

The Technology Foundations of VMware® vShield

Overview

VMware® vShield is a suite of security virtual appliances built for VMware vSphere™ 4.1 (“vSphere”). It is a critical security component for protecting virtualized datacenters from attacks and misuse. vShield App and vShield Edge are the two products in the suite that address network security. The goal of this document is to provide details on the key security technologies implemented in the vShield App and vShield Edge products that enable administrators to build a multitenant virtualized datacenter environment that is flexible, agile, scalable and secure. The document first discusses the challenges in using physical security to protect virtual infrastructure and then describes in detail the key new technologies in vShield products that address those challenges.

Challenges with Physical Security

Virtualization technology has enabled IT to consolidate compute resources on fewer servers while improving the utilization as well as providing flexibility and agility in the deployment of virtual resources. According to Gartner, 50 percent of enterprise datacenter workloads might be virtualized by 2012. In this virtual infrastructure, the *virtual machine* acts as the atomic unit. It gets deployed, moved or removed with a few clicks on the management interface. This combination of more workload being virtualized and workload becoming more mobile has created a complex environment that is more difficult to secure. Administrators have tried to deploy physical security for this dynamic virtual infrastructure and have encountered the following challenges.

Visibility

The VMware vSphere virtualization platform provides a software switch that enables communication between two virtual machines on the same host. This has improved the efficiency in virtual machine-to-virtual machine communication, but the external security infrastructure, such as firewall/IPS/IDS devices, is not able to monitor the traffic that is flowing between virtual machines. This inability of the physical security infrastructure to monitor and secure traffic has caused concern among IT and security administrators.

VLAN Issues

Administrators can provide traffic isolation in mixed-trust virtual environments, in which multiple tenants or organizations share the physical infrastructure, through the use of VLANs. However, there are numerous challenges with this approach with respect to the complexity of the deployment and the ability to scale. The following are the key limitations of VLAN deployments:

- 1) They are complex to design, deploy and maintain.
- 2) The inherently limited VLAN range (4K) is not sufficient for service providers who want to support multiple customers.
- 3) Large fault domains are very difficult to troubleshoot.
- 4) The pruning of VLANs to manage a spanning tree requires a high level of networking expertise.

Port Consistency Issues

Another challenge in deploying a fully virtualized and dynamic datacenter is related to the port profiles, which must migrate along with the virtual machine. This happens whenever that virtual machine is moved from one physical host to another or from one datacenter to another.

Settings such as VLAN membership, access control lists, quality of service, and security profiles are most often controlled at the port or physical access layer of the network and must be reconfigured when the virtual machine move occurs.

Firewall Choke Point and Unscalable Architecture

When administrators deploy an external physical firewall device, they must route all virtual machine traffic from multiple servers through this firewall. At the same time, servers with multicore CPUs are becoming more powerful and can support a higher consolidation ratio of virtual machines. This in turn means that traffic from multiple, concurrently running virtual machines can easily fill a 10Gb pipe. If the centralized firewall is not able to handle this large amount of traffic, it becomes the bottleneck in the overall network infrastructure and can impact the application performance.

With this centralized architecture, it also becomes difficult to scale the capacity of the firewall unless it is replaced by a higher-end, more expensive, physical firewall box. Instead of deploying one high-end firewall device, some administrators choose to architect security using load balancers and multiple firewall devices. As the need to support additional traffic capacity grows, administrators introduce an additional firewall device and distribute some traffic through it. This approach adds complexity in terms of managing and configuring separate firewall devices and their associated security rules, along with load balancer configurations.

Firewall Rule Sprawl and Management Concerns

Administrators who deploy multiple firewall devices must manage these devices separately and configure security rules across them. Having security rules for separate devices is more cumbersome and prone to error than having centrally managed security rules.

Managing security rules with a list of IP addresses and port numbers is a tedious task. In a complex environment, it can mean thousands of rules. This approach of creating security rules is more difficult to manage and maintain in a virtual infrastructure where resources are moved on the fly.

Rigid Infrastructure and Fixed-Capacity Physical Devices

The physical security infrastructure is not flexible enough to meet the demands of the virtual infrastructure, where network and security services must move with the resources as they are relocated from one server to another. Moving security services in a physical infrastructure means reconfiguring physical devices for the security rules as well as reconfiguring network infrastructure.

Another factor to consider is the capacity (throughput) of a physical security appliance, which is limited compared to that of a network switch, which uses ASIC technology. The main reason for this is the complexity of stateful firewall and application firewall. These firewalls require filtering and packet inspection functions that can be changed on the fly to detect new network-based attacks. A custom ASIC is hard coded and does not provide the flexibility to change unless it is a programmable network processor running microcode. So some security vendors rely on x86-based platforms that run software-based firewall functionality, which is easy to update and change. These x86-based platforms are custom made with predictable and reliable hardware solutions and are well tested for reliability. This approach of custom-made hardware negatively impacts the security vendor's ability to take advantage of a CPU speed curve. Because security vendors can't keep delivering new hardware as new CPUs are released, the typical hardware refresh cycle with such vendors is four to five years. However, the server blade vendors demonstrate very quick adoption of new and more powerful x86 chipsets and issue new servers soon after the release of new chipsets.

Virtual Security

After looking at the challenges of deploying physical security in the virtualized environment, let's look at how these challenges are addressed through VMware vShield products. vShield App and vShield Edge are virtual security appliances that provide, respectively, protocol-based and perimeter firewall functionality. These virtual security appliances are preinstalled, preconfigured virtual machines with hardened operating systems that run on the same hardware on which production virtual machines run. Customers now can take advantage of the latest and fastest servers with new x86 chipsets to run their virtual security appliances. In addition, the virtual security appliances can leverage the underlying virtualization primitives of virtual machine life cycle management and provide the ability to do faster installations, upgrades and deletions.

The following are the key advantages of virtualizing security functions:

- a) The hypervisor-based security mechanism provides an enforcement point that enables visibility into virtual infrastructure traffic.
- b) It enables security and compliance policies that are change aware and seamlessly follow virtual machines, with the means to leverage dynamic mobility capabilities such as live migration, automated virtual machine load balancing and automated virtual machine restart. It also ensures that the security and compliance policies are "always on" and will follow virtual machines.

