Server Virtualization

Network Implications & Best Practices

Maurizio Portolani

Session Objectives

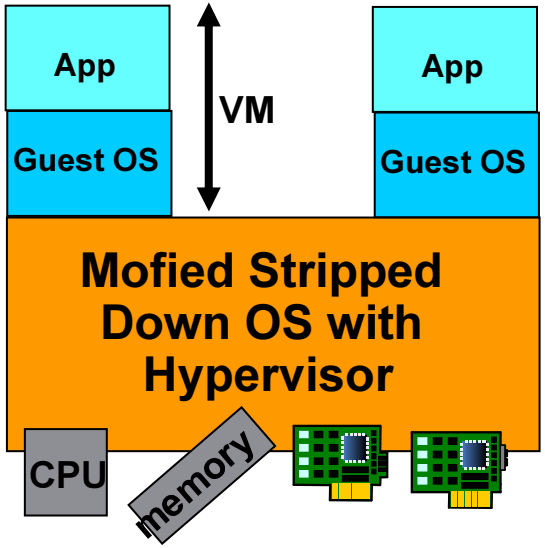
At the end of the session, the participants should be able to:

- Objective 1: Understand key concepts of server virtualization architectures as they relate to the network.
- Objective 2: Explain the impact of server virtualization on DC network design (Ethernet & Fiber Channel)
- Objective 3: Design Cisco DC networks to support server virtualization environments

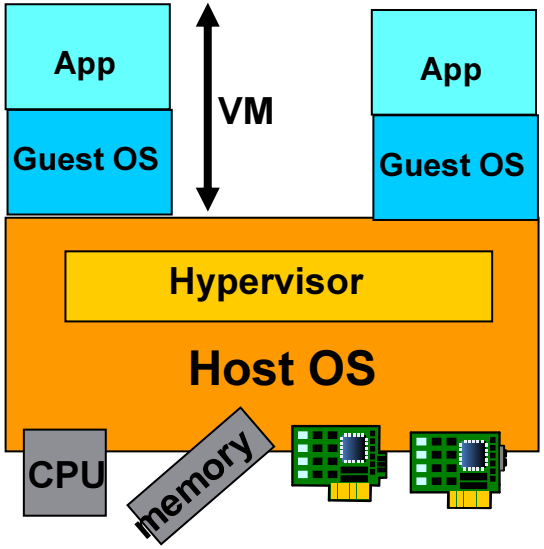
Agenda

- VMware Architecture and Components
- VMware LAN Networking
- Cisco/VMware DC DESIGNS
- Blade Server Designs
- Storage Implications of Server Virtualization

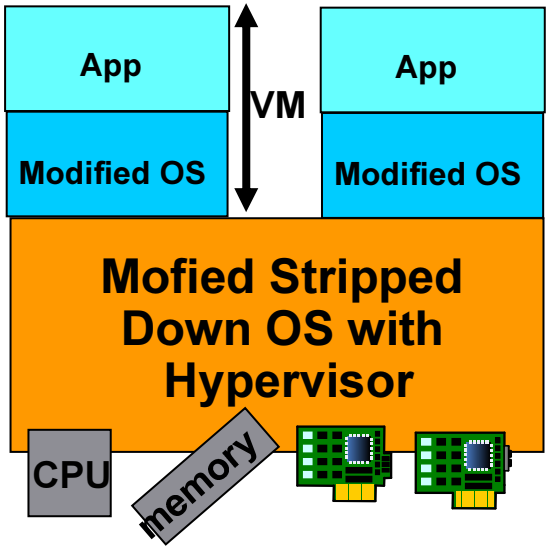
Virtualization



VMware



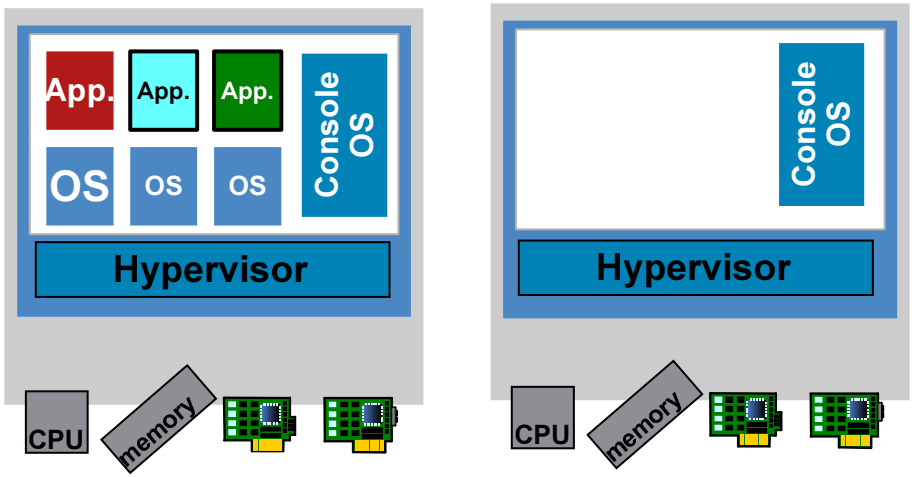
Microsoft



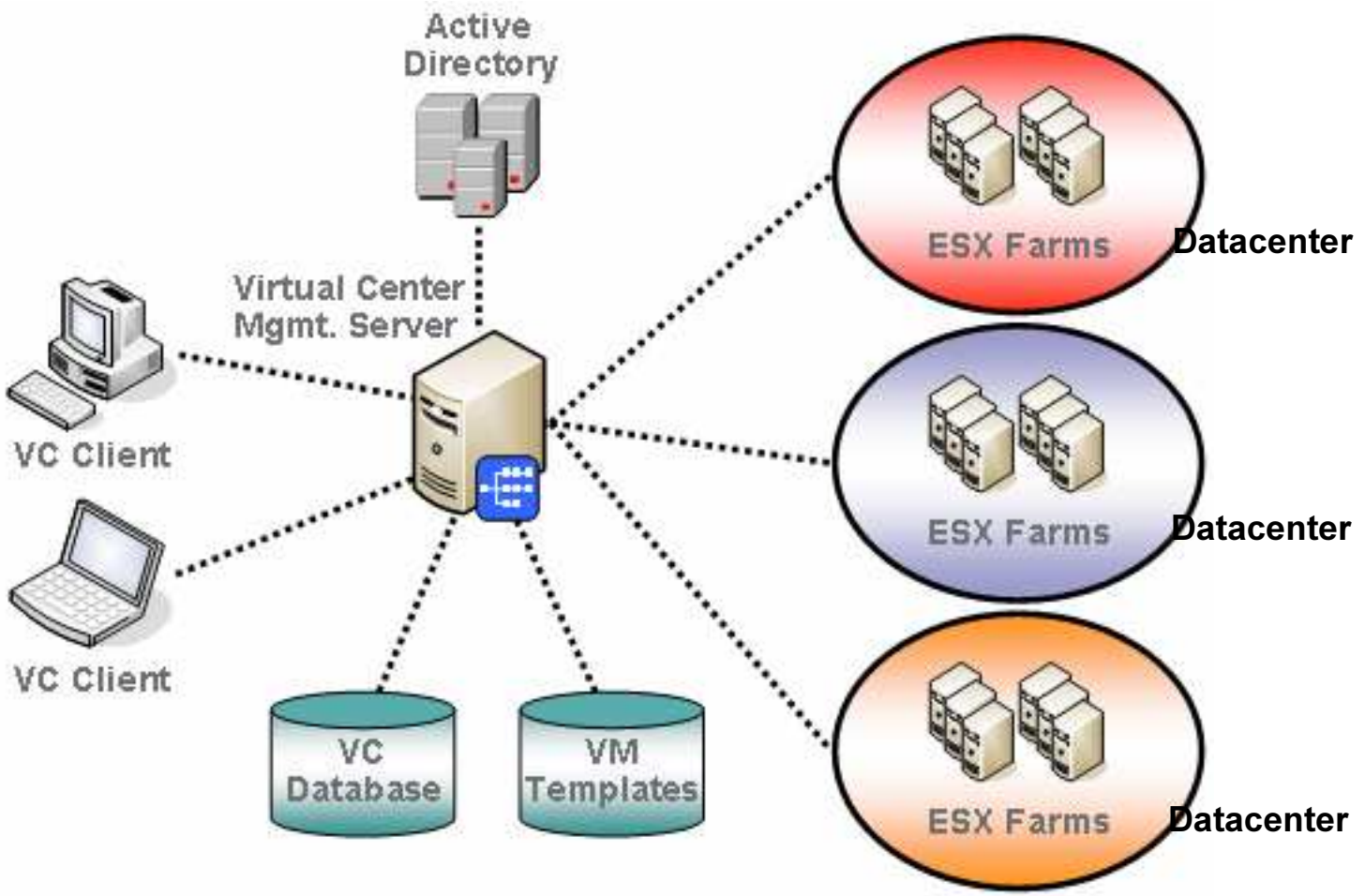
XEN aka Paravirtualization

Migration

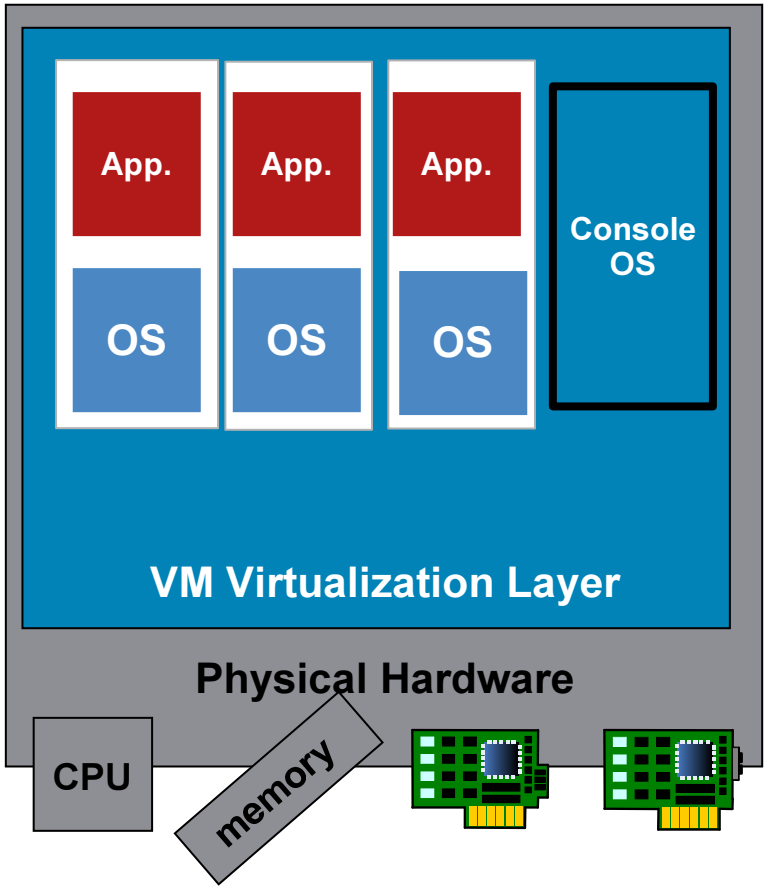
- VMotion, aka VM Migration allows a VM to be reallocated on a different Hardware without having to interrupt service.
- Downtime in the order of few milliseconds to few minutes, not hours or days
- Can be used to perform *Maintenance* on a server,
- Can be used to shift workloads more efficiently
- 2 types of Migration:
  - VMotion Migration
  - Regular Migration

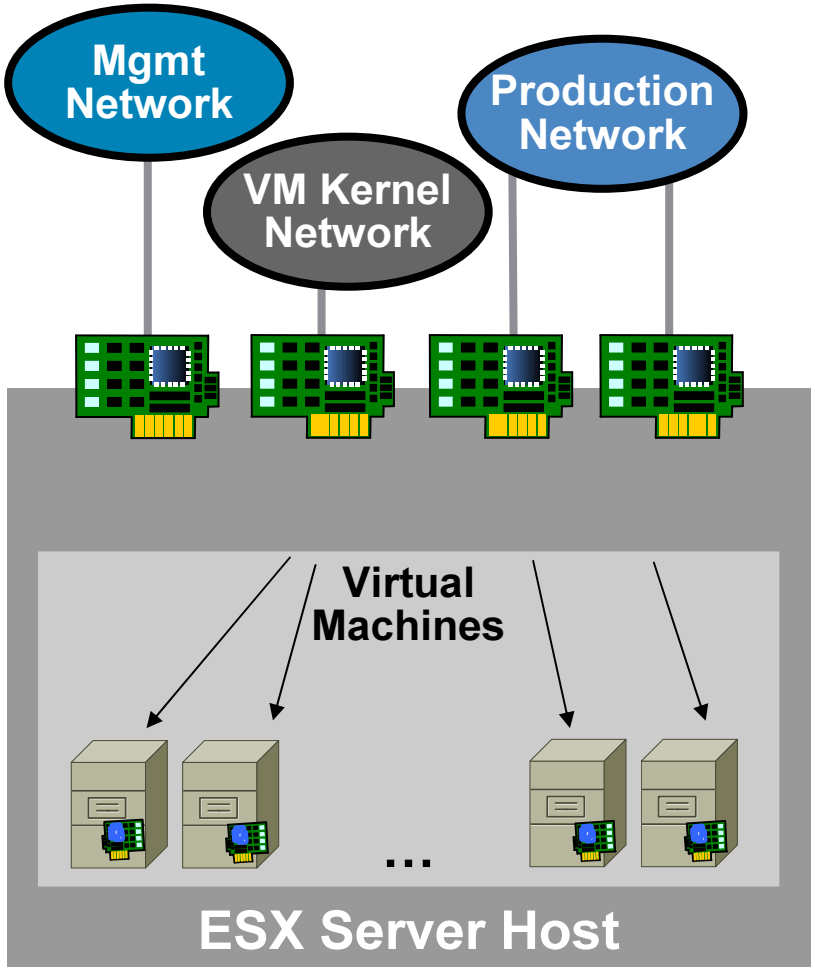


Management

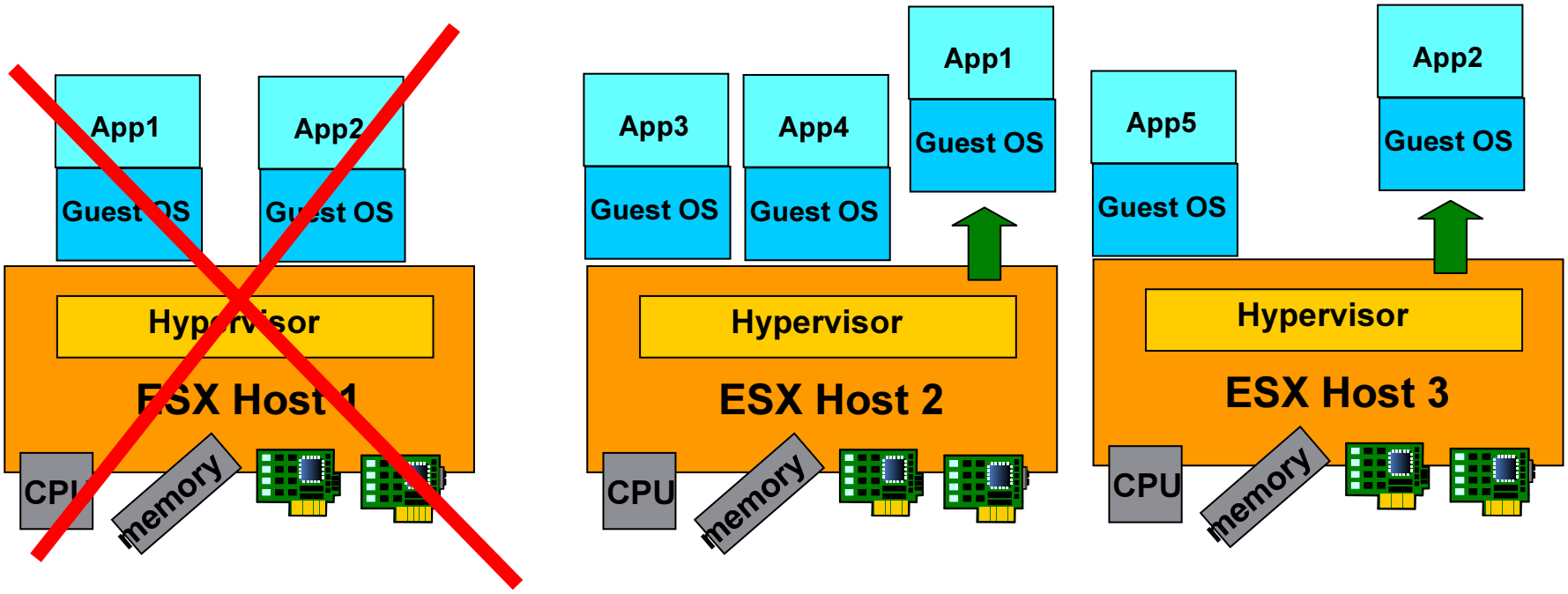


VMware Architecture in a Nutshell

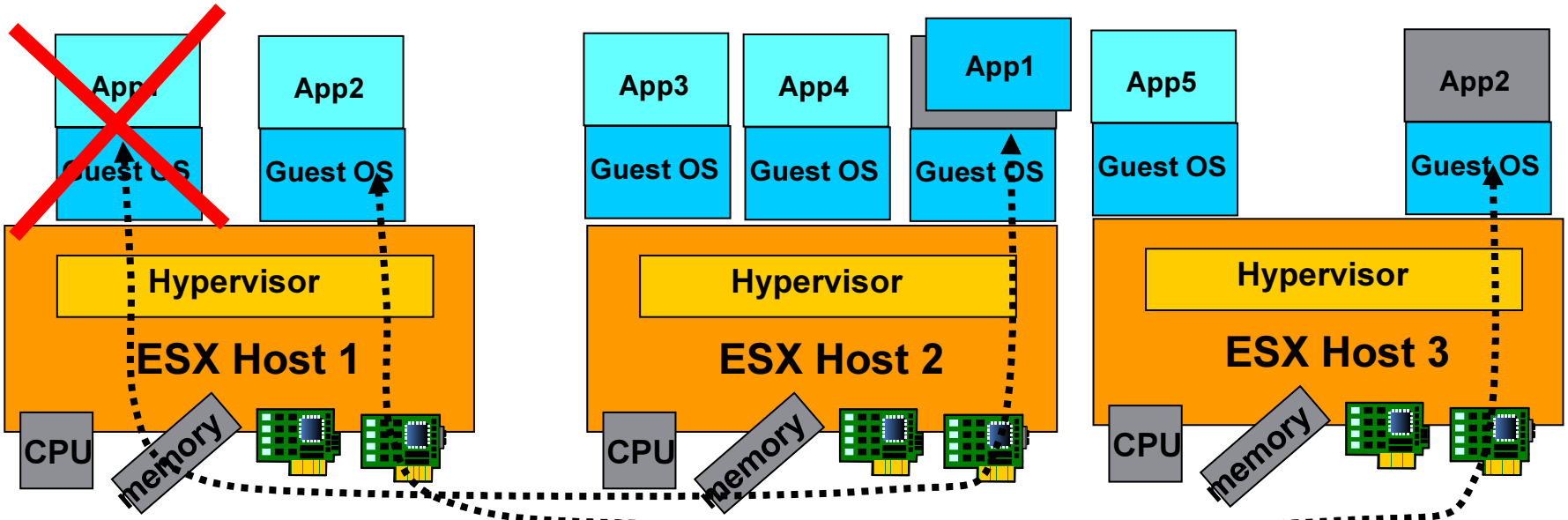




VMware HA Clustering

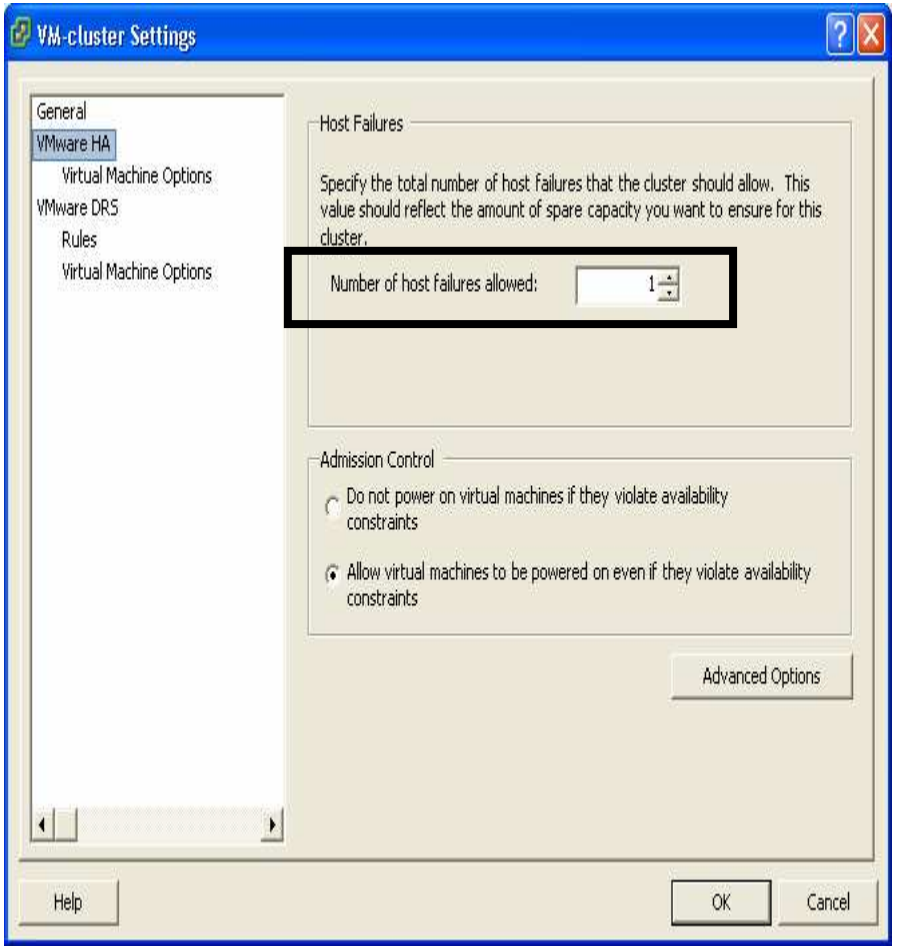


Application-level HA clustering
(Provided by MSCS, Veritas etc…)



HA + DRS

- HA takes care of Powering on VMs on available ESX hosts in the least possible time (regular migration, not VMotion based)
- DRS takes care of migrating the VMs over time to the most appropriate ESX host based on resource allocation (VMotion migration)



Questions

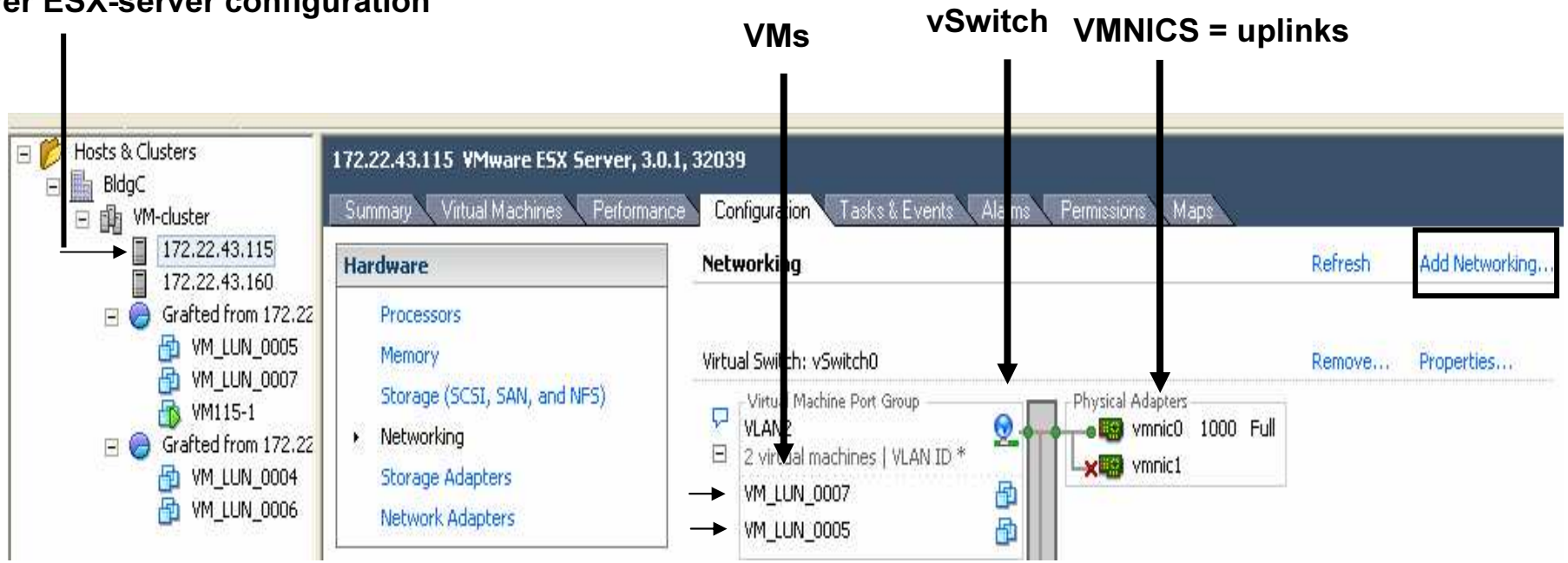
- Which ESX host “interface” is used by Virtual Center to monitor and configure VMs?
- Which ESX host “interface” is used by iSCSI?
- Can I migrate a “powered on” VM from a datacenter to a different one?
- How long does it take for VMware HA to recover from an ESX host failure?
- Does HA clustering require Vmotion?

