
TOMORROW
starts here.
CISCO
Cisco live!

Introduction to Overlay Networks and Use Cases

BRKDCT-1408

Usha Ramachandran, Technical Marketing Engineer

Cisco live!

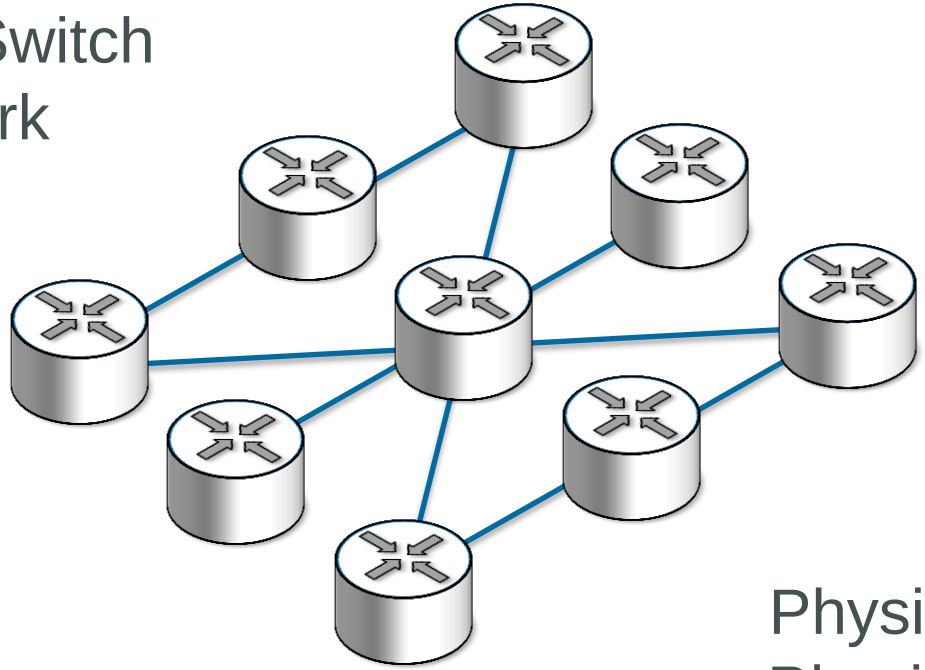
Agenda

- Virtual Overlay Technology Overview
- Why Virtual Overlays?
- VXLAN Overview
- Overlay Technologies and VXLAN Overlay comparisons
- VXLAN Enhancements
- VXLAN Gateway
- Hardware based Overlays
- Summary

Virtual Overlay Technology Overview

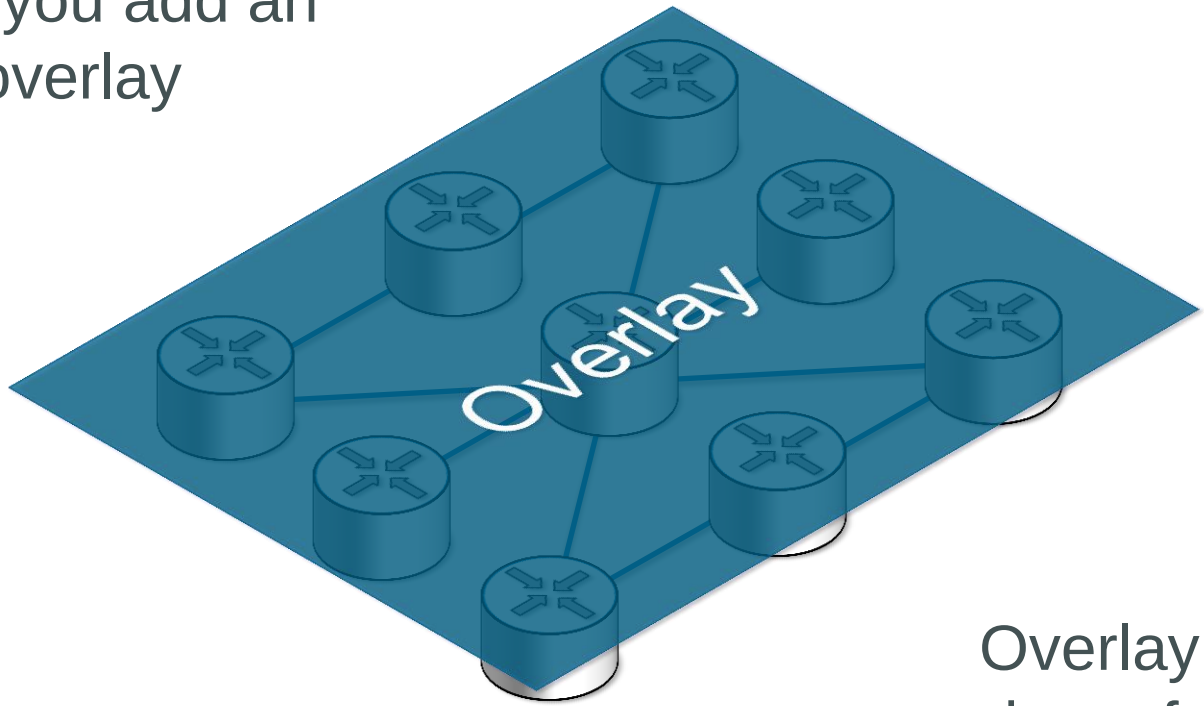
Cisco live!

You start with a Physical Switch Network



Physical Devices and Physical Connections

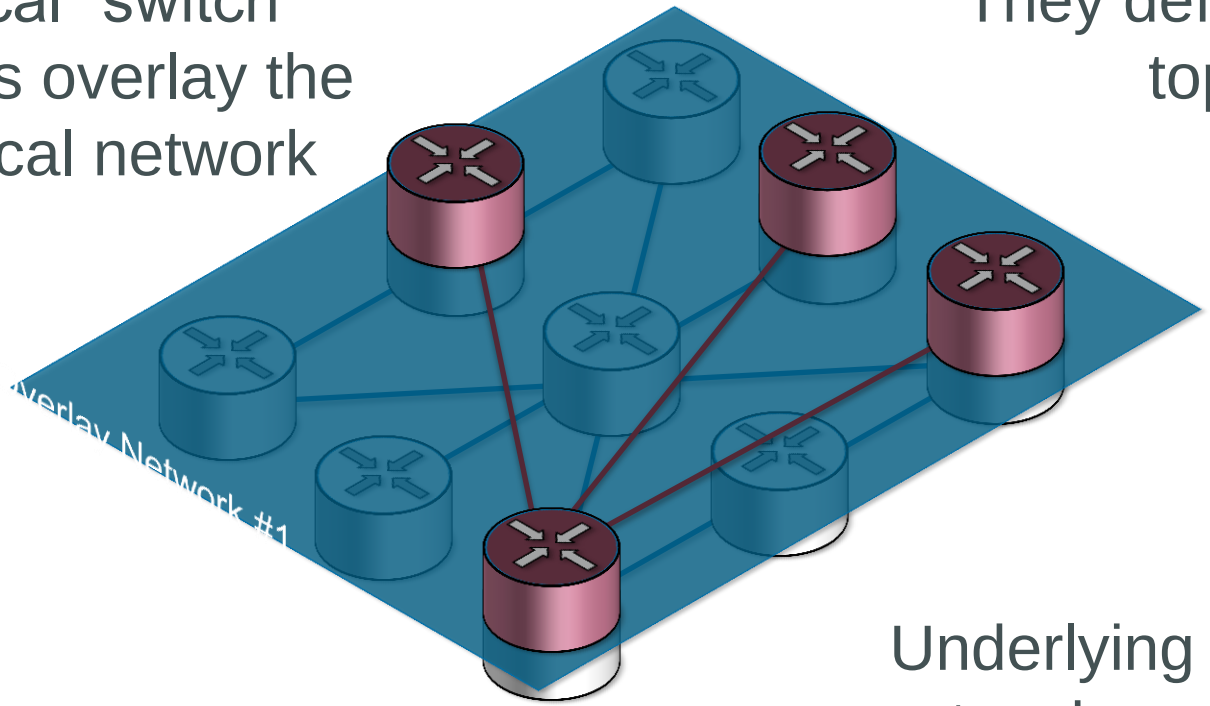
Then you add an overlay



Overlay provides base for logical network

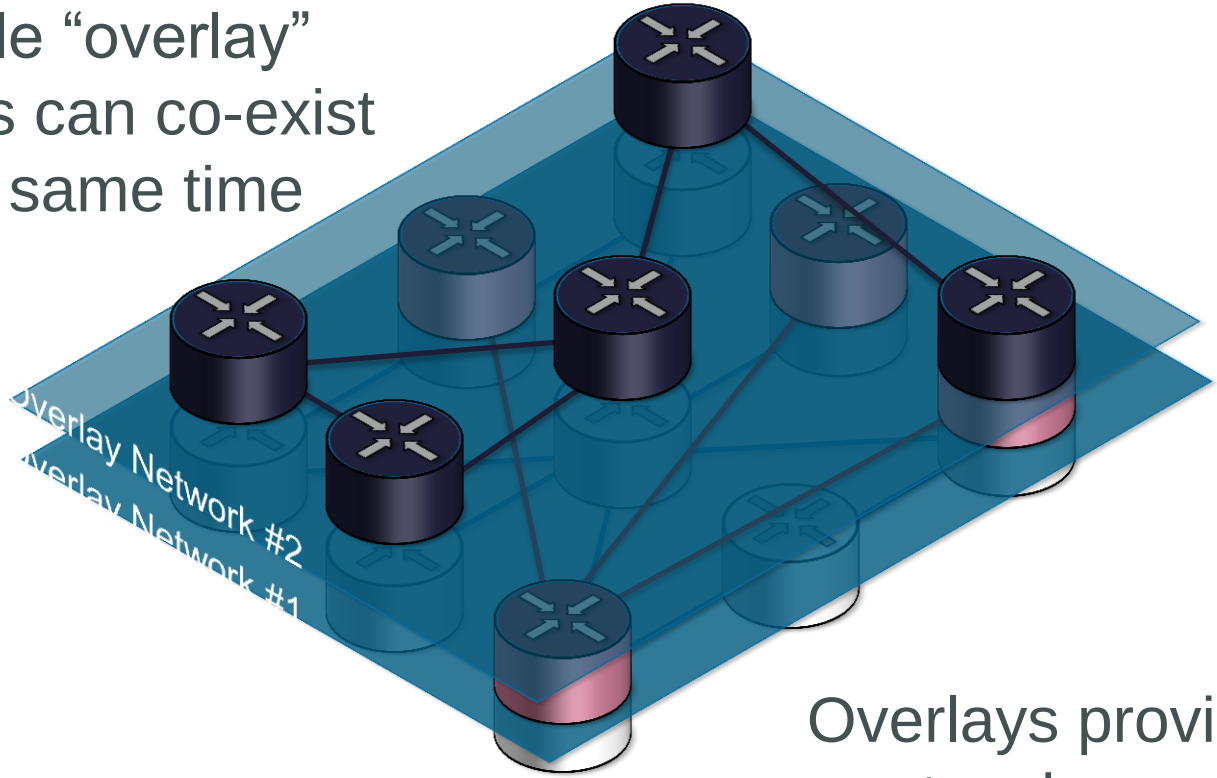
Logical “switch” devices overlay the physical network

They define their own topology



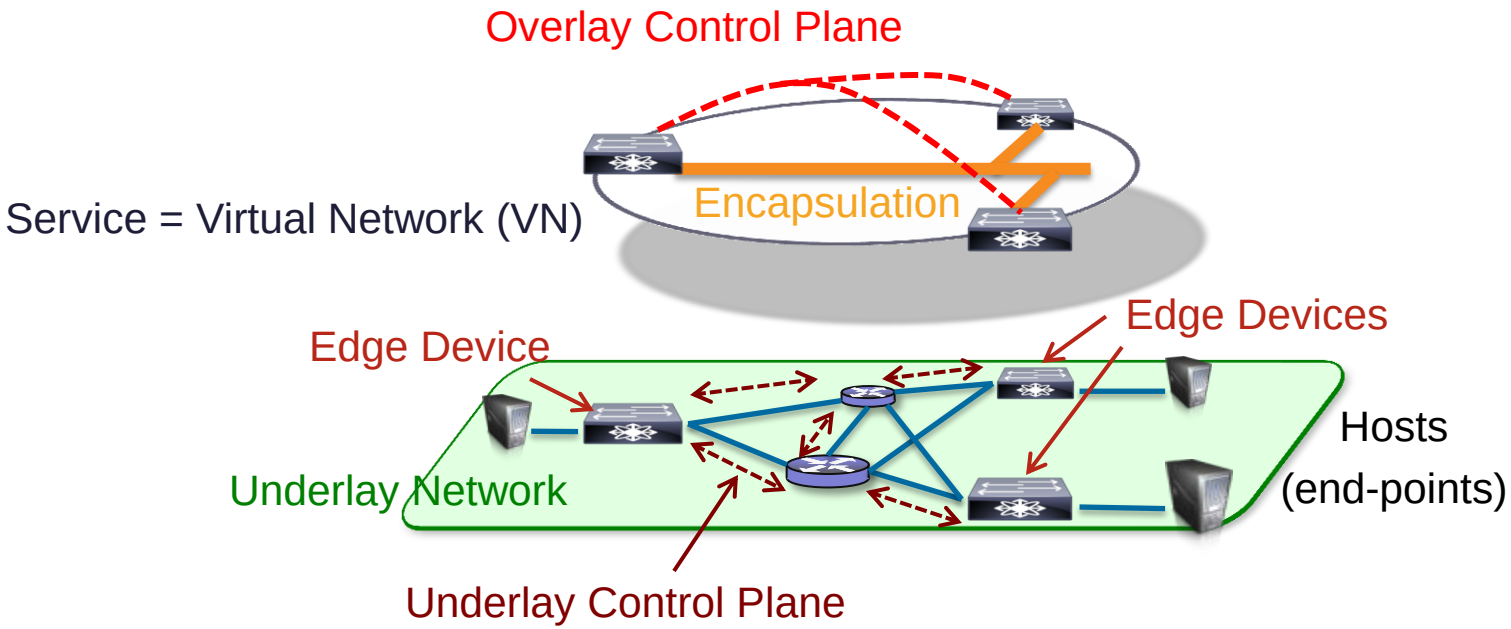
Underlying physical network carries data traffic for overlay network

Multiple “overlay” networks can co-exist at the same time

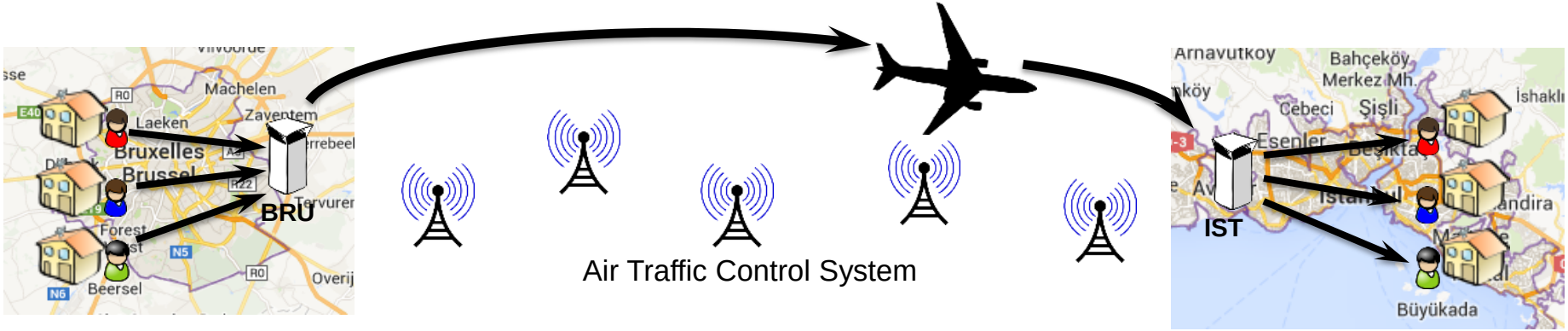


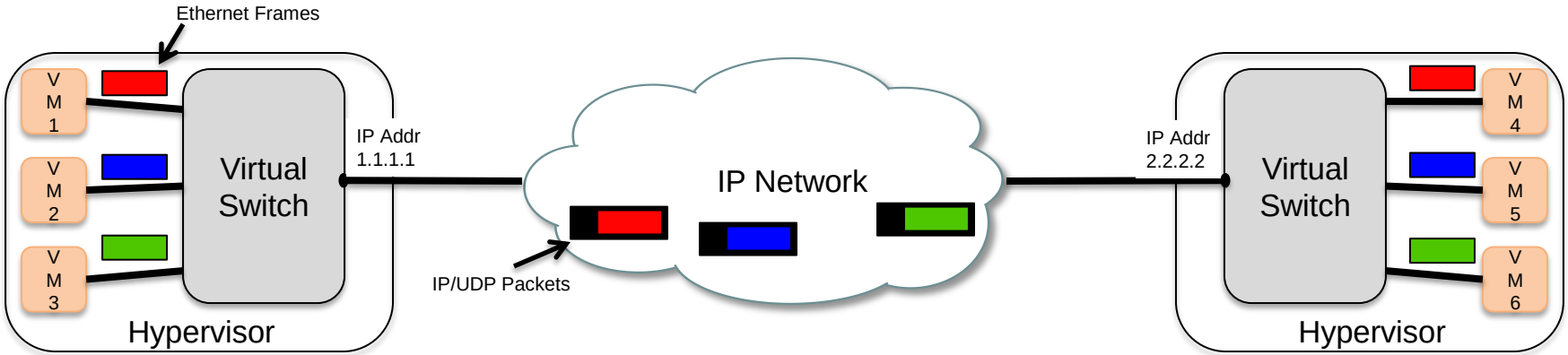
Overlays provides logical network constructs for different tenants (customers)

