VMware Networking Best Practices

Jeff Hunter, VCP
Senior Systems Engineer
VMware
jhunter@vmware.com
Currently defined per ESX host

- vmnic: Physical network adapter
- vNIC: Virtual network interface card - virtual hardware device presented to virtual machine (VM)
- virtual port: Connection point between the vNIC and the vSwitch
- vSwitch: Virtual switch - similar to physical switch. Key functions:
  - Layer 2 forwarding engine
  - VLAN tagging, stripping, and filtering
  - Security, checksum, and segmentation offload units
vNIC MAC addresses

- Virtual machines (VMs) can have up to 4 vNICs
- MAC addresses generated automatically
  - Organization Unique Identifiers (OUI) assigned by IEEE to VMware
    - 00-50-56 and 00-0c-29
  - Can also be created manually
    - 00:50:56:00:00:00 - 00:50:56:3F:FF:FF for statically assigned VM MAC addresses
- MAC address created using: VMware OUI, SMBIOS UUID for the physical ESX host, hash based on the name of the VM
- MAC address can be found in .vmx file
ESX Virtual Switch

- **Links VMs to each other locally as well as to physical network**
- **Built to order at run-time. Collection of small functional units:**
  - Core layer forwarding engine - only processes Layer 2 Ethernet headers. Completely independent of other implementation details
  - VLAN tagging, stripping, and filtering units
  - Layer 2 security, checksum, and segmentation offload units
- **ESX only loads what is needed. Modular design good for:**
  - Lowest possible complexity and demands on system performance
  - Supports temporarily loading certain components
  - VMware and third-party developers can easily incorporate modules to enhance the system
Virtual Switch Similar to Physical Switch

- Maintains a MAC address, port forwarding table
- Looks up each frame’s destination MAC when it arrives
- Forwards a frame to one or more ports for transmission
- Avoids unnecessary deliveries (in other words, it is not a hub)
- Access to single VLAN - makes “access port”
- Access to multiple VLANs - makes “trunk port”
- Admin can manage many configuration options for vSwitch as a whole and for individual ports using Virtual Infrastructure Client
  - NIC Teaming, Security, Traffic Shaping, etc.
Virtual Switch is Different Than Physical Switch

- ESX provides direct channel from virtual Ethernet adapters for configuration information
  - No need to learn unicast addresses or perform IGMP snooping to learn multicast group membership

- Spanning Tree Protocol (STP) not used on vSwitches
  - Single-tier networking topology enforced - no way to interconnect vSwitches to each other. Traffic received on an uplink is never forwarded to another uplink

- vSwitch Isolation
  - No need to cascade vSwitches - capability not provided
  - vSwitches cannot share physical NICs
  - Each vSwitch has its own forwarding table
Virtual Switch is Different Than Physical Switch

- **Uplink Ports**
  - Associated with physical NICs - provides connection between virtual network and physical network
  - Physical adapters connect to uplink ports when initialized by device driver or when teaming policies for virtual switches are reconfigured

- **Some vSwitches should not connect to physical network, do not have an uplink port**
  - Example: vSwitch that provides connections between firewall VM and VMs protected by firewall
Virtual Switch is Different Than Physical Switch

> vSwitch Correctness

- Ensures that virtual machines or other nodes on the network cannot affect the behavior of the virtual switch - two important safeguards:
  - vSwitches do not learn MAC addresses from network to populate forwarding tables - eliminates likely vector for denial-of-service (DoS) or leakage attacks
  - vSwitches make private copies of frame data used to make forwarding or filtering decisions - vSwitch does not copy entire frame (inefficient), but ESX must make sure guest OS does not have access to sensitive data when frame is passed to vSwitch
- ESX ensures that frames are contained within the appropriate VLAN
  - VLAN data is carried outside frame as it passes through the vSwitch, vSwitches have no dynamic trunking support
Virtual Local Area Networks (VLANs)

- Ethernet frames with 802.1Q tags - three different modes:
  - Virtual switch tagging (VST mode) - most common
  - One Port Group on a vSwitch for each VLAN
  - Port Group tags on outbound frames, removes tags on inbound
  - Physical switch must provide a trunk
Virtual Local Area Networks (VLANs)