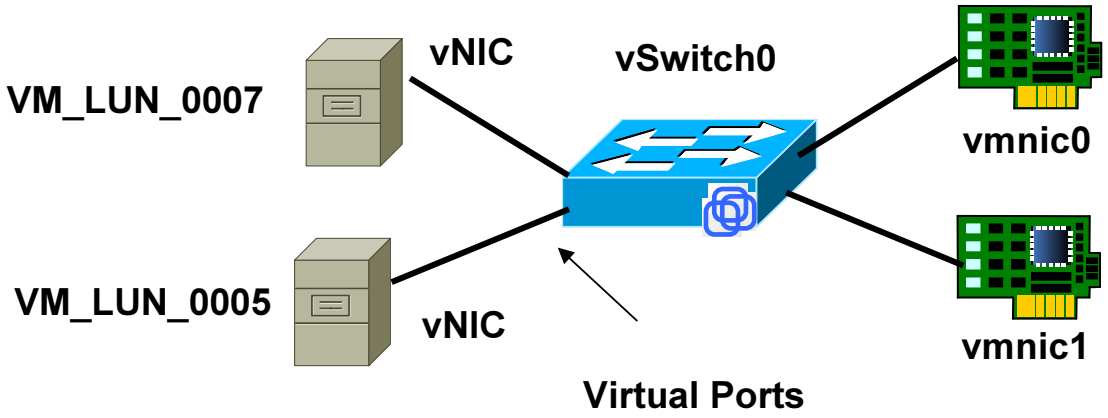
Agenda

- VMware Architecture and Components
- VMware LAN Networking
  - vSwitch Basics
  - NIC Teaming
  - vSwitch vs LAN Switch
  - Migration, HA, DRS
- Cisco/VMware DC DESIGNS
- Blade Server Designs
- Storage Implications of Server Virtualization

VMware Networking Components

Per ESX-server configuration





vNIC MAC address

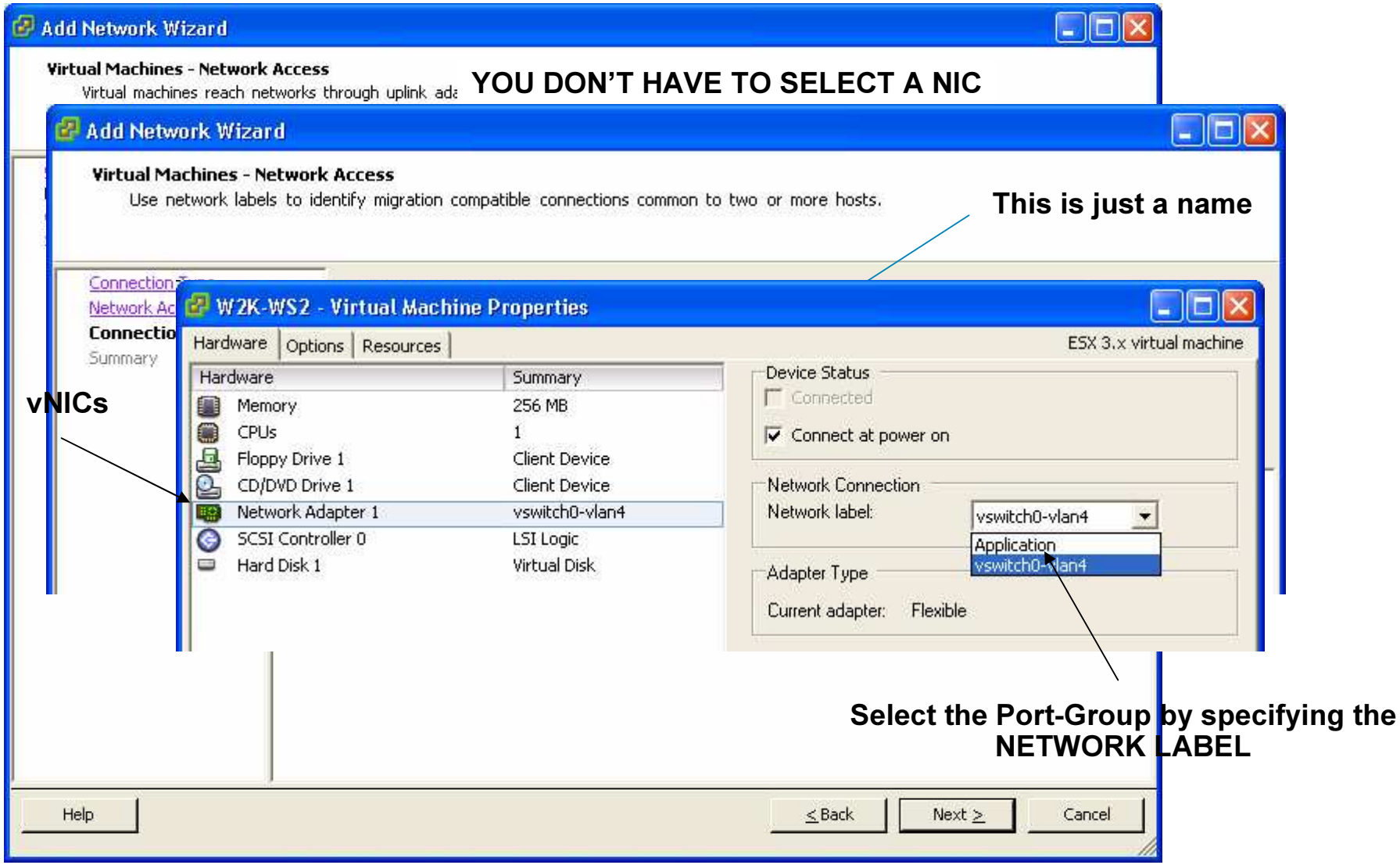
- VM's MAC address automatically generated
- Mechanisms to avoid MAC collision
- VM's MAC address doesn't change with *migration*
- VM's MAC addresses can be made static by modifying the configuration files
- ethernetN.address = 00:50:56:XX:YY:ZZ

- /vmfs/volumes/46b9d79a-2de6e23e-929d-001b78bb5a2c/VM_LUN_0005/VM_LUN_0005.vmx
- ethernet0.addressType = "vpx"
- ethernet0.generatedAddress = "00:50:56:b0:5f:24„
- ethernet0.addressType = „static“
- ethernet0.address = "00:50:56:00:00:06„

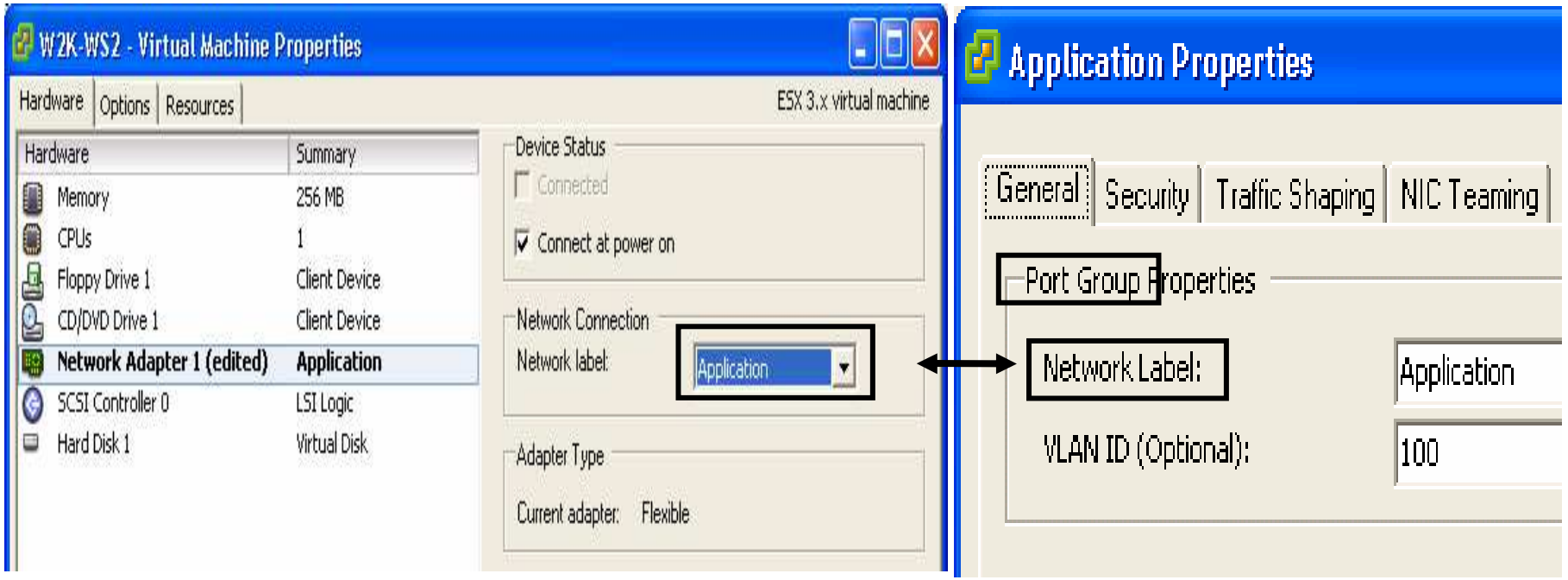
vSwitch Forwarding Characteristics

- Forwarding based on MAC address (No Learning): If traffic doesn't match a VM MAC is sent out to vmnic
- VM-to-VM traffic stays local
- Vswitches TAG traffic with 802.1q VLAN ID
- vSwitches are 802.1q Capable
- vSwitches can create Etherchannels

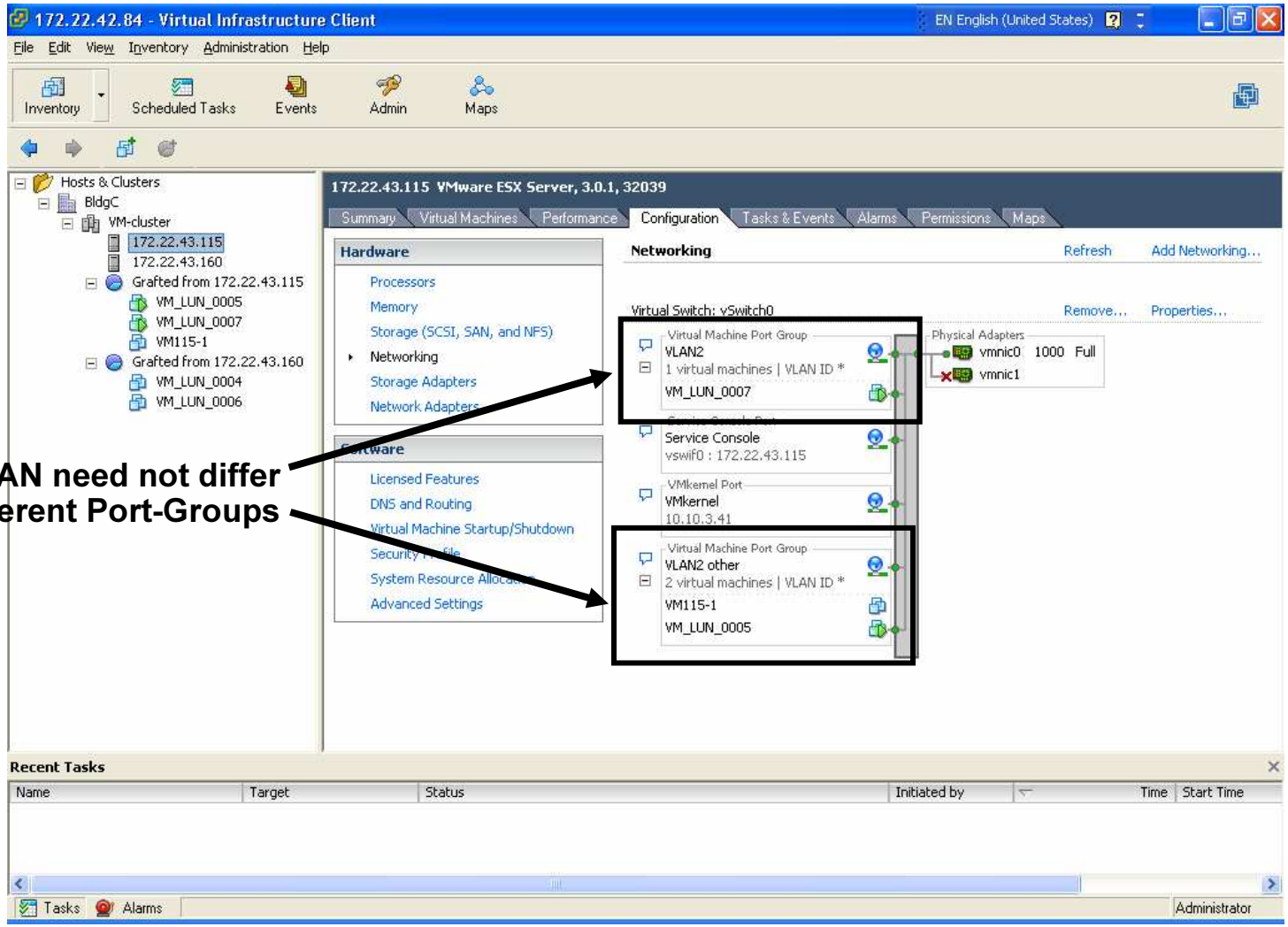
vSwitch Creation



VM ⇔ Port-Group ⇔vSwitch

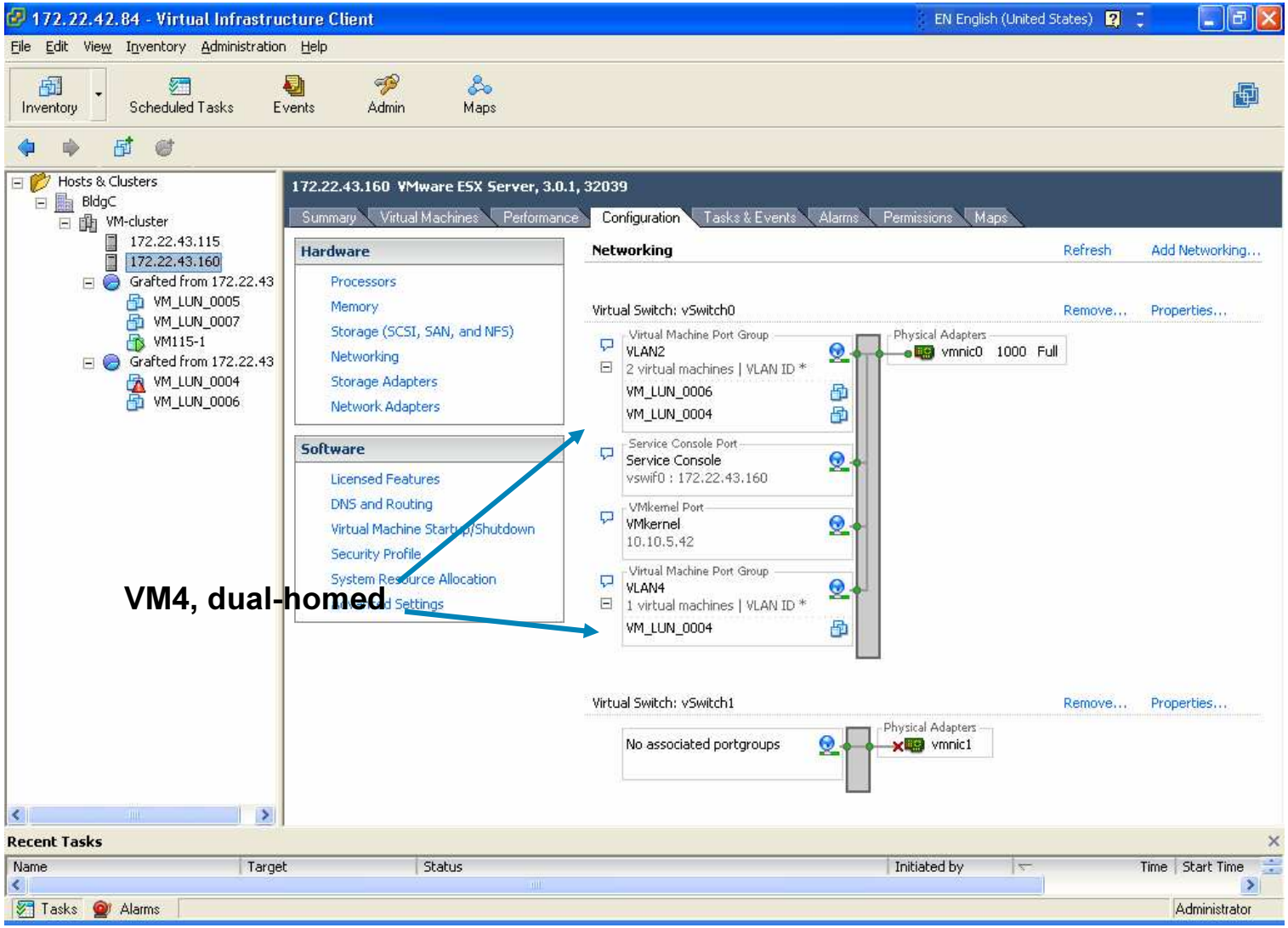


Example Configuration
Multiple Port-Groups, same VLAN



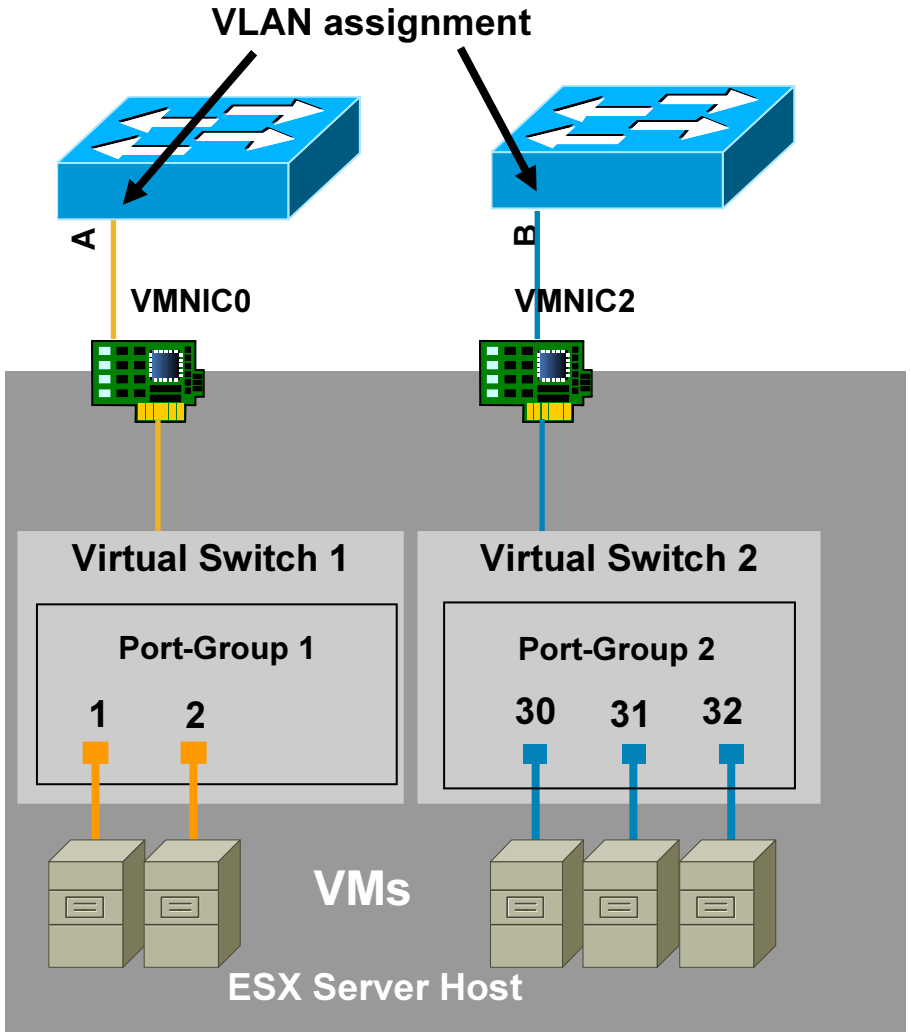
The VLAN need not differ on different Port-Groups

VM with 2 vNIC to same vSwitch



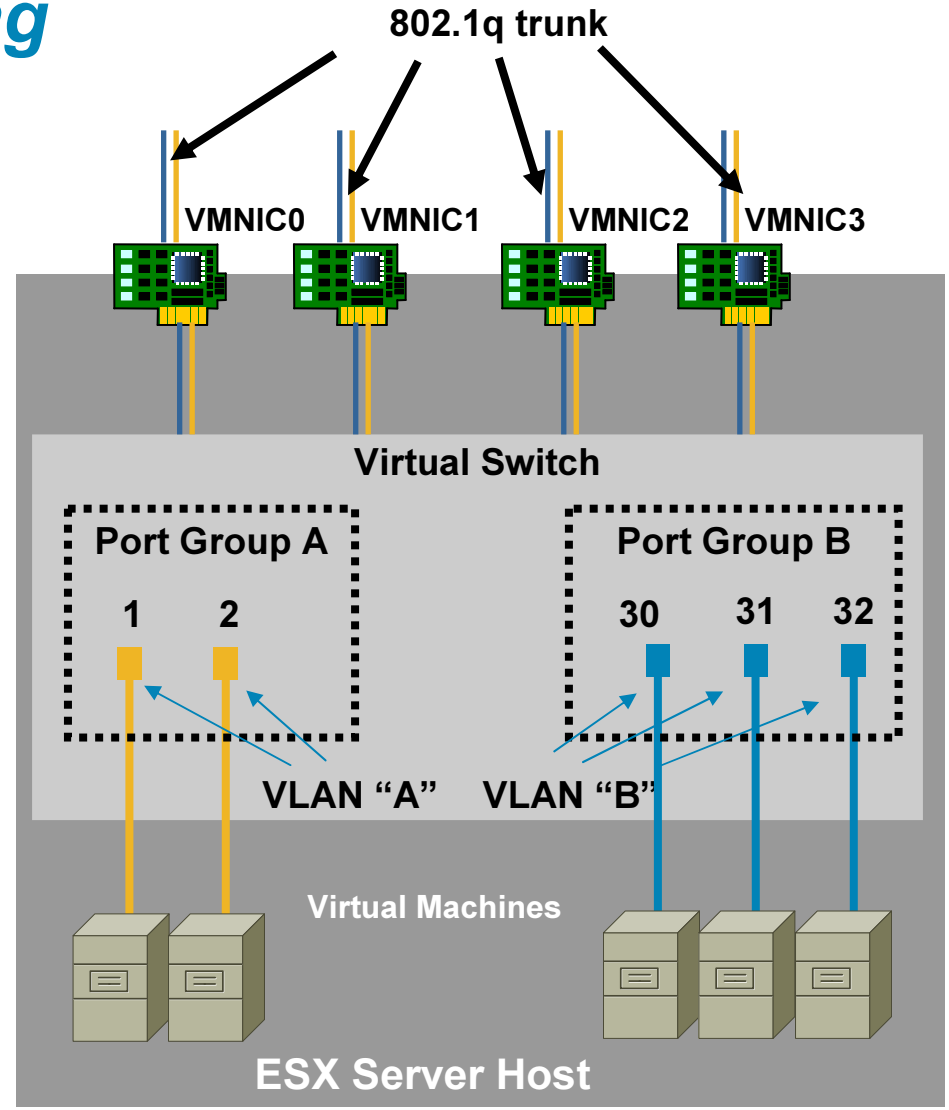
VLAN Tagging Options
External Switch Tagging

- External switch tags packet
- Configured by setting the Network Label VLAN ID to be 0
- How is VM-toVM traffic switched? (through LAN Switch or through vSwitch)
- Can use native VLAN on 802.1q trunk (as long as native VLAN is not tagged)



VLAN Tagging Options
Virtual Switch Tagging

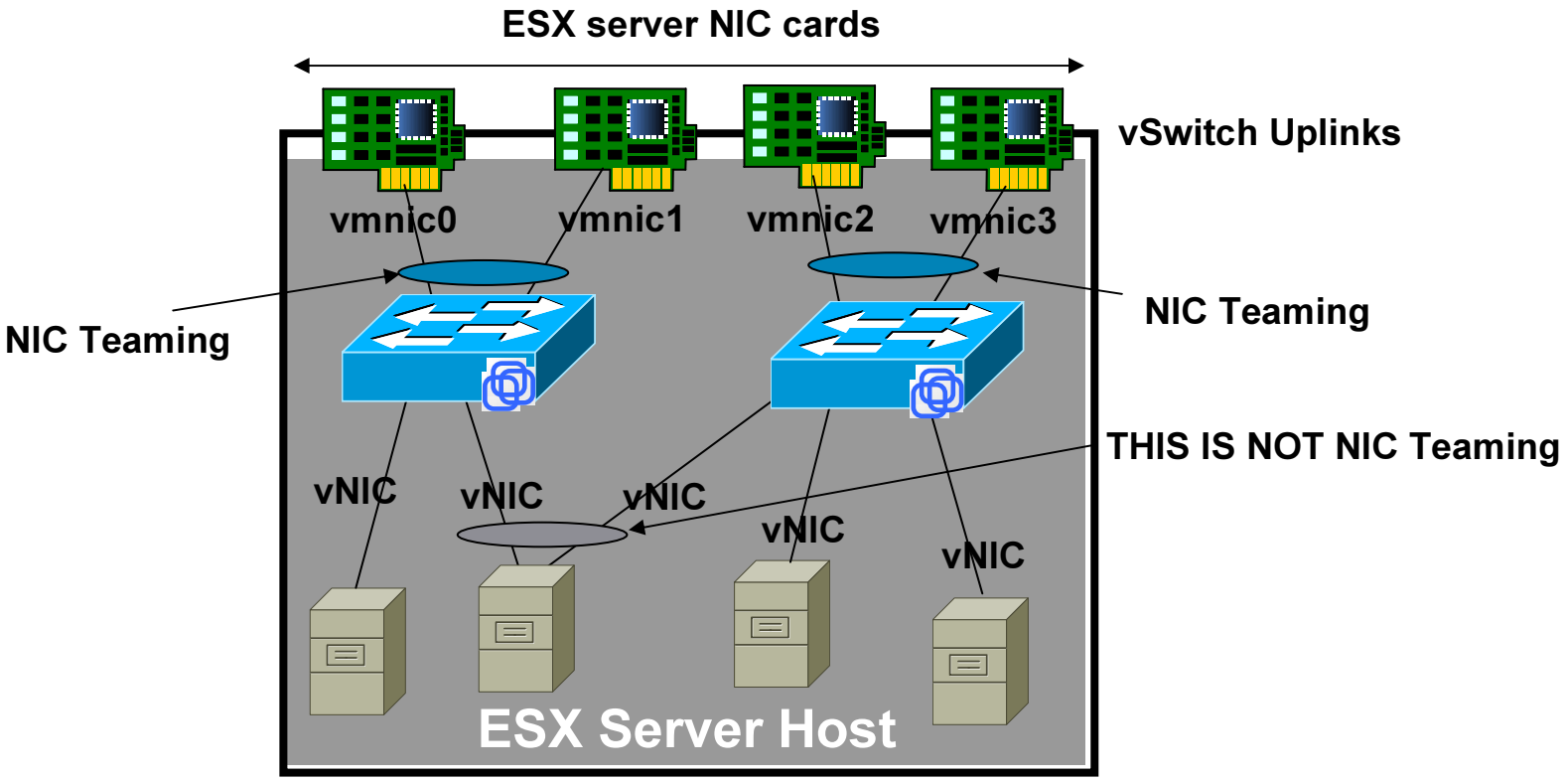
- vSwitch tagging
  - Tags outbound packets
  - Strips tag from inbound packet
- Most Common Deployment
- It is set by assigning the VLAN ID to the Network Label in the Port-Group
- Provides isolation between VLANs



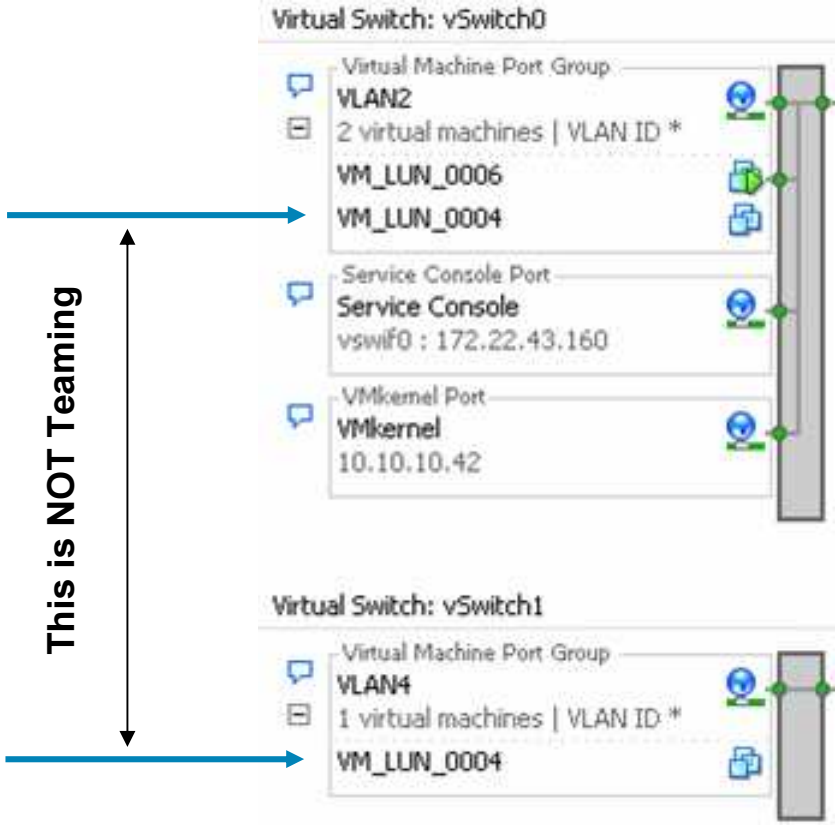
Agenda

- VMware Architecture and Components
- VMware LAN Networking
  - vSwitch Basics
  - NIC Teaming
  - vSwitch vs LAN Switch
  - Migration, HA, DRS
- Cisco/VMware DC DESIGNS
- Blade Server Designs
- Storage Implications of Server Virtualization

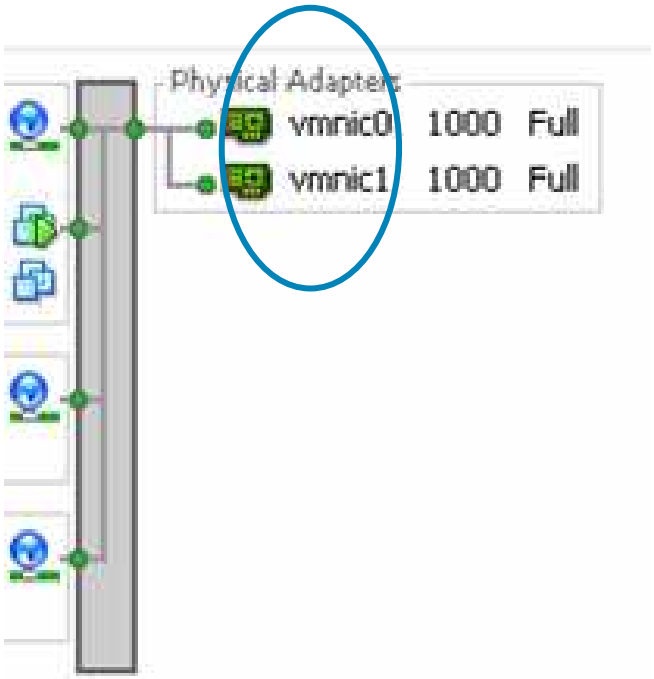
Meaning of NIC Teaming in VMware (1)



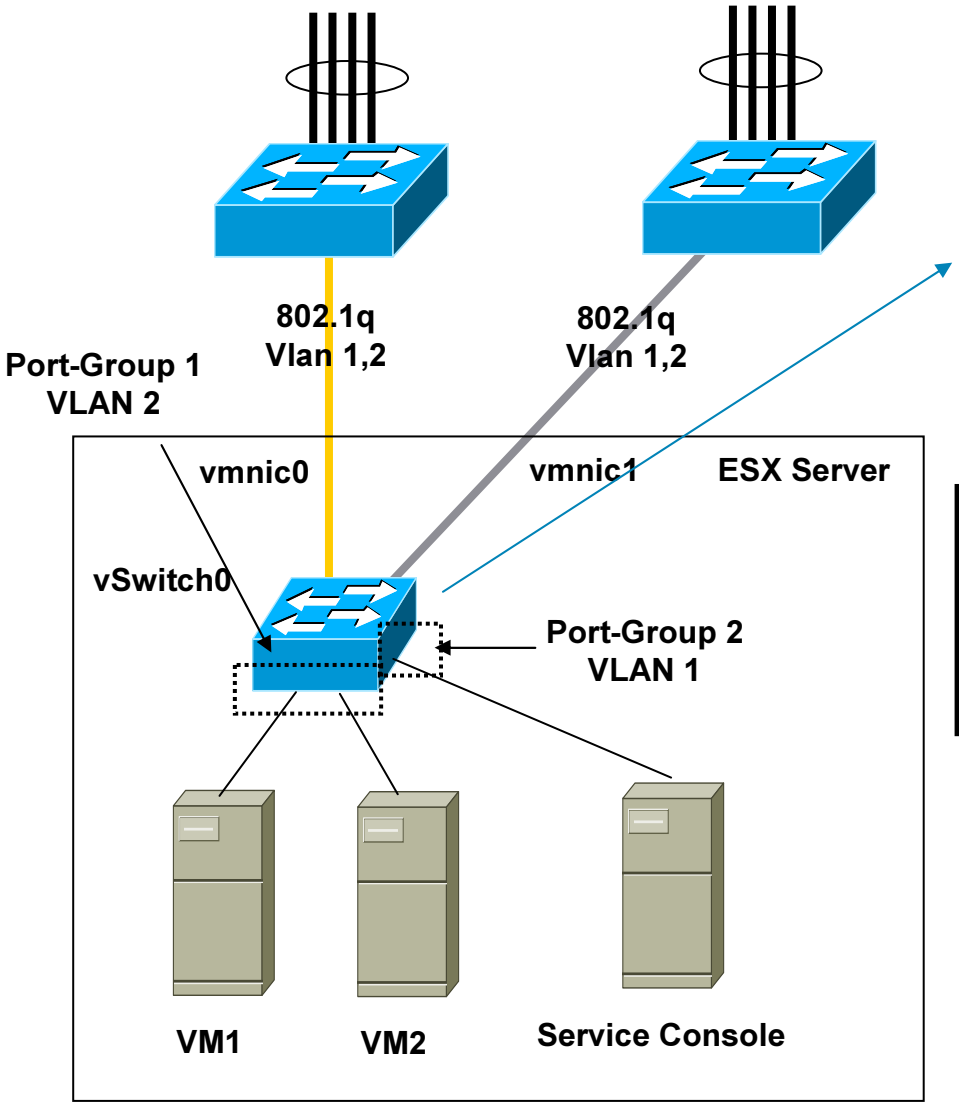
Meaning of NIC Teaming in VMware (2)

