



VMware vSphere® Distributed Switch Best Practices

TECHNICAL WHITE PAPER

Table of Contents

Introduction	4
Design Considerations	4
Infrastructure Design Goals	4
Infrastructure Component Configurations	5
Virtual Infrastructure Traffic	5
Example Deployment Components	6
Hosts	6
Clusters	6
VMware vCenter Server	7
Network Infrastructure	7
Virtual Infrastructure Traffic Types	7
Important Virtual and Physical Switch Parameters	8
VDS Parameters	8
Host Uplink Connections (vmnics) and dvuplink Parameters	8
Traffic Types and dvportgroup Parameters	9
dvportgroup Specific Configuration	9
NIOC	10
Bidirectional Traffic Shaping	10
Physical Network Switch Parameters	11
VLAN	11
Spanning Tree Protocol	11
Link Aggregation Setup	11
Link-State Tracking	12
Maximum Transmission Unit	13
Rack Server in Example Deployment	13
Rack Server with Eight 1GbE Network Adaptors	13
Design Option 1 – Static Configuration	13
dvuplink Configuration	14
dvportgroup Configuration	14
Physical Switch Configuration	15
Design Option 2 – Dynamic Configuration with NIOC and LBT	17
dvportgroup Configuration	17

Rack Server with Two 10GbE Network Adaptors	20
Design Option 1 – Static Configuration	21
dvuplink Configuration	21
dvportgroup Configuration	21
Physical Switch Configuration	22
Design Option 2 – Dynamic Configuration with NIOC and LBT	23
dvportgroup Configuration	23
Blade Server in Example Deployment	25
Blade Server with Two 10GbE Network Adaptors	25
Design Option 1 – Static Configuration	26
Design Option 2 – Dynamic Configuration with NIOC and LBT	26
Blade Server with Hardware-Assisted Logical Network Adaptors (HP Flex-10- or Cisco UCS-like Deployment)	27
Operational Best Practices	28
VMware vSphere Command-Line Interface	28
VMware vSphere API	28
Virtual Network Monitoring and Troubleshooting	29
vCenter Server on a Virtual Machine	29
Conclusion	29

Introduction

This paper provides best practice guidelines for deploying the VMware vSphere® distributed switch (VDS) in a vSphere environment. The advanced capabilities of VDS provide network administrators with more control of and visibility into their virtual network infrastructure. This document covers the different considerations that vSphere and network administrators must take into account when designing the network with VDS. It also discusses some standard best practices for configuring VDS features.

The paper describes two example deployments, one using rack servers and the other using blade servers. For each of these deployments, different VDS design approaches are explained. The deployments and design approaches described in this document are meant to provide guidance as to what physical and virtual switch parameters, options and features should be considered during the design of a virtual network infrastructure. It is important to note that customers are not limited to the design options described in this paper. The flexibility of the vSphere platform allows for multiple variations in the design options that can fulfill an individual customer's unique network infrastructure needs.

This document is intended for vSphere and network administrators interested in understanding and deploying VDS in a virtual datacenter environment. With the release of vSphere 5, there are new features as well as enhancements to the existing features in VDS. To learn more about these new features and enhancements, refer to the *What's New in Networking* paper: <http://www.vmware.com/resources/techresources/10194>.

Readers are also encouraged to refer to basic virtual and physical networking concepts before reading through this document. The following link provides technical resources for virtual networking concepts: <http://www.vmware.com/technical-resources/virtual-networking/resources.html>

For physical networking concepts, readers should refer to any physical network switch vendor's documentation.

Design Considerations

The following three main aspects influence the design of a virtual network infrastructure:

- 1) Customer's infrastructure design goals
- 2) Customer's infrastructure component configurations
- 3) Virtual infrastructure traffic requirements

Let's take a look at each of these aspects in a little more detail.

Infrastructure Design Goals

Customers want their network infrastructure to be available 24/7, to be secure from any attacks, to perform efficiently throughout day-to-day operations, and to be easy to maintain. In the case of a virtualized environment, these requirements become increasingly demanding as growing numbers of business-critical applications run in a consolidated setting. These requirements on the infrastructure translate into design decisions that should incorporate the following best practices for a virtual network infrastructure:

- Avoid any single point of failure in the network.
- Isolate each traffic type for increased resiliency and security.
- Make use of traffic management and optimization capabilities.

Infrastructure Component Configurations

In every customer environment, the utilized compute and network infrastructures differ in terms of configuration, capacity and feature capabilities. These different infrastructure component configurations influence the virtual network infrastructure design decisions. The following are some of the configurations and features that administrators must look out for:

- Server configuration: rack or blade servers
- Network adaptor configuration: 1GbE or 10GbE network adaptors; number of available adaptors; offload function on these adaptors, if any
- Physical network switch infrastructure capabilities: switch clustering

It is impossible to cover all the different virtual network infrastructure design deployments based on the various combinations of type of servers, network adaptors and network switch capability parameters. In this paper, the following four commonly used deployments that are based on standard rack server and blade server configurations are described:

- Rack server with eight 1GbE network adaptors
- Rack server with two 10GbE network adaptors
- Blade server with two 10GbE network adaptors
- Blade server with hardware-assisted multiple logical Ethernet network adaptors

It is assumed that the network switch infrastructure has standard layer 2 switch features (high availability, redundant paths, fast convergence, port security) available to provide reliable, secure and scalable connectivity to the server infrastructure.

Virtual Infrastructure Traffic

vSphere virtual network infrastructure carries different traffic types. To manage the virtual infrastructure traffic effectively, vSphere and network administrators must understand the different traffic types and their characteristics. The following are the key traffic types that flow in the vSphere infrastructure, along with their traffic characteristics:

- Management traffic: This traffic flows through a vmknic and carries VMware ESXi™ host-to-VMware vCenter™ configuration and management communication as well as ESXi host-to-ESXi host high availability (HA)-related communication. This traffic has low network utilization but has very high availability and security requirements.
- VMware vSphere® vMotion® traffic: With advancement in vMotion technology, a single vMotion instance can consume almost a full 10Gb bandwidth. A maximum of eight simultaneous vMotion instances can be performed on a 10Gb uplink; four simultaneous vMotion instances are allowed on a 1Gb uplink. vMotion traffic has very high network utilization and can be bursty at times. Customers must make sure that vMotion traffic doesn't impact other traffic types, because it might consume all available I/O resources. Another property of vMotion traffic is that it is not sensitive to throttling and makes a very good candidate on which to perform traffic management.
- Fault-tolerant traffic: When VMware Fault Tolerance (FT) logging is enabled for a virtual machine, all the logging traffic is sent to the secondary fault-tolerant virtual machine over a designated vmknic port. This process can require a considerable amount of bandwidth at low latency because it replicates the I/O traffic and memory-state information to the secondary virtual machine.
- iSCSI/NFS traffic: IP storage traffic is carried over vmknic ports. This traffic varies according to disk I/O requests. With end-to-end jumbo frame configuration, more data is transferred with each Ethernet frame, decreasing the number of frames on the network. This larger frame reduces the overhead on servers/targets and improves the IP storage performance. On the other hand, congested and lower-speed networks can cause latency issues that disrupt access to IP storage. It is recommended that users provide a high-speed path for IP storage and avoid any congestion in the network infrastructure.

- Virtual machine traffic: Depending on the workloads that are running on the guest virtual machine, the traffic patterns will vary from low to high network utilization. Some of the applications running in virtual machines might be latency sensitive as is the case with VOIP workloads.

Table 1 summarizes the characteristics of each traffic type.

TRAFFIC TYPE	BANDWIDTH USAGE	OTHER TRAFFIC REQUIREMENTS
MANAGEMENT	Low	Highly reliable and secure channel
vMOTION	High	Isolated channel
FT	Medium to high	Highly reliable, low-latency channel
ISCSI	High	Reliable, high-speed channel
VIRTUAL MACHINE	Depends on application	Depends on application

Table 1. Traffic Types and Characteristics

To understand the different traffic flows in the physical network infrastructure, network administrators use network traffic management tools. These tools help monitor the physical infrastructure traffic but do not provide visibility into virtual infrastructure traffic. With the release of vSphere 5, VDS now supports the NetFlow feature, which enables exporting the internal (virtual machine-to-virtual machine) virtual infrastructure flow information to standard network management tools. Administrators now have the required visibility into virtual infrastructure traffic. This helps administrators monitor the virtual network infrastructure traffic through a familiar set of network management tools. Customers should make use of the network data collected from these tools during the capacity planning or network design exercises.

Example Deployment Components

After looking at the different design considerations, this section provides a list of components that are used in an example deployment. This example deployment helps illustrate some standard VDS design approaches. The following are some common components in the virtual infrastructure. The list doesn't include storage components that are required to build the virtual infrastructure. It is assumed that customers will deploy IP storage in this example deployment.

Hosts

Four ESXi hosts provide compute, memory and network resources according to the configuration of the hardware. Customers can have different numbers of hosts in their environment, based on their needs. One VDS can span across 350 hosts. This capability to support large numbers of hosts provides the required scalability to build a private or public cloud environment using VDS.

Clusters

A cluster is a collection of ESXi hosts and associated virtual machines with shared resources. Customers can have as many clusters in their deployment as are required. With one VDS spanning across 350 hosts, customers have the flexibility of deploying multiple clusters with a different number of hosts in each cluster. For simple illustration purposes, two clusters with two hosts each are considered in this example deployment. One cluster can have a maximum of 32 hosts.

VMware vCenter Server

VMware vCenter Server™ centrally manages a vSphere environment. Customers can manage VDS through this centralized management tool, which can be deployed on a virtual machine or a physical host. The vCenter Server system is not shown in the diagrams, but customers should assume that it is present in this example deployment. It is used only to provision and manage VDS configuration. When provisioned, hosts and virtual machine networks operate independently of vCenter Server. All components required for network switching reside on ESXi hosts. Even if the vCenter Server system fails, the hosts and virtual machines will still be able to communicate.

Network Infrastructure

Physical network switches in the access and aggregation layer provide connectivity between ESXi hosts and to the external world. These network infrastructure components support standard layer 2 protocols providing secure and reliable connectivity.

Along with the preceding four components of the physical infrastructure in this example deployment, some of the virtual infrastructure traffic types are also considered during the design. The following section describes the different traffic types in the example deployment.

