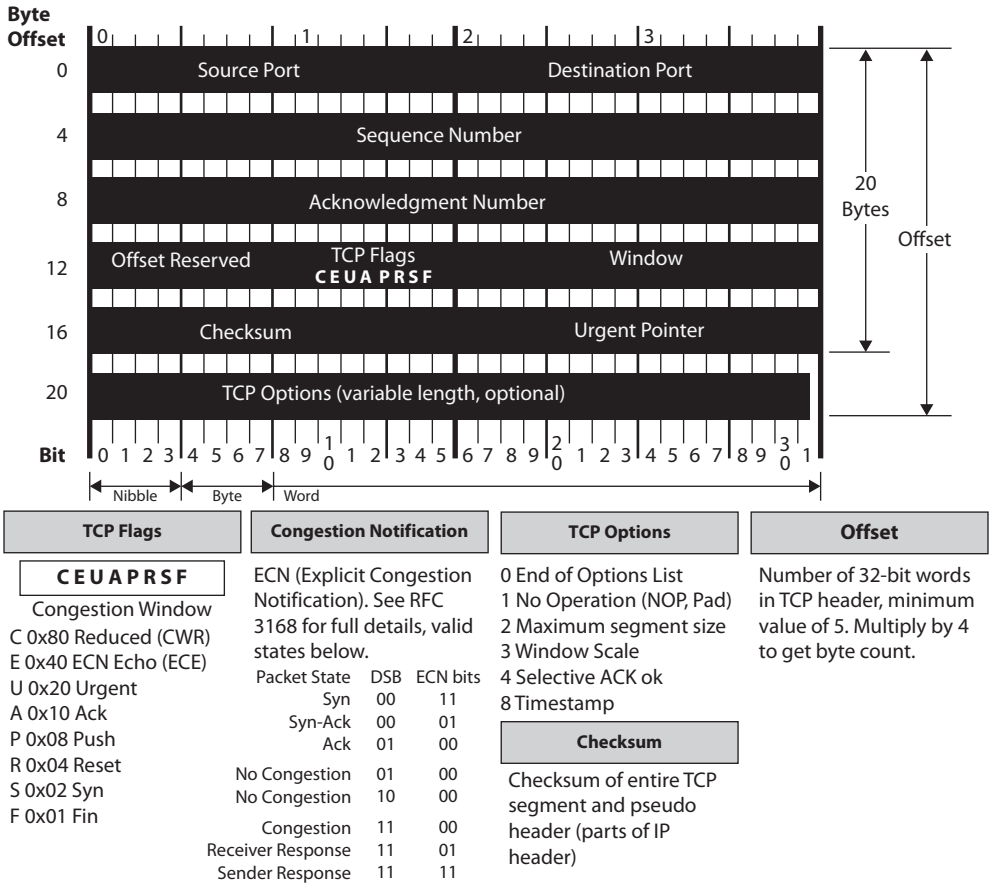


Figure 21-5 TCP header

how much data to send at a time (window size), how potential transmission errors will be identified (CRC values), and so forth. Figure 21-5 shows all of the values that make up a TCP header. So, a lot of information is going back and forth between your systems just in this one protocol—TCP. There are other protocols that are involved with networking that a stateful firewall has to be aware of and keep track of.

So "keeping state of a connection" means to keep a scorecard of all the various protocol header values as packets go back and forth between systems. The values not only have to be correct—they have to happen in the right sequence. For example, if a stateful firewall receives a packet that has all TCP flag values turned to 1, something malicious is taking place. Under no circumstances during a legitimate TCP connection should all of these values be turned on like this. Attackers send packets with all of these values turned to 1 with the hopes that the firewall does not understand or check these values and just forwards the packets onto the target system.

In another situation, if Gwen's system sends your system a SYN/ACK packet and your system did not first send a SYN packet to Gwen's system, this, too, is against the protocol rules. The protocol communication steps have to follow the proper sequence. Attackers send SYN/ACK packets to target systems in an attempt to get the firewall to interpret this as an already established connection and just allow the packets to go to the destination system without inspection. A stateful firewall will not be fooled by such actions because it keeps track of each step of the communication. It knows how protocols are supposed to work, and if something is out of order (incorrect flag values, incorrect sequence, etc.), it does not allow the traffic to pass through.

When a connection begins between two systems, the firewall investigates *all* elements of the packet (all headers, payload, and trailers). All of the necessary information about the specific connection is stored in the state table (source and destination IP addresses, source and destination ports, protocol type, header flags, sequence numbers, timestamps, etc.). Once the initial packets go through this in-depth inspection and everything is deemed safe, the firewall then just reviews the network and transport header portions for the rest of the session. The values of each header for each packet are compared to the values in the current state table, and the table is updated to reflect the progression of the communication process. Scaling down the inspection of the full packet to just the headers for each packet is done to increase performance.

TCP is considered a connection-oriented protocol, and the various steps and states this protocol operates within are very well defined. A connection progresses through a series of states during its lifetime. The states are LISTEN, SYN-SENT, SYN-RECEIVED, ESTABLISHED, FIN-WAIT-1, FIN-WAIT-2, CLOSE-WAIT, CLOSING, LAST-ACK, TIME-WAIT, and the fictional state CLOSED. A stateful firewall keeps track of each of these states for each packet that passes through, along with the corresponding acknowledgment and sequence numbers. If the acknowledgment and/or sequence numbers are out of order, this could imply that a replay attack is underway, and the firewall will protect the internal systems from this activity.

Nothing is ever simple in life, including the standardization of network protocol communication. While the previous statements are true pertaining to the states of a TCP connection, in some situations an application layer protocol has to change these basic steps. For example, FTP uses an unusual communication exchange when initializing its data channel compared to all of the other application layer protocols. FTP basically sets up two sessions just for one communication exchange between two computers. The states of the two individual TCP connections that make up an FTP session can be tracked in the normal fashion, but the state of the FTP connection follows different rules. For a stateful device to be able to properly monitor the traffic of an FTP session, it must be able to take into account the way that FTP uses one outbound connection for the control channel and one inbound connection for the data channel. If you were configuring a stateful firewall, you would need to understand the particulars of some specific protocols to ensure that each is being properly inspected and controlled.

Since TCP is a connection-oriented protocol, it has clearly defined states during the connection establishment, maintenance, and tearing-down stages. UDP is a connectionless protocol, which means that none of these steps take place. UDP holds no state, which makes it harder for a stateful firewall to keep track of. For connectionless protocols,

Stateful-Inspection Firewall Characteristics

The following lists some important characteristics of a stateful-inspection firewall:

- Maintains a state table that tracks each and every communication session
- Provides a high degree of security and does not introduce the performance hit that application proxy firewalls introduce
- Is scalable and transparent to users
- Provides data for tracking connectionless protocols such as UDP and ICMP
- Stores and updates the state and context of the data within the packets

a stateful firewall keeps track of source and destination addresses, UDP header values, and some ACL rules. This connection information is also stored in the state table and tracked. Since the protocol does not have a specific tear-down stage, the firewall will just time out the connection after a period of inactivity and remove the data being kept pertaining to that connection from the state table.

An interesting complexity of stateful firewalls and UDP connections is how ICMP comes into play. Since UDP is connectionless, it does not provide a mechanism to allow the receiving computer to tell the sending computer that data is coming too fast. In TCP, the receiving computer can alter the Window value in its header, which tells the sending computer to reduce the amount of data that is being sent. The message is basically, "You are overwhelming me and I cannot process the amount of data you are sending me. Slow down." UDP does not have a Window value in its header, so instead the receiving computer sends an ICMP packet that provides the same function. But now this means that the stateful firewall must keep track of and allow associated ICMP packets with specific UDP connections. If the firewall does not allow the ICMP packets to get to the sending system, the receiving system could get overwhelmed and crash. This is just one example of the complexity that comes into play when a firewall has to do more than just packet filtering. Although stateful inspection provides an extra step of protection, it also adds more complexity because this device must now keep a dynamic state table and remember connections.

Stateful-inspection firewalls, unfortunately, have been the victims of many types of DoS attacks. Several types of attacks are aimed at flooding the state table with bogus information. The state table is a resource, similar to a system's hard drive space, memory, and CPU. When the state table is stuffed full of bogus information, a poorly designed device may either freeze or reboot.

Proxy Firewalls

A *proxy* is a middleman. It intercepts and inspects messages before delivering them to the intended recipients. Suppose you need to give a box and a message to the president of the United States. You couldn't just walk up to the president and hand over these items.

Instead, you would have to go through a middleman, likely a Secret Service agent, who would accept the box and message and thoroughly inspect the box to ensure nothing dangerous is inside. This is what a proxy firewall does—it accepts messages either entering or leaving a network, inspects them for malicious information, and, when it decides the messages are okay, passes the data on to the destination computer.

A *proxy firewall* stands between a trusted network and an untrusted network and makes the connection, each way, on behalf of the source. What is important is that a proxy firewall breaks the communication channel; there is no *direct* connection between the two communicating devices. Where a packet-filtering device just monitors traffic as it is traversing a network connection, a proxy ends the communication session and restarts it on behalf of the sending system. Figure 21-6 illustrates the steps of a proxy-based firewall. Notice that the firewall does not simply apply ACL rules to the traffic; it stops the user connection at the internal interface of the firewall itself and then starts a new session on behalf of this user on the external interface. When the external web server replies to the request, this reply goes to the external interface of the proxy firewall and ends. The proxy firewall examines the reply information and, if it is deemed safe, starts a new session from itself to the internal system. This is just like our analogy of what the Secret Service agent does between you and the president.

A proxy technology can actually work at different layers of a network stack. A proxy-based firewall that works at the lower layers of the OSI model is referred to as a circuit-level proxy. A proxy-based firewall that works at the application layer is, strangely enough, called an application-level proxy.

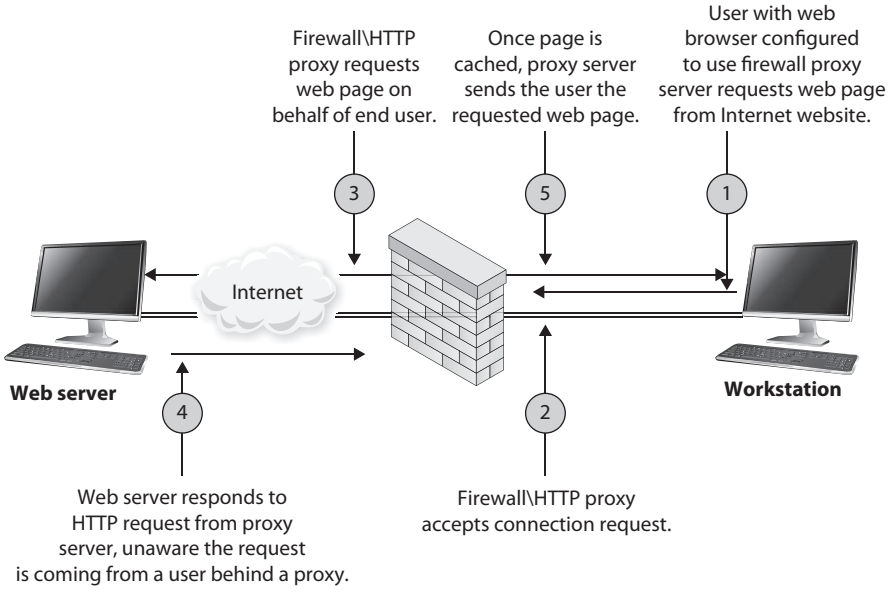


Figure 21-6 Proxy firewall breaks connection

A *circuit-level proxy* creates a connection (circuit) between the two communicating systems. It works at the session layer of the OSI model and monitors traffic from a network-based view. This type of proxy cannot "look into" the contents of a packet; thus, it does not carry out deep-packet inspection. It can only make access decisions based upon protocol header and session information that is available to it. While this means that a circuit-level proxy cannot provide as much protection as an application-level proxy, because it does not have to understand application layer protocols, it is considered application independent. So, it cannot provide the detail-oriented protection that a proxy working at a higher level can, but this allows it to provide a broader range of protection where application layer proxies may not be appropriate or available.

NOTE Traffic sent to the receiving computer through a circuit-level proxy appears to have originated from the firewall instead of the sending system. This is useful for hiding information about the internal computers on the network the firewall is protecting.

Application-level proxies inspect the packet up through the application layer. Where a circuit-level proxy only has insight up to the session layer, an application-level proxy understands the packet as a whole and can make access decisions based on the content of the packets. Application-level proxies understand various services and protocols and the commands that are used by them. An application-level proxy can distinguish between an FTP GET command and an FTP PUT command, for example, and make access decisions based on this granular level of information; on the other hand, packet-filtering firewalls and circuit-level proxies can allow or deny FTP requests only as a whole, not by the commands used within FTP.

An application-level proxy firewall has one proxy per protocol. A computer can have many types of protocols (FTP, NTP, SMTP, HTTP, and so on). Thus, one application-level proxy per protocol is required. This does not mean one proxy firewall per service is required, but rather that one portion of the firewall product is dedicated to understanding how a specific protocol works and how to properly filter it for suspicious data.

Providing application-level proxy protection can be a tricky undertaking. The proxy must totally understand how specific protocols work and what commands within that protocol are legitimate. This is a lot to know and look at during the transmission of data. As an analogy, picture a screening station at an airport that is made up of many employees, all with the job of interviewing people before they are allowed into the airport and onto an airplane. These employees have been trained to ask specific questions and detect suspicious answers and activities, and have the skill set and authority to detain suspicious individuals. Now, suppose each of these employees speaks a different language because the people they interview come from different parts of the world. So, one employee who speaks German could not understand and identify suspicious answers of a person from Italy because they do not speak the same language. This is the same for an application-level proxy firewall. Each proxy is a piece of software that has been designed to understand how a specific protocol "talks" and how to identify suspicious data within a transmission using that protocol.

NOTE If the application-level proxy firewall does not understand a certain protocol or service, it cannot protect this type of communication. In this scenario, a circuit-level proxy is useful because it does not deal with such complex issues. An advantage of a circuit-level proxy is that it can handle a wider variety of protocols and services than an application-level proxy can, but the downfall is that the circuit-level proxy cannot provide the degree of granular control that an application-level proxy provides. Life is just full of compromises.

A circuit-level proxy works similarly to a packet filter in that it makes access decisions based on address, port, and protocol type header values. It looks at the data within the packet header rather than the data at the application layer of the packet. It does not know whether the contents within the packet are safe or unsafe; it only understands the traffic from a network-based view.

An application-level proxy, on the other hand, is dedicated to a particular protocol or service. At least one proxy is used per protocol because one proxy could not properly interpret all the commands of all the protocols coming its way. A circuit-level proxy works at a lower layer of the OSI model and does not require one proxy per protocol because it does not look at such detailed information.

Application-Level Proxy Firewalls

Application-level proxy firewalls, like all technologies, have their pros and cons. It is important to fully understand all characteristics of this type of firewall before purchasing and deploying this type of solution.

Characteristics of application-level proxy firewalls:

- They have extensive logging capabilities due to the firewall being able to examine the entire network packet rather than just the network addresses and ports.
- They are capable of authenticating users directly, as opposed to packet-filtering firewalls and stateful-inspection firewalls, which can usually only carry out system authentication.
- Since they are not simply layer 3 devices, they can address spoofing attacks and other sophisticated attacks.

Disadvantages of using application-level proxy firewalls:

- They are not generally well suited to high-bandwidth or real-time applications.
- They tend to be limited in terms of support for new network applications and protocols.
- They create performance issues because of the necessary per-packet processing requirements.

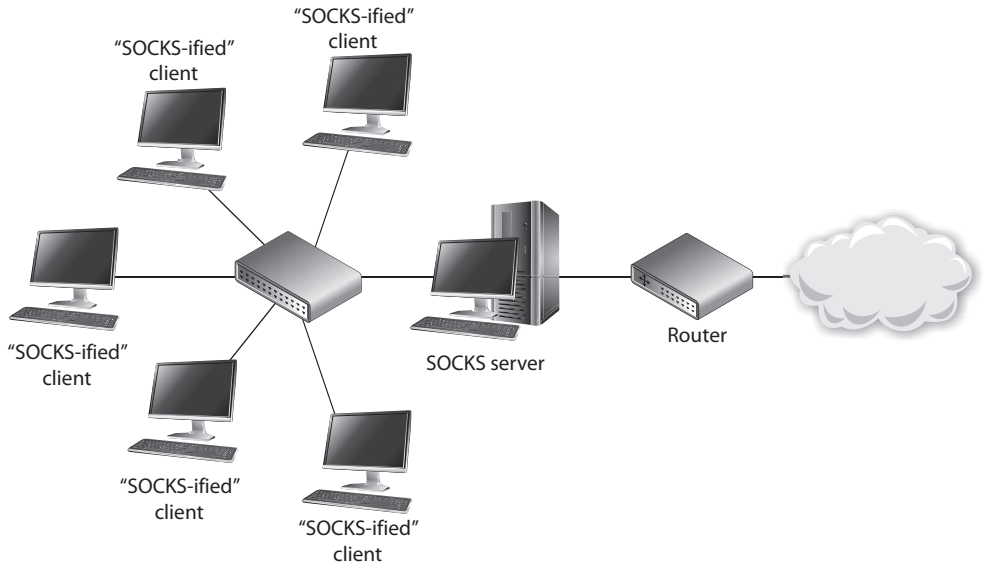


Figure 21-7 Circuit-level proxy firewall

SOCKS is an example of a circuit-level proxy gateway that provides a secure channel between two computers. When a SOCKS-enabled client sends a request to access a computer on the Internet, this request actually goes to the network's SOCKS proxy firewall, as shown in Figure 21-7, which inspects the packets for malicious information and checks its policy rules to see whether this type of connection is allowed. If the packet is acceptable and this type of connection is allowed, the SOCKS firewall sends the message to the destination computer on the Internet. When the computer on the Internet responds, it sends its packets to the SOCKS firewall, which again inspects the data and then passes the packets on to the client computer.

The SOCKS firewall can screen, filter, audit, log, and control data flowing in and out of a protected network. Because of its popularity, many applications and protocols have been configured to work with SOCKS in a manner that takes less configuration on the administrator's part, and various firewall products have integrated SOCKS software to provide circuit-based protection.

NOTE Remember that whether an application- or circuit-level proxy firewall is being used, it is still acting as a proxy. Both types of proxy firewalls deny actual end-to-end connectivity between the source and destination systems. In attempting a remote connection, the client connects to and communicates with the proxy; the proxy, in turn, establishes a connection to the destination system and makes requests to it on the client's behalf. The proxy maintains two independent connections for every one network transmission. It essentially turns a two-party session into a four-party session, with the middle process emulating the two real systems.

Application-Level vs. Circuit-Level Proxy Firewall Characteristics

Characteristics of application-level proxy firewalls:

- Each protocol that is to be monitored must have a unique proxy.
- They provide more protection than circuit-level proxy firewalls.
- They require more processing per packet and thus are slower than circuit-level proxy firewalls.

Characteristics of circuit-level proxy firewalls:

- They do not require a proxy for each and every protocol.
- They do not provide the deep-inspection capabilities of an application-level proxy firewall.
- They provide security for a wider range of protocols.

Next-Generation Firewalls

A *next-generation firewall (NGFW)* combines the best attributes of the previously discussed firewalls, but adds a number of important improvements. Most importantly, it incorporates a signature-based and/or behavioral analysis IPS engine. This means that, in addition to ensuring that the traffic is behaving in accordance with the rules of the applicable protocols, the firewall can look for specific indicators of attack even in otherwise well-behaved traffic. Some of the most advanced NGFWs include features that allow them to share signatures with a cloud-based aggregator so that once a new attack is detected by one firewall, all other firewalls manufactured by that vendor become aware of the attack signature.

Another characteristic of an NGFW is its ability to connect to external data sources such as Active Directory, whitelists, blacklists, and policy servers. This feature allows controls to be defined in one place and pulled by every NGFW on the network, which reduces the chances of inconsistent settings on the various firewalls that typically exist in large networks.

For all their power, NGFWs are not appropriate for every organization. The typical cost of ownership alone tends to make these infeasible for small or even medium-sized networks. Organizations need to ensure that the correct firewall technology is in place to monitor specific network traffic types and protect unique resource types. The firewalls also have to be properly placed; we will cover this topic in the next section.

NOTE Firewall technology has evolved as attack types have evolved. The first-generation firewalls could only monitor network traffic. As attackers moved from just carrying out network-based attacks (DoS, fragmentation, spoofing, etc.) to conducting software-based attacks (buffer overflows, injections, malware, etc.), new generations of firewalls were developed to monitor for these types of attacks.

Firewall Type	OSI Layer	Characteristics
Packet filtering	Network layer	Looks at destination and source addresses, ports, and services requested. Typically routers using ACLs to control and monitor network traffic.
Stateful	Network layer	Looks at the state and context of packets. Keeps track of each conversation using a state table.
Application-level proxy	Application layer	Looks deep into packets and makes granular access control decisions. It requires one proxy per protocol.
Circuit-level proxy	Session layer	Looks only at the header packet information. It protects a wider range of protocols and services than an application-level proxy, but does not provide the detailed level of control available to an application-level proxy.
Next-generation firewall	Multiple layers	Very fast and supportive of high bandwidth. Built-in IPS. Able to connect to external services like Active Directory.