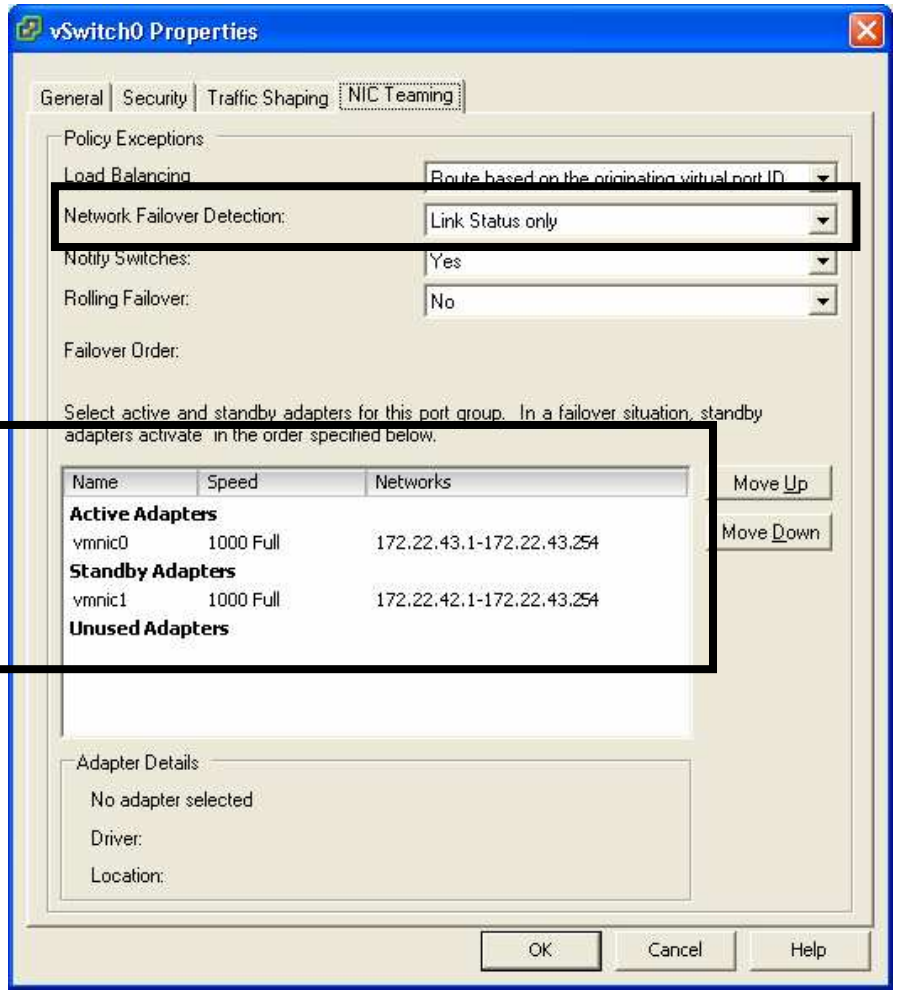


Teaming is Configured at The vmnic Level

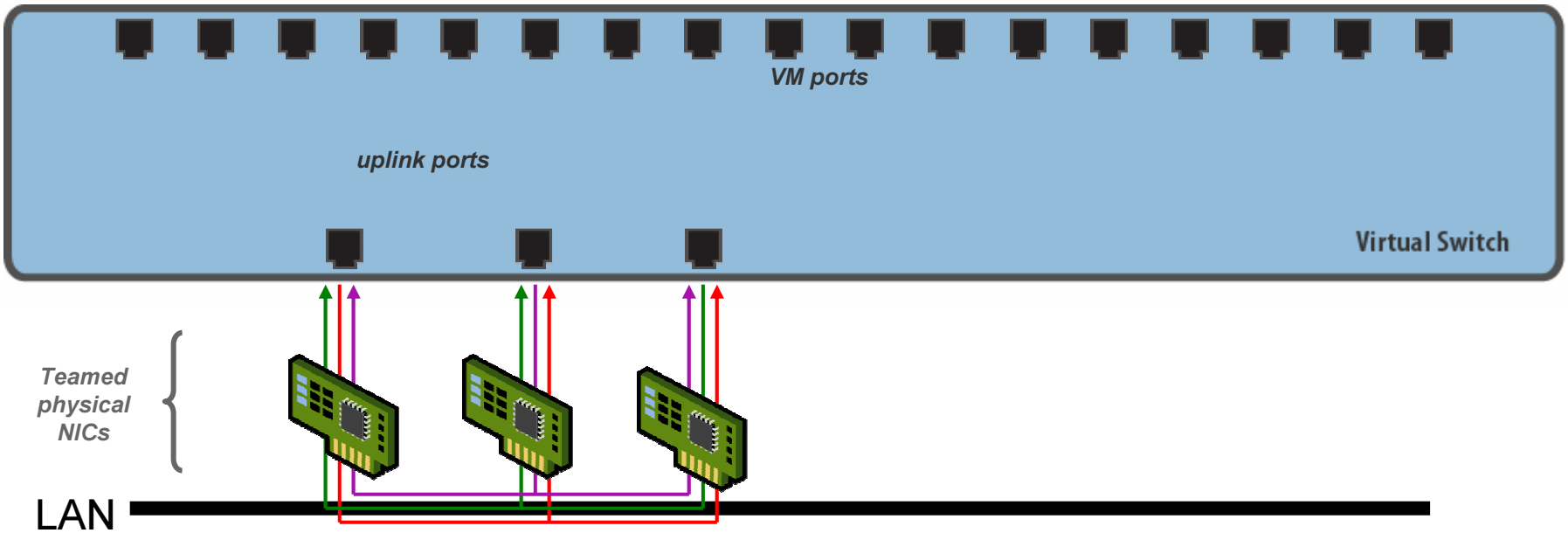


Design Example
2 NICs, VLAN 1 and 2, Active/Standby



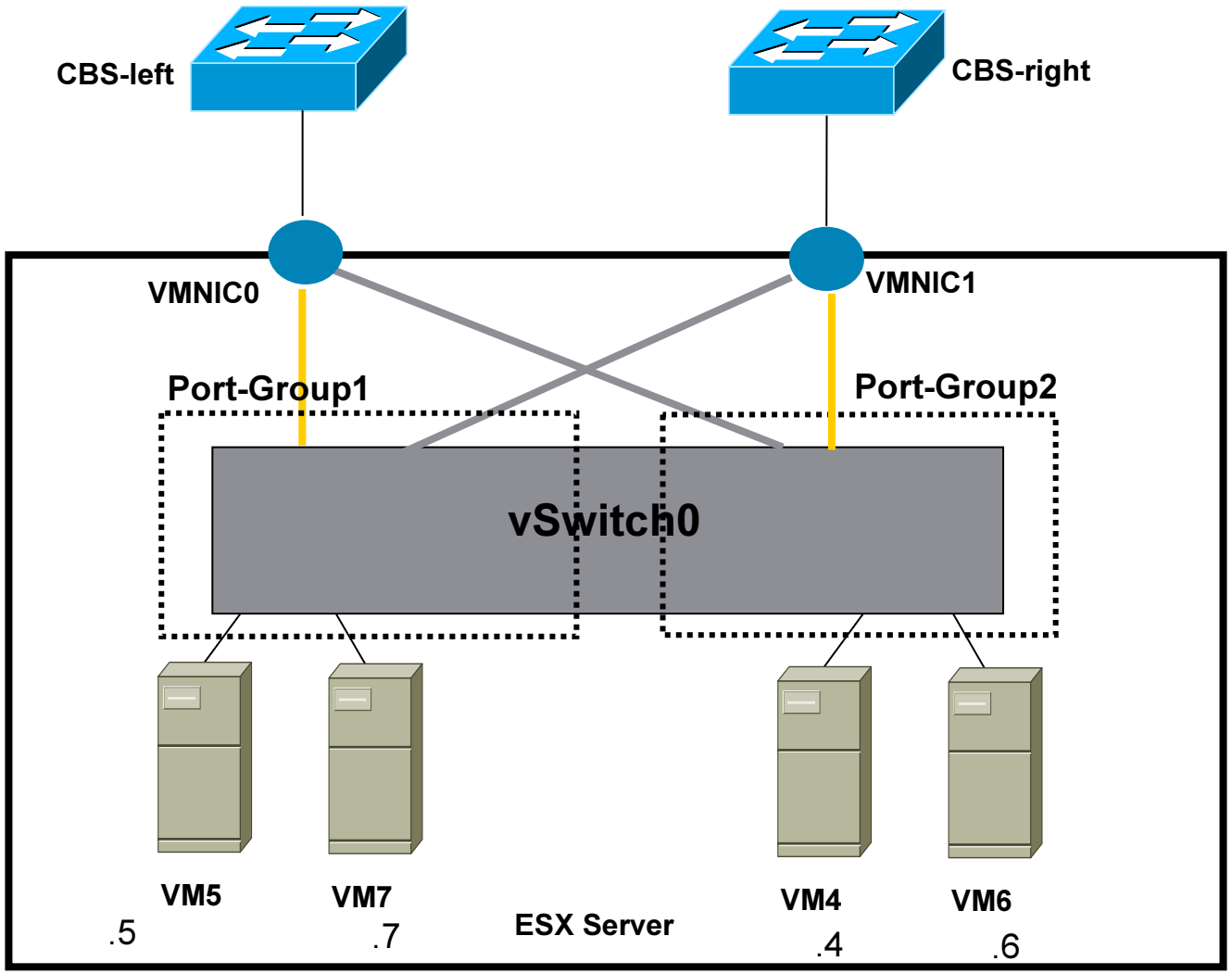


Beacon Probing

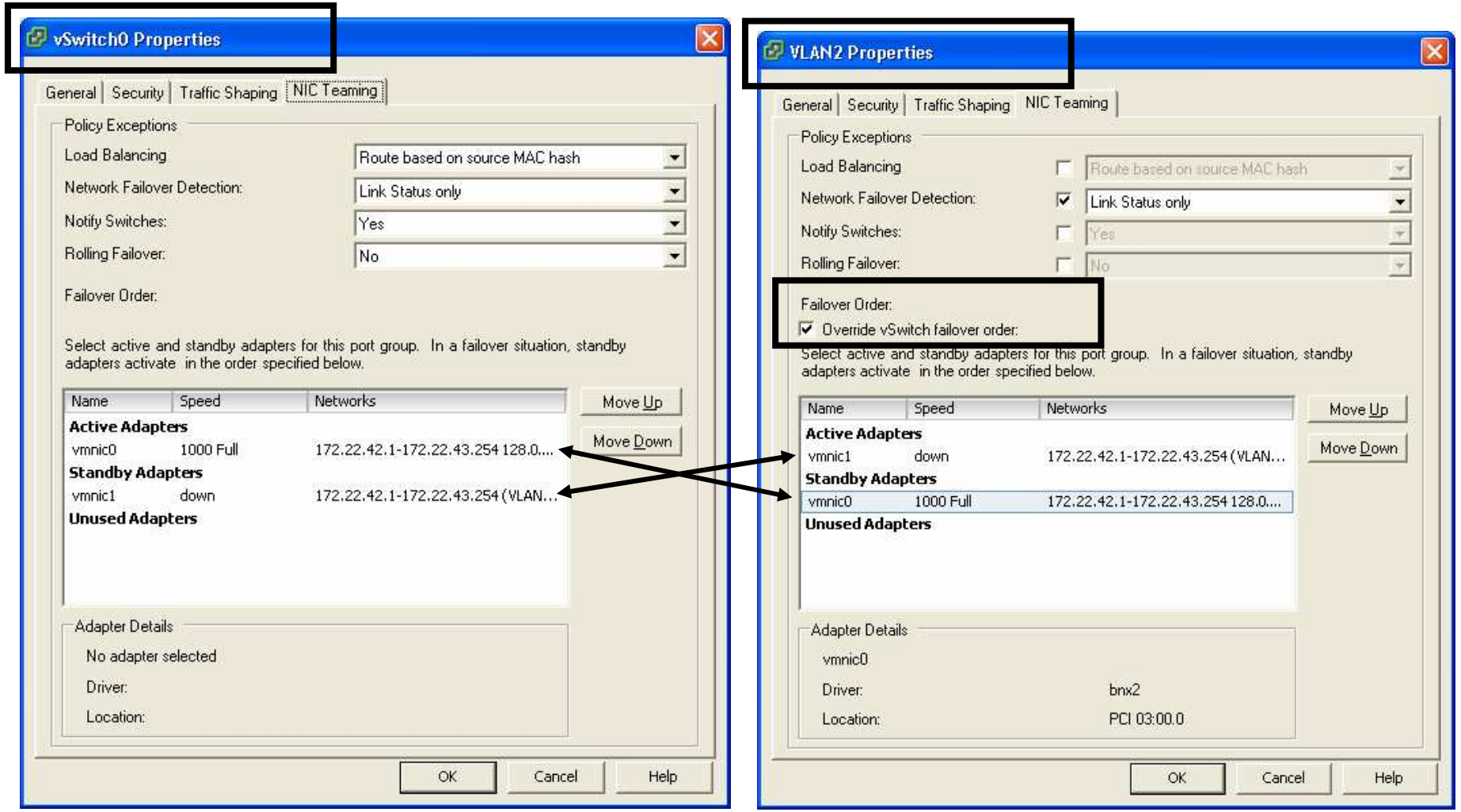


- Beacon probing attempts to detect failures which don't result in a link state failure for the NIC
- Broadcast frames sent from each NIC in the team should be seen by other NICs in the team
- Beacons are sent on each VLAN in use

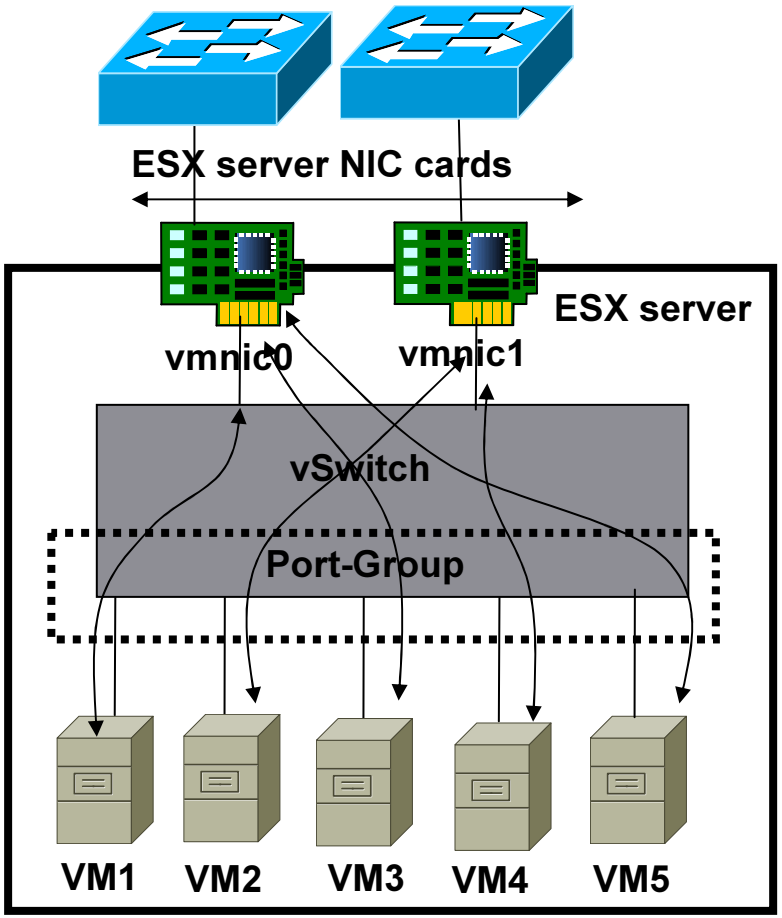
Active/Standby per-Port-Group

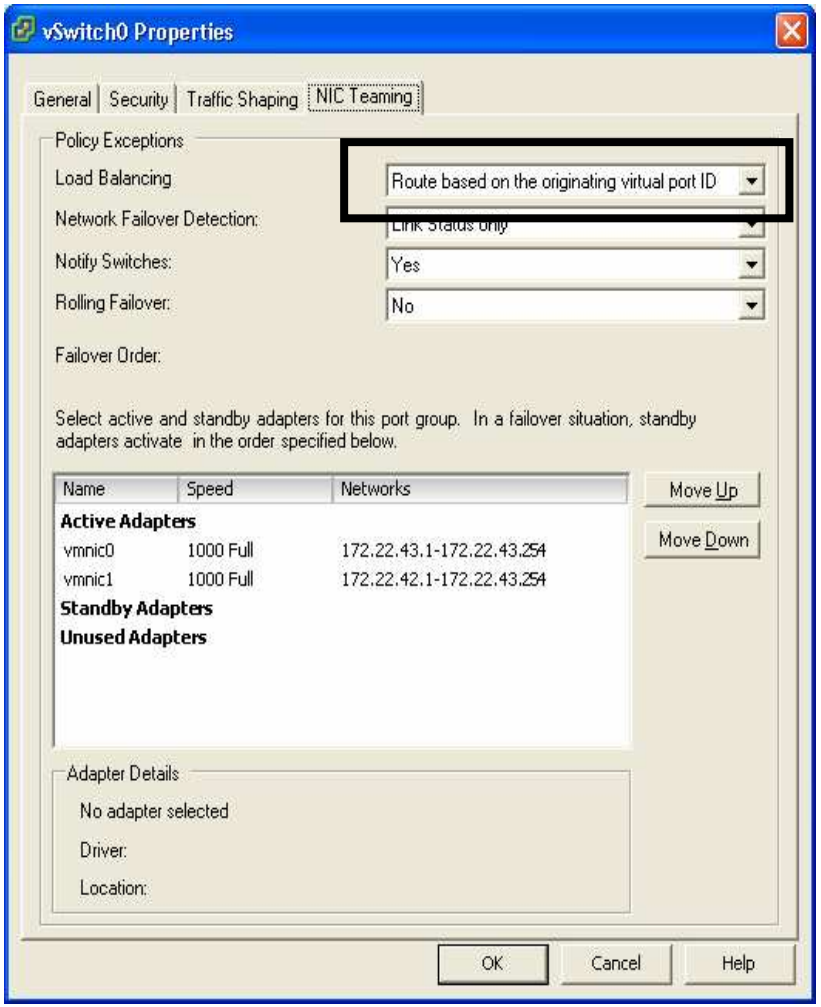


Port-Group overrides vSwitch Global Configuration



Active/Active





Active/Active
IP-based Load Balancing

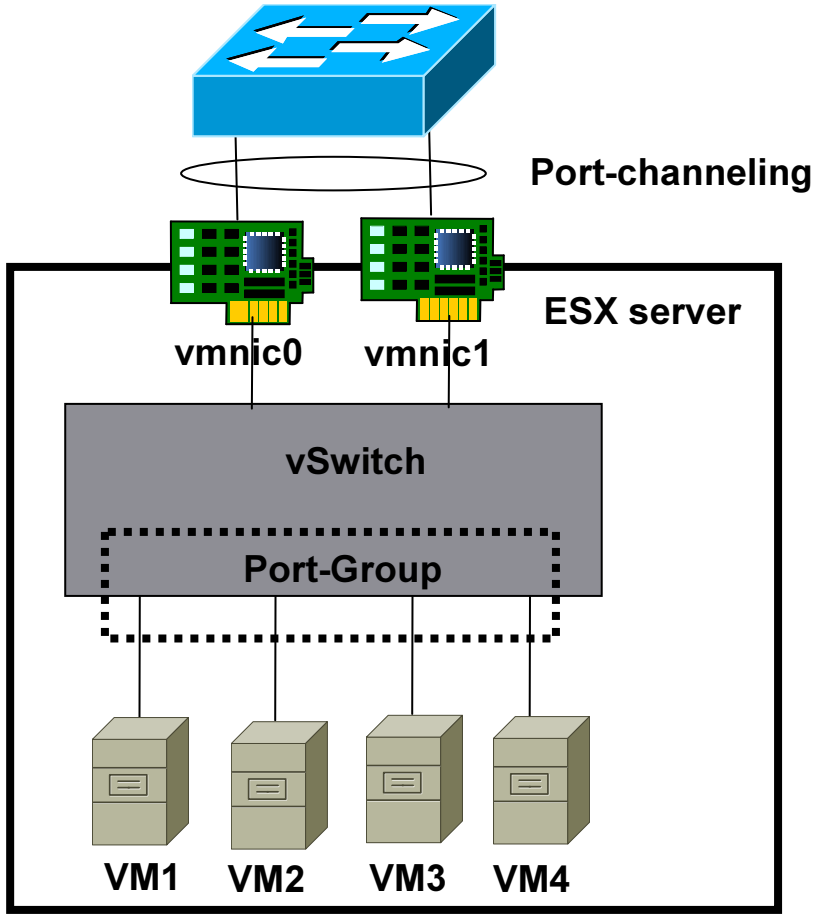
- Works with Channel-Group mode ON
- LACP is not supported (see below):

  9w0d: %LINK-3-UPDOWN: Interface GigabitEthernet1/0/14, changed state to up

  9w0d: %LINK-3-UPDOWN: Interface GigabitEthernet1/0/13, changed state to up

  9w0d: %EC-5-L3DONTBNDL2: Gi1/0/14 suspended: LACP currently not enabled on the remote port.

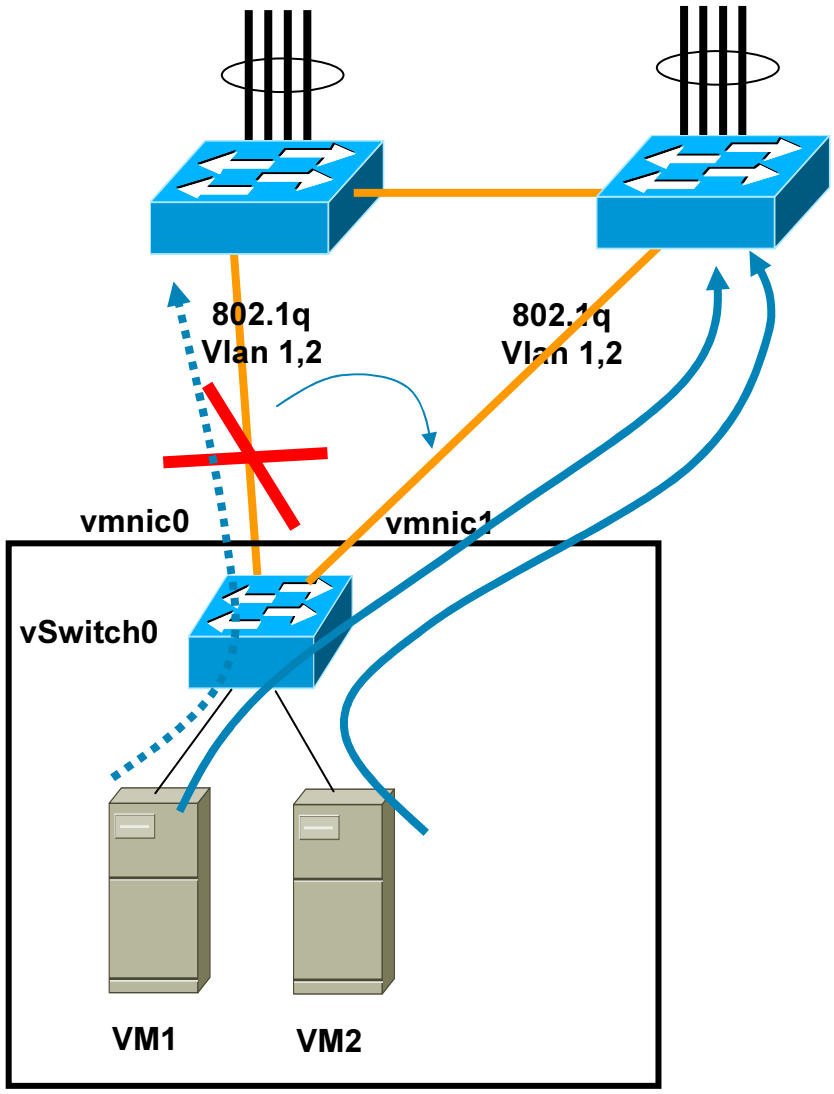
  9w0d: %EC-5-L3DONTBNDL2: Gi1/0/13 suspended: LACP currently not enabled on the remote port.

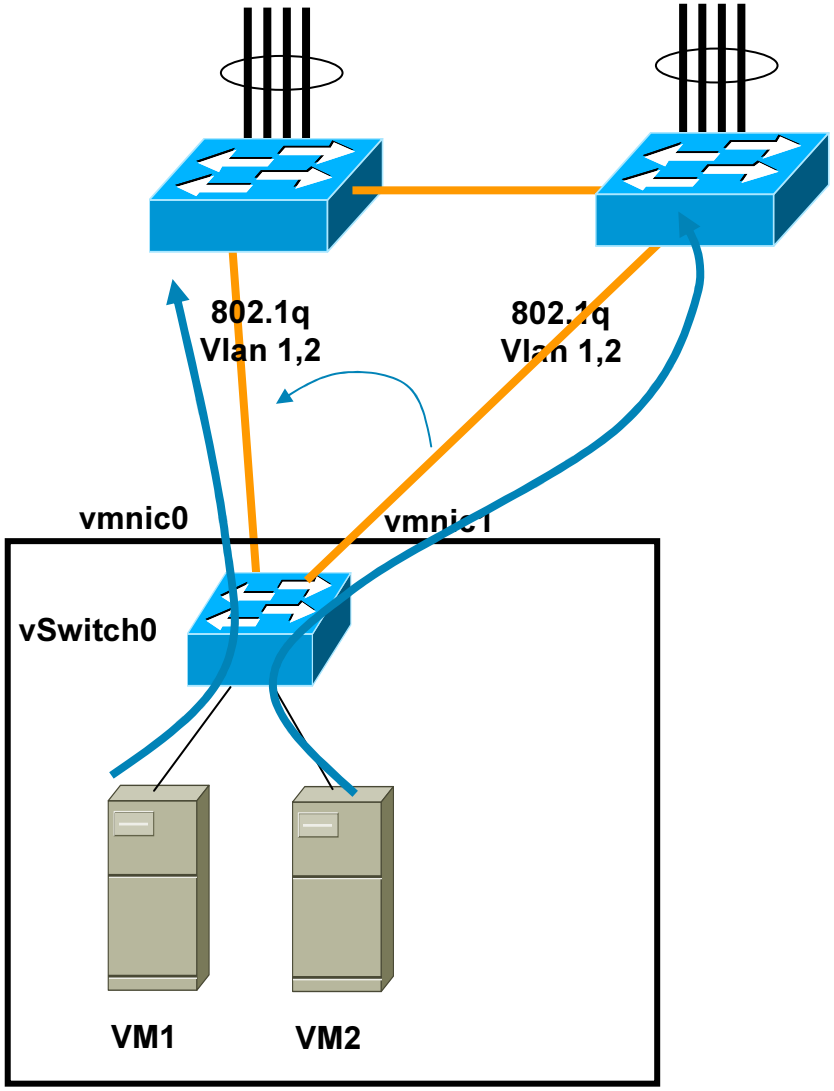


Agenda

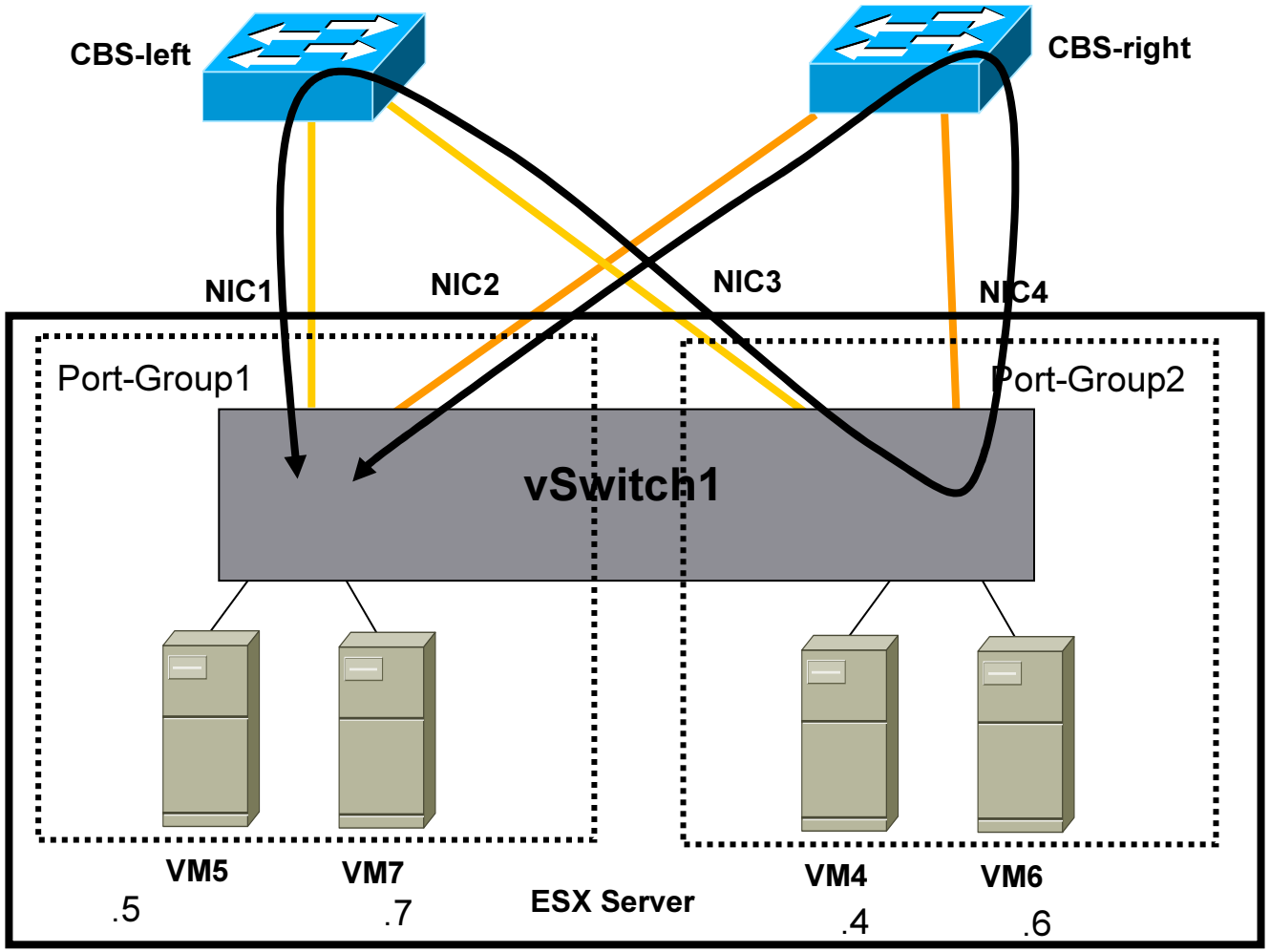
- VMware Architecture and Components
- VMware LAN Networking
  - vSwitch Basics
  - NIC Teaming
  - vSwitch vs LAN Switch
  - Migration, HA, DRS
- Cisco/VMware DC DESIGNS
- Blade Server Designs
- Storage Implications of Server Virtualization