Virtual Infrastructure Traffic Types

In this example deployment, there are standard infrastructure traffic types, including iSCSI, vMotion, FT, management and virtual machine. Customers might have other traffic types in their environment, based on their choice of storage infrastructure (FC, NFS, FCoE). Figure 1 shows the different traffic types along with associated port groups on an ESXi host. It also shows the mapping of the network adapters to the different port groups.

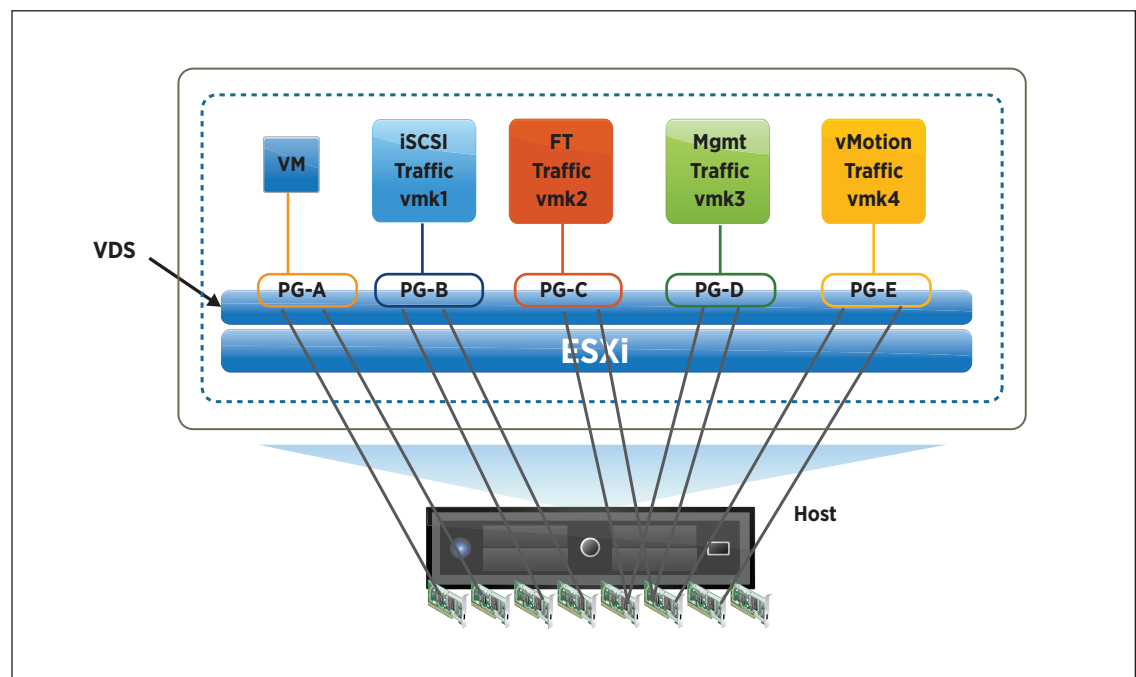


Figure 1. Different Traffic Types Running on a Host

Important Virtual and Physical Switch Parameters

Before going into the different design options in the example deployment, let's take a look at the virtual and physical network switch parameters that should be considered in all of the design options. These are some key parameters that vSphere and network administrators must take into account when designing VMware virtual networking. Because the configuration of virtual networking goes hand in hand with physical network configuration, this section will cover both the virtual and physical switch parameters.

VDS Parameters

VDS simplifies the challenges of the configuration process by providing one single pane of glass to perform virtual network management tasks. As opposed to configuring a vSphere standard switch (VSS) on each individual host, administrators can configure and manage one single VDS. All centrally configured network policies on VDS get pushed down to the host automatically when the host is added to the distributed switch. In this section, an overview of key VDS parameters is provided.

Host Uplink Connections (vmnics) and dvuplink Parameters

VDS has a new abstraction, called dvuplink, for the physical Ethernet network adaptors (vmnics) on each host. It is defined during the creation of the VDS and can be considered as a template for individual vmnics on each host. All the properties—including network adaptor teaming, load balancing and failover policies on VDS and dvportgroups—are configured on dvuplinks. These dvuplink properties are automatically applied to vmnics on individual hosts when a host is added to the VDS and when each vmnic on the host is mapped to a dvuplink. This dvuplink abstraction therefore provides the advantage of consistently applying teaming and failover configurations to all the host's physical Ethernet network adaptors (vmnics).

Figure 2 shows two ESXi hosts with four Ethernet network adaptors each. When these hosts are added to the VDS, with four dvuplinks configured on a dvuplink port group, administrators must assign the network adaptors (vmnics) of the hosts to dvuplinks. To illustrate the mapping of the dvuplinks to vmnics, Figure 2 shows one type of mapping where ESXi host vmnic0 is mapped to dvuplink1, vmnic1 to dvuplink2, and so on. Customers can choose different mapping, if required, where vmnic0 can be mapped to a different dvuplink instead of dvuplink1. VMware recommends having consistent mapping across different hosts because it reduces complexity in the environment.

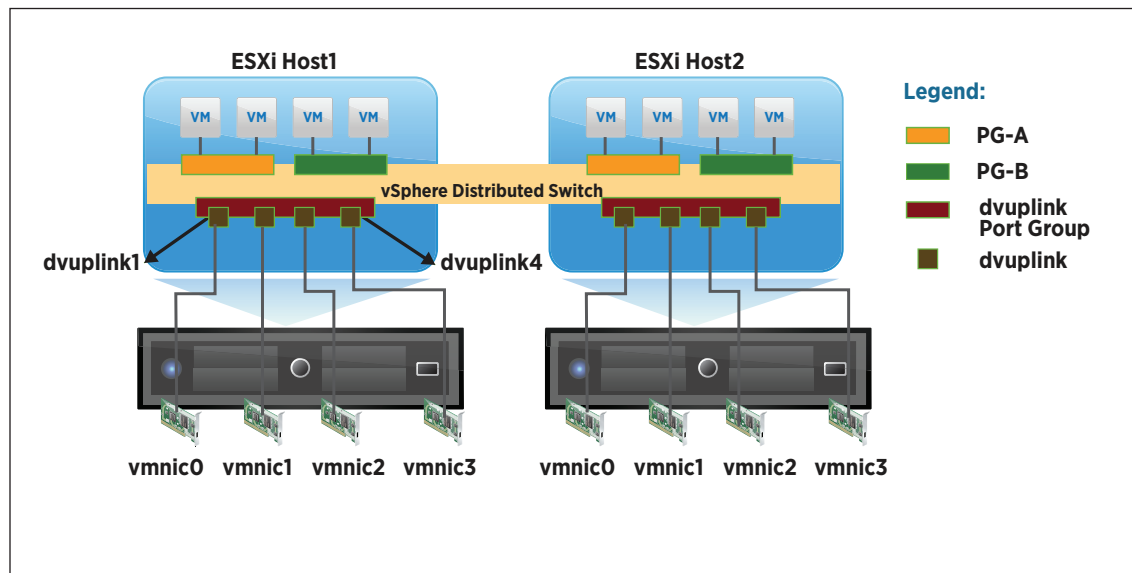


Figure 2. dvuplink-to-vmnic Mapping

As a best practice, customers should also try to deploy hosts with the same number of physical Ethernet network adaptors and with similar port speeds. Also, because the number of dvuplinks on VDS depends on the maximum number of physical Ethernet network adaptors on a host, administrators should take that into account during dvuplink port group configuration. Customers always have an option to modify this dvuplink configuration based on the new hardware capabilities.

Traffic Types and dvportgroup Parameters

Similar to port groups on standard switches, dvportgroups define how the connection is made through the VDS to the network. The VLAN ID, traffic shaping, port security, teaming and load balancing parameters are configured on these dvportgroups. The virtual ports (dvports) connected to a dvportgroup share the same properties configured on a dvportgroup. When customers want a group of virtual machines to share the security and teaming policies, they must make sure that the virtual machines are part of one dvportgroup. Customers can choose to define different dvportgroups based on the different traffic types they have in their environment or based on the different tenants or applications they support in the environment. If desired, multiple dvportgroups can share the same VLAN ID.

In this example deployment, the dvportgroup classification is based on the traffic types running in the virtual infrastructure. After administrators understand the different traffic types in the virtual infrastructure and identify specific security, reliability and performance requirements for individual traffic types, the next step is to create unique dvportgroups associated with each traffic type. As was previously mentioned, the dvportgroup configuration defined at VDS level is automatically pushed down to every host that is added to the VDS. For example, in Figure 2, the two dvportgroups, PG-A (yellow) and PG-B (green), defined at the distributed switch level are available on each of the ESXi hosts that are part of that VDS.

dvportgroup Specific Configuration

After customers decide on the number of unique dvportgroups they want to create in their environment, they can start configuring them. The configuration options/parameters are similar to those available with port groups on vSphere standard switches. There are some additional options available on VDS dvportgroups that are related to teaming setup and are not available on vSphere standard switches. Customers can configure the following key parameters for each dvportgroup.

- Number of virtual ports (dvports)
- Port binding (static, dynamic, ephemeral)
- VLAN trunking/private VLANs
- Teaming and load balancing along with active and standby links
- Bidirectional traffic-shaping parameters
- Port security

As part of the teaming algorithm support, VDS provides a unique approach to load balancing traffic across the teamed network adaptors. This approach is called load-based teaming (LBT), which distributes the traffic across the network adaptors based on the percentage utilization of traffic on those adaptors. LBT algorithm works on both ingress and egress direction of the network adaptor traffic, as opposed to the hashing algorithms that work only in egress direction (traffic flowing out of the network adaptor). Also, LBT prevents the worst-case scenario that might happen with hashing algorithms, where all traffic hashes to one network adaptor of the team while other network adaptors are not used to carry any traffic. To improve the utilization of all the links/network adaptors, VMware recommends the use of this advanced feature, LBT, of VDS. The LBT approach is recommended over the EtherChannel on physical switches and route-based IP hash configuration on the virtual switch.

Port security policies at port group level enable customer protection from certain activity that might compromise security. For example, a hacker might impersonate a virtual machine and gain unauthorized access by spoofing the virtual machine's MAC address. VMware recommends setting the MAC address "Changes" and "Forged Transmits" to "Reject" to help protect against attacks launched by a rogue guest operating system. Customers should set the "Promiscuous Mode" to "Reject" unless they want to monitor the traffic for network troubleshooting or intrusion detection purposes.

