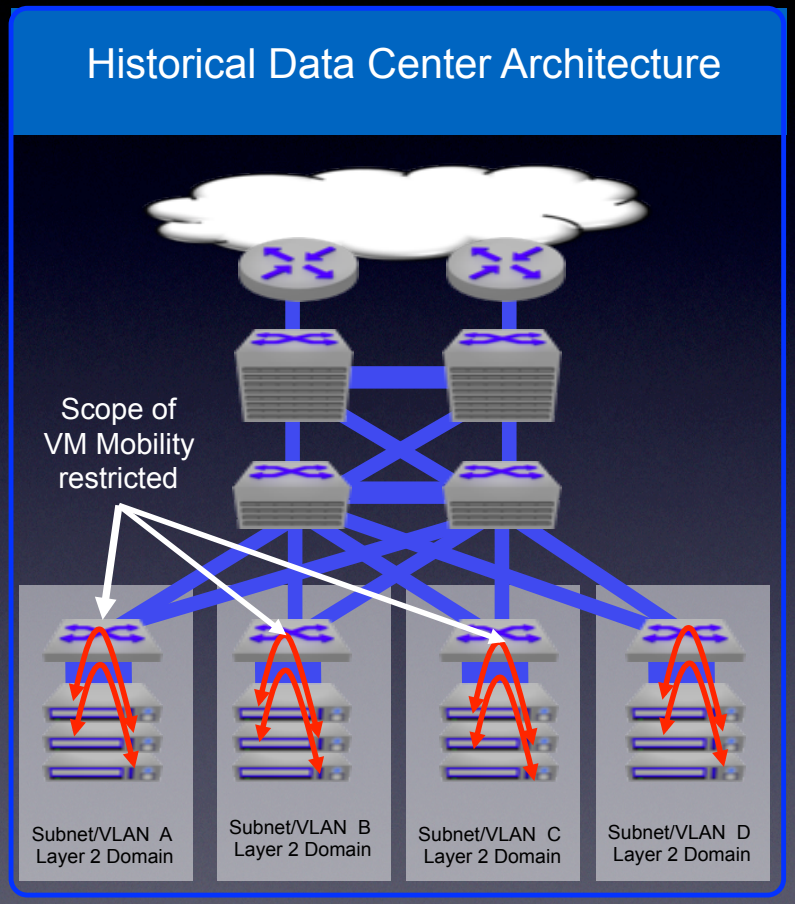
So Today, We Build Large L2 Networks!



- Use VLAN tags to segregate customer traffic
- Use Spanning Tree to create loop-free topologies
- Multi-vendor, standards-based, proven technology
- What could go wrong?

It Doesn't Scale!

Virtualization Challenges



- Experience shown large L2 domains are not optimal
  - MAC address and VLAN (4094 limit) explosion
  - Large broadcast domain
  - Single large fault domain
  - Spanning tree limitation and it's complexity
- Best practice, Silo/Segmented Layer 3 design
  - Routed traffic at the top of the rack or distribution layer
  - Reduce the size of the Layer 2 domain
  - Reducing the size of the fault and broadcast
  - Simplifying the spanning tree topology