Overlay Taxonomy

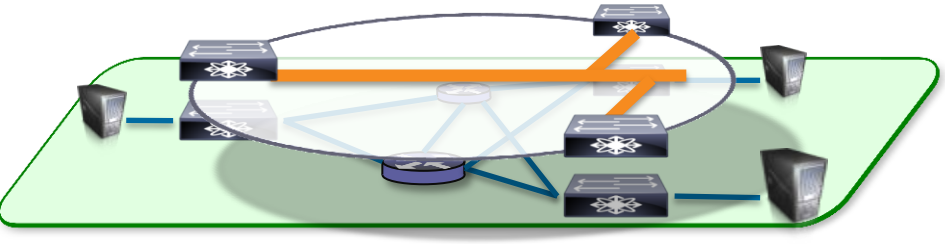


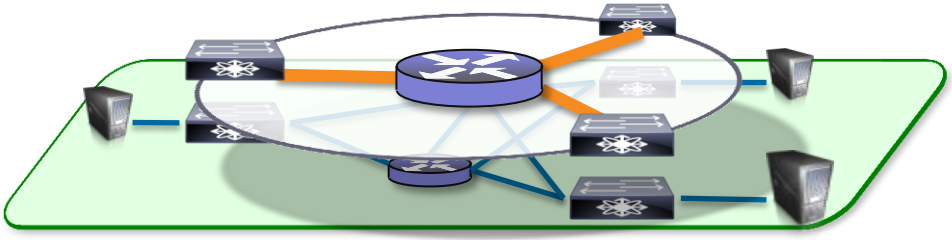
Overlay Networks





Types of Overlay Service





Layer 2 Overlays

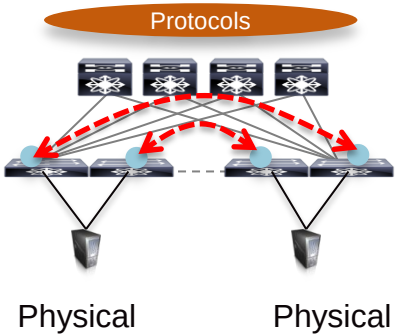
- Emulate a LAN segment
- Transport Ethernet Frames (IP and non-IP)
- Single subnet mobility (L2 domain)
- Exposure to open L2 flooding
- Useful in emulating physical topologies

Layer 3 Overlays

- Abstract IP based connectivity
- Transport IP Packets
- Full mobility regardless of subnets
- Contain network related failures (floods)
- Useful in abstracting connectivity and policy

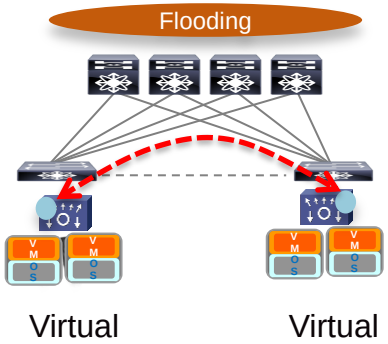
Types of Overlay Edge Devices

Network Overlays



- Router/switch end-points
- Protocols for resiliency/loops
- Traditional VPNs
- OTV, VPLS, LISP

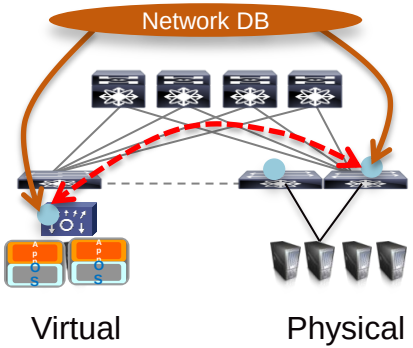
Host Overlays



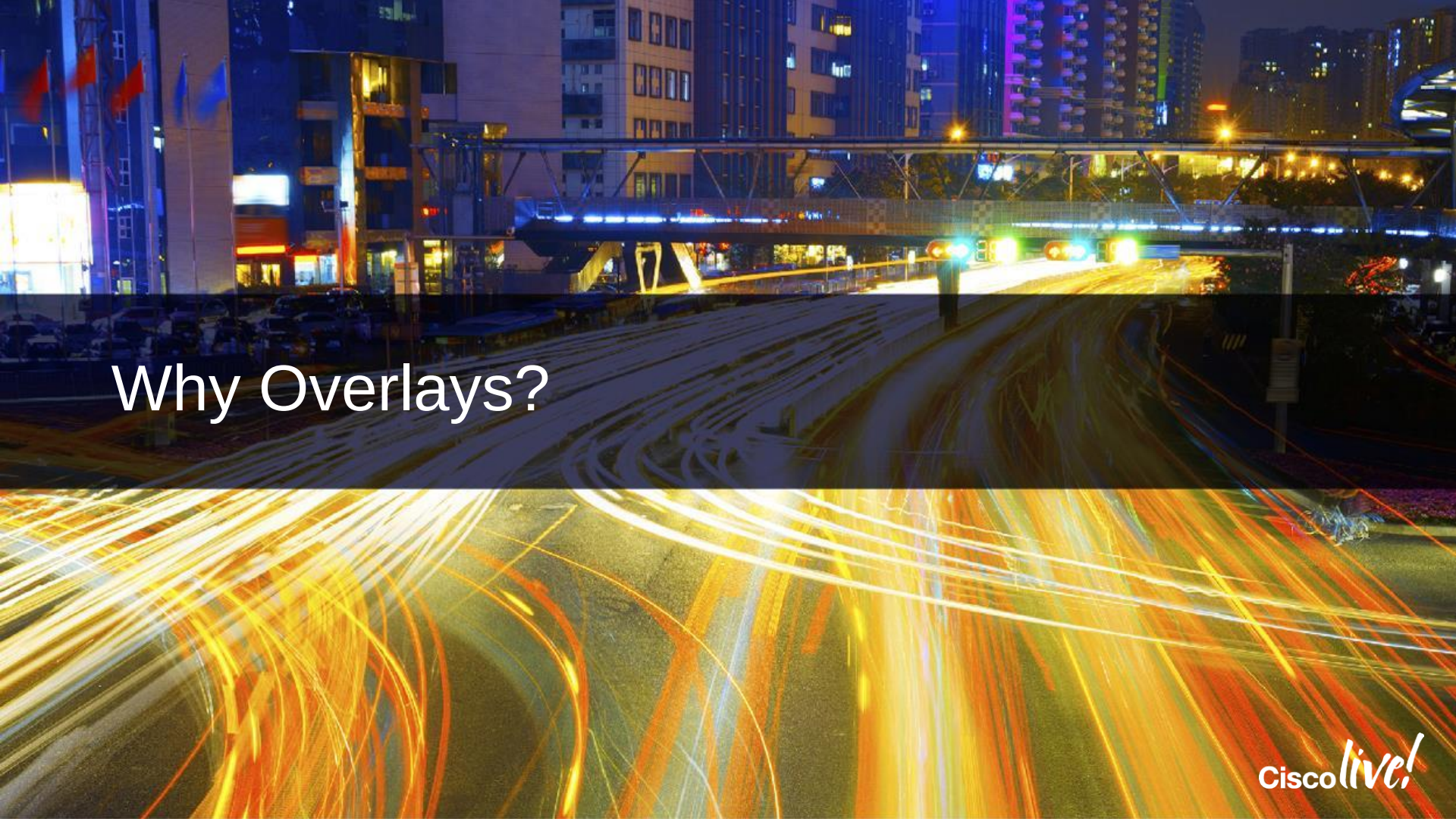
- Virtual end-points only
- Single admin domain
- VXLAN, NVGRE, STT

Tunnel End-points

Hybrid Overlays



- Physical and Virtual
- Resiliency + Scale
- x-organizations/federation
- Open Standards



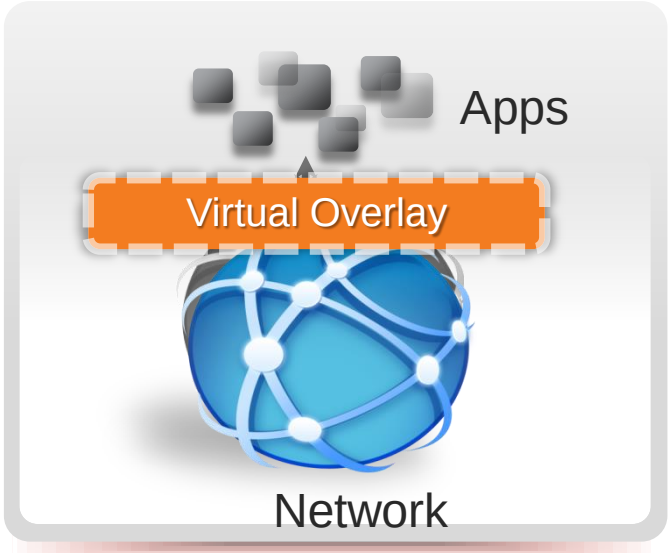
Why Overlays?

Why Use Overlay Networks?

- Simplified Workload Provisioning / Automation
  - Simplified deployment
  - Fast Provisioning of Virtual Workload by consuming the L2 network from network pool
  - Without changing the physical network (it's an Overlay!)
- Multitenant Scale
  - Provide Layer 2 networks for tenants, networks which can scale beyond 4K VLANs
- Workload anywhere (Mobility/Reachability)
  - Optimally use server resources by placing the workload anywhere
  - Yet provide Layer 2 connectivity to Workloads

Simplified Workload Provisioning / Automation

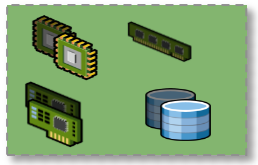
- Virtual Overlay solution providing fast provisioning of virtual workload by creating pool of overlay networks
- No change required in the physical network



Simplified Workload Provisioning / Automation

- Network Segment Pools are available to consume for cloud admin when deploying workload

Cloud Admin



Compute Resource Pool

Network Resource Pool

Cisco live!

Simplified Workload Provisioning / Automation

- No change required on physical network environment to provide network pool

Cloud Admin

Create VMs

Compute Resource Pool

Network Resource Pool

VM1

VM2

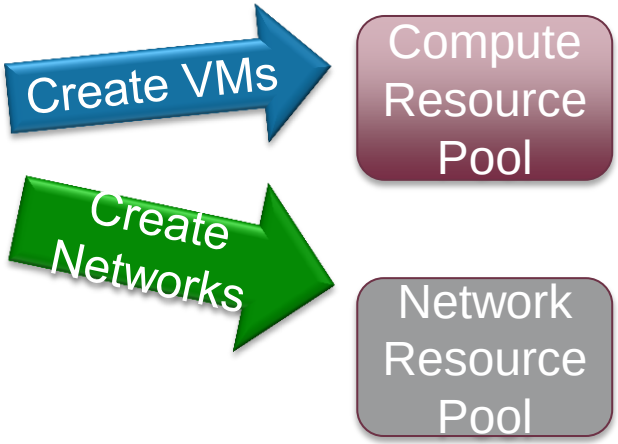
VM3

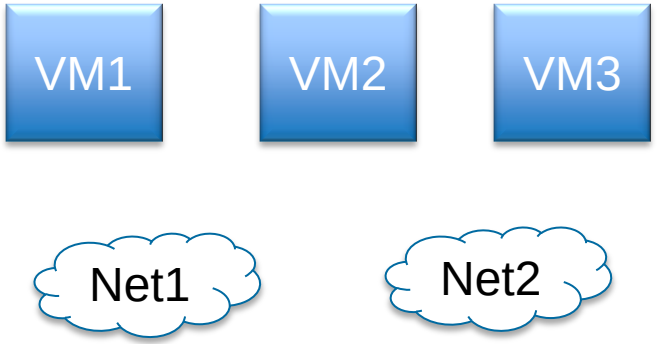
Cisco live!

Simplified Workload Provisioning / Automation

- No change required on physical network environment to provide network pool

Cloud Admin

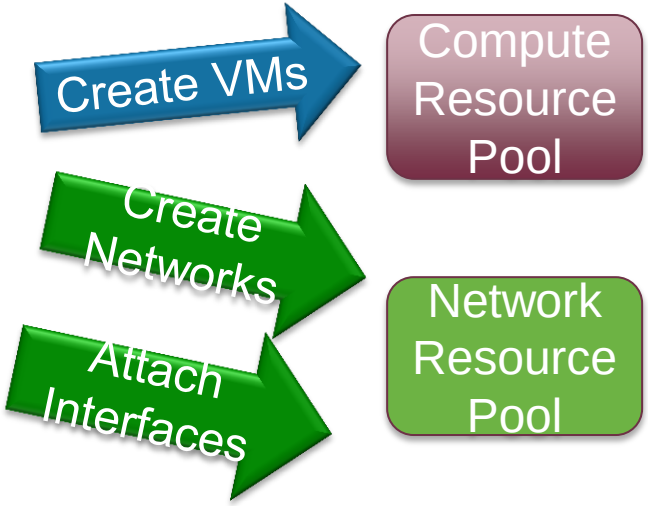


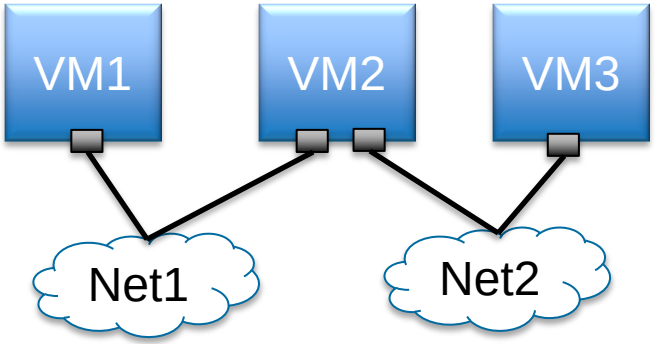


Simplified Workload Provisioning / Automation

- No change required on physical network environment to provide network pool with Overlay Networks

Cloud Admin

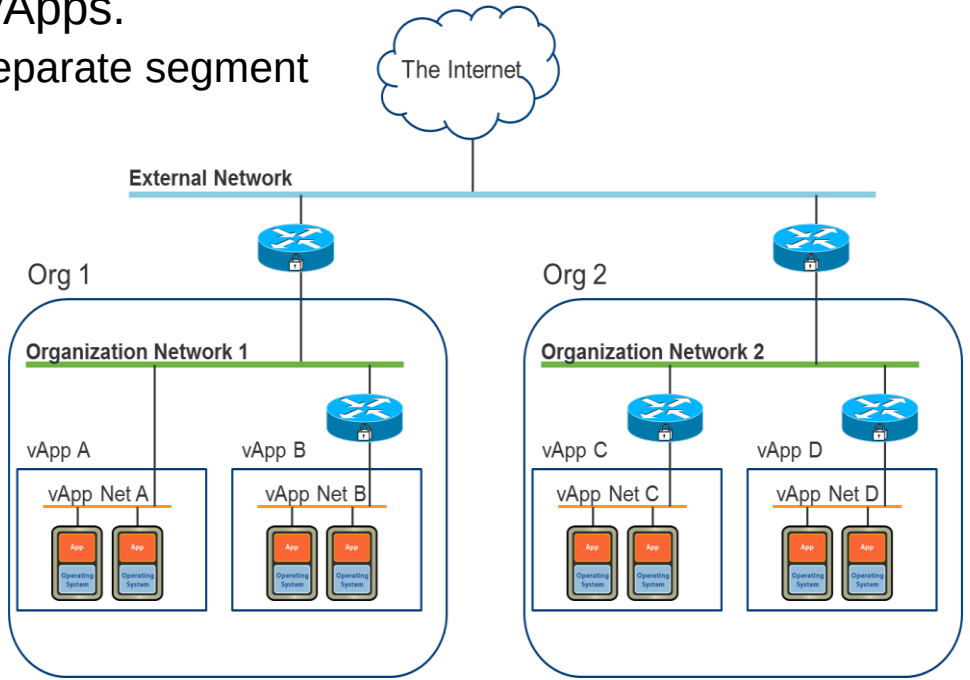




Multi-Tenancy and vApps Drive Layer2 Segments

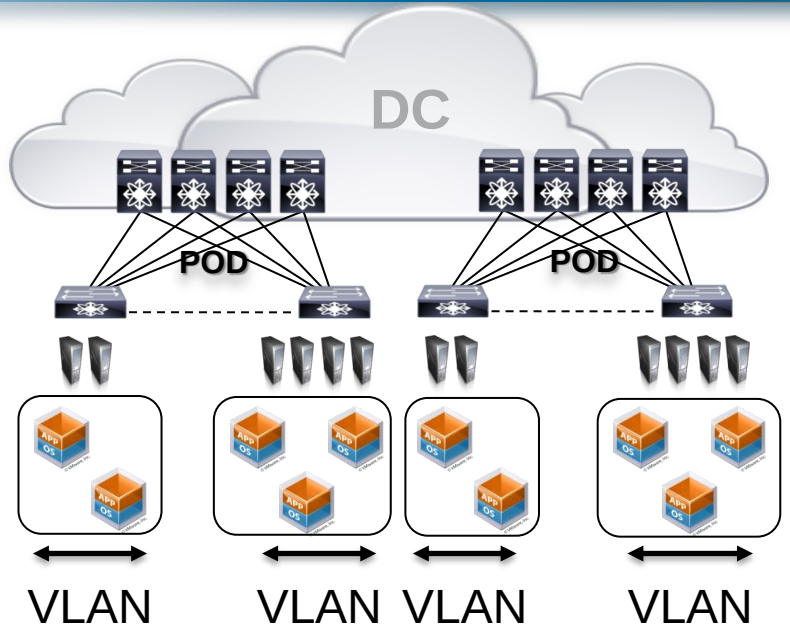
- Both MAC and IP addresses could overlap between two tenants, or even within the same tenant in different vApps.
  - Each overlapping address space needs a separate segment