- Ethernet frames with 802.1Q tags - three different modes:
  - VM guest tagging (VGT mode)
    - Install 802.1Q VLAN trunking driver inside VM
    - Physical switch must provide a trunk
  - External switch tagging (EST mode)
    - Uses external switches for VLAN tagging
    - Physical switch trunk not required
Port Groups

- Port Groups are not VLANs - they are configuration templates for the vNIC ports on the vSwitch
  - Port Groups allow administrators to group vNICs from multiple VMs and configure them simultaneously
  - Administrator can set specific QoS, security policies, and VLANs by changing Port Group configuration
  - Multiple Port Groups can point to same VLAN
Layer 2 Security Features

- vSwitch has ability to enforce security policies to prevent VMs from impersonating other nodes on network
- Three components to this feature:
  - Promiscuous mode (disabled by default)
  - MAC address change lockdown prevents VMs from changing their own unicast addresses (allowed by default)
  - Forged transmit blocking prevents VMs from sending traffic that appears to come from nodes on the network (allowed by default)
Best Practices - Preliminary Design

- Network connections should be redundant
  - Service console redundancy needed for reliable VMware HA
  - VM network(s) carry production traffic
  - vmkernel redundancy - especially if iSCSI and/or NFS is in use

- Design LAN switching infrastructure as you would in presence of a server with active/standby NIC Teaming or Port-Channeling (depending on the ESX host configuration)

- 1Gb+ required for vmkernel vMotion, iSCSI

- On ESX host, set 100Mb connections explicitly
  - 100Mb, Full-duplex (not Auto-negotiate)

- Automatic speed, duplex negotiation fine for 1GB links
Best Practices - NIC Teaming

- 2 or more physical NICs used as vSwitch uplink
  - Also referred to as bonding
- Allows teaming of heterogeneous NICs (ex. Intel and Broadcom)
- Adds redundancy - especially if connected to separate physical switches (must be same Layer 2 domain)
- DTP (Dynamic Trunking Protocol) not supported by ESX virtual switches - use static trunking on the physical switches connected to virtual switches

- NIC Teaming configuration options:
  - Active/Standby
    - Rarely used - when number of physical NICs in host is limited
Best Practices - NIC Teaming

NIC Teaming configuration options, continued:

- Active/Active with load balancing based on the hash of the source and destination IP address
- 802.3ad link aggregation required (EtherChannel in Cisco environment)
- 802.3ad LACP does not run in vSwitch
  - No dynamic configuration - must be set to static (no auto/desirable)
  - channel group set to ON
  - Use trunkfast, trunk mode ON

vSwitch EtherChannel >
Best Practices - NIC Teaming

> ESX Teaming Failback

- When set to ON, risk of black-holing traffic exists when fail-over and fail-back occurs, but port of Catalyst switch does not begin forwarding right away

> Recommendation: Set Cisco Catalyst to use trunkfast and trunk mode to ON or disable Teaming Failback on ESX host
Best Practices - NIC Teaming

- NIC Teaming configuration options, continued:
  - Active/Active based on source VM MAC-address hash ("MAC hash")
  - Active/Active based on originating VM Port-ID ("IP hash")
- Both options...
  - Distribute VM traffic across physical NICs
  - Perform load sharing (not active load balancing)
  - Confine traffic from same vNIC to same physical NIC until failover

> Recommendation: Active/Active based on originating VM Port-ID
  - Simplicity (no link aggregation required on physical switch)
  - Multi-switch redundancy (default today)
Beaconing

- Probing function in ESX - monitors availability of physical NICs - use with teams connected to more than one physical switch
- May produce false positives, cannot detect upstream failures
- Beacon frames are Layer 2, Ethertype 0x05ff, burnt-in MAC address of physical NIC (not VMware MAC address)
- Beaconing is configured per port group

Recommendation: Use Link State Tracking with Cisco switches

- Upstream failure triggers downstream link failure which ESX can detect using Link Status Only under Network Failover Detection
Best Practices - Port Group Labeling

> Develop and enforce naming standard for Port Group Network Labels

- Examples include:
  - Same as VLAN number
  - vSwitch name and VLAN number
  - Network name - i.e. Prod1, Prod2, Mgmt1, etc.
  - Application attached to Port Group

> Network label must be the same across all ESX hosts in cluster - otherwise, vMotion migrations will fail

- However, Port Group labels must be unique across all vSwitches on same ESX host
Best Practices - LAN Design

- Ensure Layer 2 path from one vSwitch NIC to other vSwitches that are in same Layer 2 domain. Layer 2 path needs to be provided by physical LAN network.