Best practice layer 3 designs are counter productive to VM mobility

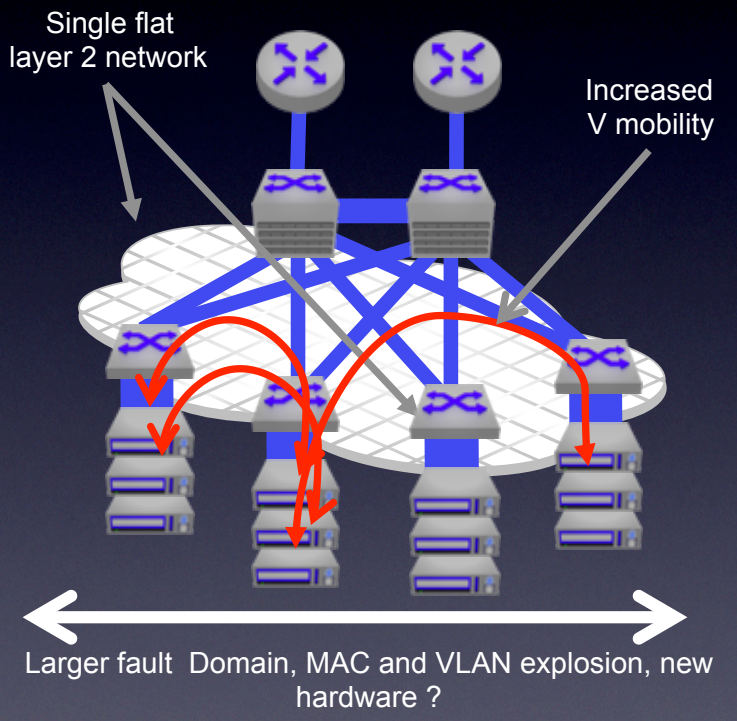
Virtualization Challenges

New protocols/technologies to scale the scope of V-mobility

- TRILL, VCS, Fabric path , short path bridging
- Remove the requirement for spanning tree
- Provide optimal active-active layer 2 traffic forwarding

The standards are based on scaling Layer 2 network topology

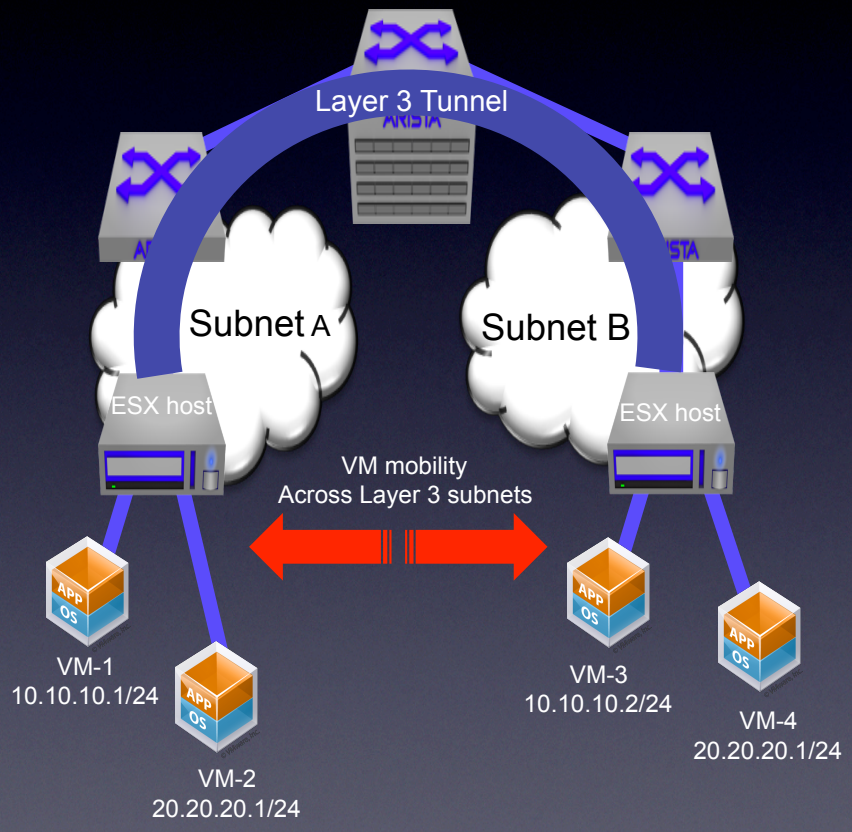
- Dramatic change from tried and trusted current L3 designs
- New hardware and operational challenges
- Don't fully address all L2 scaling concerns – MAC address/VLAN explosion, fault domain



The increased v-mobility comes at the price of needing to build a large layer 2 domain


ARISTA
VXLAN - Mobility across layer 3 boundaries

Virtual eXtensible LAN



Virtual eXtensible LAN (VXLAN)

- IETF framework proposal, co-authored by Arista, Vmware, Cisco, Citrix, Red hat and Broadcom
- Announced at VMworld 2011

Vmotion without a large L2 network

- VM mobility across Layer 3 boundaries
- Integrates seamlessly with existing infrastructure
- Supported in vSphere 5.1