- VLANs use 12 bit IDs = 4K
- VXLANs use 24 bit IDs = 16M
- NVGREs use 24 bit IDs = 16M
- STTs use 64 bit IDs

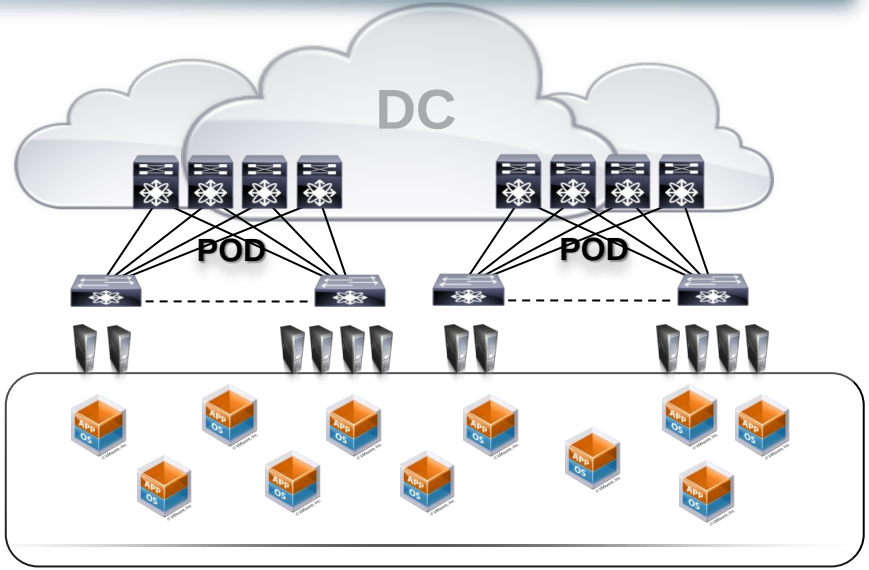


Workload Anywhere = Reachability/ Mobility

Rack-Wide VM Mobility



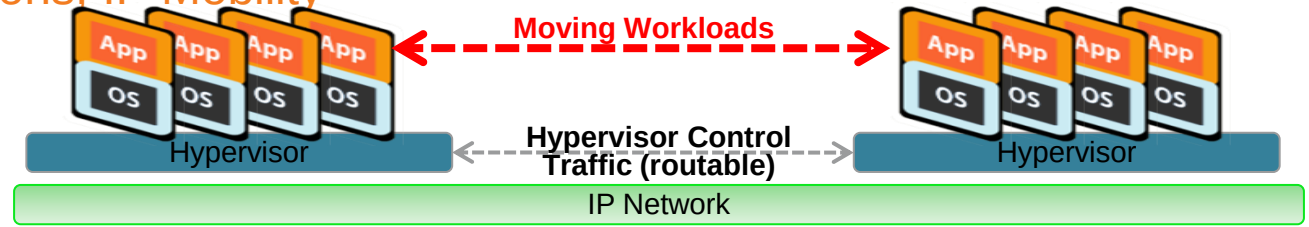
DC-Wide & Inter-DC VM Mobility



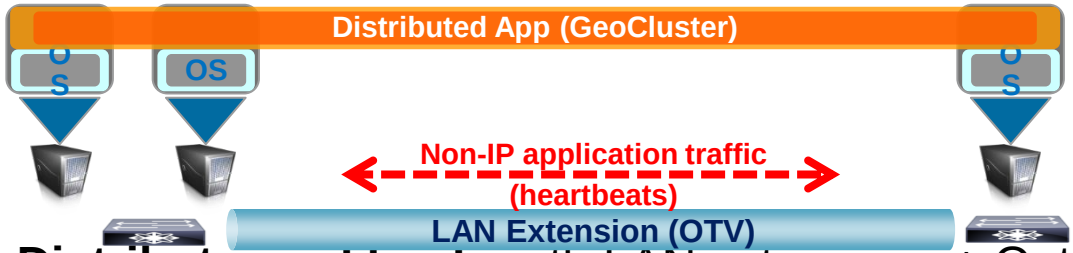
- Flexible Workload placement
- Network as a pooled resource
- Scalable multi-tenancy

Moving vs. Distributing Workloads

LAN Extensions, IP Mobility



- **Move workloads anywhere** with IP mobility solutions
  - –Can be achieved with LAN extensions
  - –IP preservation is the real requirement (LAN extensions not mandatory)



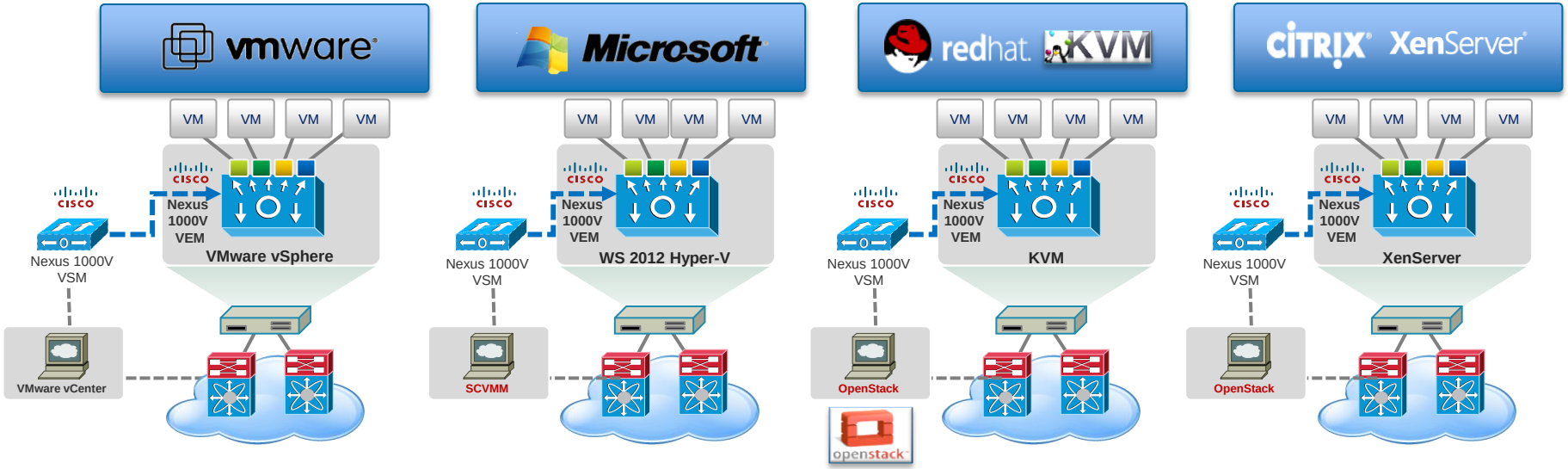
- **Distribute workloads** with LAN extensions + Optimal Routing
  - –Application High Availability with Distributed Clusters

VXLAN Overview

Cisco live!

Nexus 1000V Multi-Hypervisor Platform

- Example: Virtual Overlay Networks and Services with Nexus 1000V

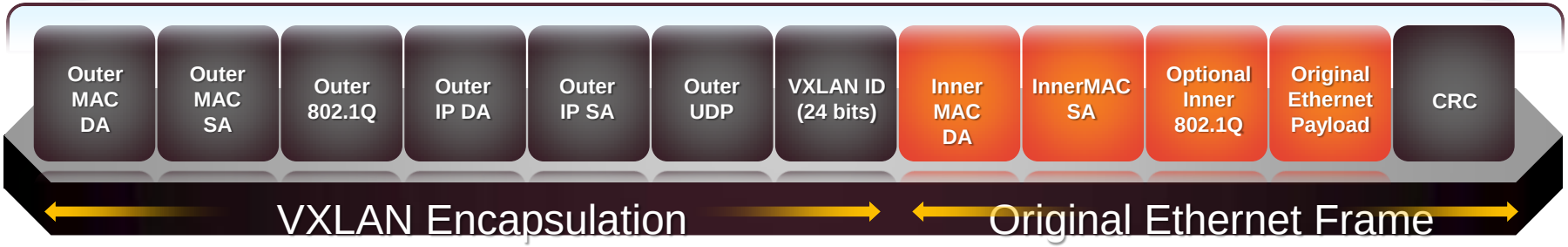


Consistent architecture, feature-set & network services ensures operational transparency across multiple hypervisors.

Virtual Extensible Local Area Network (VXLAN)

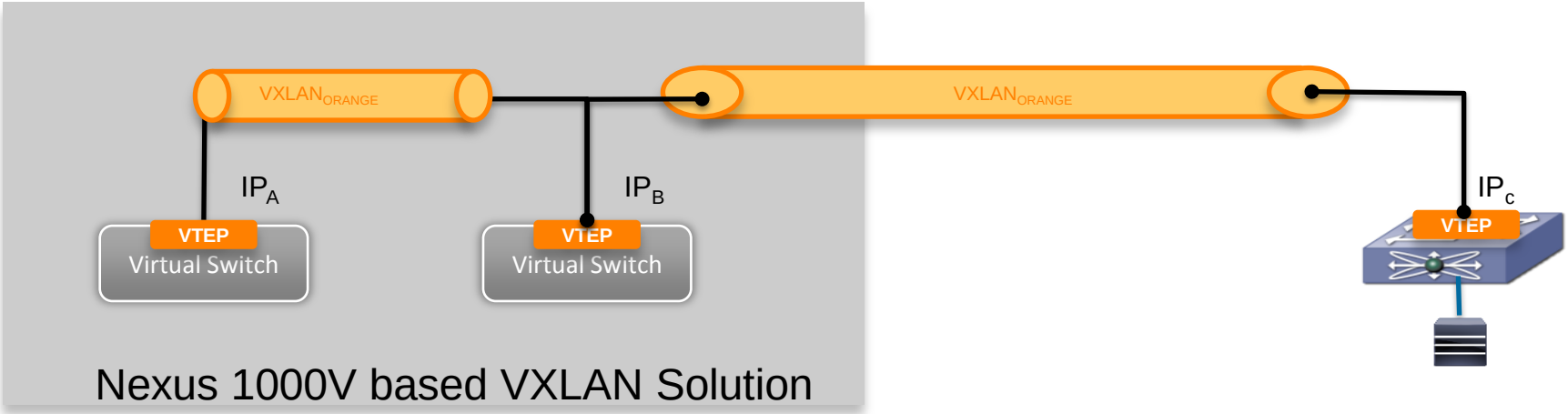
- Ethernet in IP overlay network
  - Entire Layer 2 frame encapsulated in User Datagram Protocol (UDP)
  - 50 bytes of overhead
- Include 24-bit VXLAN identifier
  - 16 M logical networks
  - Mapped into local bridge domains
- VXLAN can cross Layer 3

- Tunnel Between VEMs
  - VMs do not see VXLAN ID
- IP multicast used for Layer 2 broadcast or multicast, and unknown unicast
- Technology submitted to IETF for standardization
  - With VMware, Citrix, Red Hat, and Others



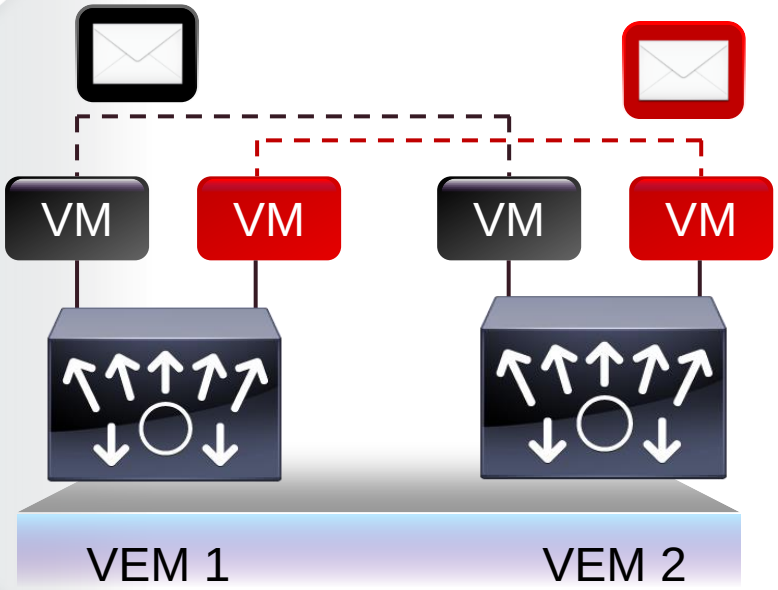
VXLAN Tunnel End Point (VTEP)

- VTEP is the IP address which defines the **Tunnel End Point**
- VTEP represents source and destination IP address on a VXLAN encapsulated tunnel
- VTEP would reside on a switch (physical or virtual) which would perform the VXLAN encapsulation



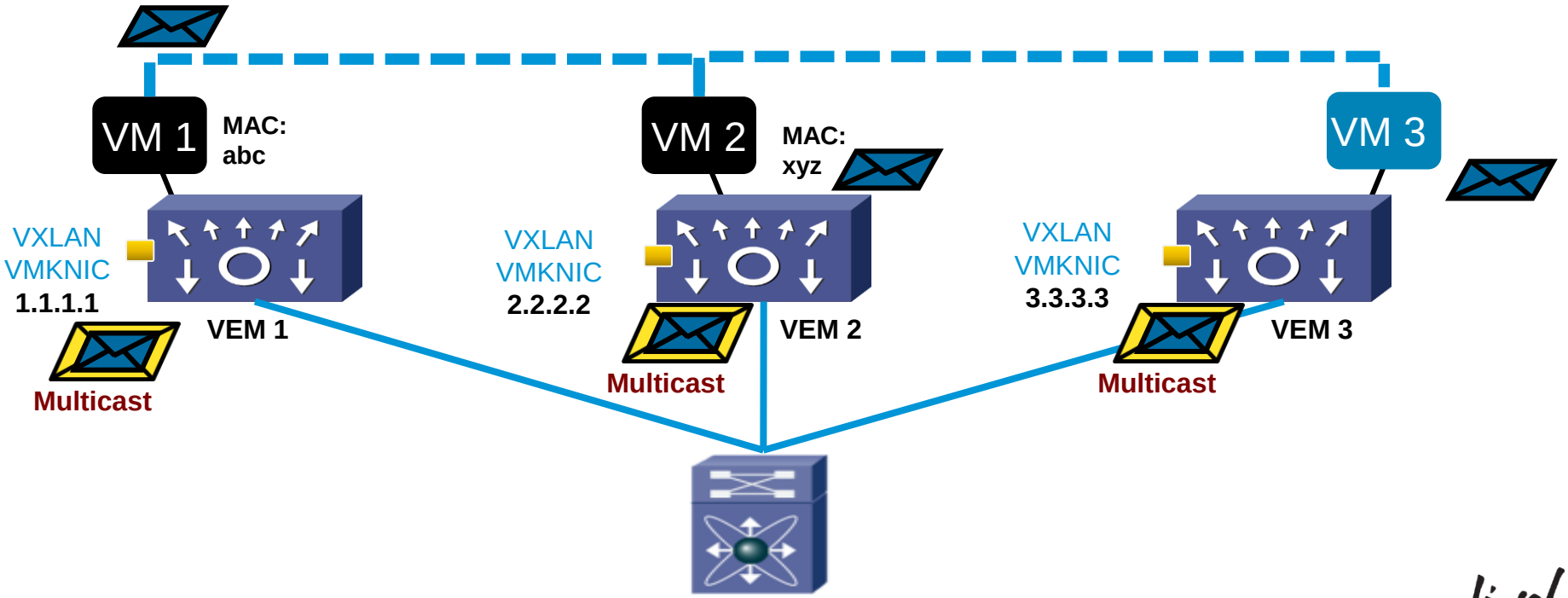
VXLAN Forwarding Basics

- Forwarding mechanisms similar to Layer 2 bridge: Flood and learn
  - VEM learns VM's source (MAC, host VXLAN IP) tuple
- Broadcast, multicast, and unknown unicast traffic
  - VM broadcast and unknown unicast traffic are sent as multicast
- Unicast
  - Unicast packets are encapsulated and sent directly (not through multicast) to destination host VXLAN IP (destination VEM)



