CHAPTER

5

Cloud Computing Security Challenges

In these days, a man who says a thing cannot be done is quite apt to be interrupted by some idiot doing it.

—Elbert Green Hubbard (1865–1915) U. S. author, editor, printer

The introduction of cloud services presents many challenges to an organization. When an organization migrates to consuming cloud services, and especially public cloud services, much of the computing system infrastructure will now be under the control of a third-party Cloud Services Provider (CSP).

Many of these challenges can and should be addressed through management initiatives. These management initiatives will require clearly delineating the ownership and responsibility roles of both the CSP (which may or may not be the organization itself) and the organization functioning in the role as customer.

Security managers must be able to determine what detective and preventative controls exist to clearly define the security posture of the organization. Although proper security controls must be implemented based on asset, threat, and vulnerability risk assessment matrices, and are contingent upon the level of data protection needed, some general management processes will be required regardless of the nature of the organization's business. These include the following:

- Security policy implementation
- Computer intrusion detection and response
- Virtualization security management

Let's look at each of these management initiatives.

Security Policy Implementation

Security policies are the foundation of a sound security implementation. Often organizations will implement technical security solutions without first creating this foundation of policies, standards, guidelines, and procedures, unintentionally creating unfocused and ineffective security controls.

A *policy* is one of those terms that can mean several things. For example, there are security policies on firewalls, which refer to the access control and routing list information. Standards, procedures, and guidelines are also referred to as policies in the larger sense of a global information security policy.

A good, well-written policy is more than an exercise created on white paper — it is an essential and fundamental element of sound security practice. A policy, for example, can literally be a lifesaver during a disaster, or it might be a requirement of a governmental or regulatory function. A policy can also provide protection from liability due to an employee's actions, or it can control access to trade secrets.

Figure 5-1 shows how the policies relate to each other hierarchically.

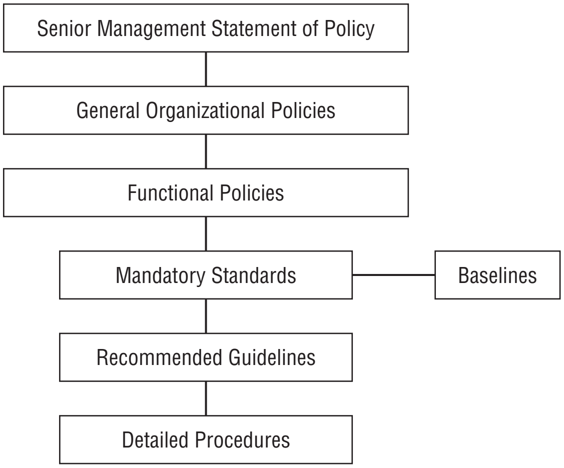

Figure 5-1: Security policy hierarchy

Policy Types

In the corporate world, when we refer to specific polices, rather than a group policy, we generally mean those policies that are distinct from the standards, procedures, and guidelines. Policies are considered the first and highest level of documentation, from which the lower-level elements of standards, procedures, and guidelines flow.

This is not to say, however, that higher-level policies are more important than the lower elements. These higher-level policies, which reflect the more general policies and statements, should be created first in the process, for strategic reasons, and then the more tactical elements can follow.

Management should ensure the high visibility of a formal security policy. This is because nearly all employees at all levels will in some way be affected, major organizational resources will be addressed, and many new terms, procedures, and activities will be introduced.

Including security as a regular topic at staff meetings at all levels of the organization can be helpful. In addition, providing visibility through such avenues as management presentations, panel discussions, guest speakers, question/answer forums, and newsletters can be beneficial.

Senior Management Statement of Policy

The first policy of any policy creation process is the *senior management statement of policy*. This is a general, high-level policy that acknowledges the importance of the computing resources to the business model; states support for information security throughout the enterprise; and commits to authorizing and managing the definition of the lower-level standards, procedures, and guidelines.

Regulatory Policies

Regulatory policies are security policies that an organization must implement due to compliance, regulation, or other legal requirements. These companies might be financial institutions, public utilities, or some other type of organization that operates in the public interest. Such policies are usually very detailed and specific to the industry in which the organization operates.

Advisory Policies

Advisory policies are security policies that are not mandated but strongly suggested, perhaps with serious consequences defined for failure to follow them (such as termination, a job action warning, and so forth). A company with such policies wants most employees to consider these policies mandatory. Most policies fall under this broad category.

Informative Policies

Informative policies are policies that exist simply to inform the reader. There are not implied or specified requirements, and the audience for this information could be certain internal (within the organization) or external parties. This does not mean that the policies are authorized for public consumption but that they are general enough to be distributed to external parties (vendors accessing an extranet, for example) without a loss of confidentiality.

Computer Security Incident Response Team (CSIRT)

As you read in Chapter 7, as part of a structured incident-handling program of intrusion detection and response, a Computer Emergency Response Team (CERT) or computer security incident response team (CSIRT) is commonly created. The main tasks of a CSIRT are as follows:

- Analysis of an event notification
- Response to an incident if the analysis warrants it
- Escalation path procedures
- Resolution, post-incident follow-up, and reporting to the appropriate parties

The prime directive of every CIRT is incident response management, which reflects a company's response to events that pose a risk to its computing environment. This management often consists of the following:

- Coordinating the notification and distribution of information pertaining to the incident to the appropriate parties (those with a need to know) through a predefined escalation path
- Mitigating risk to the enterprise by minimizing the disruptions to normal business activities and the costs associated with remediating the incident (including public relations)
- Assembling teams of technical personnel to investigate the potential vulnerabilities and resolve specific intrusions

Additional examples of CIRT activities are:

- Management of the network logs, including collection, retention, review, and analysis of data
- Management of an incident's resolution, management of a vulnerability's remediation, and post-event reporting to the appropriate parties

Response includes notifying the appropriate parties to take action in order to determine the extent of an incident's severity and to remediate the incident's effects. According to NIST, an organization should address computer security incidents by developing an incident-handling capability. The incident-handling capability should be used to do the following:

- Provide the ability to respond quickly and effectively
- Contain and repair the damage from incidents. When left unchecked, malicious software can significantly harm an organization's computing

resources, depending on the technology and its connectivity. Containing the incident should include an assessment of whether the incident is part of a targeted attack on the organization or an isolated incident.

- Prevent future damage. An incident-handling capability should assist an organization in preventing (or at least minimizing) damage from future incidents. Incidents can be studied internally to gain a better understanding of the organization's threats and vulnerabilities.

Virtualization Security Management

Although the global adoption of virtualization is a relatively recent event, threats to the virtualized infrastructure are evolving just as quickly. Historically, the development and implementation of new technology has preceded the full understanding of its inherent security risks, and virtualized systems are no different. The following sections examine the threats and vulnerabilities inherent in virtualized systems and look at some common management solutions to those threats.

VIRTUALIZATION TYPES

The Virtual Machine (VM), Virtual Memory Manager (VMM), and hypervisor or host OS are the minimum set of components needed in a virtual environment. They comprise virtual environments in a few distinct ways:

- **Type 1 virtual environments are considered "full virtualization" environments and have VMs running on a hypervisor that interacts with the hardware (see Figure 5-2).**
- **Type 2 virtual environments are also considered "full virtualization" but work with a host OS instead of a hypervisor (see Figure 5-3).**
- **Para-virtualized environments offer performance gains by eliminating some of the emulation that occurs in full virtualization environments.**
- **Other type designations include hybrid virtual machines (HVMs) and hardware-assisted techniques.**

These classifications are somewhat ambiguous in the IT community at large. The most important thing to remember from a security perspective is that there is a more significant impact when a host OS with user applications and interfaces is running outside of a VM at a level lower than the other VMs (i.e., a Type 2 architecture). Because of its architecture, the Type 2 environment increases the potential risk of attacks against the host OS. For example, a laptop running VMware with a Linux VM on a Windows XP system inherits the attack surface of both OSs, plus the virtualization code (VMM).[1]

VIRTUALIZATION MANAGEMENT ROLES

Typically, the VMware Infrastructure is managed by several users performing different roles. The roles assumed by administrators are the Virtualization Server Administrator, Virtual Machine Administrator, and Guest Administrator. VMware Infrastructure users may have different roles and responsibilities, but some functional overlap may occur. The roles assumed by administrators are configured in VMS and are defined to provide role responsibilities:

- **Virtual Server Administrator — This role is responsible for installing and configuring the ESX Server hardware, storage, physical and virtual networks, service console, and management applications.**
- **Virtual Machine Administrator — This role is responsible for creating and configuring virtual machines, virtual networks, virtual machine resources, and security policies. The Virtual Machine Administrator creates, maintains, and provisions virtual machines.**
- **Guest Administrator — This role is responsible for managing a guest virtual machine or machines. Tasks typically performed by Guest Administrators include connecting virtual devices, adding system updates, and managing applications that may reside on the operating system.**

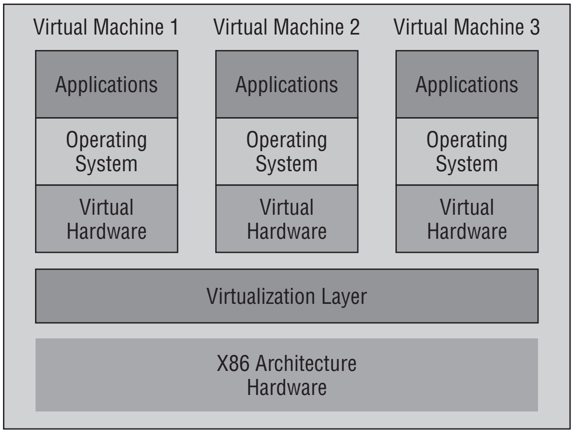

Figure 5-2: Type 1 virtualized environment

Virtual Threats

Some threats to virtualized systems are general in nature, as they are inherent threats to all computerized systems (such as denial-of-service, or DoS, attacks). Other threats and vulnerabilities, however, are unique to virtual machines. Many VM vulnerabilities stem from the fact that a vulnerability in one VM system can be exploited to attack other VM systems or the host systems, as multiple virtual machines share the same physical hardware, as shown in Figure 5-4.

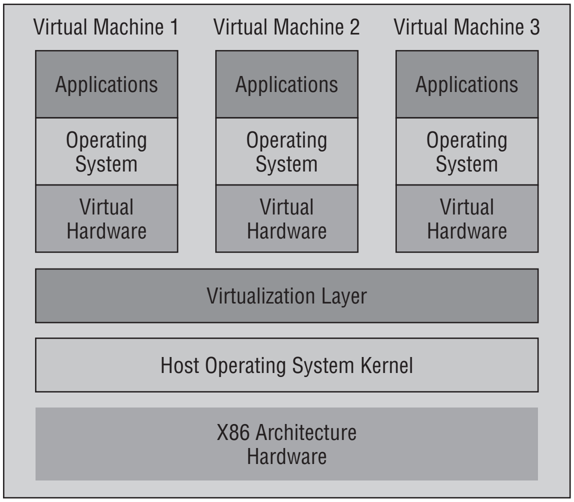

Figure 5-3: Type 2 virtualized environment

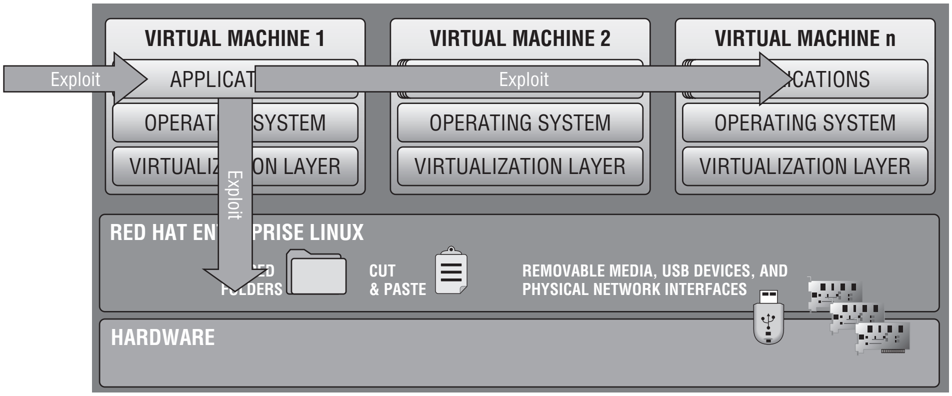

Figure 5-4: Basic VM system vulnerability

Various organizations are currently conducting security analysis and proof-of-concept (PoC) attacks against virtualized systems, and recently published research regarding security in virtual environments highlights some of the vulnerabilities exposed to any malicious-minded individuals:

- **Shared clipboard** — Shared clipboard technology allows data to be transferred between VMs and the host, providing a means of moving data between malicious programs in VMs of different security realms.

- **Keystroke logging** — Some VM technologies enable the logging of keystrokes and screen updates to be passed across virtual terminals in the virtual machine, writing to host files and permitting the monitoring of encrypted terminal connections inside the VM.
- **VM monitoring from the host** — Because all network packets coming from or going to a VM pass through the host, the host may be able to affect the VM by the following:
  - Starting, stopping, pausing, and restart VMs
  - Monitoring and configuring resources available to the VMs, including CPU, memory, disk, and network usage of VMs
  - Adjusting the number of CPUs, amount of memory, amount and number of virtual disks, and number of virtual network interfaces available to a VM
  - Monitoring the applications running inside the VM
  - Viewing, copying, and modifying data stored on the VM's virtual disks
- **Virtual machine monitoring from another VM** — Usually, VMs should not be able to directly access one another's virtual disks on the host. However, if the VM platform uses a virtual hub or switch to connect the VMs to the host, then intruders may be able to use a hacker technique known as "ARP poisoning" to redirect packets going to or from the other VM for sniffing.
- **Virtual machine backdoors** — A backdoor, covert communications channel between the guest and host could allow intruders to perform potentially dangerous operations.[2]

Table 5-1 shows how VMware's ESX server vulnerabilities can be categorized, as interpreted by the DoD (see also Figure 5-5).

According to the Burton Group five immutable laws of virtualization security must be understood and used to drive security decisions:

Law 1: All existing OS-level attacks work in the exact same way.

Law 2: The hypervisor attack surface is additive to a system's risk profile.