Similar standards proposed by Microsoft

- NVGRE for Microsoft HyperV

VM mobility within a best practice layer 3 network Architecture

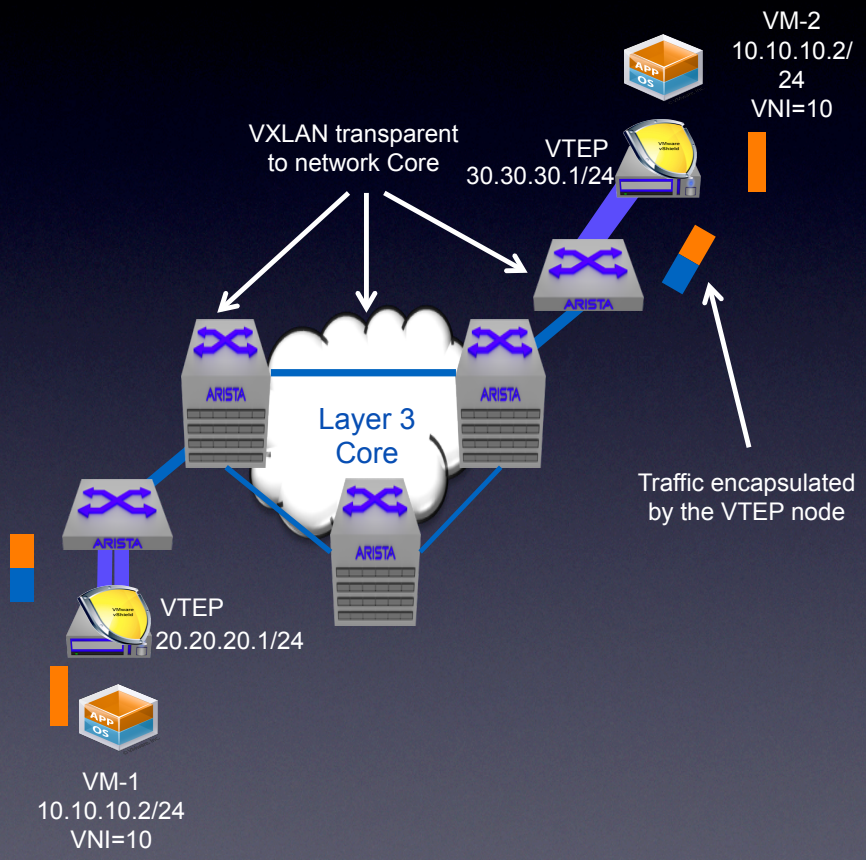
VXLAN: How does it work?

VXLAN creates logical L2 domains over standard layer 3 infrastructure

- VM traffic encapsulated inside a UDP/IP frame plus VNI identifier
- The VNI defines the layer 2 domain
- Encapsulation done by a VTEP node, VXLAN tunnel endpoint
- VTEP is a software (Vshield) or a physical switch at the ToR

The encapsulated frame routed to the remote VTEP

- Remote VTEP strips the IP/UDP header
- Original frame forwards to the local VM
- Network core transparent, not aware of the VXLAN,
- Only edge VTEP nodes need to be VXLAN aware