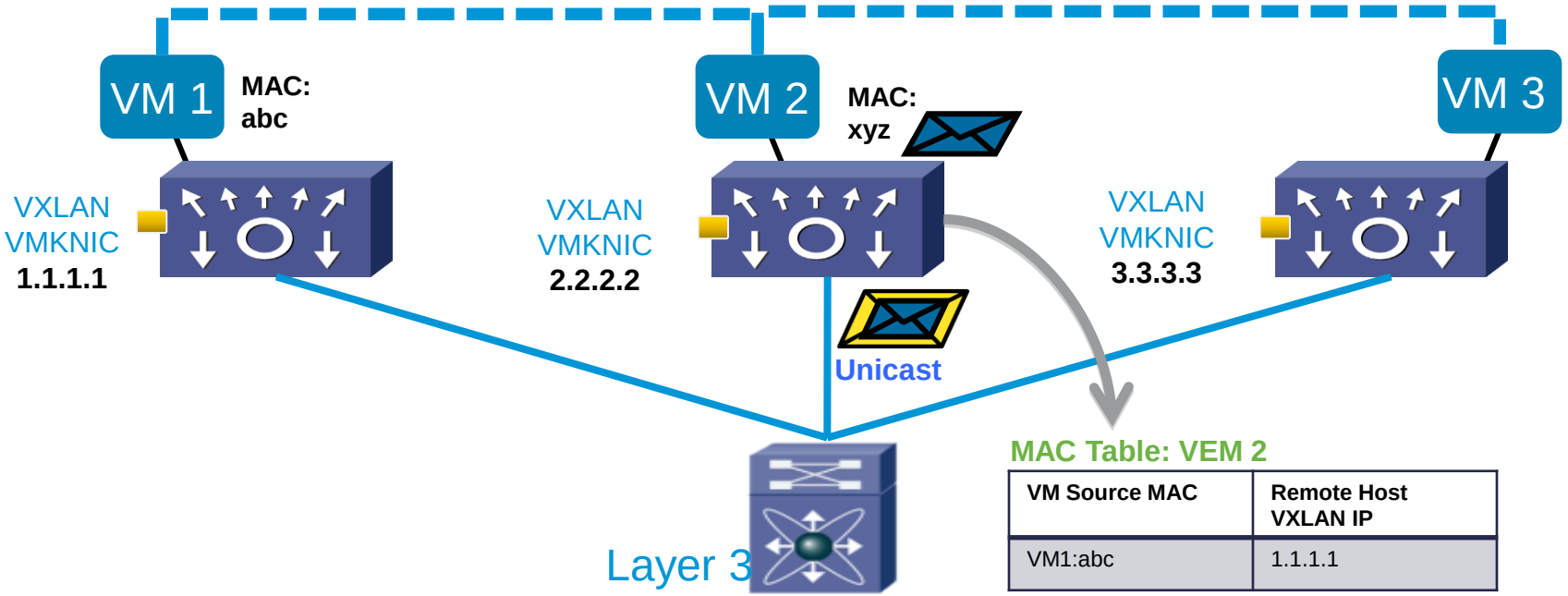
VXLAN Example Data Flow

VM1 Communicating with VM2 in a VXLAN



VXLAN Example Data Flow

VM1 Communicating with VM2 in a VXLAN

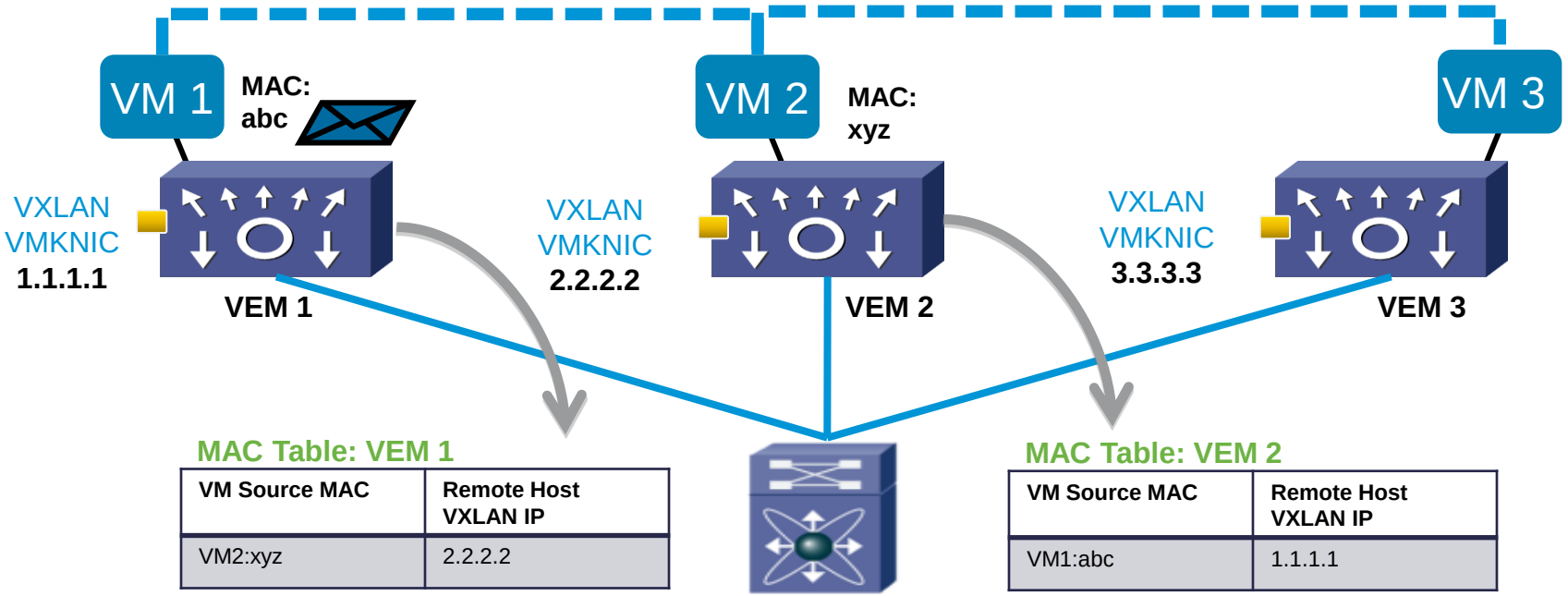


MAC Table: VEM 2

VM Source MAC	Remote Host VXLAN IP
VM1:abc	1.1.1.1

VXLAN Example Data Flow

VM1 Communicating with VM2 in a VXLAN



MAC Table: VEM 1

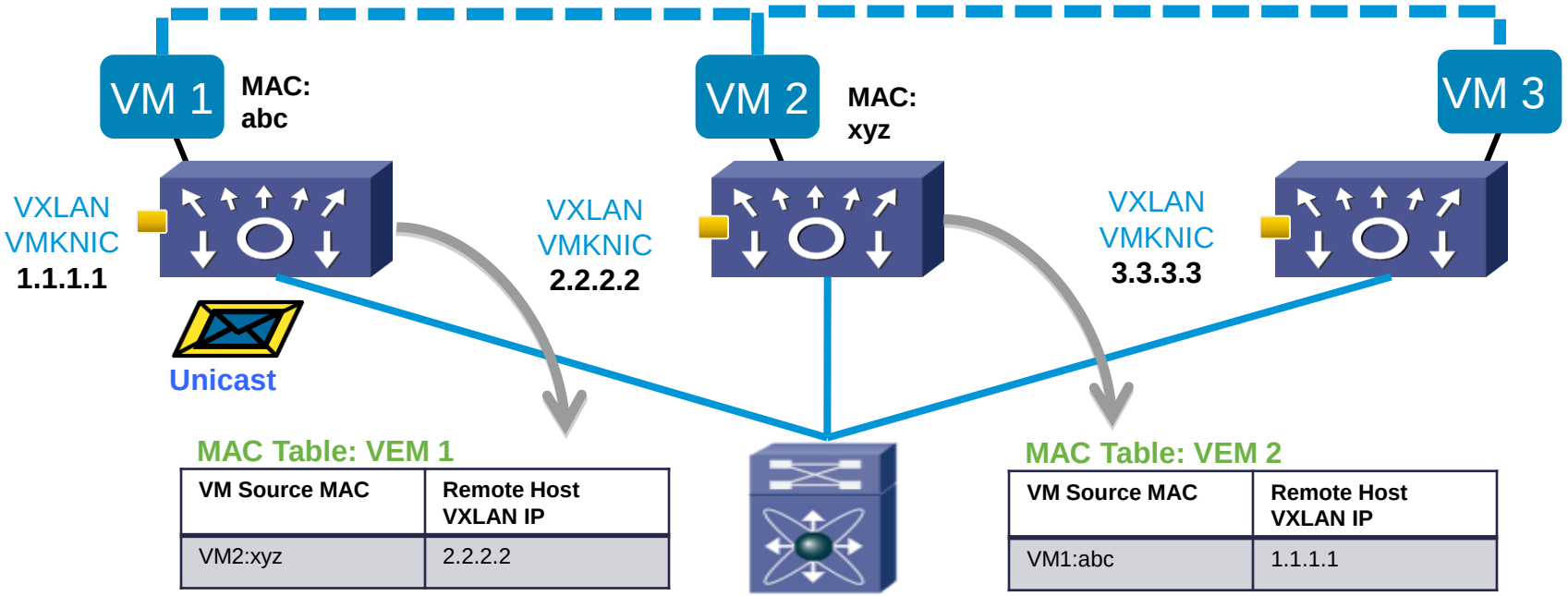
VM Source MAC	Remote Host VXLAN IP
VM2:xyz	2.2.2.2

MAC Table: VEM 2

VM Source MAC	Remote Host VXLAN IP
VM1:abc	1.1.1.1

VXLAN Example Data Flow

VM1 Communicating with VM2 in a VXLAN



MAC Table: VEM 1

VM Source MAC	Remote Host VXLAN IP
VM2:xyz	2.2.2.2

MAC Table: VEM 2

VM Source MAC	Remote Host VXLAN IP
VM1:abc	1.1.1.1

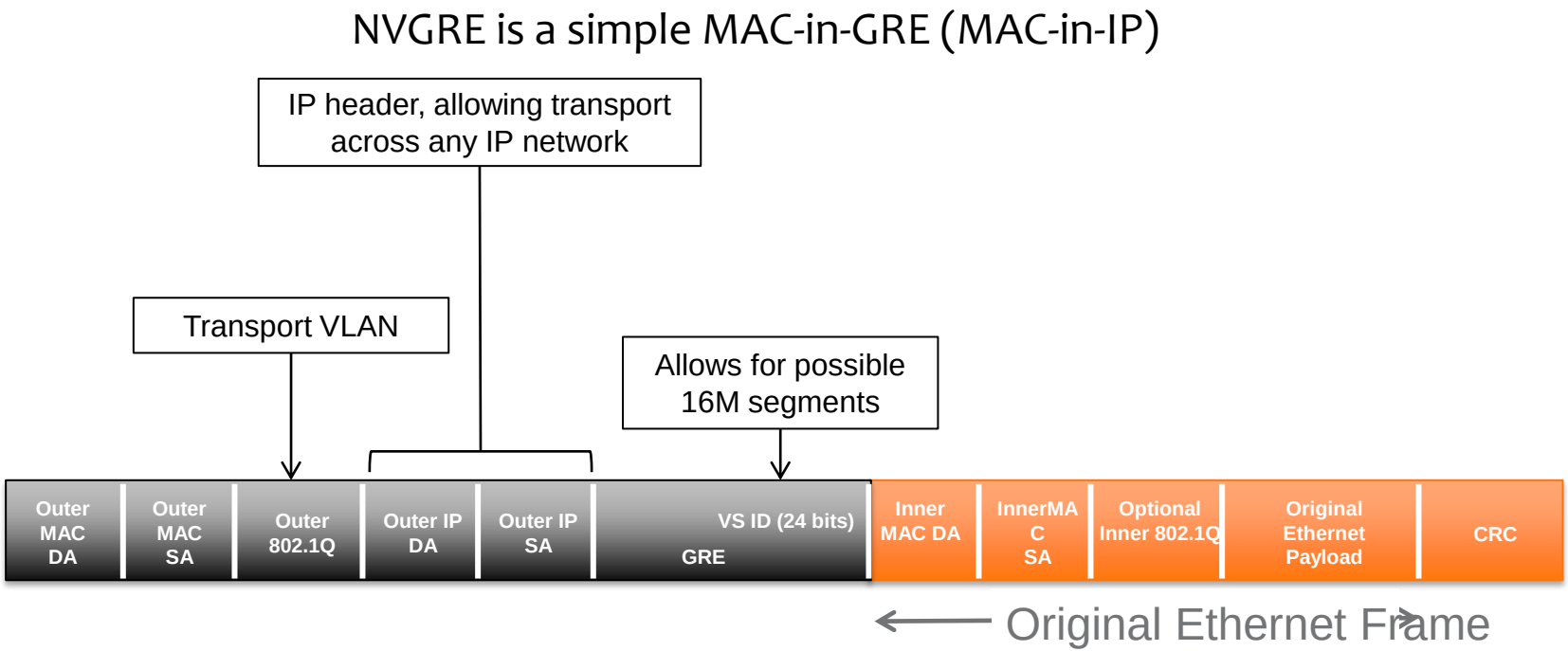
Overlay Technologies and VXLAN Overlay Comparisons

Cisco live!

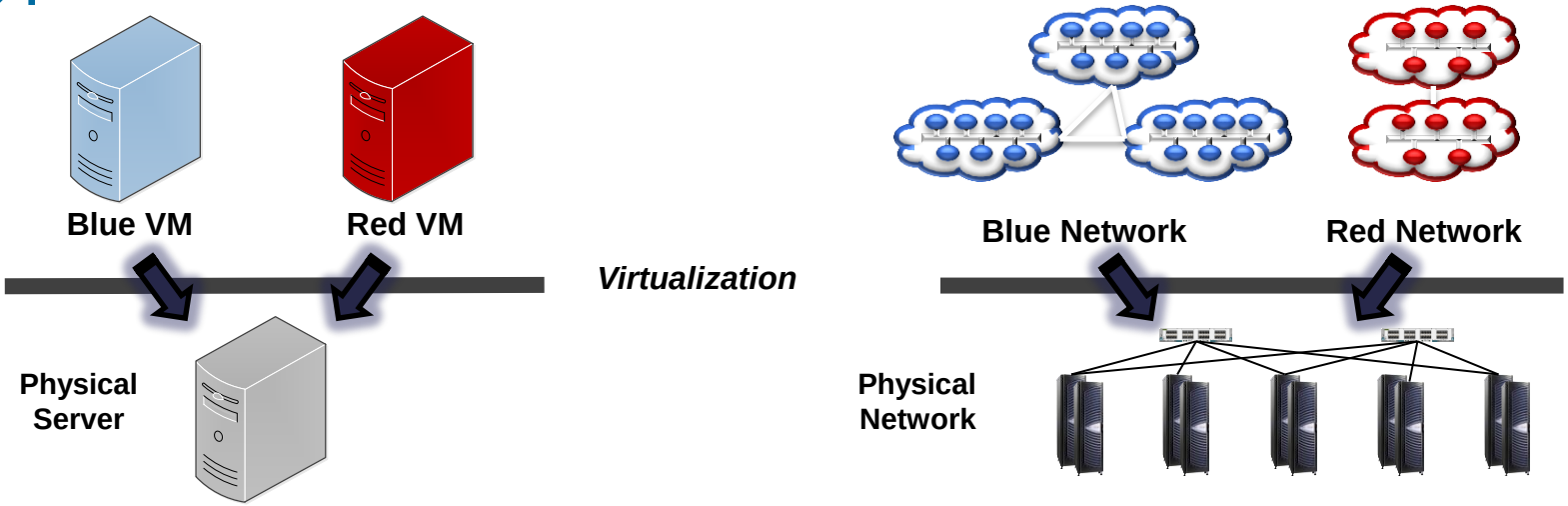
Network Virtualization over GRE (NVGRE)

- L2 frame encapsulated in GRE 50 bytes of overhead
- Include 24 bit VSID Identifier 16 M logical networks
- NVGRE can cross Layer 3
- VMs do NOT see VSID
- Technology submitted to IETF for standardization With Microsoft, Arista, Intel, Dell, HP, Broadcom and Emulex

NVGRE Frame Format



Hyper-V Network Virtualization



Server Virtualization

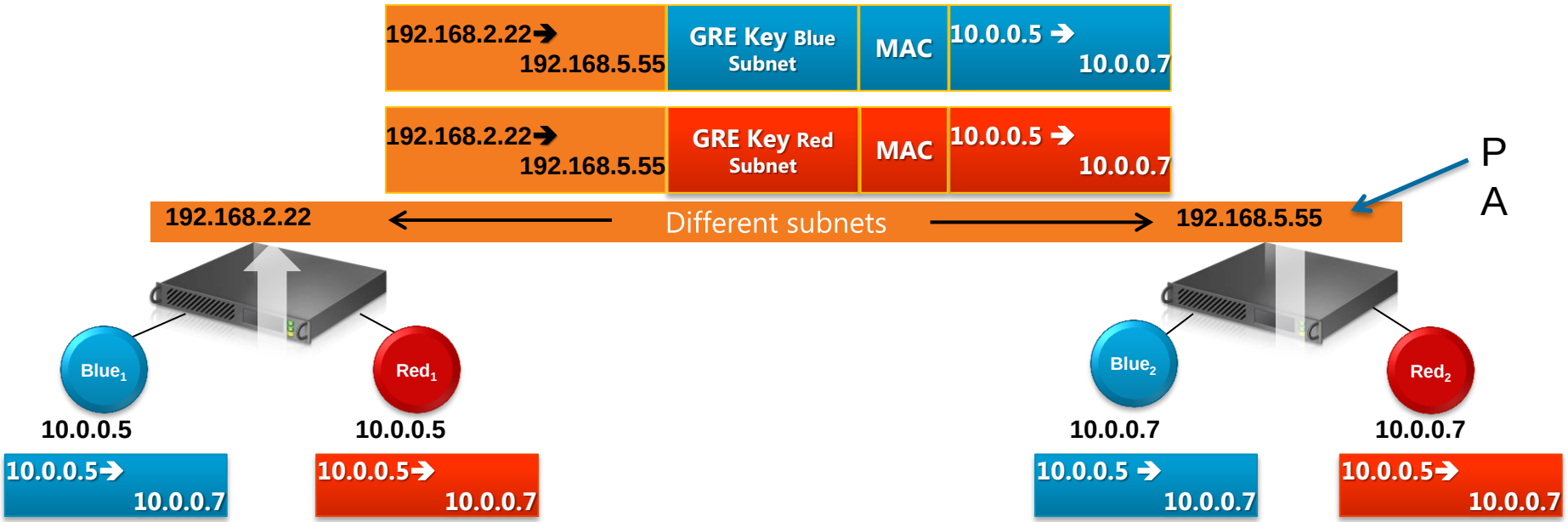
- Run multiple virtual servers on a physical server
- Each VM has illusion it is running as a physical server

