Advanced Topics & Future Directions in Network Virtualization with NSX

Bruce Davie, VMware, Inc
Disclaimer

• This presentation may contain product features that are currently under development.
• This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.
• Features are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.
• Technical feasibility and market demand will affect final delivery.
• Pricing and packaging for any new technologies or features discussed or presented have not been determined.
Objectives

• Provide an update on latest NSX capabilities
• Provide some insight into future NSX direction
• Deepen your understanding of network virtualization and its value
Overview

- Network Virtualization in One Slide
- Physical Network Integration
- Encapsulations
- Service Chaining
- Multi-site Network Virtualization
- Summary
Network Virtualization – an Analogy

- **x86 Environment**
  - Virtual Machine
  - Virtual Machine
  - Virtual Machine
  - Decoupled

- **Hypervisor**

<table>
<thead>
<tr>
<th>Application</th>
<th>Application</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Compute &amp; Memory</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Network Virtualization Platform**

<table>
<thead>
<tr>
<th>Workload</th>
<th>Workload</th>
<th>Workload</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirement: IP Transport</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **L2, L3, L4-7 Network Services**

<table>
<thead>
<tr>
<th>Virtual Network</th>
<th>Virtual Network</th>
<th>Virtual Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Network</td>
<td>Physical Network</td>
<td>Physical Network</td>
</tr>
</tbody>
</table>
NSX – Network Virtualization Platform

Controller Cluster

Virtual Network

L2

L3

VLAN

L2

NSX – Network Virtualization Platform

硬件

软件

VTEP API

NSX vSwitch

NSX vSwitch

Open vSwitch

Open vSwitch

NSX Edge

vSphere Host

vSphere Host

KVM

Xen Server

NSX Edge

北向接口

NSX API

Cloud Management Platform

物理网络

VLAN

VM

VM

北向接口

北向接口

VMworld 2014
Connecting the Physical to the Virtual

Controller Cluster

API (OVSDDB)

IP Underlay
(no multicast required)
Connecting the Physical to the Virtual

Controller Cluster

API (OVSDB)

VM MACS

DB

Physical Workloads

Logical network

IP Underlay
(no multicast required)
Connecting the Physical to the Virtual

Controller Cluster

API (OVSDB)

Logical network

VM MACs

DB

IP Underlay
(no multicast required)

Physical Workloads
Connecting the Physical to the Virtual

Controller Cluster

API (OVSDB)

Logical network

IP Underlay (no multicast required)

DB
VM MACS

PHYMACS

Physical Workloads

vmworld'2014
Connecting the Physical to the Virtual

Controller Cluster

API (OVSDB)

IP Underlay (no multicast required)

Logical network

DB
VM MACS
PHYMACS

Physical Workloads
Connecting the Physical to the Virtual

Controller Cluster

API (OVSDB)

Tunnels (VXLAN)

IP Underlay (no multicast required)

vmworld 2014
Distributed Logical Routing (P → V)

Logical View

- 192.168.1.254
- 192.168.2.254
- 192.168.1.1
- 192.168.2.1

Physical View

- 192.168.1.1
- 192.168.2.254
- vSwitch
- Hypervisor
- 192.168.2.1

vmworld 2014
Packet Walk

192.168.1.1

vSwitch
Hypervisor

192.168.2.1
Packet Walk

ARP REP: IP=192.168.1.254
MAC=LogicalRouter_A
Packet Walk

ARP: IP=192.168.2.1
SRCMAC=Hypervisor
VNI=2
Packet Walk

ARP_REP: IP=192.168.2.1
MAC=Physical
ARP REPL: IP=192.168.2.1
MAC=Physical
VNI=2
Packet Walk

192.168.1.1

Hypervisor

vSwitch

192.168.2.1
Packet Walk

Hypervisor

vSwitch

192.168.1.1

192.168.2.1
Distributed L3

• The other paths (P→V, V→V, P→P) are similar
  – Router’s ARP reply always comes from nearby VTEP or vswitch
  – That node then ARPs toward the ultimate destination

• Note that the LR is fully distributed among VTEPs and vswitches
  – Any E-W traffic will travel directly between hypervisors
  – No single device does all routing
VTEP Futures

- BFD health monitoring
  - Mitigate service node failures
  - Provide overlay health monitoring/troubleshooting
- ACL configuration
- QoS – DSCP setting
- Higher layer services (e.g. ADCs)
Handling Elephant Flows

1. Detect Elephants
   - Must be long-lived and high-bandwidth
   - vSwitch ideally suited for task, maybe combine with central control

2. Do something with them:
   - Mark the outer DSCP
   - Put them in a queue separated from mice
   - Route along their own path or network
   - Convert to mice
Results – flow statistic detection & alternate queue reaction

Mice vs Elephants (Detection off)

Latency (ms)

Latency (ms)

Bandwidth (Mbps)

Bandwidth (Mbps)

Time (Secs)

Time (Secs)
Results – flow statistic detection & alternate queue reaction

Mice vs Elephants (Detection on)

Bandwidth (Mbps)

Latency (ms)

Time (Secs)