Rolling Failover (aka Preemption)
By default Preemption is on

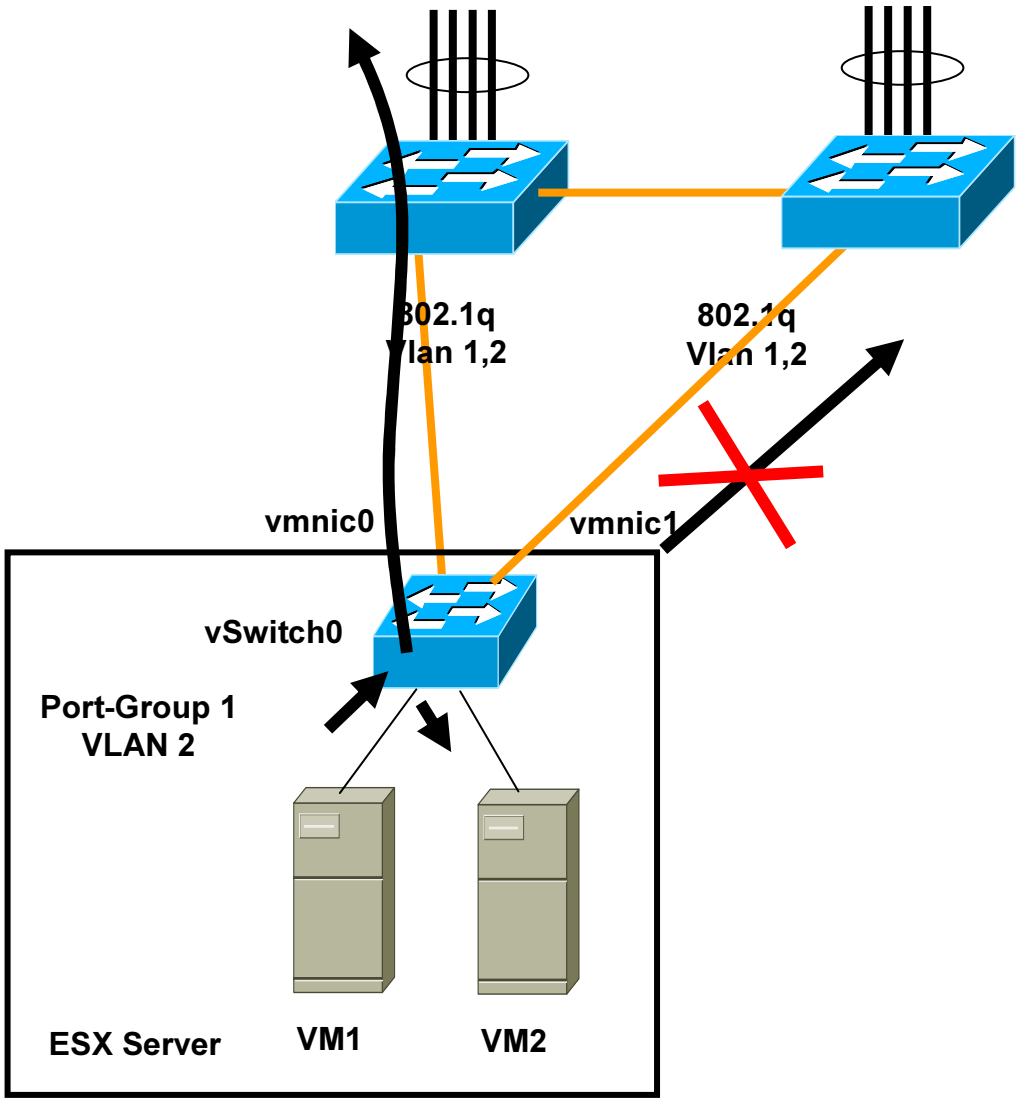




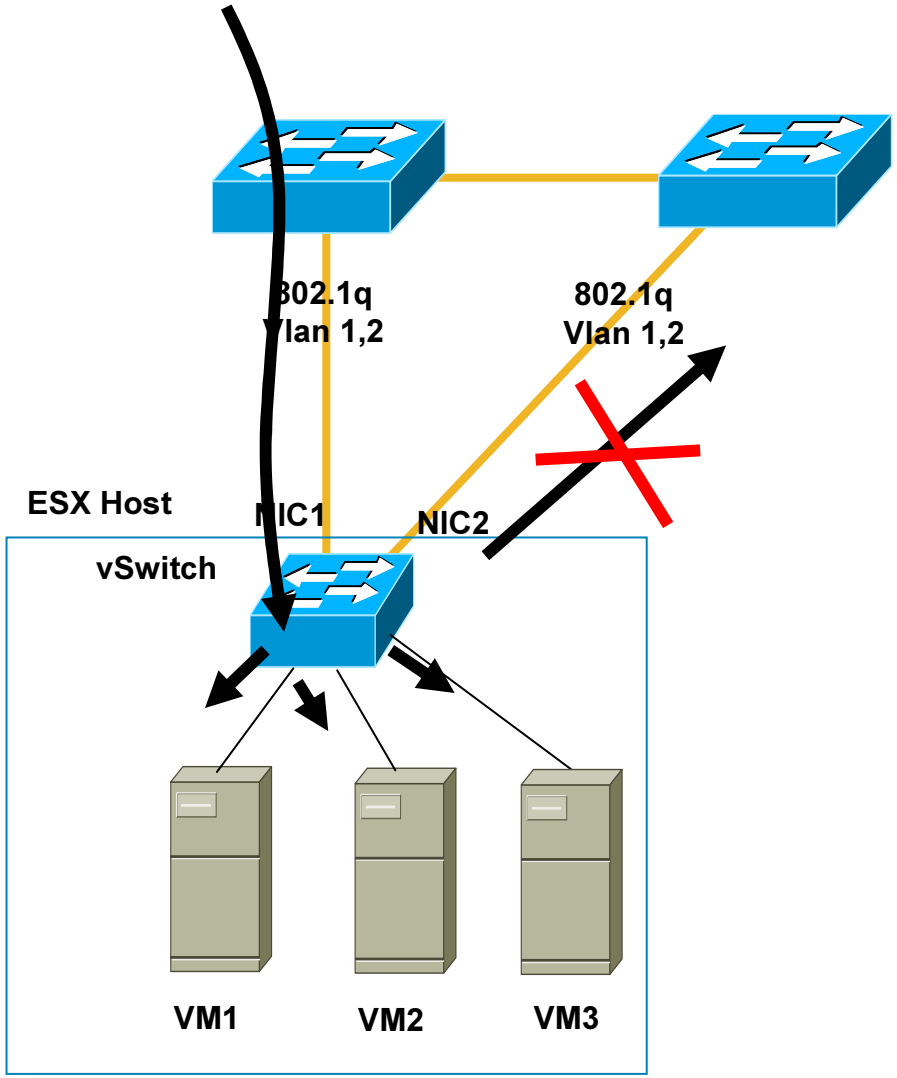
All Links Active, No Spanning-Tree
Is there a Loop?



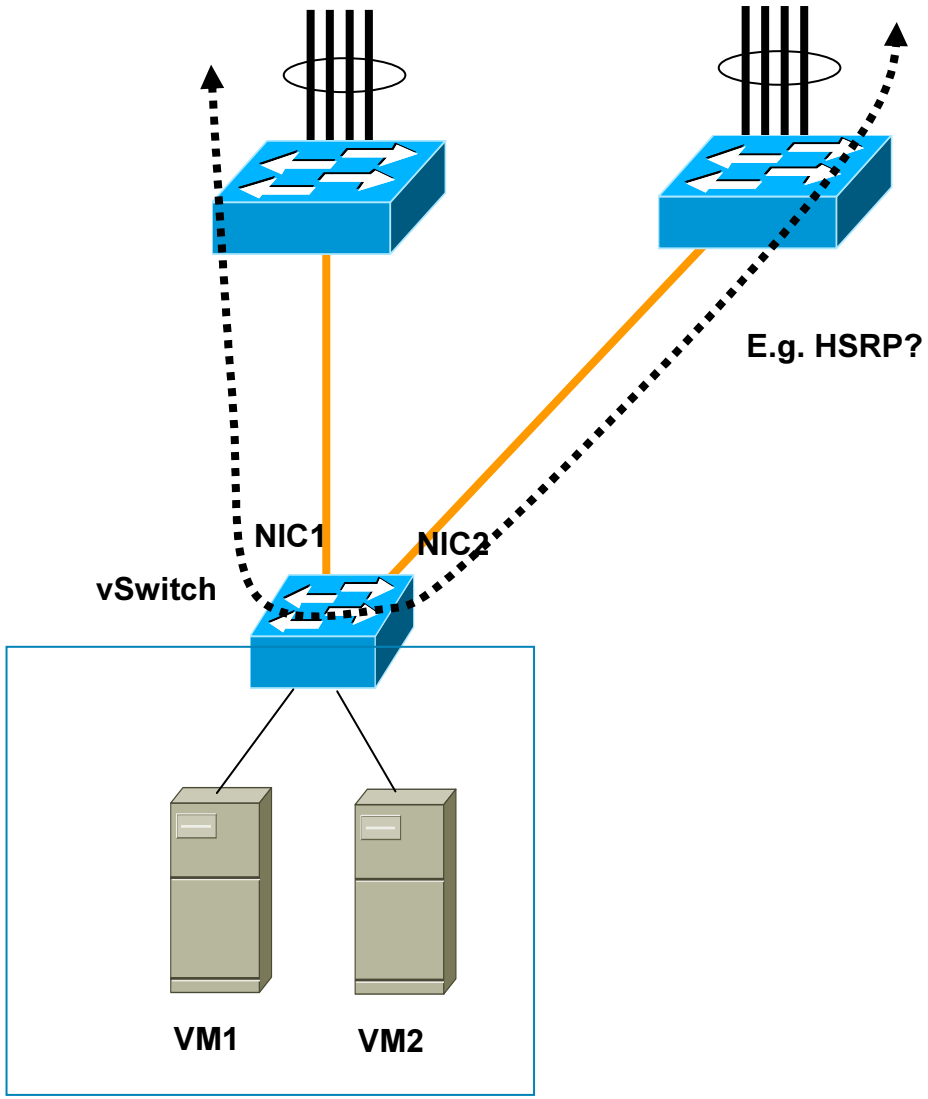
Broadcast/Multicast/Unknown Unicast Forwarding in Active/Active (1)



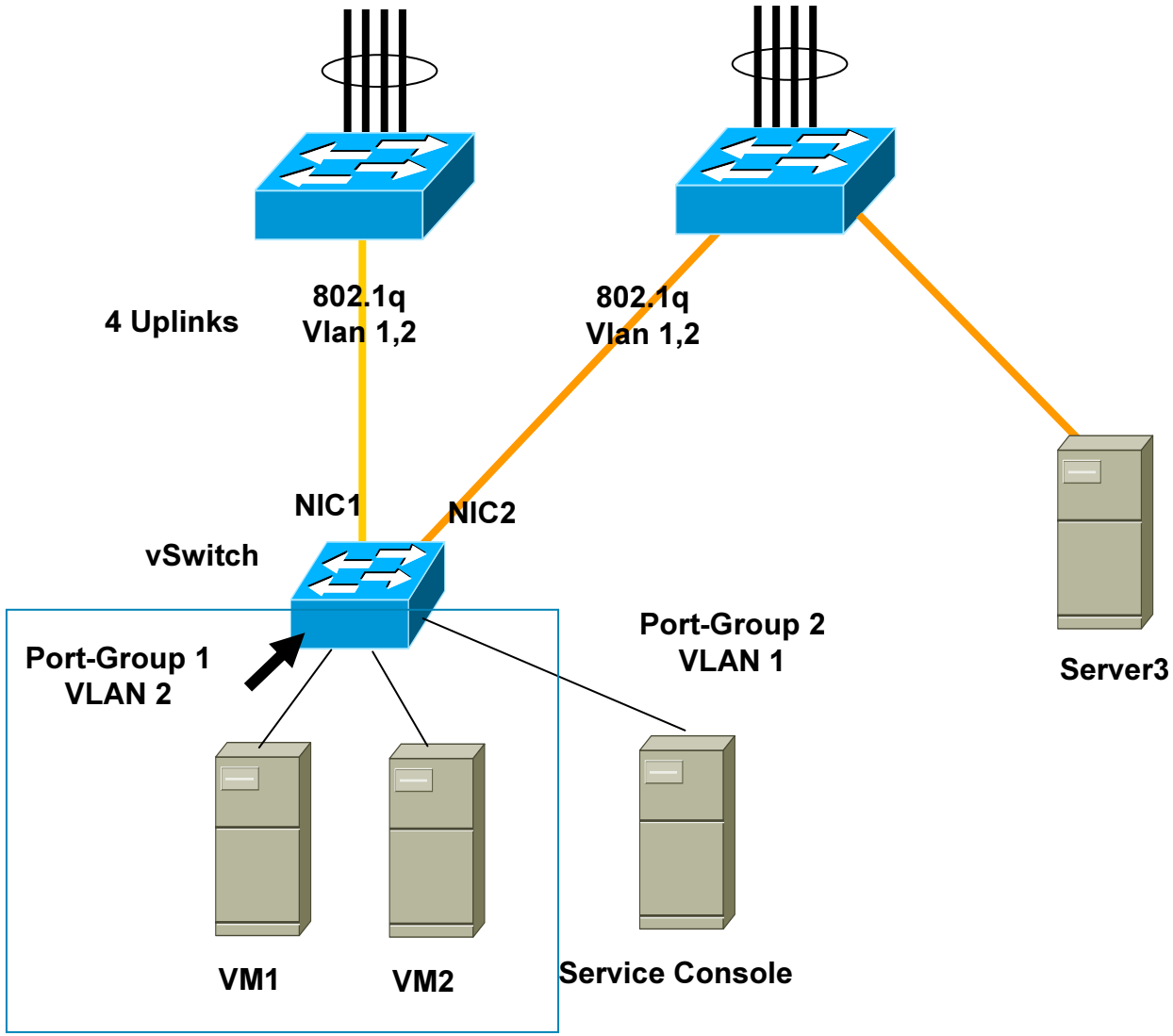
Broadcast/Multicast/Unknown Unicast Forwarding in Active/Active (2)



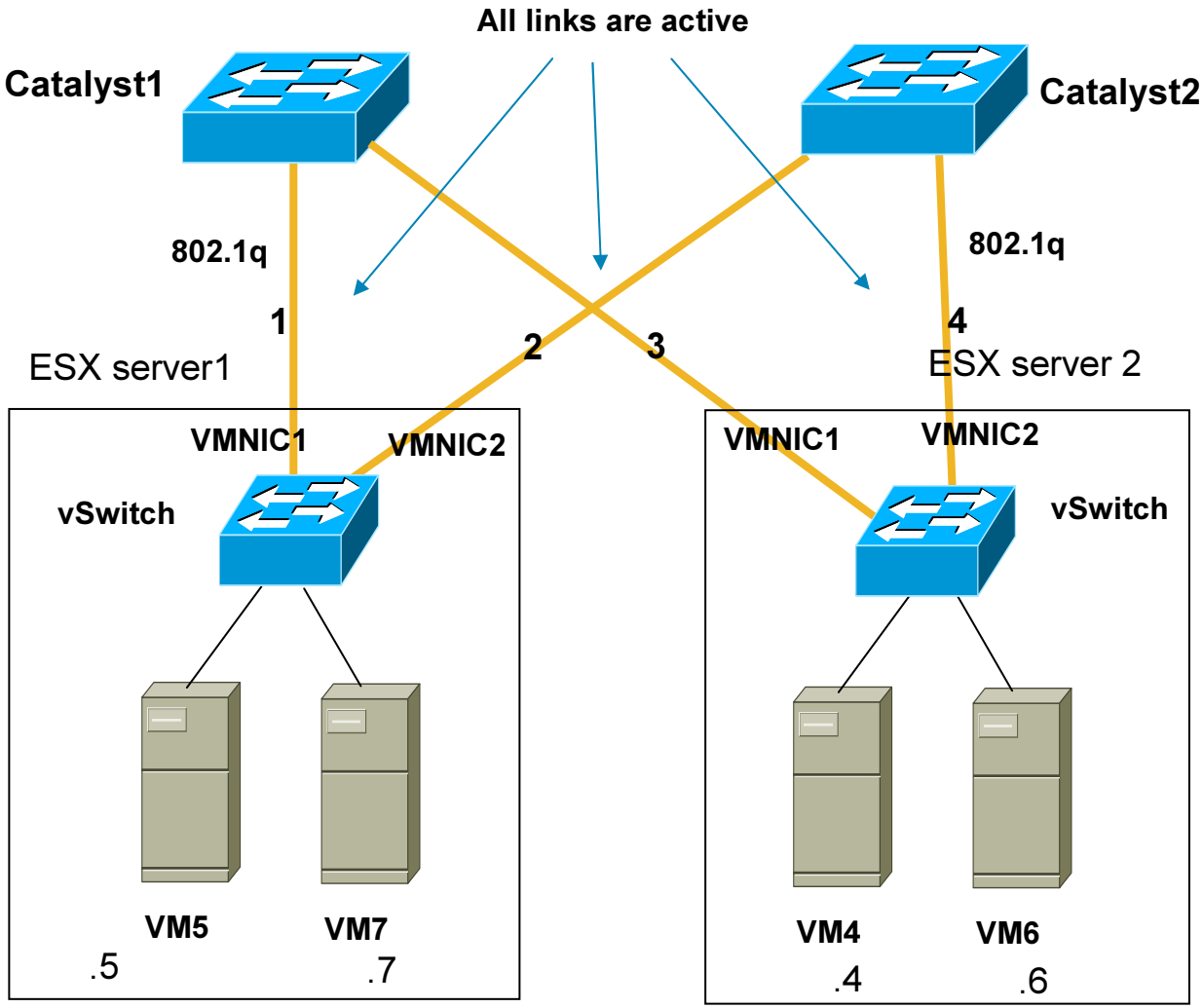
Can the vSwitch pass traffic through?



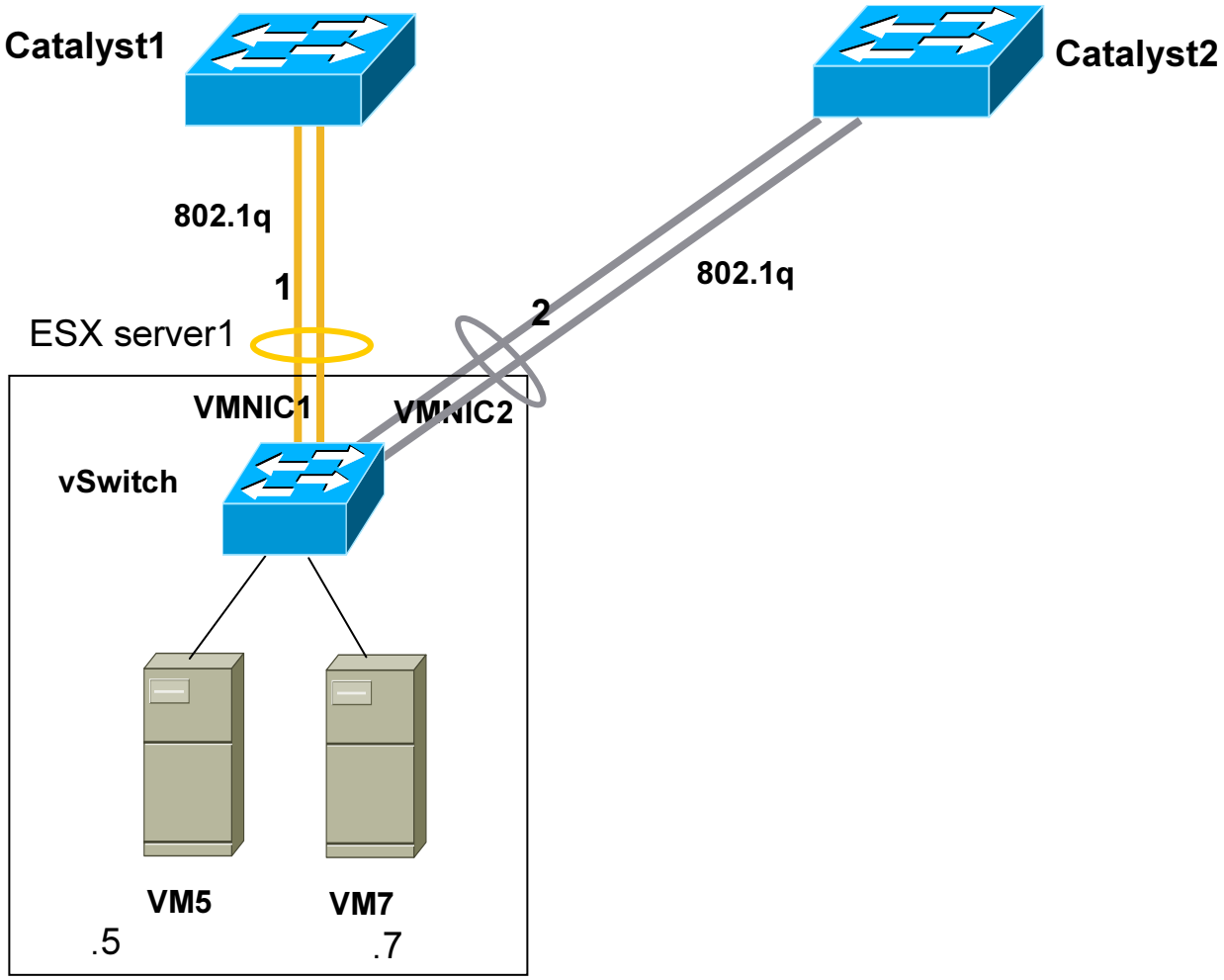
Can VM1 talk to Server3?



Can VM5 talk to VM4?

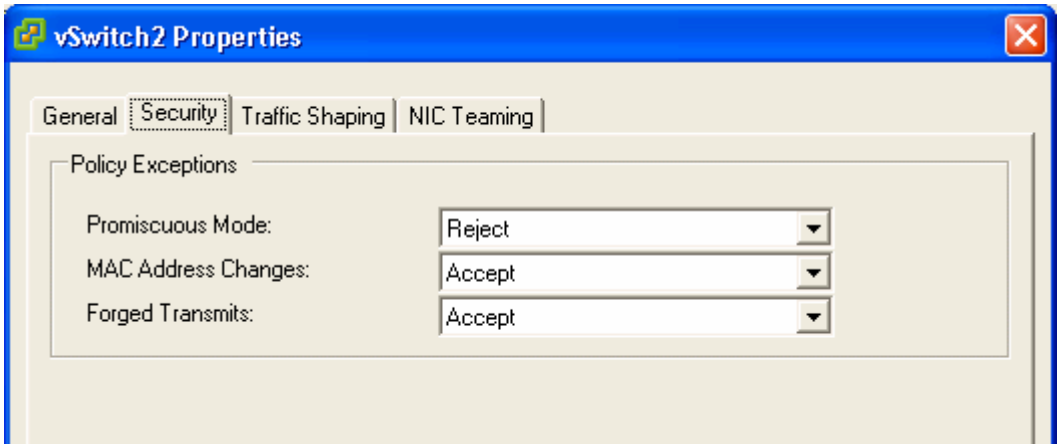


Is this design possible?



vSwitch Security

- Promiscuous mode Reject prevents a port from capturing traffic whose address is not the VM's address
- MAC Address Change, prevents the VM from modifying the vNIC address
- Forget Transmits prevents the VM from sending out traffic with a different MAC (e.g NLB)



vSwitch vs LAN Switch

- Similarly to a LAN Switch:
  - Forwarding based on MAC address
  - VM-to-VM traffic stays local
  - Vswitches TAG traffic with 802.1q VLAN ID
  - vSwitches are 802.1q Capable
  - vSwitches can create Etherchannels
  - **Preemption Configuration (similar to Flexlinks, but no delay preemption)**

- **Differently from a LAN Switch**
  - **No Learning**
  - **No Spanning-Tree protocol**
  - **No Dynamic trunk negotiation (DTP)**
  - **No 802.3ad LACP**
  - **Certain designs can isolate VMs**
  - **2 Etherchannel backing up each other is not possible**
  - **vSwitch doesn't have the equivalent of UPLINK TRACKING**
  - **No SPAN/mirroring capabilities: *Traffic capturing is by far not the equivalent of SPAN***
  - **Beaconing doesn't seem to add much value**
  - ***Port Security very limited***

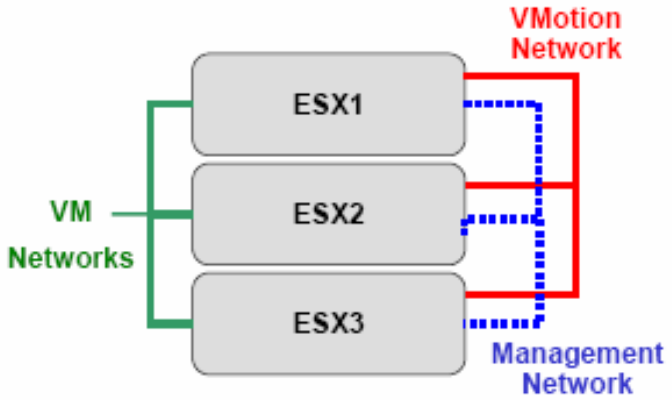
Agenda

- VMware Architecture and Components
- VMware LAN Networking
  - vSwitch Basics
  - NIC Teaming
  - vSwitch vs LAN Switch
  - Migration, HA, DRS
- Cisco/VMware DC DESIGNS
- Storage Implications of Server Virtualization

VMotion Migration Requirements

VMotion Requirements – Networking

- VMotion requires a Gigabit Ethernet network to ensure rapid migrations
  - A dedicated network is recommended to keep VM memory state secure
  - 2 hosts with cross-connected GigE cards can be used for demos
- VirtualCenter and ESX console management operations flow through the management network
  - Including VM deployment and cloning
- Network labels for each virtual NIC must be created through the ESX Server MUI:
  - Network labels are global across farms
  - VMotion automatically maps VMs to appropriate virtual NICs based on network labels
- VMs must have access to the necessary subnets on the target ESX Server