Hyper-V Network Virtualization

- Run multiple virtual networks on a physical network
- Each virtual network has illusion it is running as a physical network

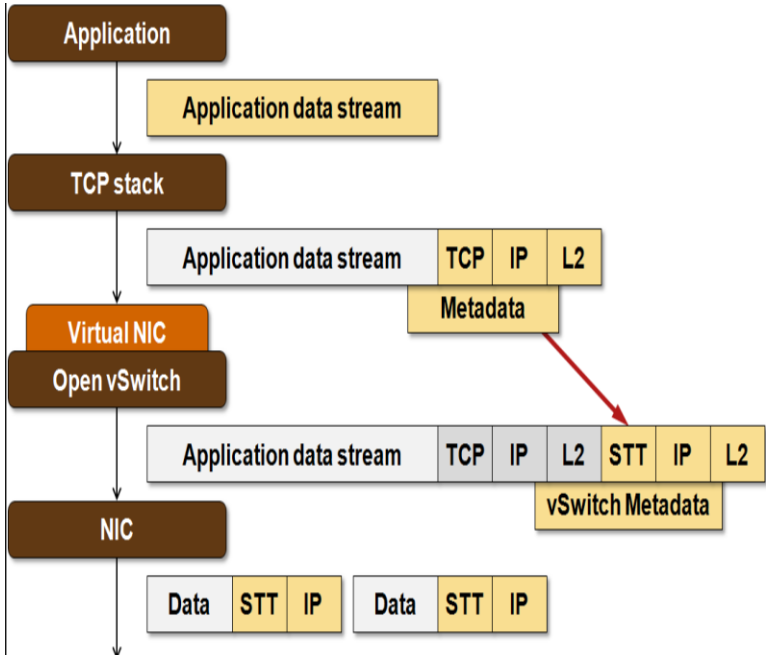
NVGRE Packet Forwarding

- Better network scalability by sharing Provider Address (PA) among VMs
- Explicit Virtual Subnet ID for better multi-tenancy support



Stateless Transport Tunneling (STT) Overview

- L2 encapsulation over IP/TCP
- Addressing the same use cases as VXLAN/ NVGRE Based Host Overlay solution are providing
- STT is explicitly designed to leverage TCP Segment Offload (TSO) capabilities of currently available NICs
- NSX Solution



VXLAN Versus NVGRE

Similarities

- Both are Overlay approach to Network Virtualization
- IP Transport (L2 Encap over L3 Network)
- 24 Bit Segment ID

Differences

- Encapsulation
  - VXLAN: UDP NVGRE: GRE
- Port Channel Load Distribution
  - UDP: 5 Tuple Hashing
  - Most (if not all) switches do not hash on GRE header
- Firewall ACL can act on VXLAN UDP port
- Forwarding Logic
  - VXLAN: Flooding/Learning (Not with enhanced VXLAN)
  - NVGRE: SCVMM acting as a controller (IP based Forwarding)

VXLAN Versus STT

Similarities

- Both carry Ethernet Frames
- Both use IP Transport
- Both can use IP Multicast
  - For broadcast and multicast frames
- Port Channel Load Distribution

  5 Tuple Hashing (UDP vs TCP)

Differences

- Encapsulation
  - VXLAN: UDP with 50 bytes
  - STT: "TCP-like" with 76 to 58 bytes (not uniform) *
- Segment ID Size
  - VXLAN: 24 bit
  - STT: 64 bit
- Firewall ACL can act on VXLAN UDP port
  - Firewalls will likely block STT since it has no TCP state machine handshake
- Forwarding Logic
  - VXLAN: Flooding/Learning
  - STT: Controller-based

Note: STT uses the TCP header, but not the protocol state machine. TCP header fields are repurposed.
* The STT header does not exist in every packet. Only the first packet of a large segment, therefore reassembly is required.

VXLAN Versus OTV

Overlay Transport Virtualization

Similarities

- Both carry Ethernet frames
- Same UDP based encapsulation header
  - VXLAN does not use the OTV Overlay ID field
- Both can use IP Multicast
  - For broadcast and multicast frames (optional for OTV)

Differences

- Forwarding Logic
  - VXLAN: Flooding/Learning
  - OTV: Uses the IS-IS protocol to advertise the MAC address to IP bindings
- OTV can locally terminate ARP and doesn't flood unknown MACs
- OTV can use an adjacency server to eliminate the need for IP multicast
- OTV is optimized for Data Center Interconnect to extend VLANs between or across data centers
- VXLAN is optimized for intra-DC and multi-tenancy

VXLAN Versus LISP

Locator / ID Separation Protocol

Similarities

- Same UDP based encapsulation header
  - VXLAN does not use the control flag bits or Nonce/Map-Version field
- 24 Bit Segment ID

Differences

- LISP carries IP packets, while VXLAN carries Ethernet frames
- Forwarding Logic
  - VXLAN: Flooding/Learning
  - LISP: Uses a mapping system to register/resolve inner IP to outer IP mappings
- For LISP, IP Multicast is only required to carry host IP multicast traffic
- LISP is designed to give IP address (Identifier) mobility / multi-homing and IP core route scalability
- LISP can provide optimal traffic routing when Identifier IP addresses move to a different location

VXLAN Enhancements

Cisco live!

Current VXLAN Challenges

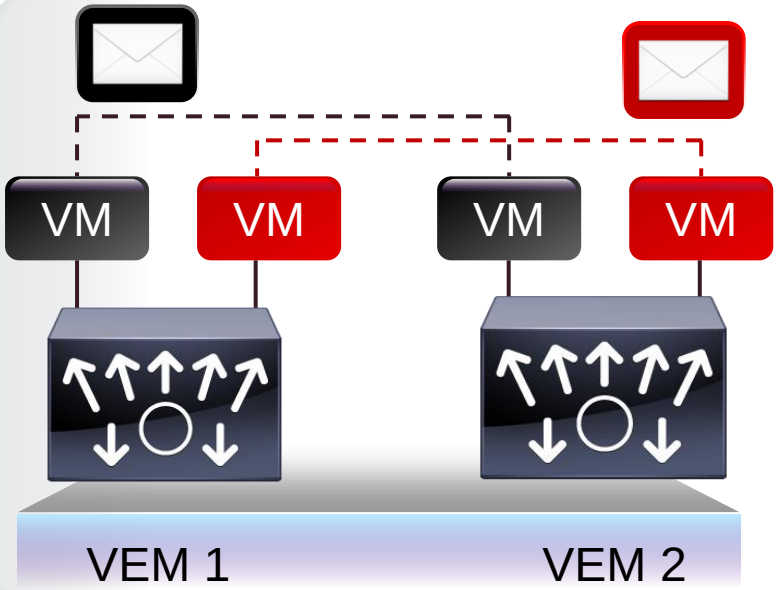
Multicast Dependency	• Multicast may not be enabled in the infrastructure • Multicast scaling
Flood and Learn based Learning	• Flooding required to handle BUM (Broadcast/Unknown Unicast/Multicast) traffic • Unknown floods can cause network meltdowns
External Connectivity	• Need the ability to connect to external nodes

VXLAN Enhancements

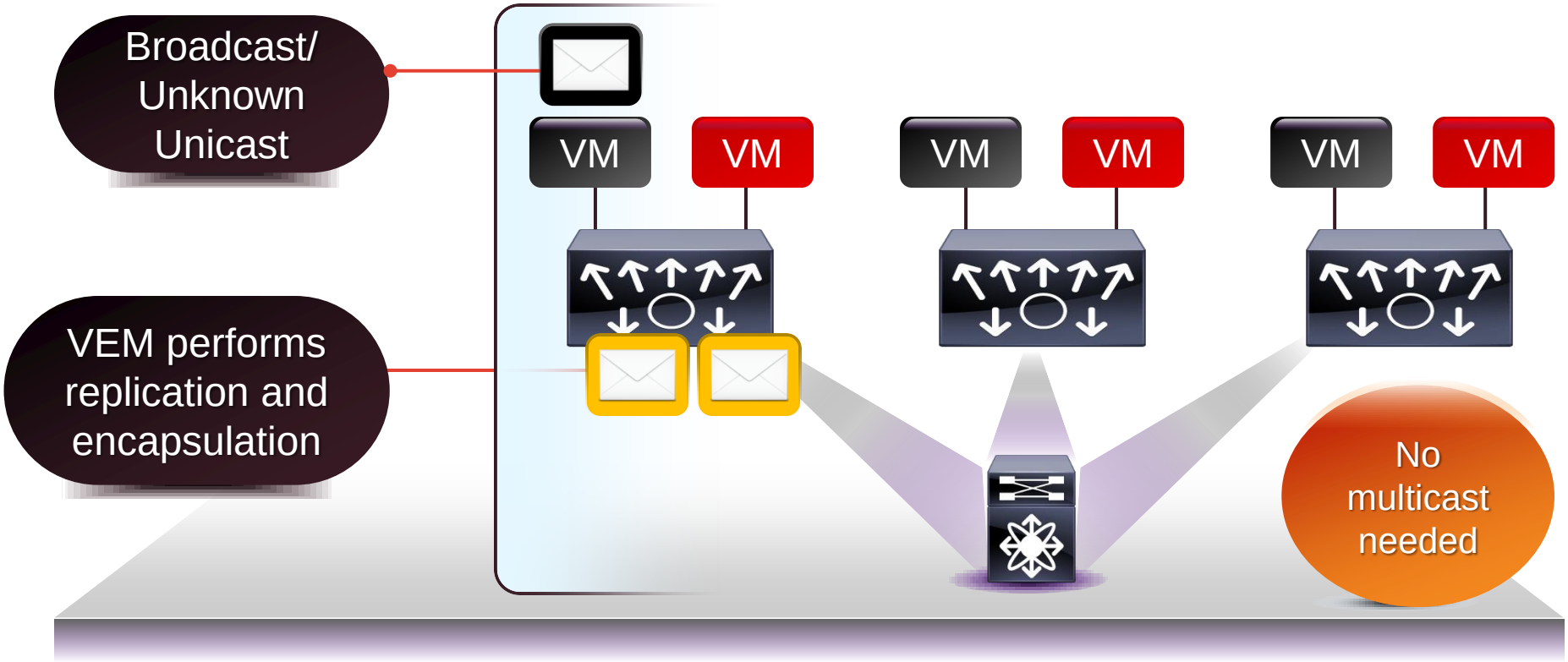
Multicast Dependency	• Head-end replication to allow unicast-mode only operation • Introduce a control plane to allow for dynamic VTEP discovery
Flood and Learn based Learning	• Workload MAC addresses are known once they are connected to the VXLAN capable devices • Leverage the control plane also to exchange L2/L3 address-to-VTEP association information
External Connectivity	• Introduce VXLAN Gateways

Enhanced VXLAN - Forwarding Basics

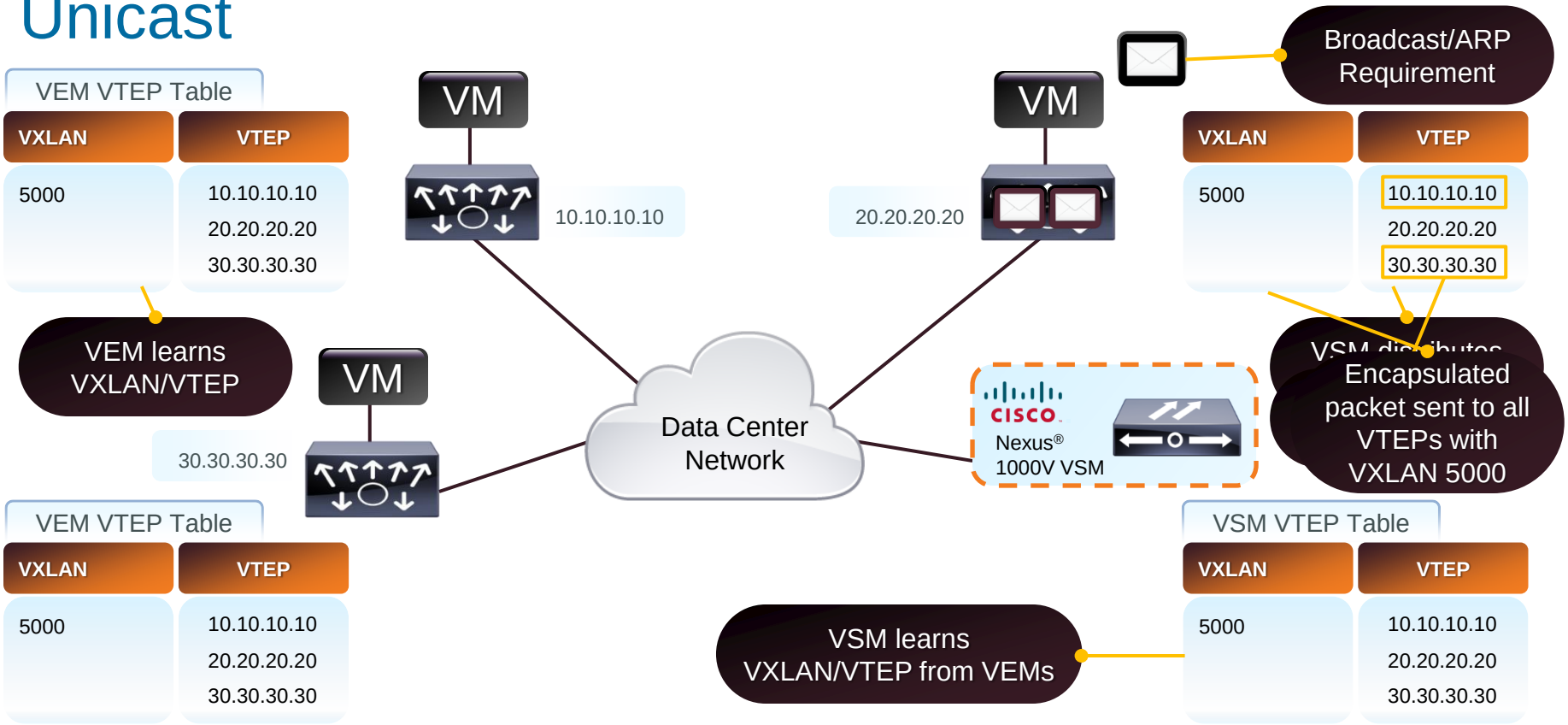
- Forwarding mechanisms similar to Layer 2 bridge: Flood and learn
  - VEM learns VM's source (MAC, host) tuple
- Broadcast, multicast, and unknown unicast traffic
  - VM BUM traffic is replicated for each host having VMs in same VXLAN. Packet is encapsulated with destination IP set to the host's VXLAN IP.
- Unicast User
  - Unicast packets are encapsulated and sent directly (not through multicast) to destination host VXLAN IP (destination VEM)



Broadcast and Unknown Unicast in Enhanced VXLAN



Enhanced VXLAN - Broadcast, Multicast, Unknown Unicast



Enhanced VXLAN – Forwarding Enhancements

MAC Distribution

Security enhancement that prevents malicious VMs from causing ”unknown unicast” broadcast storms

VEM learns all (VXLAN, MAC) from VSM

When VEM receives a MAC from VM in a VXLAN, if MAC is not found in the MAC table the frame is dropped.