- c) Security functions can be created and deleted in the same flexible manner as virtualized compute workloads, without a need to rack, stack and manage separate physical appliances.
- d) It offers a scaled-out architecture for security functions, with a capacity that can be dialed up or down as needed, with the ability to ride x86 performance curves with the latest servers.
- e) It provides a logical segmentation and isolation capability that is simple to deploy and manage.

Let's look in detail at some of the key technologies and features behind the VMware vShield products.

vShield App: Hypervisor-Based Firewall

Hypervisor-based firewalls address the blind spot that occurs in the virtual infrastructure where traffic flowing across virtual machines on the same host is not monitored. A hypervisor-based firewall, sometimes referred to as a virtual network adaptor-level firewall, provides visibility into the virtual network traffic through its ability to tap into virtual network paths on a host. As shown in Figure 1, a virtual network adaptor-level firewall is placed between a network adaptor of a virtual machine and a virtual switch (vSwitch).

A virtual network adaptor-level firewall contains a kernel module in the VMkernel of the VMware ESX® machine. The module directly intercepts packets in the virtual network packet stream. In conjunction, a vShield App firewall that resides in a dedicated virtual machine provides more sophisticated processing functionality.

In the vShield architecture, the vShield App firewall appliance acts as the decision point where all allow or drop packet actions are based on the firewall rules. While the virtual network adaptor-level firewall determines which packets go to the vShield App firewall appliance. The decision algorithm in virtual network adaptor-level firewall can be as simple as passing every packet to the appliance or as complex as necessary for the solution. In the current vShield App security solution, all packets intercepted by the virtual network adaptor-level firewall are forwarded to the vShield App firewall appliance for firewall processing.

Distributed Architecture

As described in the previous section, the vShield App appliance, along with a virtual network adaptor-level firewall module, creates a virtual firewall device that provides security to the virtual infrastructure. The combination of hypervisor module (virtual network adaptor-level firewall) and firewall service virtual appliance is deployed on every VMware ESX host.

As shown in Figure 1, the vShield App virtual machine is deployed along with a vShield hypervisor module per host. The vShield App virtual machine is a preinstalled, preconfigured virtual machine with a hardened operating system (OS) specialized for handling firewall operations. The hypervisor module effectively places a network packet filter between the virtual network adaptor and the virtual switch.¹ In the diagram, it is referred to as a virtual network adaptor-level firewall. As explained previously, it allows the traffic coming in and out of virtual network adaptors to be efficiently inspected and, if required, directed to the vShield App virtual machine for further processing, as depicted by a red dotted line in Figure 1.

Every network packet, including those that are destined for virtual machines on the same host, is seen by the vShield App appliance. However, instead of sending all traffic from all hosts to a centralized physical firewall device, vShield deploys a distributed architecture with a separate vShield App firewall appliance placed in every host, providing enforcement points positioned near the traffic sources. This distributed approach offers the following advantages:

- 1) Unwanted traffic is immediately dropped at the vShield App appliance through firewall rules, reducing traffic on uplink interfaces.
- 2) When there is a need for additional firewall capacity, administrators simply can add CPU or memory resources to the vShield App appliance. If the cluster is resource limited, administrators can add another host to the cluster along with the vShield App appliance and the hypervisor module.

1. The virtual switch can be a vSphere Network Standard Switch, a vSphere Network Distributed Switch or a Cisco Nexus 1000K switch.

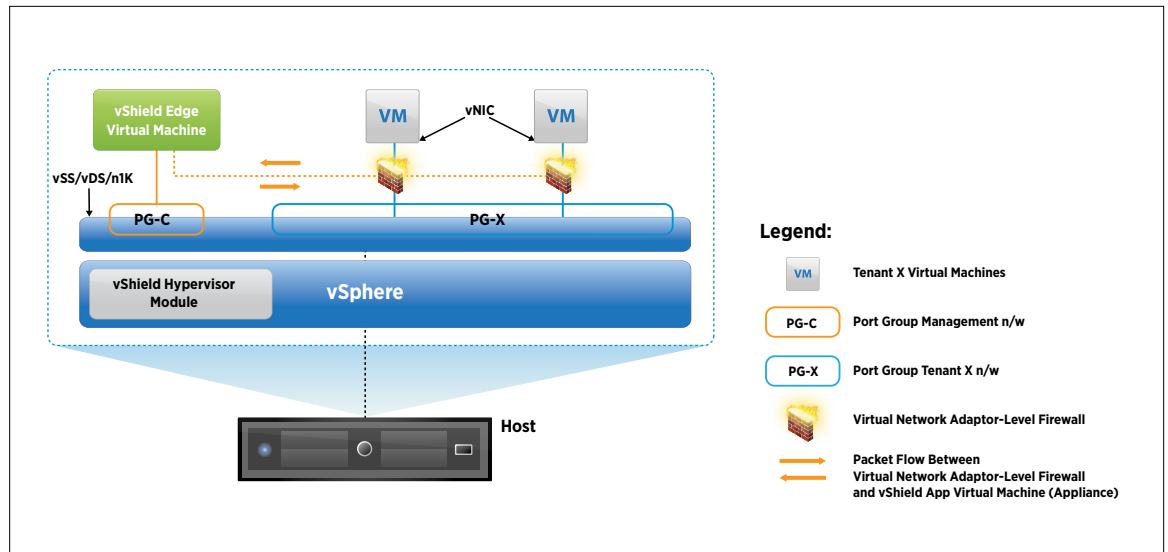


Figure 1. vShield Firewall Components

Centralized Management

The distributed, scale-out model demands a centralized policy configuration point, because it is impossible to individually manage firewall rules for every instance of vShield App appliance on every host. In the vShield environment, the vShield manager is responsible for all the management, which includes creating, deploying, upgrading and deleting vShield firewall appliances. This is different from the management of physical appliances where they cannot be programmatically deployed/created or removed/deleted. To provide a reliable communication channel that will allow the vShield manager to communicate with vShield App appliances on different hosts, it is important to create a separate management network or use the existing management network that is used for the vSphere deployment. A centralized vShield manager might need to manage hundreds of vShield App appliances, depending on the size of deployment. Figure 2 shows port group PG-C, associated with the management network and vShield App appliances of Host 1 and Host 2 connected to that port group.