NIOC

Network I/O control (NIOC) is the traffic management capability available on VDS. The NIOC concept revolves around resource pools that are similar in many ways to the ones existing for CPU and memory. vSphere and network administrators now can allocate I/O shares to different traffic types similarly to allocating CPU and memory resources to a virtual machine. The share parameter specifies the relative importance of a traffic type over other traffic and provides a guaranteed minimum when the other traffic competes for a particular network adaptor. The shares are specified in abstract units numbered 1 to 100. Customers can provision shares to different traffic types based on the amount of resources each traffic type requires.

This capability of provisioning I/O resources is very useful in situations where there are multiple traffic types competing for resources. For example, in a deployment where vMotion and virtual machine traffic types are flowing through one network adaptor, it is possible that vMotion activity might impact the virtual machine traffic performance. In this situation, shares configured in NIOC provide the required isolation to the vMotion and virtual machine traffic type and prevent one flow (traffic type) from dominating the other flow. NIOC configuration provides one more parameter that customers can utilize if they want to put any limits on a particular traffic type. This parameter is called "the limit." The limit configuration specifies the absolute maximum bandwidth for a traffic type on a host. The configuration of the limit parameter is specified in Mbps. NIOC limits and shares parameters work only on the outbound traffic, i.e., traffic that is flowing out of the ESXi host.

VMware recommends that customers utilize this traffic management feature whenever they have multiple traffic types flowing through one network adaptor, a situation that is more prominent with 10 Gigabit Ethernet (GbE) network deployments but can happen in 1GbE network deployments as well. The common use case for using NIOC in 1GbE network adaptor deployments is when the traffic from different workloads or different customer virtual machines is carried over the same network adaptor. As multiple-workload traffic flows through a network adaptor, it becomes important to provide I/O resources based on the needs of the workload. With the release of vSphere 5, customers now can make use of the new user-defined network resource pools capability and can allocate I/O resources to the different workloads or different customer virtual machines, depending on their needs. This user-defined network resource pools feature provides the granular control in allocating I/O resources and meeting the service-level agreement (SLA) requirements for the virtualized tier 1 workloads.

Bidirectional Traffic Shaping

Besides NIOC, there is another traffic-shaping feature that is available in the vSphere platform. It can be configured on a dvportgroup or dvport level. Customers can shape both inbound and outbound traffic using three parameters: average bandwidth, peak bandwidth and burst size. Customers who want more granular traffic-shaping controls to manage their traffic types can take advantage of this capability of VDS along with the NIOC feature. It is recommended that network administrators in your organization be involved while configuring these granular traffic parameters. These controls make sense only when there are oversubscription scenarios—caused by the oversubscribed physical switch infrastructure or virtual infrastructure—that are causing network performance issues. So it is very important to understand the physical and virtual network environment before making any bidirectional traffic-shaping configurations.

Physical Network Switch Parameters

The configurations of the VDS and the physical network switch should go hand in hand to provide resilient, secure and scalable connectivity to the virtual infrastructure. The following are some key switch configuration parameters the customer should pay attention to.

VLAN

If VLANs are used to provide logical isolation between different traffic types, it is important to make sure that those VLANs are carried over to the physical switch infrastructure. To do so, enable virtual switch tagging (VST) on the virtual switch, and trunk all VLANs to the physical switch ports. For security reasons, it is recommended that customers not use the VLAN ID 1 (default) for any VMware infrastructure traffic.

Spanning Tree Protocol

Spanning Tree Protocol (STP) is not supported on virtual switches, so no configuration is required on VDS. But it is important to enable this protocol on the physical switches. STP makes sure that there are no loops in the network. As a best practice, customers should configure the following:

- Use PortFast on an ESXi host-facing physical switch ports. With this setting, network convergence on these switch ports will take place quickly after the failure because the port will enter the STP forwarding state immediately, bypassing the listening and learning states.
- Use the PortFast Bridge Protocol Data Unit (BPDU) guard feature to enforce the STP boundary. This configuration protects against any invalid device connection on the ESXi host-facing access switch ports. As was previously mentioned, VDS doesn't support STP, so it doesn't send any BPDU frames to the switch port. However, if any BPDU is seen on these ESXi host-facing access switch ports, the BPDU guard feature puts that particular switch port in error-disabled state. The switch port is completely shut down and prevents affecting the Spanning Tree Topology.

The recommendation of enabling PortFast and the BPDU guard feature on the switch ports is valid only when customers connect nonswitching/bridging devices to these ports. The switching/bridging devices can be hardware-based physical boxes or servers running a software-based switching/bridging function. Customers should make sure that there is no switching/bridging function enabled on the ESXi hosts that are connected to the physical switch ports.

However, in the scenario where the ESXi host has a guest virtual machine that is configured to perform a bridging function, the virtual machine will generate BPDU frames and send them out to the VDS, which then forwards the BPDU frames through the network adaptor to the physical switch port. When the switch port configured with BPDU guard receives the BPDU frame, the switch will disable the port and the virtual machine will lose connectivity. To avoid this network failure scenario when running the software bridging function on an ESXi host, customers should disable the PortFast and BPDU guard configuration on the physical switch port and run STP.

If customers are concerned about hacks that can generate BPDU frames, they should make use of VMware vShield App™, which can block the frames and protect the virtual infrastructure from such layer 2 attacks. Refer to VMware vShield™ product documentation for more details on how to secure your vSphere virtual infrastructure: <http://www.vmware.com/products/vshield/overview.html>.

Link Aggregation Setup

Link aggregation is used to increase throughput and improve resiliency by combining multiple network connections. There are various proprietary solutions on the market along with vendor-independent IEEE 802.3ad (LACP) standard-based implementation. All solutions establish a logical channel between the two endpoints, using multiple physical links. In the vSphere virtual infrastructure, the two ends of the logical channel are the VDS and physical switch. These two switches must be configured with link aggregation parameters before the logical channel is established. Currently, VDS supports static link aggregation configuration and does not provide support for dynamic LACP. When customers want to enable link aggregation on a physical switch, they should configure static link aggregation on the physical switch and select IP hash as network adaptor teaming on the VDS.

When establishing the logical channel with multiple physical links, customers should make sure that the Ethernet network adaptor connections from the host are terminated on a single physical switch. However, if customers have deployed clustered physical switch technology, the Ethernet network adaptor connections can be terminated on two different physical switches. The clustered physical switch technology is referred to by different names by networking vendors. For example, Cisco calls their switch clustering solution Virtual Switching System; Brocade calls theirs Virtual Cluster Switching. Refer to the networking vendor guidelines and configuration details when deploying switch clustering technology.

Link-State Tracking

Link-state tracking is a feature available on Cisco switches to manage the link state of downstream ports, ports connected to servers, based on the status of upstream ports, ports connected to aggregation/core switches. When there is any failure on the upstream links connected to aggregation or core switches, the associated downstream link status goes down. The server connected on the downstream link is then able to detect the failure and reroute the traffic on other working links. This feature therefore provides the protection from network failures due to the failed upstream ports in nonmesh topologies. Unfortunately, this feature is not available on all vendors' switches, and even if it is available, it might not be referred to as link-state tracking. Customers should talk to the switch vendors to find out whether a similar feature is supported on their switches.

Figure 3 shows the resilient mesh topology on the left and a simple loop-free topology on the right. VMware highly recommends deploying the mesh topology shown on the left, which provides highly reliable redundant design and doesn't need a link-state tracking feature. Customers who don't have high-end networking expertise and are also limited in number of switch ports might prefer the deployment shown on the right. In this deployment, customers don't have to run STP because there are no loops in the network design. The downside of this simple design is seen when there is a failure in the link between the access and aggregation switches. In that failure scenario, the server will continue to send traffic on the same network adaptor even when the access layer switch is dropping the traffic at the upstream interface. To avoid this blackholing of server traffic, customers can enable link-state tracking on the virtual and physical switches and indicate any failure between access and aggregation switch layers to the server through link-state information.

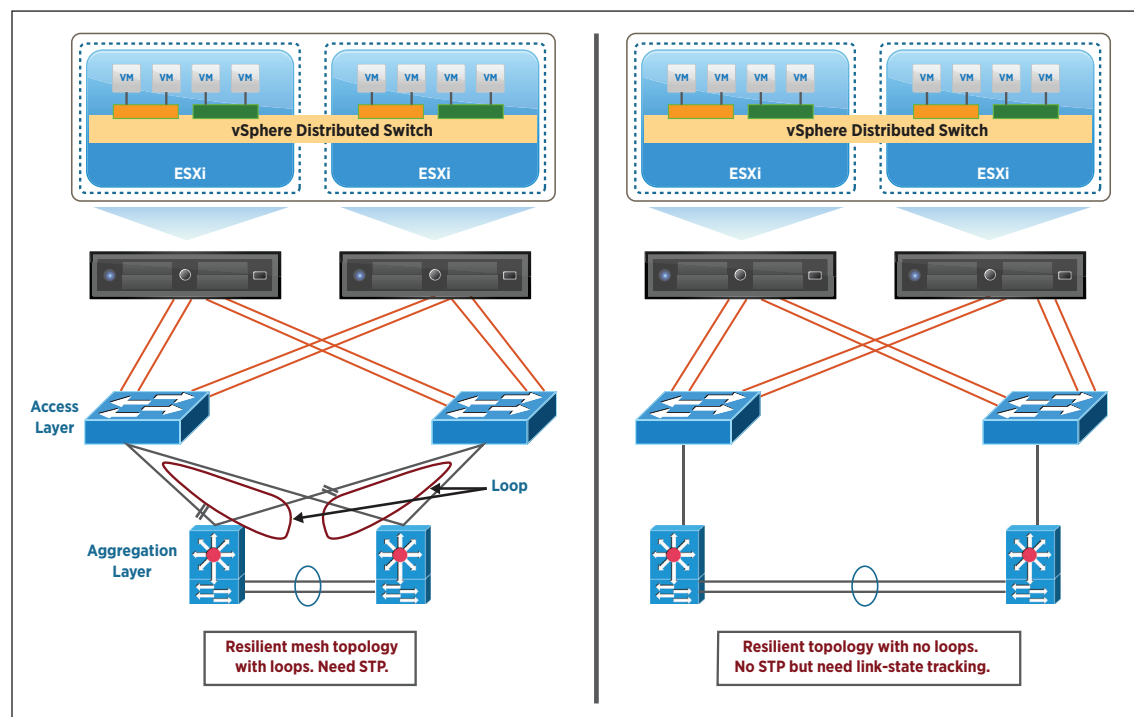


Figure 3. Resilient Loop and No-Loop Topologies

VDS has default network failover detection configuration set as “link status only.” Customers should keep this configuration if they are enabling the link-state tracking feature on physical switches. If link-state tracking capability is not available on physical switches, and there are no redundant paths available in the design, customers can make use of the beacon probing feature available on VDS. The beacon probing function is a software solution available on virtual switches for detecting link failures upstream from the access layer physical switch to the aggregation/core switches. Beacon probing is most useful with three or more uplinks in a team.