Local ARP Termination

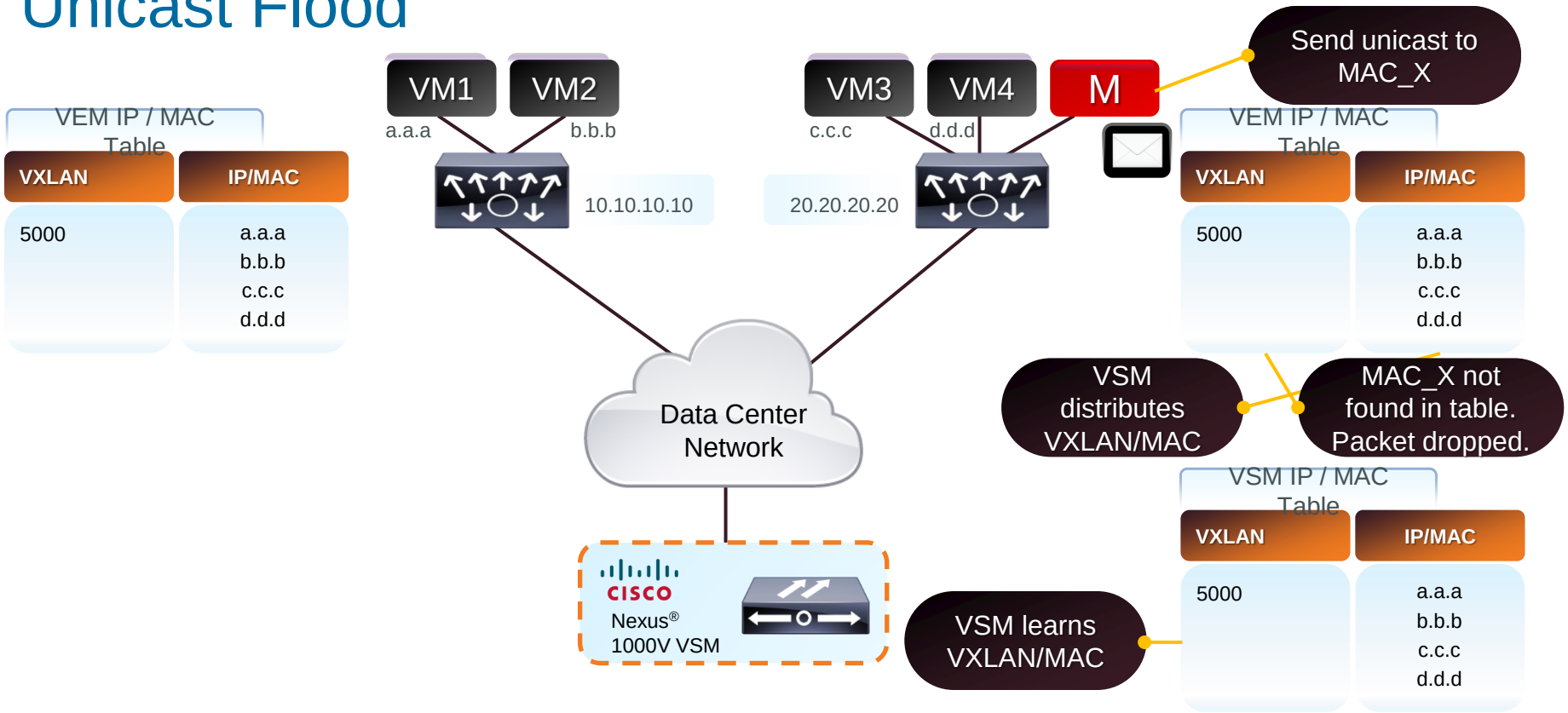
VEM terminates ARP locally for VMs in VXLAN reducing ARP broadcast traffic

VSM aggregates and distributes (VXLAN, IP, MAC) entries to VEMs

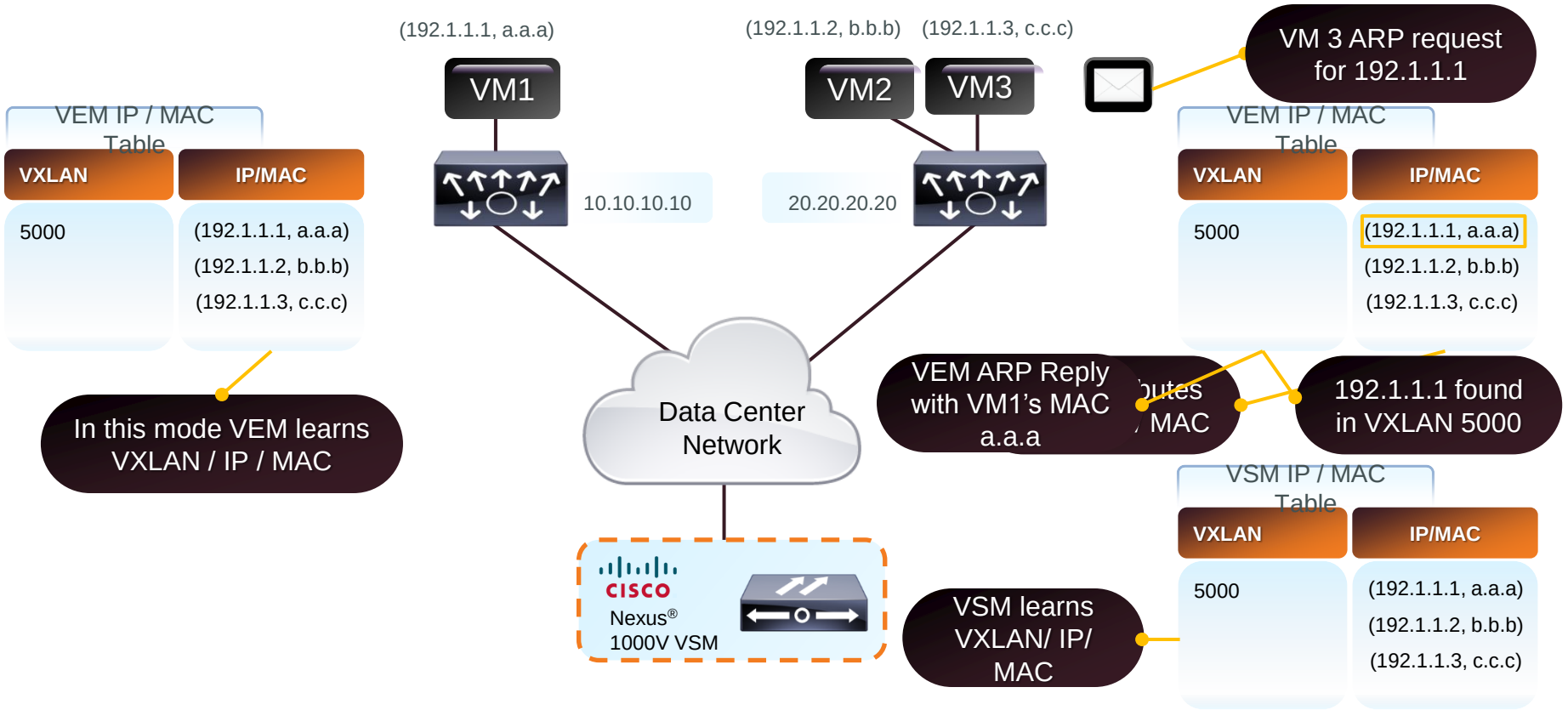
When VEM receives an ARP request, VEM looks up the MAC/IPDB for MAC address of host

VEM replies to ARP request with MAC address of the destination VM

VXLAN MAC Distribution – Prevents Unknown Unicast Flood



VXLAN ARP Termination – Reduces ARP broadcast



VXLAN Encapsulations

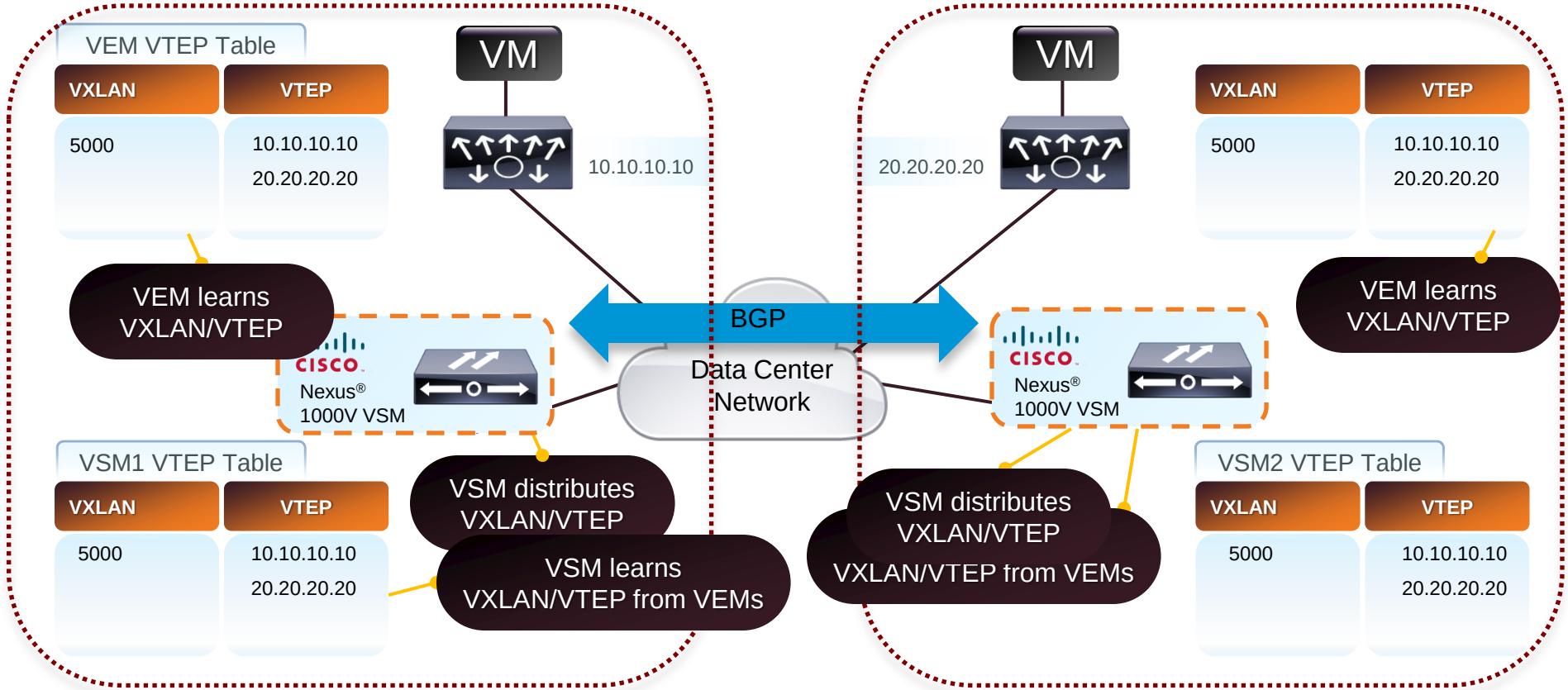
Packet \ VXLAN Mode	VXLAN (multicast mode)	Enhanced VXLAN (unicast mode)	Enhanced VXLAN MAC Distribution	Enhanced VXLAN ARP Termination
Broadcast / Multicast	Multicast Encapsulation	Replication plus Unicast Encap	Replication plus Unicast Encap	Replication plus Unicast Encap
Unknown Unicast	Multicast Encapsulation	Replication plus Unicast Encap	Drop	Drop
Known Unicast	Unicast Encapsulation	Unicast Encap	Unicast Encap	Unicast Encap
ARP	Multicast Encapsulation	Replication plus Unicast Encap	Replication plus Unicast Encap	VEM ARP Reply

For your reference

Extending VXLAN to multiple domains using MP-BGP

- Using MP-BGP to distribute overlay information
  - Use MP-BGP with EVPN Address Family for Tunnel Endpoint Discovery and Host Reachability
  - Distributed protocol – no single point of failure
  - Proven BGP scalability
  - Physical and virtual endpoints can become BGP peers to participate
- Why not use a Central Controller?
  - Single point of control could become a single point of failure
  - Scalability
  - Scope of the controller is limited to devices that understand the communication protocol used by the controller

VXLAN BGP Control-Plane – VTEP Distribution



VXLAN Gateway

VXLAN Gateway Use Cases

1. Virtualized WebServer VM communicating to Bare Metal DB Server

2. Data Center Services such as Firewall, WAN Accelerator deployed as Physical Boxes or Service Modules on Aggregation Switch

VXLAN Gateway on Nexus 1010/1110

What?

- A Layer 2 Gateway that extends the VXLAN Layer 2 domain to physical servers and services deployed on a VLAN

When?

- Created when a Layer 2 adjacency is required between VMs on a VXLAN and physical servers / services on a VLAN

How?

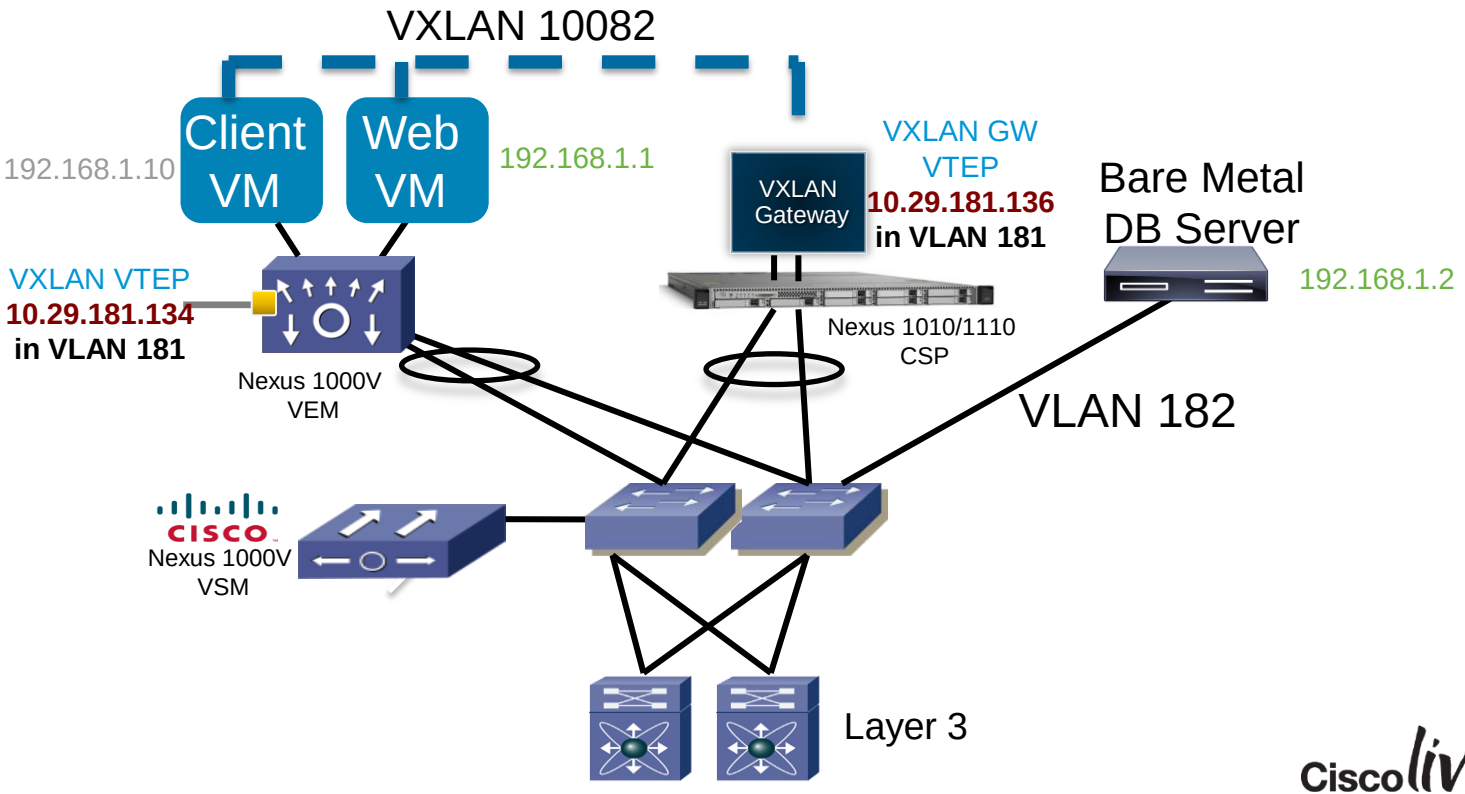
- The VXLAN gateway is managed as a VEM from the Nexus 1000V VSM
- VXLAN Gateway is a feature of the Advanced Edition, no additional cost
- Define a mapping between a VXLAN and VLAN on VSM

Where?