The vShield Manager also centrally manages the security profiles or vShield App firewall rules and distributes them to vShield App appliances across the environment. The vShield App firewall rules are hierarchical and can be managed at the datacenter, cluster, resource pool, vApp and port group levels, to provide a consistent set of rules across multiple vShield App instances.² In this paper, we refer to these objects as containers. Often these containers can be nested. For example, a cluster container can have multiple resource pools, with several virtual machines in each resource pool.

Customers can create vShield App firewall rules based on traffic to or from a specific container. A rule that is created for a container applies to all resources in that container. For example, a rule that denies any traffic from inside of a cluster to a specific destination outside that cluster will apply to all the virtual machines in that cluster.

As membership to these containers can change dynamically, the vShield Manager maintains the state of current memberships and propagates that to the vShield App firewall appliances on each host. In this way, the vShield App firewall on each VMware ESX host detects the changes and applies the appropriate security rules to the members of various containers. For example, when a virtual machine is instantiated in a resource pool container, the security profile applied to that container is automatically enforced on that virtual machine. If a virtual machine is moved from one host to another, the security profile moves along with that virtual machine. This addresses the port-consistency issue prevalent in the physical appliance-based security environment.

² It is assumed that the reader is familiar with vSphere objects such as cluster, resource pools, and so on, and has sufficient knowledge about different methods of carving compute/memory resources. Refer to vSphere documentation for more details.

Figure 2 illustrates how container rules defined at the centralized vShield Manager are pushed down to every vShield App appliance in the environment and how these rules get translated to IP address-based firewall rules. The vShield App appliance uses these IP-based rules to make comparisons with traffic that is sent or received from different network adaptors. In the following example, the container rules defined at cluster level specify that only telnet (port 23) traffic flow from resource pool 1 (RP1) is allowed and all other types of traffic between RP1 and RP2 are denied. Similarly, traffic flow from RP1 to an outside cluster is allowed, and RP2 traffic to an outside cluster is denied.

For the security administrator, such container-level rules are very easy to define and manage. In this scenario, administrators don't need to worry about punching holes in the firewall based on IP addresses. Once these container rules are applied, they automatically get translated to the IP address-based rules that are required for packet-level comparison.

For example, Figure 2 shows a table with container rules and associated IP address rules. This mapping is also described as follows, where "RP1-any" indicates resource pool 1, with a source TCP port of any number:

- 1) RP1-any → RP2-23 **ALLOW** ≈ x.x.1.2 → x.x.1.3 **ALLOW** Telnet only (with dst. TCP port 23)
 - 2) RP1-any → External **ALLOW** ≈ x.x.1.2 → x.x.2.x **ALLOW** All
 - 3) Default → Default **DENY** ≈ x.x.1.3 → x.x.2.x **DENY** All
- x.x.1.2 → x.x.1.3 **DENY** All except Telnet

The preceding IP-based rules are applied to the vShield App appliance on Host 1 and Host 2. As previously mentioned, the vShield App appliance on a particular host receives all traffic that is sent or received from all virtual network adaptor-level firewalls on that host. The vShield App appliance then compares each packet with the IP rules and determines whether to allow or drop that packet.