Law 3: Separating functionality and/or content into VMs will reduce risk.

Law 4: Aggregating functions and resources onto a physical platform will increase risk.

Law 5: A system containing a "trusted" VM on an "untrusted" host has a higher risk level than a system containing a "trusted" host with an "untrusted" VM.[3]

The current major virtualization vendors are VMware, Microsoft Hyper-V, and Citrix Systems XenServer (based on the Xen open-source hypervisor).

Table 5-1: ESX Server Application Vulnerability Severity Code Definitions

CATEGORY	ESX SERVER APPLICATION
Category I — Vulnerabilities that allow an attacker immediate access into a machine, allow super-user access, or bypass a firewall	Vulnerabilities that may result in malicious attacks on virtual infrastructure resources or services. Attacks may include, but are not limited to, malware at the VMM, virtual machine–based rootkit (SubVirt), Trojan, DOS, and executing potentially malicious actions.
Category II — Vulnerabilities that provide information that have a high potential of giving access to an intruder	Vulnerabilities that may result in unauthorized users accessing and modifying virtual infrastructure resources or services.
Category III — Vulnerabilities that provide information that potentially could lead to compromise	Vulnerabilities that may result in unauthorized users viewing or possibly accessing virtual infrastructure resources or services.

Source: ESX Server V1R1 DISA Field Security Operations, 28 April 2008, Developed by DISA for the DoD.

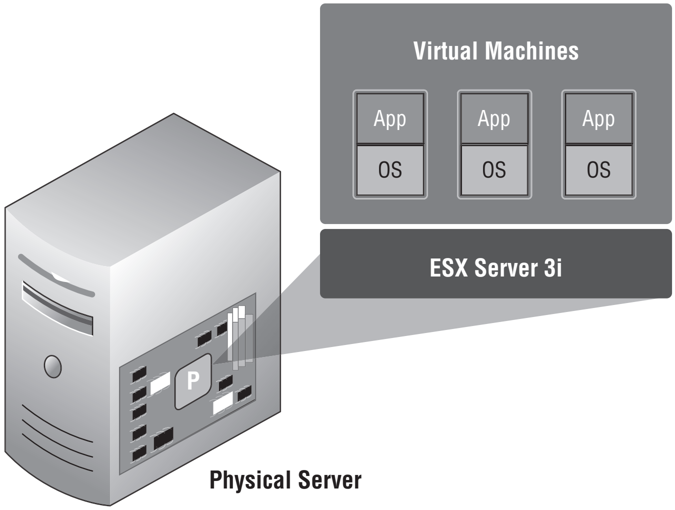

Figure 5-5: VMware ESX Server 3i

Figure 5-6 shows VMware's approach to virtualized infrastructure, and Figure 5-7 shows a little more detail into VMware's ESX server architecture.

VM THREAT LEVELS

When categorizing the threat posed to virtualized environments, often the vulnerability/threat matrix is classified into three levels of compromise:

- **Abnormally terminated — Availability to the virtual machine is compromised, as the VM is placed into an infinite loop that prevents the VM administrator from accessing the VM's monitor.**

(continued)

VM THREAT LEVELS *(continued)*

- **Partially compromised — The virtual machine allows a hostile process to interfere with the virtualization manager, contaminating stet checkpoints or over-allocating resources.**
- **Totally compromised — The virtual machine is completely overtaken and directed to execute unauthorized commands on its host with elevated privileges.[4]**

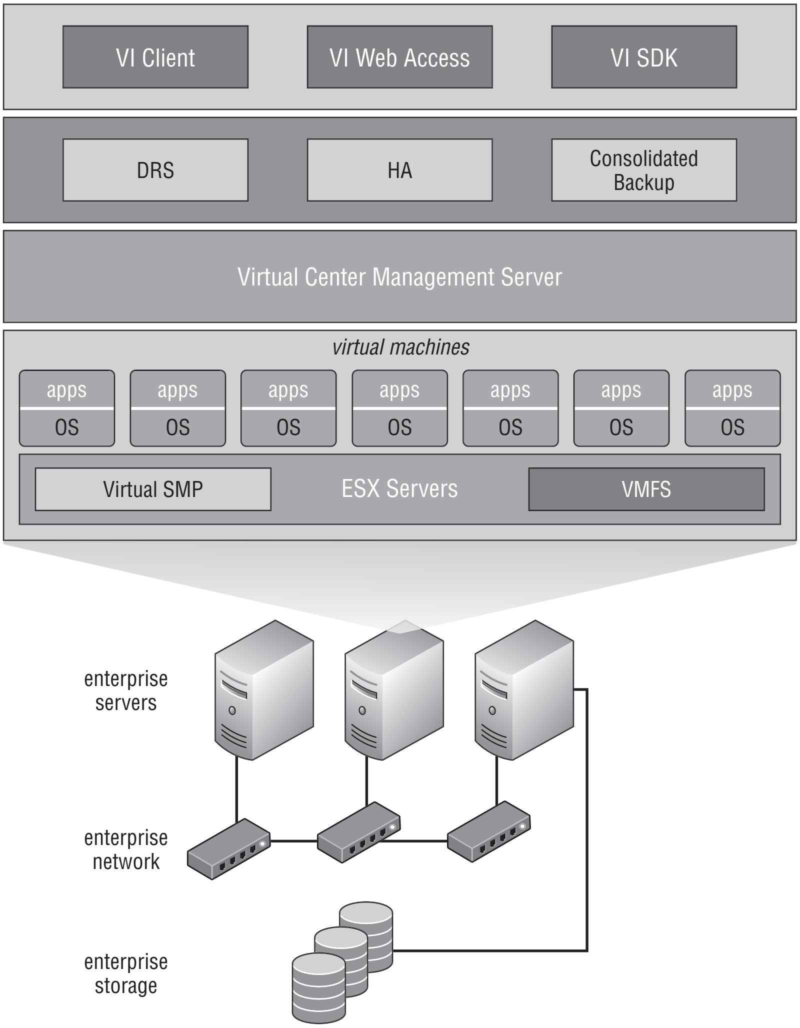

Figure 5-6: VMwARE Infrastructure

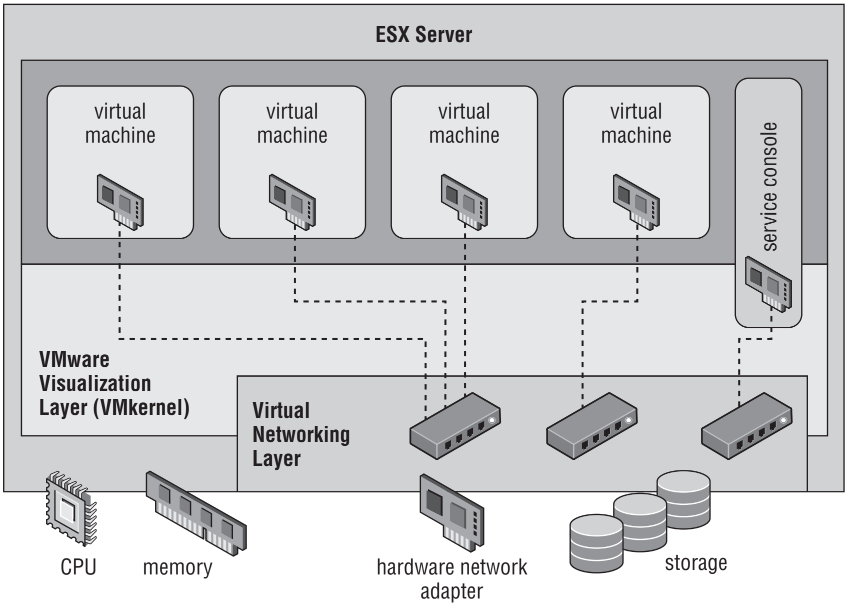

Figure 5-7: ESX server architecture

Hypervisor Risks

The *hypervisor* is the part of a virtual machine that allows host resource sharing and enables VM/host isolation. Therefore, the ability of the hypervisor to provide the necessary isolation during intentional attack greatly determines how well the virtual machine can survive risk.

One reason why the hypervisor is susceptible to risk is because it's a software program; risk increases as the volume and complexity of application code increases. Ideally, software code operating within a defined VM would not be able to communicate or affect code running either on the physical host itself or within a different VM; but several issues, such as bugs in the software, or limitations to the virtualization implementation, may put this isolation at risk. Major vulnerabilities inherent in the hypervisor consist of rogue hypervisor rootkits, external modification to the hypervisor, and VM escape.

Rogue Hypervisors

As you've seen in previous chapters, in a normal virtualization scenario, the guest operating system (the operating system that is booted inside of a virtualized environment) runs like a traditional OS managing I/O to hardware and network traffic, even though it's controlled by the hypervisor. The hypervisor, therefore, has a great level of control over the system, not only in the VM but also on the host machine.

Rootkits that target virtualization, and in particular the hypervisor, have been gaining traction in the hacker community. VM-based rootkits can hide from normal malware detection systems by initiating a "rogue" hypervisor and creating a cover channel to dump unauthorized code into the system.

Proof-of-concept (PoC) exploits have demonstrated that a hypervisor rootkit can insert itself into RAM, downgrade the host OS to a VM, and make itself invisible. A properly designed rootkit could then stay "undetectable" to the host OS, resisting attempts by malware detectors to discover and remove it.[5]

This creates a serious vulnerability in all virtualized systems. Detectability of malware code lies at the heart of intrusion detection and correction, as security researchers analyze code samples by running the code and viewing the result.

In addition, some malware tries to avoid detection by anti-virus processes by attempting to identify whether the system it has infected is traditional or virtual. If found to be a VM, it remains inactivated and hidden until it can penetrate the physical host and execute its payload through a traditional attack vector.

External Modification of the Hypervisor

In additional to the execution of the rootkit payload, a poorly protected or designed hypervisor can also create an attack vector. Therefore, a self-protected virtual machine may allow direct modification of its hypervisor by an external intruder. This can occur in virtualized systems that don't validate the hypervisor as a regular process.

VM Escape

Due to the host machine's fundamentally privileged position in relationship to the VM, an improperly configured VM could allow code to completely bypass the virtual environment, and obtain full root or kernel access to the physical host. This would result in a complete failure of the security mechanisms of the system, and is called *VM escape*.

Virtual machine escape refers to the attacker's ability to execute arbitrary code on the VM's physical host, by "escaping" the hypervisor. Sometimes called the "Holy Grail" of virtualization security research, VM escape has been the subject of a series of PoCs conducted by security researchers such as Tavis Ormandy of Google, and Tom Liston and Ed Skoudis at Intelguardians Network Intelligence.

Liston and Ormandy showed that VM escapes could occur through virtual machine shared resources called VMchat, VMftp, VMcat, and VMdrag-n-sploit.[6]

Increased Denial of Service Risk

The threat of denial-of-service (DoS) attacks against a virtualized system is as prevalent as it is against nonvirtualized systems; but because the virtual machines share the host's resources, such as memory, processor, disk, I/O devices, and so on, a denial-of-service attack risk against another VM, the host, or an external service is actually greatly increased.

Because the VM has more complex layers of resource allocation than a traditional system, DoS prevention techniques can become equally more complex. Like IT protections traditionally implemented against denial-of-service attacks, limiting the consumption of host resources through resource allocation may help lessen the exposure to DoS attacks.

VM Security Recommendations

As we've just described a host of security issues inherent in virtualized computing, let's examine some ways to protect the virtual machine. First we'll look at standard best practice security techniques that apply to traditional computer systems, and then we'll examine security techniques that are unique to virtualized systems.

Best Practice Security Techniques

The following security implementation techniques are required for most computer systems, and are still best practices for virtualized systems. These areas include physical security, patching, and remote management techniques.

Hardening the Host Operating System

Vulnerabilities inherent in the operating system of the host computer can flow upward into the virtual machine operating system. While a compromise on the VM OS would hopefully only compromise the guest domain, a compromise of the underlying host OS would give an intruder access to all services on all virtual machines hosted by the machine.

Therefore, best practice hardening techniques must be implemented to maintain the security posture of the underlying technology. Some of these techniques include the following:

- Use strong passwords, such as lengthy, hard to guess passwords with letters, numbers, and symbol combinations, and change them often.
- Disable unneeded services or programs, especially networked services.
- Require full authentication for access control.
- The host should be individually firewalled.
- Patch and update the host regularly, after testing on a nonproduction unit.

Use vendor-supplied best practice configuration guides for both the guest and host domains, and refer to some of the published standards in this area, such as the following:

- NIST Computer Resource Center (`http://csrc.nist.gov`)
- Defense Information Systems Agency (DISA) Security Technical Implementation Guides (STIGS) (`http://iase.disa.mil/stigs/index.html`)
- Center for Internet Security (`http://cisecurity.org`)
- SANS Institute (`http://www.sans.org`)
- National Security Agency (NSA) (`http://www.nsa.gov`)

We'll describe some of these techniques in detail.

Limiting Physical Access to the Host

Basic physical host security is required to prevent intruders from attacking the hardware of the virtual machine. When attackers can access a host they can do the following:

- Use OS-specific keystrokes to kill processes, monitor resource usage, or shut down the machine, commonly without needing a valid login account and password
- Reboot the machine, booting to external media with a known root password
- Steal files using external media (floppy, CD/DVD-RW, USB/flash drives, etc.)
- Capture traffic coming into or out of the network interfaces
- Remove one or more disks, mounting them in a machine with a known administrator or root password, potentially providing access to the entire contents of the host and guest VMs
- Simply remove the entire machine

Standard physical controls must also be implemented on the server room itself:

- Require card or guard access to the room with the machines.
- Use locks to anchor the machines to the building, and/or lock the cases to prevent removal of the hard drives.
- Remove floppy and CD drives after initial setup.
- In the BIOS, disable booting from any device except the primary hard drive. Also, password protect the BIOS so the boot choice cannot be changed.
- Control all external ports through host and guest system configuration or third-party applications.[7]

Using Encrypted Communications

Encryption technologies, such as Secure HTTP (HTTPS), encrypted Virtual Private Networks (VPNs), Transport Layer Security (TLS), Secure Shell (SSH), and so on should be used to provide secure communications links between the host domain and the guest domain, or from hosts to management systems. Encryption will help prevent such exploits as man-in-the-middle (MITM), spoofed attacks, and session hijacking.