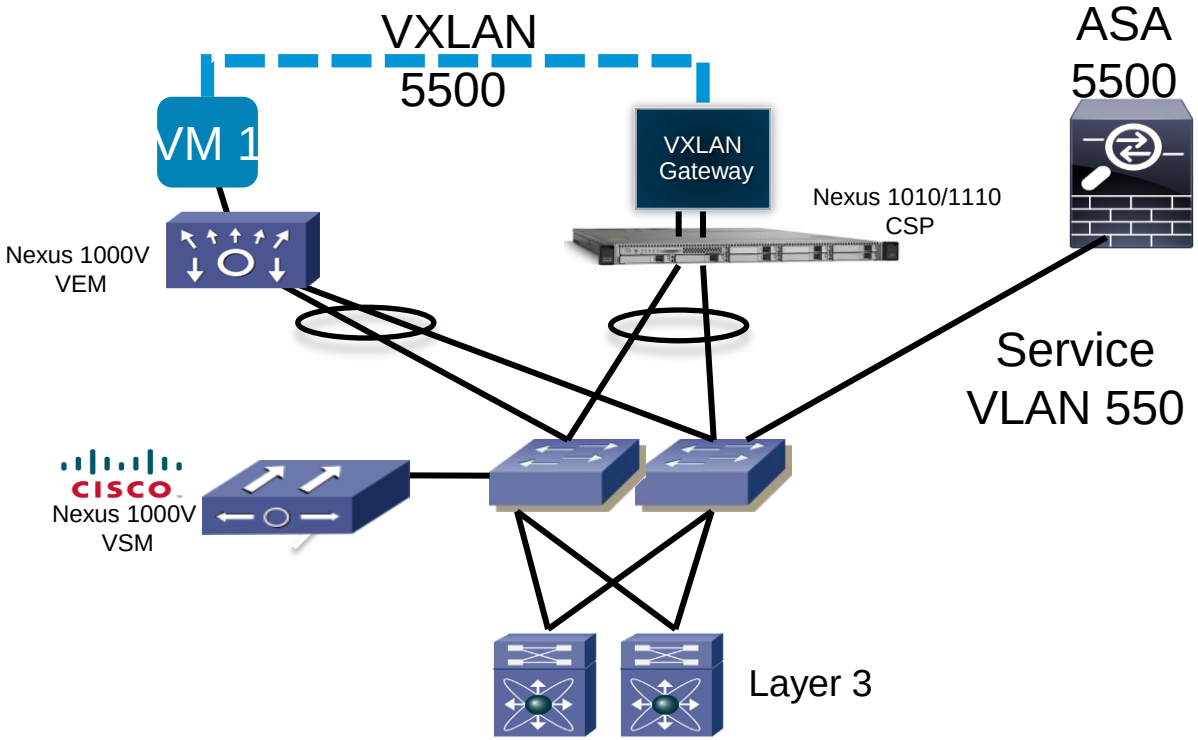
Nexus 1110 Nexus 1010

VXLAN Gateway is created as Virtual Service Blade on the Nexus 1110/1010

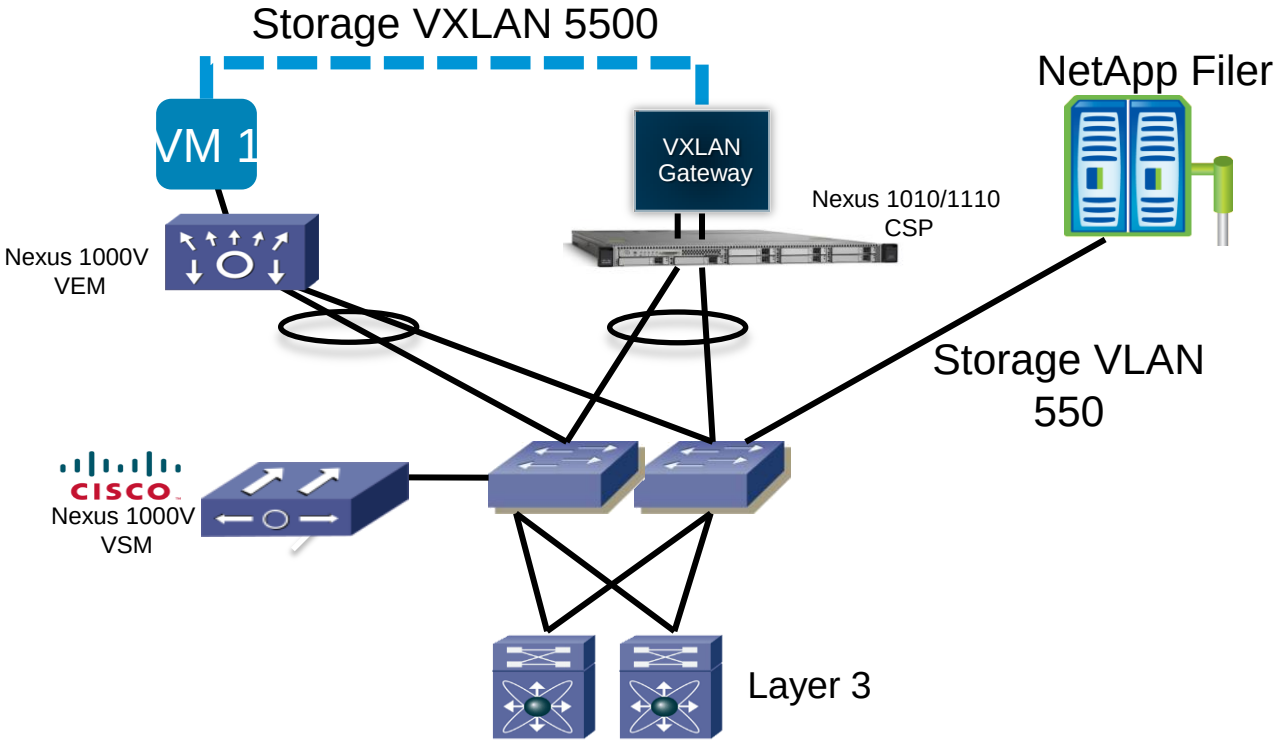
Use Case 1 – Connecting virtual and physical workloads on the same Layer 2 Segment



Use Case 2 – Connecting physical data center services to virtual workloads



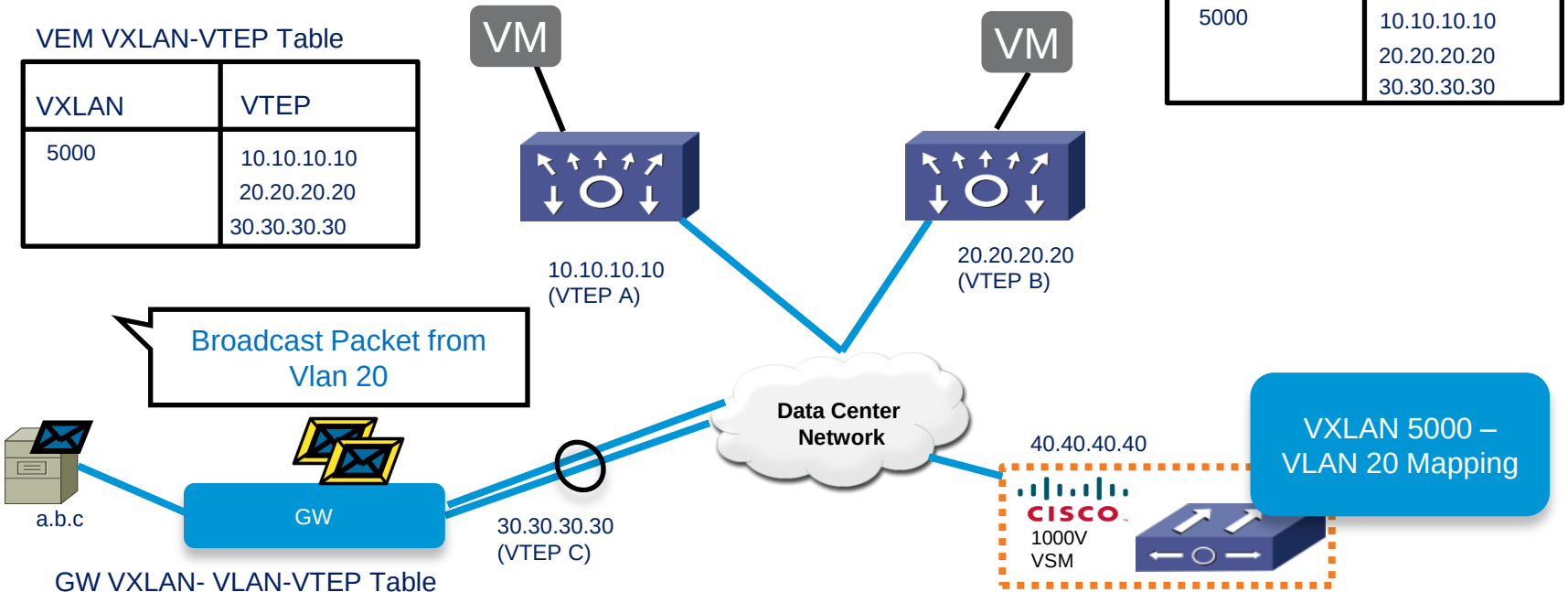
Use Case 3 – Connecting to physical storage from virtual workloads on VXLAN



Packet Flow – VLAN to VXLAN

VEM VXLAN-VTEP Table

VXLAN	VTEP
5000	10.10.10.10 20.20.20.20 30.30.30.30



VEM VXLAN-VTEP Table

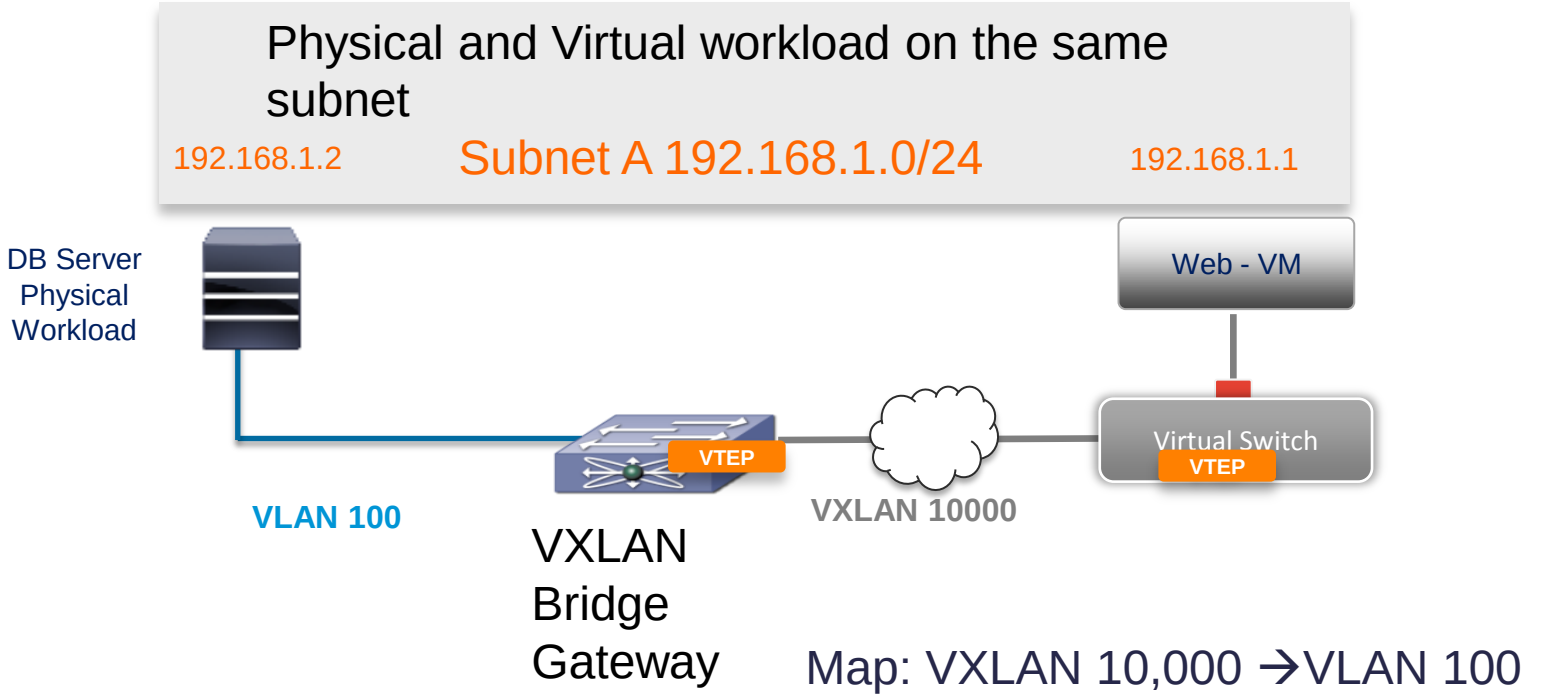
VXLAN	VTEP
5000	10.10.10.10 20.20.20.20 30.30.30.30

GW VXLAN- VLAN-VTEP Table

VXLAN	VLAN	VTEP
5000	20	10.10.10.10 20.20.20.20 30.30.30.30

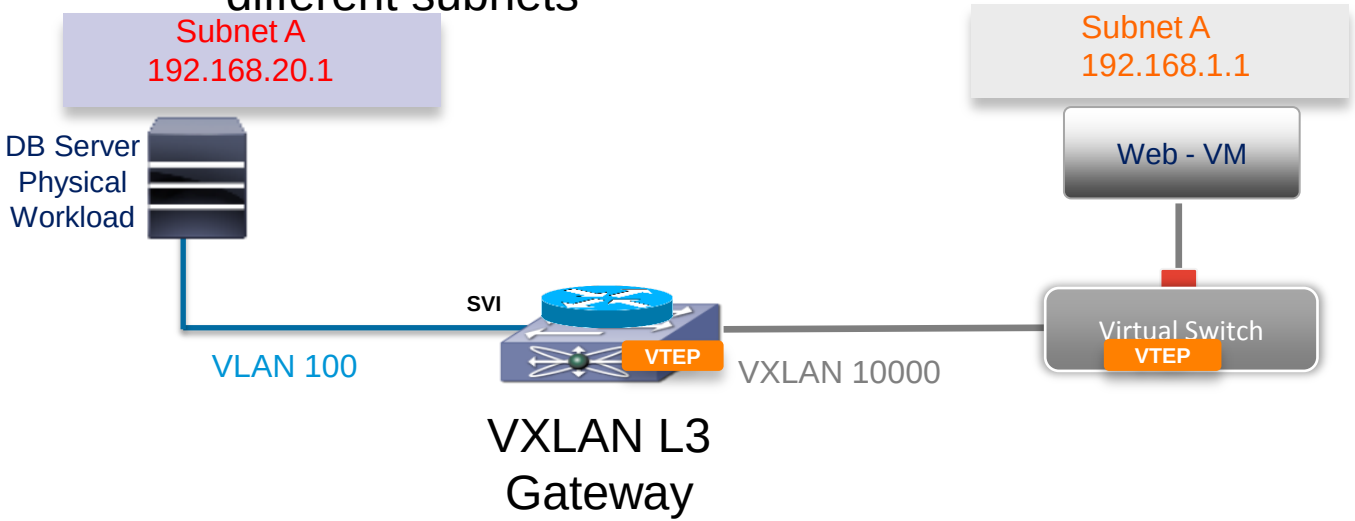
VXLAN	VTEP
5000	10.10.10.10 20.20.20.20 30.30.30.30

VXLAN-VLAN Bridge Gateway Use Case



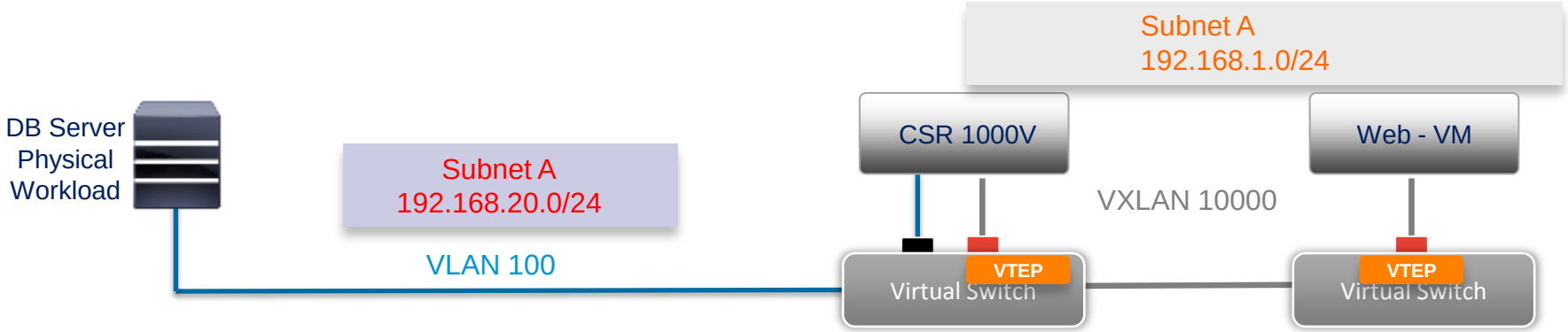
VXLAN L3 IP Gateway Use Case

Physical and Virtual workload on the different subnets

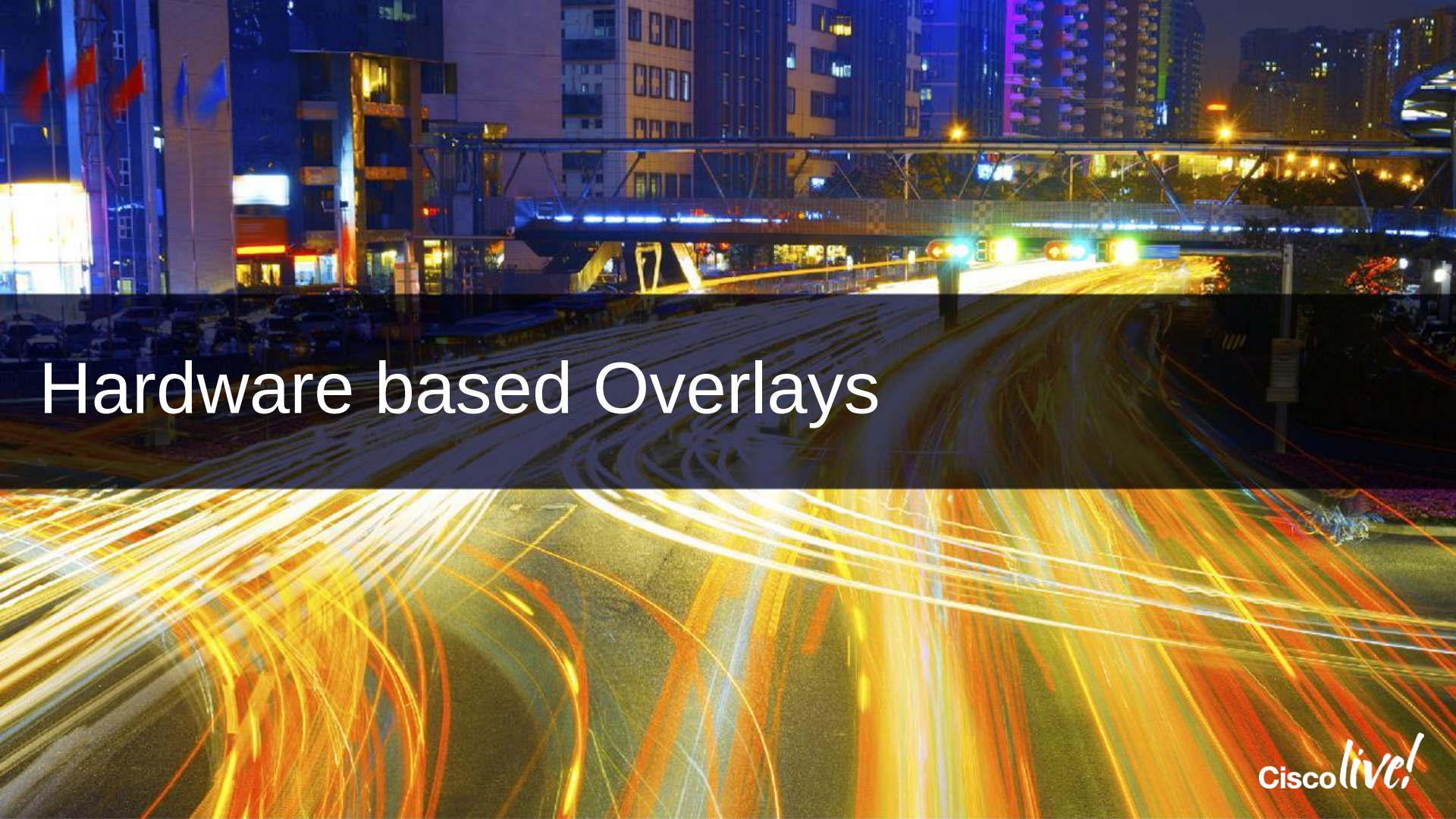


CSR 1000V as VXLAN IP Gateway

Physical and Virtual workload on the different subnets



Virtual L3 Gateway connected to Virtual Switch gets the IP Packet in clear
Virtual Switch (VTEP Configured) is responsible to encapsulation and decap.



