Distributed Overlay Virtual Ethernet (DOVE) Networks

Renato Recio
IBM Fellow & System Networking CTO

© 2012 IBM Corporation
Early Ethernet Campus Evolution

- In the beginning, Ethernet was used to interconnect stations (e.g. dumb terminals), initially through repeater & hub topologies, eventually through switched topologies.

- Ethernet campuses evolved into a structured network typically divided into a Core, Service (e.g. firewall), Aggregation and Access Layer.
  - Traffic pattern is mostly North-South (directed outside campus vs peer-peer).
  - To avoid spanning tree problems, campus networks typically are divided at access.
The same structured network topologies were used in the data center, but...

- Traffic pattern is mostly East-West (e.g. Application to Database tier).
- Large layer-2 domains needed for clustering and Virtual Machine mobility.

Partly due to Ethernet limitations (e.g. lack of flow control), the data center used additional networks, such as

- Fibre Channel Storage Area Networks (SAN)
- InfiniBand cluster networks.
Traditional Data Center Network Issues

Discrete & Decoupled
- Discrete components and piece parts
- Multiple managers and management domains
- Box level point Services

Manual & Painful
- Dynamic workload management complexity
- Multi-tenancy complications
- SLAs & security are error-prone

Limited Scale
- Too many network types, with too many nodes & tiers
- Inefficient switching
- Expensive network resources
Clients are looking for smarter Data Center Infrastructure that solves these issues.
Smarter Data Center Infrastructure

Integrated
- Expandable Integrated System
- Simple, consolidated management
- Software Driven Network stack

Automated
- Workload Aware Networking
- Dynamic provisioning

Optimized
- Converged network
- Single, flat fabric
- Grow as you need architecture
Optimized

Traditional Arcane DCN → Mesh Ahead

- Multi-tiered tree topologies
- High oversubscription
- Expensive, high bandwidth uplinks
- Small layer-2 fabric
- Robustness of higher tier products has been a concern

- Mesh, Clos, Jellyfish topologies
- Oversubscription only to WAN/core
- High cross sectional B/W (cheap TOR B/W)
- Layer-2 scaling options (more next)
- Robust, HA topologies
Standard Fabric Technology Options

**Optimized**

**Layer-3**
- Established technology
- Standards based
- Distributed control plane
- Large scalability
- HA with fast convergence
- Small layer 2 without DOVE network
- Many devices to manage

**OpenFlow**
- Large layer-2
- Distributed control plane
- Large scalability
- HA with fast convergence
- Enables network functions delivered as Services → e.g. disjoint multi-pathing (more later)
- Emerging technology
- Client acceptance barrier

**TRILL**
- Large layer-2
- Distributed control plane
- Large scalability
- HA with fast convergence
- Emerging technology (some proprietary)
- Single disjoint multi-path fabric may need new RFC

**OpenFlow Controllers**

**Open-Flow Switches**
Key fabric requirements
- Storage & cluster require full path redundancy and efficient multi-pathing.
- Large layer-2 for VMs.

Completely redundant TRILL fabrics meet these requirements, but with administration burden.
- 2x SAN configuration & maintenance

Dual-TRILL fabrics are emerging today.
- Single TRILL fabric, with disjoint multi-pathing simplifies administration.
OpenFlow Based Fabric Overview

- Each switch has Layer-2 forwarding turned off.
- Each switch connects to OpenFlow Controller (OFC).
- OFC discovers switches and switch adjacencies.
- OFC computes shared or disjoint physical paths and configures switch forwarding tables.

OpenFlow can also be used to create a disjoint multi-pathing fabric.
Virtualization increased network management complexity

### Before Virtualization

- Static workloads ran on bare-metal OS
- Each workload had network state associated with it.
- Physical network was static & simple (configured once)

### Physical Network with vSwitches

- Server virtualization = dynamic workloads
- VM’s network state resides in vSwitch/DCN
- Physical network is dynamic and more complex (VMs come up dynamically & move)
IBM VMready and DVS-5000V

VMready™ Phase 2
Shipping since Nov 2008

VMready™ Phase 3
with DVS-5000V

1. Per-VM switching in HW
2. Hypervisor Vendor Agnostic
3. Management Platform Integration with VMware vCenter
4. IEEE 802.1Qbg standard based
5. IBM Distributed Virtual Switch 5000V for VMware
6. Network state migrates ahead of VM

Closed loop verification with vCenter
IBM System Networking OS feature set on a distributed, virtual switch for VMware

Seamless integration with VMware vCenter

Standards (Qbg) based network virtualization coordination between Hypervisor & physical switch

Optimizes East-West traffic between VMs in a single server (VEB) & VMs within a rack/chassis (VEPA)

Administration simplicity (VEPA)
Number of VMs per socket is rapidly growing (10x every 10 years).