In addition, standard traditional authentication techniques, such as failed login timeouts, strong passwords, BIOS passwords, warning banners, and password masking should be implemented.

Disabling Background Tasks

Most traditional server operating systems have multiple low-priority processes that are scheduled to run after primary business hours, when the server is expected to be less busy. Disabling, limiting, or off-loading these processes to other servers may be advisable if the host is beginning to suffer from resource contention.

The primary problem with background task detection on a virtual machine is that the virtual idle process is not fully aware of the state of the other virtual machines, and may not be able to make an accurate determination as to whether the host processor is really idle or not. This may lead to a situation where the background task demands more processor cycles than was initially intended.

In addition, several hacker exploits are designed to piggyback off of these processes, in an attempt to be less detectable to malware detection. Some of these processes may include file indexing tools, logging tools, and defragmenters.

Updating and Patching

Most standards organizations enforce the concept of timely patching and updating of systems. Unfortunately, the proliferation of VMs in an organization adds complexity to the patch control process. This means that not only must you patch and update the host OS promptly, but each of the virtual machines requires the same patching schedule. This is one reason standardization of an operating system throughout an enterprise is very important, if at all possible.

The patch schedule also requires management of the shutdown process, as most patches require rebooting after the patch is applied, and the administrator may have a very narrow maintenance window in which to apply the patch. Now you're shutting down not only the host, but every system that's on that host, and updating every VM on the host, as well as the host itself.

It's also imperative that the patch be tested on a nonproduction system representative of the system to be updated. While it's important that both the host and guest VMs receive the latest security patch, a research and testing control process must be implemented to demonstrate what effect an update may have on your specific configuration. A large part of integration resource is expended when an update has unforeseen consequences, and must be rolled back or results in required patching of other system components.

Keep up to date via mailing lists and news groups for information about the latest patches for your systems, and to research update implementation issues, especially for organizations that have systems comparable to yours. Also, some patches must be specifically modified by the virtualization vendor prior to implementation, so keep in close contact with your virtualization vendor through the patching and updating process.

Enabling Perimeter Defense on the VM

Perimeter defense devices are some of the oldest and most established ways of enforcing the security policy, by regulating data traffic ingress and egress. In fact, a common error of IT management is allocating too many resources (time and money) to purely perimeter defense, in the form of firewalls and hardened DMZ routers, while neglecting hardening the internal, trusted network. This often creates what's referred to as an *M&M* network security posture: crunchy on the outside but soft on the inside. The network is difficult to get into, but it lacks adequate controls once an intruder succeeds in penetrating the perimeter.

One advantage of enabling firewalls or intrusion detection systems through virtual machines on the host OS is that successful compromise of the guest domain may not compromise the host domain if the VM has been configured properly. Because the host domain controls the actual network traffic and makes final routing determinations after the VM has communicated, network-based intrusion detection or firewall products can very successfully be implemented at this choke point, and further helps the organization to implement a "defense-in-depth" strategy.

Implementing File Integrity Checks

One of the tenets of information systems security is the preservation of file integrity — that is, the guarantee that the contents of a file haven't been subjected to unauthorized alterations, either intentionally or unintentionally. File integrity checking is the process of verifying that the files retain the proper consistency, and serves as a check for intrusion into the system.

While the security classification level of the data to be hosted in the VM will determine the intensity and focus of the checking, it's recommended that file integrity checking be implemented at the host operating system level.

One way of implementing file integrity checking is by storing hash values of core OS files offline, as these files should not change often. Tripwire (`www.tripwire.com`), is one of the most established vendors of file integrity checking, and has recently begun focusing on virtualized environments in addition to traditional physical environments.

Maintaining Backups

We shouldn't even have to tell you this, but unfortunately we do. Perform image backups frequently for all production VMs. This will aid recovery of both individual files or the complete server image.

Protection of the physical backup is also a part of best practices. This includes protection of the data stream of the backup, which should be encrypted to prevent the interception of a server image by capturing the packets in the backup, as well as physical control of the backup media transport and storage.

THE ATTACK SURFACE

Attack surface **is a term that refers to the all of a host's running services that expose it to attack. The security profession tries to shrink the attack surface to as small a footprint as possible, while still maintaining business functionality. Shrinking reduces the vulnerability exposure the attack surface provides an attacker, and has the added benefit of lowering the complexity and resources needed to secure a system.**

AUDITING VM

It's very important that system auditors and assessors understand the inherent risks of any virtualized system that engages a connection to public networks (such as the Internet). Many standards and guidelines are being built to guide auditors in assessing the security posture of a virtualized environment, including guidelines from the U.S. Department of Defense (DoD) Defense Information Systems Agency (DISA),[8] the Center for Internet Security (CIS),[9] and various consulting organizations, such as the Burton Group.[10] These guidelines also provide recommendations for implementing the controls necessary to secure virtual machines and their hypervisors.

VM-Specific Security Techniques

A fundamental requirement for a successful virtualization security process is recognizing the dynamic nature of virtual machines. Therefore, many of the following security techniques are fairly unique to virtualized systems, and should be implemented in addition to the traditional best practice techniques just described.

Hardening the Virtual Machine

Virtual machines need to be configured securely, according to vendor-provided or industry best practices. Because this hardening may vary according to the vendor's implementation of virtualization, follow the vendor recommendations for best practice in this area.

This hardening can include many steps, such as the following:

- Putting limits on virtual machine resource consumption
- Configuring the virtual network interface and storage appropriately
- Disabling or removing unnecessary devices and services
- Ensuring that components that might be shared across virtual network devices are adequately isolated and secured
- Keeping granular and detailed audit logging trails for the virtualized infrastructure

It's important to use vendor supplied best practice configuration guides for both the guest and host domains, and refer to some of the published standards in this area, such as:

- NIST Computer Resource Center (`http://csrc.nist.gov/`)
- Defense Information Systems Agency (DISA) Security Technical Implementation Guides (STIGS) (`http://iase.disa.mil/stigs/index.html`)
- Center for Internet Security (`http://cisecurity.org`)
- SANS Institute (`http://www.sans.org/`)
- National Security Agency (NSA) (`http://www.nsa.gov/`)

Let's look at some important VM hardening techniques.

Harden the Hypervisor

It is critical to focus on the hypervisor as an attack vector, and strive to ensure that the hypervisor is deployed securely. Even before this stage, when you are evaluating various vendors' virtualization technology, place a premium on a vendor's track record of identifying vulnerabilities to its technology and the frequency of patch distribution.

Employ change and configuration controls to manage the virtual system patches and configuration changes to the hypervisor, and implement a testing process to test for publish vulnerabilities. Engaging a third-party testing service is standard best practice also.

Root Secure the Monitor

Because most operating systems can be compromised through privilege escalation, the VM monitor should be "root secure." This means that no level of privilege within the virtualized guest environment permits interference with the host system.

Implement Only One Primary Function per VM

While contemporary servers and virtual machines are adept at multi-tasking many functions, it's a lot easier to maintain secure control if the virtual machine is configured with process separation. It greatly complicates the hacker's ability to compromise multiple system components if the VM is implemented with one primary function per virtual server or device.

Firewall Any Additional VM Ports

The virtual machine may open multiple ports linked to the host's external IP address, besides the usual ports opened by the host. These ports are used to connect remotely to the virtual machine layer to view or configure virtual machines, share drives, or perform other tasks.

Therefore, the host system should be independently firewalled with a minimum of access allowed. Remote management of the host and VM will likely be required, but this communication should only take place on a separate NIC for administrative access only.

Harden the Host Domain

The Center for Internet Security (CIS) recently published a Xen benchmark study[11] that incorporates a lot of valuable security advice for hardening the host domain: "Before any virtual machines can be secure, the Host Domain of the host Linux operating system must be secure. A compromise of the Host Domain makes compromising the Guest Domains a simple task. Thus steps should be taken to reduce the attack surface of the Host Domain. These include but are not limited to:

- Remove unnecessary accounts and groups.
- Disable unnecessary services.
- Remove unnecessary binaries, libraries, and files.
- Firewall network access to the host.
- Install monitoring or Host Intrusion Detection Systems.
- Ensure that the Host Domain is not accessible from the Guest Domains.
- Ensure that monitoring or remote console interfaces for the Host Domain are not accessible via the Guest Domains.
- Ensure that the Guest Domains cannot directly affect any network storage or other resources that the Host Domain relies on for boot, configuration, or authentication.

The Host Domain host should only be used as a resource for virtualizing other operating environments. The Host Domain system should not host any other services or resources itself, including web, email and file servers. If such services are required, migrate the services to another system or consider creating a virtual machine to host them inside of a Guest Domain."

Use Unique NICs for Sensitive VMs

If possible, VMs that contain confidential databases and encrypted or sensitive information should have their network interface address bound to distinct and separate physical network interfaces (NICs). This external NIC would be the primary attack vector for intrusion, and isolation can help protect the VM.

Disconnect Unused Devices

It's advisable to disconnect the unneeded default virtual machine device connections when configuring the VM. Because the VM can control physical devices on the host, it's possible to insert media with undesired code into the device, enabling the code to execute when the VM mounts. Enable host access to devices only when explicitly required by the VM.

Additional VM Recommendations

Tavis Ormandy[12] also has additional recommendations for hardening virtualized systems:

- Treat Virtual Machines like services that can be compromised; use chroot, systrace, acls, least privileged users, etc.
- Disable emulated hardware you don't need, and external services you don't use (DHCP daemons, etc.) to reduce the attack surface exposed to hostile users.
- Xen is worth watching in future; separating domains should limit the impact of a compromise.
- Maintain the integrity of guest operating systems, protect the kernel using standard procedures of disabling modules: `/dev/mem`, `/dev/port`, etc.
- Keep guest software up-to-date with published vulnerabilities. If an attacker cannot elevate their privileges within the guest, the likelihood of compromising the VMM is significantly reduced.
- Keep Virtual Machine software updated to ensure all known vulnerabilities have been corrected.
- Avoid guests that do not operate in protected mode, and make use of any security features offered, avoid running untrusted code with root-equivalent privileges within the guest.

Securing VM Remote Access

Many virtual machine systems are rack-mounted, and may be located in a server farm physically distinct from the administration location. This usually requires the system administrator to access the virtualized system remotely for management tasks. This requires secure remote communications techniques.

Although each vendor's implementation of virtualization technology may differ, some general standard best practices exist when using remote services to access a system for administration. Most systems utilize a dedicated management NIC, and running service processes that are used to create a secure connection with the remote administrator.

Standard practices for remote administration include the following:

- Strong authentication practices should be employed:
  - Two-factor authentication
  - Strong passwords
  - One-time passwords
  - Private/public PKI key pairs
- Use encrypted communications only, such as a SSH or VPNs.
- MAC address or IP address filtering should be employed.
- Telnet access to the unit should be denied, as it does not encrypt the communications channel.

THE VALUE OF SSH

SSH (Secure Shell) is a terminal connection emulator that resembles Telnet but is a more secure tool for running management tasks remotely. SSH is cross-platform and can run both purely text-based sessions as well as X-Windows graphical applications. SSH is flexible enough to enable administrators to run the same set of management tools used in the nonvirtual, traditional environment, and it includes a wealth of various add-on tools built upon the SSH technology, such as SFTP (secure FTP) and PuTTY (see `http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html`**).**

It is best practice in SSH implementation to disable the less secure version 1 of the SSH protocol (SSH-1) and use only SSH-2. In addition, employ role-based access control (RBAC), or another access control mechanism, that forces users to use defined login accounts, to enforce accountability.

Summary

With the adoption of cloud technology comes many challenges to an organization, especially in the area of secure computing. Managing the security of the organization's private cloud as well as supervising the actions of the Cloud Services Provider can well become a monumental task.

To help lessen the size of the task, clearly defined management initiatives must be instituted which delineate clear ownership and responsibility of the data security. Therefore in this chapter we examined detective, preventative, and best practice controls to ensure that virtualization doesn't break the security posture of the company.

To this end we looked at the need and function of security policies, and gave some examples of what types of polices are usually developed. We also touched upon computer intrusion detection and response and the creation of a Computer Security Incident Response Team (CSIRT).

We spent the rest of the chapter examining various virtualization security management best practices. We looked first at some specific threats to the virtual environment, then examined a few general security best practices, and then ended with details of hardening techniques that are unique to virtualized systems.

Notes

1. "Attacking and Defending Virtual Environments," the Burton Group, http://www.burtongroup.com/Guest/Srms/AttackingDefendingVirtual.aspx
2. Virtual Machine Security Guidelines Version 1.0, September 2007, the Center for Internet Security
3. "Attacking and Defending Virtual Environments," the Burton Group, http://www.burtongroup.com/Guest/Srms/AttackingDefendingVirtual.aspx
4. "An Empirical Study into the Security Exposure to Hosts of Hostile Virtualized Environments," Tavis Ormandy, Google, Inc.
5. http://theinvisiblethings.blogspot.com/2006/06/introducing-blue-pill.html
6. "Attacking and Defending Virtual Environments," the Burton Group, http://www.burtongroup.com/Guest/Srms/AttackingDefendingVirtual.aspx
7. Virtual Machine Security Guidelines, Version 1.0, the Center for Internet Security
8. "Virtual Machine Security Technical Implementation Guide," U.S. Department of Defense Information Systems Agency, http://iase.disa.mil/stigs/stig/vm_stig_v2r2.pdf
9. "CIS Level 1 Benchmark for Virtual Machines," Center for Internet Security, http://www.cisecurity.org/bench_vm.htm

10. “Attacking and Defending Virtual Environments,” the Burton Group, `http://www.burtongroup.com/Guest/Srms/AttackingDefendingVirtual.aspx`

11. Benchmark for Xen 3.2 Version 1.0, May 2008, the Center for Internet Security (CIS)

12. “An Empirical Study into the Security Exposure to Hosts of Hostile Virtualized Environments, ” Tavis Ormandy, Google, Inc.

CHAPTER

6

Cloud Computing Security Architecture

It is much more secure to be feared than to be loved.