Table 21-1 Comparison of Different Types of Firewalls

Table 21-1 lists the important concepts and characteristics of the firewall types discussed in the preceding sections. Although various firewall products can provide a mix of these services and work at different layers of the OSI model, it is important you understand the basic definitions and functionalities of these firewall types.

Appliances

A firewall may take the form of either software installed on a regular computer using a regular operating system or a dedicated hardware appliance that has its own operating system. The second choice is usually more secure, because the vendor uses a stripped-down version of an operating system (usually Linux or BSD Unix). Operating systems are full of code and functionality that are not necessary for a firewall. This extra complexity opens the doors for vulnerabilities. If a hacker can exploit and bring down a company's firewall, then the company is very exposed and in danger.

In today's jargon, dedicated hardware devices that have stripped-down operating systems and limited and focused software capabilities are called *appliances*. Where an operating system has to provide a vast array of functionality, an appliance provides very focused functionality—as in just being a firewall.

If a software-based firewall is going to run on a regular system, then the unnecessary user accounts should be disabled, unnecessary services deactivated, unused subsystems disabled, unneeded ports closed, and so on. If firewall software is going to run on a regular system and not a dedicated appliance, then the system needs to be fully locked down.

Firewall Architecture

Firewalls can be placed in a number of areas on a network to meet particular needs. They can protect an internal network from an external network and act as a choke point for all traffic. A firewall can be used to segment and partition network sections and enforce access controls between two or more subnets. Firewalls can also be used to provide a DMZ architecture. And as covered in the previous section, the right firewall type needs to be placed in the right location. Organizations have common needs for firewalls; hence, they keep them in similar places on their networks. We will see more on this topic in the following sections.

Dual-Homed Firewall *Dual-homed* refers to a device that has two interfaces: one connected to one network and the other connected to a different network. If firewall software is installed on a dual-homed device—and it usually is—the underlying operating system should have packet forwarding and routing turned off for security reasons. If they are enabled, the computer may not apply the necessary ACLs, rules, or other restrictions required of a firewall. When a packet comes to the external NIC from an untrusted network on a dual-homed firewall and the operating system has forwarding enabled, the operating system forwards the traffic instead of passing it up to the firewall software for inspection.

Many network devices today are *multihomed*, which just means they have several NICs that are used to connect several different networks. Multihomed devices are commonly used to house firewall software, since the job of a firewall is to control the traffic as it goes from one network to another. A common multihomed firewall architecture allows an organization to have several DMZs. One DMZ may hold devices that are shared between organizations in an extranet, another DMZ may house the organization's DNS and mail servers, and yet another DMZ may hold the organization's web servers. Different DMZs are used for two reasons: to control the different traffic types (for example, to ensure HTTP traffic only goes toward the web servers and ensure DNS requests go toward the DNS server), and to ensure that if one system on one DMZ is compromised, the other systems in the rest of the DMZs are not accessible to this attacker.

If a company depends solely upon a multihomed firewall with no redundancy, this system could prove to be a single point of failure. If it goes down, then all traffic flow stops. Some firewall products have embedded redundancy or fault-tolerance capabilities. If a company uses a firewall product that does not have these capabilities, then the network should have redundancy built into it.

Along with potentially being a single point of failure, another security issue that is posed by relying on a single firewall is the lack of defense in depth. If the company depends on just one firewall, no matter what architecture is being used or how many interfaces the device has, there is only one layer of protection. If an attacker can compromise the one firewall, then she can gain direct access to company network resources.

Screened Host A *screened host* is a firewall that communicates directly with a perimeter router and the internal network. Figure 21-8 shows this type of architecture.

Traffic received from the Internet is first filtered via packet filtering on the outer router. The traffic that makes it past this phase is sent to the screened-host firewall, which applies

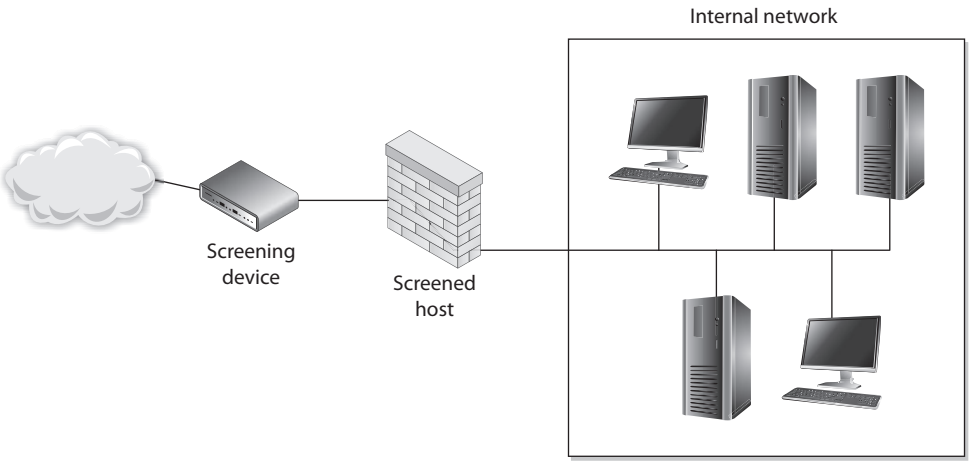


Figure 21-8 A screened host is a firewall that is screened by a router.

more rules to the traffic and drops the denied packets. Then the traffic moves to the internal destination hosts. The screened host (the firewall) is the only device that receives traffic directly from the router. No traffic goes directly from the Internet, through the router, and to the internal network. The screened host is always part of this equation.

If the firewall is an application-based system, protection is provided at the network layer by the router through packet filtering, and at the application layer by the firewall. This arrangement offers a high degree of security, because for an attacker to be successful, she would have to compromise two systems.

What does the word "screening" mean in this context? As shown in Figure 21-8, the router is a screening device and the firewall is the screened host. This just means there is a layer that scans the traffic and gets rid of a lot of the "junk" before the traffic is directed toward the firewall. A screened host is different from a screened subnet, which is described next.

Screened Subnet A *screened-subnet* architecture adds another layer of security to the screened-host architecture. The external firewall screens the traffic entering the DMZ network. However, instead of the firewall then redirecting the traffic to the internal network, an interior firewall also filters the traffic. The use of these two physical firewalls creates a DMZ.

In an environment with only a screened host, if an attacker successfully breaks through the firewall, nothing lies in her way to prevent her from having full access to the internal network. In an environment using a screened subnet, the attacker would have to hack through another firewall to gain access. In this layered approach to security, the more layers provided, the better the protection. Figure 21-9 shows a simple example of a screened subnet.

The examples shown in the figures are simple in nature. Often, more complex networks and DMZs are implemented in real-world systems. Figures 21-10 and 21-11 show some other possible architectures of screened subnets and their configurations.

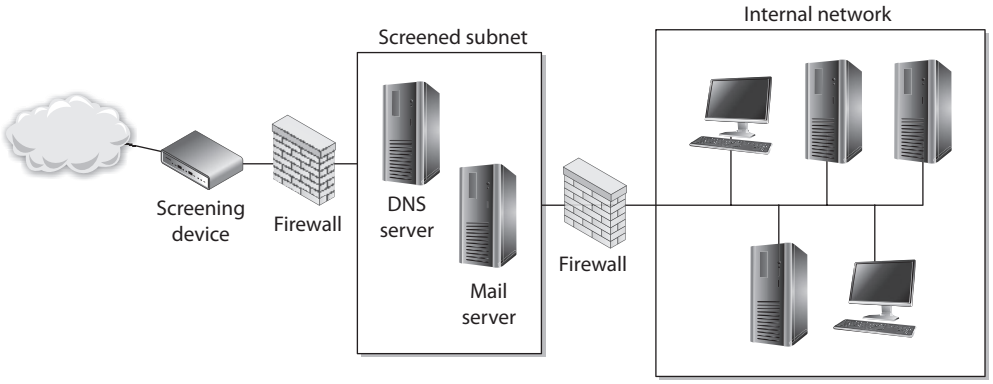


Figure 21-9 With a screened subnet, two firewalls are used to create a DMZ.

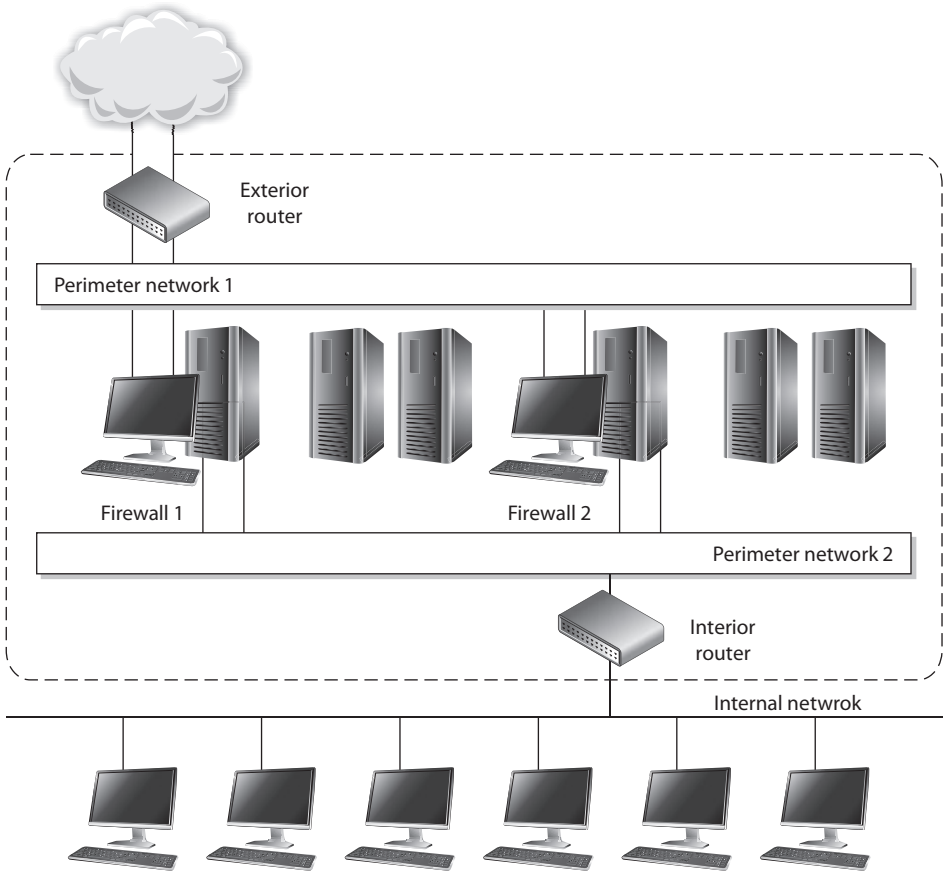


Figure 21-10 A screened subnet can have different networks within it and different firewalls that filter for specific threats.

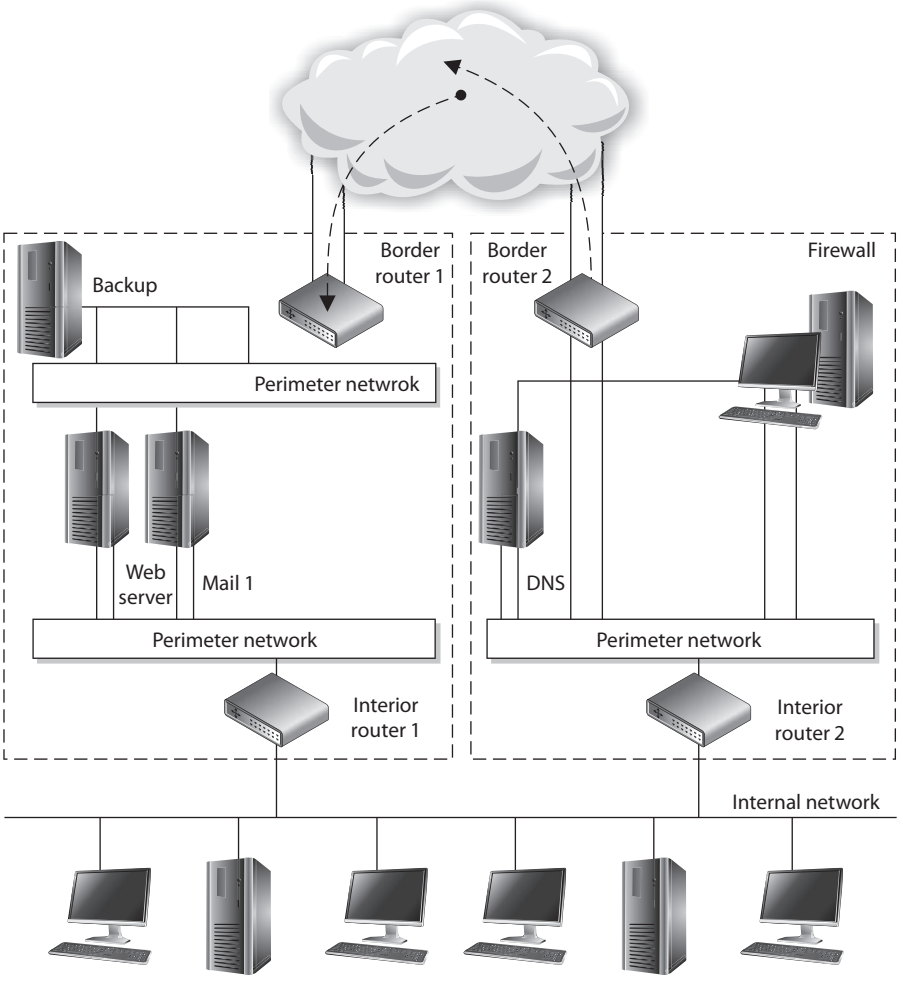


Figure 21-11 Some architectures have separate screened subnets with different server types in each.

The screened-subnet approach provides more protection than a stand-alone firewall or a screened-host firewall because three devices are working together and an attacker must compromise all three devices to gain access to the internal network. This architecture also sets up a DMZ between the two firewalls, which functions as a small network isolated among the trusted internal and untrusted external networks. The internal users usually

Firewall Architecture Characteristics

It is important to understand the following characteristics of these firewall architecture types:

Dual-homed:

- A single computer with separate NICs connected to each network.
- Used to divide an internal trusted network from an external untrusted network.
- Must disable a computer's forwarding and routing functionality so the two networks are truly segregated.

Screened host:

- A router filters (screens) traffic before it is passed to the firewall.

Screened subnet:

- An external router filters (screens) traffic before it enters the subnet. Traffic headed toward the internal network then goes through two firewalls.

have limited access to the servers within this area. Web, e-mail, and other public servers often are placed within the DMZ. Although this solution provides the highest security, it also is the most complex. Configuration and maintenance can prove to be difficult in this setup, and when new services need to be added, three systems may need to be reconfigured instead of just one.

TIP Sometimes a screened-host architecture is referred to as a single-tiered configuration and a screened subnet is referred to as a two-tiered configuration. If three firewalls create two separate DMZs, this may be called a three-tiered configuration.

Organizations used to deploy a piece of hardware for every network function needed (DNS, mail, routers, switches, storage, web), but today many of these items run within virtual machines on a smaller number of hardware machines. This reduces software and hardware costs and allows for more centralized administration, but these components still need to be protected from each other and external malicious entities. As an analogy, let's say that 15 years ago each person lived in their own house and a police officer was placed between each house so that the people in the houses could not attack each other. Then last year, many of these people moved in together so that now at least five

people live in the same physical house. These people still need to be protected from each other, so some of the police officers had to be moved inside the houses to enforce the laws and keep the peace. Analogously, virtual firewalls have "moved into" the virtualized environments to provide the necessary protection between virtualized entities.

As illustrated in Figure 21-12, a network can have a traditional physical firewall on the physical network and *virtual firewalls* within the individual virtual environments.

Virtual firewalls can provide bridge-type functionality in which individual traffic links are monitored between virtual machines, or they can be integrated within the hypervisor. The hypervisor is the software component that carries out virtual machine management and oversees guest system software execution. If the firewall is embedded within the hypervisor, then it can "see" and monitor all the activities taking place within the system.

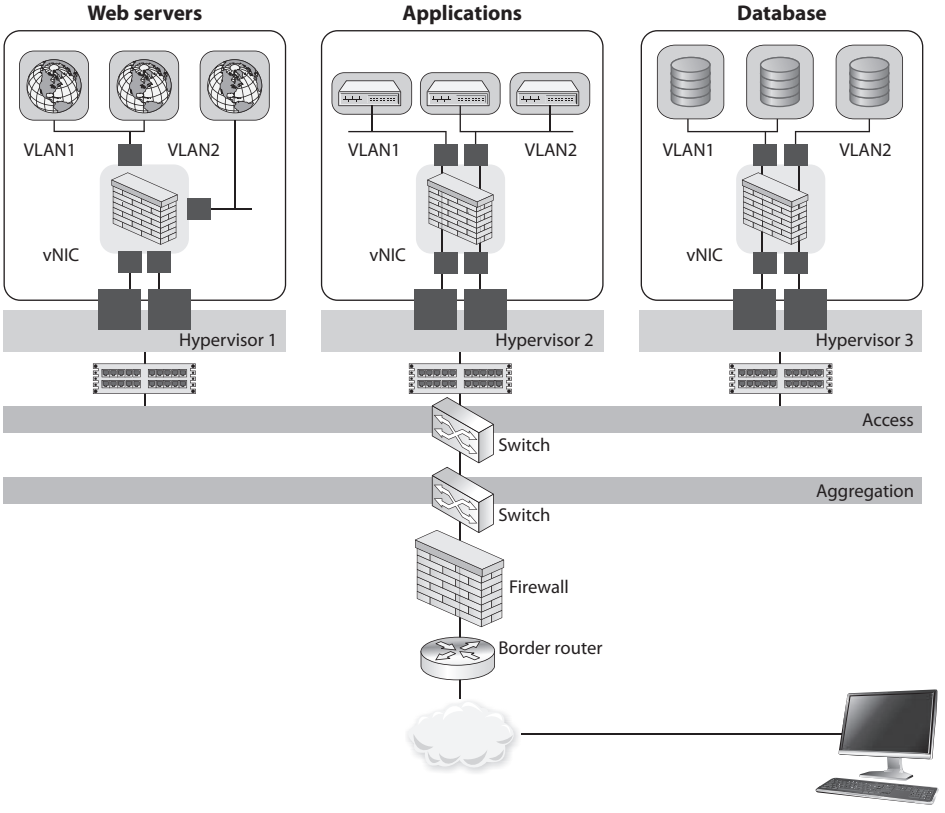


Figure 21-12 Virtual firewalls

Bastion Host

A system is considered a *bastion host* if it is a highly exposed device that is most likely to be targeted by attackers. The closer any system is to an untrusted network, such as the Internet, the more it is considered a target candidate since it has a smaller number of layers of protection guarding it. If a system is on the public side of a DMZ or is directly connected to an untrusted network, it is considered a bastion host; thus, it needs to be extremely locked down.

The system should have all unnecessary services disabled, unnecessary accounts disabled, unneeded ports closed, unused applications removed, unused subsystems and administrative tools removed, and so on. The attack surface of the system needs to be reduced, which means the number of potential vulnerabilities needs to be reduced as much as possible.

A bastion host does not have to be a firewall—the term just relates to the position of the system in relation to an untrusted environment and its threat of attack. Different systems can be considered bastion hosts (mail, web, DNS, etc.) if they are placed on the outer edges of networks.

The "Shoulds" of Firewalls

The default action of any firewall should be to implicitly deny any packets not explicitly allowed. This means that if no rule states that the packet can be accepted, that packet should be denied, no questions asked. Any packet entering the network that has a source address of an internal host should be denied. *Masquerading*, or *spoofing*, is a popular attacking trick in which the attacker modifies a packet header to have the source address of a host inside the network she wants to attack. This packet is spoofed and illegitimate. There is no reason a packet coming from the Internet should have an internal source network address, so the firewall should deny it. The same is true for outbound traffic. No traffic should be allowed to leave a network that does not have an internal source address. If this occurs, it means someone, or some program, on the internal network is spoofing traffic. This is how *zombies* work—the agents used in distributed DoS (DDoS) attacks. If packets are leaving a network with different source addresses, these packets are spoofed and the network is most likely being used as an accomplice in a DDoS attack.

Firewalls should reassemble fragmented packets before sending them on to their destination. In some types of attacks, the hackers alter the packets and make them seem to be something they are not. When a fragmented packet comes to a firewall, the firewall is seeing only part of the picture. It makes its best guess as to whether this piece of a packet is malicious or not. Because these fragments contain only a part of the full packet, the firewall is making a decision without having all the facts. Once all fragments are allowed through to a host computer, they can be reassembled into malicious packages that can cause a lot of damage. A firewall should accept each fragment, assemble the fragments

into a complete packet, and then make an access decision based on the whole packet. The drawback to this, however, is that firewalls that do reassemble packet fragments before allowing them to go on to their destination computer cause traffic delay and more overhead. It is up to the organization to decide whether this configuration is necessary and whether the added traffic delay is acceptable.

Many organizations choose to deny network entrance to packets that contain source routing information, which was mentioned earlier. Source routing means that the packet decides how to get to its destination, not the routers in between the source and destination computer. Source routing moves a packet throughout a network on a predetermined path. The sending computer must know about the topology of the network and how to route data properly. This is easier for the routers and connection mechanisms in between, because they do not need to make any decisions on how to route the packet. However, it can also pose a security risk. When a router receives a packet that contains source routing information, the router assumes the packet knows what needs to be done and passes the packet on. In some cases, not all filters may be applied to the packet, and a network administrator may want packets to be routed only through a certain path and not the route a particular packet dictates. To make sure none of this misrouting happens, many firewalls are configured to check for source routing information within the packet and deny it if it is present.