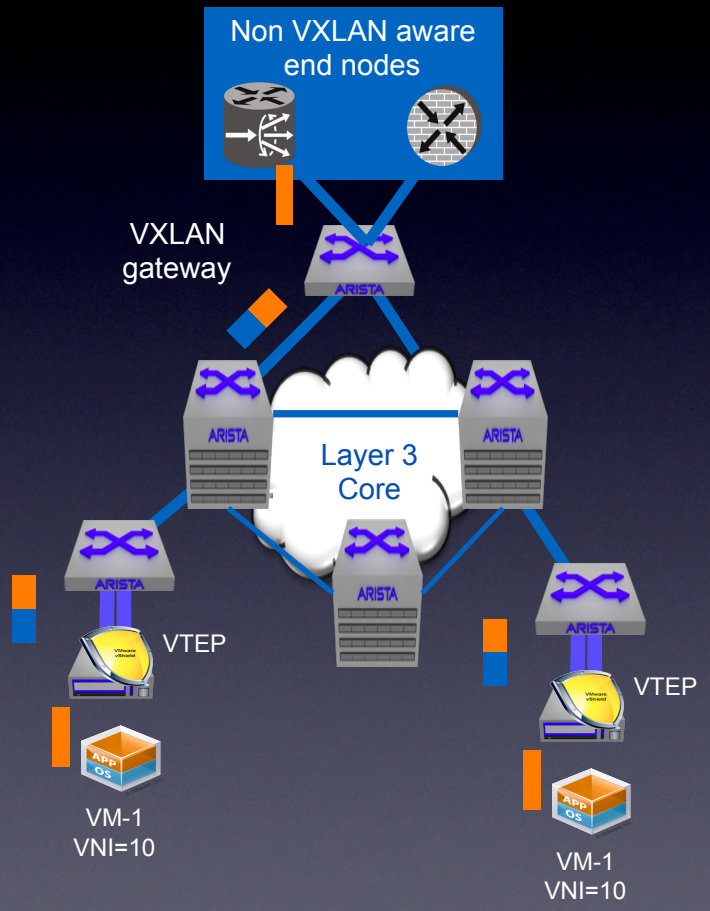
Virtual Tunnel End-Point

VMware Vshield provides a software VTEP functionality

- Encapsulation and de-encapsulation for VM traffic
- Connectivity into the VXLAN environment

VXLAN gateway functionality

- Physical VTEP functionality
- Resides on a standard 10Gbe/1Gbe ToR switch
- Encapsulation and de-encapsulation for physical servers and hardware appliances
- Providing connectivity into the VXLAN environment for non- VM machines



VXLAN Framing Format

Outer MAC Header

Outer DMAC	Outer SMAC	Ethertype (Opt)	OVLAN	Ethertype 0x800

Outer IP Header

Version	IHL	TOS	Length	ID	Flags	Fragment Offset	TTL	Protocol	Header Checksum	Outer SA	Outer DA