Maximum Transmission Unit

Make sure that the maximum transmission unit (MTU) configuration matches across the virtual and physical network switch infrastructure.

Rack Server in Example Deployment

After looking at the major components in the example deployment and key virtual and physical switch parameters, let's take a look at the different types of servers that customers can have in their environment. Customers can deploy an ESXi host on either a rack server or a blade server. This section discusses a deployment in which the ESXi host is running on a rack server. Two types of rack server configuration will be described in the following section:

- Rack server with eight 1GbE network adaptors
- Rack server with two 10GbE network adaptors

The various VDS design approaches will be discussed for each of the two configurations.

Rack Server with Eight 1GbE Network Adaptors

In a rack server deployment with eight 1GbE network adaptors per host, customers can either use the traditional static design approach of allocating network adaptors to each traffic type or make use of advanced features of VDS such as NIOC and LBT. The NIOC and LBT features help provide a dynamic design that efficiently utilizes I/O resources. In this section, both the traditional and new design approaches are described, along with their pros and cons.

Design Option 1 – Static Configuration

This design option follows the traditional approach of statically allocating network resources to the different virtual infrastructure traffic types. As shown in Figure 4, each host has eight Ethernet network adaptors. Four are connected to one of the first access layer switches; the other four are connected to the second access layer switch, to avoid single point of failure. Let's look in detail at how VDS parameters are configured.

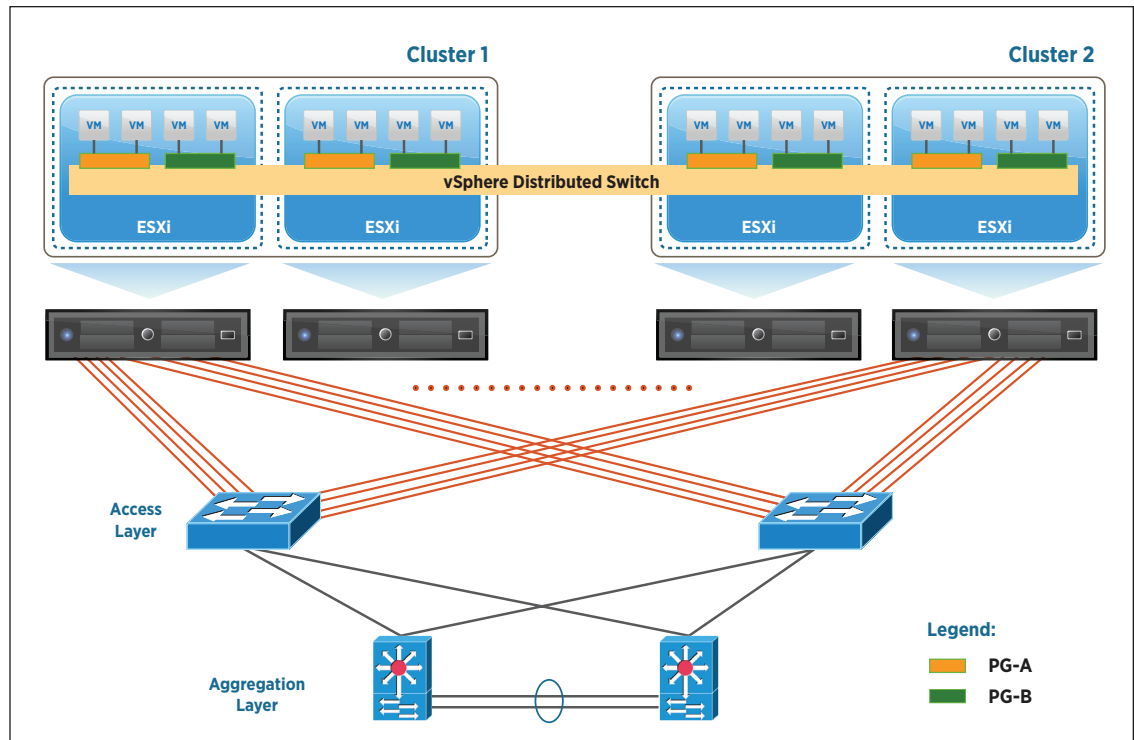


Figure 4. Rack Server with Eight 1GbE Network Adaptors

dvuplink Configuration

To support the maximum of eight 1GbE network adaptors per host, the dvuplink port group is configured with eight dvuplinks (dvuplink1–dvuplink8). On the hosts, dvuplink1 is associated with vmnic0, dvuplink2 is associated with vmnic1, and so on. It is a recommended practice to change the names of the dvuplinks to something meaningful and easy to track. For example, dvuplink1, which gets associated with vmnic on a motherboard, can be renamed as “LOM-uplink1”; dvuplink2, which gets associated with vmnic on an expansion card, can be renamed as “Expansion-uplink1.”

If the hosts have some Ethernet network adaptors as LAN on motherboard (LOM) and some on expansion cards, for a better resiliency story, VMware recommends selecting one network adaptor from LOM and one from an expansion card when configuring network adaptor teaming. To configure this teaming on a VDS, administrators must pay attention to the dvuplink and vmnic association along with dvportgroup configuration where network adaptor teaming is enabled. In the network adaptor teaming configuration on a dvportgroup, administrators must choose the various dvuplinks that are part of a team. If the dvuplinks are named appropriately according to the host vmnic association, administrators can select “LOM-uplink1” and “Expansion-uplink1” when configuring the teaming option for a dvportgroup.

dvportgroup Configuration

As described in Table 2, there are five different port groups that are configured for the five different traffic types. Customers can create up to 5,000 unique port groups per VDS. In this example deployment, the decision on creating different port groups is based on the number of traffic types.

According to Table 2, dvportgroup PG-A is created for the management traffic type. There are other dvportgroups defined for the other traffic types. The following are the key configurations of dvportgroup PG-A:

- Teaming option: Explicit failover order provides a deterministic way of directing traffic to a particular uplink. By selecting dvuplink1 as an active uplink and dvuplink2 as a standby uplink, management traffic will be carried over dvuplink1 unless there is a failure on dvuplink1. All other dvuplinks are configured as unused. Configuring the failback option to “No” is also recommended, to avoid the flapping of traffic between two network adaptors.

The failback option determines how a physical adaptor is returned to active duty after recovering from a failure. If failback is set to “No,” a failed adaptor is left inactive, even after recovery, until another currently active adaptor fails and requires a replacement.

- VMware recommends isolating all traffic types from each other by defining a separate VLAN for each dvportgroup.
- There are several other parameters that are part of the dvportgroup configuration. Customers can choose to configure these parameters based on their environment needs. For example, customers can configure PVLAN to provide isolation when there are limited VLANs available in the environment.

As you follow the dvportgroups configuration in Table 2, you can see that each traffic type is carried over a specific dvuplink, with the exception of the virtual machine traffic type. The virtual machine traffic type uses two active links, dvuplink7 and dvuplink8, and these links are utilized through the LBT algorithm. As was previously mentioned, the LBT algorithm is much more efficient than the standard hashing algorithm in utilizing link bandwidth.

TRAFFIC TYPE	PORT GROUP	TEAMING OPTION	ACTIVE UPLINK	STANDBY UPLINK	UNUSED UPLINK
MANAGEMENT	PG-A	Explicit Failover	dvuplink1	dvuplink2	3, 4, 5, 6, 7, 8
vMOTION	PG-B	Explicit Failover	dvuplink3	dvuplink4	1, 2, 5, 6, 7, 8
FT	PG-C	Explicit Failover	dvuplink4	dvuplink3	1, 2, 5, 6, 7, 8
ISCSI	PG-D	Explicit Failover	dvuplink5	dvuplink6	1, 2, 3, 4, 7, 8
VIRTUAL MACHINE	PG-E	LBT	dvuplink7/ dvuplink8	None	1, 2, 3, 4, 5, 6

Table 2. Static Design Configuration

Physical Switch Configuration

The external physical switch—where the rack servers’ network adaptors are connected to—is configured with trunk configuration with all the appropriate VLANs enabled. As described in the “Physical Network Switch Parameters” section, the following switch configurations are performed based on the VDS setup described in Table 2.

- Enable STP on the trunk ports facing the ESXi hosts, along with the PortFast mode and BPDU guard feature.
- The teaming configuration on VDS is static, so no link aggregation is configured on the physical switches.
- Because of the mesh topology deployment, as shown in Figure 4, the link-state tracking feature is not required on the physical switches.

In this design approach, resiliency to the infrastructure traffic is achieved through active/standby uplinks, and security is accomplished by providing separate physical paths for the different traffic types. However, with this design, the I/O resources are underutilized because the dvuplink2 and dvuplink6 standby links are not used to send or receive traffic. Also, there is no flexibility to allocate more bandwidth to a traffic type when it needs it.

There is another variation to the static design approach that addresses the need of some customers to provide higher bandwidth to the storage and vMotion traffic type. In the static design that was previously described, iSCSI and vMotion traffic is limited to 1GB. If a customer wants to support higher bandwidth for iSCSI, they can make use of the iSCSI multipathing solution. Also, with the release of vSphere 5, vMotion traffic can be carried over multiple Ethernet network adaptors through the support of multi-network adaptor vMotion, thereby providing higher bandwidth to the vMotion process.

For more details on how to set up iSCSI multipathing, refer to the VMware vSphere Storage guide:
<https://www.vmware.com/support/pubs/vsphere-esxi-vcenter-server-pubs.html>.

The configuration of multi-network adaptor vMotion is quite similar to the iSCSI multipath setup, where administrators must create two separate vmkernel interfaces and bind each one to a separate dvportgroup. This configuration with two separate dvportgroups provides the connectivity to two different Ethernet network adaptors or dvuplinks.