Firewalls are not effective "right out of the box." You really need to understand the type of firewall being implemented and its configuration ramifications. For example, a firewall may have implied rules, which are used before the rules you configure. These implied rules might contradict your rules and override them. In this case, you may think that a certain traffic type is being restricted, but the firewall allows that type of traffic into your network by default.

The following list addresses some of the issues that need you need to understand as they pertain to firewalls:

- Most of the time a distributed approach needs to be used to control all network access points, which cannot happen through the use of just one firewall.
- Firewalls can present a potential bottleneck to the flow of traffic and a single point of failure threat.
- Some firewalls do not provide protection from malware and can be fooled by the more sophisticated attack types.
- Firewalls do not protect against sniffers or rogue wireless access points and provide little protection against insider attacks.

The role of firewalls is becoming more and more complex as they evolve and take on more functionality and responsibility. At times, this complexity works against security professionals because it requires them to understand and properly implement additional functionality. Without an understanding of the different types of firewalls and architectures available, many more security holes can be introduced, which lays out the welcome mat for attackers.

Intrusion Detection and Prevention Systems

The options for intrusion detection and prevention include host-based intrusion detection systems (HIDSs), network-based intrusion detection systems (NIDSs), and wireless intrusion detection systems (WIDSs). Each may operate in detection or prevention mode depending on the specific product and how it is employed. As a refresher, the main difference between an intrusion detection system (IDS) and an intrusion prevention system (IPS) is that an IDS only detects and reports suspected intrusions, while an IPS detects, reports, and stops suspected intrusions. How do they do this? There are two basic approaches: rule-based or anomaly-based.

Rule-Based IDS/IPS

Rule-based intrusion detection and prevention is the simplest and oldest technology. Essentially, we write rules (or subscribe to a service that writes them for us) and load those onto the system. The IDS/IPS monitors the environment in which it is placed, looking for anything that matches a rule. For example, suppose you have a signature for a particular piece of malware. You could create a rule that looks for any data that matches that signature and either raise an alert (IDS) or drop the data and generate the alert (IPS). Rule-based approaches are very effective when we know the telltale signs of an attack. But what if the attacker changes tools or procedures?

The main drawback of rule-based approaches to detecting attacks is that we need to have a rule that accurately captures the attack. This means someone got hacked, investigated the compromise, generated the rule, and shared it with the community. This process takes time and, until the rule is finalized and loaded, the system won't be effective against that specific attack. Of course, there's nothing stopping the adversary from slightly modifying tools or techniques to bypass your new rule either.

Anomaly-Based IDS/IPS

Anomaly-based intrusion detection and prevention uses a variety of approaches to detect things that don't look right. One basic approach is to observe the environment for some time to figure out what "normal" looks like. This is called the *training mode*. Once it has created a baseline of the environment, the IDS/IPS can be switched to *testing mode*, in which it compares observations to the baselines created earlier. Any observation that is significantly different generates an alert. For example, a particular workstation has a pattern of behavior during normal working hours and never sends more than, say, 10MB of data to external hosts during a regular day. One day, however, it sends out 100MB. That is pretty anomalous, so the IDS/IPS raises an alert (or blocks the traffic). But what if that was just the annual report being sent to the regulators?

The main challenge with anomaly-based approaches is that of *false positives*; that is, detecting intrusions when none happened. False positives can lead to fatigue and desensitizing the personnel who need to examine each of these alerts. Conversely, *false negatives* are events that the system incorrectly classifies as benign, delaying the response until the intrusion is detected through some other means. Obviously, both are bad outcomes.

EDR, NDR, and XDR

HIDS and antimalware features are increasingly being bundled into comprehensive *endpoint detection and response (EDR)* platforms. Similarly, NIDSs are evolving into *network detection and response (NDR)* products. These newer solutions do everything that HIDSs and NIDSs do, but also offer a host of other features such as combining rule-based and anomaly detection capabilities. *Extended detection and response (XDR)* platforms take this one step further by correlation of events across multiple sensors, both in the cloud and on premises, to get a more holistic view of what is going on in an environment.

Perhaps the most important step toward reducing errors is to baseline the system. *Baselining* is the process of establishing the normal patterns of behavior for a given network or system. Most of us think of baselining only in terms of anomaly-based IDSs because these typically have to go through a period of learning before they can determine what is anomalous. However, even rule-based IDSs should be configured in accordance with whatever is normal for an organization. There is no such thing as a one-size-fits-all *set* of IDS/IPS rules, though some *individual* rules may very well be applicable to all (e.g., detecting a known specimen of malware).

NOTE The term "perimeter" has lost some of its importance of late. While it remains an important concept in terms of security architecting, it can mislead some into imagining a wall separating us from the bad guys. A best practice is to assume the adversaries are already "inside the wire," which downplays the importance of a perimeter in security operations.

Whitelisting and Blacklisting

One of the most effective ways to tune detection platforms like IDS/IPS is to develop lists of things that are definitely benign and those that are definitely malicious. The platform, then, just has to figure out the stuff that is not on either list. A *whitelist* (more inclusively called an *allow list*) is a set of known-good resources such as IP addresses, domain names, or applications. Conversely, a *blacklist* (also known as a *deny list*) is a set of known-bad resources. In a perfect world, you would only want to use whitelists, because nothing outside of them would ever be allowed in your environment. In reality, we end up using them in specific cases in which we have complete knowledge of the acceptable resources. For example, whitelisting applications that can execute on a computer is an effective control because users shouldn't be installing arbitrary software on their own. Similarly, we can whitelist devices that are allowed to attach to our networks.

Things are different when we can't know ahead of time all the allowable resources. For example, it is a very rare thing for an organization to be able to whitelist websites for every user. Instead, we would rely on blacklists of domain and IP addresses. The problem with blacklists is that the Internet is such a dynamic place that the only thing we can

be sure of is that our blacklist will always be incomplete. Still, blacklisting is better than nothing, so we should always try to use whitelists first, and then fall back on blacklists when we have no choice.

Antimalware Software

Traditional antimalware software uses signatures to detect malicious code. Signatures, sometimes referred to as fingerprints, are created by antimalware vendors. A *signature* is a set of code segments that a vendor has extracted from a malware sample. Similar to how our bodies have antibodies that identify and go after specific pathogens by matching segments of their genetic codes, antimalware software has an engine that scans files, e-mail messages, and other data passing through specific protocols and then compares them to its database of signatures. When there is a match, the antimalware software carries out whatever activities it is configured to do, which can be to quarantine the item, attempt to clean it (remove the malware), provide a warning message dialog box to the user, and/or log the event.

Signature-based detection (also called *fingerprint detection*) is a reasonably effective way to detect conventional malware, but it has a delayed response time to new threats. Once malware is detected in the wild, the antimalware vendor must study it, develop and test a new signature, release the signature, and all customers must download it. If the malicious code is just sending out silly pictures to all of your friends, this delay is not so critical. If the malicious software is a new variant of TrickBot (a versatile Trojan behind many ransomware attacks), this amount of delay can be devastating.

Since new malware is released daily, it is hard for the signature-based vendors to keep up. Another technique that almost all antimalware software products use is referred to as *heuristic detection*. This approach analyzes the overall structure of the malicious code, evaluates the coded instructions and logic functions, and looks at the type of data within the virus or worm. So, it collects a bunch of information about this piece of code and assesses the likelihood of it being malicious in nature. It has a type of "suspiciousness counter," which is incremented as the program finds more potentially malicious attributes. Once a predefined threshold is met, the code is officially considered dangerous and the antimalware software jumps into action to protect the system. This allows antimalware software to detect unknown malware, instead of just relying on signatures.

As an analogy, let's say Barney is the town cop who is employed to root out the bad guys and lock them up (quarantine). If Barney uses a signature method, he compares a stack of photographs of bad actors to each person he sees on the street. When he sees a match, he quickly throws the bad guy into his patrol car and drives off. By contrast, if he uses a heuristic method, he watches for suspicious activity. So if someone with a ski mask is standing outside a bank, Barney assesses the likelihood of this being a bank robber against it just being a cold guy in need of some cash.

Some antimalware products create a simulated environment, called a *virtual machine* or *sandbox*, and allow some of the logic within the suspected code to execute in the protected environment. This allows the antimalware software to see the code in question in action, which gives it more information as to whether or not it is malicious.

NOTE The virtual machine or sandbox is also sometimes referred to as an *emulation buffer*. They are all the same thing—a piece of memory that is segmented and protected so that if the code is malicious, the system is protected.

Reviewing information about a piece of code is called *static analysis*, while allowing a portion of the code to run in a virtual machine is called *dynamic analysis*. They are both considered heuristic detection methods.

Now, even though all of these approaches are sophisticated and effective, they are not 100 percent effective because malware writers are crafty. It is a continual cat-and-mouse game that is carried out every day. The antimalware industry comes out with a new way of detecting malware, and the very next week the malware writers have a way to get around this approach. This means that antimalware vendors have to continually increase the intelligence of their products and you have to buy a new version every year.

The next phase in the antimalware software evolution is referred to as behavior blockers. Antimalware software that carries out *behavior blocking* actually allows the suspicious code to execute within the operating system unprotected and watches its interactions with the operating system, looking for suspicious activities. The antimalware software watches for the following types of actions:

- Writing to startup files or the Run keys in the Windows registry
- Opening, deleting, or modifying files
- Scripting e-mail messages to send executable code
- Connecting to network shares or resources
- Modifying an executable logic
- Creating or modifying macros and scripts
- Formatting a hard drive or writing to the boot sector

If the antimalware program detects some of these potentially malicious activities, it can terminate the software and provide a message to the user. The newer-generation behavior blockers actually analyze sequences of these types of operations before determining the system is infected. (The first-generation behavior blockers only looked for individual actions, which resulted in a large number of false positives.) The newer-generation software can intercept a dangerous piece of code and not allow it to interact with other running processes. They can also detect rootkits. In addition, some of these antimalware programs can allow the system to roll back to a state before an infection took place so the damages inflicted can be "erased."

While it sounds like behavior blockers might bring us our well-deserved bliss and utopia, one drawback is that the malicious code must actually execute in real time; otherwise, our systems can be damaged. This type of constant monitoring also requires a high level of system resources. We just can't seem to win.

EXAM TIP Heuristic detection and behavior blocking are considered proactive and can detect new malware, sometimes called "zero-day" attacks. Signature-based detection cannot detect new malware.

Most antimalware vendors use a blend of all of these technologies to provide as much protection as possible. The individual antimalware attack solutions are shown in Figure 21-13.

NOTE Another antimalware technique is referred to as *reputation-based protection*. An antimalware vendor collects data from many (or all) of its customers' systems and mines that data to search for patterns to help identify good and bad files. Each file type is assigned a reputation metric value, indicating the probability of it being "good" or "bad." These values are used by the antimalware software to help it identify "bad" (suspicious) files.

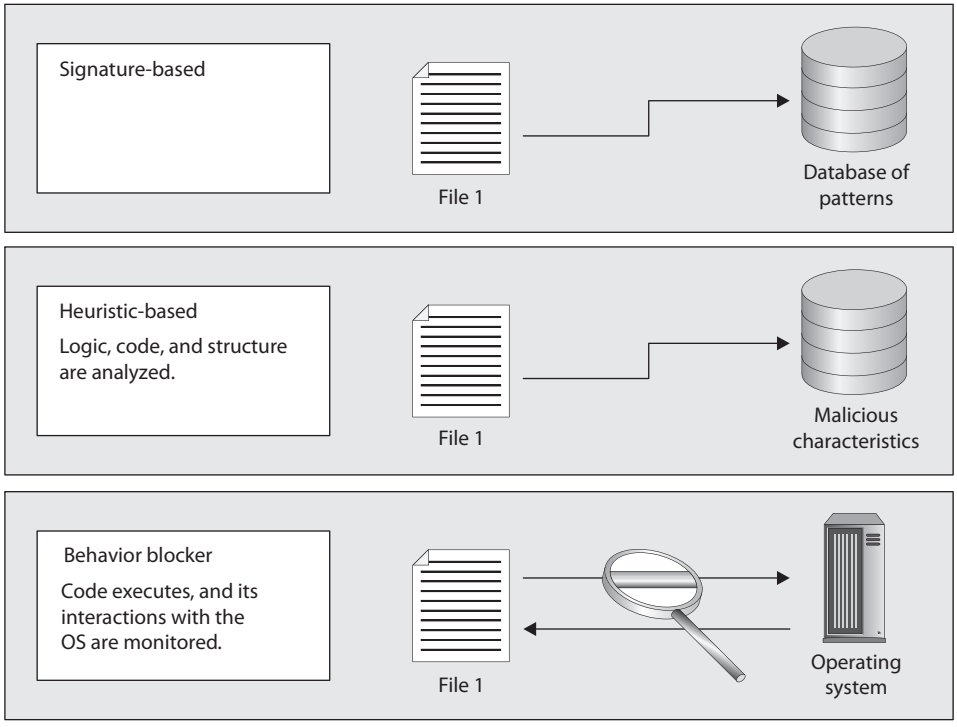


Figure 21-13 Antimalware vendors use various types of malware detection.

Detecting and protecting an enterprise from the long list of malware requires more than just rolling out antimalware software. Just as with other pieces of a security program, certain administrative, physical, and technical controls must be deployed and maintained.

The organization should either have a stand-alone antimalware policy or have one incorporated into an existing security policy. It should include standards outlining what type of antimalware software and antispyware software should be installed and how they should be configured.

Antimalware information and expected user behaviors should be integrated into the security-awareness program, along with who users should contact if they discover a virus. A standard should cover the do's and don'ts when it comes to malware, which are listed next:

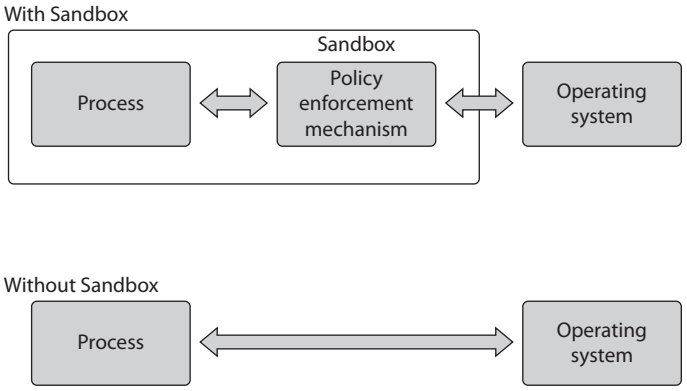
- Every workstation, server, and mobile device should have antimalware software installed.
- An automated way of updating malware signatures should be deployed on each device.
- Users should not be able to disable antimalware software.
- A preplanned malware eradication process should be developed and a contact person designated in case of an infection.
- All external disks (USB drives and so on) should be scanned automatically.
- Backup files should be scanned.
- Antimalware policies and procedures should be reviewed annually.
- Antimalware software should provide boot malware protection.
- Antimalware scanning should happen at a gateway and on each device.
- Virus scans should be automated and scheduled. Do not rely on manual scans.
- Critical systems should be physically protected so malicious software cannot be installed locally.

Since malware has cost organizations millions of dollars in operational costs and productivity hits, many have implemented antimalware solutions at network entry points. The scanning software can be integrated into a mail server, proxy server, or firewall. (The solutions are sometimes referred to as *virus walls*.) This software scans incoming traffic, looking for malware so it can be detected and stopped before entering the network. These products can scan Simple Mail Transport Protocol (SMTP), HTTP, FTP, and possibly other protocol types, but what is important to realize is that the product is only looking at one or two protocols and not *all* of the incoming traffic. This is the reason each server and workstation should also have antimalware software installed.

Sandboxing

A *sandbox* is an application execution environment that isolates the executing code from the operating system to prevent security violations. To the code, the sandbox looks just like the environment in which it would expect to run. For instance, when we sandbox

an application, it behaves as if it were communicating directly with the OS. In reality, it is interacting with another piece of software whose purpose is to ensure compliance with security policies. Another instance is that of software (such as helper objects) running in a web browser. The software acts as if it were communicating directly with the browser, but those interactions are mediated by a policy enforcer of some sort. The power of sandboxes is that they offer an additional layer of protection when running code that we are not certain is safe to execute.



Outsourced Security Services

Nearly all of the preventive and detective measures we've discussed in the preceding subsections can be outsourced to an external service provider. Why would we want to do that? Well, for starters, many small and midsized organizations lack the resources to provide a full team of experienced security professionals. We are experiencing workforce shortages that are not likely to be solved in the near term. This means that hiring, training, and retaining qualified personnel is not feasible in many cases. Instead, many organizations have turned to managed security services providers (MSSPs) for third-party provided security services.

EXAM TIP Outsourced security services are what (ISC)^2 refers to as *third-party provided security*.

MSSPs typically offer a variety of services ranging from point solutions to taking over the installation, operation, and maintenance of all technical (and some cases physical) security controls. (Sorry, you still have to provide policies and many administrative controls.) Your costs will vary depending on what you need but, in many cases, you'll get more than you could've afforded if you were to provide these services in-house. Still, there are some issues that you should consider before hiring an MSSP:

- **Requirements** Before you start interviewing potential MSSPs, make sure you know your requirements. You can outsource the day-to-day activities, but you can't outsource your responsibility to understand your own security needs.

- **Understanding** Does the MSSP understand your business processes? Are they asking the right questions to get there? If your MSSP doesn't know what it is that your organization does (and how), they will struggle to provide usable security. Likewise, you need to understand their qualifications and processes. Trust is a two-way street grounded on accurate information.
- **Reputation** It is hard to be a subpar service provider and not have customers complain about you. When choosing an MSSP, you need to devote some time to reading online reviews and asking other security professionals about their experiences with specific companies.
- **Costing** You may not be able to afford the deluxe version of the MSSP's services, so you will likely have to compromise and address only a subset of your requirements. When you have trimmed down your requirements, is it still more cost-effective to go with this provider? Should you go with another? Should you just do it yourself?
- **Liability** Any reasonable MSSP will put limits on their liability if your organization is breached. Read the fine print on the contract and consult your attorneys, particularly if you are in an industry that is regulated by the government.

Honeypots and Honeynets

A *honeypot* is a network device that is intended to be exploited by attackers, with the administrator's goal being to gain information on the attackers' tactics, techniques, and procedures (TTPs). Honeypots can work as early detection mechanisms, meaning that the network staff can be alerted that an intruder is attacking a honeypot system, and they can quickly go into action to make sure no production systems are vulnerable to that specific attack type. A honeypot usually sits in the screened subnet, or DMZ, and attempts to lure attackers to it instead of to actual production computers. Think of honeypots as marketing devices; they are designed to attract a segment of the market, get them to buy something, and keep them coming back. Meanwhile, threat analysts are keeping tabs on their adversaries' TTPs.

To make a honeypot system alluring to attackers, administrators may enable services and ports that are popular to exploit. Some honeypot systems *emulate* services, meaning the actual services are not running but software that acts like those services is available. Honeypot systems can get an attacker's attention by advertising themselves as easy targets to compromise. They are configured to look like the organization's regular systems so that attackers will be drawn to them like bears are to honey.

Another key to honeypot success is to provide the right kind of bait. When someone attacks your organization, what is it that they are after? Is it credit card information, patient files, intellectual property? Your honeypots should look like systems that would allow the attacker to access the assets for which they are searching. Once compromised, the directories and files containing this information must appear to be credible. It should also take a long time to extract the information, so that we maximize the contact time with our "guests."

A *honeynet* is an entire network that is meant to be compromised. While it may be tempting to describe honeynets as networks of honeypots, that description might be a bit misleading. Some honeynets are simply two or more honeypots used together. However, others are designed to ascertain a specific attacker's intent and dynamically spawn honeypots that are designed to be appealing to that particular attacker. As you can see, these very sophisticated honeynets are not networks of preexisting honeypots, but rather adaptive networks that interact with the adversaries to keep them engaged (and thus under observation) for as long as possible.

NOTE *Black holes* are sometimes confused with honeynets, when in reality they are almost the opposite of them. Black holes typically are routers with rules that silently drop specific (typically malicious) packets without notifying the source. They normally are used to render botnet and other known-bad traffic useless. Whereas honeypots and honeynets allow us to more closely observe our adversaries, black holes are meant to make them go away for us.

Wrapping up the honey collection, *honeyclients* are synthetic applications meant to allow an attacker to conduct a client-side attack while also allowing the threat analysts an opportunity to observe the TTPs being used by their adversaries. Honeyclients are particularly important in the honey family, because most of the successful attacks happen on the client side, and honeypots are not particularly well suited to track client-side attacks. Suppose you have a suspected phishing or spear phishing attack that you'd like to investigate. You could use a honeyclient to visit the link in the e-mail and pretend it is a real user. Instead of getting infected, however, the honeyclient safely catches all the attacks thrown at it and reports them to you. Since it is not really the web browser it is claiming to be, it is impervious to the attack and provides you with information about the actual tools the attacker is throwing at you. Honeyclients come in different flavors, with some being highly interactive (meaning a human has to operate them), while others involve low interaction (meaning their behavior is mostly or completely automated).

Organizations use these systems to identify, quantify, and qualify specific traffic types to help determine their danger levels. The systems can gather network traffic statistics and return them to a centralized location for better analysis. So as the systems are being attacked, they gather intelligence information that can help the network staff better understand what is taking place within their environment.