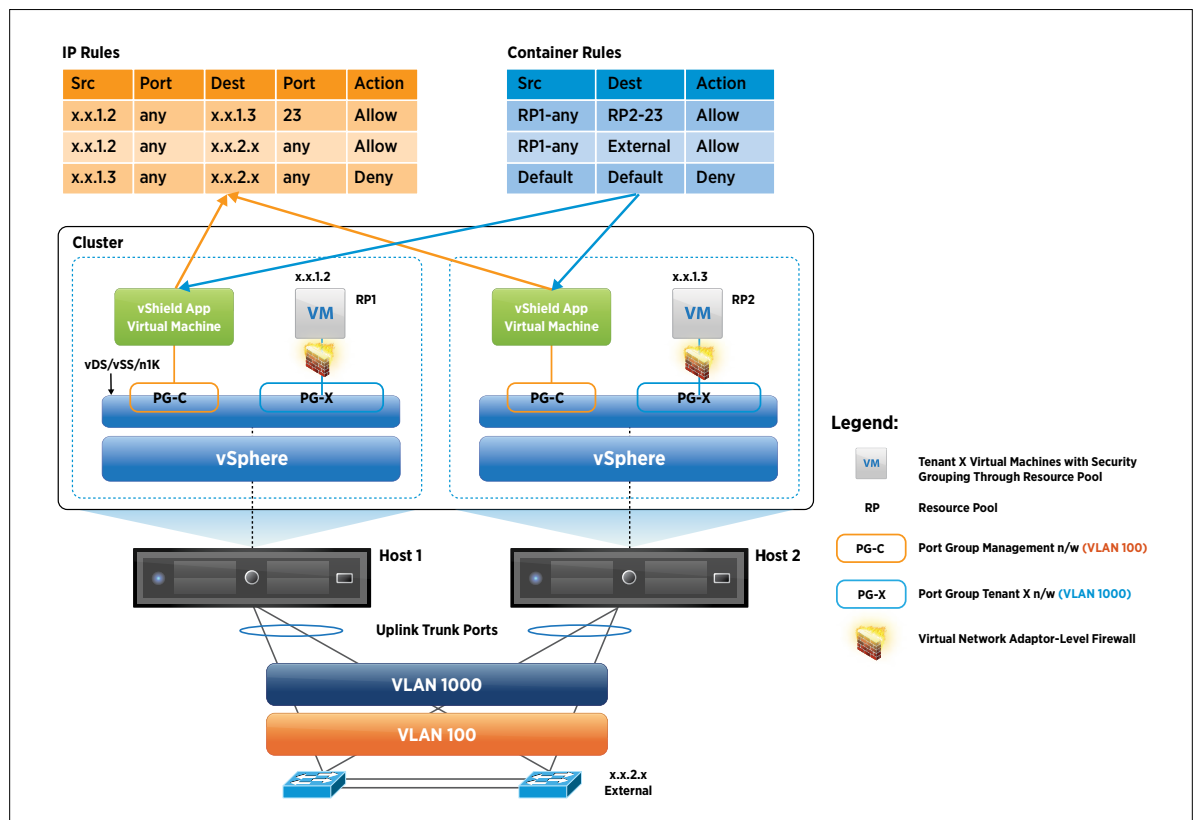


Figure 2. Secure Rule Mapping

This approach of deploying security per container is different from the traditional method of creating IP-based security rules. This new method for security in a virtualized infrastructure is simpler than the traditional one but also demands strong collaboration between security and virtual infrastructure (VI) administrators. Instead of security administrators' creating security rules based on the IP parameter groupings, the rules are now based on containers that the VI administrator creates. The rules also can be based on user-defined security groups³ described in the following section.

Custom Security Groups

In the previous section, we discussed how administrators can define security rules for the following vSphere objects, also termed as *containers*:

- Datacenter
- Cluster
- Resource pool
- vApp
- Port group
- VLAN

In this section, we will talk about another custom container known as a *security group*. A security group is a trust zone that the customer can create and assign resources to for vShield App firewall protection. Security groups also enable administrators to create a container by assigning resources such as virtual machines and virtual network adaptors. Similar to the vSphere objects, security groups can be used as a basis for firewall rules. The rules defined for a security group are applied to all virtual machines/virtual network adaptors in that security group.

This specialized approach enables customers to maintain separation of duties between VI and security administrators. For example, when VI administrators deploy a new virtual machine in the virtual infrastructure, they inform the security administrators about this action. Based on the virtual machine profile, security administrators decide in which security group that virtual machine belongs. Then the Security administrator associates the network adaptor of that virtual machine to the identified security group. All the security rules defined for that particular security group are now applied to the new deployed virtual machine.

vShield App Deployment

Let's look at a typical vShield App-based security deployment to understand different traffic flows in this environment. As shown in Figure 3, a cluster of two hosts provides compute and memory resources to a Tenant X. Host 1 and Host 2 provide resources to two resource pools, RP1 and RP2. Each resource pool has one virtual machine, which is protected with security policies defined at resource pool level. A vShield App appliance, deployed one per host, is used to protect the resource pools from each other and from the external world. For the purpose of simplifying the diagram, the management port group interface of the vShield App appliance is not shown in Figure 3. The following are additional details about the network configuration in the environment:

- 1) All virtual machines are connected to port group X (PG-X).
- 2) PG-X is configured with VLAN 1000. This means all virtual machines are in the same layer-2 domain.
- 3) A virtual network adaptor-level firewall module sits between the vSwitch and the network adaptor driver on every host. It is responsible for intercepting and forwarding traffic to the vShield App appliance.

3. Refer to vShield product documentation for more details on user-defined security groups.

Traffic Flows

The following section describes standard traffic flows in the virtual infrastructure and how the virtual network adaptor-level firewall forwards traffic to the vShield App appliance in the respective hosts for security checks. There are two types of traffic flows:

- 1) Virtual machine-to-virtual machine communication on different hosts
- 2) Virtual machine-to-virtual machine communication on the same host

Virtual machine-to-virtual machine communication on different hosts is described as follows. The two virtual machines are part of different security containers (RP1 and RP2). Only Telnet traffic is allowed between these two virtual machines. All other traffic is blocked.

Traffic Flow Within a Tenant with Different Security Group Virtual Machines

Let's see how the packet flows when "virtual machine in RP1" on Host 1 creates a Telnet session with "virtual machine in RP2" on Host 2. In Figure 3, the numbered blue circles and orange arrows provide the direction of flow.

The order of the flow is described in the following text. Each of the following numbers correlates with a blue circle with the same number in the diagram.

- 1) When a virtual machine on Host 1 sends the packet out as part of the Telnet protocol, it is intercepted by the virtual network adaptor-level firewall and is forwarded to the vShield App appliance on that host.
- 2) The vShield App appliance inspects the packet. If the security profile allows the packet to flow through, the packet is sent back to the virtual network adaptor-level firewall.
- 3) The packet is received by the virtual network adaptor-level firewall from the vShield App appliance.
- 4) The virtual network adaptor-level firewall sends the packet to vSwitch port group PG-X.
- 5) The vSwitch looks up the MAC address and accordingly sends the traffic out on the uplink port of Host 1.
- 6) The external infrastructure that involves physical switches will carry this packet on VLAN 1000.
- 7) The external switch sends the packet to the Host 2 network adaptor based on the MAC address table.
- 8) The vSwitch on Host 2 receives the packet.
- 9) The vSwitch looks up the MAC address and accordingly sends the traffic out to the virtual machine on Host 2.
- 10) The virtual network adaptor-level firewall intercepts the packet and forwards it to the vShield App appliance on Host 2.
- 11) The vShield App appliance inspects the packet. If the security profile allows the packet to flow through, the packet is sent back to the virtual network adaptor-level firewall.
- 12) The virtual network adaptor-level firewall sends the packet to the virtual machine on Host 1.