Outer UDP Header

SRC Port = xxxx	DST Port = VXLAN Port	UDP Length	UDP Checksum

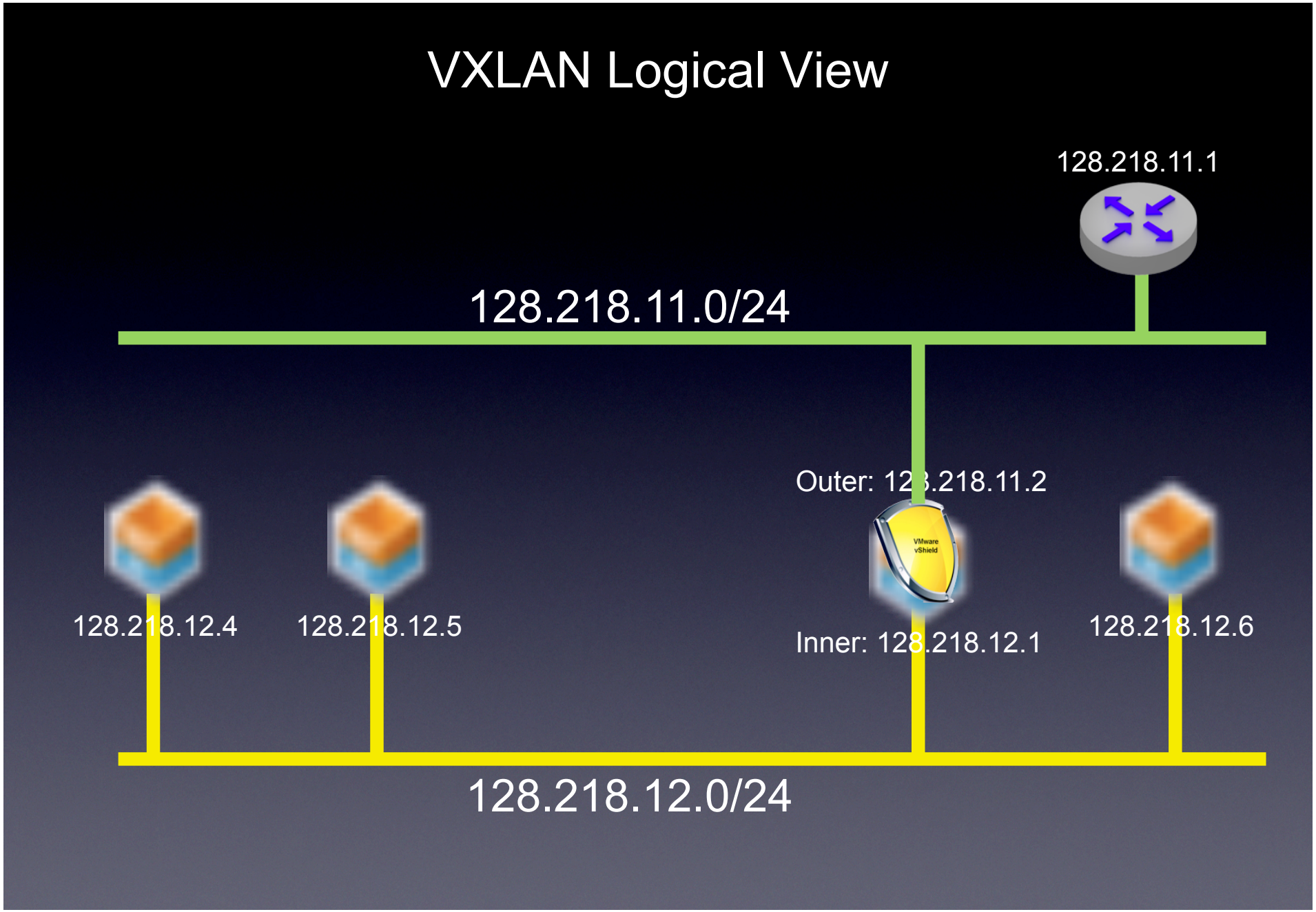
VXLAN Header

8 Reserved Flags	Reserved (24)	24-Bits VXLAN Network Identifier : (VNI)	Reserved (8)

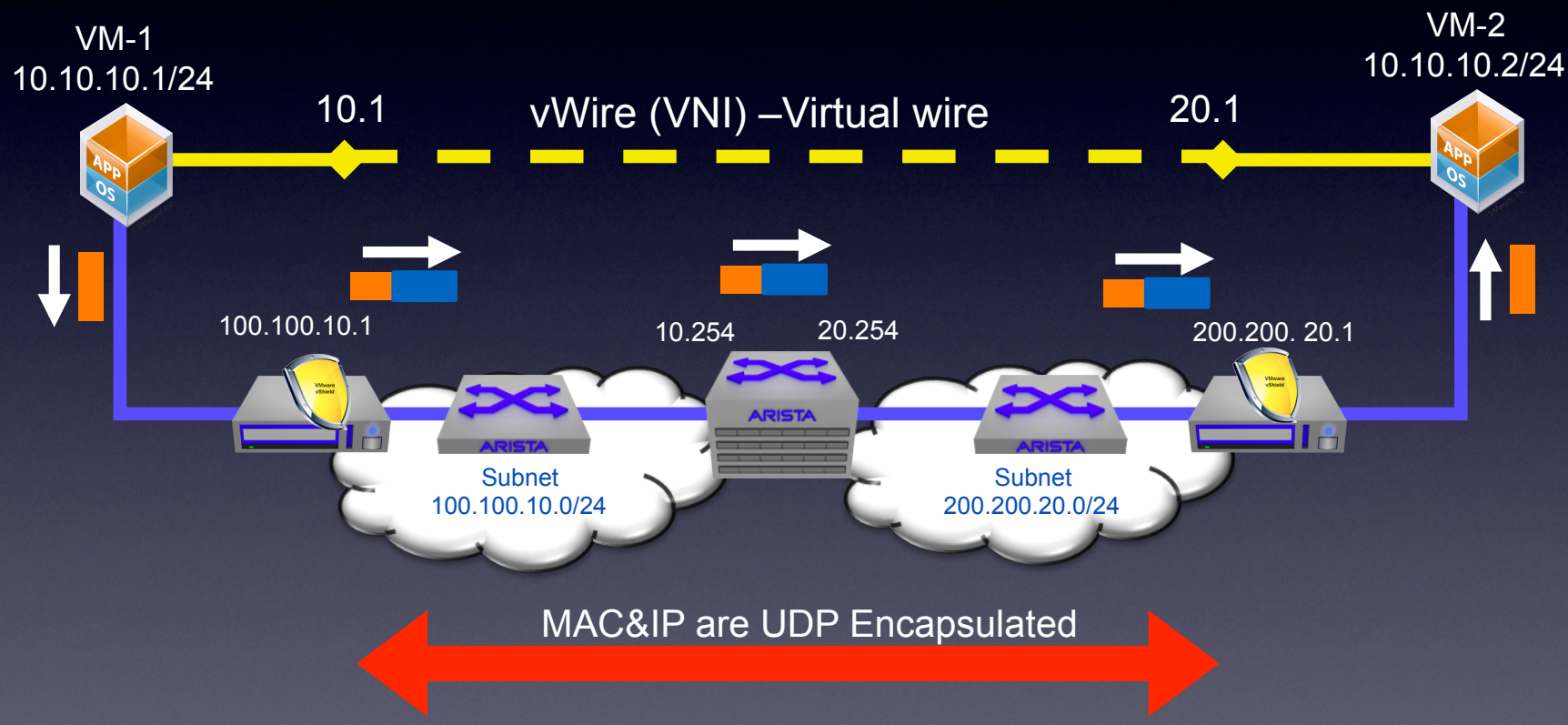
Virtual eXtensible LAN

VXLAN encapsulated frame format

- Ethernet header constructed from the local VTEP MAC and default router MAC (18 bytes)
- IP address header contains the SRC and DEST of the local and remote VTEP (20-24 bytes)
- UDP header, SRC port hash of the inner Ethernets header, dst port IANA defined (8 bytes)
  - Allows ECMP load-balancing across the network core which is VXLAN unaware.
- 24-bit VNI to scale up to 16 million for the Layer 2 domain/ vWires (8 bytes)


VXLAN Logical View
128.218.11.1
128.218.11.0/24
Outer: 128.218.11.2
VMware vShield
128.218.12.4
128.218.12.5
Inner: 128.218.12.1
128.218.12.6
128.218.12.0/24