It should be clear from the foregoing that honeypots and honeynets are not defensive controls like firewalls and IDSs, but rather help us collect threat intelligence. To be effective, they must be closely monitored by a competent threat analyst. By themselves, honeypots and honeynets do not improve your security posture. However, they can give your threat intelligence team invaluable insights into your adversaries' methods and capabilities.

It is also important to make sure that the honeypot systems are not connected to production systems and do not provide any "jumping off" points for the attacker. There have been instances where companies improperly implemented honeypots and they were exploited by attackers, who were then able to move from those systems to the company's

internal systems. The honeypots need to be properly segmented from any other live systems on the network.

On a smaller scale, organizations may choose to implement *tarpits*, which are similar to honeypots in that they appear to be easy targets for exploitation. A tarpit can be configured to appear as a vulnerable service that attackers commonly attempt to exploit. Once the attackers start to send packets to this "service," the connection to the victim system seems to be live and ongoing, but the response from the victim system is slow and the connection may time out. Most attacks and scanning activities take place through automated tools that require quick responses from their victim systems. If the victim systems do not reply or are very slow to reply, the automated tools may not be successful because the protocol connection times out.

NOTE Deploying honeypots and honeynets has potential liability issues. Be sure to consult your legal counsel before starting down this road.

Artificial Intelligence Tools

Artificial intelligence (AI) is a multidisciplinary field primarily associated with computer science, with influences from mathematics, cognitive psychology, philosophy, and linguistics (among others). At a high level, AI can be divided into two different approaches, as shown in Figure 21-14: symbolic and non-symbolic; the key difference is in how each represents knowledge. Both approaches are concerned with how knowledge is organized, how inference proceeds to support decision-making, and how the system learns.

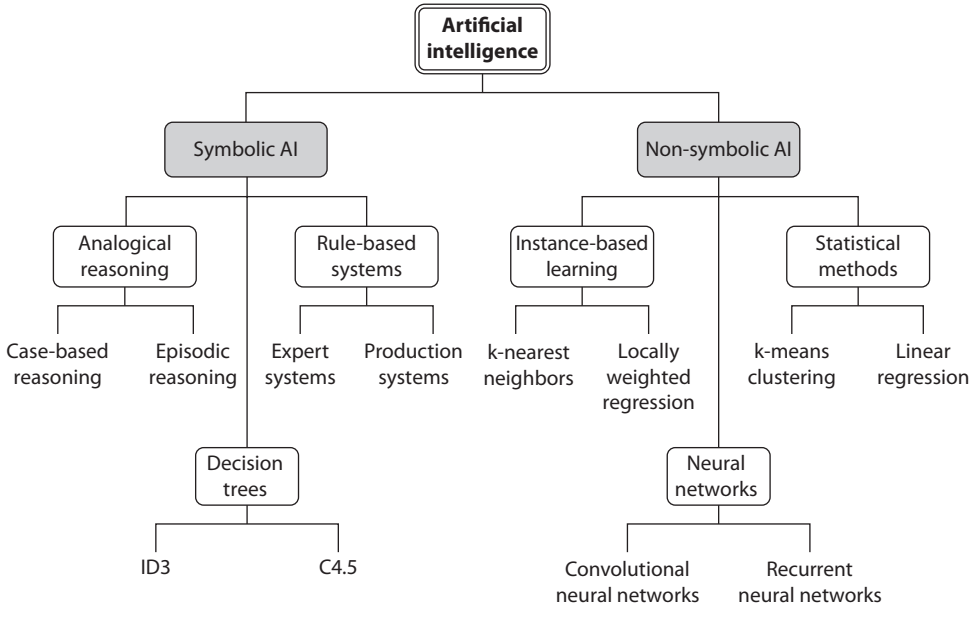


Figure 21-14 A partial taxonomy of artificial intelligence

In symbolic approaches to AI, system developers model real-world concepts, their relationships, and how they interact to solve a set of problems using a set of symbols (e.g., words or tokens). Symbolic AI requires considerable knowledge engineering of both the problem and solution domains, which makes it labor-intensive. However, it yields results that are inherently explainable to humans since the results are derived from human knowledge models in the first place. Symbolic AI systems include the expert systems that became prolific in the 1980s. These relied on extensive interviewing of subject matter experts and time-consuming encoding of their expertise in a series of conditional structures. Unsurprisingly, these early systems were unable to adapt or learn absent human intervention, which is a problem when we consider the number of exceptions that apply to almost all processes.

Another approach to AI departs from the use of symbolic representations of human knowledge and focuses instead on learning patterns in data for classifying objects, predicting future results, or clustering similar sets of data. These non-symbolic AI approaches are where many of the most recent advances have occurred, primarily in classification tasks such as image and voice recognition. In the current vernacular, these non-symbolic approaches are commonly called *machine learning (ML)* even though symbolic systems may also learn. As with symbolic approaches, non-symbolic ML systems also incorporate knowledge representations and reasoning. The knowledge representation is typically quantitative vectors (i.e., non-symbolic) with features from the dataset that describe the input (e.g., pixels from an image, frequencies from an audio file, word vectors, etc.).

Whereas symbolic AI requires considerable knowledge engineering, non-symbolic AI generally requires significant data acquisition and data curating, which can be labor-intensive even for domains where data is readily available. However, rather than having to program the knowledge, as in a symbolic system, the non-symbolic ML system acquires its knowledge in the form of numeric parameters (i.e., weights) through offline training with datasets with millions of examples. As training progresses, the ML model learns the correct parameters that minimize a cost function. That function typically deals with classifying some sample (helpful for finding malware) or making a prediction (allowing us to detect anomalies like spikes in outbound traffic).

Classification determines the class of a new sample based on what is known about previous samples. A common example of this is an algorithm called k-nearest neighbors (KNN), which is a supervised learning technique in which the nearest k neighbors influence the classification of the new point (e.g., if more than half of its k nearest neighbors are in one class, then the new point also belongs in that class). For cybersecurity, this is helpful when trying to determine whether a binary file is malware or detecting whether an e-mail is spam.

Prediction compares previous data samples and determines what the next sample(s) should be. If you have ever taken a statistics class, you may recall a type of analysis called *regression*, in which you try to determine the line (or curve) that most closely approximates a sequence of data points. We use the same approach to prediction in ML by learning from previous observations to determine where the next data point(s) should appear, which is useful for network flow analysis.

On the other hand, there is also unsupervised learning such as clustering, where we do not have a preconception of which classes (or even how many) exist; we determine where the samples naturally clump together. One of the most frequently used clustering algorithms is k-means clustering, in which new data points are added to one of the k clusters based on which one is closest to the new point. Clustering is useful for anomaly detection.

Finally, reinforcement learning tunes decision-making parameters toward choices that lead to positive outcomes in the environment. For example, one might have a security analyst provide feedback to an anomaly detector when it incorrectly classifies a malicious file or event (i.e., a false positive). This feedback adjusts the internal model's weights, so that its anomaly classification improves.

AI has shortcomings that you must consider before employing it. Neither symbolic nor non-symbolic AI approaches cope well with novel situations, and both require a human to re-engineer (symbolic) or retrain (non-symbolic) the algorithms. Symbolic, knowledge-engineered systems may contain underlying biases of the individual(s) who encode the system. Training data sets for non-symbolic approaches may contain biases that are not representative of the operational environment. These biases lead to either false positives or, worse, false negatives when the system is deployed. The best way forward is to combine both approaches, using each other's strengths to offset the other's weaknesses.

Logging and Monitoring

Logging and monitoring are two key activities performed by a SOC using the various tools we just discussed (and probably a few others). These two tasks go hand in hand, since you can't really monitor (at least not very effectively) if you are not logging and, conversely, logging makes little sense if you aren't monitoring. In the sections that follow, we first address how to collect and manage logs, and then discuss the ways in which you should be monitoring those logs (as well as other real-time data feeds).

Log Management

We discussed log reviews and how to prevent log tampering in Chapter 18. To understand how logs support day-to-day security operations, however, we need to take a step back and review why we might be logging system events in the first place. After all, if you don't have clear goals in mind, you will likely collect the wrong events at least some of the time.

Logging Requirements

Earlier in this chapter, we discussed cyberthreat intelligence and, in particular, the collection management framework (CMF). That section on the CMF is a great one to review when you're thinking about what your logging goals should be. After all, logs are data sources that can (and probably should) feed your threat intelligence. Just like intelligence requirements are meant to answer questions from decision-makers, logs should do the same for your SOC analysts. There should be specific questions your security team routinely asks, and those are the questions that should drive what you log and how. For

example, you may be concerned about data leaks of your sensitive research projects to overseas threat actors. What events from which system(s) would you need to log in order to monitor data egress? How often will you be checking logs (which determines how long you must retain them)? If you simply go with default logging settings, you may be ill informed when it comes to monitoring.

Log Standards

Another best practice is to standardize the format of your logs. If you are using a security information and event management (SIEM) system (which we'll discuss shortly), then that platform will take care of normalizing any logs you forward to it. Otherwise, you'll have to do it yourself using either the configuration settings on the system that's logging (if it allows multiple formats) or by using a data processing pipeline such as the open-source Logstash.

NOTE It is essential that you standardize the timestamps on all logs across your environment. If your organization is small, you can use local time; otherwise, we recommend you always use Coordinated Universal Time (UTC).

Something else to consider as you standardize your logs is who will be consuming them. Many SOCs leverage tools for automation, such as some of the AI techniques we discussed earlier. These automated systems may have their own set of requirements for formatting, frequency of updates, or log storage. You should ensure that your standards address the needs of all stakeholders (even non-human ones).

Logging Better

Finally, as with anything else you do in cybersecurity, you want to evaluate the effectiveness of your log management efforts and look for ways to sustain what you're doing well and improve the rest. Establishing and periodically evaluating metrics is an excellent approach to objectively determine opportunities for improvement. For example, how often do analysts lack information to classify an event because of incomplete logging? What logs, events, and fields are most commonly used when triaging alerts? Which are never needed? These questions will point to metrics, and the metrics, in turn, will tell you how well your logging supports your goals.

Security Information and Event Management

A *security information and event management (SIEM)* system is a software platform that aggregates security information (like asset inventories) and security events (which could become incidents) and presents them in a single, consistent, and cohesive manner. SIEMs collect data from a variety of sensors, perform pattern matching and correlation of events, generate alerts, and provide dashboards that allow analysts to see the state of the network. One of the best-known commercial solutions is Splunk, while on the open-source side the Elastic Stack (formerly known as the Elasticsearch-Logstash-Kibana, or ELK, stack) is very popular. It is worth noting that, technically, both of these systems are

data analytics platforms and not simply SIEMs. Their ability to ingest, index, store, and retrieve large volumes of data applies to a variety of purposes, from network provisioning to marketing to enterprise security.

Among the core characteristics of SIEMs is the ability to amass all relevant security data and present it to the security analyst in a way that makes sense. Before these devices became mainstream, security personnel had to individually monitor a variety of systems and manually piece together what all this information might mean. Most SIEMs now include features that group together information and events that seem to be related to each other (or "correlated" in the language of statistics). This allows the analyst to quickly determine the events that are most important or for which there is the most evidence.

SIEM correlations require a fair amount of fine-tuning. Most platforms, out of the box, come with settings that are probably good enough to get you started. You'll have to let your SIEM tool run for a while (one week or longer) for it to start making sense of your environment and giving you meaningful alerts. Inevitably, you'll find that your analysts are drowning in false positives (sadly, a very common problem with automated platforms) that consume their time and joy. This is where you start tuning your settings using things like whitelists and analyst ratings that will make the platform more accurate. You may also discover blind spots (that is, incidents that your SIEM did not pick up) due to insufficient logging or inadequate sensor placement, so you tune a bit there too.

NOTE SIEM fine-tuning should follow your established configuration management processes.

Security Orchestration, Automation, and Response

A tool that is becoming increasingly popular in SOCs is the security orchestration, automation, and response (SOAR) platform. SOAR is an integrated system that enables more efficient security operations through automation of various workflows. The following are the three key components of a SOAR solution:

- **Orchestration** This refers to the integration and coordination of other security tools such as firewalls, IDS/IPS, and SIEM platforms. Orchestration enables automation.
- **Automation** SOAR platforms excel at automating cybersecurity playbooks and workflows, driving significant efficiency gains where those processes exist (or are created).
- **Response** Incident response workflows can involve dozens (or even hundreds) of distinct tasks. A SOAR platform can automatically handle many of those, freeing up the incident responders to work on what humans do best.

Egress Monitoring

A security practice that is oftentimes overlooked by smaller organizations is *egress monitoring*, which is keeping an eye on (and perhaps restricting) the information that is flowing *out* of our networks. Chapter 6 introduced data loss prevention (DLP), which is a very specific use case of this. Beyond DLP, we should be concerned about ensuring that our platforms are not being used to attack others and that our personnel are not communicating (knowingly or otherwise) with unsavory external parties.

A common approach to egress monitoring is to allow only certain hosts to communicate directly with external destinations. This allows us to focus our attention on a smaller set of computers that presumably would be running some sort of filtering software. A good example of this approach is the use of a web gateway, which effectively implements a man-in-the-middle "attack" on all of our organization's web traffic. It is not uncommon to configure these devices to terminate (and thus decrypt) all HTTPS traffic and to do deep packet inspection (DPI) before allowing information to flow out of the network.

User and Entity Behavior Analytics

While most attacks historically are caused by external threat actors, we must not neglect to monitor the activities of users and entities within our organizations. Even if we never encounter a malicious insider, our users are oftentimes unwitting accomplices when they visit the wrong site, click the wrong link, or open the wrong attachment. *User and entity behavior analytics (UEBA)* is a set of processes that determines normal patterns of behavior so that abnormalities can be detected and investigated. For example, if a user hardly ever sends large amounts of data out to the Internet and then one day starts sending megabytes' worth, that would trigger a UEBA alert. Maybe the transmission was perfectly legitimate, but perhaps it was the early part of a data loss incident.

UEBA can exist as a stand-alone product or as a feature in some other tool, such as an EDR or NDR platform. Either way, UEBA uses machine learning to predict future behaviors based on past observations, and statistical analyses to determine when a deviation from the norm is significant enough to raise an alert. As with any other type of solution that offers behavioral analytics, UEBA solutions are prone to false positives. This means that you would probably need to put some effort into fine-tuning a UEBA solution, even after its training period.

EXAM TIP UEBA is a good choice for detecting both malicious insiders and benign user accounts that have been taken over by a malicious actor.

Continuous Monitoring

NIST Special Publication 800-137, *Information Security Continuous Monitoring (ISCM) for Federal Information Systems and Organizations*, defines *information security continuous monitoring* as "maintaining ongoing awareness of information security, vulnerabilities, and threats to support organizational risk management decisions." Think of ISCM as an

ongoing and structured verification of security controls. Are the existing controls still the right ones? Are they still effective? If not, why? These are some of the questions to which continuous monitoring provides answers. It is a critical part of the risk management framework we covered in Chapter 2.

There is a distinction here between logging, monitoring, and continuous monitoring. Your logging policies should be pretty expansive. Data storage is cheap and you want to capture as much data as you can in case you ever need it. Monitoring is more limited because it typically requires a human to personally do it, or at least to deal with the reports (such as SIEM alerts) that come out of it. You would, for example, monitor traffic on a certain port when it looks suspicious and then move on to monitoring something else when you determine that traffic is benign. Continuous monitoring is much more prescriptive. It is a deliberate, risk-based process to determine what gets monitored, how it is monitored, and what to do with the information you gather.

In the end, the whole point of continuous monitoring is to determine if the controls remain effective (in the face of changing threat and organizational environments) at reducing risk to acceptable levels. To do this, you need to carefully consider which metrics would allow you to say "yes" or "no" for each control. For example, suppose you are concerned about the risk of malware infections in your organization, so you implement antimalware controls. As part of continuous monitoring for those controls, you could measure the number of infections in some unit of time (day, week, month).

The metrics and measurements provide data that must be analyzed in order to make it actionable. Continuing our malware example, if your controls are effective, you would expect the number of infections to remain steady over time or (ideally) decrease. You would also want to consider other information in the analysis. For example, your malware infections could go up if your organization goes through a growth spurt and hires a bunch of new people, or the infections could go down during the holidays because many employees are taking vacation. The point is that the analysis is not just about understanding what is happening, but also why.

Finally, continuous monitoring involves deciding how to respond to the findings. If your organization's malware infections have increased and you think this is related to the surge in new hires, should you provide additional security awareness training or replace the antimalware solution? Deciding what to do about controls that are no longer sufficiently effective must take into account risk, cost, and a host of other organizational issues.

Continuous monitoring is a deliberate process. You decide what information you need, then collect and analyze it at a set frequency, and then make business decisions with that information. Properly implemented, this process is a powerful tool in your prevention kit.

Chapter Review

Most of the time spent by the typical organization conducting security operations is devoted to emplacing and maintaining the preventive and detective measures, and then using those to log events and monitor the environment. Entire books have been written

on these topics, so in this chapter we just covered the essentials. A key takeaway is that tools alone will never be enough to give you the visibility you need to detect attacks; you need the integration of people, processes, and technology. We may have put a bit more focus on technology in this chapter, but we wanted to close it by highlighting the fact that well-trained people, working as a team and following existing processes, are essential components of security operations. This is particularly true when things go wrong and we need to respond to incidents, which we're about to cover in the next chapter.

Quick Review

- The security operations center (SOC) encompasses the people, processes, and technology that allow logging and monitoring of preventive controls, detection of security events, and incident response.
- Tier 1 security analysts spend most of their time monitoring security tools and other technology platforms for suspicious activity.
- Tier 2 security analysts dig deeper into the alerts, declare security incidents, and coordinate with incident responders and intelligence analysts to further investigate, contain, and eradicate the threats.
- Threat intelligence is evidence-based knowledge about an existing or emerging menace or hazard to assets that can be used to inform decisions regarding responses to that menace or hazard.
- Threat intelligence is commonly derived from three types of sources: threat data feeds, open-source intelligence (OSINT), and internal systems.
- Cyberthreat hunting is the practice of proactively looking for threat actors in your networks.
- Firewalls support and enforce the organization's network security policy by restricting access to one network from another network.
- Packet-filtering firewalls make access decisions based upon network-level protocol header values using access control lists (ACLs).
- Stateful firewalls add to the capabilities of packet-filtering firewalls by keeping track of the state of a connection between two endpoints.
- Proxy firewalls intercept and inspect messages before delivering them to the intended recipients.
- A next-generation firewall (NGFW) combines the attributes of the previously discussed firewalls, but adds a signature-based and/or behavioral analysis IPS engine, as well as cloud-based threat data sharing.
- Intrusion detection and prevention systems (IDS/IPS) can be categorized as either host-based (HIDS) or network-based (NIDS) and rule-based or anomaly-based.
- A whitelist is a set of known-good resources such as IP addresses, domain names, or applications. Conversely, a blacklist is a set of known-bad resources.

- Antimalware software is most effective when it is installed in every entry and end point and covered by a policy that delineates user training as well as software configuration and updating.
- A sandbox is an application execution environment that isolates the executing code from the operating system to prevent security violations.
- A honeypot is a network device that is intended to be exploited by attackers, with the administrator's goal being to gain information on the attackers' tactics, techniques, and procedures.
- A honeynet is an entire network that is meant to be compromised.
- Honeyclients are synthetic applications meant to allow an attacker to conduct a client-side attack while also allowing the security analysts an opportunity to observe the techniques being used by their adversaries.
- Machine learning (ML) systems acquire their knowledge in the form of numeric parameters (i.e., weights), through training with datasets consisting of millions of examples. In supervised learning, ML systems are told whether or not they made the right decision. In unsupervised training, they learn by observing an environment. Finally, in reinforcement learning they get feedback on their decisions from the environment.
- Effective logging requires a standard time zone for all timestamps.
- A security information and event management (SIEM) system is a software platform that aggregates security information (like asset inventories) and security events (which could become incidents) and presents them in a single, consistent, and cohesive manner.
- Security orchestration, automation, and response (SOAR) platforms are integrated systems that enable more efficient security operations through automation of various workflows.
- Egress monitoring is the process of scanning (and perhaps restricting) the information that is flowing out of our networks.
- User and entity behavior analytics (UEBA) is a set of processes that determines normal patterns of behavior so that abnormalities can be detected and investigated.
- Continuous monitoring allows organizations to maintain ongoing awareness of information security, vulnerabilities, and threats to support organizational risk management decisions.

Questions

Please remember that these questions are formatted and asked in a certain way for a reason. Keep in mind that the CISSP exam is asking questions at a conceptual level. Questions may not always have the perfect answer, and the candidate is advised against always looking for the perfect answer. Instead, the candidate should look for the best answer in the list.

Use the following scenario to answer Questions 1–3. The startup company at which you are the director of security is going through a huge growth spurt and the CEO has decided it's time to let you build out a security operations center (SOC). You already have two cybersecurity analysts (one is quite experienced), a brand-new security information and event management (SIEM) platform, and pretty good security processes in place.