Hardware based Overlays

Cisco live!

Host vs. Network Based Overlays

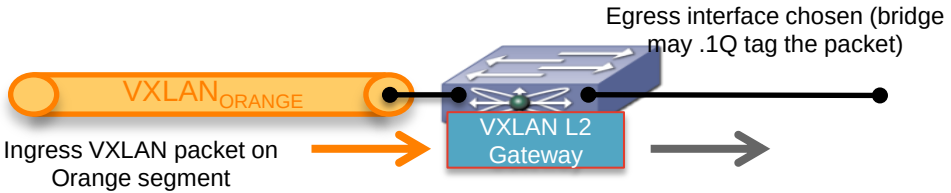
Host Based Overlays	Network Based Overlays
Centralized Control and point of Management	Distributed Control
VTEP is initiated at the Server	VTEP is initiated at the Edge of Network
Server performs encapsulation and forwarding	Edge switch performs encapsulation and forwarding
End Points are Virtual	End Points are Physical Switches/Routers
Nexus 1000V, CSR 1000V	Nexus 3K/5K/6K/7K, Nexus 9K standalone, ASR 1K/9K

Hybrid Overlays

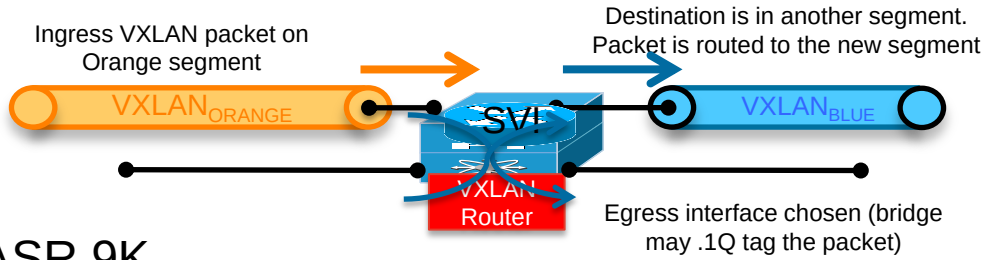
Combination of both Host and Network based overlays where Virtual and Physical worlds interconnect

VXLAN on HW Platforms - Supported Functionality

- VXLAN to VLAN Bridging (L2 Gateway)
  - ✓ N5600, N6K-X, N7K (F3), N9K, N31XX, ASR 9K

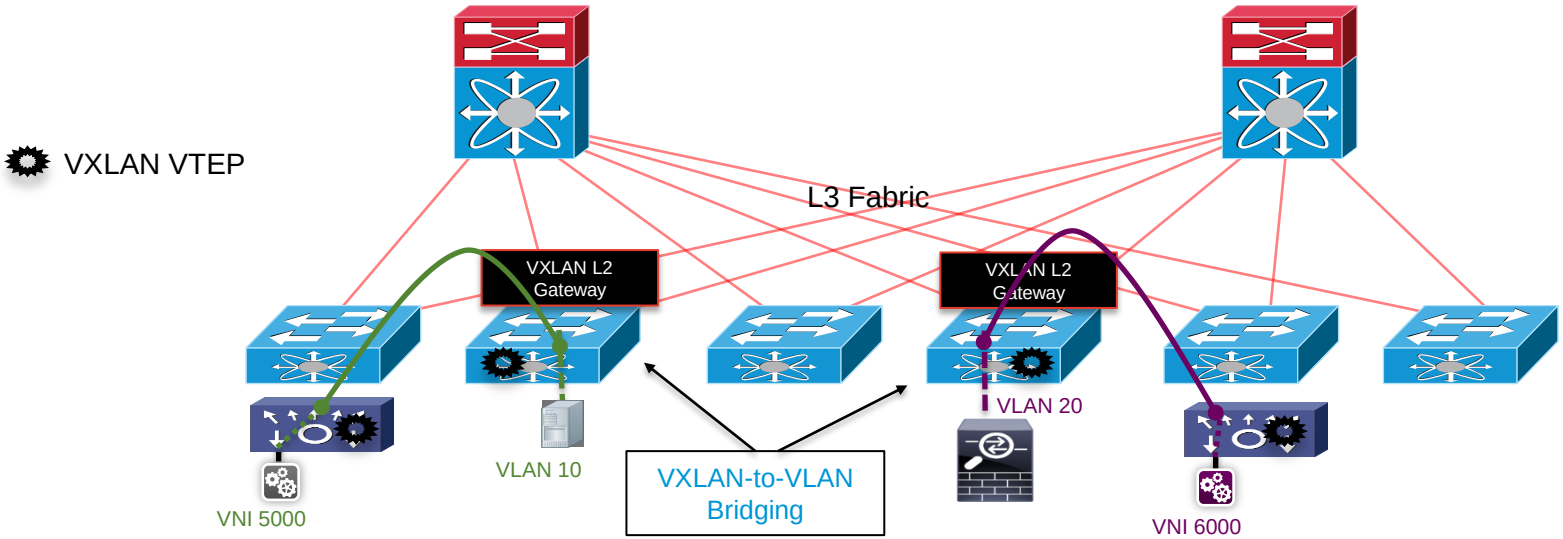


- V(X)LAN-to-V(X)LAN Routing (L3 Gateway)
  - ✓ N5600, N6K-X, N7K (F3), N9K, ASR 9K



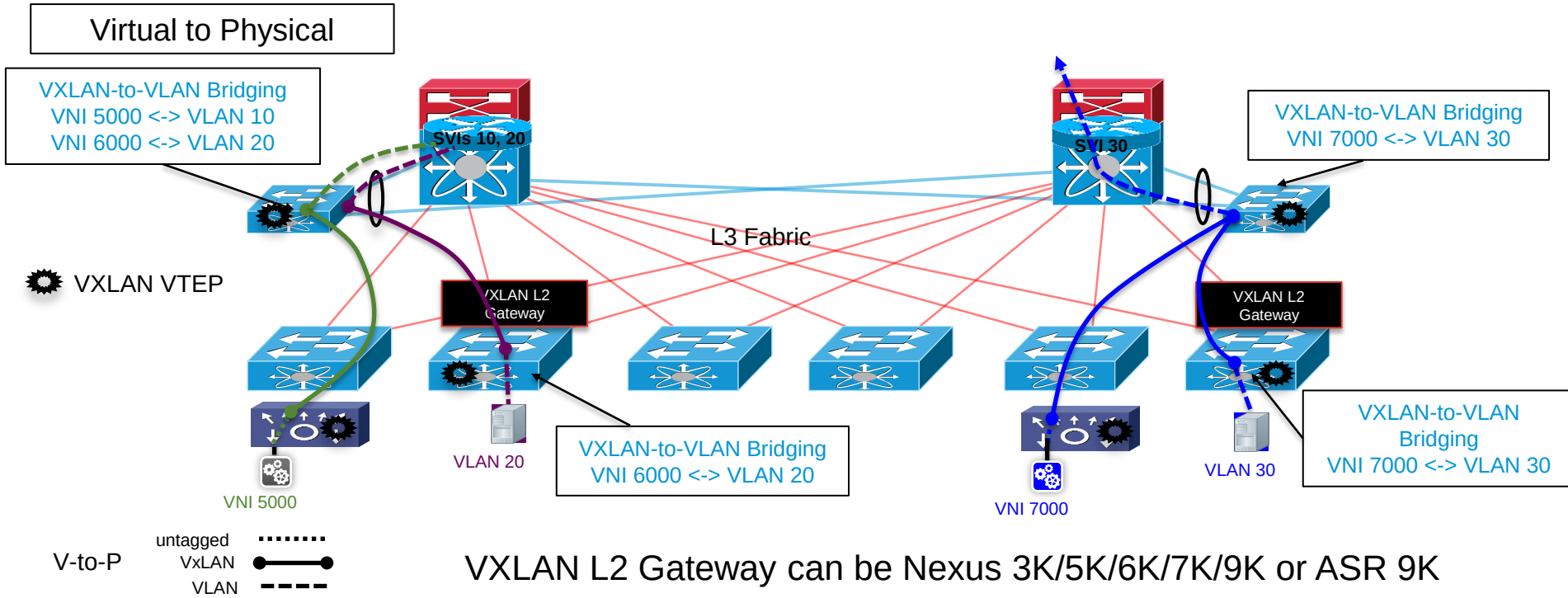
HW VXLAN Bridging - Intra-Subnet Communication

Virtual to Physical



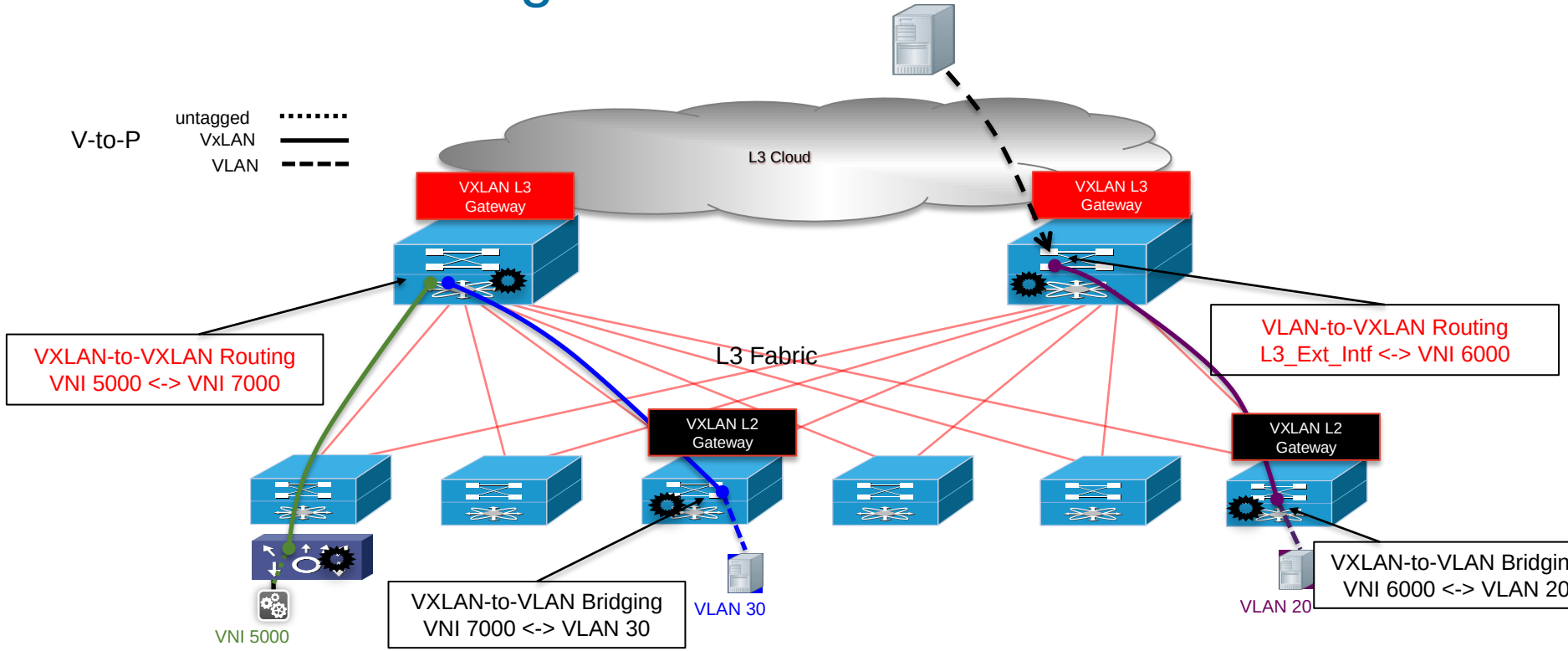
VXLAN L2 Gateway can be Nexus 3K/5K/6K/7K/9K or ASR 9K

HW VXLAN Bridging - Inter-Subnets Communication



VXLAN L2 Gateway can be Nexus 3K/5K/6K/7K/9K or ASR 9K

HW VXLAN Routing - Inter-Subnets Communication



VXLAN L3 Gateway can be Nexus 5K/6K/7K/9K or ASR 9K

Summary

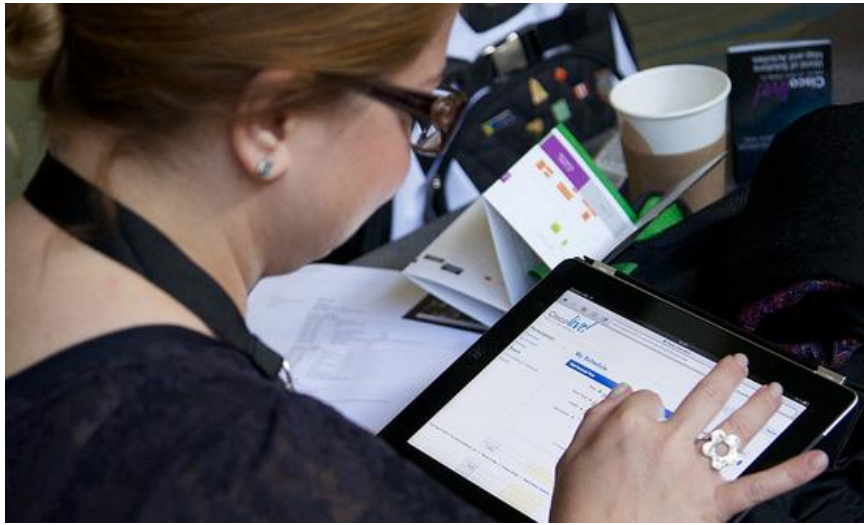
Cisco live!

Virtual Overlay - Summary

- Types of Virtual Overlays- Host vs. Network vs. Hybrid
- VXLAN is one of the Encapsulation types to provide Virtual Overlays
- Overlay Use Cases include Simplified Workload Provisioning / Automation, scale and workload mobility
- Enhancements to VXLAN include:
  - Unicast-mode VXLAN
  - MAC Distribution and ARP termination
  - Using MP-BGP as a control plane to expand the VXLAN domain
- Nexus 1000V provides a host-based overlay
- Hardware platforms also support VXLAN and provide L2 and L3 gateway functionality

Complete Your Online Session Evaluation

- Give us your feedback and you could win fabulous prizes. Winners announced daily.
- Complete your session evaluation through the Cisco Live mobile app or visit one of the interactive kiosks located throughout the convention center.



Don’t forget: Cisco Live sessions will be available for viewing on-demand after the event at CiscoLive.com/Online

Continue Your Education

- Related Sessions
  - BRKDCT-2328 - Evolution of Network Overlays in Data Center Clouds
  - BRKDCT-2404 - VXLAN deployment models - A practical perspective
- Walk-in Self-Paced Labs - Cisco Nexus 1000V: Implementing Overlay Networks in OpenStack/KVM environment
- Table Topics
- Meet the Engineer 1:1 meetings


Thank you.


CISCO