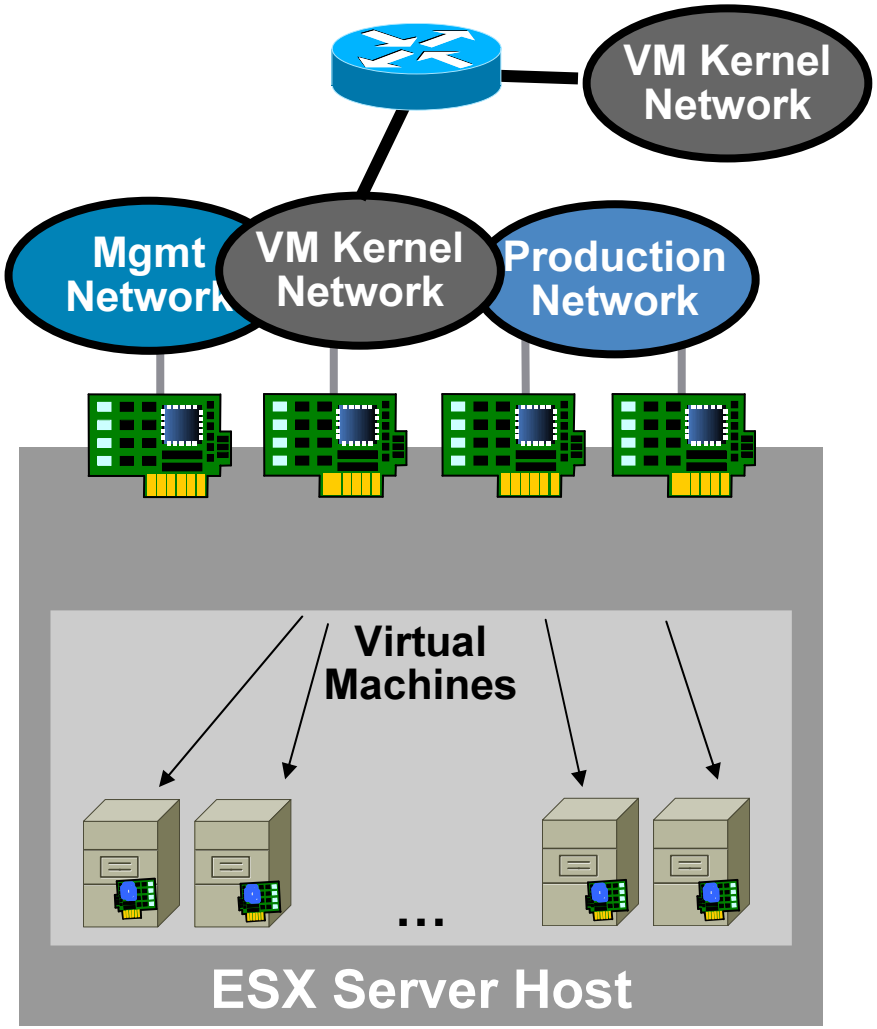
Minimum Network

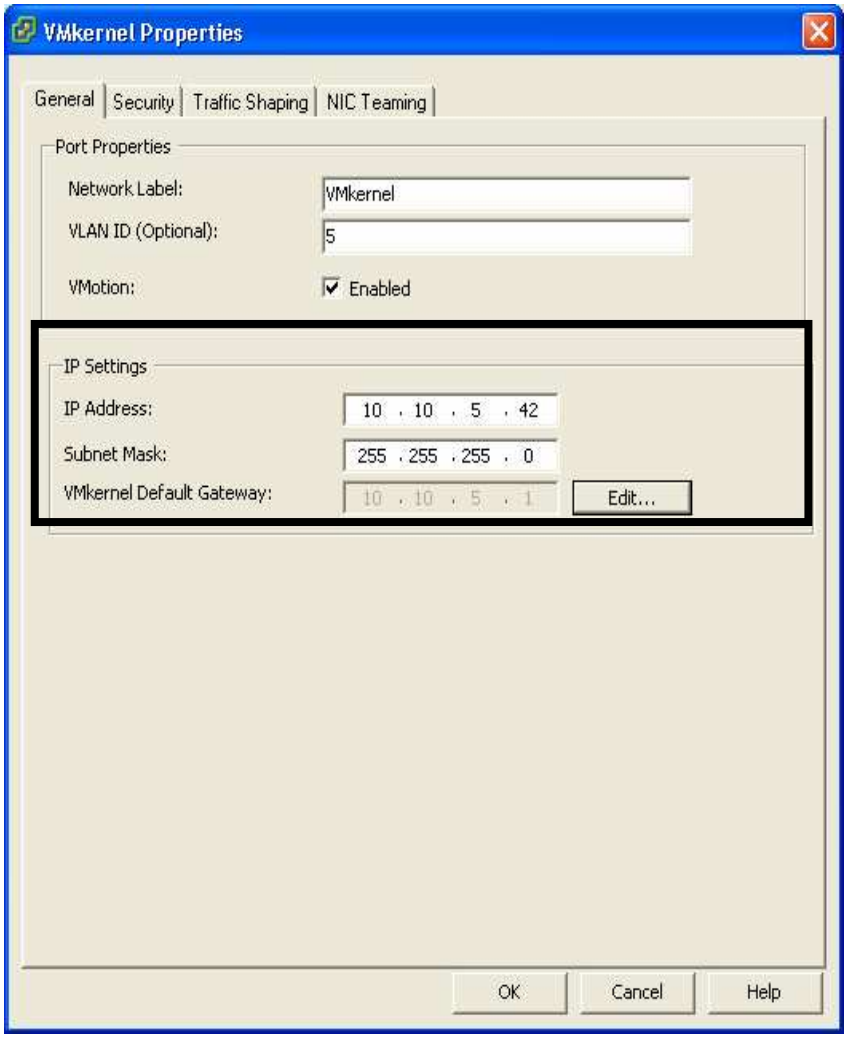
- 2 NICs with at least one GigE NIC for VMotion.
- For best security, dedicate the GigE NIC to VMotion and use VLANs to divide the VM & Mgmt traffic on the other NIC
- For best availability, combine both NICs into a bond, and use VLANS to divide traffic into at least 3 networks (1 or more for VMs, 1 for COS, and 1 for VMotion)

Best Practice Network

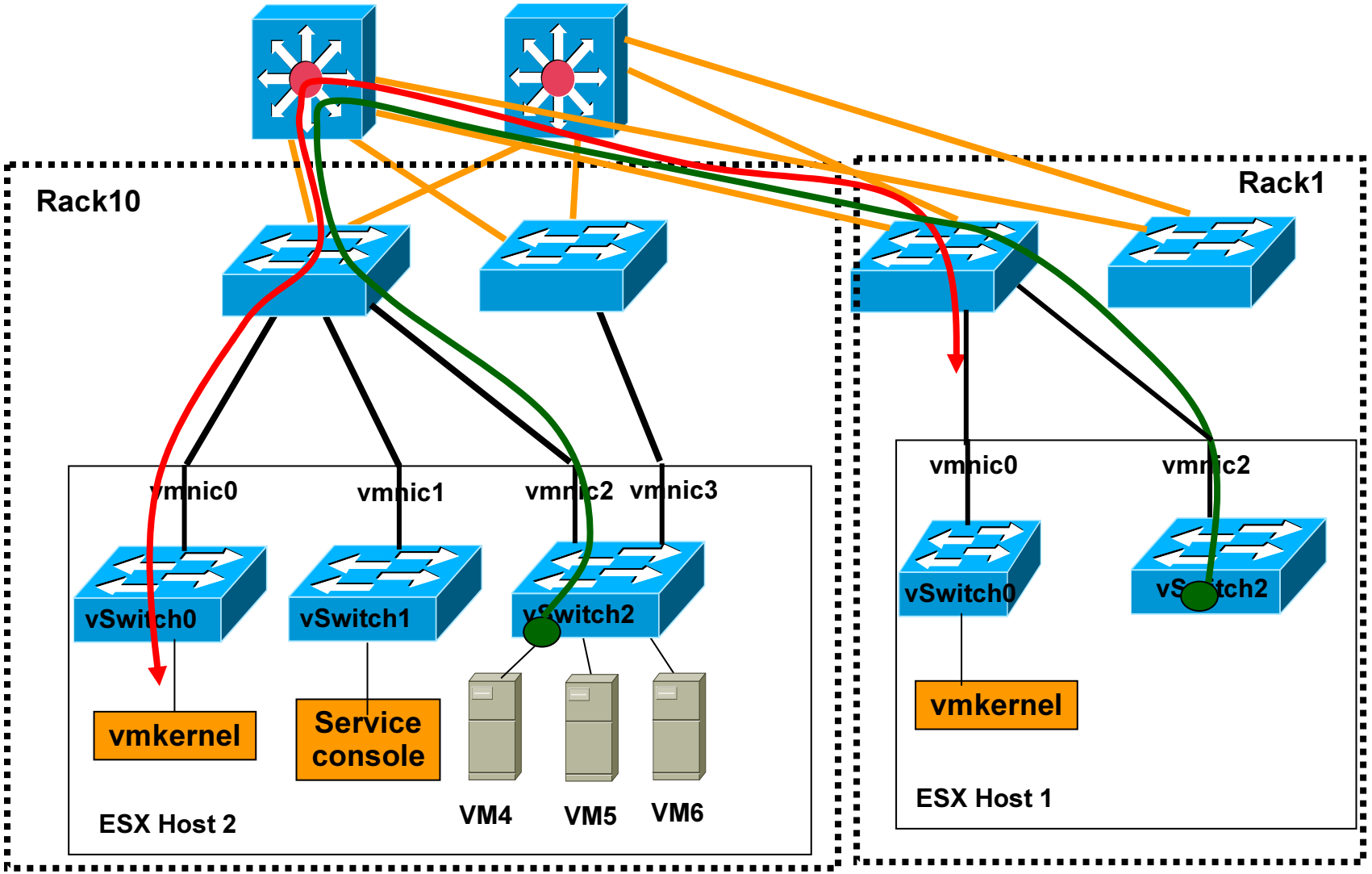
- **1 dedicated NIC for Console OS (10/100 or GigE)**
- **1 dedicated NIC for VMotion (GigE)**
- **1 or more NICs for VMs (10/100 or GigE)**

VMKernel Network can be routed



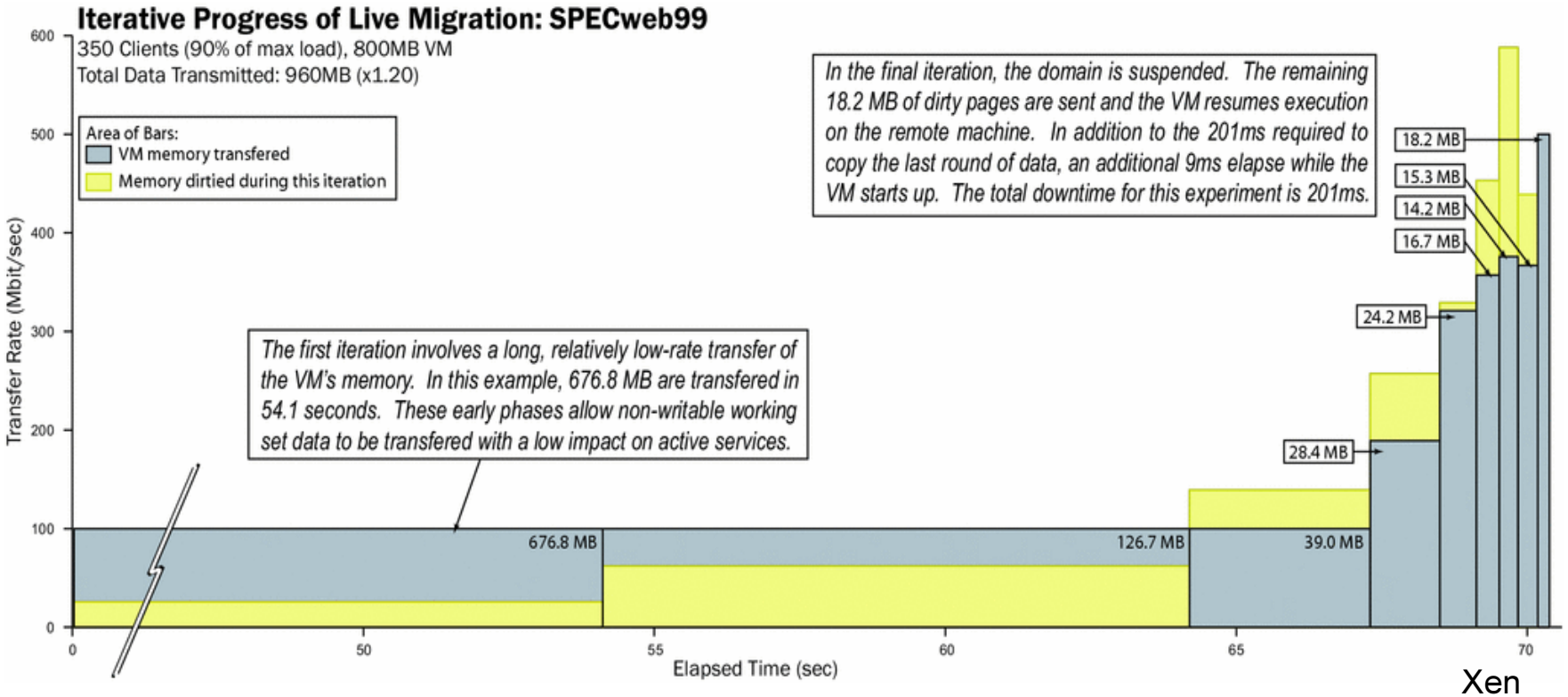


VMotion L2 Design



VM Migration—SpecWeb99 Migration

- SPECweb99 is the SPEC benchmark for evaluating the performance of WWW Servers
- The Standard Performance Evaluation Corporation (SPEC) is a non-profit corporation formed to establish, maintain and endorse a standardized set of relevant benchmarks that can be applied to the newest generation of high-performance computers



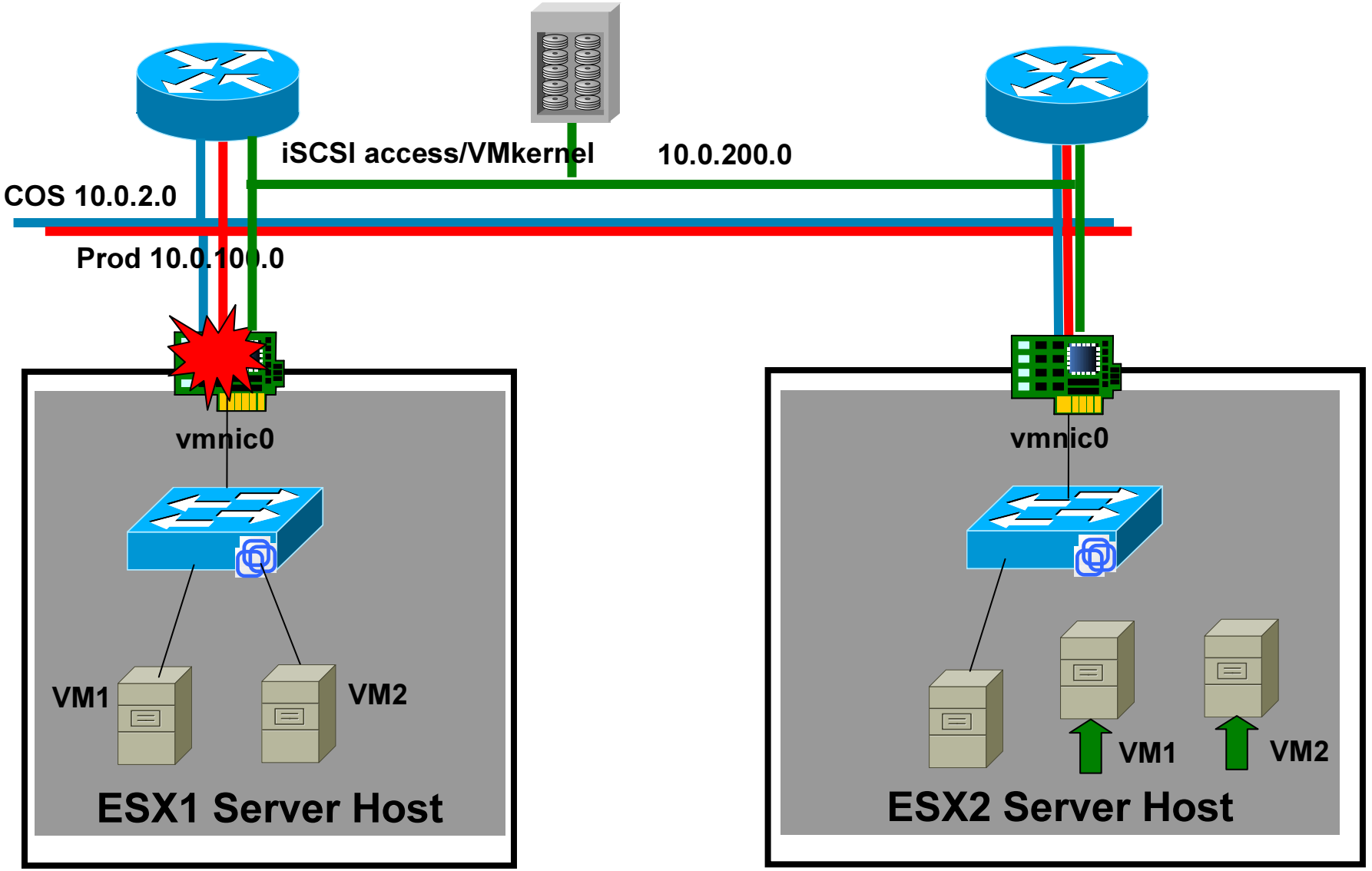
Xen

HA clustering (1)

- EMC/Legato AAM based
- HA Agent runs in every host
- Heartbeats Unicast UDP port ~8042 (4 UDP ports opened)
- Hearbeats run on the Service Console ONLY
- When a Failure Occurs, the ESX Host pings the gateway (on the SERVICE CONSOLE ONLY) to verify Network Connectivity
- If ESX Host is isolated, it shuts down the VMs thus releaseing locks on the SAN

- Recommendations:
  - Have 2 Service Console on redundant paths
  - Avoid losing SAN access (e.g. via iSCSI)
  - *Make sure you know before hand if DRS is activated too!*
- Caveats:
  - Losing Production VLAN connectivity only, ISOLATES VMs (there's no equivalent of uplink tracking on the vswitch)
- Solution:
  - NIC TEAMING

HA clustering (2)



Questions

- What is a Network Label?
- Do vSwitches always have vmnics? Yes/No
- What does the vNIC attach to? A vswitch? A VLAN?
- Can 2 Port-Groups be in the same VLAN on the same vSwitch?
- Does NIC teaming require NIC vendor driver installation in VM? Yes/No
- Does a VM MAC address change during a migration?

Agenda

- VMware Architecture and Components
- VMware LAN Networking
  - vSwitch Basics
  - NIC Teaming
  - vSwitch vs LAN Switch
  - Migration, HA, DRS
- Cisco/VMware DC DESIGNS
- Blade Server Designs
- Storage Implications of Server Virtualization

vSwitch and NIC Teaming Best Practices

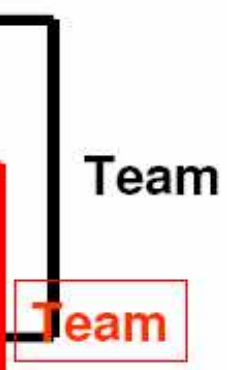
- Q: Should I use multiple vSwitches or multiple Port-Groups to isolate traffic?
- A: We didn't see any advantage in using multiple vSwitches, multiple Port-Groups with different VLANs give you enough flexibility to isolate servers
- Q: Should I use EST or VST?
- A: Always use VST, i.e. assign the VLAN from the vSwitch
- Q: Can I use native VLAN for VMs?
- A: Yes you can, but to make it simple don't. If you do, do not TAG VMs with the native VLAN

- Q: Which NIC Teaming configuration should I use?
- A: Active/Active, Virtual Port-ID based
- Q: Do I have to attach all NICs in the team to the same switch or to different switches?
- A: with Active/Active Virtual Port-ID based, it doesn't matter
- Q: Should I use Beaconing?
- A: No
- Q: Should I use Rolling Failover (i.e. no preemption)
- A: No, default is good, just enable *trunkfast* on the Cisco switch

NIC Team Across Hardware

Common 4 Port Design Example:

- On-board NIC 1 (vSwitch1 uplink)
  - SC (preferred NIC in team) + VMotion
- On-board NIC 2 (vSwitch2 uplink)
  - Default virtual machine port group
- PCI based NIC 1 (vSwitch1 uplink)
  - SC + VMotion (preferred NIC in team)
- PCI based NIC 2 (vSwitch2 uplink)
  - Default virtual machine port group



Design Goal:

- Prevent single points of failure (PCI bus/motherboard ASIC)

Cisco Switchport Configuration

- Make it a Trunk
- Enable Trunkfast
- Can the Native VLAN be used for VMs?
- Yes, but IF you do, you have 2 options
  - Configure VLAN ID = 0 for the VMs that are going to use the native VLAN (preferred)
  - Configure "vlan dot1q tag native" on the 6k (not recommended)
- Do not enable Port Security (see next slide)
- Make sure that "teamed" NICs are in the same Layer 2 domain
- Provide a Redundant Layer 2 path

- interface GigabitEthernetX/X
- description <<** VM Port **>>
- no ip address
- switchport
- switchport trunk encapsulation dot1q
- *switchport trunk native vlan <id>*
- switchport trunk allowed vlan xx,yy-zz
- switchport mode trunk
- *switchport nonegotiate*
- no cdp enable
- *spanning-tree portfast trunk*
- !

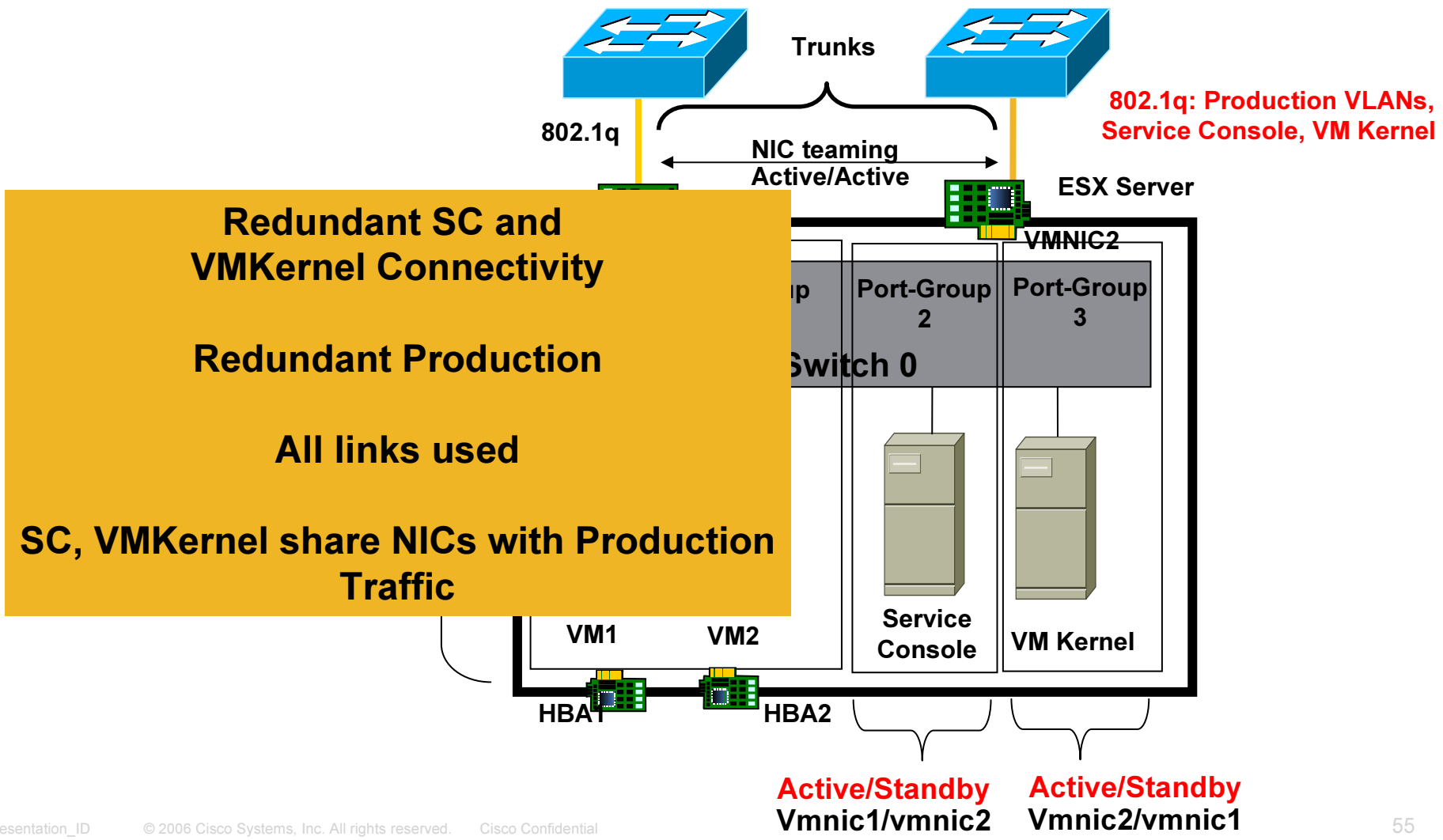
Typically: SC, VMKernel, VM Production

Port Security and VMware *Incompatible*

- **http://www.cisco.com/en/US/partner/products/hw/switches/ps5023/products_configuration_guide_chapter09186a00808b0210.html#wp1170581**
- **3750-STACK-top-R1(config-if)#switchport port-security maximum <number> vlan <vlan_number>**
- **How many MACs do you have to count?**
  - **SC, SC iSCSI, VMKernel, VMotion = 4**
  - **+ 1 MAC per VM**
  - **+ BIA MAC**
- **maximum 5, violation restrict**

- If a MAC moves (i.e. Vmotion migration or NIC Teaming)
- 3750-STACK-top-R1#
- 9w0d: %PORT_SECURITY-2-PSECURE_VIOLATION: Security violation occurred, caused
- Port goes down or traffic is dropped