  Example: In this topology, everything is Layer 2 adjacent. However, it is still possible to lose connectivity between VMs 5, 7 and/or VMs 4, 6:
  - VMs 5, 7 hash to VMNIC1
  - VMs 4, 6 hash to VMNIC2
  - VM 5 sends traffic to VM4
  - Catalyst1 floods Link 3
  - vSwitch on Host 2 would not accept traffic
Best Practices - LAN Design

Ensure Layer 2 path from one vSwitch NIC to other vSwitches that are in same Layer 2 domain. Layer 2 path needs to be provided by physical LAN network.

Example: In this topology, Layer 2 adjacency between the NIC cards via the Cisco Catalyst switches.
Best Practices - vSwitch Configuration

- No need to create multiple vSwitches to segment VM traffic or Service Console (SC), vmkernel traffic
  - Use Port Groups with different VLAN IDs and override the global teaming configuration to implement Port Group specific traffic load balancing

- The preferred 802.1q VLAN tagging mechanism is Virtual Switch Tagging (VST) - assign specific VLAN ID to a Port Group

- Avoid using native VLAN as explicit VLAN ID - if needed, specify VLAN ID = 0 to the relevant Port Group

- Create NIC teaming configurations across different NIC chipsets

- Cross team PCI based NICs with on-board motherboard NICs
Sample Design: Host with 4 Physical NICs

- 2 vSwitches - 2 physical NICs per vSwitch
- 3 Port Groups: SC, vMotion, VM
- VLAN trunking recommended
- vSwitch 1
  - On-board NIC 1 teamed with PCI NIC 1
  - Service Console (SC) Port Group - prefer NIC 1 in team
  - vMotion Port Group - prefer NIC 2 in team
- vSwitch 2
  - On-board NIC 2 teamed with PCI NIC 2
  - 1 Port Group per VLAN
  - Connect VMs to appropriate Port Groups
Sample Design: Host with 2 Physical NICs (Blades)

- 1 vSwitch with both physical NICs
- 2 Port Groups: SC-vMotion, VM
  - Each with different active/preferred NIC in team
- VLAN trunking highly recommended

In other words...
- SC-vMotion traffic is active/preferred on NIC 1
- VM traffic is always active/preferred on NIC 2
- Service Console and vMotion traffic is separated from VM traffic, but falls back to sharing in the event of a network or NIC failure
Sample Design: Host with 5+ Physical NICs

- Without trunks (no VLAN tagging)
  - 2 dedicated, teamed NICs for Service Console
  - 2 dedicated, teamed NICs for each VM group/network
  - 1 dedicated NIC for vMotion
  - If using iSCSI, 2 dedicated, teamed NICs for iSCSI storage connectivity

- With trunks (VLAN tagging enabled)
  - Use one of the previous approaches and scale up to meet bandwidth and redundancy requirements

- Without VLAN tagging, it is easy to consume many physical NICs. Recommendation: Use VLAN tagging (VST)
Additional Noteworthy Items

- Cisco Discovery Protocol (CDP) is supported
- Traffic shaping is possible
  - Average Bandwidth, Burst Size, and Peak Bandwidth settings
  - Default is disabled (no limits)
- TCP Segmentation Offload (TSO) and Jumbo Frame support were added to TCP/IP stack in ESX 3.5
- For performance and security reasons...
  - Service Console traffics should be isolated from VM traffic
  - vMotion should have its own dedicated network
  - iSCSI, NFS should have its own dedicated network
Additional Resources

VMware VI Product Documentation Online
>  http://www.vmware.com/support/pubs/vi_pubs.html

VMware Virtual Networking Concepts

VMware VI 3.x in a Cisco Network Environment

Highly recommended