—Niccolo Machiavelli

With all the advantages of the cloud paradigm and its potential for decreasing costs and reducing the time required to start new initiatives, cloud security will always be a major concern. Virtualized resources, geographically dispersed servers, and co-location of processing and storage pose challenges and opportunities for cloud providers and users.

The security posture of a cloud system is based on its security architecture. While there is no standard definition for security architecture, the Open Security Alliance (OSA) defines *security architecture* as "the design artifacts that describe how the security controls (= security countermeasures) are positioned, and how they relate to the overall IT Architecture. These controls serve the purpose to maintain the system's quality attributes, among them confidentiality, integrity, availability, accountability and assurance" (`http://www.opensecurityarchitecture.org/cms/definitions`).

A second definition developed by the Information Security Society Switzerland (ISSS) describes a *security architecture* as "a cohesive security design, which addresses the requirements (e.g., authentication, authorization, etc.) and in particular the risks of a particular environment/scenario, and specifies what security controls are to be applied where. The design process should be reproducible" (`http://www.isss.ch/fileadmin/publ/agsa/Security_Architecture.pdf`).

In this chapter, the general security architecture issues involved, the architectural components of trusted cloud computing, core security architectural functions, and the potential of autonomic systems to implement secure architectures will be presented.

Architectural Considerations

A variety of factors affect the implementation and performance of cloud security architecture. There are general issues involving regulatory requirements, adherence to standards, security management, information classification, and security awareness. Then there are more specific architecturally related areas, including trusted hardware and software, providing for a secure execution environment, establishing secure communications, and hardware augmentation through microarchitectures. These important concepts are addressed in this section.

General Issues

A variety of topics influence and directly affect the cloud security architecture. They include such factors as compliance, security management, administrative issues, controls, and security awareness.

Compliance with legal regulations should be supported by the cloud security architecture. As a corollary, the cloud security policy should address classification of information, what entities can potentially access information, under what conditions the access has to be provided, the geographical jurisdiction of the stored data, and whether or not the access is appropriate. Proper controls should be determined and verified with assurance methods, and appropriate personnel awareness education should be put in place.

Compliance

In a public cloud environment, the provider does not normally inform the clients of the storage location of their data. In fact, the distribution of processing and data storage is one of the cloud's fundamental characteristics. However, the cloud provider should cooperate to consider the client's data location requirements. In addition, the cloud vendor should provide transparency to the client by supplying information about storage used, processing characteristics, and other relevant account information. Another compliance issue is the accessibility of a client's data by the provider's system engineers and certain other employees. This factor is a necessary part of providing and maintaining cloud services, but the act of acquiring sensitive information should be monitored, controlled, and protected by safeguards such as separation of duties. In situations where information is stored in a foreign jurisdiction, the ability of local law enforcement agencies to access a client's sensitive data is a concern. For example, this scenario might occur when a government entity conducts a computer forensics investigation of a cloud provider under suspicion of illegal activity.

The cloud provider's claims for data protection and compliance must be backed up by relevant certifications, logging, and auditing. In particular, at a minimum, a cloud provider should undergo a Statement on Auditing Standard # 70 (SAS 70) "Service Organizations" Type II Audit (`www.SaS70.com`). This audit evaluates a service organization's internal controls to determine whether accepted best practices are being applied to protect client information. Cloud vendors are required to undergo subsequent audits to retain their SAS 70 Type II Audit certification.

Another source of direction for the cloud provider is given in Domain 4 of the "Cloud Security Alliance Security Guidance for Critical Areas of Focus in Cloud Computing" (`http://www.cloudsecurityalliance.org/`). Domain 4 stresses the roles of cloud customers, cloud providers, and auditors with respect to compliance responsibilities, the requirements of compliance evidence, and the need to acquaint assessors with the unique characteristics of cloud computation.

A related issue is the management policy associated with data stored in the cloud. When a client's engagement with the cloud provider is terminated, compliance and privacy requirements have to be considered. In some cases, information has to be preserved according to regulatory requirements and in other instances the provider should not hold a client's data in primary or backup storage if the client believes it has been destroyed. If stored in a foreign jurisdiction, the data might be subject to that country's privacy laws and not the laws applicable in the client's geographic location.

The evolution and application of appropriate cloud standards focused on legal requirements will also serve to meet the cloud's compliance requirements and provide the necessary protections. A number of standards organizations have joined forces under the title of the Cloud Standards Coordination Working Group to develop a cloud computing standardization approach. The Working Group includes the Object Management Group, the Distributed Management Task Force, the TeleManagement (TM) Forum, the Organization for the Advancement of Structured Information Standards, the Open Grid Forum, the Cloud Security Alliance, the Open Cloud Consortium, the Storage and Network Industry Association, and the Cloud Computing Interoperability Forum. Standards efforts are discussed in more detail in Chapter 7.

Security Management

Security architecture involves effective security management to realize the benefits of cloud computation. Proper cloud security management and administration should identify management issues in critical areas such as access control, vulnerability analysis, change control, incident response, fault tolerance, and disaster recovery and business continuity planning. These areas are enhanced and supported by the proper application and verification of cloud security controls.

Controls

The objective of cloud security controls is to reduce vulnerabilities to a tolerable level and minimize the effects of an attack. To achieve this, an organization must determine what impact an attack might have, and the likelihood of loss. Examples of loss are compromise of sensitive information, financial embezzlement, loss of reputation, and physical destruction of resources. The process of analyzing various threat scenarios and producing a representative value for the estimated potential loss is known as a *risk analysis (RA)*. Controls function as countermeasures for vulnerabilities. There are many kinds of controls, but they are generally categorized into one of the following four types:[1]

- **Deterrent controls** — Reduce the likelihood of a deliberate attack.
- **Preventative controls** — Protect vulnerabilities and make an attack unsuccessful or reduce its impact. Preventative controls inhibit attempts to violate security policy.
- **Corrective controls** — Reduce the effect of an attack.
- **Detective controls** — Discover attacks and trigger preventative or corrective controls. Detective controls warn of violations or attempted violations of security policy and include such controls as intrusion detection systems, organizational policies, video cameras, and motion detectors.

OMB CIRCULAR A-130

The U.S. Office of Management and Budget Circular A-130, revised November 30, 2000, requires that a review of the security controls for each major U.S. government application be performed at least every three years. For general support systems, OMB Circular A-130 requires that the security controls are either reviewed by an independent audit or self-reviewed. Audits can be self-administered or independent (either internal or external). The essential difference between a self-audit and an independent audit is objectivity; however, some systems may require a fully independent review.

Complementary Actions

Additional activities involved in cloud security management include the following:

- Management and monitoring of service levels and service-level agreements
- Acquisition of adequate data to identify and analyze problem situations through instrumentation and dashboards

- Reduction of the loss of critical information caused by lack of controls.
- Proper management of data on an organization's distributed computing resources. Data centralized on the cloud reduces the potential for data loss in organizations with large numbers of laptop computers and other personal computing devices.
- Monitoring of centrally stored cloud information, as opposed to having to examine data distributed throughout an organization on a variety of computing and storage devices.
- Provisioning for rapid recovery from problem situations.

Cloud security management should also foster improved capabilities to conduct forensic analysis on cloud-based information using a network forensic model. This model will provide for more rapid acquisition and verification of evidence, such as taking advantage of automatic hashing that is applied when storing data on a cloud.

Cloud security management can also be enhanced by the selective use of automation and by the application of emerging cloud management standards to areas such as interoperable security mechanisms, quality of service, accounting, provisioning, and API specifications. APIs provide for control of cloud resources through program interfaces, and remote APIs should be managed to ensure that they are documented and consistent.

Cloud security management should address applications with the goal of enterprise cost containment through scalability, pay as you go models, on-demand implementation and provisioning, and reallocation of information management operational activities to the cloud.

Information Classification

Another major area that relates to compliance and can affect the cloud security architecture is information classification. The information classification process also supports disaster recovery planning and business continuity planning.

Information Classification Objectives

There are several good reasons to classify information. Not all data has the same value to an organization. For example, some data is more valuable to upper management, because it aids them in making strategic long-range or short-range business direction decisions. Some data, such as trade secrets, formulas, and new product information, is so valuable that its loss could create a significant problem for the enterprise in the marketplace — either by creating public embarrassment or by causing a lack of credibility.

For these reasons, it is obvious that information classification has a higher, enterprise-level benefit. Information stored in a cloud environment can have an impact on a business globally, not just on the business unit or line operation levels. Its primary purpose is to enhance confidentiality, integrity, and availability (the CIA triad described in Chapter 3), and minimize risks to the information. In addition, by focusing the protection mechanisms and controls on the information areas that most need it, you achieve a more efficient cost-to-benefit ratio.

Information classification has the longest history in the government sector. Its value has long been established, and it is a required component when securing trusted systems. In this sector, information classification is used primarily to prevent the unauthorized disclosure of information and the resultant failure of confidentiality.

Information classification supports privacy requirements and enables regulatory compliance. A company might wish to employ classification to maintain a competitive edge in a tough marketplace. There might also be sound legal reasons for an organization to employ information classification on the cloud, such as to minimize liability or to protect valuable business information.

Information Classification Benefits

In addition to the aforementioned reasons, employing information classification has several clear benefits to an organization engaged in cloud computing. Some of these benefits are as follows:

- It demonstrates an organization's commitment to security protections.
- It helps identify which information is the most sensitive or vital to an organization.
- It supports the tenets of confidentiality, integrity, and availability as it pertains to data.
- It helps identify which protections apply to which information.
- It might be required for regulatory, compliance, or legal reasons.

Information Classification Concepts

The information that an organization processes must be classified according to the organization's sensitivity to its loss or disclosure. The information system owner is responsible for defining the sensitivity level of the data. Classification according to a defined classification scheme enables security controls to be properly implemented.

The following classification terms are typical of those used in the private sector and are applicable to cloud data**:**

- **Public data** — Information that is similar to unclassified information; all of a company's information that does not fit into any of the next categories can be considered public. While its unauthorized disclosure may be against policy, it is not expected to impact seriously or adversely the organization, its employees, and/or its customers.

- **Sensitive data** — Information that requires a higher level of classification than normal data. This information is protected from a loss of confidentiality as well as from a loss of integrity due to an unauthorized alteration. This classification applies to information that requires special precautions to ensure its integrity by protecting it from unauthorized modification or deletion. It is information that requires a higher-than-normal assurance of accuracy and completeness.
- **Private data** — This classification applies to personal information that is intended for use within the organization. Its unauthorized disclosure could seriously and adversely impact the organization and/or its employees. For example, salary levels and medical information are considered private.
- **Confidential data** — This classification applies to the most sensitive business information that is intended strictly for use within the organization. Its unauthorized disclosure could seriously and adversely impact the organization, its stockholders, its business partners, and/or its customers. This information is exempt from disclosure under the provisions of the Freedom of Information Act or other applicable federal laws or regulations. For example, information about new product development, trade secrets, and merger negotiations is considered confidential.

An organization may use a high, medium, or low classification scheme based upon its CIA needs and whether it requires high, medium, or low protective controls. For example, a system and its information may require a high degree of integrity and availability, yet have no need for confidentiality.

The designated owners of information are responsible for determining data classification levels, subject to executive management review. Table 6-1 shows a simple High/Medium/Low (H/M/L) data classification schema for sensitive information.

Table 6-1: High/Medium/Low Classifications

CLASSIFICATION	IMPACT
High	Could cause loss of life, imprisonment, major financial loss, or require legal remediation if the information is compromised
Medium	Could cause noticeable financial loss if the information is compromised
Low	Would cause only minor financial loss or require minor administrative action for correction if the information is compromised

From NIST SP 800-26, “Security Self-Assessment Guide for Information Technology Systems.”

Classification Criteria

Several criteria may be used to determine the classification of an information object:

- **Value** — Value is the number one commonly used criteria for classifying data in the private sector. If the information is valuable to an organization or its competitors, then it needs to be classified.
- **Age** — The classification of information might be lowered if the information's value decreases over time. In the U.S. Department of Defense, some classified documents are automatically declassified after a predetermined time period has passed.
- **Useful life** — If the information has been made obsolete due to new information, substantial changes in the company, or other reasons, the information can often be declassified.
- **Personal association** — If information is personally associated with specific individuals or is addressed by a privacy law, it might need to be classified. For example, investigative information that reveals informant names might need to remain classified.

Information Classification Procedures

There are several steps in establishing a classification system. These are the steps in priority order:

1. Identify the appropriate administrator and data custodian. The data custodian is responsible for protecting the information, running backups, and performing data restoration.
2. Specify the criteria for classifying and labeling the information.
3. Classify the data by its owner, who is subject to review by a supervisor.
4. Specify and document any exceptions to the classification policy.
5. Specify the controls that will be applied to each classification level.
6. Specify the termination procedures for declassifying the information or for transferring custody of the information to another entity.
7. Create an enterprise awareness program about the classification controls.

Distribution of Classified Information

External distribution of sensitive or classified information stored on a cloud is often necessary, and the inherent security vulnerabilities need to be addressed. Some of the instances when this distribution is required are as follows:

- **Court order** — Classified or sensitive information might need to be disclosed to comply with a court order.
- **Government contracts** — Government contractors might need to disclose classified or sensitive information in accordance with (IAW) the procurement agreements related to a government project.
- **Senior-level approval** — A senior-level executive might authorize the release of classified or sensitive information to external entities or organizations. This release might require the signing of a confidentiality agreement by the external party.

Employee Termination

It is important to understand the impact of employee terminations on the integrity of information stored in a cloud environment. This issue applies to employees of the cloud client as well as the cloud provider. Typically, there are two types of terminations, friendly and unfriendly, and both require specific actions.

Friendly terminations should be accomplished by implementing a standard set of procedures for outgoing or transferring employees. This activity normally includes the following:[2]

- The removal of access privileges, computer accounts, authentication tokens.
- The briefing on the continuing responsibilities of the terminated employee for confidentiality and privacy.
- The return of company computing property, such as laptops.
- Continued availability of data. In both the manual and the electronic worlds, this may involve documenting procedures or filing schemes, such as how documents are stored on the hard disk and how they are backed up. Employees should be instructed whether or not to "clean up" their PC before leaving.
- If cryptography is used to protect data, the availability of cryptographic keys to management personnel must be ensured.