TRAFFIC TYPE	PORT GROUP	TEAMING OPTION	ACTIVE UPLINK	STANDBY UPLINK	UNUSED UPLINK
MANAGEMENT	PG-A	Explicit Failover	dvuplink1	dvuplink2	3, 4, 5, 6, 7, 8
vMOTION	PG-B1	None	dvuplink3	dvuplink4	1, 2, 5, 6, 7, 8
vMOTION	PG-B2	None	dvuplink4	dvuplink3	1, 2, 5, 6, 7, 8
FT	PG-C	Explicit Failover	dvuplink2	dvuplink1	3, 4, 5, 6, 7, 8
iSCSI	PG-D1	None	dvuplink5	None	1, 2, 3, 4, 6, 7, 8
iSCSI	PG-D2	None	dvuplink6	None	1, 2, 3, 4, 5, 7, 8
VIRTUAL MACHINE	PG-E	LBT	dvuplink7/ dvuplink8	None	1, 2, 3, 4, 5, 6

Table 3. Static Design Configuration with iSCSI Multipathing and Multi-Network Adaptor vMotion

As shown in Table 3, there are two entries each for the vMotion and iSCSI traffic types. Also shown is a list of the additional dvportgroup configurations required to support the multi-network adaptor vMotion and iSCSI multipathing processes. For multi-network adaptor vMotion, dvportgroups PG-B1 and PG-B2 are listed, configured with dvuplink 3 and dvuplink4 respectively as active links. And for iSCSI multipathing, dvportgroups PG-D1 and PG-D2 are connected to dvuplink5 and dvuplink6 respectively as active links. Load balancing across the multiple dvuplinks is performed by the multipathing logic in the iSCSI process and by the ESXi platform in the vMotion process. Configuring the teaming policies for these dvportgroups is not required.

FT, management and virtual machine traffic-type dvportgroup configuration and physical switch configuration for this design remain the same as those described in “Design Option 1” of the previous section.

This static design approach improves on the first design by using advanced capabilities such as iSCSI multipathing and multi-network adaptor vMotion. But at the same time, this option has the same challenges related to underutilized resources and inflexibility in allocating additional resources on the fly to different traffic types.

Design Option 2 – Dynamic Configuration with NIOC and LBT

After looking at the traditional design approach with static uplink configurations, let's take a look at the VMware-recommended design option that takes advantage of the advanced VDS features such as NIOC and LBT.

In this design, the connectivity to the physical network infrastructure remains the same as that described in the static design option. However, instead of allocating specific dvuplinks to individual traffic types, the ESXi platform utilizes those dvuplinks dynamically. To illustrate this dynamic design, each virtual infrastructure traffic type's bandwidth utilization is estimated. In a real deployment, customers should first monitor the virtual infrastructure traffic over a period of time, to gauge the bandwidth utilization, and then come up with bandwidth numbers for each traffic type. The following are some bandwidth numbers estimated by traffic type:

- Management traffic (<1GB)
- vMotion (1GB)
- FT (1GB)
- iSCSI (1GB)
- Virtual machine (2GB)

Based on this bandwidth information, administrators can provision appropriate I/O resources to each traffic type by using the NIOC feature of VDS. Let's take a look at the VDS parameter configurations for this design, as well as the NIOC setup. The dvuplink port group configuration remains the same, with eight dvuplinks created for the eight 1 GbE network adaptors. The dvportgroup configuration is described in the following section.

dvportgroup Configuration

In this design, all dvuplinks are active and there are no standby and unused uplinks, as shown in Table 4. All dvuplinks are therefore available for use by the teaming algorithm. The following are the key parameter configurations of dvportgroup PG-A:

- Teaming option: LBT is selected as the teaming algorithm. With LBT configuration, the management traffic initially will be scheduled based on the virtual port ID hash. Depending on the hash output, management traffic is sent out over one of the dvuplinks. Other traffic types in the virtual infrastructure can also be scheduled on the same dvuplink initially. However, when the utilization of the dvuplink goes beyond the 75 percent threshold, the LBT algorithm will be invoked and some of the traffic will be moved to other underutilized dvuplinks. It is possible that management traffic will be moved to other dvuplinks when such an LBT event occurs.
- The fallback option means going from using a standby link to using an active uplink after the active uplink comes back into operation after a failure. This fallback option works when there are active and standby dvuplink configurations. In this design, there are no standby dvuplinks. So when an active uplink fails, the traffic flowing on that dvuplink is moved to another working dvuplink. If the failed dvuplink comes back, the LBT algorithm will schedule new traffic on that dvuplink. This option is left as the default.
- VMware recommends isolating all traffic types from each other by defining a separate VLAN for each dvportgroup.
- There are several other parameters that are part of the dvportgroup configuration. Customers can choose to configure these parameters based on their environment needs. For example, they can configure PVLAN to provide isolation when there are limited VLANs available in the environment.

As you follow the dvportgroups configuration in Table 4, you can see that each traffic type has all dvuplinks active and that these links are utilized through the LBT algorithm. Let's now look at the NIOC configuration described in the last two columns of Table 4.

The NIOC configuration in this design helps provide the appropriate I/O resources to the different traffic types. Based on the previously estimated bandwidth numbers per traffic type, the shares parameter is configured in the NIOC shares column in Table 4. The shares values specify the relative importance of specific traffic types, and NIOC ensures that during contention scenarios on the dvuplinks, each traffic type gets the allocated bandwidth. For example, a shares configuration of 10 for vMotion, iSCSI and FT allocates equal bandwidth to these traffic types. Virtual machines get the highest bandwidth with 20 shares and management gets lower bandwidth with 5 shares.

To illustrate how share values translate to bandwidth numbers, let's take an example of 1Gb capacity dvuplink carrying all five traffic types. This is a worst-case scenario where all traffic types are mapped to one dvuplink. This will never happen when customers enable the LBT feature, because LBT will balance the traffic based on the utilization of uplinks. This example shows how much bandwidth each traffic type will be allowed on one dvuplink during a contention or oversubscription scenario and when LBT is not enabled.

- Total shares: management (5) + vMotion (10) + FT (10) + iSCSI (10) + virtual machine (20) = 55
- Management: 5 shares; $(5/55) * 1\text{Gb} = 90.91\text{Mbps}$
- vMotion: 10 shares; $(10/55) * 1\text{Gb} = 181.18\text{Mbps}$
- FT: 10 shares; $(10/55) * 1\text{Gb} = 181.18\text{Mbps}$
- iSCSI: 10 shares; $(10/55) * 1\text{Gb} = 181.18\text{Mbps}$
- Virtual machine: 20 shares; $(20/55) * 1\text{Gb} = 363.64\text{Mbps}$

To calculate the bandwidth numbers during contention, you should first calculate the percentage of bandwidth for a traffic type by dividing its share value by the total available share number (55). In the second step, the total bandwidth of the dvuplink (1Gb) is multiplied with the percentage of bandwidth number calculated in the first step. For example, 5 shares allocated to management traffic translate to 90.91Mbps of bandwidth to management process on a fully utilized 1Gb network adaptor. In this example, custom share configuration is discussed, but a customer can make use of predefined high (100), normal (50) and low (25) shares when assigning them to different traffic types.

The vSphere platform takes these configured share values and applies them per uplink. The schedulers running at each uplink are responsible for making sure that the bandwidth resources are allocated according to the shares. In the case of an eight 1GbE network adaptor deployment, there are eight schedulers running. Depending on the number of traffic types scheduled on a particular uplink, the scheduler will divide the bandwidth among the traffic types, based on the share numbers. For example, if only FT (10 shares) and management (5 shares) traffic are flowing through dvuplink 5, FT traffic will get double the bandwidth of management traffic, based on the shares value. Also, when there is no management traffic flowing, all bandwidth can be utilized by the FT process. This flexibility in allocating I/O resources is the key benefit of the NIOC feature.

The NIOC limits parameter of Table 4 is not configured in this design. The limits value specifies an absolute maximum limit on egress traffic for a traffic type. Limits are specified in Mbps. This configuration provides a hard limit on any traffic, even if I/O resources are available to use. Using limits configuration is not recommended unless you really want to control the traffic, even though additional resources are available.

There is no change in physical switch configuration in this design approach, even with the choice of the new LBT algorithm. The LBT teaming algorithm doesn't require any special configuration on physical switches. Refer to the physical switch settings described in "Design Option 1."

TRAFFIC TYPE	PORT GROUP	TEAMING OPTION	ACTIVE UPLINK	STANDBY UPLINK	NIOC SHARES	NIOC LIMITS
MANAGEMENT	PG-A	LBT	1, 2, 3, 4, 5, 6, 7, 8	None	5	–
vMOTION	PG-B	LBT	1, 2, 3, 4, 5, 6, 7, 8	None	10	–
FT	PG-C	LBT	1, 2, 3, 4, 5, 6, 7, 8	None	10	–
ISCSI	PG-D	LBT	1, 2, 3, 4, 5, 6, 7, 8	None	10	–
VIRTUAL MACHINE	PG-E	LBT	1, 2, 3, 4, 5, 6, 7, 8	None	20	–

Table 4. Dynamic Design Configuration with NIOC and LBT

This design does not provide higher than 1Gb bandwidth to the vMotion and iSCSI traffic types as is the case with static design using multi-network adaptor vMotion and iSCSI multipathing. The LBT algorithm cannot split the infrastructure traffic across multiple dvuplink ports and utilize all the links. So even if vMotion dvportgroup PG-B has all eight 1GbE network adaptors as active uplinks, vMotion traffic will be carried over only one of the eight uplinks. The main advantage of this design is evident in the scenarios where the vMotion process is not using the uplink bandwidth, and other traffic types are in need of the additional resources. In these situations, NIOC makes sure that the unused bandwidth is allocated to the other traffic types that need it.

This dynamic design option is the recommended approach because it takes advantage of the advanced VDS features and utilizes I/O resources efficiently. This option also provides active-active resiliency where no uplinks are in standby mode. In this design approach, customers allow the vSphere platform to make the optimal decisions on scheduling traffic across multiple uplinks.

Some customers who have restrictions in the physical infrastructure in terms of bandwidth capacity across different paths and limited availability of the layer 2 domain might not be able to take advantage of this dynamic design option. When deploying this design option, it is important to consider all the different traffic paths that a traffic type can take and to make sure that the physical switch infrastructure can support the specific characteristics required for each traffic type. VMware recommends that vSphere and network administrators work together to understand the impact of the vSphere platform's traffic scheduling feature over the physical network infrastructure before deploying this design option.