1. The number of alerts on your SIEM is overwhelming your two analysts and many alerts go uninvestigated each day. How can you correct this?
   - **A.** Hire an intelligence analyst to help you focus your collection efforts.
   - **B.** Tune the SIEM platform to reduce false-positive alerts.
   - **C.** Establish a threat hunting program to find attackers before they trigger alerts.
   - **D.** Establish thresholds below which events will not generate alerts.
2. You hire an intelligence analyst and want her to start addressing intelligence requirements. Which of the following should be her first step?
   - **A.** Finding out what questions decision-makers need answered
   - **B.** Establishing a collection management framework
   - **C.** Identifying data sources
   - **D.** Subscribing to a threat data feed
3. Your SOC is maturing rapidly and you are ready to start a cyberthreat hunting program. Which of the following describes the crux of this effort?
   - **A.** Proving or negating hypotheses of threat actions based on threat intelligence
   - **B.** Neutralizing threat actors before they can breach your organization
   - **C.** Digging deeper into the alerts to determine if they constitute security incidents
   - **D.** Allowing hunters an opportunity to observe techniques used by their adversaries
4. A firewall that can only make decisions based on examining a single network layer header is called a
   - **A.** Stateful firewall
   - **B.** Screened host
   - **C.** Packet filter
   - **D.** Next-generation firewall
5. A firewall that understands the three-step handshake of a TCP connection is called a
   - **A.** Packet filter
   - **B.** Proxy firewall
   - **C.** Transport-layer proxy
   - **D.** Stateful firewall

6. What is the main challenge with anomaly-based approaches to intrusion detection and prevention?
   - **A.** False positives
   - **B.** Needing a rule that accurately captures the attack
   - **C.** Cost
   - **D.** Immaturity of the technology
7. Which of the following is an effective technique for tuning automated detection systems like IDS/IPS and SIEMs?
   - **A.** Access control lists
   - **B.** State tables
   - **C.** Whitelists
   - **D.** Supervised machine learning
8. Which of the following terms would describe a system designed to ascertain a specific attacker's intent and dynamically spawn multiple virtual devices that are designed to be appealing to that particular attacker?
   - **A.** Honeypot
   - **B.** Honeyclient
   - **C.** Honeyseeker
   - **D.** Honeynet
9. Which of the following is *not* a typical application of machine learning?
   - **A.** Classification
   - **B.** Prediction
   - **C.** Clustering
   - **D.** Knowledge engineering
10. Which of the following is *not* true about continuous monitoring?
    - **A.** It involves ad hoc processes that provide agility in responding to novel attacks.
    - **B.** Its main goal is to support organizational risk management.
    - **C.** It helps determine whether security controls remain effective.
    - **D.** It relies on carefully chosen metrics and measurements.

Answers

1. **B.** False positives are a very common problem with automated platforms like SIEMs, but they can be alleviated by fine-tuning the platform. An intelligence analyst could help a little bit but would clearly not be the best answer, while threat hunting would be a distractor for such a young SOC that still needs to get alerts

under control. Ignoring low-scoring alerts as a matter of policy would be a very dangerous move when dealing with stealthy attackers.

2. **A.** Threat intelligence is meant to help decision-makers choose what to do about a threat. It answers a question that these leaders may have. The CMF and data sources are all important, of course, but they are driven by the requirements that come out of leaders' questions. After the requirements are known, the intelligence analyst may (or may not) need to subscribe to a threat data feed.
3. **A.** The crux of threat hunting is to develop a hypothesis of adversarial action based on threat intelligence, and then to prove or negate the hypothesis. Inherent in this description are two factors: a) the adversary is already inside the network, and b) no alerts tipped off the defenders to the adversary's presence. These factors negate answers B and C. Answer D describes the purpose of a honeypot, not threat hunting.
4. **C.** Packet filtering is a firewall technology that makes access decisions based upon network-level protocol header values. The device that is carrying out packet-filtering processes is configured with access control lists (ACLs), which dictate the type of traffic that is allowed into and out of specific networks.
5. **D.** Stateful firewalls keep track of the state of a protocol connection, which means they understand the three-step handshake a TCP connection goes through (SYN, SYN/ACK, ACK).
6. **A.** The main challenge with anomaly-based approaches is that of false positives—detecting intrusions when none happened. These can lead to fatigue and desensitizing the personnel who need to examine each of these alerts. Despite this shortcoming, anomaly-based approaches are mature and cost-effective technologies that are differentiated from rule-based systems by not needing rules that accurately capture attacks.
7. **C.** One of the most effective ways to tune detection platforms like IDS/IPS is to develop lists of things that are definitely benign and those that are definitely malicious. The platform, then, just has to figure out the stuff that is not on either list. A whitelist (more inclusively called an allow list) is a set of known-good resources such as IP addresses, domain names, or applications.
8. **D.** Some honeynets are designed to ascertain a specific attacker's intent and dynamically spawn honeypots that are designed to be appealing to that particular attacker. These very sophisticated honeynets are not networks of preexisting honeypots, but rather adaptive networks that interact with the adversaries to keep them engaged (and thus under observation) for as long as possible.
9. **D.** Machine learning (ML), which is a non-symbolic approach to artificial intelligence (AI), is typically used for classification and prediction (using supervised or semi-supervised learning) as well as clustering (using unsupervised learning). Knowledge engineering is a requirement for symbolic forms for AI, such as expert systems, which are not ML in the common sense of the term.

10. A. Continuous monitoring is a deliberate, data-driven process supporting organizational risk management. One of the key questions it answers is whether controls are still effective at mitigating risks. Continuous monitoring could potentially lead to a decision to implement specific ad hoc processes, but these would not really be part of continuous monitoring.

CHAPTER 22

Security Incidents

This chapter presents the following:

- Incident management
- Incident response planning
- Investigations

It takes 20 years to build a reputation and few minutes of cyber-incident to ruin it.

—Stephane Nappo

No matter how talented your security staff may be, or how well everyone in your organization complies with your excellent security policies and procedures, or what cutting-edge technology you deploy, the sad truth is that the overwhelming odds are that your organization will experience a major compromise (if it hasn't already). What then? Having the means to manage incidents well can be just as important as anything else you do to secure your organization. In this chapter, we will cover incident management in general and then drill down into the details of incident response planning.

Although ISC² differentiates incident management and incident investigations, for many organizations, the latter is part of the former. This differentiation is useful to highlight the fact that some investigations involve suspects who may be our own colleagues. While many of us would enjoy the challenge of figuring out how an external threat actor managed to compromise our defenses, there is nothing fun about substantiating allegations that someone we work with did something wrong that caused losses to the organization. Still, as security professionals, we must be ready for whatever threats emerge and deal with the ensuing incidents well and rapidly.

Overview of Incident Management

There are many incident management models, but all share some basic characteristics. They all require that we identify the event, analyze it to determine the appropriate countermeasures, correct the problem(s), and, finally, take measures to keep the event from happening again. (ISC)² has broken out these four basic actions and prescribes seven phases in the incident management process: detection, response, mitigation, reporting, recovery, remediation, and lessons learned. Your own organization will have a unique approach, but it is helpful to baseline it off the industry standard.

Although we commonly use the terms "event" and "incident" interchangeably, there are subtle differences between the two. A *security event* is any occurrence that can be observed, verified, and documented. These events are not necessarily harmful. For example, a remote user login, changes to the Windows Registry on a host, and system reboots are all security events that could be benign or malicious depending on the context. A *security incident* is one or more related events that negatively affect the organization and/or impact its security posture. That remote login from our previous example could be a security incident if it was a malicious user logging in. We call reacting to these issues "incident response" (or "incident handling") because something is negatively affecting the organization and causing a security breach.

EXAM TIP A security event is not necessarily a security violation, whereas a security incident is.

Many types of security incidents (malware, insider attacks, terrorist attacks, and so on) exist, and sometimes an incident is just human error. Indeed, many incident response individuals have received a frantic call in the middle of the night because a system is acting "weird." The reasons could be that a deployed patch broke something, someone misconfigured a device, or the administrator just learned a new scripting language and rolled out some code that caused mayhem and confusion.

Many organizations are at a loss as to who to call or what to do right after they have been the victim of a cybercrime. Therefore, all organizations should have an *incident management policy (IMP)*. This document indicates the authorities and responsibilities regarding incident response for everyone in the organization. Though the IMP is frequently drafted by the CISO or someone on that person's team, it is usually signed by whichever executive "owns" organizational policies. This could be the chief information officer (CIO), chief operations officer (COO), or chief human resources officer (CHRO). It is supported by an incident response plan that is documented and tested before an incident takes place. (More on this plan later.) The IMP should be developed with inputs from all stakeholders, not just the security department. Everyone needs to work together to make sure the policy covers all business, legal, regulatory, and security (and any other relevant) issues.

The IMP should be clear and concise. For example, it should indicate whether systems can be taken offline to try to save evidence or must continue functioning at the risk of destroying evidence. Each system and functionality should have a priority assigned to it. For instance, if a file server is infected, it should be removed from the network, but not shut down. However, if the mail server is infected, it should not be removed from the network or shut down, because of the priority the organization attributes to the mail server over the file server. Tradeoffs and decisions such as these have to be made when formulating the IMP, but it is better to think through these issues before the situation occurs, because better logic is usually possible before a crisis, when there's less emotion and chaos.

Incident Management

Incident management includes proactive and reactive processes. Proactive measures need to be put into place so that incidents can be prevented or, failing that, detected quickly. Reactive measures need to be put into place so that detected incidents are dealt with properly.

Most organizations have only reactive management processes, which walk through how an incident should be handled. A more holistic approach is an incident management program that includes both proactive and reactive incident management processes, ensuring that triggers are monitored to make sure all incidents are actually uncovered. This commonly involves log aggregation, a security information and event management (SIEM) system, and user education. Having clear ways of dealing with incidents is not necessarily useful if you don't have a way to find out if incidents are indeed taking place.

All organizations should develop an *incident response team*, as mandated by the incident management policy, to respond to the large array of possible security incidents. The purpose of having an incident response (IR) team is to ensure that the organization has a designated group of people who are properly skilled, who follow a standard set of procedures, and who jump into action when a security incident takes place. The team should have proper reporting procedures established, be prompt in their reaction, work in coordination with law enforcement, and be recognized (and funded) by management as an important element of the overall security program. The team should consist of representatives from various business units, such as the legal department, HR, executive management, the communications department, physical/corporate security, IS security, and information technology.

There are three different types of incident response teams that an organization can choose to put into place. A *virtual* team is made up of experts who have other duties and assignments within the organization. It is called "virtual" because its members are not full-time incident responders but instead are called in as needed and may be physically remote. This type of team introduces a slower response time, and members must neglect their regular duties should an incident occur. However, a *permanent* team of folks who are dedicated strictly to incident response can be cost prohibitive to smaller organizations. The third type is a *hybrid* of the virtual and permanent models. Certain core members are permanently assigned to the team, whereas others are called in as needed.

Regardless of the type, the incident response team should have the following basic items available:

- A list of outside agencies and resources to contact or report to.
- An outline of roles and responsibilities.
- A call tree to contact these roles and outside entities.
- A list of computer or forensic experts to contact.
- A list of steps to take to secure and preserve evidence.

- A list of items that should be included in a report for management and potentially the courts.
- A description of how the different systems should be treated in this type of situation. (For example, remove the systems from both the Internet and the network and power them down.)

When a suspected crime is reported, the incident response team should follow a set of predetermined steps to ensure uniformity in their approach and that no steps are skipped. First, the IR team should investigate the report and determine whether an actual crime has been committed. If the team determines that a crime has been committed, they should inform senior management immediately. If the suspect is an employee, the team should contact a human resources representative right away. The sooner the IR team begins documenting events, the better. If someone is able to document the starting time of the crime, along with the employees and resources involved, that provides a good foundation for evidence. At this point, the organization must decide if it wants to conduct its own forensic investigation or call in experts. If experts are going to be called in, the system that was attacked should be left alone in order to try and preserve as much evidence of the attack as possible. If the organization decides to conduct its own forensic investigation, it must deal with many issues and address tricky elements. (Forensics will be discussed later in this chapter.)

Computer networks and business processes face many types of threats, each requiring a specialized type of recovery. However, an incident response team should draft and enforce a basic outline of how *all* incidents are to be handled. This is a much better approach than the way many organizations deal with these threats, which is usually in an ad hoc, reactive, and confusing manner. A clearly defined incident-handling process is more cost-effective, enables recovery to happen more quickly, and provides a uniform approach with certain expectation of its results.

Incident handling should be closely related to disaster recovery planning (covered in Chapter 23) and should be part of the organization's disaster recovery plan, usually as an appendix. Both are intended to react to some type of incident that requires a quick response so that the organization can return to normal operations. Incident handling is a recovery plan that responds to malicious technical threats. The primary goal of incident handling is to contain and mitigate any damage caused by an incident and to prevent any further damage. This is commonly done by detecting a problem, determining its cause, resolving the problem, and documenting the entire process.

Without an effective incident-handling program, individuals who have the best intentions can sometimes make the situation worse by damaging evidence, damaging systems, or spreading malicious code. Many times, the attacker booby-traps the compromised system to erase specific critical files if a user does something as simple as list the files in a directory. A compromised system can no longer be trusted because the internal commands listed in the path could be altered to perform unexpected activities. The system could now have a back door for the attacker to enter when he wants, or could

have a logic bomb silently waiting for a user to start snooping around, only to destroy any and all evidence.

Incident handling should also be closely linked to the organization's security training and awareness program to ensure that these types of mishaps do not take place. Past issues that the incident response team encountered can be used in future training sessions to help others learn what the organization is faced with and how to improve response processes.

Employees need to know how to report an incident. Therefore, the incident management policy should detail an escalation process so that employees understand when evidence of a crime should be reported to higher management, outside agencies, or law enforcement. The process must be centralized, easy to accomplish (or the employees won't bother), convenient, and welcomed. Some employees feel reluctant to report incidents because they are afraid they will get pulled into something they do not want to be involved with or accused of something they did not do. There is nothing like trying to do the right thing and getting hit with a big stick. Employees should feel comfortable about the process, and not feel intimidated by reporting suspicious activities.

The incident management policy should also dictate how employees should interact with external entities, such as the media, government, and law enforcement. This, in particular, is a complicated issue influenced by jurisdiction, the status and nature of the crime, and the nature of the evidence. Jurisdiction alone, for example, depends on the country, state, or federal agency that has control. Given the sensitive nature of public disclosure, communications should be handled by communications, human resources, or other appropriately trained individuals who are authorized to publicly discuss incidents. Public disclosure of a security incident can lead to two possible outcomes. If not handled correctly, it can compound the negative impact of an incident. For example, given today's information-driven society, denial and "no comment" may result in a backlash. On the other hand, if public disclosure is handled well, it can provide the organization with an opportunity to win back public trust. Some countries and jurisdictions either already have or are contemplating breach disclosure laws that require organizations to notify the public if a security breach involving personally identifiable information (PII) is even suspected. So, being open and forthright with third parties about security incidents often is beneficial to organizations.

A sound incident-handling program works with outside agencies and counterparts. The members of the team should be on the mailing list of the Computer Emergency Response Team (CERT) so they can keep up-to-date about new issues and can spot malicious events, hopefully before they get out of hand. CERT is a division of the Software Engineering Institute (SEI) that is responsible for monitoring and advising users and organizations about security preparation and security breaches.

NOTE Resources for CERT can be found at https://www.cert.org/incident-management/.

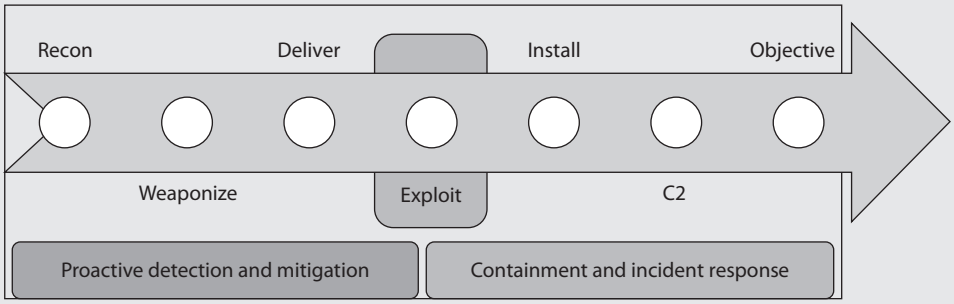
The Cyber Kill Chain

Even as we think about how best to manage incidents, it is helpful to consider a model that describes the stages attackers must complete to achieve their objectives. In their seminal 2011 white paper titled "Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains," Eric Hutchins, Michael Cloppert, and Rohan Amin (employees of Lockheed Martin Corporation, publisher of the white paper) describe a seven-stage intrusion model that has become an industry standard known as the Cyber Kill Chain framework. The seven stages are described here:

1. **Reconnaissance** The adversary has developed an interest in your organization as a target and begins a deliberate information-gathering effort to find vulnerabilities.
2. **Weaponization** Armed with detailed-enough information, the adversary determines the best way into your systems and begins preparing and testing the weapons to be used against you.
3. **Delivery** The cyber weapon is delivered into your system. In over 95 percent of the published cases, this delivery happens via e-mail.
4. **Exploitation** The malicious software is executing on a CPU within your network. This may have launched when the target user clicked a link, opened an attachment, visited a website, or plugged in a USB thumb drive. It could also (in somewhat rare cases) be the result of a remote exploit. One way or another, the attacker's software is now running in your systems.
5. **Installation** Most malicious software is delivered in stages. First, there is the exploit that compromised the system in the prior step. Then, some other software is installed in the target system to ensure persistence, ideally with a good measure of stealth.
6. **Command and Control (C2)** Once the first two stages of the software (exploit and persistence) have been executed, most malware will "phone home" to the attackers to let them know the attack was successful and to request updates and instructions.
7. **Actions on Objectives** Finally, the malware is ready to do whatever it was designed to do. Perhaps the intent is to steal intellectual property and send it to an overseas server. Or perhaps this particular effort is an early phase in a grander attack, so the malware will pivot off the compromised system. Whatever the case, the attacker has won at this point.

As you can probably imagine, the earlier in the kill chain we identify the attack, the greater our odds are of preventing the adversaries from achieving their objectives.

This is a critical concept in this model: if you can thwart the attack before stage four (exploitation), you stand a better chance of winning. Early detection, then, is the key to success.



Incident response is the component of incident management that is executed when a security incident takes place. It starts with detecting the incident and eventually leads to the application of lessons learned during the response. Let's take a closer look at each of the steps in the incident response process.

Detection

The first and most important step in responding to an incident is to realize that you have a problem in the first place. The organization's incident response plan should have specific criteria and a process by which the security staff declares that an incident has occurred. The challenge, of course, is to separate the wheat from the chaff and zero in on the alerts or other indicators that truly represent an immediate danger to the organization.

Detection boils down to having a good sensor network implemented throughout your environment. There are three types of sensors: technical, human, and third-party. *Technical sensors* are, perhaps, the type most of us are used to dealing with. They are provided by the previously mentioned SIEM systems and the other types of systems introduced in Chapter 21: detection and response (EDR), network detection and response (NDR), and security orchestration, automation, and response (SOAR). *Human sensors* can be just as valuable if everyone in your organization has the security awareness to notice odd events and promptly report them to the right place. Many organizations use a special e-mail address to which anyone can send an e-mail report. *Third-party sensors* (technical or human) exist in other organizations. For example, maybe you have a really good relationship with your supply chain partners, and they will alert you to incidents in their environments that appear related to you. That third party could also be a government agency letting you know you've been hacked, which is never a good way to start your day, but is better than not knowing.

Despite this abundance of sensors, detecting incidents can be harder than it sounds, for a variety of reasons. First, sophisticated adversaries may use tools and techniques that you are unable to detect (at least at first). Even if the tools or techniques are known to you, they may very well be hiding under a mound of false positives in your SIEM system. In some (improperly tuned) systems, the ratio of false positives to true positives can be ten to one (or higher). This underscores the importance of tuning your sensors and analysis platforms to reduce the rate of false positives as much as possible.

Response

Having detected the incident, the next step is to respond by containing the damage that has been or is about to be done to your most critical assets. The goal of containment during the response phase is to prevent or reduce any further damage from this incident so that you can begin to mitigate and recover. Done properly, mitigation buys the IR team time for a proper investigation and determination of the incident's root cause. The response strategy should be based on the category of the attack (e.g., internal or external), the assets affected by the incident, and the criticality of those assets. So, what kind of mitigation strategy is best? Well, it depends.

When complete isolation or containment is not a viable solution, you may opt to use boundary devices to stop one system from infecting another. This involves temporarily changing firewall/filtering router rule configuration. Access control lists can be applied to minimize exposure. These response strategies indicate to the attacker that his attack has been noticed and countermeasures are being implemented. But what if, in order to perform a root cause analysis, you need to keep the affected system online and not let on that you've noticed the attack? In this situation, you might consider installing a honeynet or honeypot to provide an area that will contain the attacker but pose minimal risk to the organization. This decision should involve legal counsel and upper management because honeynets and honeypots can introduce liability issues, as discussed in Chapter 21. Once the incident has been contained, you need to figure out what just happened by putting the available pieces together.

This is the substage of analysis, where more data is gathered (audit logs, video captures, human accounts of activities, system activities) to try and figure out the root cause of the incident. The goals are to figure out who did this, how they did it, when they did it, and why. Management must be continually kept abreast of these activities because they will be making the big decisions on how this situation is to be handled.

EXAM TIP Watch out for the context in which the term "response" is used. It can refer to the entire seven-phase incident management process or to the second phase of it. In the second usage, you can think of it as *initial* response aimed at containment.

Mitigation

Having "stopped the bleeding" with the initial containment response, the next step is to determine how to properly mitigate the threat. Though the instinctive reaction may be to clean up the infected workstation or add rules to your firewalls and IDS/IPS,

this well-intentioned response could lead you on an endless game of whack-a-mole or, worse yet, blind you to the adversary's real objective. What do you know about the adversary? Who is it? What are they after? Is this tool and its use consistent with what you have already seen? Part of the mitigation stage is to figure out what information you need in order to restore security.