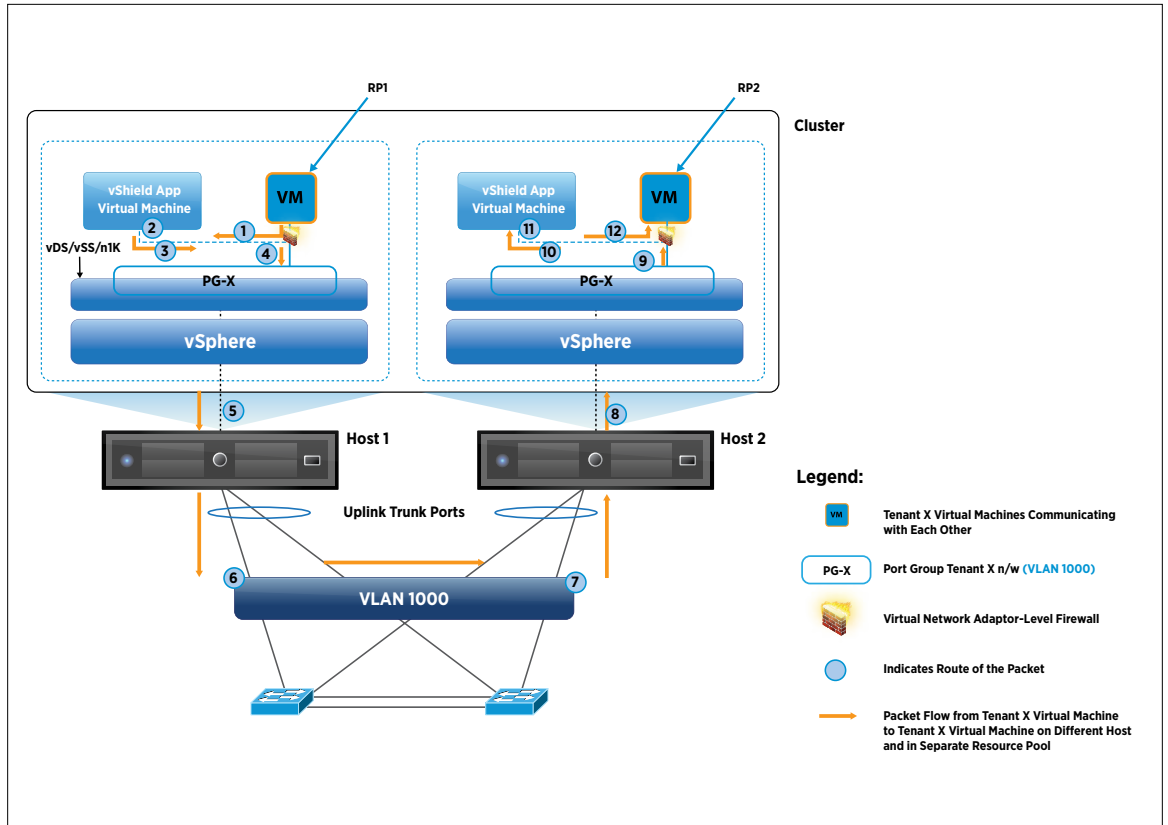


Figure 3. Same Tenant/Different Security Group Virtual Machine Communication

Performance

Performance is one of the key parameters that administrators take into account while evaluating various products. In this section, we will look into the performance of a vShield App firewall appliance. Its performance depends on the amount of CPU and memory resources allocated to the virtual machine, as well as to the following other parameters:

- Traffic rate
- Packet size
- Number of security rules

Testing on a VMware ESX host running on an Intel Xeon Processor X5560 (Nehalem) at 2.8GHz with dual quad-core CPUs and two vCPUs, and 2GB DRAM allocated to a vShield App firewall appliance, yields the following results:

- 1) Greater than 1Gbps throughput with 128 byte packets and 5K security rules
- 2) Greater than 7Gbps throughput with 16K byte packets and 5K security rules
- 3) Almost 9Gbps throughput with 128K byte packets and 5K security rules

These numbers demonstrate that the virtual appliance firewall performance is comparable to a physical appliance-based firewall using a similar x86 configuration platform. There is an added flexibility with a virtual appliance in on-demand deployment of security in a scale-out fashion. Another benefit is that a CPU is not exclusively dedicated to security processing. For a given workload that is more compute than networking intensive, the CPU on the host would be more efficiently utilized for computing than for security services.

High Availability

VMware vSphere platform features such as VMware High Availability (VMware HA) and VMware Distributed Resource Scheduler (VMware DRS) provide resiliency to the design and increase availability of the solution. The following different types of failures should be considered when using the vShield App appliance for securing the infrastructure:

- **Host failure:** If the host dies, the vShield App appliance is no longer available, nor are the virtual machines that it was protecting. However, through the vSphere platform's VMware HA feature, the virtual machines running on the failed host are restarted on other hosts in the cluster, where they will be protected by the vShield App virtual machine running on those hosts.
- **Virtual machine failure:** A heartbeat monitoring mechanism helps detect OS failures on a virtual machine. When such failures are detected, the virtual machine is automatically restarted. Security rules defined for that virtual machine do not change. It is continuously protected through virtual network adaptor-level firewall.
- **vShield App virtual machine failure:** With the failure of a vShield App appliance, all traffic to and from that host will be impacted. Until the appliance is brought back into action, all communication of the virtual machines on that host will be interrupted.
- **vShield Manager failure:** When vShield Manager goes down, the vShield App appliance will continue to provide security to the infrastructure, but no new virtual machine can be added to the security groups. In addition, the flow-monitoring data might be lost, depending on the duration of the failure.

The following are the constraints on the vShield App appliance:

- 1) **Pinning the vShield App virtual machine to the deployed host:** The vShield App firewall correlates with the security rules and data path state for virtual machines on that host. This mandates that a vShield App virtual machine should never be moved from that host.
- 2) **Configuring startup order:** To make sure that virtual machines on a particular host are protected as of the first time they boot up, it is important to power-on the vShield App appliance before any other virtual machine on that host. This is done by the vShield infrastructure when a host is restarted after an upgrade cycle or any failure.
- 3) **Restricting permissions on a vShield App appliance:** VI administrators should not be allowed to make arbitrary virtual machine operations (delete, move) on the vShield App virtual machine.

Flow Monitoring

Flow monitoring is a traffic analysis tool that provides a detailed view of the traffic on a virtual network that passes through a vShield App appliance. The flow monitoring output defines which machines are exchanging data and over which application. This data includes the number of sessions, packets and bytes transmitted per session. Session details include sources, destinations, direction of sessions, applications and ports being used. Session details can be used to create vShield App firewall allow or deny rules. Security Administrator can use flow monitoring as a forensic tool to detect rogue services and examine outbound sessions.

The following are the key advantages over traditional monitoring approaches:

- 1) It provides flow data of virtual machine-to-virtual machine traffic, which is not visible to external monitoring devices.
- 2) It has the ability to dynamically generate security rules based on observed flow.
- 3) Because all virtual network infrastructure traffic flows are logged, it also helps in debugging any network issues.