Every customer environment is different, and the requirements for the traffic types are also different. Depending on the need of the environment, a customer can modify these design options to fit their specific requirements. For example, customers can choose to use a combination of static and dynamic design options when they need higher bandwidth for iSCSI and vMotion activities. In this hybrid design, four uplinks can be statically allocated to iSCSI and vMotion traffic types while the remaining four uplinks are used dynamically for the remaining traffic types. Table 5 shows the traffic types and associated port group configurations for the hybrid design. As shown in the table, management, FT and virtual machine traffic will be distributed on dvuplink1 to dvuplink4 through the vSphere platform's traffic scheduling features, LBT and NIOC. The remaining four dvuplinks are statically assigned to vMotion and iSCSI traffic types.

TRAFFIC TYPE	PORT GROUP	TEAMING OPTION	ACTIVE UPLINK	STANDBY UPLINK	NIOC SHARES	NIOC LIMITS
MANAGEMENT	PG-A	LBT	1, 2, 3, 4	None	5	–
vMOTION	PG-B1	None	5	6	–	–
vMOTION	PG-B2	None	6	5	–	–
FT	PG-C	LBT	1, 2, 3, 4	None	10	–
ISCSI	PG-D1	None	7	None	–	–
ISCSI	PG-D2	None	8	None	–	–
VIRTUAL MACHINE	PG-E	LBT	1, 2, 3, 4	None	20	–

Table 5. Hybrid Design Configuration

Rack Server with Two 10GbE Network Adaptors

The two 10GbE network adaptors deployment model is becoming very common because of the benefits they provide through I/O consolidation. The key benefits include better utilization of I/O resources, simplified management and reduced CAPEX and OPEX. Although this deployment provides these benefits, there are some challenges when it comes to the traffic management aspects. Especially in highly consolidated virtualized environments where more traffic types are carried over fewer 10GbE network adaptors, it becomes critical to prioritize traffic types that are important and provide the required SLA guarantees. The NIOC feature available on the VDS helps in this traffic management activity. In the following sections, you will see how to utilize this feature in the different designs.

As shown in Figure 5, rack servers with two 10GbE network adaptors are connected to the two access layer switches to avoid any single point of failure. Similar to the rack server with eight 1GbE network adaptors, the different VDS and physical switch parameter configurations are taken into account with this design. On the physical switch side, the new 10Gb switches might have support for FCoE that enables convergence for SAN and LAN traffic. This document covers only the standard 10Gb deployments that support IP storage traffic (iSCSI/NFS) and not FCoE.

In this section, two design options are described; one is a traditional approach and the other one is a VMware-recommended approach.

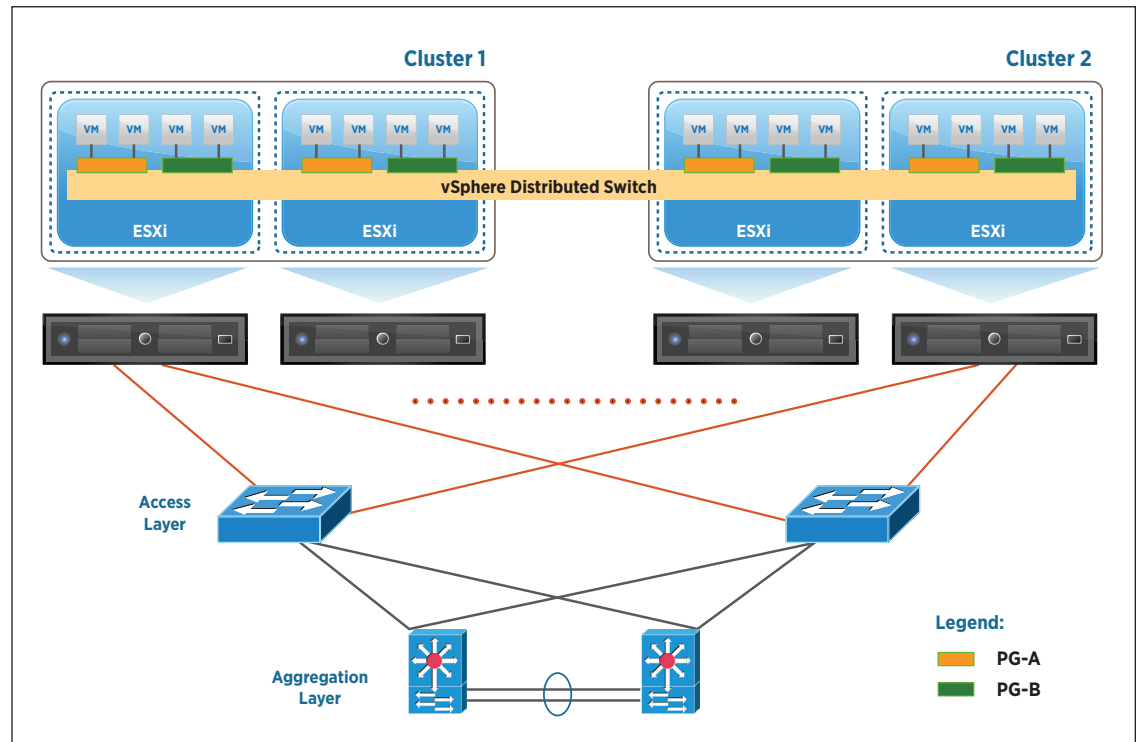


Figure 5. Rack Server with Two 10GbE Network Adaptors

Design Option 1 – Static Configuration

The static configuration approach for rack server deployment with 10GbE network adaptors is similar to the one described in “Design Option 1” of rack server deployment with eight 1GbE adaptors. There are a few differences in the configuration where the numbers of dvuplinks are changed from eight to two, and dvportgroup parameters are different. Let’s take a look at the configuration details on the VDS front.

dvuplink Configuration

To support the maximum two Ethernet network adaptors per host, the dvuplink port group is configured with two dvuplinks (dvuplink1, dvuplink2). On the hosts, dvuplink1 is associated with vmnic0 and dvuplink2 is associated with vmnic1.

dvportgroup Configuration

As described in Table 6, there are five different dvportgroups that are configured for the five different traffic types. For example, dvportgroup PG-A is created for the management traffic type. The following are the other key configurations of dvportgroup PG-A:

- Teaming option: An explicit failover order provides a deterministic way of directing traffic to a particular uplink. By selecting dvuplink1 as an active uplink and dvuplink2 as a standby uplink, management traffic will be carried over dvuplink1 unless there is a failure with it. Configuring the failback option to “No” is also recommended, to avoid the flapping of traffic between two network adaptors. The failback option determines how a physical adaptor is returned to active duty after recovering from a failure. If failback is set to “No,” a failed adaptor is left inactive, even after recovery, until another currently active adaptor fails, requiring its replacement.
- VMware recommends isolating all traffic types from each other by defining a separate VLAN for each dvportgroup.

- There are various other parameters that are part of the dvportgroup configuration. Customers can choose to configure these parameters based on their environment needs.

Table 6 provides the configuration details for all the dvportgroups. According to the configuration, dvuplink1 carries management, iSCSI and virtual machine traffic; dvuplink2 handles vMotion, FT and virtual machine traffic. As you can see, the virtual machine traffic type makes use of two uplinks, and these uplinks are utilized through the LBT algorithm.

With this deterministic teaming policy, customers can decide to map different traffic types to the available uplink ports, depending on environment needs. For example, if iSCSI traffic needs higher bandwidth and other traffic types have relatively low bandwidth requirements, customers can decide to keep only iSCSI traffic on dvuplink1 and move all other traffic to dvuplink2. When deciding on these traffic paths, customers should understand the physical network connectivity and the paths' bandwidth capacities.

Physical Switch Configuration

The external physical switch, which the rack servers' network adaptors are connected to, has trunk configuration with all the appropriate VLANs enabled. As described in the physical network switch parameters sections, the following switch configurations are performed based on the VDS setup described in Table 6.

- Enable STP on the trunk ports facing ESXi hosts, along with the PortFast mode and BPDU guard feature.
- The teaming configuration on VDS is static and therefore no link aggregation is configured on the physical switches.
- Because of the mesh topology deployment shown in Figure 5, the link state-tracking feature is not required on the physical switches.

TRAFFIC TYPE	PORT GROUP	TEAMING OPTION	ACTIVE UPLINK	STANDBY UPLINK	UNUSED UPLINK
MANAGEMENT	PG-A	Explicit Failover	dvuplink1	dvuplink2	None
vMOTION	PG-B	Explicit Failover	dvuplink2	dvuplink1	None
FT	PG-C	Explicit Failover	dvuplink2	dvuplink1	None
ISCSI	PG-D	Explicit Failover	dvuplink1	dvuplink2	None
VIRTUAL MACHINE	PG-E	LBT	dvuplink1/ dvuplink2	None	None

Table 6. Static Design Configuration

This static design option provides flexibility in the traffic path configuration, but it cannot protect against one traffic type's dominating others. For example, there is a possibility that a network-intensive vMotion process might take away most of the network bandwidth and impact virtual machine traffic. Bidirectional traffic-shaping parameters at port group and port levels can provide some help in managing different traffic rates. However, using this approach for traffic management requires customers to limit the traffic on the respective dvportgroups. Limiting traffic to a certain level through this method puts a hard limit on the traffic types, even when the bandwidth is available to utilize. This underutilization of I/O resources because of hard limits is overcome through the NIOC feature, which provides flexible traffic management based on the shares parameters. "Design Option 2," described in the following section, is based on the NIOC feature.

Design Option 2 – Dynamic Configuration with NIOC and LBT

This dynamic design option is the VMware-recommended approach that takes advantage of the NIOC and LBT features of the VDS.

Connectivity to the physical network infrastructure remains the same as that described in “Design Option 1.” However, instead of allocating specific dvuplinks to individual traffic types, the ESXi platform utilizes those dvuplinks dynamically. To illustrate this dynamic design, each virtual infrastructure traffic type's bandwidth utilization is estimated. In a real deployment, customers should first monitor the virtual infrastructure traffic over a period of time to gauge the bandwidth utilization, and then come up with bandwidth numbers.