Once you have a hypothesis about the adversary's goals and plans, you can test it. If this particular actor is usually interested in PII on your high-net-worth clients but the incident you detected was on a (seemingly unrelated) host in the warehouse, was that an initial entry or pivot point? If so, then you may have caught the attacker before they worked their way further along the kill chain. But what if you got your attribution wrong? How could you test for that? This chain of questions, combined with quantifiable answers from your systems, forms the basis for an effective response. To quote the famous hockey player Wayne Gretzky, we should all "skate to where the puck is going to be, not where it has been."

NOTE It really takes a fairly mature threat intelligence capability to determine who is behind an attack (attribution), what are their typical tactics, techniques, and procedures (TTPs), and what might be their ultimate objective. If you do not have this capability, you may have no choice but to respond only to what you're detecting, without regard for what the adversary may actually be trying to do.

Once you are comfortable with your understanding of the facts of the incident, you move to eradicate the adversary from the affected systems. It is important to gather evidence before you recover systems and information. The reason is that, in many cases, you won't know that you will need legally admissible evidence until days, weeks, or even months after an incident. It pays, then, to treat each incident as if it will eventually end up in a court of justice.

Once all relevant evidence is captured, you can begin to fix all that was broken. The mitigation phase ends when you have affected systems that, while still isolated from the production networks, are free from adversarial control. For hosts that were compromised, the best practice is to simply reinstall the system from a gold master image and then restore data from the most recent backup that occurred prior to the attack. You may also have to roll back transactions and restore databases from backup systems. Once you are done, it is as if the incident never happened. Well, almost.

CAUTION An attacked or infected system should never be trusted, because you do not necessarily know all the changes that have taken place and the true extent of the damage. Some malicious code could still be hiding somewhere. Systems should be rebuilt to ensure that all of the potential bad mojo has been released by carrying out a proper exorcism.

Reporting

Though we discuss reporting at this point in order to remain consistent with the incident response process that (ISC)^2 identifies, incident reporting and documentation occurs at various stages in the response process. In many cases involving sophisticated attackers,

the IR team first learns of the incident because someone else reports it. Whether it is an internal user, an external client or partner, or even a government entity, this initial report becomes the starting point of the entire process. In more mundane cases, we become aware that something is amiss thanks to a vigilant member of the security staff or one of the sensors deployed to detect attacks. However we learn of the incident, this first report starts what should be a continuous process of documentation.

According to NIST Special Publication 800-61, Revision 2, *Computer Security Incident Handling Guide*, the following information should be reported for each incident:

- Summary of the incident
- Indicators
- Related incidents
- Actions taken
- Chain of custody for all evidence (if applicable)
- Impact assessment
- Identity and comments of incident handlers
- Next steps to be taken

Recovery

Once the incident is mitigated, you must turn your attention to the recovery phase, in which the aim is to restore full, trustworthy functionality to the organization. It is one thing to restore an individual affected device, which is what we do in mitigation, and another to restore the functionality of business processes, which is the goal of recovery. For example, suppose you have a web service that provides business-to-business (B2B) logistic processes for your organization and your partner organizations. The incident to which you're responding affected the database and, after several hours of work, you mitigated that system and are ready to put it back online. In this recovery stage, you would certify the system as trustworthy and then integrate it back into the web service, thus restoring the business capability.

It is important to note that the recovery phase is characterized by significant testing to ensure the following:

- The affected system is really trustworthy
- The affected system is properly configured to support whatever business processes it did previously
- No compromises exist in those processes

The third characteristic of this phase is assured by close monitoring of all related systems to ensure that the compromise did not persist. Doing this during off-peak hours helps ensure that, should we discover anything else malicious, the impact to the organization is reduced.

Remediation

It is not enough to put the pieces of Humpty Dumpty back together again. You also need to ensure that the attack is never again successful. In the remediation phase, which can (and should) run concurrently with the other phases, you decide which security controls (e.g., updates, configuration changes, firewall/IDS/IPS rules) need to be put in place or modified. There are two steps to this. First, you may have controls that are hastily put into effect because, even if they cause some other issues, their immediate benefit outweighs the risks. Later on, you should revisit those controls and decide which should be made permanent (i.e., through your change management process) and what others you may want to put in place.

NOTE For best results, the remediation phase should start right after detection and be conducted in parallel with the other phases.

Another aspect of remediation is the identification of indicators of attack (IOAs) that can be used in the future to detect this attack in real time (i.e., as it is happening) as well as indicators of compromise (IOCs), which tell you when an attack has been successful and your security has been compromised. Typical indicators of both attack and compromise include the following:

- Outbound traffic to a particular IP address or domain name
- Abnormal DNS query patterns
- Unusually large HTTP requests and/or responses
- DDoS traffic
- New registry entries (in Windows systems)

At the conclusion of the remediation phase, you have a high degree of confidence that this particular attack will never again be successful against your organization. Ideally, you should incorporate your IOAs and IOCs into the following lessons learned stage and share them with the community so that no other organization can be exploited in this manner. This kind of collaboration with partners (and even competitors) makes the adversary have to work harder.

EXAM TIP Mitigation, recovery, and remediation are conveniently arranged in alphabetical order. First you stop the threat, then you get back to business as usual, and then you ensure the threat is never again able to cause this incident.

Lessons Learned

Closure of an incident is determined by the nature or category of the incident, the desired incident response outcome (for example, business resumption or system restoration), and the team's success in determining the incident's source and root cause. Once you have

determined that the incident is closed, it is a good idea to have a team briefing that includes all groups affected by the incident to answer the following questions:

- What happened?
- What did we learn?
- How can we do it better next time?

The team should review the incident and how it was handled and carry out a postmortem analysis. The information that comes out of this meeting should indicate what needs to go into the incident response process and documentation, with the goal of continuous improvement. Instituting a formal process for the briefing provides the team with the ability to start collecting data that can be used to track its performance metrics.

Incident Response Planning

Incident management is implemented through two documents: the incident management policy (IMP) and the incident response plan (IRP). As discussed in the previous section, the IMP establishes authorities and responsibilities across the entire organization. The IMP identifies the IR lead for the organization and describes what every staff member is required to do with regard to incidents. For example, the IMP describes how employees are to report suspected incidents, to whom the report should be directed, and how quickly it should be done.

The IRP gets into the details of what should be done when responding to suspected incidents. The key sections of the IRP cover roles and responsibilities, incident classification, notifications, and operational tasks, all of which are described in the sections that follow. Normally, the IRP does not include detailed procedures for responding to specific incidents (e.g., phishing, data leak, ransomware), but establishes the framework within which all incidents will be addressed. Specific procedures are usually documented in *runbooks*, which are step-by-step scripts developed to deal with incidents that are either common enough or damaging enough to require this level of detailed documentation. Runbooks are described after the IRP sections.

Roles and Responsibilities

The group of individuals who make up the incident response team must have a variety of skills. They must also have a solid understanding of the systems affected by the incident, the system and application vulnerabilities, and the network and system configurations. Although formal education is important, real-world applied experience combined with proper training is key for these folks.

Many organizations divide their IR teams into two sub-teams. The first is the core team of incident responders, who come from the IT and security departments. These individuals are technologists who handle the routine incidents like restoring a workstation whose user inadvertently clicked the wrong link and caused self-infected damage. The second, or extended, team consists of individuals in other departments

who are activated for more complex incidents. The extended team includes attorneys, public relations specialists, and human resources staff (to name a few). The exact makeup of this extended team will vary based on the specifics of the incident, but the point is that these are individuals whose day-to-day duties don't involve IT or security, and yet they are essential to a good response. Table 22-1 shows some examples of the roles and responsibilities in these two teams.

Role	Responsibilities
Core IR Team	
Chief information security officer (CISO)	• Develops and maintains the IR plan • Communicates with senior organizational leadership • Directs security controls before and after incidents
Director of security operations	• Directs execution of the IR plan • Communicates with applicable law enforcement agencies • Declares security incidents
IR team lead	• Overall responsibility for the IR plan • Communicates with senior organizational leadership • Maintains repository of incident response lessons learned
Cybersecurity analyst	• Monitors and analyzes security events • Nominates events for escalation to security incidents • Performs additional analyses for IR team lead as required
IT support specialist	• Manages security platforms • Implements mitigation, recovery, and remediation measures as directed by the IR team lead
Threat intelligence analyst	• Provides intelligence products related to incidents • Maintains repository of incident facts to support future intelligence products
Extended IR Team	
Human resources manager	• Provides oversight for incident-related human resource requirements (e.g., employee relations, labor agreements)
Legal counsel	• Provides oversight for incident-related legal requirements (e.g., liability issues, requirement for law enforcement reporting/coordination) • Ensures evidence collected maintains its forensic value in the event the organization chooses to take legal action
Public relations	• Ensures communications during an incident protect the confidentiality of sensitive information • Prepares communications to stockholders and the press
Business unit lead	• Balances IR actions and business requirements • Ensures business unit support to the IR team

Table 22-1 IR Team Roles and Responsibilities

In addition to these two teams, most organizations rely on third parties when the requirements of the incident response exceed the organic capabilities of the organization. Unless you have an exceptionally well-resourced internal IR team, odds are that you'll need help at some point. The best course of action is to enter into an IR services agreement with a reputable provider *before* any incidents happen. By taking care of the contract and nondisclosure agreement (NDA) beforehand, the IR service provider will be able to jump right into action when time is of the essence. Another time-saving measure is to coordinate a familiarization visit with your IR provider. This will allow the folks who may one day come to your aid to become familiar with your organization, infrastructure, policies, and procedures. They will also get a chance to meet your staff, so everyone learns everyone else's capabilities and limitations.

Incident Classification

The IR team should have a way to quickly determine whether the response to an incident requires that everyone be activated 24/7 or the response can take place during regular business hours over the next couple of days. There is, obviously, a lot of middle ground between these two approaches, but the point is that incident classification criteria should be established, understood by the whole team, and periodically reviewed to ensure that it remains relevant and effective.

There is no one-size-fits-all approach to developing an incident classification framework, but regardless of how you go about it, you should consider three incident dimensions:

- **Impact** If you have a risk management program in place, classifying an incident according to impact should be pretty simple since you've already determined the losses as part of your risk calculations. All you have to do is establish the thresholds that differentiate a bad day from a terrible one.
- **Urgency** The urgency dimension speaks to how quickly the incident needs to be mitigated. For example, an ongoing exfiltration of sensitive data needs to be dealt with immediately, whereas a scenario where a user caused self-infected damage with a bitcoin mining browser extension shouldn't require IR team members to get out of bed in the middle of the night.
- **Type** This dimension helps the team identify the resources that need to be notified and mobilized to deal with the incident. The team that handles the data exfiltration incident mentioned earlier is probably going to be different than the one that handles the infected browser.

Not all organizations explicitly call out each of these dimensions (and some organizations have more dimensions), but it is important to at least consider them. The simplest approach to incident classification simply uses severity and assigns various levels to this parameter depending on whether certain conditions are met. Table 22-2 shows a simple classification matrix for a small to medium-sized organization.

Severity	Criteria	Initial Response Time
Severity 1 (critical)	• Confirmed incident compromising mission-critical systems • Active exfiltration, alteration, or destruction of sensitive data • Incident requiring notification to government regulators • Life-threatening ongoing physical situation (e.g., suspicious package on site, unauthorized/hostile person, credible threat)	1 hour
Severity 2 (high)	• Confirmed incident compromising systems that are not mission-critical • Active exfiltration of non-sensitive data • Time-sensitive investigation of employees • Non-life-threatening but serious, ongoing physical situation (e.g., unauthorized person, theft of property)	4 hours
Severity 3 (moderate)	• Possible incident affecting any systems • Security policy violations • Long-term employee investigations requiring extensive collection and analysis • Non-life-threatening past physical situation (e.g., sensitive area left unsecured overnight)	48 hours

Table 22-2 Sample Incident Classification Matrix

The main advantage of formally classifying incidents is that it allows the preauthorized commitment of resources within specific timeframes. For example, if one of your SOC tier 2 analysts declares a severity 1 (critical) incident, she could be authorized to call the external IR service provider, committing the organization to pay the corresponding fees. There would be no need to get a hold of the CISO and get permission.

Notifications

Another benefit of classifying incidents is that it lets the IR team know who they need to inform and how frequently. Obviously, we don't want to call the CISO at home whenever an employee violates a security policy. On the other hand, we really don't want the CEO to find out the organization had an incident from reading the morning news. Keeping the right decision-makers informed at the right cadence enables everyone to do their jobs well, engenders trust, and leads to unified external messaging.

Table 22-3 shows an example notification matrix that builds on the classification shown previously in Table 22-2.

Notifications to external parties such as customers, partners, government regulators, and the press should be handled by communications professionals and not by the cybersecurity staff. The technical members of the IR team provide the facts to these communicators, who then craft messages (in coordination with the legal and marketing teams) that do not make things worse for the organization either legally or reputationally. Properly handled, IR communications can help improve trust and loyalty to the

Stakeholder	Severity Level	Notification
Executive leaders	S1	Immediate via e-mail and phone
	S2	On the next daily operational report
	S3	None
CISO	S1	Immediate via e-mail and phone
	S2	Within 4 hours via e-mail and phone
	S3	On the next daily operational report
Affected business units	S1	Immediate via e-mail and phone
	S2	Within 4 hours via e-mail
	S3	On the next daily operational report
Affected customers/partners	S1	Within 8 hours via e-mail
	S2	Within 72 hours via e-mail
	S3	None

Table 22-3 Sample Incident Notification Matrix

organization. Improperly handled, however, these notifications (or the lack thereof) can ruin (and have ruined) organizations.

Operational Tasks

Keeping stakeholders informed is just one of the many tasks involved in incident response. Just like any other complex endeavor, we should leverage structured approaches to ensure that all required tasks are performed, and that they are done consistently and in the right order. Now, of course, different types of incidents require different procedures. Responding to a ransomware attack requires different procedures than the procedures for responding to a malicious insider trying to steal company secrets. Still, all incidents follow a very similar pattern at a high level. We already saw this in the discussion of the seven phases in the incident management process that you need to know for the CISSP exam, which apply to all incidents.

Many organizations deal with the need for completeness and consistency in IR by spelling out operational tasks in the IRP, sometimes with a field next to each task to indicate when the task was completed. The IR team lead can then just walk down this list to ensure the right things are being done in the right order. Table 22-4 shows a sample operational tasks checklist.

Table 22-4 is not meant to be all-inclusive but it does capture the most common tasks that apply to every IR in most organizations. As mentioned earlier, different types of incidents require different approaches. While the task list should be general enough to accommodate these specialized procedures, we also want to keep it specific enough to serve as an overall execution plan.

Operational Task	Date/Time Completed
Pre-Execution	
Identify assets affected	
Obtain access (physical and logical) to all affected assets	
Determine forensic evidence requirements	
Review compliance requirements (e.g., GDPR, HIPAA, PCI DSS)	
Initiate communications plan	
Response	
Perform immediate actions to mitigate the impact of the incident	
Validate detection mechanisms	
Request relevant intelligence from threat intelligence team	
Gather and preserve incident-related data (e.g., PCAP, log files)	
Develop an initial timeline of incident-related activity	
Develop mitigation plan based on initial assessment	
Mitigation	
Verify availability of backup/redundant system (if mission-critical system was compromised)	
Activate backup/redundant systems for continuity of operations (if mission-critical system was compromised)	
Isolate affected assets	
Collect forensic evidence from compromised systems (if applicable)	
Remove active threat mechanisms to limit further activity	
Initiate focused monitoring of the environment for additional activity	
Recovery	
Restore affected systems' known-good backups or gold masters	
Validate additional controls on restored systems prevent reoccurrence	
Reconnect restored systems to production networks	
Verify no additional threat activity exists on restored systems	
Remediation	
Finalize root cause, threat mechanisms, and incident timeline	
Identify IOCs and IOAs	
Initiate change management processes to prevent reoccurrence	
Implement preventive and detective controls to prevent reoccurrence	

Table 22-4 Sample Operational Tasks List

Runbooks

When we need specialized procedures, particularly when we expect a certain type of incident to happen more than once, we want to document those procedures to ensure we don't keep reinventing the wheel every time a threat actor gets into our systems. A *runbook* is a collection of procedures that the IR team will follow for specific types of incidents. Think of a runbook as a cookbook. If you feel like having a bean casserole for dinner, you open your cookbook and look up that recipe. It'll tell you what ingredients you need and what the step-by-step procedure is to make it. Similarly, a runbook has tabs for the most likely and/or most dangerous incidents you may encounter. Once the incident is declared by the SOC (or whoever is authorized to declare an incident has occurred), the IR team lead opens the runbook and looks up the type of incident that was declared. The runbook specifies what resources are needed (e.g., specific roles and tools) and how to apply them.

When developing runbooks, you have to be careful that the documentation doesn't take more time and resources to develop than you would end up investing in responding to that incident type. As with any other control, the cost of a runbook cannot exceed the cost of doing nothing (and figuring things out on the fly). For that reason, most organizations focus their runbooks on incidents that require complex responses and those that are particularly sensitive. Other incidents can be (and usually are) added to the runbook, but those additions are deliberate decisions of the SOC manager based on the needs of the organization. For example, if an organization experiences high turnover rates, it might be helpful for new staff to have a more comprehensive runbook to which they can turn.

Another aspect to consider is that runbooks are only good if they are correct, complete, and up to date. Even if you do a great job when you first write runbooks, you'll have to invest time periodically in keeping them updated. For best results, incorporate runbooks into your change management program so that, whenever an organizational change is made, the change advisory board (CAB) asks the question: does this require an update to the IR runbooks?

Investigations

Whatever type of security incident we're facing, we should treat the systems and facilities that it affects as potential crime scenes. The reason is that what may at first appear to have been a hardware failure, a software defect, or an accidental fire may have in fact been caused by a malicious actor targeting the organization. Even acts of nature like storms or earthquakes may provide opportunities for adversaries to victimize us. Because we are never (initially) quite sure whether an incident may have a criminal element, we should treat all incidents as if they do (until proven otherwise).

Since computer crimes are only increasing and will never really go away, it is important that all security professionals understand how computer investigations should be carried out. This includes understanding legal requirements for specific situations, the chain of custody for evidence, what type of evidence is admissible in court, incident response procedures, and escalation processes.

Cops or No Cops?

Management needs to make the decision as to whether law enforcement should be called during an incident response. The following are some of the issues to understand if law enforcement is brought in:

- You may not have a choice in certain cases (e.g., cases involving national security, child pornography, etc.).
- Law enforcement agencies bring significant investigative capability.
- The organization may lose control over where the investigation leads once law enforcement is involved.
- Secrecy of compromise is not promised; it could become part of public record.
- Evidence will be collected and may not be available for a long period of time.

Successfully prosecuting a crime requires solid evidence. Computer forensics is the art of retrieving this evidence and preserving it in the proper ways to make it admissible in court. Without proper computer forensics, few computer crimes could ever be properly and successfully presented in court. The most common reasons evidence is deemed inadmissible in court are lack of qualified staff handling it, lack of established procedures, poorly written policy, or a broken chain of custody.

When a potential computer crime takes place, it is critical that the investigation steps are carried out properly to ensure that the evidence will be admissible to the court (if the matter goes that far) and can stand up under the cross-examination and scrutiny that will take place. As a security professional, you should understand that an investigation is not just about potential evidence on a disk drive. The context matters during an investigation, including the people, network, connected internal and external systems, applicable laws and regulations, management's stance on how the investigation is to be carried out, and the skill set of whoever is carrying out the investigation. Messing up just one of these components could make your case inadmissible or at least damage it if it is brought to court.

Motive, Opportunity, and Means

Today's computer criminals are similar to their traditional counterparts. To understand the "why" in crime, it is necessary to understand the motive, opportunity, and means—or MOM. This is the same strategy used to determine the suspects in a traditional, non-computer crime.

Motive is the "who" and "why" of a crime. The motive may be induced by either internal or external conditions. A person may be driven by the excitement, challenge, and adrenaline rush of committing a crime, which would be an internal condition. Examples of external conditions might include financial trouble, a sick family member, or other dire straits. Understanding the motive for a crime is an important piece in figuring out who

would engage in such an activity. For example, financially motivated attackers such as those behind ransomware want to get your money. In the case of ransomware purveyors, they realize that if they don't decrypt a victim's data after payment of the ransom, the word will get out and no other victims will pay the ransom. For this reason, most modern ransomware actors reliably turn over decryption keys upon payment. Some ransomware gangs even go the extra mile and set up customer service operations to help victims with payment and decryption issues.

Opportunity is the "where" and "when" of a crime. Opportunities usually arise when certain vulnerabilities or weaknesses are present. If an organization does not regularly patch systems (particularly public-facing ones), attackers have all types of opportunities within that network. If an organization does not perform access control, auditing, and supervision, employees may have many opportunities to embezzle funds and defraud the organization. Once a crime fighter finds out why a person would want to commit a crime (motive), she will look at what could allow the criminal to be successful (opportunity).

Means pertains to the abilities a criminal would need to be successful. Suppose a crime fighter was asked to investigate a case of fraud facilitated by a subtle but complex modification made to a software system within a financial institution. If the suspects were three people and two of them just had general computer knowledge, but the third one was a programmer and system analyst, the crime fighter would realize that this person is much likelier to have the means to commit this crime than the other two individuals.