Logging and Auditing

The vShield App keeps track of system events as well as audit logs. System events are those related to vShield operation. They detail every operational event, such as vShield App reboot or a break in communication between a vShield App and the vShield Manager. Events might relate to basic operation (informational) or to a critical error (critical). In addition to system events and audit logs, the vShield App appliance has the ability to log traffic that was denied or allowed.

The vShield Manager aggregates system events into a report that can be filtered by a particular vShield App or event severity.

The audit logs provide a view into the actions performed by all vShield Manager users. The vShield Manager retains audit log data for one year, after which time the data is discarded.

Spoof Guard

Spoof Guard is a feature that helps prevent IP and MAC spoofing. This is achieved by authoritatively maintaining a list of MAC and IP addresses for every network adaptor. The list can be seeded with the MAC and IP address of a network adaptor the first time such information is seen by the virtualized management layer, and it can also be manually overridden. The list is then pushed down to the vShield App firewall appliance, which inspects every packet originating out of a network adaptor for the prescribed MAC and IP. If these do not match, the packet is simply dropped. This prevents malicious virtual machines from spoofing other MACs and IPs.

REST API Support

REST APIs enable the remote configuring of security services on the vShield Manager. A secure virtual datacenter (vDC) can be provisioned rapidly through the use of API calls over HTTPS to vShield Manager, which then forwards the commands to the appropriate vShield Edge or vShield App virtual machines for enforcement. It takes minutes to deploy virtual security appliances, as compared to days and weeks for installing and configuring physical firewall devices.

The following are the advantages of a REST API:

- 1) It Enables the retrieval of configuration details from the devices.
- 2) It Simplifies and automates the deployment of devices and the service configuration process.
- 3) It Provides an easy-to-deploy workflow with request and response operations.

vShield works seamlessly with existing enterprise IT security measures through REST APIs. Administrators can get integration of vShield capabilities into third-party security solutions, including existing antivirus and antimalware solutions.

vShield Edge: Perimeter Firewall

A perimeter firewall is a device that provides security around the edge of a network and protects the internal infrastructure from the outside world. In the vShield portfolio, vShield Edge acts as a perimeter firewall and provides network edge security and gateway services. It helps isolate the virtual machines connected to a vSS port group, vDS port group or Cisco Nexus 1000V port group from shared uplink networks. It also provides common gateway services such as DHCP, VPN, NAT and load balancing. Common deployments of vShield Edge include DMZ, VPN extranets and multitenant cloud environments in which the vShield Edge provides perimeter security for virtual datacenters.

The vShield Edge firewall supports two approaches in creating an isolated internal network:

1. Traditional VLAN-based layer-2 isolation method
2. A method based on an innovative port group isolation (PGI) technology that VMware has developed

In this document, we will focus on explaining PGI technology. For more details on the architecture surrounding the VLAN and PGI approaches, refer to the *VMware vShield Edge Reference Design Guide*.

Port Group Isolation

Administrators are facing challenges with the VLAN-based isolation method, due to the complexity of the solution as well as limitations on the number of available VLANs. PGI technology addresses these challenges through an encapsulation method. It also simplifies the deployment while removing the limitation on the number of tenants/users that a service provider can serve. PGI technology is sometimes referred to as cross-host fencing (CHF) or vCloud Director Network Isolation (vCDNI).

The vSphere platform makes use of the knowledge of virtual infrastructure and its resources while implementing this technology. When a virtual machine is deployed in the virtual infrastructure, it is assigned a MAC address and an IP address. In addition, it is connected to a specific port group of a virtual switch. So any virtual machine that gets deployed must be associated with a port group on a virtual switch. This fact is used to isolate a group of virtual machines from other virtual machines. Administrators now can use one subnet (single layer-2 domain) to place all their virtual machines and can provide isolation within a group of virtual machines by assigning them to different port groups. Traffic between virtual machines in the same port group is allowed, but traffic between virtual machines across different port groups is not allowed by a virtual switch. This port group isolation feature is supported on a distributed virtual switch (vDS), but not on a standard switch (vSS) or Cisco Nexus 1000V. In a vDS, a port group definition can span across the entire datacenter, so an isolated virtual network through port groups can also span across the complete datacenter.

To provide this isolation across the existing physical switching infrastructure, an overlay approach is employed. This overlay is achieved through MAC-in-MAC encapsulation, which is provided through the vDS infrastructure. The external physical switch infrastructure is transparent to this process and continues to make switching decisions based on the outer MAC address. The outer MAC header is also referred to as PMAC. This PMAC is not visible to virtual machines because the vDS removes this header before it forwards the data to virtual machines. Similarly, the inner MAC header, which is also referred to as VMAC, is not visible to the external physical network. The following section describes the packet format details and possible configuration changes required in the physical switch infrastructure.

Packet Format

Figure 4 describes the MAC-in-MAC header format for packets that are not fragmented because they are equal to or less than 1,500 bytes MTU.

Outer Destination MAC Address (PMAC)				
Outer Destination MAC Address		Outer Source MAC Address		
Outer Source MAC Address				
EtherType = 802.1Q [Optional]		802.1Q Tag [Optional]		
EtherType = CHF/vCDNI		Frag ID 4b	F 1b	V 2b Fence ID[2]
Fence ID[1]	Fence ID[0]	Original Ethernet Frame (VMAC) + Payload		
....				

Figure 4. Header Format

- Outer destination MAC address (48 bits)
- Outer source MAC address (48 bits)
- 802.1Q EtherType and tag (32 bits): (optional)
- CHF (cross-host fencing)/vCD Director Network Isolation EtherType (16 bits)
 - 0x88de indicates the EtherType for CHF/vCDNI

- V (version): (2 bits)
- F (fragmentation): (1 bit)
 - 0 indicates no fragmentation of the original Ethernet packet or the second fragment if FragID is non-zero.
 - 1 indicates encapsulated original Ethernet packet is fragmented and it's the first fragment.
- FragID (fragmentation ID): (5 bits)
- FenceID (24 bits)

Because the MAC-in-MAC encapsulation process adds additional (18–22) bytes of header to each packet that goes out on the physical switch infrastructure, it is important that the MTU restriction is taken into account before making the determination to transmit a packet. This is done by the vDS and associated logic. When the original Ethernet packet length plus the fence header length (18–22 bytes) exceeds the provisioned physical MTU, the vDS and associated logic will fragment the frames into two. The fragment ID is used to facilitate the defragmentation on the receiving side.