Configuration with 2 NIC
SC, VMKernel, Production share NICs



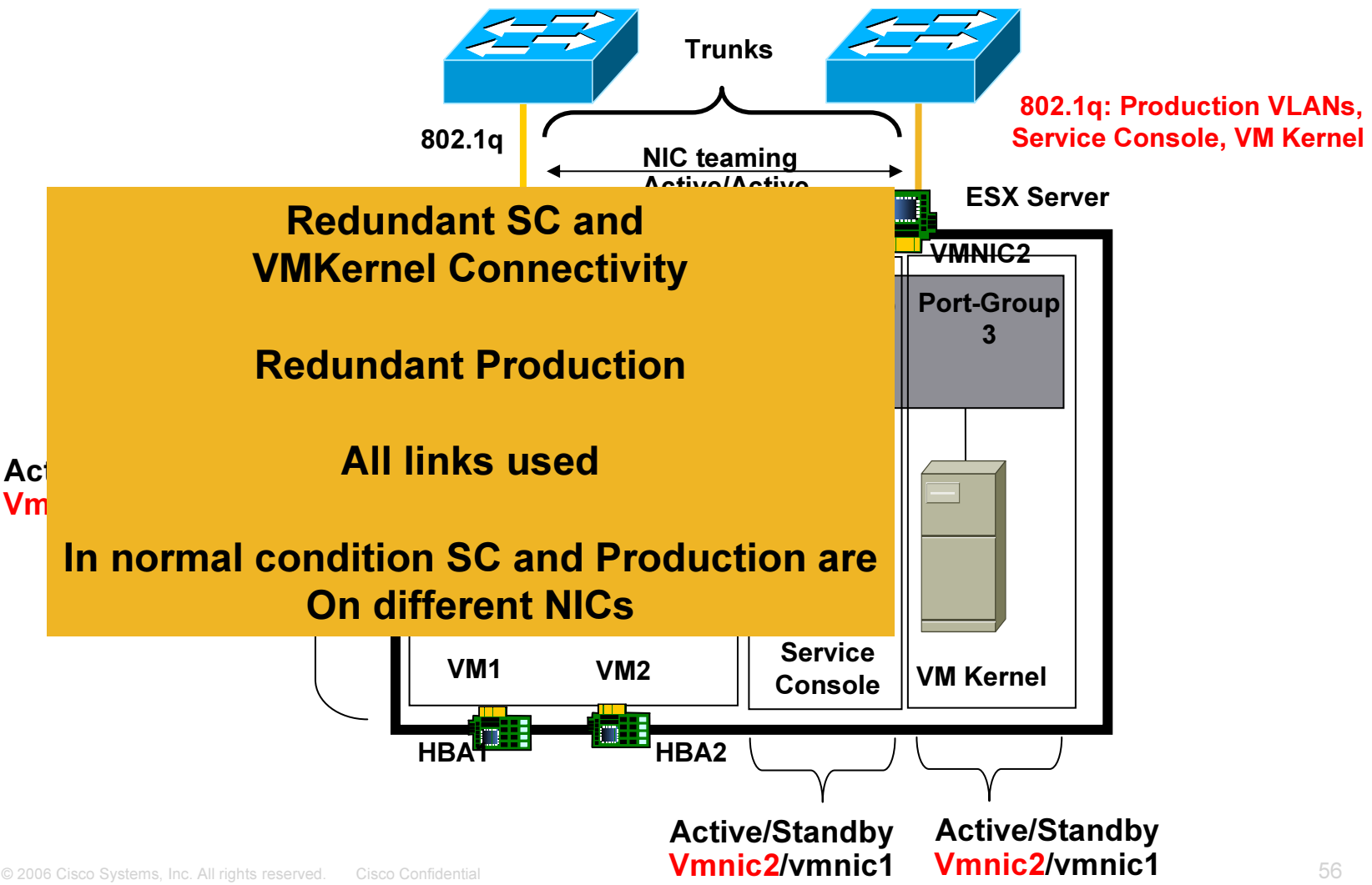
Redundant SC and VMKernel Connectivity

Redundant Production

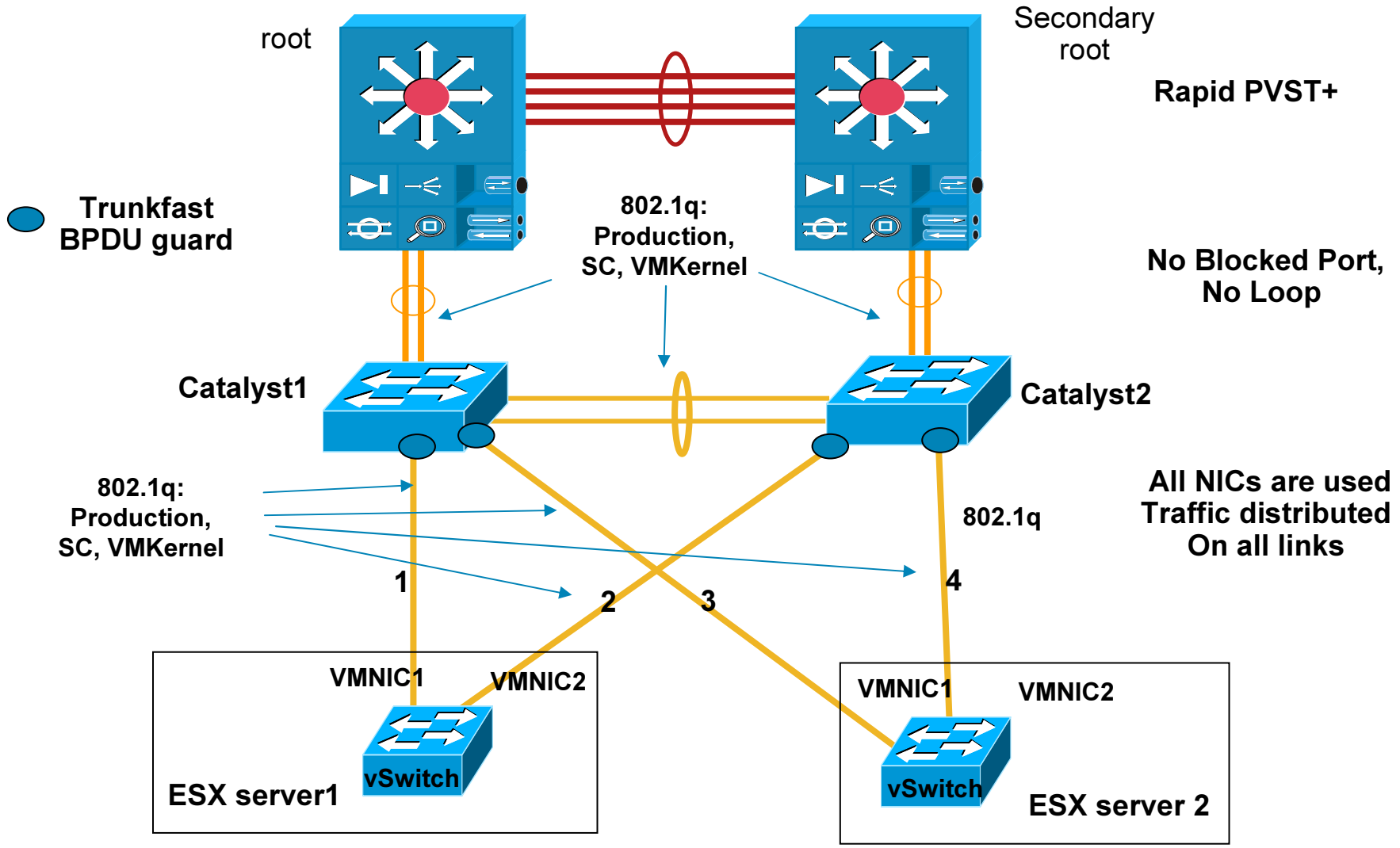
All links used

SC, VMKernel share NICs with Production Traffic

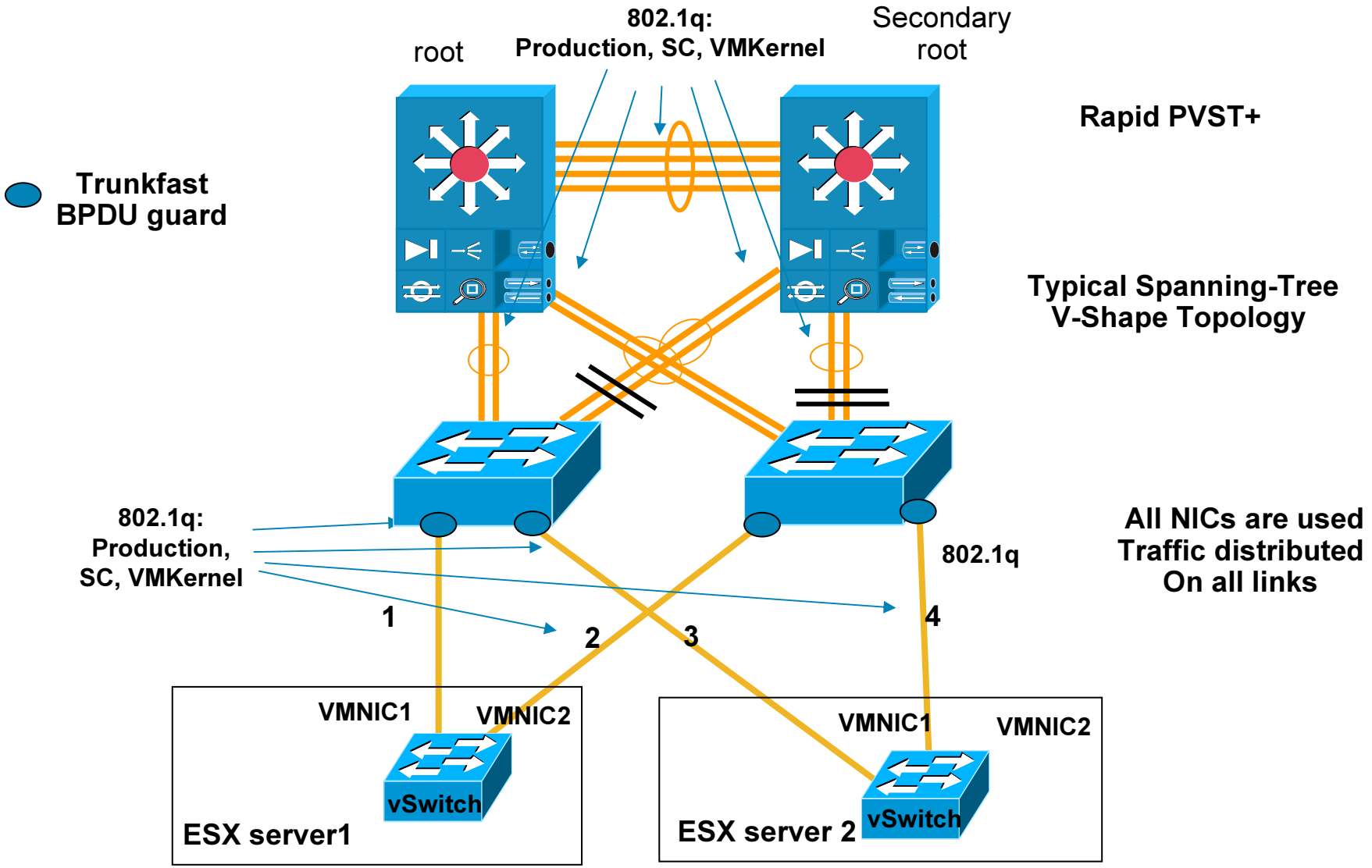
Configuration with 2 NICs
Dedicated NIC to SC, VMKernel, Separate NIC for Production



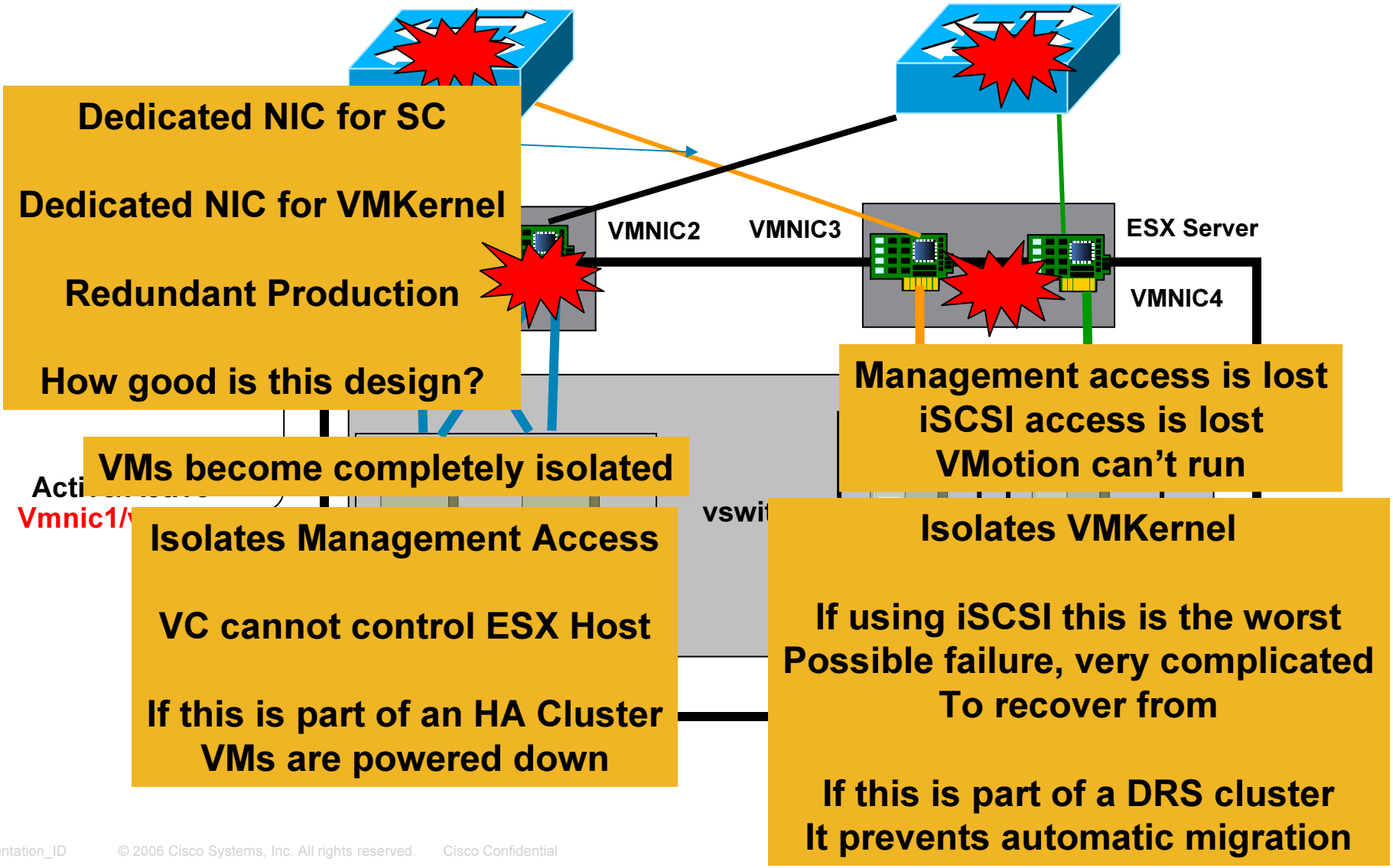
Network Attachment (1)



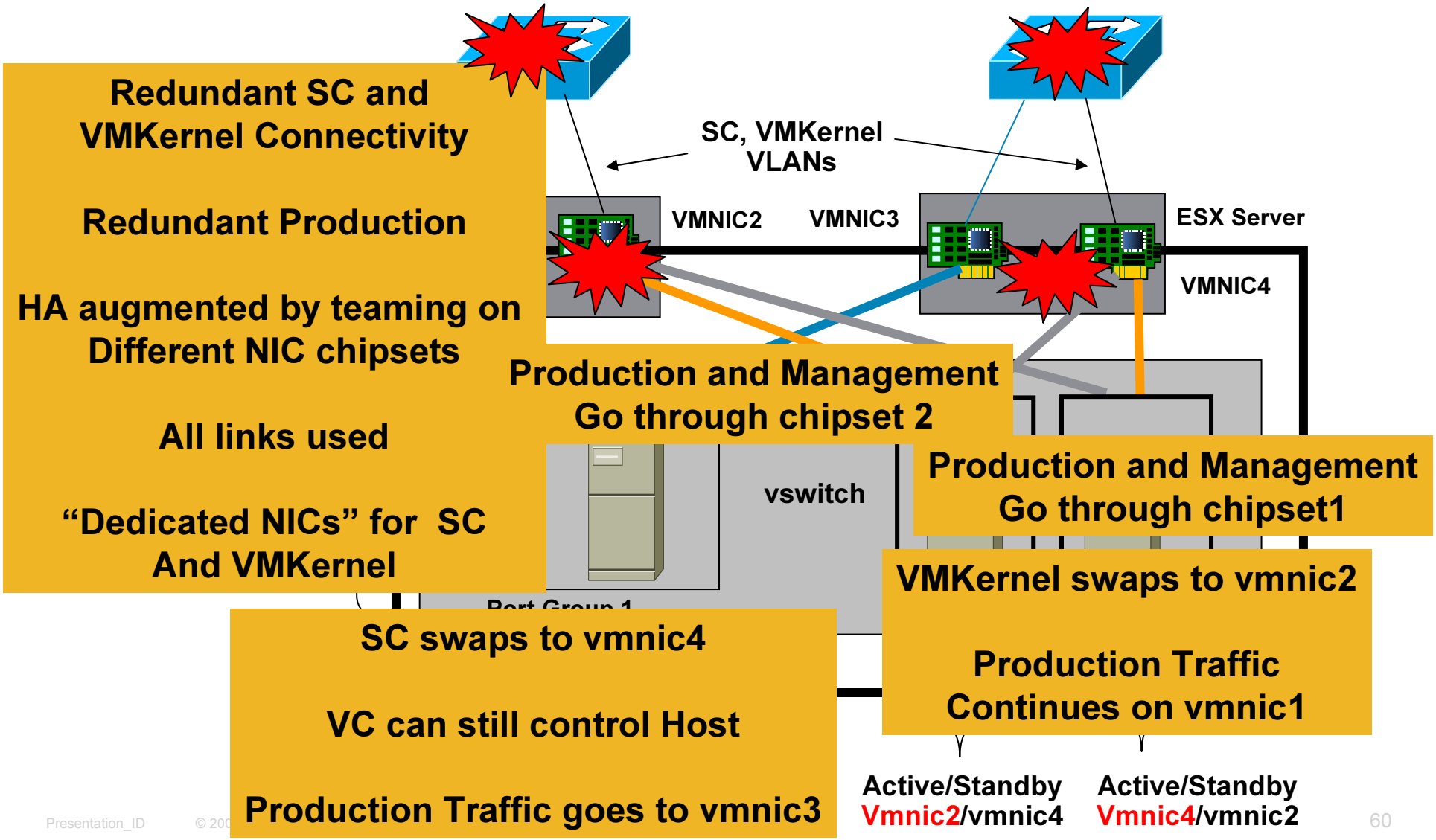
Network Attachment (2)



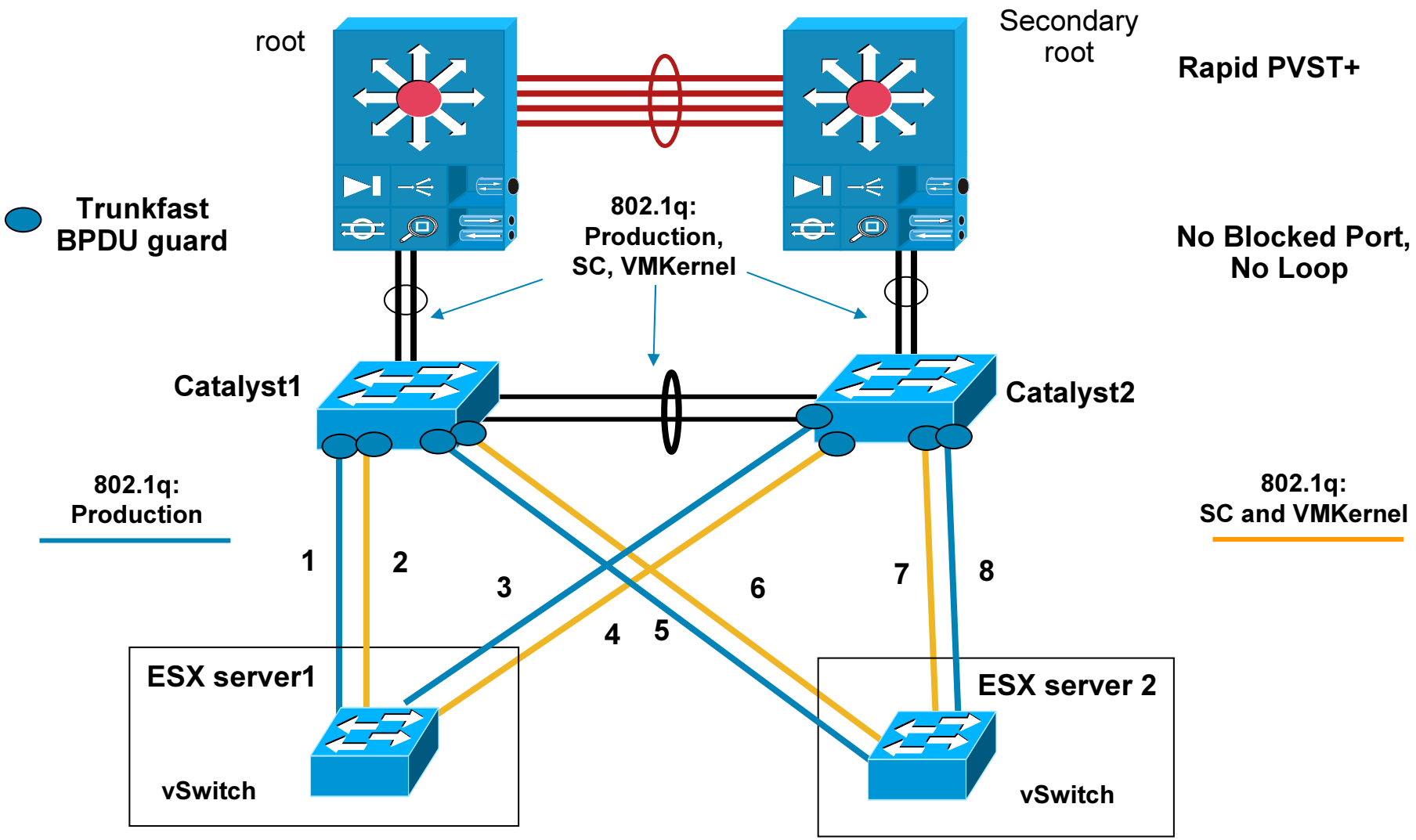
Configuration with 4 NICs
Dedicated NICs for SC and VMKernel



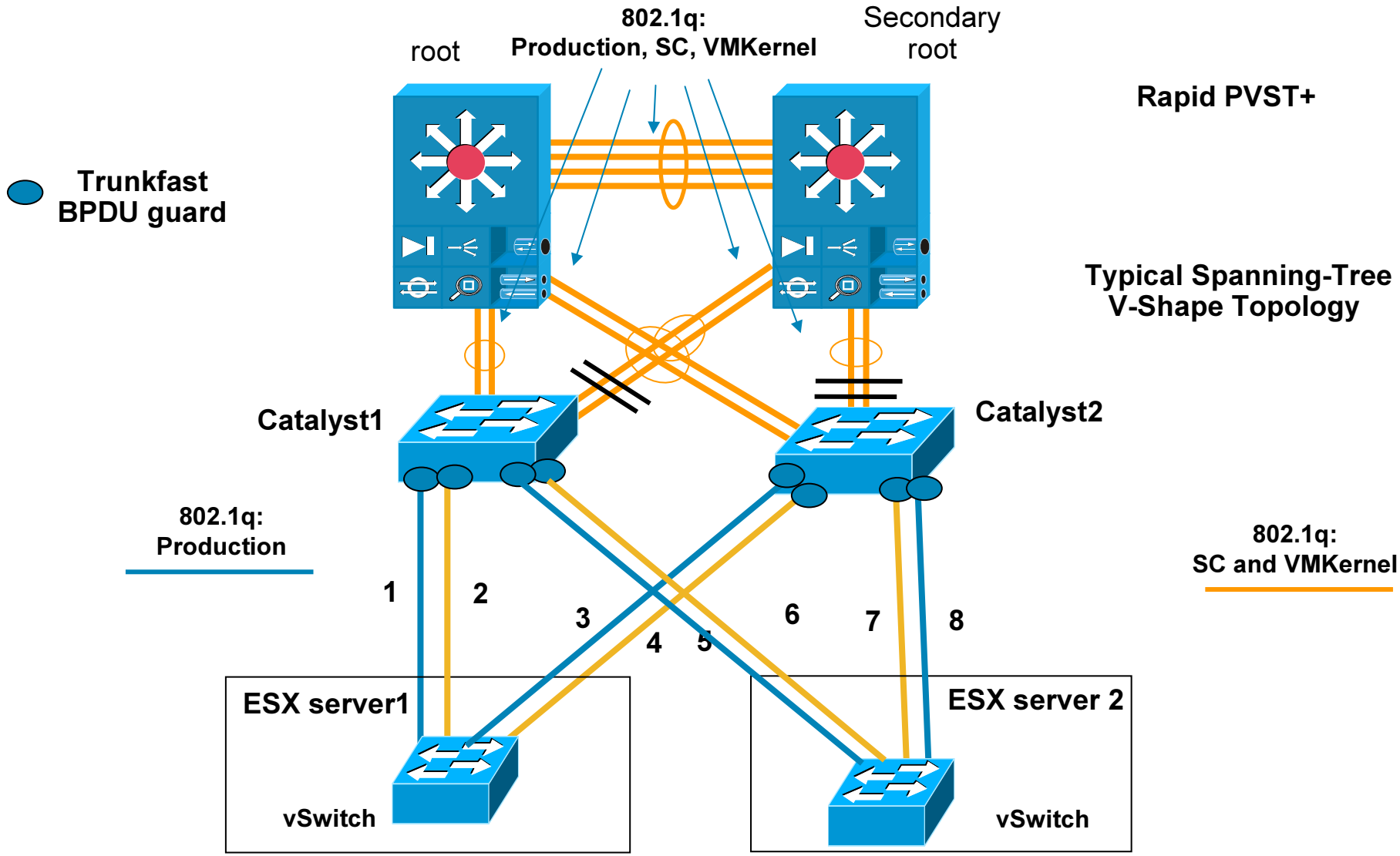
Configuration with 4 NICs



Network Attachment (1)



Network Attachment (2)



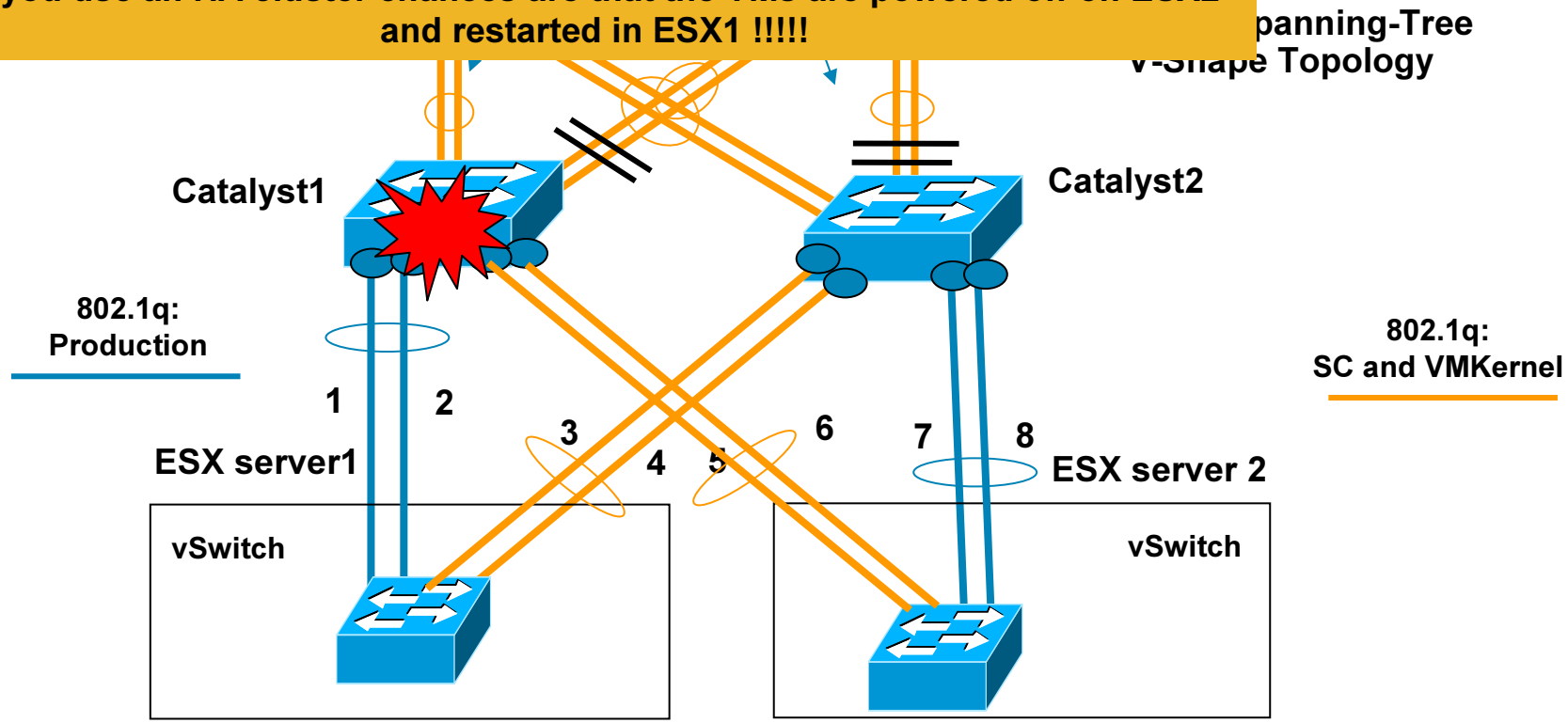
How About?

Production on ESX 1 is Completely isolated

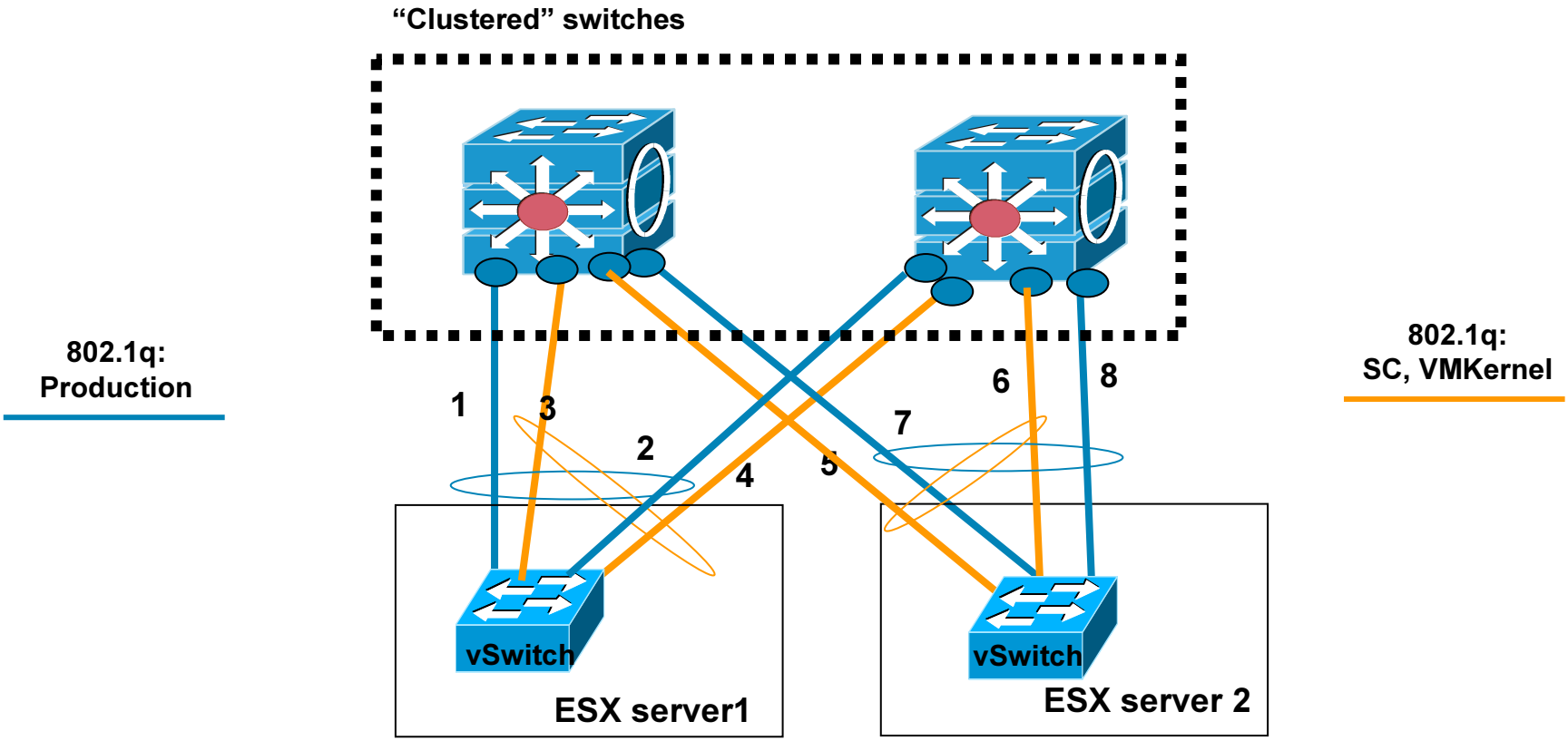
HA doesn't do anything for ESX1, VMs are isolated

Management and VMKernel are isolated. On ESX2, if you use iSCSI, this is not easy to recover from

If you use an HA cluster chances are that the VMs are powered off on ESX2 and restarted in ESX1 !!!!!