The following are some bandwidth numbers estimated by traffic type:

- Management traffic (<1GB)
- vMotion (2GB)
- FT (1GB)
- iSCSI (2GB)
- Virtual machine (2GB)

These bandwidth estimates are different from the one considered with rack server deployment with eight 1GbE network adaptors. Let's take a look at the VDS parameter configurations for this design. The dvuplink port group configuration remains the same, with two dvuplinks created for the two 10GbE network adaptors. The dvportgroup configuration is as follows.

dvportgroup Configuration

In this design, all dvuplinks are active and there are no standby and unused uplinks, as shown in Table 7. All dvuplinks are therefore available for use by the teaming algorithm. The following are the key configurations of dvportgroup PG-A:

- Teaming option: LBT is selected as the teaming algorithm. With LBT configuration, management traffic initially will be scheduled based on the virtual port ID hash. Based on the hash output, management traffic will be sent out over one of the dvuplinks. Other traffic types in the virtual infrastructure can also be scheduled on the same dvuplink with LBT configuration. Subsequently, if the utilization of the uplink goes beyond the 75 percent threshold, the LBT algorithm will be invoked and some of the traffic will be moved to other underutilized dvuplinks. It is possible that management traffic will get moved to other dvuplinks when such an event occurs.
- There are no standby dvuplinks in this configuration, so the fallback setting is not applicable for this design approach. The default setting for this fallback option is “Yes.”
- VMware recommends isolating all traffic types from each other by defining a separate VLAN for each dvportgroup.
- There are several other parameters that are part of the dvportgroup configuration. Customers can choose to configure these parameters based on their environment needs.

As you follow the dvportgroups configuration in Table 7, you can see that each traffic type has all the dvuplinks as active and these uplinks are utilized through the LBT algorithm. Let's take a look at the NIOC configuration.

The NIOC configuration in this design not only helps provide the appropriate I/O resources to the different traffic types but also provides SLA guarantees by preventing one traffic type from dominating others.

Based on the bandwidth assumptions made for different traffic types, the shares parameters are configured in the NIOC shares column in Table 7. To illustrate how share values translate to bandwidth numbers in this deployment, let's take an example of a 10Gb capacity dvuplink carrying all five traffic types. This is a worst-case scenario in which all traffic types are mapped to one dvuplink. This will never happen when customers enable the LBT feature, because LBT will move the traffic type based on the uplink utilization.

The following example shows how much bandwidth each traffic type will be allowed on one dvuplink during a contention or oversubscription scenario and when LBT is not enabled:

- Total shares: management (5) + vMotion (20) + FT (10) + iSCSI (20) + virtual machine (20) = 75
- Management: 5 shares; $(5/75) * 10Gb = 667Mbps$
- vMotion: 20 shares; $(20/75) * 10Gb = 2.67Gbps$
- FT: 10 shares; $(10/75) * 10Gb = 1.33Gbps$
- iSCSI: 20 shares; $(20/75) * 10Gb = 2.67Gbps$
- Virtual machine: 20 shares; $(20/75) * 10Gb = 2.67Gbps$

For each traffic type, first the percentage of bandwidth is calculated by dividing the share value by the total available share number (75), and then the total bandwidth of the dvuplink (10Gb) is used to calculate the bandwidth share for the traffic type. For example, 20 shares allocated to vMotion traffic translate to 2.67Gbps of bandwidth to the vMotion process on a fully utilized 10GbE network adaptor.

In this 10GbE deployment, customers can provide bigger pipes to individual traffic types without the use of trunking or multipathing technologies. This was not the case with an eight-1GbE deployment.

There is no change in physical switch configuration in this design approach, so refer to the physical switch settings described in “Design Option 1” in the previous section.

TRAFFIC TYPE	PORT GROUP	TEAMING OPTION	ACTIVE UPLINK	STANDBY UPLINK	NIOC SHARES	NIOC LIMITS
MANAGEMENT	PG-A	LBT	dvuplink1, 2	None	5	–
vMOTION	PG-B	LBT	dvuplink1, 2	None	20	–
FT	PG-C	LBT	dvuplink1, 2	None	10	–
ISCSI	PG-D	LBT	dvuplink1, 2	None	20	–
VIRTUAL MACHINE	PG-E	LBT	dvuplink1, 2	None	20	–

Table 7. Dynamic Design Configuration

This design option utilizes the advanced VDS features and provides customers with a dynamic and flexible design approach. In this design, I/O resources are utilized effectively and SLAs are met based on the shares allocation.

Blade Server in Example Deployment

Blade servers are server platforms that provide higher server consolidation per rack unit as well as lower power and cooling costs. Blade chassis that host the blade servers have proprietary architectures and each vendor has its own way of managing resources in the blade chassis. It is difficult in this document to look at all of the various blade chassis available on the market and to discuss their deployments. In this section, we will focus on some generic parameters that customers should consider when deploying VDS in a blade chassis environment.

From a networking point of view, all blade chassis provide the following two options:

- Integrated switches: With this option, the blade chassis enables built-in switches to control traffic flow between the blade servers within the chassis and the external network.
- Pass-through technology: This is an alternative method of network connectivity that enables the individual blade servers to communicate directly with the external network.

In this document, the integrated switch option is described as “where the blade chassis has a built-in Ethernet switch.” This Ethernet switch acts as an access layer switch, as shown in Figure 6.

This section discusses a deployment in which the ESXi host is running on a blade server. The following two types of blade server configuration will be described in the next section:

- Blade server with two 10GbE network adaptors
- Blade server with hardware-assisted multiple logical network adaptors

For each of these two configurations, various VDS design approaches will be discussed.

Blade Server with Two 10GbE Network Adaptors

This deployment is quite similar to that of a rack server with two 10GbE network adaptors in which each ESXi host is provided with two 10GbE network adaptors. As shown in Figure 6, an ESXi host running on a blade server in the blade chassis is also provided with two 10GbE network adaptors.

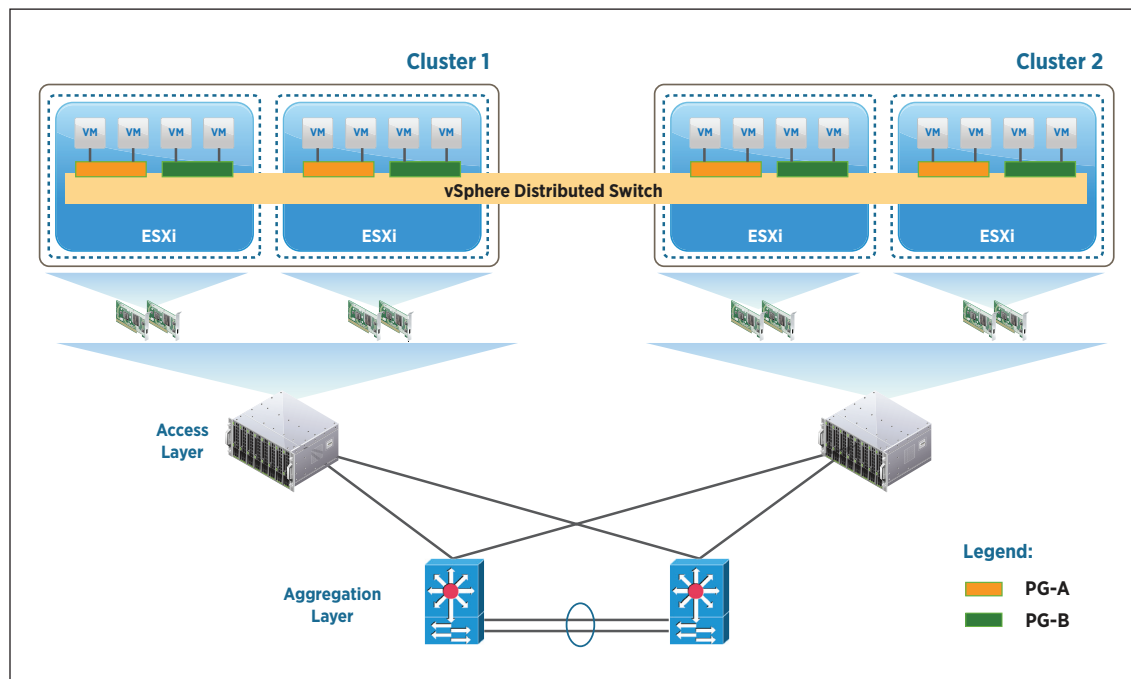


Figure 6. Blade Server with Two 10GbE Network Adaptors

In this section, two design options are described. One is a traditional static approach and the other one is a VMware-recommended dynamic configuration with NIOC and LBT features enabled. These two approaches are exactly the same as the deployment described in the “Rack Server with Two 10GbE Network Adaptors” section. Only blade chassis-specific design decisions will be discussed as part of this section. For all other VDS and switch-related configurations, refer to the “Rack Server with Two 10GbE Network Adaptors” section of this document.

Design Option 1 – Static Configuration

The configuration of this design approach is exactly the same as that described in the “Design Option 1” section under “Rack Server with Two 10GbE Network Adaptors.” Refer to Table 6 for dvportgroup configuration details. Let’s take a look at the blade server-specific parameters that require attention during the design.

Network and hardware reliability considerations should be incorporated during the blade server design as well. In these blade server designs, customers must focus on the following two areas:

- High availability of blade switches in the blade chassis
- Connectivity of blade server network adaptors to internal blade switches

High availability of blade switches can be achieved by having two Ethernet switching modules in the blade chassis. And the connectivity of two network adaptors on the blade server should be such that one network adaptor is connected to the first Ethernet switch module, and the other network adaptor is hooked to the second switch module in the blade chassis.

Another aspect that requires attention in the blade server deployment is the network bandwidth availability across the midplane of the blade chassis and between the blade switches and aggregation layer. If there is an oversubscription scenario in the deployment, customers must think about utilizing traffic shaping and prioritization (802.1p tagging) features available in the vSphere platform. The prioritization feature enables customers to tag the important traffic coming out of the vSphere platform. These high-priority-tagged packets are then treated according to priority by the external switch infrastructure. During congestion scenarios, the switch will drop lower-priority packets first and avoid dropping the important, high-priority packets.

This static design option provides customers with the flexibility to choose different network adaptors for different traffic types. However, when doing the traffic allocation on a limited, two 10GbE network adaptors, administrators ultimately will schedule multiple traffic types on a single adaptor. As multiple traffic types flow through one adaptor, the chances of one traffic type’s dominating others increases. To avoid the performance impact of the “noisy neighbors” (dominating traffic type), customers must utilize the traffic management tools provided in the vSphere platform. One of the traffic management features is NIOC, and that feature is utilized in “Design Option 2,” which is described in the following section.