Given the potential for adverse consequences during an unfriendly termination, organizations should do the following:

- System access should be terminated as quickly as possible when an employee is leaving a position under less-than-friendly terms. If employees are to be fired, system access should be removed at the same time (or just before) the employees are notified of their dismissal.

- When an employee resigns and it can be reasonably assumed that it is on unfriendly terms, system access should be immediately terminated, or as soon as feasible.
- During the *notice of termination* period, it may be necessary to restrict the individual to a given area and function. This may be particularly true for employees capable of changing programs or modifying the system or applications.
- In some cases, physical removal from the offices may be necessary.

In either scenario, network access and system rights must be strictly controlled.

Security Awareness, Training, and Education

Security awareness is often overlooked as an element affecting cloud security architecture because most of a security practitioner's time is spent on controls, intrusion detection, risk assessment, and proactively or reactively administering security. Employees must understand how their actions, even seemingly insignificant actions, can greatly impact the overall security position of an organization.

Employees of both the cloud client and the cloud provider must be aware of the need to secure information and protect the information assets of an enterprise. Operators need ongoing training in the skills that are required to fulfill their job functions securely, and security practitioners need training to implement and maintain the necessary security controls, particularly when using or providing cloud services.

All employees need education in the basic concepts of security and its benefits to an organization. The benefits of the three pillars of security awareness training — awareness, training, and education — will manifest themselves through an improvement in the behavior and attitudes of personnel and through a significant improvement in an enterprise's security.

The purpose of computer security awareness, training, and education is to enhance security by doing the following:

- Improving awareness of the need to protect system resources
- Developing skills and knowledge so computer users can perform their jobs more securely
- Building in-depth knowledge, as needed, to design, implement, or operate security programs for organizations and systems

An effective computer security awareness and training program requires proper planning, implementation, maintenance, and periodic evaluation. In general, a computer security awareness and training program should encompass the following seven steps:[3]

1. Identify program scope, goals, and objectives.
2. Identify training staff.
3. Identify target audiences.
4. Motivate management and employees.
5. Administer the program.
6. Maintain the program.
7. Evaluate the program.

Making cloud system users and providers aware of their security responsibilities and teaching them correct practices helps change their behavior. It also supports individual accountability because without knowledge of the necessary security measures and to how to use them, personnel cannot be truly accountable for their actions.

Security Awareness

As opposed to training, the security awareness of an organization refers to the degree to which its personnel are collectively aware of the importance of security and security controls. In addition to the benefits and objectives previously mentioned, security awareness programs also have the following benefits:

- They can reduce the unauthorized actions attempted by personnel.
- They can significantly increase the effectiveness of the protection controls.
- They help to prevent the fraud, waste, and abuse of computing resources.

Personnel are considered "security aware" when they clearly understand the need for security, how security affects viability and the bottom line, and the daily risks to cloud computing resources.

It is important to have periodic awareness sessions to orient new employees and refresh senior employees. The material should always be direct, simple, and clear. It should be fairly motivational and should not contain a lot of techno-jargon, and you should convey it in a style that the audience easily understands. These sessions are most effective when they demonstrate how the security interests of the organization parallel the interests of the audience.

The following activities can be used to improve security within an organization without incurring large costs or draining resources:

- **Live/interactive presentations** — Lectures, videos, and computer-based training (CBT)
- **Publishing/distribution** — Posters, company newsletters, bulletins, and the intranet
- **Incentives** — Awards and recognition for security-related achievements

- **Reminders** — Log-in banner messages and marketing paraphernalia such as mugs, pens, sticky notes, and mouse pads

Training and Education

Training is different from awareness in that it provides security information in a more formalized manner, such as classes, workshops, or individualized instruction. The following types of training are related to cloud security:

- Security-related job training for operators and specific users
- Awareness training for specific departments or personnel groups with security-sensitive positions
- Technical security training for IT support personnel and system administrators
- Advanced training for security practitioners and information systems auditors
- Security training for senior managers, functional managers, and business unit managers

In-depth training and education for systems personnel, auditors, and security professionals is critical, and typically necessary for career development. In addition, specific product training for cloud security software and hardware is vital to the protection of the enterprise.

Motivating the personnel is always the prime directive of any training, and their understanding of the value of security's impact to the bottom line is also vital. A common training technique is to create hypothetical cloud security vulnerability scenarios and then solicit input on possible solutions or outcomes.

Trusted Cloud Computing

Trusted cloud computing can be viewed as a computer security architecture that is designed to protect cloud systems from malicious intrusions and attacks, and ensure that computing resources will act in a specific, predictable manner as intended. A trusted cloud computing system will protect data in use by hypervisors and applications, protect against unauthorized access to information, provide for strong authentication, apply encryption to protect sensitive data that resides on stolen or lost devices, and support compliance through hardware and software mechanisms.

Trusted Computing Characteristics

In a cloud computational system, multiple processes might be running concurrently. Each process has the capability to access certain memory locations and to execute a subset of the computer's instruction set. The execution and memory space assigned to each process is called a *protection domain*. This domain can be extended to virtual memory, which increases the apparent size of real memory by using disk storage. The purpose of establishing a protection domain is to protect programs from all unauthorized modification or executional interference.

A *trusted computing base (TCB)* is the total combination of protection mechanisms within a computer system, which includes the hardware, software, and firmware that are trusted to enforce a security policy. Because the TCB components are responsible for enforcing the security policy of a computing system, these components must be protected from malicious and untrusted processes. The TCB must also provide for memory protection and ensure that the processes from one domain do not access memory locations of another domain. The *security perimeter* is the boundary that separates the TCB from the remainder of the system. A *trusted path* must also exist so that users can access the TCB without being compromised by other processes or users. Therefore, a *trusted computer system* is one that employs the necessary hardware and software assurance measures to enable its use in processing multiple levels of classified or sensitive information. This system meets the specified requirements for reliability and security.

Another element associated with trusted computing is the *trusted platform module (TPM)*. The TPM stores cryptographic keys that can be used to attest to the operating state of a computing platform and to verify that the hardware and software configuration has not been modified. However, the standard TPM cannot be used in cloud computing because it does not operate in the virtualized cloud environment. To permit a TPM version to perform in the cloud, specifications have been generated for a virtual TPM (VTM)[4] that provides software instances of TPMs for each virtual machine operating on a trusted server.

Trusted computing also provides the capability to ensure that software that processes information complies with specified usage policies and is running unmodified and isolated from other software on the system. In addition, a trusted computing system must be capable of enforcing mandatory access control (MAC) rules. MAC rules are discussed in more detail later in this chapter.

Numerous trust-related issues should be raised with, and satisfied by, a cloud provider. They range from concerns about security, performance, cost, control, availability, resiliency, and vendor lock in. Following are some of the critical questions that should be asked to address these concerns:

- Do I have any control or choice over where my information will be stored? Where will my data reside and what are the security and privacy laws in effect in those locations?
- Are your cloud operations available for physical inspection?
- Can you provide an estimate of historical downtimes at your operation?

- Are there any exit charges or penalties for migrating from your cloud to another vendor's cloud operation? Do you delete all my data from your systems if I move to another vendor?
- Can you provide documentation of your disaster recovery policies and procedures and how they are implemented?

These questions related to basic trust issues associated with cloud computing arise from the characteristics and architecture of cloud resources. The cloud handles multi-party, co-located applications, and this capability brings with it corresponding security issues and requirements to minimize risk. The cloud provider must conduct quality risk assessments at regular, known intervals to meet the trust expectations of clients and auditors, and demonstrate that risk is being managed effectively. Additional factors that inspire trust include the following:

- Use of industry-accepted standards.
- Provision for interoperability and transparency.
- Robust authentication and authorization mechanisms in access control.
- Management of changing personnel and relationships in both the cloud client and provider organizations.
- Establishment of accountability with respect to security and privacy requirements in a multi-party, flexible service delivery setting.
- Use of information system security assurance techniques and metrics to establish the effectiveness of hardware and software protection mechanisms.
- Establishment of effective policies and procedures to address multiple legal jurisdictions associated with cloud international services and compliance requirements.
- Application of Information Rights Management (IRM) cryptographic techniques to protect sensitive cloud-based documents and provide an audit trail of accesses and policy changes. IRM prevents protected documents from screen capture, being printed, faxed, or forwarded, and can prohibit messages and attachments from being accessed after a specified period of time.

Also, because of the high volume of data that is being moved around in various locations, authorization privileges and rights management constraints must be attached to the data itself to restrict access only to authorized users.

Because of legal and forensic requirements, a trusted cloud provider should also have a Security Information and Event Management (SIEM) capability that can manage records and logs in a manner that meets legal constraints. An SEIM is a software mechanism that provides for centralized acquisition, storage, and analysis of recorded events and logs generated by other tools on an enterprise network.

Information stored in a SEIM can be used for data mining to discover significant trends and occurrences, and to provide for reliable and legally acceptable storage of information. It can also be used by report generators, and provide for backup of log data that might be lost at the source of the data.

Secure Execution Environments and Communications

In a cloud environment, applications are run on different servers in a distributed mode. These applications interact with the outside world and other applications and may contain sensitive information whose inappropriate access would be harmful to a client. In addition, cloud computing is increasingly being used to manage and store huge amounts of data in database applications that are also co-located with other users' information. Thus, it is extremely important for the cloud supplier to provide a secure execution environment and secure communications for client applications and storage.

Secure Execution Environment

Configuring computing platforms for secure execution is a complex task; and in many instances it is not performed properly because of the large number of parameters that are involved. This provides opportunities for malware to exploit vulnerabilities, such as downloading code embedded in data and having the code executed at a high privilege level.

In cloud computing, the major burden of establishing a secure execution environment is transferred from the client to the cloud provider. However, protected data transfers must be established through strong authentication mechanisms, and the client must have practices in place to address the privacy and confidentiality of information that is exchanged with the cloud. In fact, the client's port to the cloud might provide an attack path if not properly provisioned with security measures. Therefore, the client needs assurance that computations and data exchanges are conducted in a secure environment. This assurance is affected by trust enabled by cryptographic methods. Also, research into areas such as compiler-based virtual machines promises a more secure execution environment for operating systems.

Another major concern in secure execution of code is the widespread use of "unsafe" programming languages such as C and C++ instead of more secure languages such as object-oriented Java and structured, object-oriented C#.

Secure Communications

As opposed to having managed, secure communications among the computing resources internal to an organization, movement of applications to the cloud requires a reevaluation of communications security. These communications apply to both data in motion and data at rest.

Secure cloud communications involves the structures, transmission methods, transport formats, and security measures that provide confidentiality, integrity, availability, and authentication for transmissions over private and public communications networks. Secure cloud computing communications should ensure the following:

- **Confidentiality** — Ensures that only those who are supposed to access data can retrieve it. Loss of confidentiality can occur through the intentional release of private company information or through a misapplication of network rights. Some of the elements of telecommunications used to ensure confidentiality are as follows:
  - Network security protocols
  - Network authentication services
  - Data encryption services
- **Integrity** — Ensures that data has not been changed due to an accident or malice. Integrity is the guarantee that the message sent is the message received and that the message is not intentionally or unintentionally altered. Integrity also contains the concept of nonrepudiation of a message source. Some of the constituents of integrity are as follows:
  - Firewall services
  - Communications Security Management
  - Intrusion detection services
- **Availability** — Ensures that data is accessible when and where it is needed, and that connectivity is accessible when needed, allowing authorized users to access the network or systems. Also included in that assurance is the guarantee that security services for the security practitioner are usable when they are needed. Some of the elements that are used to ensure availability are as follows:
  - Fault tolerance for data availability, such as backups and redundant disk systems
  - Acceptable logins and operating process performances
  - Reliable and interoperable security processes and network security mechanisms

APIs

Common vulnerabilities such as weak antivirus software, unattended computing platforms, poor passwords, weak authentication mechanisms, and inadequate intrusion detection that can impact communications must be more stringently analyzed, and proper APIs must be used.

For example, in using IaaS, a cloud client typically communicates with cloud server instances through Representational State Transfer (REST) client/server model or Simple Object Access Protocol (SOAP) APIs. REST is a software architecture such as used in the World Wide Web and was developed with the HTTP/1.1 protocol. With SOAP, applications running on different operating systems and using different programming languages can communicate with each other.

Virtual Private Networks

Another important method to secure cloud communications is through a virtual private network (VPN). A VPN is created by building a secure communications link between two nodes by emulating the properties of a point-to-point private link. A VPN can be used to facilitate secure remote access into the cloud, securely connect two networks together, or create a secure data tunnel within a network.

The portion of the link in which the private data is encapsulated is known as the *tunnel*. It may be referred to as a secure, encrypted tunnel, although it's more accurately defined as an encapsulated tunnel, as encryption may or may not be used. To emulate a point-to-point link, data is encapsulated, or wrapped, with a header that provides routing information. Most often the data is encrypted for confidentiality. This encrypted part of the link is considered the actual virtual private network connection. Figure 6-1 shows a common VPN configuration with example IP addresses for remote access into an organization's intranet through the Internet. Address 192.168.123.2 designates the organization's router.

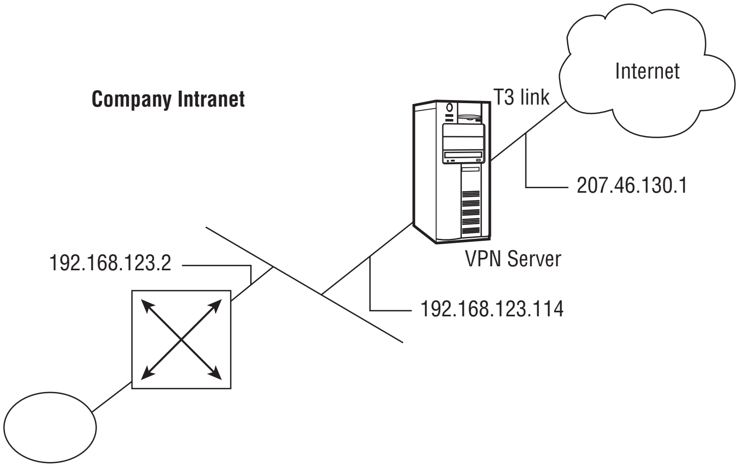

Figure 6-1: VPN configuration

The two general types of VPNs relevant to cloud computing are remote access and network-to-network. These VPN types are described in the following sections.