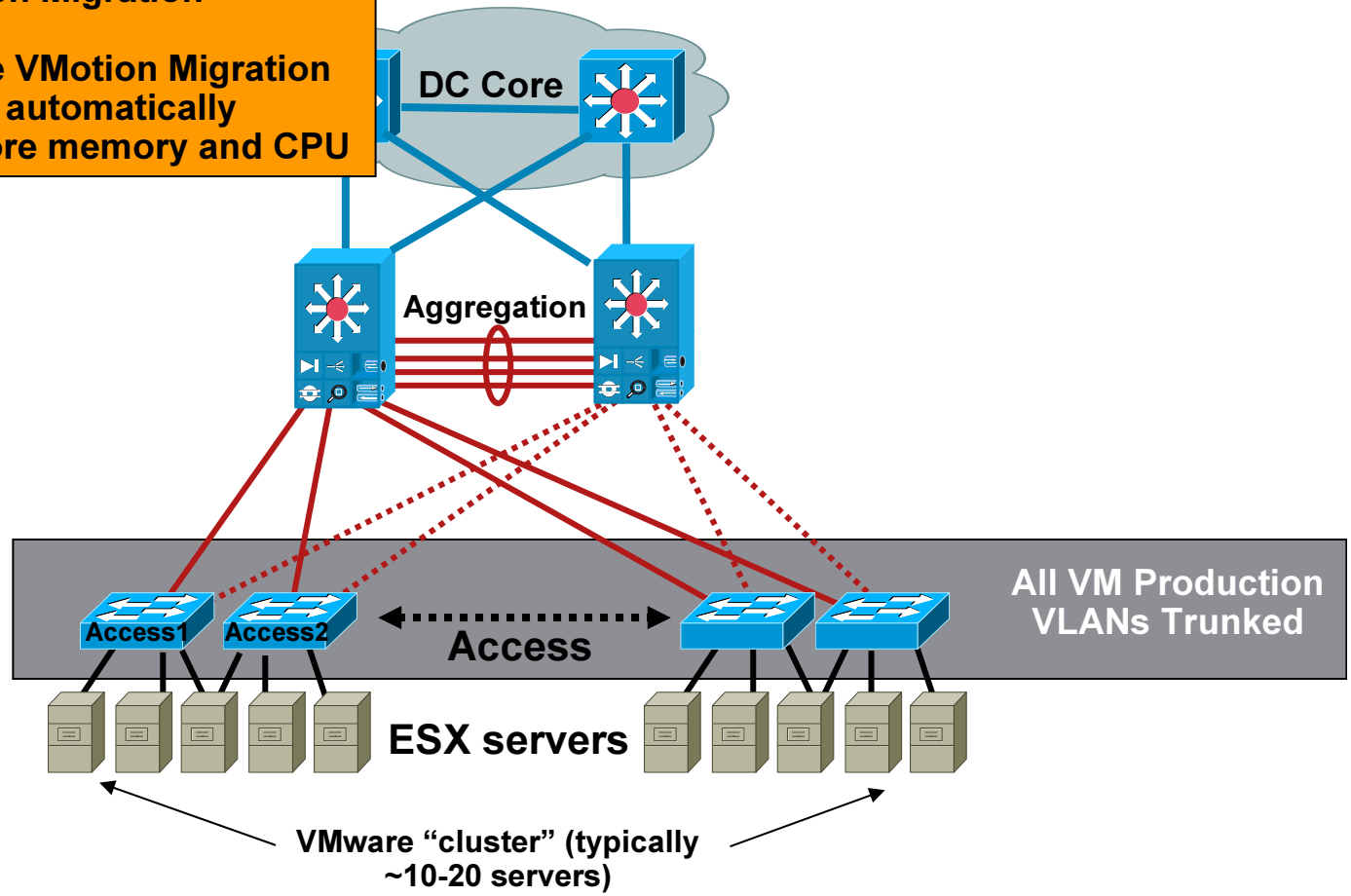
4 NICs with Etherchannel



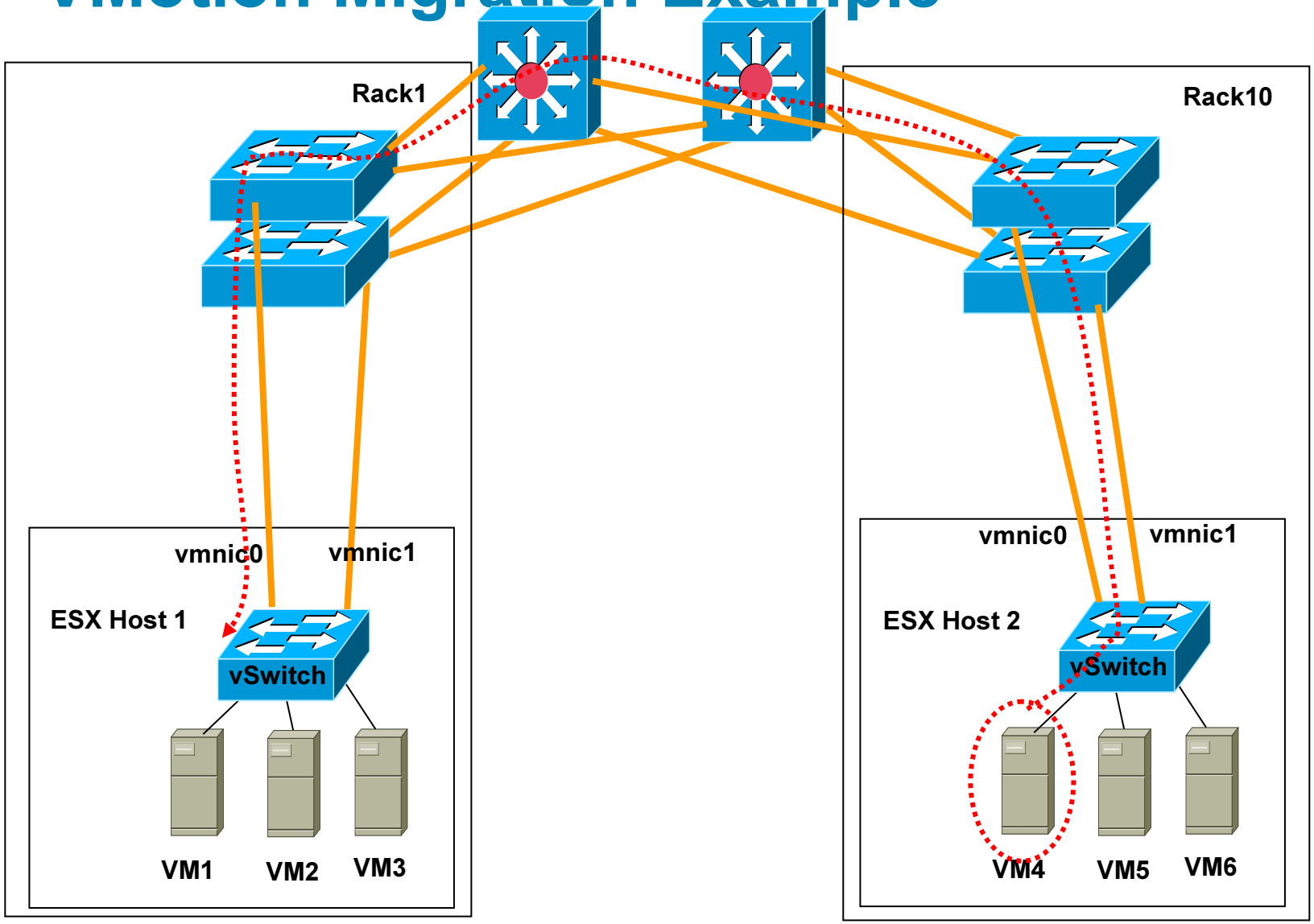
Typical ESX HA/DRS cluster design

Maintenance Mode **in an HA cluster leverages VMotion Migration**

DRS cluster **may require VMotion Migration If you want VMs to automatically Move to the Host with more memory and CPU**

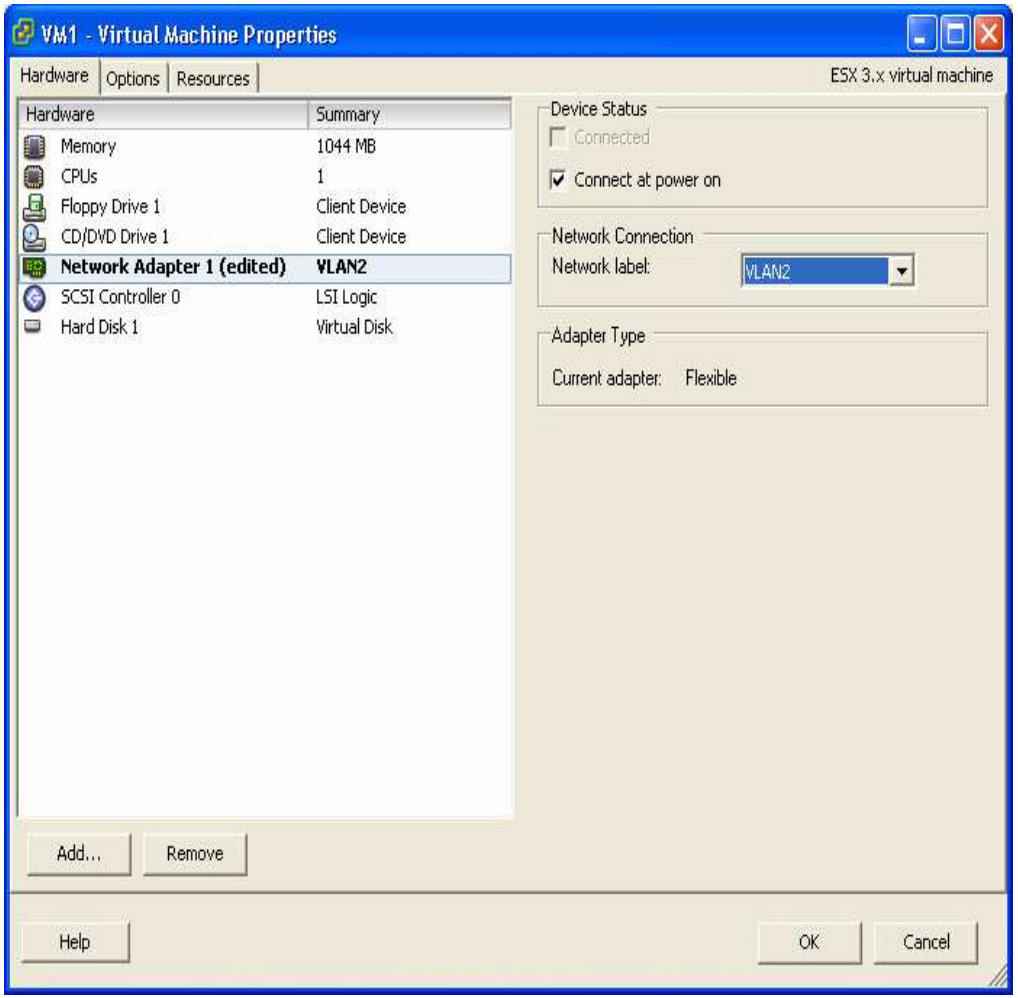


VMotion Migration Example



Network Label and VMotion

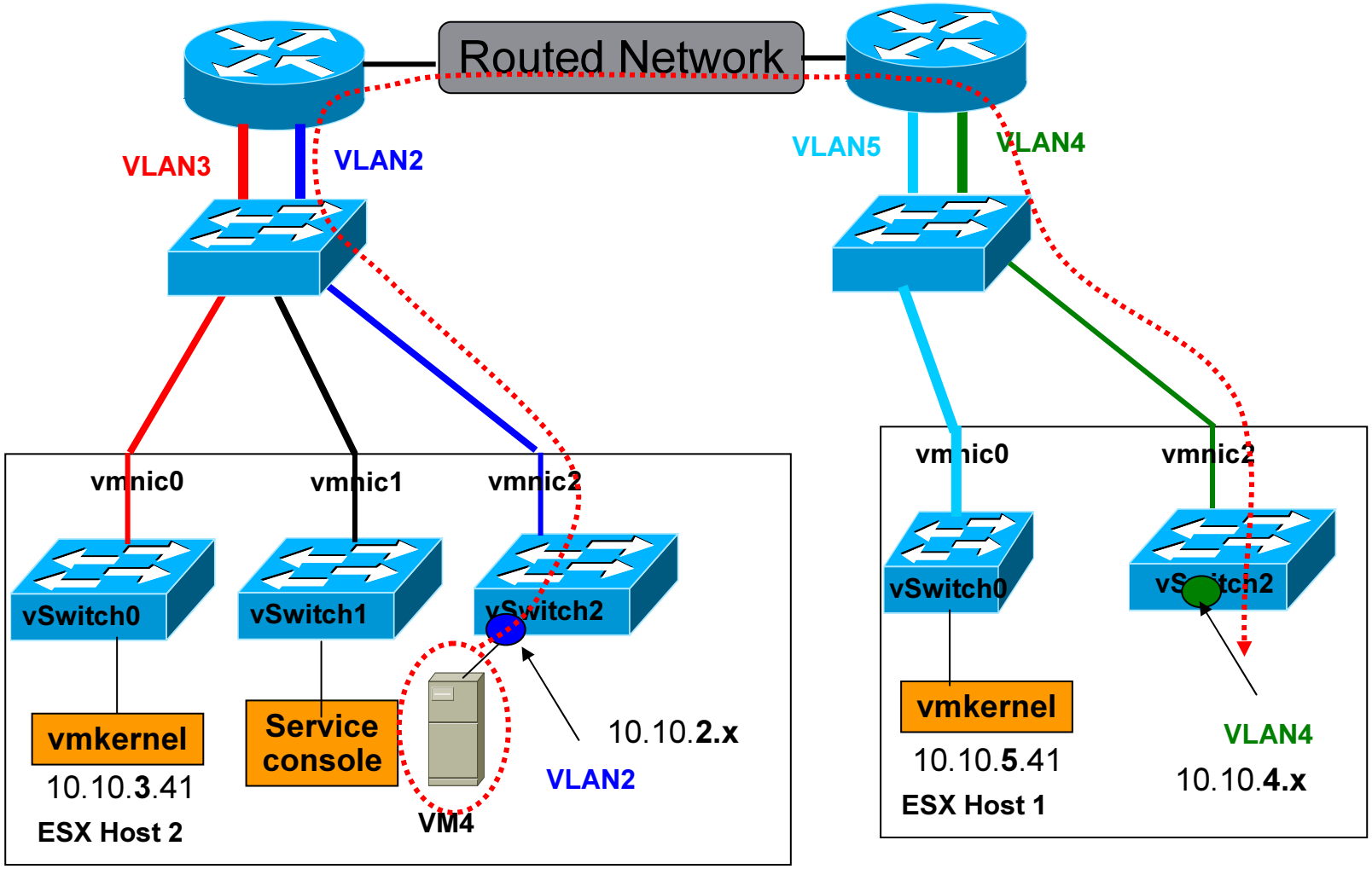
- VMs moving from one ESX server to a different one look for the same *Network Label*



VMotion Best Practices

- **Datapoints:**
  - **Migration only happens within a *VM HA/DRS cluster* and/or within a *datacenter***
  - **VMotion looks for the Network Label to be available on the Target ESX Host**
  - **VM MAC doesn't change during the migration**
- **Best Practice**
  - **Make the VMkernel network routed, extend the Layer 2 domain only for the VM production traffic**
  - **At most the Layer 2 domain needs to encompass ~10-20 machines, set the Layer 2 boundary within the Data center accordingly**
  - **Enable the option "Notify Switch" in the vswitch so that target vswitch sends out a RARP to update the mac-forwarding tables**

Experimental Routed Network for Powered-off Migration or HA Cluster

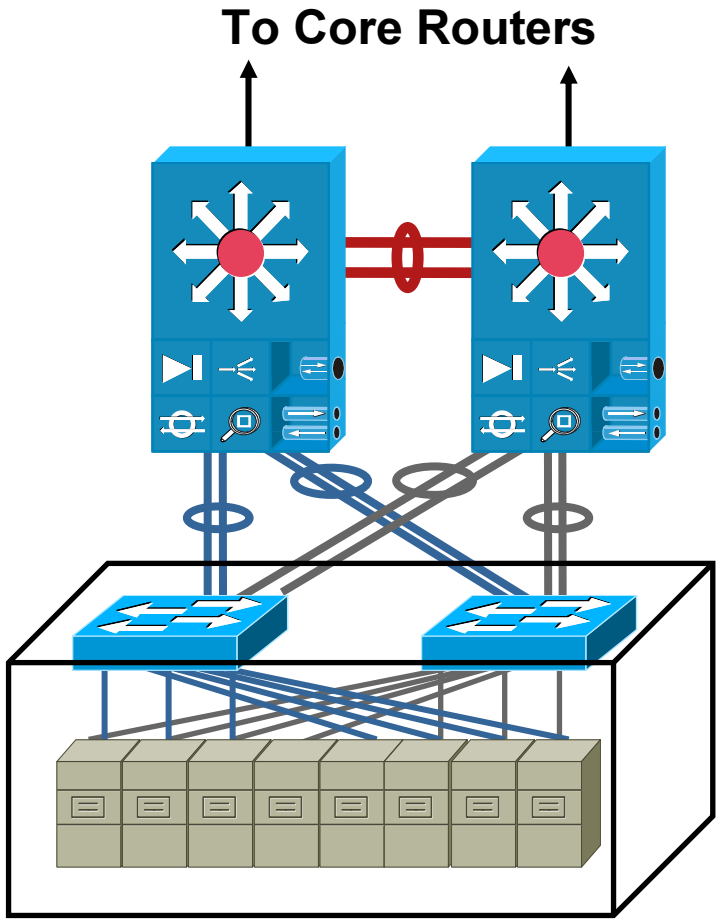


Agenda

- VMware Architecture and Components
- VMware LAN Networking
  - vSwitch Basics
  - NIC Teaming
  - vSwitch vs LAN Switch
  - Migration, HA, DRS
- Cisco/VMware DC DESIGNS
- Blade Server Designs
- Storage Implications of Server Virtualization

Handy Features for Large Scale Deployments
Flexlinks

- Flexlinks keeps one set of ports in forwarding state and a backup set of ports are non forwarding for the same set of VLANs
- No Spanning-Tree is involved so it's very light weight on the Control Plane
- You can have half VLANs active on one set of links and half VLANs active on the other set of links
- Preemption configurable
- Failover <100ms
- Preemption Delay Configurable
  - off
  - forced
  - bandwidth

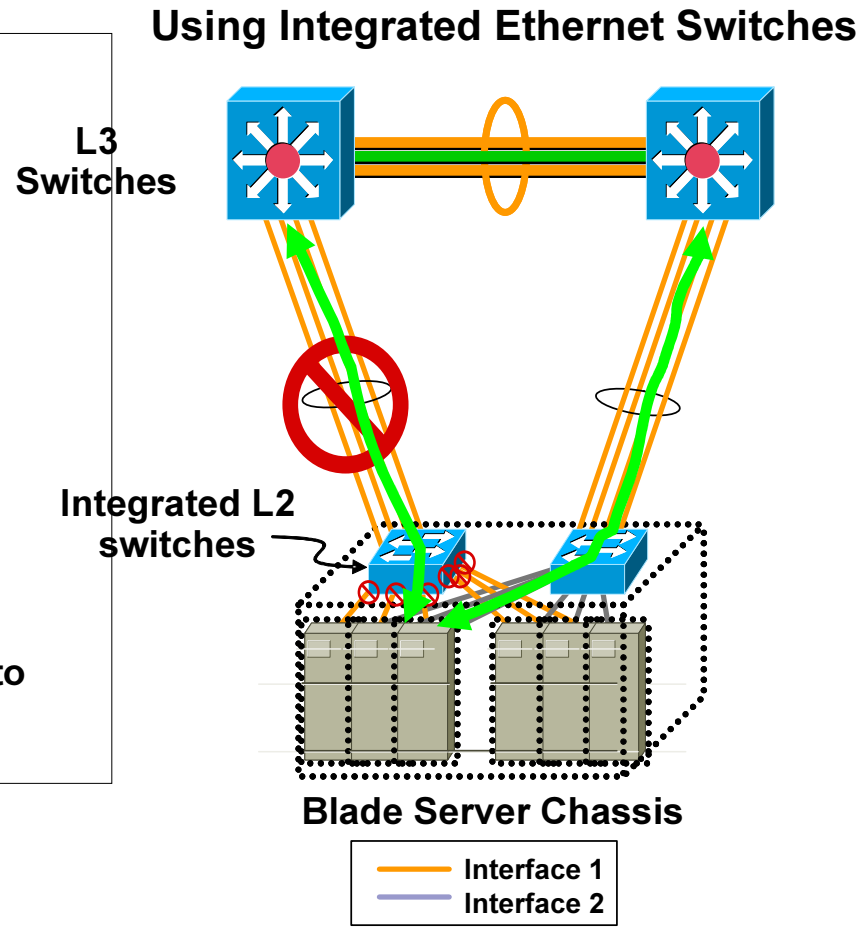


Design with the Integrated Switch
Uplink Tracking / Trunk Resiliency

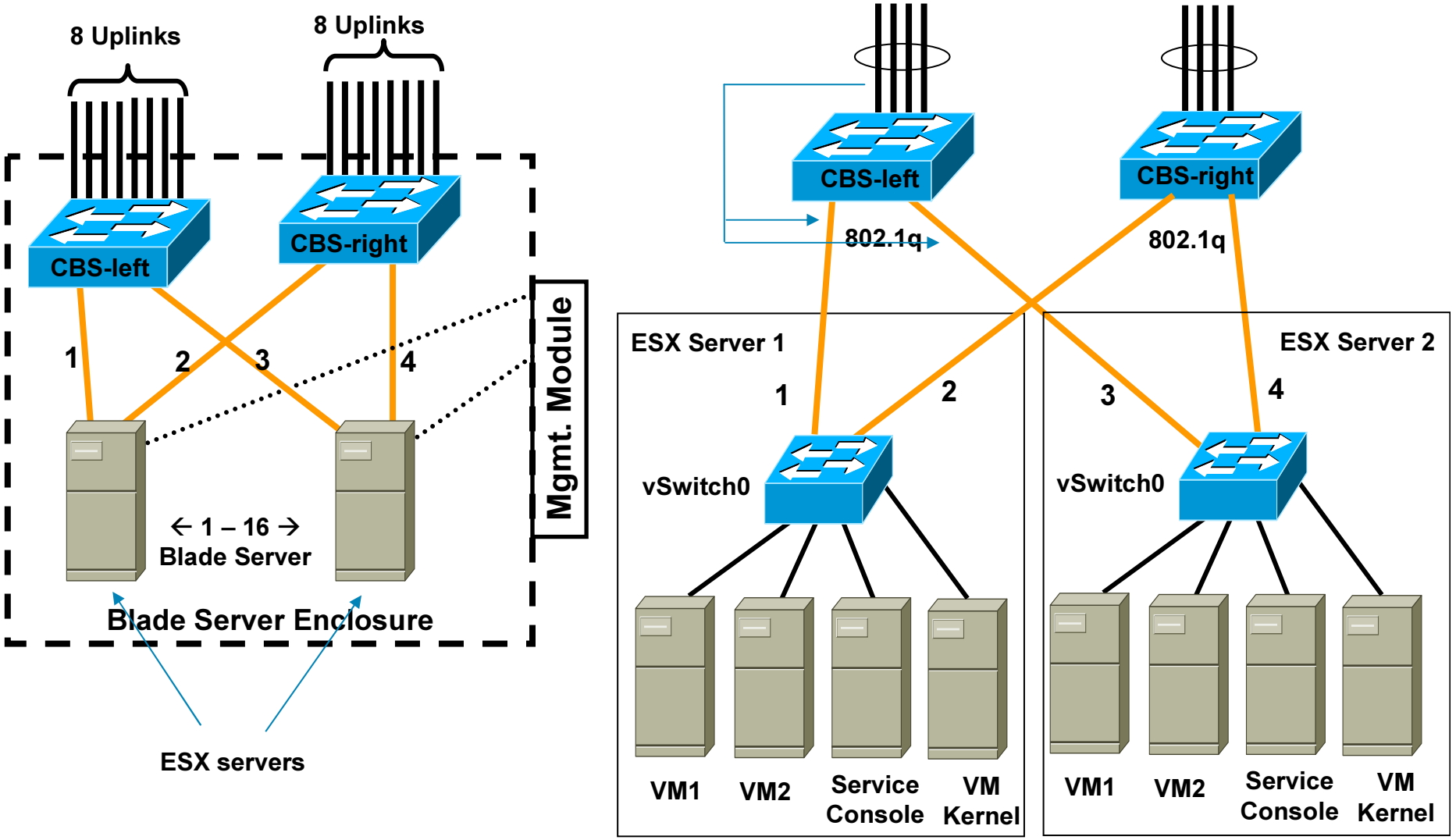
- switch(config)# link state track 1
- switch(config)# int range PO1, PO2
- switch(config-if-range)# link state group 1 upstream
- switch(config-if-range)# int range gig0/1 - 16
- switch(config-if-range)# link state group 1 downstream
- switch(config-if-range)# end

- Note:
- PO1 is composed of gig ports 21 and 22
- PO2 is composed of gig ports 23 and 24

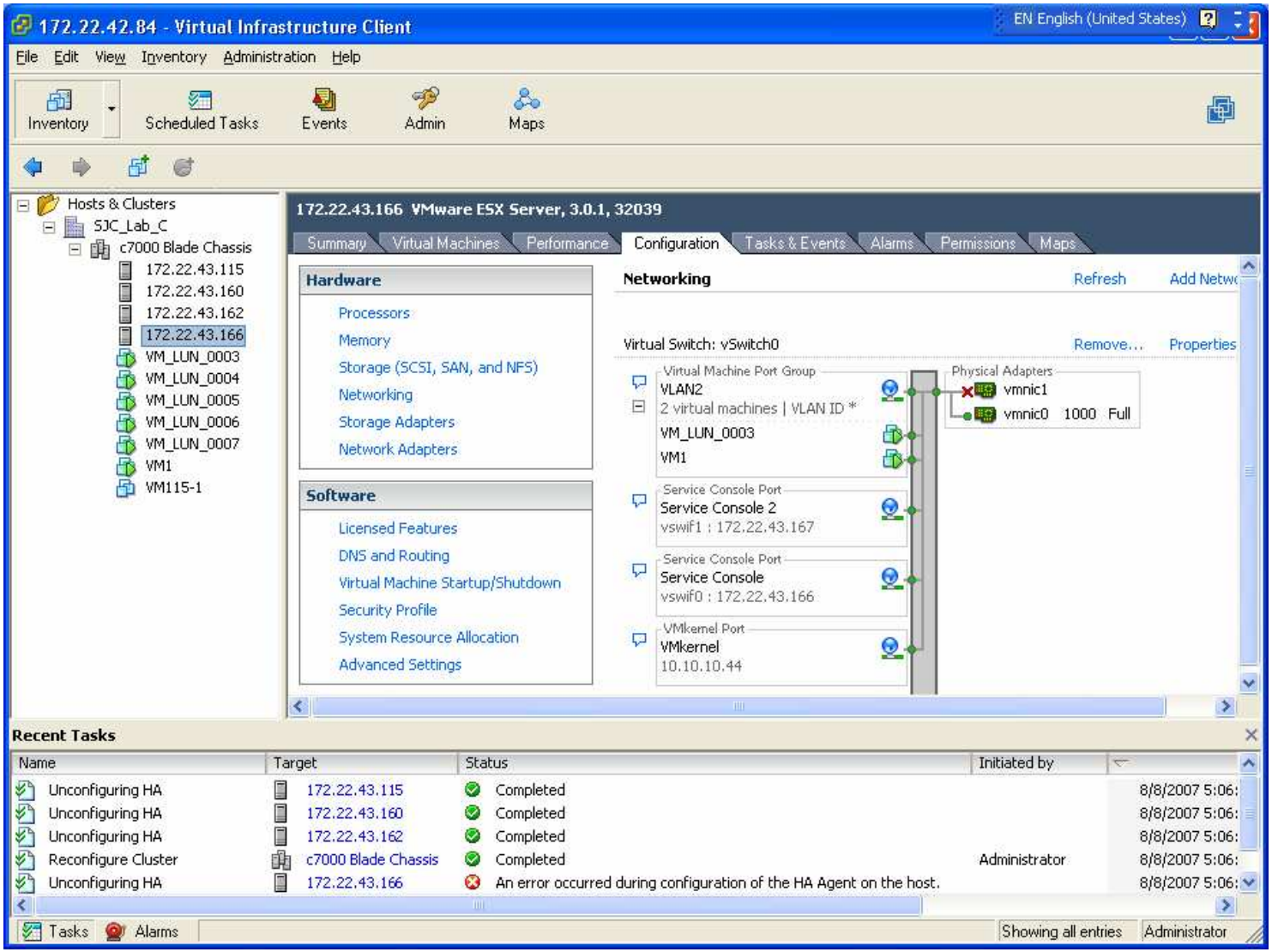
- These Etherchannels must be created separately prior to creating the Layer 2 Trunk Failover Feature.