Computer Criminal Behavior

Like traditional criminals, computer criminals have a specific *modus operandi* (*MO*, pronounced "em-oh"). In other words, each criminal typically uses a distinct method of operation to carry out their crime, and that method can be used to help identify them. The difference with computer crimes is that the investigator, obviously, must have knowledge of technology. For example, the MO of a particular computer criminal may include the use of specific tools or targeting specific systems or networks. The method usually involves repetitive signature behaviors, such as sending e-mail messages or programming syntax. Knowledge of the criminal's MO and signature behaviors can be useful throughout the investigative process. Law enforcement can use the information to identify other offenses by the same criminal, for example. The MO and signature behaviors can also provide information that is useful during interviews (conducted by authorized staff members or law enforcement agencies) and potentially a trial.

Psychological crime scene analysis (profiling) can also be conducted using the criminal's MO and signature behaviors. Profiling provides insight into the thought processes of the attacker and can be used to identify the attacker or, at the very least, the tool he used to conduct the crime.

Evidence Collection and Handling

Good evidence is the bedrock on which any sound investigation is built. When dealing with any incident that might end up in court, digital evidence must be handled in a careful fashion so that it can be admissible no matter what jurisdiction is prosecuting

a defendant. Within the United States, the *Scientific Working Group on Digital Evidence (SWGDE)* aims to ensure consistency across the forensic community. The principles developed by SWGDE for the standardized recovery of computer-based evidence are governed by the following attributes:

- Consistency with all legal systems
- Allowance for the use of a common language
- Durability
- Ability to cross international and state boundaries
- Ability to instill confidence in the integrity of evidence
- Applicability to all forensic evidence
- Applicability at every level, including that of individual, agency, and country

The international standard on digital evidence handling is ISO/IEC 27037: *Guidelines for Identification, Collection, Acquisition, and Preservation of Digital Evidence*. This document identifies four phases of digital evidence handling, which are identification, collection, acquisition, and preservation. Let's take a closer look at each.

NOTE You must ensure that you have the legal authority to search for and seize digital evidence before you do so. If in doubt, consult your legal counsel.

Identification

The first phase of digital evidence handling is to identify the digital crime scene. Rarely does only one device comprise the scene of the crime. More often than not, digital evidence exists on a multitude of other devices such as routers, network appliances, cloud services infrastructure, smartphones, and even IoT devices. Whether or not you have to secure a court order to seize evidence, you want to be very deliberate about determining what you think you need to collect and where it might exist.

When you arrive at the crime scene (whether it be physical or virtual), you want to carefully document everything you see and do. If you're dealing with a physical crime scene, photograph it from every possible angle before you touch anything. Label wires and cables and then snap a photo of the labeled system before it is disassembled. Remember that you want to instill confidence in the integrity of evidence and how it was handled from the very onset.

Identifying evidence items at a crime scene may not be straightforward. You could discover wireless networks that would allow someone to remotely tamper with the evidence. This would require you to consider ways to isolate the evidence from radio frequency (RF) signals in order to control the crime scene. There may also be evidence in devices (e.g., thumb drives) that are hidden either deliberately or unintentionally. Law enforcement agents sometimes resort to using specially trained dogs that can sniff out

Controlling the Crime Scene

Whether the crime scene is physical or digital, it is important to control who comes in contact with the evidence of the crime to ensure its integrity. The following are just some of the steps that should take place to protect the crime scene:

- Only allow authorized individuals access to the scene. These individuals should have knowledge of basic crime scene analysis.
- Document who is at the crime scene. In court, the integrity of the evidence may be in question if too many people were milling around the crime scene.
- Document who were the last individuals to interact with the systems.
- If the crime scene does become contaminated, document it. The contamination may not negate the derived evidence, but it will make investigating the crime more challenging.

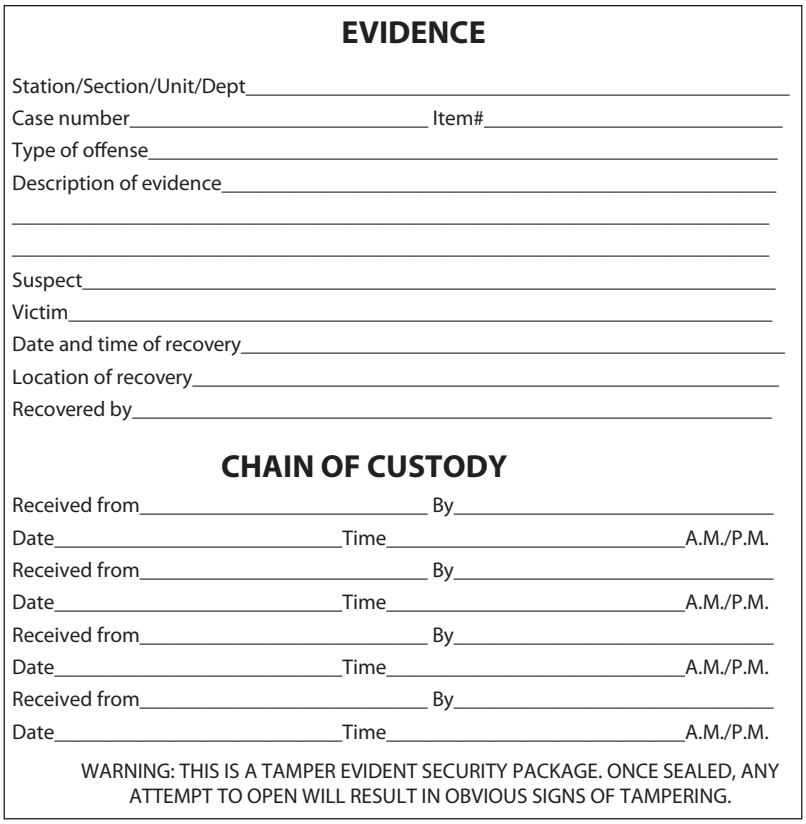
electronics. Thoroughness in identifying evidence is the most important consideration in this phase, and this may require you to think outside the box to ensure you don't miss or lose a critical evidentiary item.

Collection

Once you've identified the evidence you need, you can begin collecting it. Evidence collection is the process of gaining physical control over items that could potentially have evidentiary value. This is where you walk into someone's office and collect their computer, external hard drives, thumb drives, and so on. It is critical that you have the legal authority to do this and that you document what you take, where you take it from, and what its condition is at the time.

Each piece of evidence should be labeled in some way with the date, time, initials of the collector, and a case number if one has been assigned. The piece of evidence should then be placed in a container, which should be sealed (ideally with evidence tape) so that tampering can be detected. An example of the data that should be collected and displayed on each evidence container is shown in Figure 22-1.

After everything is properly labeled, a chain of custody log should be made for each container and an overall log should be made capturing all events. A *chain of custody* documents each person that has control of the evidence at every point in time. In large investigations, one person may collect evidence, another may transport it, and a third may store it. Keeping track of all these individuals' possession of the evidence is critical to proving in court that the evidence was not tampered with. It is not hard for a good defense attorney to get evidence dismissed from court because of improper handling. For this reason, the chain of custody should follow evidence through its entire life cycle, beginning with identification and ending with its destruction, permanent archiving, or return to owner.



EVIDENCE

Station/Section/Unit/Dept_____

Case number_____ Item#_____

Type of offense_____

Description of evidence_____

_____

_____

Suspect_____

Victim_____

Date and time of recovery_____

Location of recovery_____

Recovered by_____

CHAIN OF CUSTODY

Received from_____ By_____

Date_____Time_____A.M./P.M.

Received from_____ By_____

Date_____Time_____A.M./P.M.

Received from_____ By_____

Date_____Time_____A.M./P.M.

Received from_____ By_____

Date_____Time_____A.M./P.M.

WARNING: THIS IS A TAMPER EVIDENT SECURITY PACKAGE. ONCE SEALED, ANY ATTEMPT TO OPEN WILL RESULT IN OBVIOUS SIGNS OF TAMPERING.

Figure 22-1 Evidence container data

Evidence collection activities can get tricky depending on what is being searched for and where. For example, American citizens are protected by the Fourth Amendment against unlawful search and seizure, so law enforcement agencies must have probable cause and request a search warrant from a judge or court before conducting such a search. The actual search can take place only in the areas outlined by the warrant. The Fourth Amendment does not apply to actions by private citizens unless they are acting as police agents. So, for example, if Kristy's boss warned all employees that the management could remove files from their computers at any time, and her boss is not a police officer or acting as a police agent, she could not successfully claim that her Fourth Amendment rights were violated. Kristy's boss may have violated some specific privacy laws, but he did not violate Kristy's Fourth Amendment rights.

In some circumstances, a law enforcement agent is legally permitted to seize evidence that is not included in the search warrant, such as if the suspect tries to destroy the evidence. In other words, if there is an impending possibility that evidence might be destroyed, law enforcement may quickly seize the evidence to prevent its destruction.

This is referred to as *exigent circumstances*, and a judge will later decide whether the seizure was proper and legal before allowing the evidence to be admitted. For example, if a police officer had a search warrant that allowed him to search a suspect's living room but no other rooms and then he saw the suspect putting a removable drive in his pocket while standing in another room, the police officer could seize the drive even though it was outside the area covered under the search warrant.

EXAM TIP Always treat an investigation, regardless of type, as if it would ultimately end up in a courtroom.

Acquisition

In most corporate investigations involving digital evidence, the sort of Crime TV collection we just described will not take place unless law enforcement is involved. Instead, the IR team will probably be able to piece together a timeline of activities from various network resources and you may have to collect only a single laptop. In many cases you can probably acquire the evidence you need remotely without seizing any devices at all. Whatever the case, you ultimately need to get a hold of the data that will confirm or deny the claim that is being investigated, and you must do it in a forensically sound manner.

Acquisition means creating a forensic image of digital data for examination. Generally, speaking, there are two types of acquisition: physical and logical. In *digital acquisition*, the investigator makes a bit-for-bit copy of the contents of a physical storage device, bypassing the operating system. This includes all files, of course, but also free space and previously deleted data. In *logical acquisition*, on the other hand, the forensic image is of the files and folders in a file system, which means we rely on the operating system. This approach is sometimes necessary when dealing with evidence that exists in cloud services, where physical acquisition is normally not possible.

Before creating a forensic image, the investigator must have a medium onto which to copy the data, and ensure this medium has been properly purged, meaning it does not contain any preexisting data. (In some cases, hard drives that were thought to be new and right out of the box contained old data not purged by the vendor.) Two copies are normally created: a *primary image* (a control copy that is stored in a library) and a *working image* (used for analysis and evidence collection). To ensure that the original image is not modified, it is important to compute the cryptographic hashes (e.g., SHA-1) for files and directories before and after the analysis to prove the integrity of the original image.

The investigator works from the duplicate image because it preserves the original evidence, prevents inadvertent alteration of original evidence during examination, and allows re-creation of the duplicate image if necessary.

Acquiring evidence on live systems and those using network storage further complicates matters because you cannot turn off the system to make a copy of the hard drive. Imagine the reaction you'd receive if you were to tell an IT manager that you need to shut down a primary database or e-mail system. It wouldn't be favorable. So these systems and others, such as those using on-the-fly encryption, must be imaged while they are running.

In fact, some evidence is very volatile and can only be collected from a live system. Examples of volatile data that could have evidentiary value include

- Registers and cache
- Process tables and ARP cache
- System memory (RAM)
- Temporary file systems
- Special disk sectors

Preservation

To preserve evidence in a forensically sound manner, you must have established procedures based on legally accepted best practices, and your staff must follow those procedures to the letter. We've already covered two crucial steps in the chain of evidence and the use of hashes to verify that the evidence has not been altered. Another element of preserving digital evidence is ensuring that only a small group of qualified individuals have access to the evidence, and then only to perform specific functions. Again, this access needs to be part of your established procedures. In some cases, organizations implement two-person control of digital evidence to minimize the risk of tampering.

We introduced the topic of evidence storage in Chapter 10, but it bears pointing out that storage of media evidence should be dust-free and kept at room temperature without much humidity, and, of course, the media should not be stored close to any strong magnets or magnetic fields. Even if you don't have a dedicated evidence storage area, you should ensure that whatever space you commandeer is used strictly for this purpose, at least for the life of the investigation.

What Is Admissible in Court?

There are limits to what evidence can be introduced into a legal proceeding. Though the details will be different in each jurisdiction around the world, generally, digital evidence is admissible in court if it meets three criteria:

- **Relevance** Evidence must be relevant to the case, meaning it must help to prove facts being alleged. If a suspect is accused of murder, then a web search history for favorite vacationing spots is probably irrelevant. Judges typically rule on relevance of evidence.
- **Reliability** Evidence must be acquired using a sound forensic methodology that prevents alteration and ensures the evidence remains unaltered during the forensic examination. Multiple high-profile cases in recent years have had evidence rendered inadmissible because the chain of custody was broken.
- **Legality** The persons acquiring and presenting the evidence must have the legal authority to do so. If you have a court-issued search warrant, you must limit collection to whatever is spelled out in it. If you are conducting a workplace investigation, you must limit your collection to organization-owned assets, and only after legal counsel agrees.

The reliability of evidence is most often established by chains of custody and cryptographic hashing. But there is another element to reliability that excludes evidence deemed to be hearsay. *Hearsay evidence* is any statement made outside of the court proceeding that is offered into evidence to prove the truth of the matter asserted in the statement. Suppose that David is accused of fraud and Eliza tells Frank that David told her he was stealing from the company. Eliza's testimony in court would be admissible, but Frank normally wouldn't be allowed to testify about what Eliza claims to have heard because, coming from him, it would be considered hearsay.

Hearsay evidence can also include many computer-generated documents such as log files. In some countries, such as the United States, when computer logs are to be used as evidence in court, they must satisfy a legal exception to the hearsay rule of the Federal Rules of Evidence (FRE) called the business records exception rule or business entry rule. Under this rule, a party could admit any records of a business (1) that were made in the regular course of business; (2) that the business has a regular practice to make such records; (3) that were made at or near the time of the recorded event; and (4) that contain information transmitted by a person with knowledge of the information within the document.

It is important to show that the logs, and all evidence, have not been tampered with in any way, which is the reason for the chain of custody of evidence. Several tools are available that run checksums or hashing functions on the logs, which will allow the team to be alerted if something has been modified.

When evidence is being collected, one issue that can come up is the user's expectation of privacy. If an employee is suspected of, and charged with, a computer crime, he might claim that his files on the computer he uses are personal and not available to law enforcement and the courts. This is why it is important for organizations to conduct security awareness training, have employees sign documentation pertaining to the acceptable use of the organization's computers and equipment, and have legal banners pop up on every employee's computer when they log on. These are key elements in establishing that a user has no right to privacy when he is using organization equipment. The following banner is suggested by CERT Advisory:

> *This system is for the use of authorized users only. Individuals using this computer system without authority, or in excess of their authority, are subject to having all of their activities on this system monitored and recorded by system personnel.*
>
> *In the course of monitoring an individual improperly using this system, or in the course of system maintenance, the activities of authorized users may also be monitored.*
>
> *Anyone using this system expressly consents to such monitoring and is advised that if such monitoring reveals possible evidence of criminal activity, system personnel may provide the evidence of such monitoring to law enforcement officials.*

This explicit warning strengthens a legal case that can be brought against an employee or intruder, because the continued use of the system after viewing this type of warning implies that the person acknowledges the security policy and gives permission to be monitored.

NOTE Don't dismiss the possibility that as an information security professional you will be responsible for entering evidence into court. Most tribunals, commissions, and other quasi-legal proceedings have admissibility requirements. Because these requirements can change between jurisdictions, you should seek legal counsel to better understand the specific rules for your jurisdiction.

Digital Forensics Tools, Tactics, and Procedures

Digital forensics is a science and an art that requires specialized techniques for the recovery, authentication, and analysis of electronic data for the purposes of a digital criminal investigation. It is a fusion of computer science, IT, engineering, and law. When discussing computer forensics with others, you might hear the terms computer forensics, network forensics, electronic data discovery, cyberforensics, and forensic computing.

Forensics Field Kits

When a forensics team is deployed, the forensic investigators should be properly equipped with all the tools and supplies that they'll need to conduct the investigation. The following are some of the common items in forensics field kits:

- **Documentation tools** Tags, labels, forms, and written procedures
- **Disassembly and removal tools** Antistatic bands, pliers, tweezers, screwdrivers, wire cutters, and so on
- **Package and transport supplies** Antistatic bags, evidence bags and tape, cable ties, and others
- **Cables and adapters** Enough to connect to every physical interface you may come across

(ISC)^2 uses *digital forensics* as a synonym for all of these other terms, so that's what you'll see on the CISSP exam.

Anyone who conducts a forensic investigation must be properly skilled in this trade and know what to look for. If someone reboots the attacked system or inspects various files, this could corrupt viable evidence, change timestamps on key files, and erase footprints the criminal may have left. Most digital evidence has a short lifespan and must be collected quickly and in the *order of volatility*. In other words, the most volatile or fragile evidence should be collected first. In some situations, it is best to remove the system from the network, dump the contents of the memory, power down the system, and make a sound image of the attacked system and perform forensic analysis on this copy. Working on the copy instead of the original drive ensures that the evidence stays unharmed on the original system in case some steps in the investigation actually corrupt or destroy data. Dumping the memory contents to a file before doing any work on the system or powering it down is a crucial step because of the information that could be stored there. This is another method of capturing fragile information. However, this creates a sticky situation: capturing RAM or conducting live analysis can introduce changes to the crime scene because various state changes and operations take place. Whatever method the forensic investigator chooses to use to collect digital evidence, that method must be documented. This is the most important aspect of evidence handling.

Forensic Investigation Techniques

To ensure that forensic investigations are carried out in a standardized manner and the evidence collected is admissible, it is necessary for the investigative team to follow specific laid-out steps so that nothing is missed. Figure 22-2 illustrates the phases through a common investigation process and lists various techniques that fall under each phase. Each team or organization may come up with its own steps, but all should be essentially accomplishing the same things:

- Identification
- Preservation
- Collection
- Examination
- Analysis
- Presentation
- Decision

NOTE The principles of criminalistics are included in the forensic investigation process. They are identification of the crime scene, protection of the environment against contamination and loss of evidence, identification of evidence and potential sources of evidence, and the collection of evidence. In regard to minimizing the degree of contamination, it is important to understand that it is impossible not to change a crime scene—be it physical or digital. The key is to minimize changes and document what you did and why, and how the crime scene was affected.

Identification	Preservation	Collection	Examination	Analysis	Presentation
Event/crime detection	Case management	Preservation	Preservation	Preservation	Documentation
Resolve signature	Imaging technologies	Approved methods	Traceability	Traceability	Expert testimony
Profile detection	Chain of custody	Approved software	Validation techniques	Statistical	Clarification
Anomalous detection	Time synchronization	Approved hardware	Filtering techniques	Protocols	Mission impact statement
Complaints		Legal authority	Pattern matching	Data mining	Recommended countermeasure
System monitoring		Lossless compression	Hidden data discovery	Timeline	Statistical interpretation
Audit analysis		Sampling	Hidden data extraction	Link	
		Data reduction		Spatial	
		Recovery techniques			

Figure 22-2 Characteristics of the different phases through an investigation process

During the examination and analysis process of a forensic investigation, it is critical that the investigator work from an image that contains *all* of the data from the original disk. It should be a bit-level copy, sector by sector, to capture deleted files, slack spaces, and unallocated clusters. These types of images can be created through the use of a specialized tool such as Forensic Toolkit (FTK), EnCase Forensic, or the dd Unix utility. A file copy tool does not recover all data areas of the device necessary for examination. Figure 22-3 illustrates a commonly used tool in the forensic world for evidence collection.

The next step is the analysis of the evidence. Forensic investigators use a scientific method that involves

- Determining the characteristics of the evidence, such as whether it's admissible as primary or secondary evidence, as well as its source, reliability, and permanence
- Comparing evidence from different sources to determine a chronology of events
- Event reconstruction, including the recovery of deleted files and other activity on the system

This can take place in a controlled lab environment or, thanks to hardware write-blockers and forensic software, in the field. When investigators analyze evidence in a lab, they are dealing with "dead forensics"; that is, they are working only with static data. Live forensics, which takes place in the field, includes volatile data. If evidence is lacking, then an experienced investigator should be called in to help complete the picture.

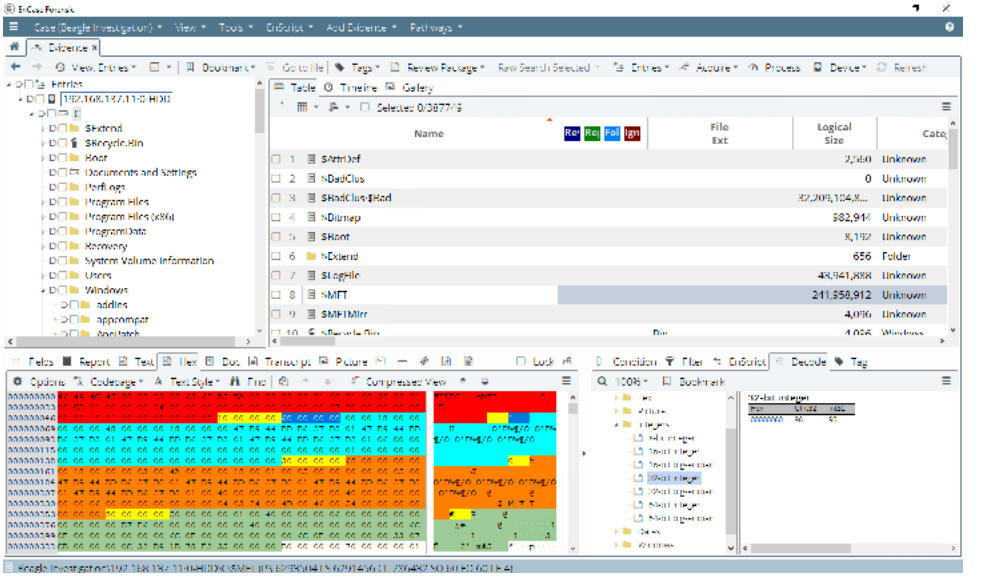


Figure 22-3 EnCase Forensic can be used to collect digital forensic data.