Elephant
Mice

cumulusnetworks.com
Tunneling

• Networking people love to argue about tunnel formats
• Primarily a low-level detail of the implementation
• But tunnel format matters:
  – Interoperability (HW + SW endpoints)
  – ECMP on current switches
  – Extensibility
  – Performance
  – Visibility
• Current options (VXLAN, NVGRE, STT) all fall short somewhere
• Enter Geneve (Generic Network Virtualization Encapsulation)
  – VMware, Microsoft, Red Hat, Intel (the x86 world)
Tunnels are like cables

Physical

**Copper** Cable

Third party hardware

Virtual Network

WORLD

Controller

VXLAN Cable

VXLAN Cable

Hypervisor

STT Cable

Hypervisor

VM

VM

VM

VM

vmworld 2014
Tunnels are like cables

Physical

Copper Cable

Third party hardware

VXLAN Cable

Geneve

Virtual Network

WORLD

Controller

Hypervisor

Third party hardware

Hypervisor

Geneve

VM

VM

VM

VM

VM

VM

VM

VM
Tunnels are like cables
Tunnels are like cables

Physical

Copper Cable

Geneve

Virtual Network

WORLD

Controller

Third party hardware

Hypervisor

Geneve

VM

VM

VM

VM

vmworld'2014
Geneve Header

<table>
<thead>
<tr>
<th>MAC</th>
<th>IP</th>
<th>UDP</th>
<th>Geneve Options</th>
<th>Inner Eth</th>
<th>Inner IP</th>
<th>Inner L4</th>
<th>Payload</th>
</tr>
</thead>
<tbody>
<tr>
<td>+-----------------+---------------------------------+-----------------+-------------------+-----------------+-----------------+-----------------+-----------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ver</td>
<td>Opt Len</td>
<td>O</td>
<td>C</td>
<td>Rsvd.</td>
<td>Protocol Type</td>
<td>Virtual Network Identifier (VNI)</td>
<td>Reserved</td>
</tr>
<tr>
<td>+-----------------+---------------------------------+-----------------+-------------------+-----------------+-----------------+-----------------+-----------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How the Options Are Used

• <Type, length, value> structure
  – Type is structured to allow vendor-specific options
• “C” bit indicates “critical” options
• Example use:
  – Convey the source or dest of a packet when that info can’t be determined from other fields
    • E.g. ARP request from a logical router could be from anywhere physically
    • Mirrored packets might be sent somewhere other than dest address
  – Indicate traceflow packets
  – Carry logical port info for egress policy
  – State versioning
  – Service chaining
  – Etc.
What about VXLAN, STT, etc.?

- Hardware that supports VXLAN and STT will be around for a long time
- If you’re buying switches today, they’ll support VXLAN
- VXLAN NIC offloads also available today
- Of course we’ll continue to support VXLAN & STT
  - Easy for us to support multiple encapsulation types
  - We mix & match STT & VXLAN (and GRE) today
- Geneve goal is that we don’t need another encap for a long time
Service Chaining

- Creating a graph of services (e.g. load balance, firewall, WAN optimize, etc.)
- Network virtualization provides a natural way to do this in automated manner
  - Creating virtual topologies
- Often need to pass metadata along the chain
  - e.g. make the results of a classification step available to a later node
  - Ongoing argument about how to pass this metadata – Geneve provides a reasonable option
Service Chaining Example: E-W Firewall & Routing
Multi-Site Network Virtualization

- We support some multi-site scenarios today (see NET1974)
  - E.g. stretched metro cluster
  - Snapshot, clone, restore across locations
- Important to think of the full picture, not just networking
  - E.g. do you want to migrate a VM across the WAN without its data?
  - Where does your Cloud Management Platform live? How many CMP instances?
- Lots of distinct use cases ➔ plenty of work ongoing
The Multi-Site Spectrum

Single DC Federation
- Sub-ms latency
- High BW

Metro Area DCs
- Low-ms latency
- High BW

Geographically Dispersed DCs
- 100-ms latency
- Constrained BW
Connecting Virtualized Data Centers to the WAN
Using “Option B” to Map Logical Networks to MPLS Labels

Logical Network Prefixes advertised in MP-BGP with MPLS labels

MPLS Core

To Customer Sites

ASBR

MPLS Labelled Packets mapped to/from logical networks

Treat interface like inter-AS (RFC 4364)

NSX Edge
Multi-site using MP-BGP
Multi-site using MP-BGP
Controller State Distribution

- All nodes active
- Workload sliced among nodes
- Logical network state – semantically rich
Controller State Distribution

- Transport Network
- WebService API
- Persistent Storage
- Logical Network

Controller Cluster

Node1
Node2
Node3
Node4
Node5

NSX Controller

WebService

Persistent Storage

Logical Network

Transport Network

Logical Network

Controller Cluster

Node1
Node2
Node3
Node4
Node5

NSX Controller

WebService

Persistent Storage

Logical Network

Transport Network
Summary

• Network virtualization – not just for the bleeding edge
• Physical networks are part of the story
  – Control the physical edge for non-virtualized workloads and north-south traffic
  – Communicate with the underlay for congestion/elephant flow mitigation
  – Keep moving up the stack
• Tunneling – a detail, but an important one
• Multi-site
  – Consider use case & complete system
  – Some solutions today, more soon
• Exciting times for networking!
Thank You

Bruce Davie
bdavie@vmware.com
Thank You
Fill out a survey

Every completed survey is entered into a drawing for a $25 VMware company store gift certificate
Advanced Topics & Future Directions in Network Virtualization with NSX

Bruce Davie, VMware, Inc