At the receiver, these fragmented packets will go through the defrag process only when the fence ID, inner destination MAC, inner source MAC and fragment ID all match up. In cases where there is incorrect defragmentation, the transport layer in the virtual machine can drop the packet based on mismatched checksums or other errors. Increasing the MTU of the underlying physical infrastructure and the hosts by 24 bytes is recommended, to improve performance and avoid the fragmentation and defragmentation operation at the vDS level.

vShield Edge and PGI

To better understand how PGI technology is adopted in the vShield Edge product, the following section explains the deployment of a multitenant virtual datacenter using the PGI isolation method.

Multitenant Deployment

As depicted in Figure 5, the deployment has two tenants, X and Y, provisioned on a two-host cluster. The virtual machines of each tenant are isolated through the vShield Edge perimeter firewall appliance. The following details elaborate on the network configuration in this environment.

- 1) The Tenant X blue virtual machine is connected to port group X (PG-X) on Host 2.
- 2) The Tenant Y green virtual machine is connected to port group Y (PG-Y) on Host 1.
- 3) Both PG-X and PG-Y are configured with VLAN 1000. The isolation between two tenant resources, despite the use of the same VLAN ID, is maintained through PGI technology.
- 4) PG-C is configured with VLAN 100. This is the shared uplink network, which corresponds with a provisioned corporate network.
- 5) The green vShield Edge virtual machine appliance on Host 1 is the perimeter firewall between the PG-Y isolated network (also called Tenant Y private network) and the PG-C corporate network. This appliance acts as the gateway for all the green virtual machines of Tenant Y. So when any Tenant Y virtual machine must communicate with the outside world or other Tenant virtual machines, the traffic will be directed to the vShield Edge appliance.
- 6) The blue vShield Edge virtual machine appliance on Host 2 is the perimeter firewall between the PG-X isolated network (also called Tenant X private network) and the PG-C corporate network. This appliance acts as the gateway for all the blue virtual machines of Tenant X. So when any Tenant X virtual machine must communicate with the outside world or other tenant virtual machines, the traffic will be directed to the vShield Edge appliance.

Traffic Flow

The following section describes standard traffic flows in a multitenant virtual infrastructure deployment and how the vShield Edge firewall provides isolation between these tenants. The following are three types of traffic flows:

- 1) Tenant X virtual machine and Tenant Y virtual machine on the same host
- 2) Tenant X virtual machine and Tenant Y virtual machine on different hosts
- 3) Virtual machines from the same tenant on different hosts

The following describes the “Tenant X virtual machine and Tenant Y virtual machine on different hosts” scenario. The two virtual machines are part of different tenants, and only Telnet traffic is allowed between these virtual machines. All other traffic is blocked. The vShield Edge firewall is configured to open only the Telnet port, and all other ports are blocked.

Traffic Flow Between Different Host and Different Port Group

The following section describes how a packet flows when a “green virtual machine with a red border” on Host 1 creates a Telnet session with a “blue virtual machine with a red border” on Host 2. In Figure 5, the numbered green circles and orange arrows provide the direction of flow.

The order of flow is described in the following text. Each of the following numbers correlates with a green circle with the same number in the diagram.

- 1) When a green virtual machine on Host 1 sends the packet out as part of the Telnet protocol, the packet is sent to the vDS on port group PG-Y. Because this communication belongs to a different port group’s virtual machine (blue virtual machine on PG-X), the packet is forwarded to the gateway device, the green vShield Edge appliance on Host 1. There is no encapsulation here, as the communication remains in the same host.
- 2) A green vShield Edge appliance on Host 1 receives the packet and checks it against the configured security rules.
- 3) A green vShield Edge appliance sends the packet out to the vDS on port group PG-C, because Telnet port traffic is allowed between two tenants. The packet is sent to the next hop, the gateway blue vShield Edge on Host 2. The green vShield Edge on Host 1 detects all other gateway (FIREWALL/NAT) devices in the network. It also detects the internal private network IP address range. The following are the virtual machine MAC addresses on the packet that comes out of the green vShield Edge appliance:
 - a. Source virtual machine MAC: green vShield Edge virtual machine MAC.
 - b. Destination virtual machine MAC: blue vShield Edge virtual machine MAC external.
- 4) The vDS receives the packet and checks whether the destination MAC is within the same host or a different host. It utilizes the knowledge of the vSphere deployment inventory to determine the MAC, IP and virtual machine associations, and whether or not the packet is to be encapsulated. This determination on encapsulation is based on whether or not the destination MAC is on a different host. In this case, the blue vShield Edge belongs to a different host, so the packet is to be encapsulated. A lookup is performed to get the host MAC (also called PMAC) address where the destination virtual machine MAC (blue vShield Edge virtual machine) is deployed. The following are the new MAC header parameters:
 - a. Source host MAC: Host 1 MAC address (PMAC 1).
 - b. Destination host MAC: Host 2 MAC address (PMAC 2).
 - c. Fence ID: This is the unique ID per internal network or fence. Assume it is 1 for PG-C, 2 for PG-X, and 3 for PG-Y. So in this case, the fence ID will be 1 because the packet was sent on to vDS on PG-C.

- 5) The external infrastructure that involves physical switches will carry this packet on VLAN 100.
- 6) The external switch sends the packet to Host 2 network adaptor, based on the MAC address table.
- 7) The vDS looks up the fence ID (1). Because the packet is on the same PG-C, it removes the encapsulation and forwards the packet to the blue vShield Edge virtual machine. The following are the MAC header parameters:
 - a. Source virtual machine MAC: green vShield Edge virtual machine MAC.
 - b. Destination virtual machine MAC: blue vShield Edge virtual machine MAC external.
- 8) The blue vShield Edge firewall forwards the packet after the rule checking to vDS on port group PG-X. The following are the MAC header parameters:
 - a. Source virtual machine MAC: blue vShield Edge virtual machine MAC internal.
 - b. Destination virtual machine MAC: blue virtual machine with red border MAC.
- 9) The vDS looks up the destination MAC and detects that it is on the same host and same port group, PG-X. So the packet is sent to the blue virtual machine without encapsulation.