Remote Access VPNs

A VPN can be configured to provide remote access to corporate resources over the public Internet to maintain confidentiality and integrity. This configuration enables the remote user to utilize whatever local ISP is available to access the Internet without forcing the user to make a long-distance or 800 call to a third-party access provider. Using the connection to the local ISP, the VPN software creates a virtual private network between the dial-up user and the corporate VPN server across the Internet. Figure 6-2 shows a remote user VPN connection.

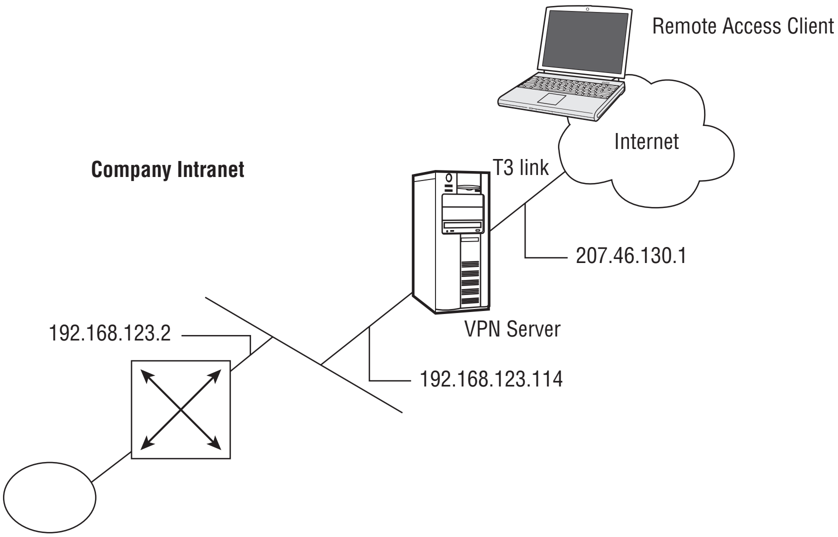

Figure 6-2: Remote access VPN configuration

Network-to-Network VPNs

A VPN is commonly used to connect two networks, perhaps the main corporate LAN and a remote branch office LAN, through the Internet. This connection can use either dedicated lines to the Internet or dial-up connections to the Internet. However, the corporate hub router that acts as a VPN server must be connected to a local ISP with a dedicated line if the VPN server needs to be available 24/7. The VPN software uses the connection to the local ISP to create a VPN tunnel between the branch office router and the corporate hub router across the Internet. Figure 6-3 shows a remote branch office connected to the corporate main office using a VPN tunnel through the Internet.

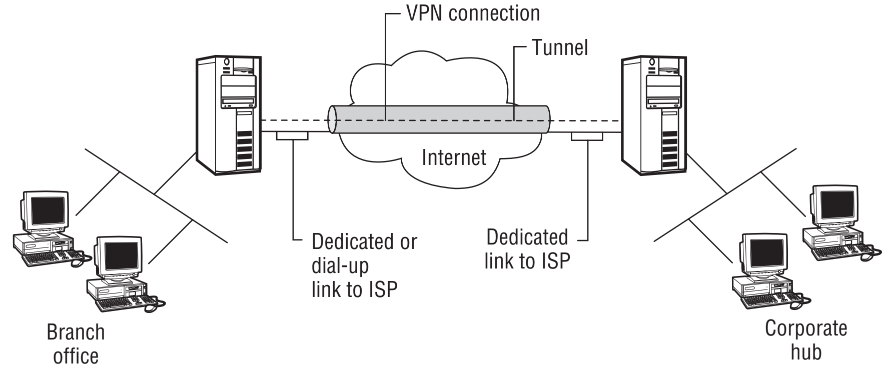

Figure 6-3: A network-to-network VPN configuration

VPN Tunneling

Tunneling is a method of transferring data from one network to another network by encapsulating the packets in an additional header. The additional header provides routing information so that the encapsulated payload can traverse the intermediate networks, as shown in Figure 6-4.

For a tunnel to be established, both the tunnel client and the tunnel server must be using the same tunneling protocol. Tunneling technology can be based on either a Layer 2 or a Layer 3 tunneling protocol. These layers correspond to the Open Systems Interconnection (OSI) Reference Model.

Tunneling, and the use of a VPN, is not intended as a substitute for encryption/ decryption. In cases where a high level of security is necessary, the strongest possible encryption should be used within the VPN itself, and tunneling should serve only as a convenience.

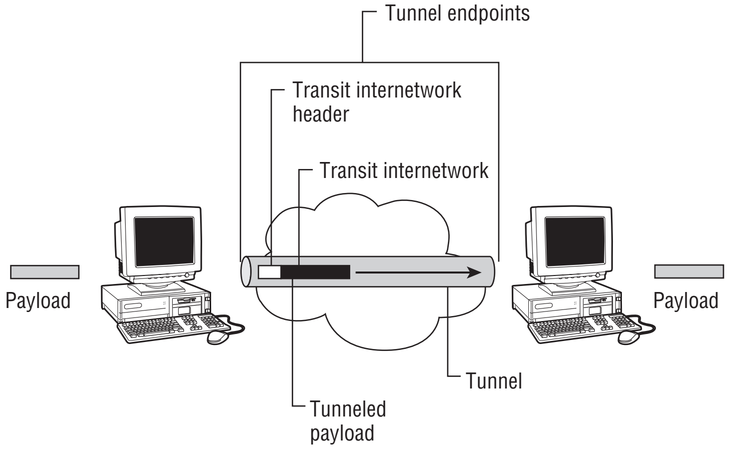

Figure 6-4: A VPN tunnel and payload

A popular tunneling protocol for network-to-network connectivity is IPSec, which encapsulates IP packets in an additional IP header. IPSec operates at the Network Layer of the OSI Reference Model and allows multiple simultaneous tunnels. IPSec contains the functionality to encrypt and authenticate IP data. It is built into the new IPv6 standard and is used as an add-on to the current IPv4. IPSec tunnel mode allows IP packets to be encrypted and then encapsulated in an IP header to be sent across a corporate IP Intranetwork or a public IP Internetwork, such as the Internet.

IPSec uses an authentication header (AH) to provide source authentication and integrity without encryption, and it uses the Encapsulating Security Payload (ESP) to provide authentication and integrity along with encryption. With IPSec, only the sender and recipient know the key. If the authentication data is valid, then the recipient knows that the communication came from the sender and was not changed in transit.

Public Key Infrastructure and Encryption Key Management

To secure communications, data that is being exchanged with a cloud should be encrypted, calls to remote servers should be examined for imbedded malware, and digital certificates should be employed and managed. A certification process can be used to bind individuals to their public keys as used in public key cryptography. A *certificate authority (CA)* acts as notary by verifying a person's identity and issuing a certificate that vouches for a public key of the named individual. This certification agent signs the certificate with its own private key. Therefore, the individual is verified as the sender if that person's public key opens the data.

The certificate contains the subject's name, the subject's public key, the name of the certificate authority, and the period in which the certificate is valid. To verify the CA's signature, its public key must be cross-certified with another CA. (The X.509 standard defines the format for public key certificates.) This certificate is then sent to a repository, which holds the certificates and *certificate revocation lists (CRLs)* that denote the revoked certificates. Figure 6-5 illustrates the use of digital certificates in a transaction between a subscribing entity and a transacting party. Digital certificates are discussed in more detail in the following sections.

The integration of digital signatures and certificates and the other services required for e-commerce is called the *public key infrastructure (PKI)*. These services provide integrity, access control, confidentiality, authentication, and nonrepudiation for electronic transactions. The PKI includes the following elements:

- Digital certificates
- Certificate authority (CA)
- Registration authorities

- Policies and procedures
- Certificate revocation
- Nonrepudiation support
- Timestamping
- Lightweight Directory Access Protocol (LDAP)
- Security-enabled applications

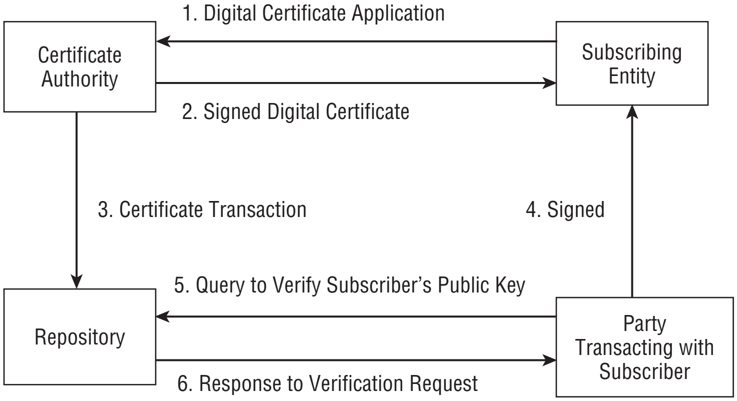
1. Digital Certificate Application
Certificate Authority
Subscribing Entity
2. Signed Digital Certificate
3. Certificate Transaction
4. Signed
5. Query to Verify Subscriber's Public Key
Repository
Party Transacting with Subscriber
6. Response to Verification Request

Figure 6-5: A transaction with digital certificates

Digital Certificates

The digital certificate and management of the certificate are major components of PKI. Remember: The purpose of a digital certificate is to verify to all that an individual's public key — posted on a public "key ring" — is actually his or hers. A trusted, third-party CA can verify that the public key is that of the named individual and then issue a certificate attesting to that fact. The CA accomplishes the certification by digitally signing the individual's public key and associated information.

Certificates and CRLs can be held in a repository, with responsibilities defined between the repository and the CA. The repository access protocol determines how these responsibilities are assigned. In one protocol, the repository interacts with other repositories, CAs, and users. The CA deposits its certificates and CRLs into the repository. The users can then access the repository for this information.

Directories and X.500

In PKI, a repository is usually referred to as a *directory*. The directory contains entries associated with an object class. An object class can refer to individuals or other computer-related entities. The class defines the attributes of the object. Attributes for PKI are defined in RFC 2587, "Internet X.509 Public Key Infrastructure LDAP v2 Schema," by Boeyen, Howes, and Richard, published in April 1999. Additional information on attributes can be found in RFC 2079, "Definition of an X.500 Attribute Type and an Object Class to Hold Uniform Resource Identifiers (URLs)," by M. Smith, published in January 1997.

The X.509 certificate standard defines the authentication bases for the X.500 directory. The X.500 directory stores information about individuals and objects in a distributed database residing on network servers. Some of the principal definitions associated with X.500 include the following:

- Directory user agents (DUAs) — Clients
- Directory server agents (DSAs) — Servers
- Directory Service Protocol (DSP) — Enables information exchanges between DSAs
- Directory Access Protocol (DAP) — Enables information exchanges from a DUA to a DSA
- Directory Information Shadowing Protocol (DISP) — Used by a DSA to duplicate or "shadow" some or all of its contents

DSAs accept requests from anonymous sources as well as authenticated requests. They share information through a *chaining* mechanism.

The Lightweight Directory Access Protocol

The Lightweight Directory Access Protocol (LDAP) was developed as a more efficient version of the DAP and has evolved into a second version (see RFC 1777, "Lightweight Directory Access Protocol," by Yeong, Y., T. Howes, and S. Killie, 1995). LDAP servers communicate through referrals — that is, a directory receiving a request for information it does not have will query the tables of remote directories. If it finds a directory with the required entry, it sends a referral to the requesting directory.

LDAP provides a standard format to access the certificate directories. These directories are stored on network LDAP servers and provide public keys and corresponding X.509 certificates for the enterprise. A directory contains information such as individuals' names, addresses, phone numbers, and public key certificates. The standards under X.500 define the protocols and information models for computer directory services that are independent of the platforms and other related entities. LDAP servers are subject to attacks that affect availability and integrity. For example, denial-of-service attacks on an LDAP server could prevent access to the CRLs and thus permit the use of a revoked certificate.

The DAP protocol in X.500 was unwieldy and led to most client implementations using LDAP. LDAP version 3 provides extensions that offer shadowing and chaining capabilities.

X.509 Certificates

The original X.509 certificate (CCITT, *The Directory-Authentication Framework*, Recommendation X.509, 1988) was developed to provide the authentication foundation for the X.500 directory. Since then, a version 2 and a version 3 have been developed. Version 2 of the X.509 certificate addresses the reuse of names, and version 3 provides for certificate extensions to the core certificate fields. These extensions can be used as needed by different users and different applications. A version of X.509 that takes into account the requirements of the Internet was published by the IETF (see RFC 2459, "Internet X.509 Public Key Infrastructure Certificate and CRL Profile," by Housley, R., W. Ford, W. Polk, and D. Solo, 1999).

The Consultation Committee, International Telephone and Telegraph, International Telecommunications Union (CCITT-ITU)/International Organization for Standardization (ISO) has defined the basic format of an X.509 certificate. This structure is outlined in Figure 6-6.

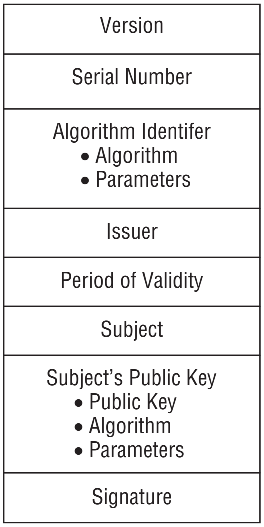

Figure 6-6: The CCITT-ITU/ ISO X.509 certificate format

If version 3 certificates are used, the optional extensions field can be used. It comes before the signature field components in the certificate. Some typical extensions are the entity's name and supporting identity information, the attributes of the key, certificate policy information, and the type of the subject. The digital signature serves as a tamper-evident envelope.

Some of the different types of certificates that are issued include the following:

- **CA certificates** — Issued to CAs, these certificates contain the public keys used to verify digital signatures on CRLs and certificates.

- **End entity certificates** — Issued to entities that are not CAs, these certificates contain the public keys that are needed by the certificate's user in order to perform key management or verify a digital signature.
- **Self-issued certificates** — These certificates are issued by an entity to itself to establish points of trust and to distribute a new signing public key.
- **Rollover certificates** — These certificates are issued by a CA to transition from an old public key to a new one.