Virtual eXtensible LAN

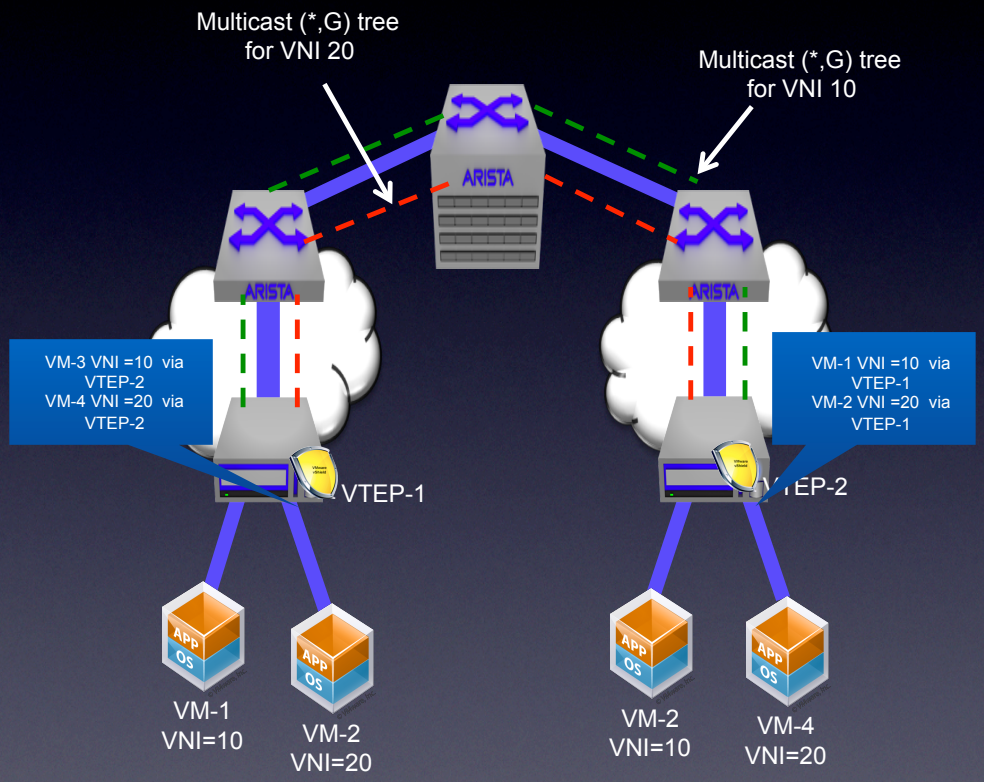


Encapsulation is transparent to traditional switch/router nodes

Virtual eXtensible LAN

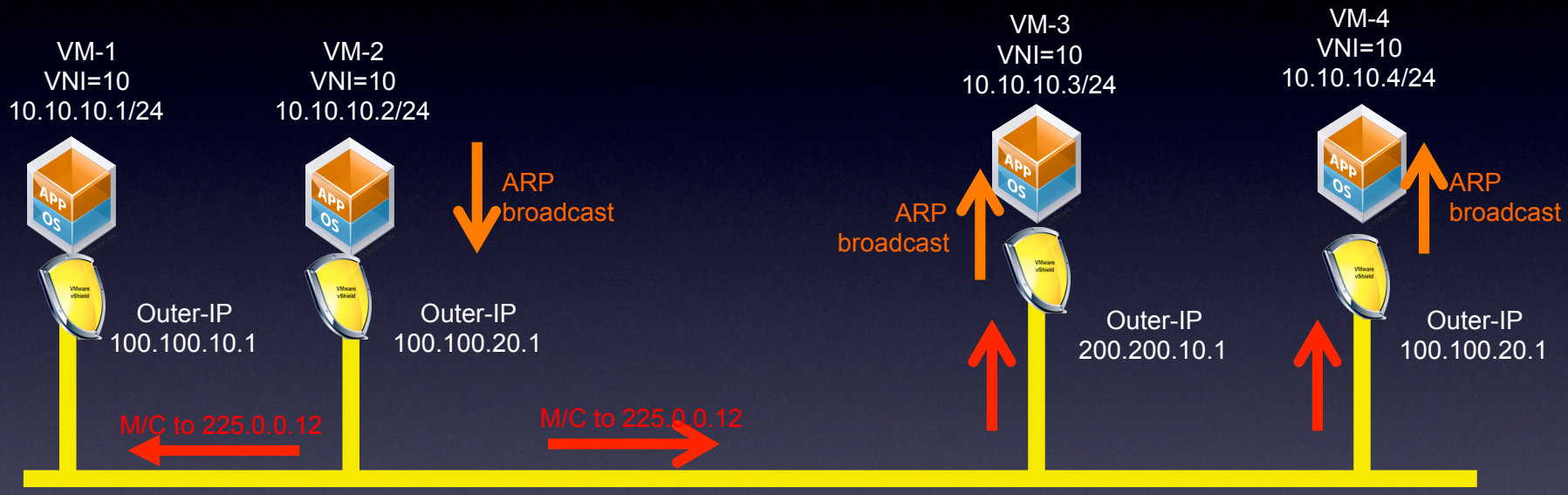
VXLAN Forwarding and learning

- Inner VM MAC learned via IP multicast
- VTEP member within a vWire/VNI joins the associated IP multicast group
- Broadcast traffic within the VNI flooded to the IP Multicast group
- Remote VM MAC bond to the remote VTEP IP on he local VTEP
- Once remote MAC to VTEP binding are created traffic forwarded using standard Layer 3 protocols