- Increases amount of VM-VM traffic in Enterprise Data Centers (e.g. co-resident Web, Application & Database).
- VM growth increases network complexity associated with creating/migrating: layer-2 (VLANs, ACLs...) & layer-3 (e.g. Firewall, IPS) attributes.
Layer-2 vSwitch features, plus:

1. Layer-3 Distributed Overlay Virtual Ethernet (DOVE)
2. Simple “configure once” network (physical network doesn’t have to be configured per VM).
3. De-couples virtual from physical
4. Multi-tenant aware
5. Enables cross-subnet virtual appliance services (e.g. Firewall, IPS)
DOVE Technology
VXLAN based Encapsulation Example

Original Packet

Encapsulation Protocol (EP) Header (e.g. VXLAN based)
(VXLAN extension in Yellow → necessary IETF version field)
Overview of DOVE Technology Elements

**DOVE Controller**
- Performs management & a portion of control plane functions across DOVE Switches

**DOVE Switches (DOVES)**
- Provides layer-2 over UDP overlay (e.g. based on OTV/VXLAN)
- Performs data and some control plane functions
- Runs in Hypervisor vSwitch or gateways
- Provides interfaces for Virtual Appliances to plug into (Analogous to appliance line-cards on a modular switch)
DOVE network simplifies virtual machine network
  – Enables multi-tenancy all the way to the VM
  – Enables single MAC Address per physical server (2 for HA)
  – Significantly reduces size of physical network TCAM & ACL tables
  – Increases layer-2 scale within Data Center and across Data Centers, by decoupling VM’s layer-2 from physical network
  – Qbg automates layer-2 provisioning, DOVE automates layer 3-7 provisioning

Standards based multi-pathed physical network
System Networking Element Manager

Perform Efficient Firmware or Configuration Updates to Multiple Switches

Automate VM network resident port profiles and converged fabric Quality of Service

Performance trend & root-cause analysis, fault management, ..
Software Defined Networking Technologies

- Network functions delivered as services
  - Multi-tenant VM security
  - Virtualized load balancing

- Network API’s provides an abstract interface into underlying controller
  - Distributes, configures & controls state between services & controllers
  - Provides multiple abstract views

- Network Operating System drives set of devices
  - Physical devices (e.g. TOR)
  - Virtual devices (e.g. DVS 5000v)
## Summary of Technology Trends

<table>
<thead>
<tr>
<th>CAPEX</th>
<th>OPEX</th>
<th>Technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SDN Stack (OpenFlow, DOVE, Services)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Integrated Management</td>
</tr>
<tr>
<td></td>
<td></td>
<td>vSwitch with Qbg</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DOVE Switch</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multi-pathing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Converged Enhanced Ethernet</td>
</tr>
</tbody>
</table>
Thank You!

QUESTIONS
Back-ups

- Automating coordination of layer-2 state
- IBM DVS 5000v overview
- Examples of some of the values associated with DOVE Technology
  - Multi-tenancy
  - Efficiency
- Summary of Technology Trends
Automating Layer-2 State using IEEE 802.1Qbg (VM Creation)

1. Create set of Virtual Port Profiles
2. Query available Port Profile types & select one
3. Push VM & Port Profile info to server’s virtualization infrastructure
4. VSI Discovery and Configuration Protocol (VDP)
5. Retrieve Port Profile Information
6. VM is brought on-line after VDP completes

System Admin

VM Manager

Network Admin

VSI Manager

Database

Switch (a.k.a. Bridge)

vSwitch

Physical End Station
### IBM DVS 5000v Capabilities

<table>
<thead>
<tr>
<th>Feature</th>
<th>Internal</th>
<th>External</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VM-VM Virtual Switching Mode</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Virtual Switch Provider</strong></td>
<td>✓IBM</td>
<td>✓IBM</td>
</tr>
<tr>
<td>Flexibility to use a mix of internal and external switching modes</td>
<td>✓Yes</td>
<td></td>
</tr>
<tr>
<td>Eliminates the need to manage the virtual switch</td>
<td></td>
<td>✓Yes</td>
</tr>
<tr>
<td><strong>Standard based network virtualization coordination</strong></td>
<td>✓IEEE 802.1 de-facto version 0 VEPA &amp; VDP</td>
<td>✓IEEE 802.1 de-facto version 0 VEPA &amp; VDP</td>
</tr>
<tr>
<td>Requires new hardware</td>
<td>✓No</td>
<td>✓No</td>
</tr>
<tr>
<td>Sophisticated switch attributes (e.g. ACLs, QoS)</td>
<td>✓Yes</td>
<td>✓Yes (External switch’s)</td>
</tr>
<tr>
<td><strong>Automated migration of port profiles</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Internal Switch</td>
<td>✓Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>B. External Switches</td>
<td>✓Yes</td>
<td>✓Yes - With external switch that supports Qbg</td>
</tr>
<tr>
<td>Cross Data Center VM migration</td>
<td>✓Yes - Through partner’s standard based approaches (e.g. MPLS/VPLS)</td>
<td></td>
</tr>
</tbody>
</table>
Hypervisor vSwitches enable addition of virtual appliances (vAppliances), which provide secure communication across subnets (e.g. APP to Database tier).

- However, all traffic must be sent to an external Layer-3 switch, which is inefficient considering VM/socket growth rates and integrated servers.

To solve this issue requires cross-subnet communications in Hypervisor’s vSwitch.
Multi-tenant, Cloud environments require multiple IP address spaces within the same server, within a Data Center and across Data Centers (see above).

- Layer-3 Distributed Overlay Virtual Ethernet (DOVE) switches enable multi-tenancy all the way into the Server/Hypervisor, with overlapping IP Address spaces for the Virtual Machines.

Note, vSwitches and vAppiances are not shown.