Certificate Revocation Lists

Users check the certificate revocation list (CRL) to determine whether a digital certificate has been revoked. They check for the serial number of the signature. The CA signs the CRL for integrity and authentication purposes. A CRL is shown in Figure 6-7 for an X.509 version 2 certificate.

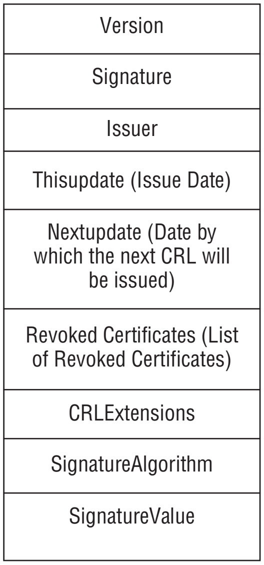
Version
Signature
Issuer
Thisupdate (Issue Date)
Nextupdate (Date by which the next CRL will be issued)
Revoked Certificates (List of Revoked Certificates)
CRLExtensions
SignatureAlgorithm
SignatureValue

Figure 6-7: CRL format (version 2)

The CA usually generates the CRLs for its population. If the CA generates the CRLs for its entire population, the CRL is called a *full CRL*.

Key Management

Obviously, when dealing with encryption keys, the same precautions must be used as with physical keys to secure the areas or the combinations to the safes. The following sections describe the components of key management.

Key Distribution

Because distributing secret keys in symmetric key encryption poses a problem, secret keys can be distributed using asymmetric key cryptosystems. Other means of distributing secret keys include face-to-face meetings to exchange keys, sending the keys by secure messenger, or some other secure alternate channel. Another method is to encrypt the secret key with another key, called a *key encryption key*, and send the encrypted secret key to the intended receiver. These key encryption keys can be distributed manually, but they need not be distributed often. The X9.17 Standard (ANSI X9.17 [Revised], "American National Standard for Financial Institution Key Management [Wholesale]," American Bankers Association, 1985) specifies key encryption keys as well as data keys for encrypting the plain-text messages.

Key distribution can also be accomplished by splitting the keys into different parts and sending each part by a different medium.

In large networks, key distribution can become a serious problem because in an N-person network, the total number of key exchanges is $N(N–1)/2$. Using public key cryptography or the creation and exchange of session keys that are valid only for a particular session and length of time are useful mechanisms for managing the key distribution problem.

Keys can be *updated* by generating a new key from an old key. If, for example, Alice and Bob share a secret key, they can apply the same transformation function (a hash algorithm) to their common secret key and obtain a new secret key.

Key Revocation

A digital certificate contains a timestamp or period for which the certificate is valid. Also, if a key is compromised or must be made invalid because of business- or personnel-related issues, it must be revoked. The CA maintains a CRL of all invalid certificates. Users should regularly examine this list.

Key Recovery

A system must be put in place to decrypt critical data if the encryption key is lost or forgotten. One method is *key escrow*. In this system, the key is subdivided into different parts, each of which is encrypted and then sent to a different trusted individual in an organization. Keys can also be escrowed onto smart cards.

Key Renewal

Obviously, the longer a secret key is used without changing it, the more it is subject to compromise. The frequency with which you change the key is a direct function of the value of the data being encrypted and transmitted. Also, if the same secret key is used to encrypt valuable data over a relatively long period of time, you risk compromising a larger volume of data when the key is broken. Another important concern if the key is not changed frequently is that an attacker can intercept and change messages and then send different messages to the receiver.

Key encryption keys, because they are not used as often as encryption keys, provide some protection against attacks. Typically, private keys used for digital signatures are not frequently changed and may be kept for years.

Key Destruction

Keys that have been in use for long periods of time and are replaced by others should be destroyed. If the keys are compromised, older messages sent with those keys can be read.

Keys that are stored on disks, EEPROMS, or flash memory should be overwritten numerous times. One can also destroy the disks by shredding and burning them. However, in some cases, it is possible to recover data from disks that were put into a fire. Any hardware device storing the key, such as an EPROM, should also be physically destroyed.

Older keys stored by the operating system in various locations in memory must also be searched out and destroyed.

Multiple Keys

Usually, an individual has more than one public/private key pair. The keys may be of different sizes for different levels of security. A larger key size may be used for digitally signing documents, whereas a smaller key size may be used for encryption. A person may also have multiple roles or responsibilities wherein they want to sign messages with a different signature. One key pair may be used for business matters, another for personal use, and another for some other activity, such as being a school board member.

Distributed versus Centralized Key Management

A CA is a form of centralized key management. It is a central location that issues certificates and maintains CRLs. An alternative is *distributed key management*, in which a "chain of trust" or "web of trust" is set up among users who know each other. Because they know each other, they can trust that each one's public key is valid. Some of these users may know other users and can thus verify their public key. The chain spreads outward from the original group. This arrangement results in an informal verification procedure that is based on people knowing and trusting each other.

Further Considerations

Additional mechanisms that can be applied to network connections to provide for secure cloud communications include the following:

- Layered security
- Segmentation of virtual local area networks and applications

- Clustering of DNS servers for fault tolerance
- Load balancers
- Firewalls

Microarchitectures

The term *computer architecture* refers to the organization of the fundamental elements composing the computer. From another perspective, it refers to the view a programmer has of the computing system when viewed through its instruction set. The main hardware components of a digital computer are the central processing unit (CPU), memory, and input/output devices. A basic CPU of a general-purpose digital computer consists of an arithmetic logic unit (ALU), control logic, one or more accumulators, multiple general-purpose registers, an instruction register, a program counter, and some on-chip local memory. The ALU performs arithmetic and logical operations on the binary words of the computer.

The design elements of the microprocessor hardware and firmware that provide for the implementation of the higher-level architecture are referred to as *microarchitecture*. As an example, a microarchitecture design might incorporate the following:

- **Pipelining** — Increases the performance of a computer by overlapping the steps of different instructions. For example, if the instruction cycle is divided into three parts — fetch, decode, and execute — instructions can be overlapped (as shown in Figure 6-8) to increase the execution speed of the instructions.

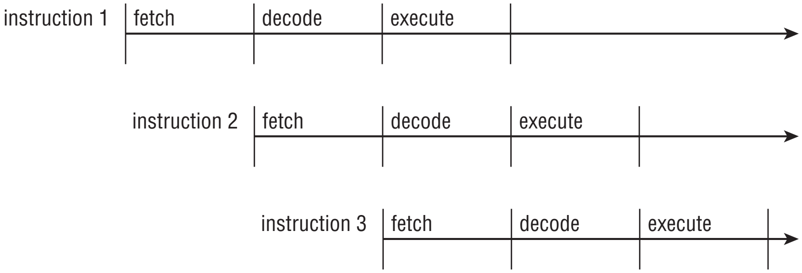

Figure 6-8: Instruction pipelining

- **Superscalar processor** — A processor that enables the concurrent execution of multiple instructions in both the same pipeline stage as well as different pipeline stages.
- **Very-long instruction word (VLIW) processor** — A processor in which a single instruction specifies more than one concurrent operation. For

example, the instruction might specify and concurrently execute two operations in one instruction. VLIW processing is illustrated in Figure 6-9.

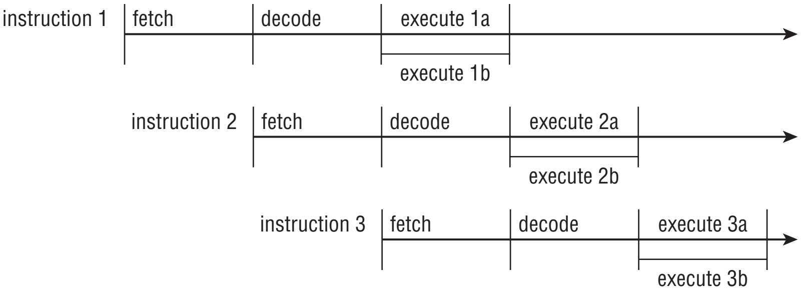

Figure 6-9: VLIW processing

- **Multi-programming** — Executes two or more programs simultaneously on a single processor (CPU) by alternating execution among the programs.
- **Multi-tasking** — Executes two or more subprograms or tasks at the same time on a single processor (CPU) by alternating execution among the tasks.
- **Multi-processing** — Executes two or more programs at the same time on multiple processors. In symmetric multi-processing, the processors share the same operating system, memory, and data paths, while in massively parallel multi-processing, large numbers of processors are used. In this architecture, each processor has its own memory and operating system but communicates and cooperates with all the other processors.
- **Multi-threading** — Concurrent tasks that share resources and run inside a process. In a multi-processing system, threads run in parallel.
- **Simultaneous multi-threading (SMT)** — Multiple threads running on a single core. SMT is especially valuable in enhancing the speed of RSA encryption computations that are widely used in securing cloud transactions.

Microarchitectures can be designed as hardware accelerators for functions such as encryption, arithmetic, and secure Web transactions to support cloud computing.

Identity Management and Access Control

Identity management and access control are fundamental functions required for secure cloud computing. The simplest form of identity management is logging on to a computer system with a user ID and password. However, true identity management, such as is required for cloud computing, requires more robust authentication, authorization, and access control. It should determine what resources are authorized to be accessed by a user or process by using technology such as biometrics or smart cards, and determine when a resource has been accessed by unauthorized entities.

Identity Management

Identification and authentication are the keystones of most access control systems. Identification is the act of a user professing an identity to a system, usually in the form of a username or user logon ID to the system. Identification establishes user accountability for the actions on the system. User IDs should be unique and not shared among different individuals. In many large organizations, user IDs follow set standards, such as first initial followed by last name, and so on. In order to enhance security and reduce the amount of information available to an attacker, an ID should not reflect the user's job title or function.

Authentication is verification that the user's claimed identity is valid, and it is usually implemented through a user password at logon. Authentication is based on the following three factor types:

- **Type 1** — Something you know, such as a personal identification number (PIN) or password
- **Type 2** — Something you have, such as an ATM card or smart card
- **Type 3** — Something you are (physically), such as a fingerprint or retina scan

Sometimes a fourth factor, something you do, is added to this list. Something you do might be typing your name or other phrases on a keyboard. Conversely, something you do can be considered something you are.

Two-factor authentication requires two of the three factors to be used in the authentication process. For example, withdrawing funds from an ATM machine requires two-factor authentication in the form of the ATM card (something you have) and a PIN number (something you know).

Passwords

Because passwords can be compromised, they must be protected. In the ideal case, a password should be used only once. This "one-time password," or OTP, provides maximum security because a new password is required for each new logon. A password that is the same for each logon is called a *static password*. A password that changes with each logon is termed a *dynamic password*. The changing of passwords can also fall between these two extremes. Passwords can be required to change monthly, quarterly, or at other intervals, depending on the criticality of the information needing protection and the password's frequency of use. Obviously, the more times a password is used, the more chance there is of it being compromised. A *passphrase* is a sequence of characters that is usually longer than the allotted number for a password. The passphrase is converted into a virtual password by the system.

In all these schemes, a front-end authentication device or a back-end authentication server, which services multiple workstations or the host, can perform the authentication.

Passwords can be provided by a number of devices, including tokens, memory cards, and smart cards.

Tokens

Tokens, in the form of small, hand-held devices, are used to provide passwords. The following are the four basic types of tokens:

- Static password tokens
  1. Owners authenticate themselves to the token by typing in a secret password.
  2. If the password is correct, the token authenticates the owner to an information system.
- Synchronous dynamic password tokens, clock-based
  1. The token generates a new, unique password value at fixed time intervals that is synchronized with the same password on the authentication server (this password is the time of day encrypted with a secret key).
  2. The unique password is entered into a system or workstation along with an owner's PIN.
  3. The authentication entity in a system or workstation knows an owner's secret key and PIN, and the entity verifies that the entered password is valid and that it was entered during the valid time window.
- Synchronous dynamic password tokens, counter-based
  1. The token increments a counter value that is synchronized with a counter in the authentication server.
  2. The counter value is encrypted with the user's secret key inside the token and this value is the unique password that is entered into the system authentication server.

3. The authentication entity in the system or workstation knows the user's secret key and the entity verifies that the entered password is valid by performing the same encryption on its identical counter value.

- Asynchronous tokens, challenge-response
  1. A workstation or system generates a random challenge string, and the owner enters the string into the token along with the proper PIN.
  2. The token performs a calculation on the string using the PIN and generates a response value that is then entered into the workstation or system.
  3. The authentication mechanism in the workstation or system performs the same calculation as the token using the owner's PIN and challenge string and compares the result with the value entered by the owner. If the results match, the owner is authenticated.

Memory Cards

Memory cards provide nonvolatile storage of information, but they do not have any processing capability. A memory card stores encrypted passwords and other related identifying information. A telephone calling card and an ATM card are examples of memory cards.

Smart Cards

Smart cards provide even more capability than memory cards by incorporating additional processing power on the cards. These credit-card-size devices comprise microprocessor and memory and are used to store digital signatures, private keys, passwords, and other personal information.

Biometrics

An alternative to using passwords for authentication in logical or technical access control is *biometrics*. Biometrics is based on the Type 3 authentication mechanism — something you are. Biometrics is defined as an automated means of identifying or authenticating the identity of a living person based on physiological or behavioral characteristics. In biometrics, identification is a one-to-many search of an individual's characteristics from a database of stored images. Authentication is a one-to-one search to verify a claim to an identity made by a person. Biometrics is used for identification in physical controls and for authentication in logical controls.

There are three main performance measures in biometrics:

- **False rejection rate (FRR) or Type I Error** — The percentage of valid subjects that are falsely rejected.
- **False acceptance rate (FAR) or Type II Error** — The percentage of invalid subjects that are falsely accepted.
- **Crossover error rate (CER)** — The percentage at which the FRR equals the FAR. The smaller the CER, the better the device is performing.

In addition to the accuracy of the biometric systems, other factors must be considered, including enrollment time, throughput rate, and acceptability. *Enrollment time* is the time that it takes to initially register with a system by providing samples of the biometric characteristic to be evaluated. An acceptable enrollment time is around two minutes. For example, in fingerprint systems the actual fingerprint is stored and requires approximately 250KB per finger for a high-quality image. This level of information is required for one-to-many searches in forensics applications on very large databases.