Design Option 2 – Dynamic Configuration with NIOC and LBT

This dynamic configuration approach is exactly the same as that described in the “Design Option 2” section under “Rack Server with Two 10GbE Network Adaptors.” Refer to Table 7 for the dvportgroup configuration details and NIOC settings. The physical switch-related configuration in the blade chassis deployment is the same as that described in the rack server deployment. For the blade center-specific recommendation on reliability and traffic management, refer to the previous section.

VMware recommends this design option, which utilizes the advanced VDS features and provides customers with a dynamic and flexible design approach. With this design, I/O resources are utilized effectively and SLAs are met based on the shares allocation.

Blade Server with Hardware-Assisted Logical Network Adaptors (HP Flex-10– or Cisco UCS–like Deployment)

Some of the new blade chassis support traffic management capabilities that enable customers to carve I/O resources. This is achieved by providing logical network adaptors for the ESXi hosts. Instead of two 10GbE network adaptors, the ESXi host now sees multiple physical network adaptors that operate at different configurable speeds. As shown in Figure 7, each ESXi host is provided with eight Ethernet network adaptors that are carved out of two 10GbE network adaptors.

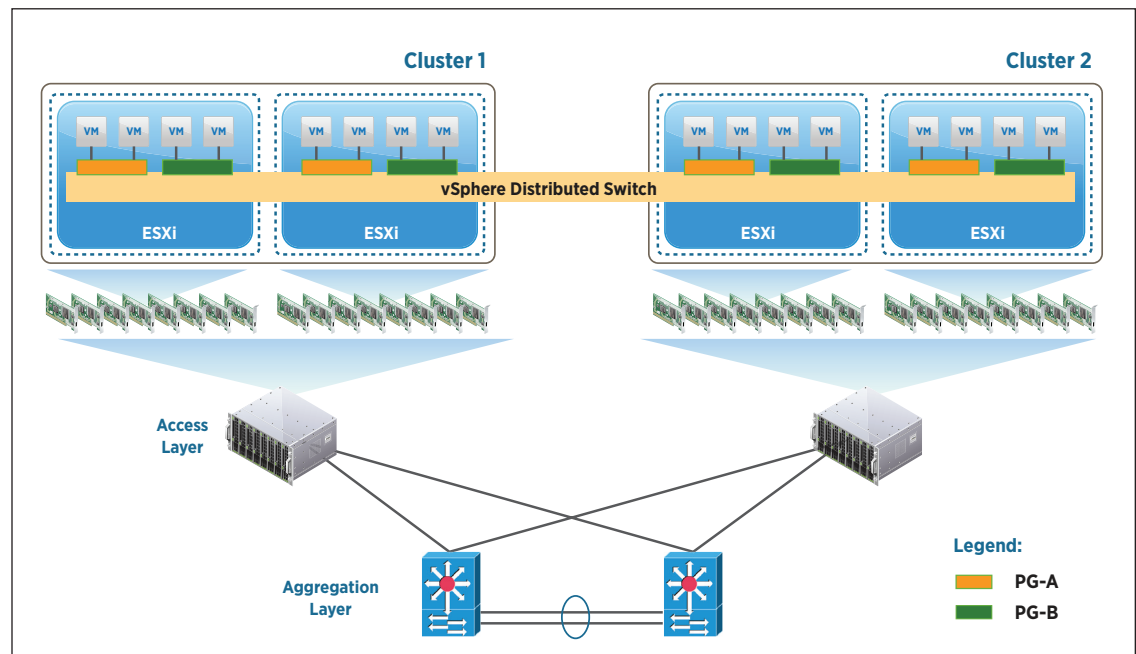


Figure 7. Multiple Logical Network Adaptors

This deployment is quite similar to that of the rack server with eight 1GbE network adaptors. However, instead of 1GbE network adaptors, the capacity of each network adaptor is configured at the blade chassis level. In the blade chassis, customers can carve out different capacity network adaptors based on the need of each traffic type. For example, if iSCSI traffic needs 2.5Gb of bandwidth, a logical network adaptor with that amount of I/O resources can be created on the blade chassis and provided for the blade server.

As for the configuration of the VDS and blade chassis switch infrastructure, the configuration described in “Design Option 1” under “Rack Server with Eight 1GbE Network Adaptors” is more relevant for this deployment. The static configuration option described in that design can be applied as is in this blade server environment. Refer to Table 2 for the dvportgroup configuration details and switch configurations described in that section for physical switch configuration details.

The question now is whether NIOC capability adds any value in this specific blade server deployment. NIOC is a traffic management feature that helps in scenarios where multiple traffic types flow through one uplink or network adaptor. If in this particular deployment only one traffic type is assigned to a specific Ethernet network adaptor, the NIOC feature will not add any value. However, if multiple traffic types are scheduled over one network adaptor, customers can make use of NIOC to assign appropriate shares to different traffic types. This NIOC configuration will ensure that bandwidth resources are allocated to traffic types and that SLAs are met.

As an example, let's consider a scenario in which vMotion and iSCSI traffic is carried over one 3Gb logical uplink. To protect the iSCSI traffic from network-intensive vMotion traffic, administrators can configure NIOC and allocate shares to each traffic type. If the two traffic types are equally important, administrators can configure shares with equal values (10 each). With this configuration, when there is a contention scenario, NIOC will make sure that the iSCSI process will get half of the 1Gb uplink bandwidth and avoid having any impact on the vMotion process.

VMware recommends that the network and server administrators work closely together when deploying the traffic management features of the VDS and blade chassis. To achieve the best end-to-end quality of service (QoS) result, a considerable amount of coordination is required during the configuration of the traffic management features.

Operational Best Practices

After a customer successfully designs the virtual network infrastructure, the next challenges are how to deploy the design and how to keep the network operational. VMware provides various tools, APIs, and procedures to help customers effectively deploy and manage their network infrastructure. The following are some key tools available in the vSphere platform:

- VMware vSphere® Command-Line Interface (vSphere CLI)
- VMware vSphere® API
- Virtual network monitoring and troubleshooting
 - NetFlow
 - Port mirroring

In the following section, we will briefly discuss how vSphere and network administrators can utilize these tools to manage their virtual network. Refer to the vSphere documentation for more details on the tools.

VMware vSphere Command-Line Interface

vSphere administrators have several ways to access vSphere components through vSphere interface options, including VMware vSphere® Client™, vSphere Web Client, and vSphere Command-Line Interface. The vSphere CLI command set enables administrators to perform configuration tasks by using a vSphere vCLI package installed on supported platforms or by using VMware vSphere® Management Assistant (vMA). Refer to the *Getting Started with vSphere CLI* document for more details on the commands: <http://www.vmware.com/support/developer/vcli>.

The entire networking configuration can be performed through vSphere vCLI, helping administrators automate the deployment process.

VMware vSphere API

The networking setup in the virtualized datacenter involves configuration of virtual and physical switches. VMware has provided APIs that enable network switch vendors to get information about the virtual infrastructure, which helps them to automate the configuration of the physical switches and the overall process. For example, vCenter can trigger an event after the vMotion process of a virtual machine is performed. After receiving this event trigger and related information, the network vendors can reconfigure the physical switch port policies such that when the virtual machine moves to another host, the VLAN/access control list (ACL) configurations are migrated along with the virtual machine. Multiple networking vendors have provided this automation between physical and virtual infrastructure configurations through integration with vSphere APIs. Customers should check with their networking vendors to learn whether such an automation tool exists that will bridge the gap between physical and virtual networking and simplify the operational challenges.

Virtual Network Monitoring and Troubleshooting

Monitoring and troubleshooting network traffic in a virtual environment require similar tools to those available in the physical switch environment. With the release of vSphere 5, VMware gives network administrators the ability to monitor and troubleshoot the virtual infrastructure through features such as NetFlow and port mirroring.

NetFlow capability on a distributed switch along with a NetFlow collector tool helps monitor application flows and measures flow performance over time. It also helps in capacity planning and ensuring that I/O resources are utilized properly by different applications, based on their needs.

Port mirroring capability on a distributed switch is a valuable tool that helps network administrators debug network issues in a virtual infrastructure. Granular control over monitoring ingress, egress or all traffic of a port helps administrators fine-tune what traffic is sent for analysis.

vCenter Server on a Virtual Machine

As mentioned earlier, vCenter Server is only used to provision and manage VDS configurations. Customers can choose to deploy it on a virtual machine or a physical host, depending on their management resource design requirements. In case of vCenter Server failure scenarios, the VDS will continue to provide network connectivity, but no VDS configuration changes can be performed.

By deploying vCenter Server on a virtual machine, customers can take advantage of vSphere platform features such as vSphere High Availability (HA) and VMware Fault Tolerance (Fault Tolerance) to provide higher resiliency to the management plane. In such deployments, customers must pay more attention to the network configurations. This is because if the networking for a virtual machine hosting vCenter Server is misconfigured, the network connectivity of vCenter Server is lost. This misconfiguration must be fixed. However, customers need vCenter Server to fix the network configuration because only vCenter Server can configure a VDS. As a work-around to this situation, customers must connect to the host directly where the vCenter Server virtual machine is running through vSphere Client. Then they must reconnect the virtual machine hosting vCenter Server to a VSS that is also connected to the management network of hosts. After the virtual machine running vCenter Server is reconnected to the network, it can manage and configure VDS.

Refer to the community article “Virtual Machine Hosting a vCenter Server Best Practices” for guidance regarding the deployment of vCenter on a virtual machine:

<http://communities.vmware.com/servlet/JiveServlet/previewBody/14089-102-1-16292/VMhostVCBestPractices.html>.

Conclusion

A VMware vSphere distributed switch provides customers with the right measure of features, capabilities and operational simplicity for deploying a virtual network infrastructure. As customers move on to build private or public clouds, VDS provides the scalability numbers for such deployments. Advanced capabilities such as NIOC and LBT are key for achieving better utilization of I/O resources and for providing better SLAs for virtualized business-critical applications and multitenant deployments. Support for standard networking visibility and monitoring features such as port mirroring and NetFlow helps administrators manage and troubleshoot a virtual infrastructure through familiar tools. VDS also is an extensible platform that enables integration with other networking vendor products through open vSphere APIs.



VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

Copyright © 2012 VMware, Inc. All rights reserved. This product is protected by U.S. and international copyright and intellectual property laws. VMware products are covered by one or more patents listed at <http://www.vmware.com/go/patents>. VMware is a registered trademark or trademark of VMware, Inc. in the United States and/or other jurisdictions. All other marks and names mentioned herein may be trademarks of their respective companies. Item No: VMW-vSPHR-DIST-SWITCH-PRCTICES-USLET-101