Virtual eXtensible LAN

Broadcast/unknown unicast forwarding

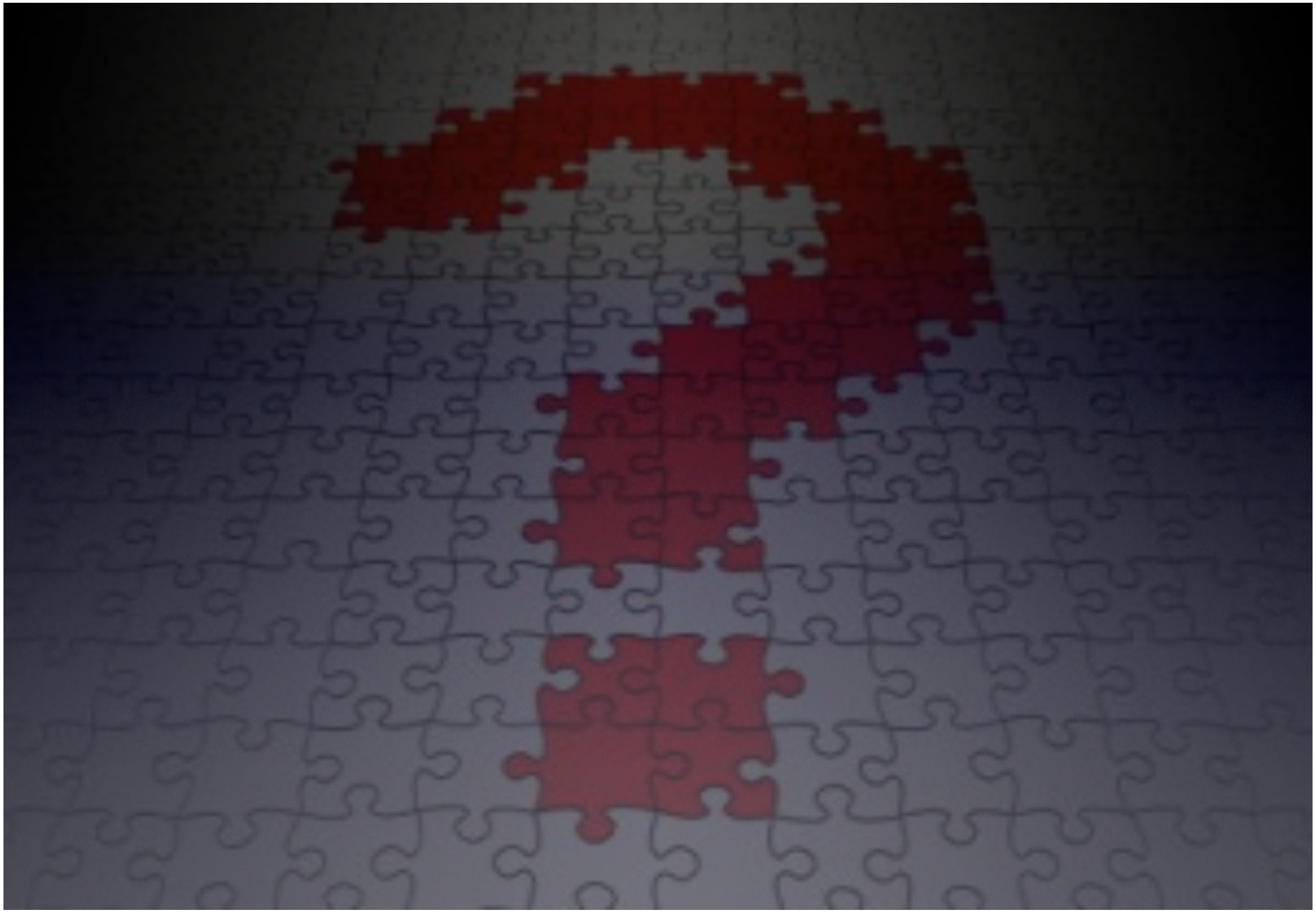


For VNI 10, broadcast sent on Multicast group 225.0.0.12

IP Multicast proven technology and require no additional hardware or software in the network core

VXLAN Summary

- For stateful vmotion, VM IP addresses need to be preserved
- Current solutions involve constructing large layer 2 domains
- VXLAN, delivers vmobility over layer 3 removing the requirement to build larger layer 2 networks.
- Uses standard UDP packet headers, technology transparent to current infrastructure
- Uses UDP to encapsulate, inner protocol controls reliable delivery.
- Layer 3 approach, overcomes MAC and VLAN limitations of Layer 2 approach and can scale to support 16.7m unique vWires




ARISTA
Thank You!