In finger-scan technology, a full fingerprint is not stored; rather, the features extracted from this fingerprint are stored by using a small template that requires approximately 500 to 1,000 bytes of storage. The original fingerprint cannot be reconstructed from this template. Finger-scan technology is used for one-to-one verification by using smaller databases. Updates of the enrollment information might be required because some biometric characteristics, such as voice and signature, might change over time.

The *throughput rate* is the rate at which the system processes and identifies or authenticates individuals. Acceptable throughput rates are in the range of 10 subjects per minute. *Acceptability* refers to considerations of privacy, invasiveness, and psychological and physical comfort when using the system. For example, a concern with retina scanning systems might be the exchange of body fluids on the eyepiece. Another concern would be disclosing the retinal pattern, which could reveal changes in a person's health, such as diabetes or high blood pressure.

Collected biometric images are stored in an area referred to as a *corpus*. The corpus is stored in a database of images. Potential sources of error include the corruption of images during collection, and mislabeling or other transcription problems associated with the database. Therefore, the image collection process and storage must be performed carefully with constant checking. These images are collected during the enrollment process and thus are critical to the correct operation of the biometric device.

The following are typical biometric characteristics that are used to uniquely authenticate an individual's identity:

- **Fingerprints** — Fingerprint characteristics are captured and stored. Typical CERs are 4–5%.

- **Retina scans** — The eye is placed approximately two inches from a camera and an invisible light source scans the retina for blood vessel patterns. CERs are approximately 1.4%.
- **Iris scans** — A video camera remotely captures iris patterns and characteristics. CER values are around 0.5%.
- **Hand geometry** — Cameras capture three-dimensional hand characteristics. CERs are approximately 2%.
- **Voice** — Sensors capture voice characteristics, including throat vibrations and air pressure, when the subject speaks a phrase. CERs are in the range of 10%.
- **Handwritten signature dynamics** — The signing characteristics of an individual making a signature are captured and recorded. Typical characteristics including writing pressure and pen direction. CERs are not published at this time.

Other types of biometric characteristics include facial and palm scans.

Implementing Identity Management

Realizing effective identity management requires a high-level corporate commitment and dedication of sufficient resources to accomplish the task. Typical undertakings in putting identity management in place include the following:

- Establishing a database of identities and credentials
- Managing users' access rights
- Enforcing security policy
- Developing the capability to create and modify accounts
- Setting up monitoring of resource accesses
- Installing a procedure for removing access rights
- Providing training in proper procedures

An identity management effort can be supported by software that automates many of the required tasks.

The Open Group and the World Wide Web Consortium (W3C) are working toward a standard for a global identity management system that would be interoperable, provide for privacy, implement accountability, and be portable. Identity management is also addressed by the XML-based eXtensible Name Service (XNS) open protocol for universal addressing. XNS provides the following capabilities:

- A permanent identification address for a container of an individual's personal data and contact information
- Means to verify whether an individual's contact information is valid
- A platform for negotiating the exchange of information among different entities

Access Control

Access control is intrinsically tied to identity management and is necessary to preserve the confidentiality, integrity, and availability of cloud data.

These and other related objectives flow from the organizational security policy. This policy is a high-level statement of management intent regarding the control of access to information and the personnel who are authorized to receive that information.

Three things that must be considered for the planning and implementation of access control mechanisms are threats to the system, the system's vulnerability to these threats, and the risk that the threats might materialize. These concepts are defined as follows:

- **Threat** — An event or activity that has the potential to cause harm to the information systems or networks
- **Vulnerability** — A weakness or lack of a safeguard that can be exploited by a threat, causing harm to the information systems or networks
- **Risk** — The potential for harm or loss to an information system or network; the probability that a threat will materialize

Controls

Controls are implemented to mitigate risk and reduce the potential for loss. Two important control concepts are *separation of duties* and the principle of *least privilege.* Separation of duties requires an activity or process to be performed by two or more entities for successful completion. Thus, the only way that a security policy can be violated is if there is collusion among the entities. For example, in a financial environment, the person requesting that a check be issued for payment should not also be the person who has authority to sign the check. Least privilege means that the entity that has a task to perform should be provided with the minimum resources and privileges required to complete the task for the minimum necessary period of time.

Control measures can be administrative, logical (also called technical), and physical in their implementation.

- Administrative controls include policies and procedures, security awareness training, background checks, work habit checks, a review of vacation history, and increased supervision.
- Logical or technical controls involve the restriction of access to systems and the protection of information. Examples of these types of controls are encryption, smart cards, access control lists, and transmission protocols.
- Physical controls incorporate guards and building security in general, such as the locking of doors, the securing of server rooms or laptops, the protection of cables, the separation of duties, and the backing up of files.

Controls provide accountability for individuals who are accessing sensitive information in a cloud environment. This accountability is accomplished through access control mechanisms that require identification and authentication, and through the audit function. These controls must be in accordance with and accurately represent the organization's security policy. Assurance procedures ensure that the control mechanisms correctly implement the security policy for the entire life cycle of a cloud information system.

In general, a group of processes that share access to the same resources is called a *protection domain,* and the memory space of these processes is isolated from other running processes.

Models for Controlling Access

Controlling access by a subject (an active entity such as an individual or process) to an object (a passive entity such as a file) involves setting up access rules. These rules can be classified into three categories or models.

Mandatory Access Control

The authorization of a subject's access to an object depends upon labels, which indicate the subject's *clearance*, and the *classification or sensitivity* of the object. For example, the military classifies documents as unclassified, confidential, secret, and top secret. Similarly, an individual can receive a clearance of confidential, secret, or top secret and can have access to documents classified at or below his or her specified clearance level. Thus, an individual with a clearance of "secret" can have access to secret and confidential documents with a restriction. This restriction is that the individual must have a *need to know* relative to the classified documents involved. Therefore, the documents must be necessary for that individual to complete an assigned task. Even if the individual is cleared for a classification level of information, the individual should not access the information unless there is a need to know. *Rule-based access control* is a type of mandatory access control because rules determine this access (such as the correspondence of clearance labels to classification labels), rather than the identity of the subjects and objects alone.

Discretionary Access Control

With discretionary access control, the subject has authority, within certain limitations, to specify what objects are accessible. For example, access control lists (ACLs) can be used. An access control list is a list denoting which users have what privileges to a particular resource. For example, a *tabular listing* would show the subjects or users who have access to the object, e.g., file X, and what privileges they have with respect to that file.

An *access control triple* consists of the user, program, and file, with the corresponding access privileges noted for each user. This type of access control is used in local, dynamic situations in which the subjects must have the discretion to specify what resources certain users are permitted to access. When a user within certain limitations has the right to alter the access control to certain objects, this is termed a *user-directed discretionary access control*. An identity-based access control is a type of discretionary access control based on an individual's identity. In some instances, a hybrid approach is used, which combines the features of user-based and identity-based discretionary access control.

Nondiscretionary Access Control

A central authority determines which subjects can have access to certain objects based on the organizational security policy. The access controls might be based on the individual's role in the organization (role-based) or the subject's responsibilities and duties (task-based). In an organization with frequent personnel changes, nondiscretionary access control is useful because the access controls are based on the individual's role or title within the organization. Therefore, these access controls don't need to be changed whenever a new person assumes that role.

Access control can also be characterized as *context-dependent* or *content-dependent*. Context-dependent access control is a function of factors such as location, time of day, and previous access history. It is concerned with the environment or context of the data. In content-dependent access control, access is determined by the information contained in the item being accessed.

Single Sign-On (SSO)

Single sign-on (SSO) addresses the cumbersome situation of logging on multiple times to access different resources. When users must remember numerous passwords and IDs, they might take shortcuts in creating them that could leave them open to exploitation. In SSO, a user provides one ID and password per work session and is automatically logged on to all the required applications. For SSO security, the passwords should not be stored or transmitted in the clear. SSO applications can run either on a user's workstation or on authentication servers. The advantages of SSO include having the ability to use stronger passwords, easier administration of changing or deleting the passwords, and less time to access resources. The major disadvantage of many SSO implementations is that once users obtain access to the system through the initial logon, they can freely roam the network resources without any restrictions.

Authentication mechanisms include items such as smart cards and magnetic badges. Strict controls must be enforced to prevent a user from changing configurations that another authority sets.

SSO can be implemented by using scripts that replay the users' multiple logins or by using authentication servers to verify a user's identity, and encrypted authentication tickets to permit access to system services.

Enterprise access management (EAM) provides access control management services to Web-based enterprise systems that include SSO. SSO can be provided in a number of ways. For example, SSO can be implemented on Web applications residing on different servers in the same domain by using nonpersistent, encrypted cookies on the client interface. This task is accomplished by providing a cookie to each application that the user wishes to access. Another solution is to build a secure credential for each user on a reverse proxy that is situated in front of the Web server. The credential is then presented each time a user attempts to access protected Web applications.

Autonomic Security

Autonomic computing refers to a self-managing computing model in which computer systems reconfigure themselves in response to changing conditions and are self-healing. The promise of autonomic computing will take a number of years to fully materialize, but it offers capabilities that can improve the security of information systems and cloud computing in particular. The ability of autonomic systems to collect and interpret data and recommend or implement solutions can go a long way toward enhancing security and providing for recovery from harmful events.

Autonomic Systems

Autonomic systems are based on the human autonomic nervous system, which is self-managing, monitors changes that affect the body, and maintains internal balances. Therefore, an autonomic computing system has the goal of performing self-management to maintain correct operations despite perturbations to the system. Such a system requires sensory inputs, decision-making capability, and the ability to implement remedial activities to maintain an equilibrium state of normal operation. Examples of events that would have to be handled autonomously include the following:

- Malicious attacks
- Hardware or software faults
- Excessive CPU utilization
- Power failures
- Organizational policies
- Inadvertent operator errors
- Interaction with other systems
- Software updates

IBM introduced the concept of autonomic computing and its eight defining characteristics[5] as follows:

- **Self-awareness** — An autonomic application/system "knows itself" and is aware of its state and its behaviors.
- **Self-configuring** — An autonomic application/system should be able configure and reconfigure itself under varying and unpredictable conditions.
- **Self-optimizing** — An autonomic application/system should be able to detect sub-optimal behaviors and optimize itself to improve its execution.
- **Self-healing** — An autonomic application/system should be able to detect and recover from potential problems and continue to function smoothly.
- **Self-protecting** — An autonomic application/system should be capable of detecting and protecting its resources from both internal and external attack and maintaining overall system security and integrity.
- **Context-aware** — An autonomic application/system should be aware of its execution environment and be able to react to changes in the environment.
- **Open** — An autonomic application/system must function in a heterogeneous world and should be portable across multiple hardware and software architectures. Consequently, it must be built on standard and open protocols and interfaces.
- **Anticipatory** — An autonomic application/system should be able to anticipate, to the extent possible, its needs and behaviors and those of its context, and be able to manage itself proactively.

The underlying concept of autonomic systems is self-management, whereby a computational system maintains proper operation in the face of changing external and internal conditions, evaluates the necessity for upgrades, installs software, conducts regression testing, performs performance tuning of middleware, and detects and corrects problem situations in general.

Autonomic Protection

Autonomic self-protection involves detecting a harmful situation and taking actions that will mitigate the situation. These systems will also be designed to predict problems from analysis of sensory inputs and initiate corrective measures.

An autonomous system security response is based on network knowledge, capabilities of connected resources, information aggregation, the complexity of the situation, and the impact on affected applications.

The decision-making element of autonomic computing, taking into account the current security posture and security context of the system to be protected, can take actions such as changing the strength of required authentications or modifying encryption keys. The security context is derived from information acquired from network and system supervising elements and then collected into a higher-level representation of the system security status.

An oft overlooked aspect of autonomic systems is that security vulnerabilities can be introduced by configuration changes and additional autonomous activities that are intended to address other computational areas.

Autonomous protection systems should, therefore, adhere to the following guidelines:

- Minimize overhead requirements.
- Be consistent with security policies.
- Optimize security-related parameters.
- Minimize impact on performance.
- Minimize potential for introducing new vulnerabilities.
- Conduct regression analysis and return to previous software versions if problems are introduced by changes.
- Ensure that reconfiguration processes are secure.

Autonomic Self-Healing

The process of diagnosing and repairing failures in IT systems can be difficult, time consuming, and usually requires intensive labor effort. Autonomic self-healing systems can provide the capability to detect and repair software problems and identify hardware faults without manual intervention.

The autonomic process would obtain logged and monitored information and perform an analysis to diagnose the problem area. This procedure is usually conducted by an autonomic manager that controls computing resource elements with well-defined interfaces that support the diagnostic and mitigation actions. The managed elements control their internal states and have defined performance characteristics and relationships with other computational elements.

The objective of the autonomous self-healing process is to keep the elements operating according to their design specifications.

Summary

Cloud computing security architecture is a critical element in establishing trust in the cloud computing paradigm. Confidence in using the cloud depends on trusted computing mechanisms, robust identity management and access control techniques, providing a secure execution environment, securing cloud communications, and supporting microarchitectures.

Autonomic computing can employ self-management, self-healing, and self-protection techniques to make cloud computing a more reliable, secure, and safe choice for the growing requirements for processing and storing large amounts of information in a cost-effective manner.

In Chapter 7, cloud computing life cycle issues are detailed, including responding to incidents, discussing cloud computing encryption issues, and cloud computing virtual machine retirement considerations.

Notes

1. NIST Special Publication 800-30, "Risk Management Guide for Information Technology Systems," July 2002.
2. NIST Special Publication 800-14, "Generally Accepted Principles and Practices for Securing Information Technology Systems," September 1996.
3. NIST Special Publication 800-14, "Generally Accepted Principles and Practices for Securing Information Technology Systems," September 1996.
4. Berger, S., Cáceres, R., Goldman, K.A., et al., "vTPM: Virtualizing the Trusted Platform Module," in Proceedings of the 15th USENIX Security Symposium, Vancouver, B.C., 2006.
5. Horn, P., "Autonomic Computing: IBM's Perspective on the State of Information Technology," `http://www.research.ibm.com/autonomic/`, IBM Corp., October 2001.