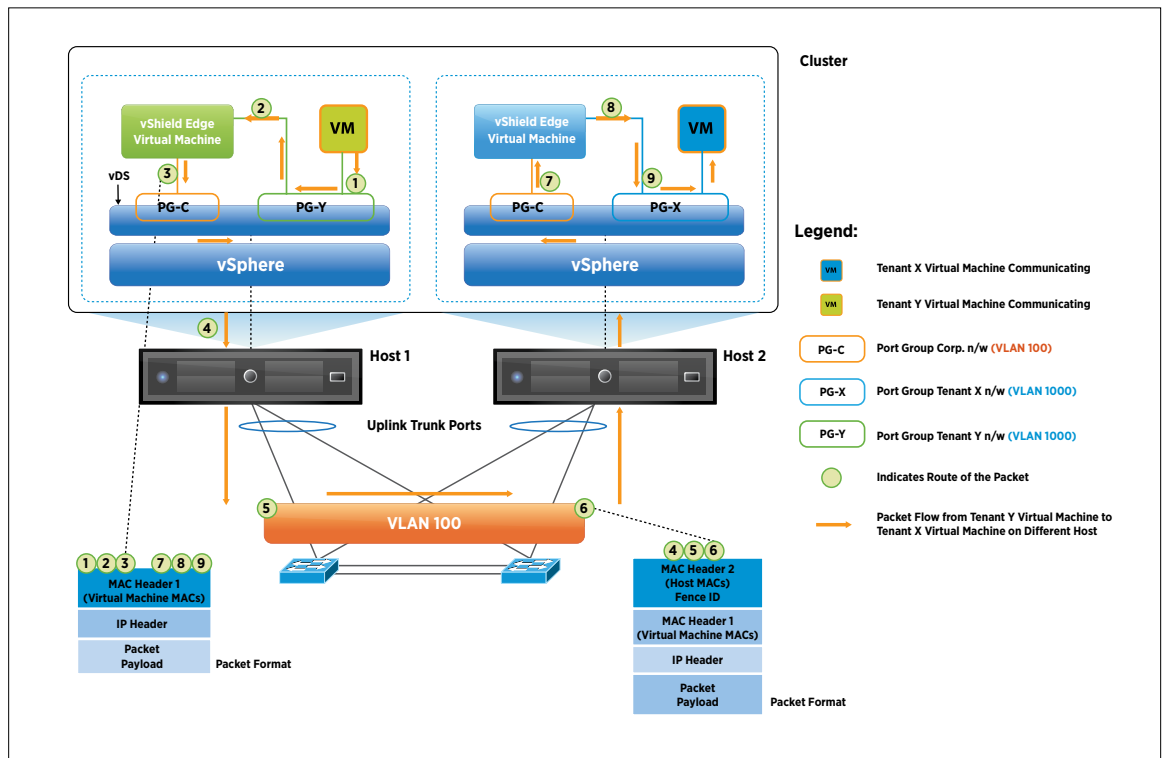


Figure 5. Different Host/Different Tenant Virtual Machine Communication

Performance

The performance of the security device is one of the key parameters that administrators take into account when evaluating different products. In this section, we will look into the performance of a vShield Edge firewall appliance. The performance numbers of vShield Edge appliances are reported for the following two functions:

- Firewall and NAT
- VPN (IPsec)

Testing on a VMware ESX host running on an Intel Xeon Processor X5560 (Nehalem) at 2.8GHz with dual quad-core CPUs and one vCPU, and 256MB DRAM allocated to a vShield Edge firewall appliance, yields the following results:

- Firewall and NAT
 - 1) UDP: 4Gbps throughput with single UDP session and 2000 security rules
 - 2) TCP: 3Gbps throughput with four sessions
 - 3) Maximum number of concurrent sessions supported: 64K
 - 4) Maximum new session rate under no background load: 9K sessions/sec
- VPN (This is purely a CPU-bound activity)
 - 1) 220Mbps with AES encryption
 - 2) 112Mbps with 3DES encryption

High Availability

VMware vSphere platform features such as VMware High Availability (VMware HA) and VMware Distributed Resource Scheduler (VMware DRS) provide resiliency to the design and increase availability of the solution. The following different types of failures should be considered when using the vShield Edge appliance for securing multitenant infrastructures.

- Host failure: Because the vShield Edge appliance is not tied to a host, the host failure scenario is different from that for the vShield App appliance. In a deployment where a vShield Edge appliance exists on a host, and if that host dies, virtual machines running on the failed host as well as the vShield Edge appliance are restarted on another host in the cluster, through the vSphere platform's VMware HA feature.
- Virtual machine failure: A heartbeat monitoring mechanism helps detect OS failures on a virtual machine. When such failures are detected, the virtual machine is automatically restarted. The virtual machine is continuously protected through the vShield Edge appliance after the restart.
- vShield Edge virtual machine failure: With the failure of the vShield Edge appliance, traffic flowing within the internal port group will not be impacted, but following traffic flows will be impacted.
 - Traffic from one port group virtual machine to another port group virtual machine
 - Traffic from virtual machine to external world (corporate network port group)

Conclusion

The vShield suite of appliances takes advantage of knowledge of the vSphere virtual infrastructure (platform) and seamlessly embeds security through unique technologies such as virtual network adaptor-level firewall and port group isolation. These new technologies not only address current challenges with physical security but also provide a flexible, scalable and simple security architecture for the dynamic cloud infrastructure, making the vision of a secure private and public cloud a reality.

References

1. *vSphere Web Access Administrator's Guide*
2. *vSphere Virtual Machine Administration Guide*
3. *vShield Administration Guide*
4. *VMware vShield Edge and vShield App Reference Design Guide*