HP Blade Server + *VM*



Tracking on VM Network



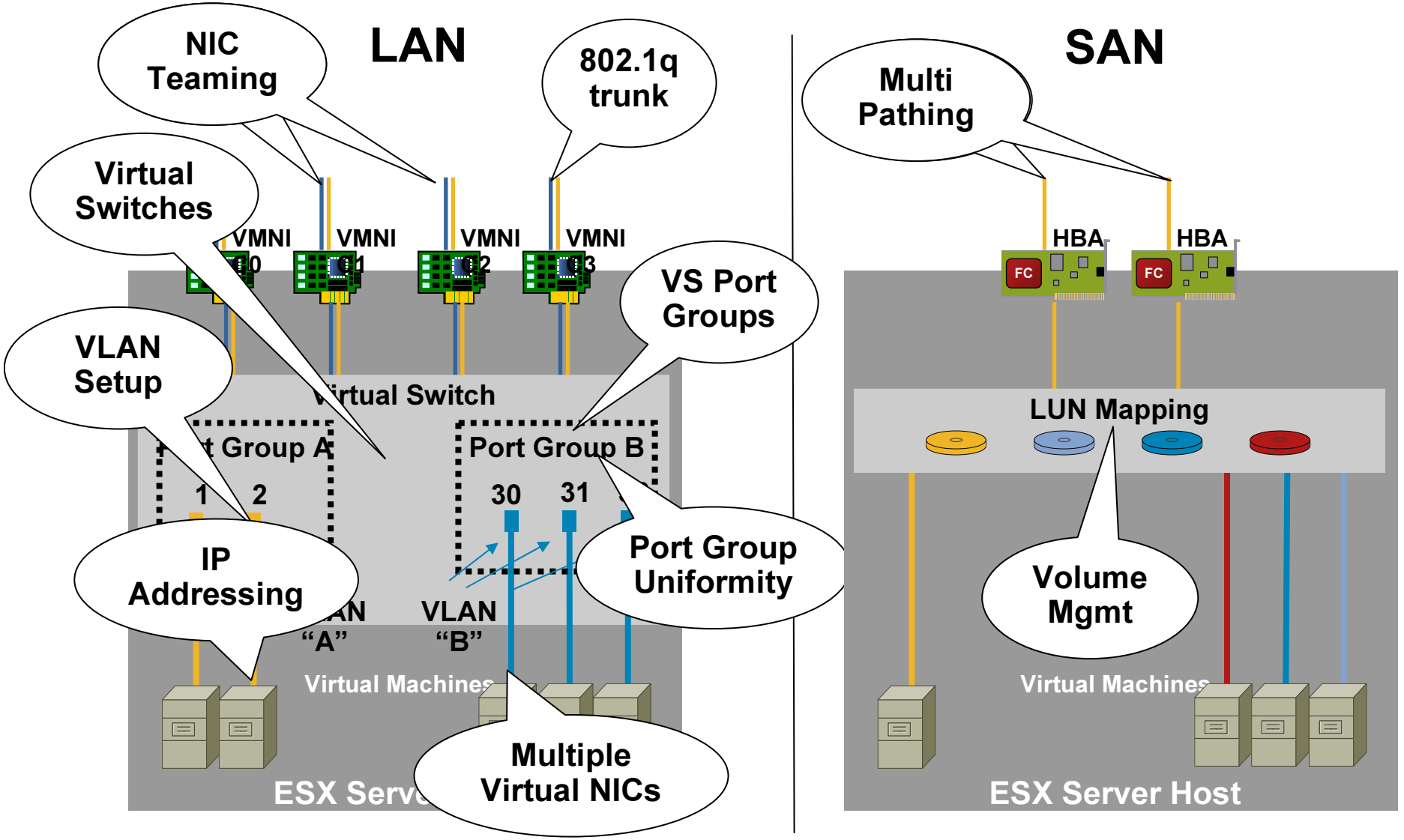
Agenda

- VMware Architecture and Components
- VMware LAN Networking
- Cisco/VMware DC DESIGNS
- Blade Server Designs
- Storage Implications of Server Virtualization

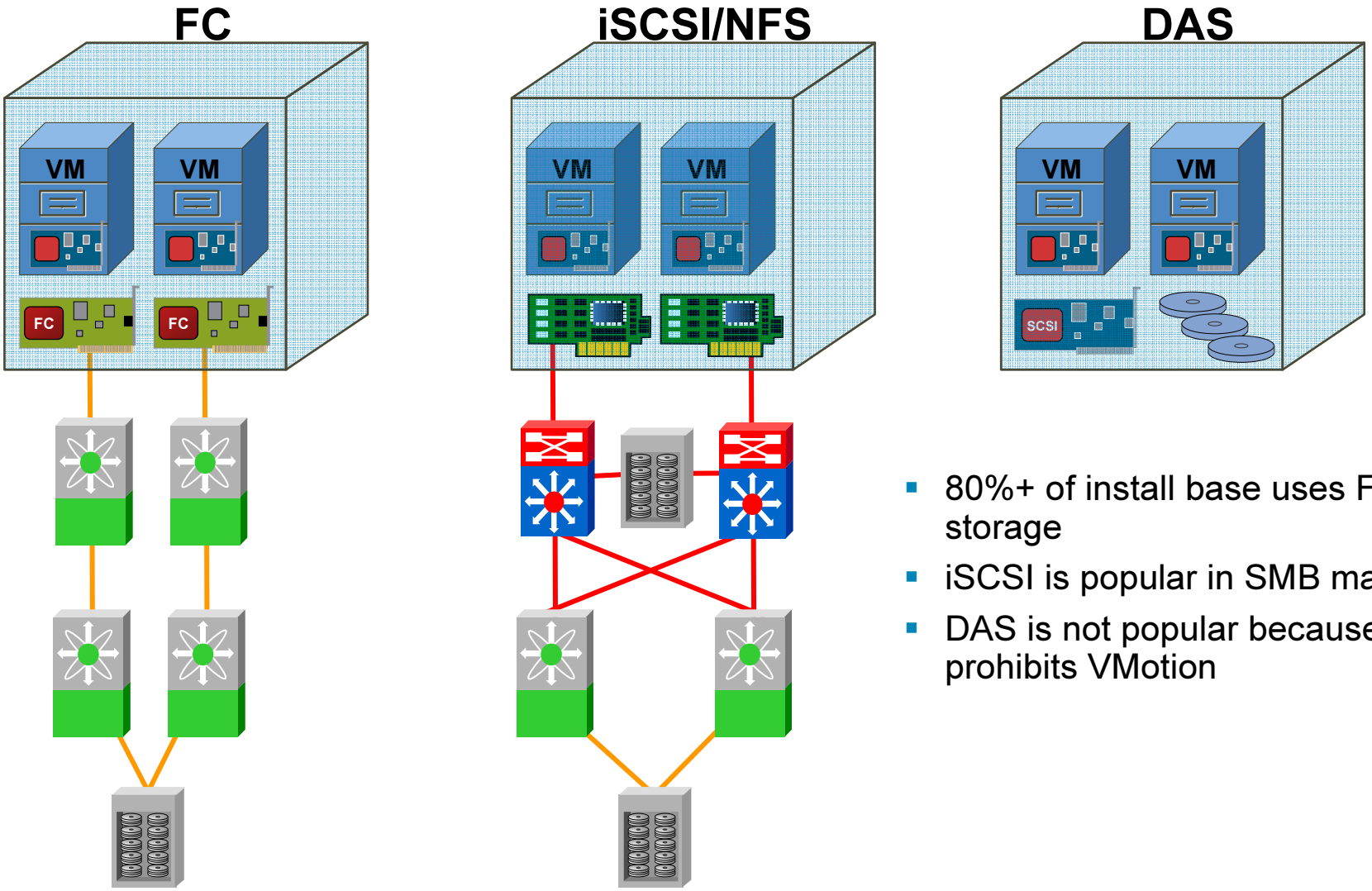
  VMWARE Storage

  Design Aspects

It's Just Another SAN Attached Host…



VMware ESX Storage Options



- 80%+ of install base uses FC storage
- iSCSI is popular in SMB market
- DAS is not popular because it prohibits VMotion

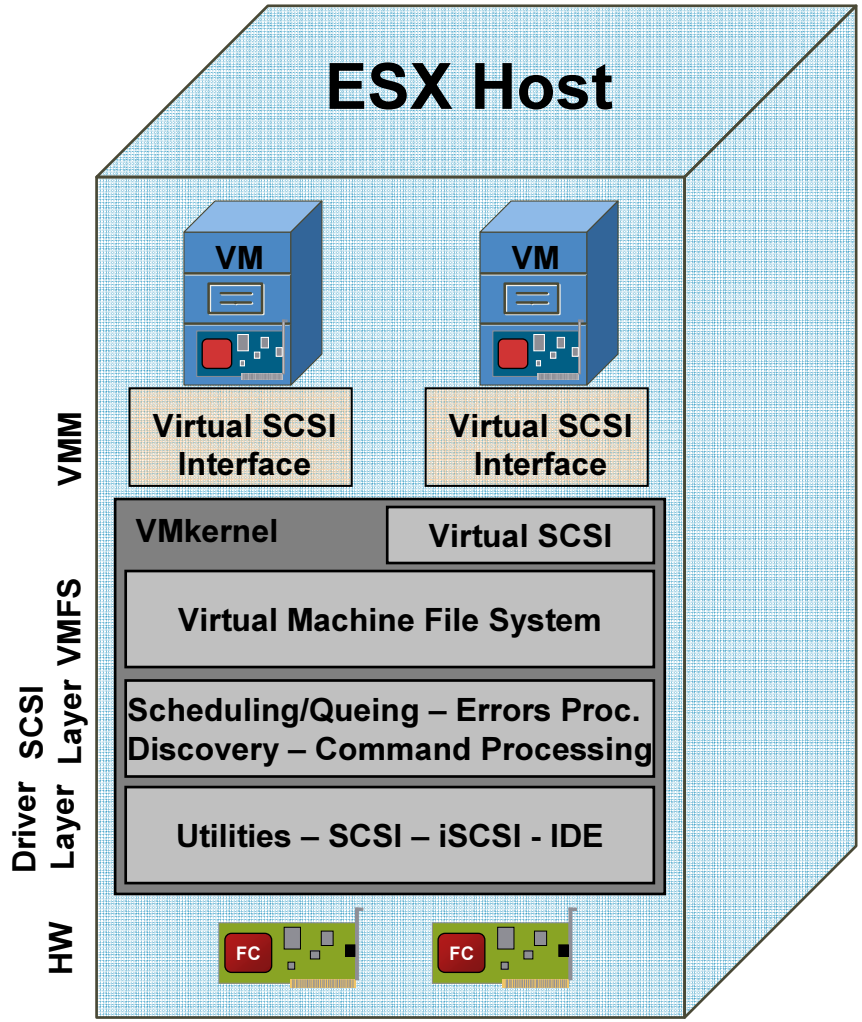
ESX Networked Storage Support

Type	Boot ESX Server	VMotion	VMFS	RDM	MSCS Support	VMware HA & DRS
FC	Yes	Yes	Yes	Yes	Yes	Yes
NFS	Yes	Yes	No	No	No	Yes
iSCSI (HW)*	Yes	Yes	Yes	Yes	No	Yes
iSCSI (SW)	No	Yes	Yes	Yes	No	Yes

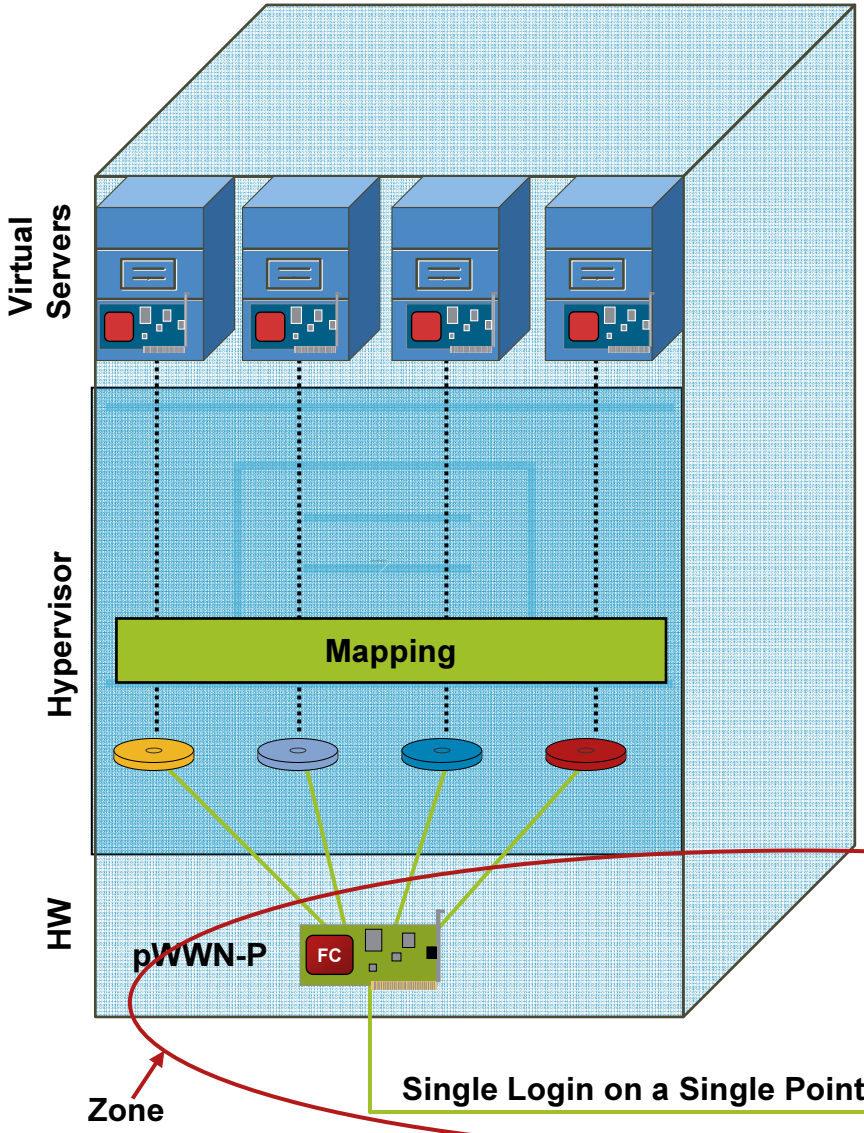
ESX FC Data Flow

- All storage shows up on virtual SCSI controller and appears as SCSI drive regardless of storage source

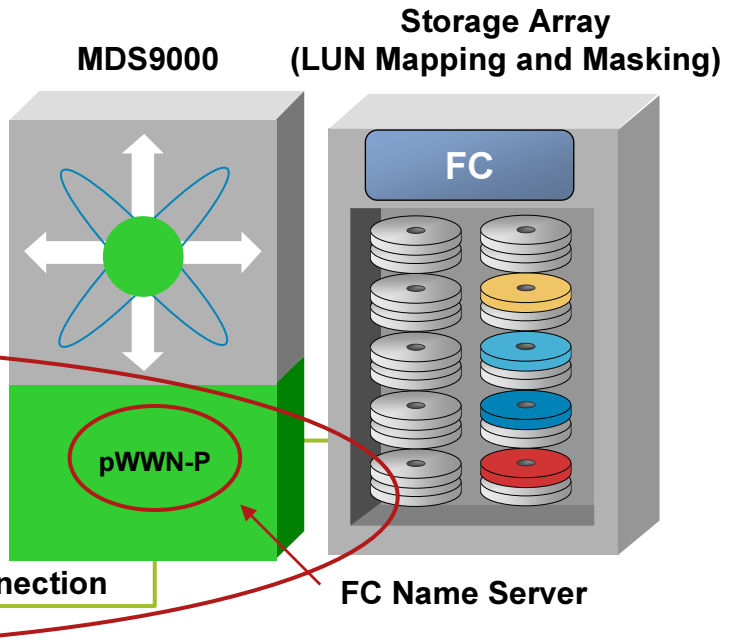
1. Virtual Machine guest OS issues Read/Write to disk
2. OS device driver sends request to the virtual SCSI controller
3. Virtual SCSI controller forwards command to the VMkernel
4. VMkernel locates VM file on VMFS, maps virtual to physical blocks, sends request to physical HBA driver
5. HBA sends FCP operations out the wire



Virtual Servers Share a Physical HBA

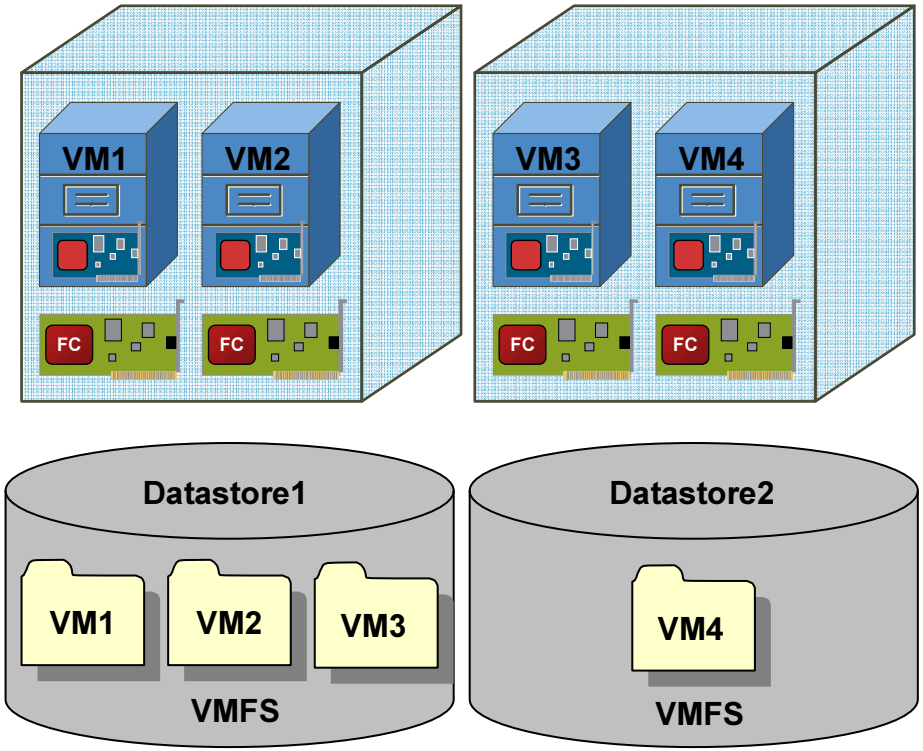


- A zone includes the physical hba and the storage array
- Access control is demanded to storage array "LUN masking and mapping", it is based on the physical HBA pWWN and it is the same for all VMs
- The hypervisor is in charge of the mapping, errors may be disastrous



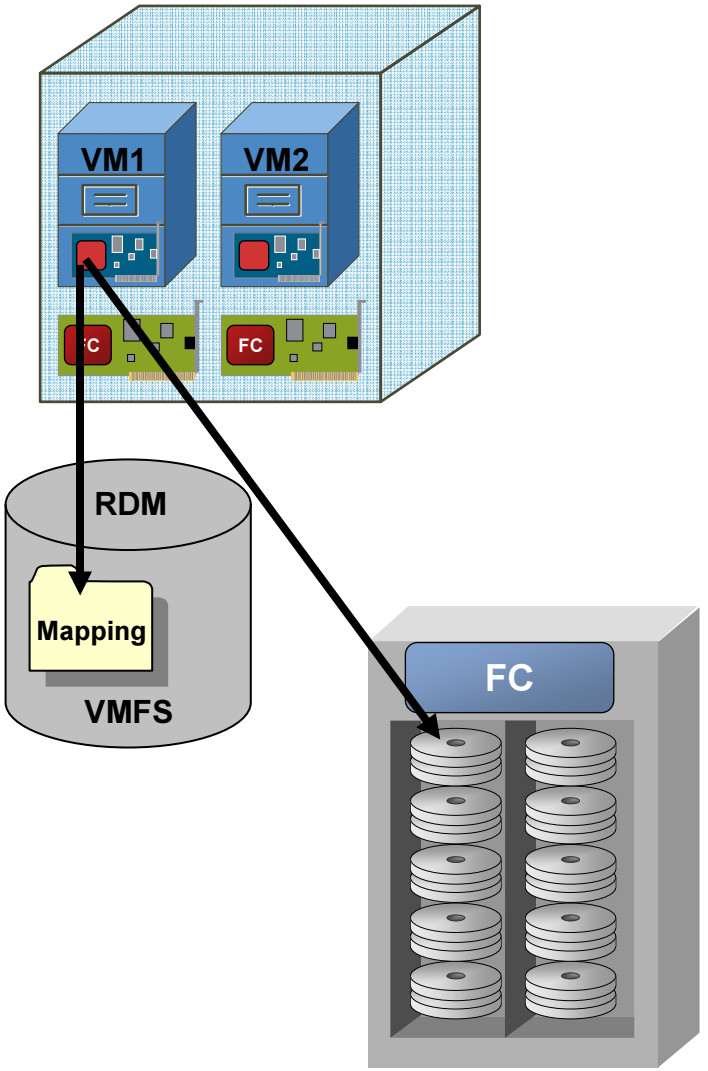
What Is a Datastore?

- A datastore is simply a pool of storage, internal or networked
- Can be VMFS-based or Raw-Mapped
- With networked storage a datastore is a cluster resource available to all ESX hosts
- To enable VMotion a datastore must be available to the source and destination ESX hosts
- Multiple datastores can be defined within a cluster



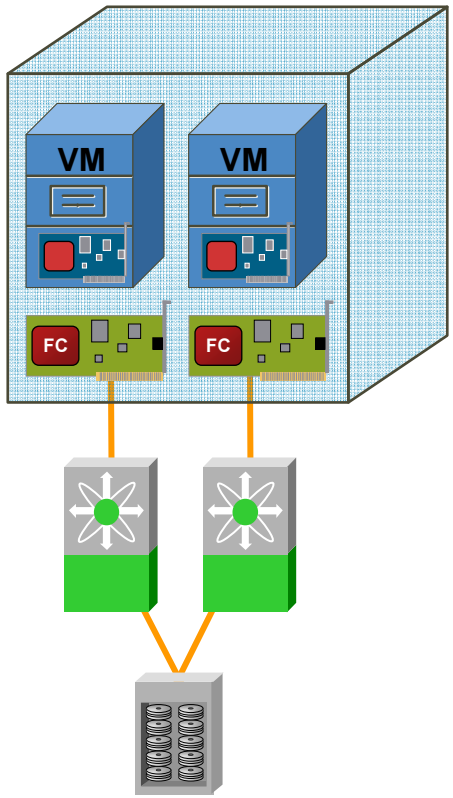
Raw Device Mapping

- RDM allows direct read/write access to disk
- Block mapping is still maintained within a VMFS file
- Rarely used but important for clustering (MSCS supported)
- Used with NPIV environments



Storage Multi-Pathing

- No storage load balancing, strictly failover
- Two modes of operation dictate behavior (Fixed and Most Recent)
- Fixed Mode
  - Allows definition of preferred paths
  - If preferred path fails a secondary path is used
  - If preferred path reappears it will fail back
- Most Recently Used
  - If current path fails a secondary path is used
  - If previous path reappears the current path is still used



ESX Storage Reference Documents

- ESX SAN Compatibility Guide
  http://www.vmware.com/pdf/vi3_san_guide.pdf
- VMware SAN Storage Design Guide

  http://www.vmware.com/pdf/vi3_san_design_deploy.pdf
- iSCSI Configuration Guide

  http://www.vmware.com/pdf/vi3_iscsi_cfg.pdf

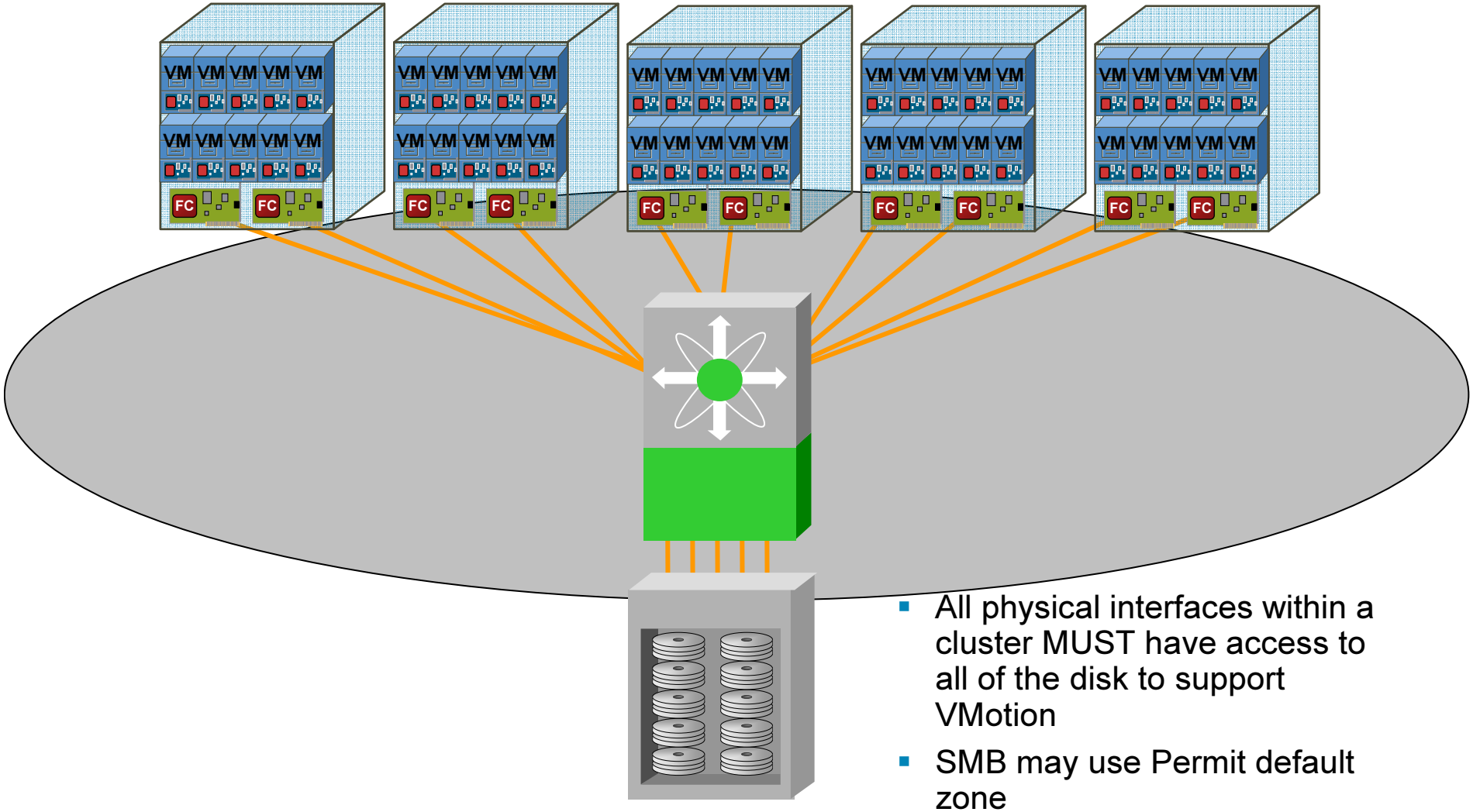
Agenda

- VMware Architecture and Components
- VMware LAN Networking
- Cisco/VMware DC DESIGNS
- Blade Server Designs
- Storage Implications of Server Virtualization

  VMWARE Storage

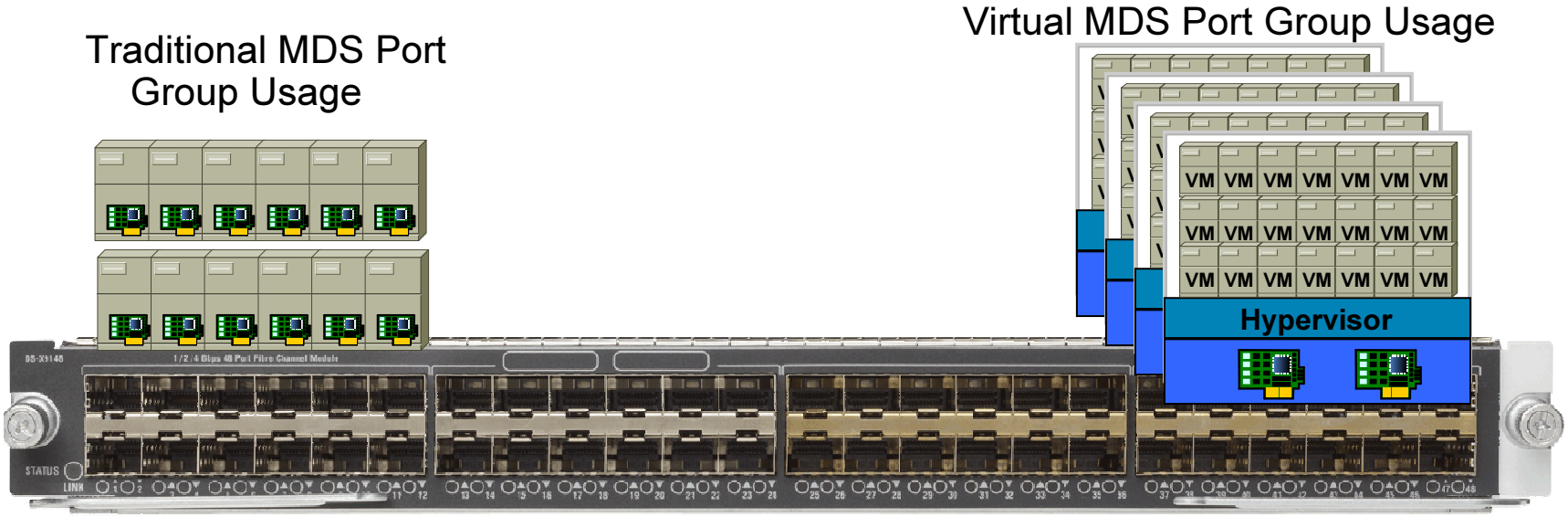
  Design Aspects

Zoning for VMotion



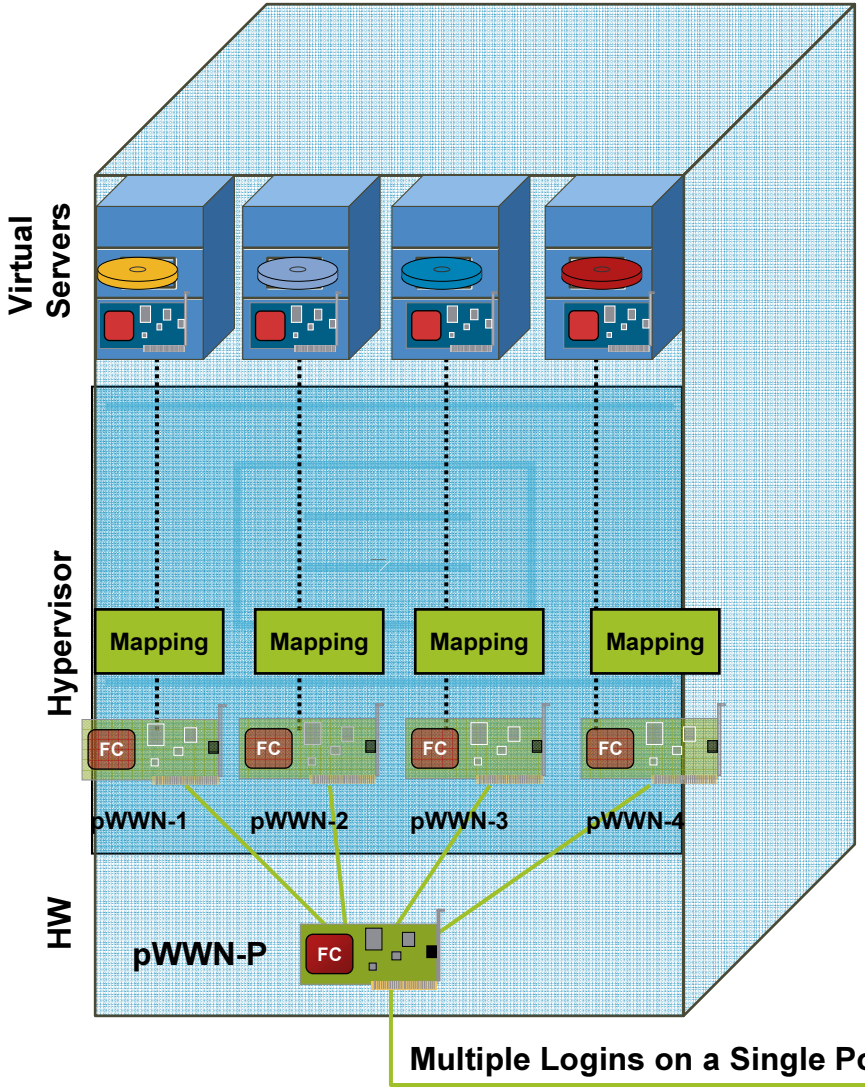
- All physical interfaces within a cluster MUST have access to all of the disk to support VMotion
- SMB may use Permit default zone
- Enterprise customers ideally will use many-to-many zone

Oversubscription Challenges



- Many customers target low I/O servers for VM consolidation but…
- Aggregation of multiple VMs on a single HBA increases bandwidth requirements on a per-port basis

Virtual Server Using NPIV and Storage Device Mapping



- Virtual HBAs can be zoned individually
- "LUN masking and mapping" is based on the virtual HBA pWWN of each VMs
- Very safe with respect to configuration errors
- Only supports RDM
- Availble in ESX 3.5