Finally, the interpretation of the analysis should be presented to the appropriate party. This could be a judge, lawyer, CEO, or board of directors. Therefore, it is important to present the findings in a format that will be understood by a nontechnical audience. As a CISSP, you should be able to explain these findings in layperson's terms using metaphors and analogies. Of course, the findings, which are top secret or company confidential, should be disclosed only to authorized parties. This may include the legal department or any outside counsel that assists with the investigation.

Other Investigative Techniques

Unless you work for a law enforcement agency, most of the investigations in which you will be involved are likely to focus on digital forensics investigative techniques. These techniques are applied when a device was compromised, or a malicious insider attempted to steal sensitive files, or something like that. All the evidence you need is probably in a device that you can get your hands on, so you can collect it, acquire it, analyze it, and get to the facts with just digital evidence. However, there may be other situations in which you'll need other types of evidence either in addition to or instead of 1's and 0's copied from some storage device. Interviews, surveillance, and undercover investigative techniques are some of the practices for acquiring evidence that you should be familiar with.

Interviews

Interviews can be effective for ascertaining facts when you have willing interviewees. Interviewing is both an art and a science, and the specific techniques you use will vary

from case to case. Typically, interviews are conducted by a business unit manager with assistance from the human resources and legal departments. This doesn't, however, completely relieve you as an information security professional from responsibility during the interviewing process. You may be asked to provide input or observe an interview in order to clarify technical information that comes up in the course of questioning.

Whether you are conducting an interview or your technical assistance is needed for an interview, keep the following best practices in mind:

- *Have a plan.* Without a plan, the interview will be ineffective. Prepare an outline beforehand that focuses on getting the information you need from each interviewee. However, you should remain flexible and not read off a script.
- *Be fair and objective.* If you are conducting an interview, it is to get to the facts of an incident, not necessarily to reinforce whatever conclusions you may have already reached. Keep an open mind, focus on the facts, and try to avoid any biases.
- *Compartmentalize information.* Your interview plan should address what information you share with each interviewee, and what you don't share. You should not tell one interviewee what another said unless it's absolutely essential and legally permissible.
- *One interviewee at a time.* Interviewing multiple individuals together can introduce problematic group dynamics such as peer pressure. It can also lead interviewees to distort or suppress information.
- *Do not record the interview.* Recording devices can have a chilling effect on interviewees. Instead, have at least one notetaker in the room and, after the interview is complete, read back the notes to the interviewee to ensure their accuracy. If you must record the interview, ensure you comply with all applicable legal requirements (e.g., consent of all parties).
- *Keep it confidential.* Do your best to keep every aspect of the investigation under wraps. Even the fact that someone is being interviewed about an incident can have a damaging reputational effect for that person.

The employee interviewer should be in a position that is senior to the employee subject. A vice president is not going to be very intimidated or willing to spill his guts to the mailroom clerk. The interview should be held in a private place, in an environment conducive to making the subject relatively comfortable and at ease. If exhibits are going to be shown to the subject, they should be shown one at a time, and otherwise kept in a folder. It is not necessary to read a person their rights before the interview unless it is performed by law enforcement officers.

Surveillance

Two main types of surveillance are used when it comes to identifying computer crimes: physical surveillance and computer surveillance. *Physical surveillance* pertains to security cameras, security guards, and closed-circuit TV (CCTV), which may capture evidence.

Physical surveillance can also be used by an undercover agent to learn about the suspect's spending activities, family and friends, and personal habits in the hope of gathering more clues for the case.

Computer surveillance pertains to passively monitoring (auditing) events by using network sniffers, keyboard monitors, wiretaps, and line monitoring. In most jurisdictions, active monitoring may require a search warrant. In most workplace environments, to legally monitor an individual, the person must be warned ahead of time that her activities may be subject to this type of monitoring.

Undercover

Undercover investigative techniques are pretty rare in most corporate investigations, but can provide information and evidence that would be difficult to acquire otherwise. The goal of undercover work is to assume an identity that allows the investigator to blend into the suspect's environment to observe, and perhaps record, the suspect's actions.

A thin line exists between enticement and entrapment when it comes to capturing a suspect's actions. *Enticement* is legal and ethical, whereas *entrapment* is neither legal nor ethical. In the world of computer crimes, a honeypot is a good example to explain the difference between enticement and entrapment. Organizations put systems in their screened subnets that either emulate services that attackers usually like to take advantage of or actually have the services enabled. The hope is that if an attacker breaks into the organization's network, she will go right to the honeypot instead of the systems that are actual production machines. The attacker will be *enticed* to go to the honeypot system because it has many open ports and services running and exhibits vulnerabilities that the attacker would want to exploit. The organization can log the attacker's actions and later attempt to prosecute.

The action in the preceding example is legal unless the organization crosses the line to entrapment. For example, suppose a web page has a link that indicates that if an individual clicks it, she could then download thousands of MP3 files for free. However, when she clicks that link, she is taken to the honeypot system instead, and the organization records all of her actions and attempts to prosecute. Entrapment does not prove that the suspect had the intent to commit a crime; it only proves she was successfully tricked.

Forensic Artifacts

One of the grandfathers of forensic science, Dr. Edmond Locard, famously stated that "every contact leaves a trace." This principle, known as Locard's exchange principle, states that criminals always leave something behind at the crime scene. This fragmentary or trace evidence is a *forensic artifact*. A forensic artifact is anything that has evidentiary value. On a typical computer, the following are examples of forensic artifacts:

- Deleted items (in the recycle bin or trash)
- Web browser search history
- Web browser cache files
- E-mail attachments

- Skype history
- Windows event logs
- Prefetch files

Forensic artifacts can also be evidentiary items relating to network traffic. Network forensics is a subdiscipline that is focused on what happened on the network rather than on the endpoints. The tools used in network forensics are unique to that subdiscipline, and so are the artifacts for which the investigator looks. Tools used in network forensics include NDR solutions, SIEM systems, and the log files of any network device or server. They also include network sniffers that can capture full network frames. The following are some of the more useful network artifacts an investigator would be interested in:

- DNS log records
- Web proxy log records
- IDS/IPS alerts
- Packet capture (pcap) files

Finally, with the proliferation of mobile devices such as smartphones, tablets, and smartwatches, we must not overlook forensic artifacts stored on them. Unlike traditional computers, mobile devices are usually carried by their users around the clock. This means mobile devices tend to document multiple aspects of a person's life, some of which can serve as evidence of criminal activity.

Though mobile devices can be a treasure trove of information for the forensic investigator, they are not always easy to acquire and analyze. For starters, there are so many different models that no single tool can acquire all evidence from all devices. Staff expertise is similarly challenged by this diversity, because an investigator who is skilled at iPhone analysis may not be able to operate at the same level given an Android device. Just to make things more interesting, there is also the issue of encryption, which is prevalent in mobile devices these days.

Still, if forensic investigators can overcome these challenges, mobile devices are excellent sources of evidence for a variety of criminal activity. Among the most useful forensic artifacts found in them are

- Call logs
- SMS messages
- E-mail messages
- Web browser history

Reporting and Documenting

We already covered reporting in a fair amount of detail in Chapter 19. When it comes to investigations, however, there are some additional issues to consider. First and foremost, the need to document *everything* you do cannot be overstated. If you cannot account for

or explain the why of any activities you undertook, it may render evidence inadmissible in court or even undermine the whole case. For this reason, many organizations assign investigators to work in teams of two, where one person documents while the other conducts the investigation. Most forensic analysis tools have a feature that automatically logs everything an investigator does with the tool.

Another issue that is particularly important in writing investigation reports is the need to remain completely logical and factual. Any conclusions you reach must follow logically from a sequence of facts that you spell out for the reader. For example, suppose that Carlos is one of your staff and is suspected of sending sensitive files to a competitor in hopes of landing a lucrative job with them. Even if you are sure he did it (after examining his computer), you should not just jump out and say so. Instead, you show how the forensic artifacts that you found, when arrayed on a timeline, substantiate the claim that Carlos sent sensitive files to a competitor. You'd start by establishing that he was logged into his computer, and then he logged into his personal e-mail account through a webmail interface, and then an e-mail was sent containing sensitive files x, y, and z, and then the e-mail was deleted from his sent items, and so on. It is ultimately up to the reader (presumably a senior manager or court official) to determine guilt or innocence. Your job is to establish the facts and determine whether or not they are consistent with the allegation.

Chapter Review

Incident management is a critical function for any organization. Odds are that if you are among the lucky few who haven't had a major incident yet, you will be faced with one in the near future. In fact, the IronNet 2021 Cybersecurity Impact Report found that 86 percent of respondents had a cybersecurity incident so severe in the previous year that it required a C-level or board meeting. Even if you've outsourced IR to a third-party service provider, you still need to have an incident management policy and an IR plan to guide the conduct of the entire organization before, during, and after an incident. The policy establishes authorities and responsibilities, while the plan specifies the procedures to be followed.

The other major topic we discussed in this chapter is investigations. Thankfully, the need to conduct investigations is fairly rare in most organizations. But therein lies the problem: if you hardly ever need to recall knowledge or practice skills, you are certain to lose them. This is why having detailed standard procedures for investigative work is absolutely essential. For example, evidence acquisition, as we saw, is a complex process that has very little room for errors, particularly if the evidence will end up in court (and we should always assume it will).

Quick Review

- A security event is any occurrence that can be observed, verified, and documented, whereas a security incident is one or more related events that negatively affect the organization and/or impact its security posture.

- A good incident response team should consist of representatives from various business units, such as the legal department, HR, executive management, the communications department, physical/corporate security, IS security, and information technology.
- Incident management encompasses seven phases according to the CISSP CBK: detection, response, mitigation, reporting, recovery, remediation, and lessons learned.
- The detection phase encompasses the search for indicators that an event has occurred and the formal declaration of the event.
- The response phase entails the initial actions undertaken to contain the damage caused by a security incident.
- The goal of the mitigation phase is to eradicate the threat actor from the affected systems.
- Incident reporting occurs at various phases of incident management.
- The aim of the recovery phase is to restore full, trustworthy functionality to the organization.
- In the remediation phase, the incident response team decides which security controls need to be deployed or changed to prevent the incident from recurring.
- The lessons learned phase is important to determine what needs to go into the incident response process and documentation, with the goal of continuous improvement.
- The incident management policy (IMP) establishes authorities and responsibilities across the entire organization, identifies the incident response (IR) lead for the organization, and describes what every staff member is required to do with regard to incidents.
- The incident response plan (IRP) gets into the details of what should be done when responding to suspected incidents, and includes roles and responsibilities, incident classification, notifications, and operational tasks.
- Incident classification criteria allow the organization to prioritize IR assets and usually consider the impact and type of the incident, and urgency with which the response must be started.
- A runbook is a collection of procedures that the IR team will follow for specific types of incidents.
- The four phases of evidence handling are identification, collection, acquisition, and preservation.
- Evidence collection is the process of gaining physical control over devices that could potentially have evidentiary value.
- A chain of custody documents each person that has control of the evidence at every point in time.
- Acquisition means creating a forensic image of digital data for examination.

- Evidence preservation requires maintaining a chain of custody and cryptographic hashes of all digital evidence, and also controlling access to the evidence.
- To be admissible in court, evidence must be relevant, reliable, and legally obtained.
- To be admissible in court, business records such as computer logs have to be made and collected in the normal course of business, not specially generated for a case in court. Business records can easily be deemed hearsay if there is no firsthand proof of their accuracy and reliability.
- Digital forensics is a science and an art that requires specialized techniques for the recovery, authentication, and analysis of electronic data for the purposes of a digital criminal investigation.
- In addition to forensic techniques, organizations sometimes use interviews, surveillance, and undercover investigation techniques.
- When looking for suspects, it is important to consider the motive, opportunity, and means (MOM).
- A forensic artifact is anything that has evidentiary value.

Questions

Please remember that these questions are formatted and asked in a certain way for a reason. Keep in mind that the CISSP exam is asking questions at a conceptual level. Questions may not always have the perfect answer, and the candidate is advised against always looking for the perfect answer. Instead, the candidate should look for the best answer in the list.

1. What are the phases of incident management?

A. Identification, collection, acquisition, and preservation

B. Detection, response, mitigation, reporting, recovery, remediation, and lessons learned

C. Protection, containment, response, remediation, and reporting

D. Analysis, classification, incident declaration, containment, eradication, and investigation

2. During which phase of incident management does the IR team contain the damage caused by a security incident?

A. Preservation

B. Response

C. Eradication

D. Remediation

3. During which phase of incident management are security controls deployed or changed to prevent the incident from recurring?
   - **A.** Preservation
   - **B.** Response
   - **C.** Eradication
   - **D.** Remediation
4. Which document establishes authorities and responsibilities with regard to incidents across the entire organization?
   - **A.** Incident management policy
   - **B.** Incident response plan
   - **C.** Incident response runbook
   - **D.** Incident classification criteria
5. After a computer forensic investigator seizes a computer during a crime investigation, what is the next step?
   - **A.** Label and put it into a container, and then label the container
   - **B.** Dust the evidence for fingerprints
   - **C.** Make an image copy of the disks
   - **D.** Lock the evidence in the safe
6. Which of the following is a necessary characteristic of evidence for it to be admissible?
   - **A.** It must be real.
   - **B.** It must be noteworthy.
   - **C.** It must be reliable.
   - **D.** It must be important.
7. Which of the following is *not* considered a best practice when interviewing willing witnesses?
   - **A.** Compartmentalize information
   - **B.** Interview one interviewee at a time
   - **C.** Be fair and objective
   - **D.** Record the interview

Use the following scenario to answer Questions 8–10. You recently improved your organization's security posture, which now includes a fully staffed security operations center (SOC), network detection and response (NDR) and endpoint detection and response (EDR) systems, centrally managed updates and data backups, and network segmentation using VLANs. It's the end of the workday and just as you are getting ready to go

home your SOC detects a ransomware infection affecting at least two workstations in your marketing department. The SOC manager declares an incident and activates the IR team.

8. What should be your IR team's first action?
   - **A.** Determine the scope of the infection across the organization
   - **B.** Isolate the marketing VLAN from the rest of the network
   - **C.** Disconnect the infected computers from the network
   - **D.** Determine why the EDR system failed to protect the workstations
9. Using your NDR system, you determine the external hosts from which the malware was downloaded and with which the infected systems were communicating. As part of the remediation phase, which of the following is the next best action to take with this information?
   - **A.** Determine whether the external hosts you identified are related to the incident
   - **B.** Block traffic to/from the external hosts that you identified
   - **C.** Visit the remote hosts using a forensic workstation to acquire evidence
   - **D.** Share the address of the hosts with your partners as indicators of compromise (IOCs)
10. Luckily, this version of ransomware is buggy, and you find a security researcher's blog with detailed instructions for how to decrypt infected systems. Which of the following approaches will best mitigate the incident and make the affected systems operational again?
    - **A.** Follow the directions to decrypt the systems and remove the malware
    - **B.** Reinstall from a golden master and restore the data from backups
    - **C.** Reinstall from a golden master even though you have no backups
    - **D.** Restore the systems from the last known-good system backup

Answers

1. **B.** Incident management encompasses seven phases according to the CISSP CBK: detection, response, mitigation, reporting, recovery, remediation, and lessons learned.
2. **B.** The goal of containment during the response phase is to prevent or reduce any further damage from this incident so that you can begin to mitigate and recover. Done properly, this buys the IR team time for a proper investigation and determination of the incident's root cause.
3. **D.** In the remediation phase, you decide which control changes (e.g., firewall or IDS/IPS rules) are needed to preclude this incident from happening again. Another aspect of remediation is the identification of indicators of attack (IOAs)

that can be used in the future to detect this attack in real time (i.e., as it is happening) as well as indicators of compromise (IOCs), which tell you when an attack has been successful and your security has been compromised.

4. **A.** The incident management policy (IMP) establishes authorities and responsibilities across the entire organization, identifies the incident response (IR) lead for the organization, and describes what every staff member is required to do with regard to incidents. The incident response plan (IRP) gets into the details of what should be done when responding to suspected incidents, and includes roles and responsibilities, incident classification, notifications, and operational tasks. A runbook is a collection of procedures that the IR team will follow for specific types of incidents.
5. **C.** Several steps need to be followed when gathering and extracting evidence from a scene. Once a computer has been confiscated, the first thing the computer forensics team should do is make an image of the hard drive. The team will work from this image instead of the original hard drive so that the original stays in a pristine state and the evidence on the drive is not accidentally corrupted or modified.
6. **C.** For evidence to be admissible, it must be relevant to the case, reliable, and legally obtained. For evidence to be reliable, it must be consistent with fact and must not be based on opinion or be circumstantial.
7. **D.** Recording devices can have a chilling effect on interviewees. Instead, have at least one notetaker in the room and, after the interview is complete, read back the notes to the interviewee to ensure their accuracy.
8. **B.** Having detected the incident, the next step is to respond by containing the damage that has been or is about to be done to your most critical assets. You could simply disconnect the infected systems from the network, but since there are multiple workstations and they are in the same department, it is probably better to isolate that entire VLAN until you can determine the true scope of the problem. Since this incident happened at the end of the workday, isolating the VLAN should have little or no impact on the marketing department.
9. **B.** In the remediation phase, you decide which security controls need to be put in place to prevent the attack from succeeding again. This includes controls that are hastily put into effect because you have high confidence that they will help in the short term. The situation in the question is a perfect example of when you bypass your change management process and quickly make changes to deal with the incident at hand. You probably want to share the IOCs with your partners (and perhaps your regional CERT), but that happens after you block the traffic.
10. **B.** You have a centralized backup system that was not affected, so you know you should have backups for all the workstations. The problem is that you may not know if any of the full-system backups also include the ransomware, so restoring systems from backups could bring you back to square one. It is best to reinstall the systems from golden masters and then restore only the data files. This process may take a bit longer, but it minimizes the risk of reinfection.

This page intentionally left blank

CHAPTER 23

Disasters

This chapter presents the following:

- Recovery strategies
- Disaster recovery processes
- Testing disaster recovery plans
- Business continuity

It wasn't raining when Noah built the ark.

—Howard Ruff

Disasters are just regular features in our collective lives. Odds are that, at some point, we will all have to deal with at least one disaster (if not more), whether it be in our personal world or professional world. And when that disaster hits, figuring out a way to deal with it in real time is probably not going to go all that well for the unprepared. This chapter is all about thinking of all the terrible things that might happen, and then ensuring we have strategies and plans to deal with them. This doesn't just mean recovering from the disaster, but also ensuring that the business continues to operate with as little disruption as possible.

As the old adage goes, no battle plan ever survived first contact with the enemy, which is the reason why we must test and exercise plans until our responses as individuals and organizations are so ingrained in our brains that we no longer need to think about them. As terrible and complex disasters unfold around us, we will do the right things reflexively. Does that sound a bit ambitious? Perhaps. Still, it is our duty as cybersecurity professionals to do what we can to get our organizations as close to that goal as realistically possible. Let's see how we go about doing this.

Recovery Strategies

In the previous chapters in this part of the book, we have discussed preventing and responding to security incidents, including various types of investigations, as part of standard security operations. These are things we do day in and day out. But what happens on those rare occasions when an incident has disastrous effects? That is the realm of disaster recovery and business continuity planning. *Disaster recovery (DR)* is the set of practices that enables an organization to minimize loss of, and restore, mission-critical

technology infrastructure after a catastrophic incident. *Business continuity (BC)* is the set of practices that enables an organization to continue performing its critical functions through and after any disruptive event. As you can see, DR is mostly in the purview of safety and contingency operations, while BC is much broader than that. Accordingly, we'll focus on DR for most of this chapter but circle back to our roles in BC as cybersecurity leaders.

EXAM TIP As CISSPs, we are responsible for disaster recovery because it deals mostly with information technology and security. We provide inputs and support for business continuity planning but normally are not the lead for it.

Before we go much further, recall that we discussed the role of *maximum tolerable downtime (MTD)* values in Chapter 2. In reality, basic MTD values are a good start, but are not granular enough for an organization to figure out what it needs to put into place to be able to absorb the impact of a disaster. MTD values are usually "broad strokes" that do not provide the details needed to pinpoint the actual recovery solutions that need to be purchased and implemented. For example, if the business continuity planning (BCP) team determines that the MTD value for the customer service department is 48 hours, this is not enough information to fully understand what redundant solutions or backup technology should be put into place. MTD in this example does provide a basic deadline that means if customer service is not up within 48 hours, the company may not be able to recover and everyone should start looking for new jobs.

As shown in Figure 23-1, more than just MTD metrics are needed to get production back to normal operations after a disruptive event. We will walk through each of these metric types and see how they are best used together.

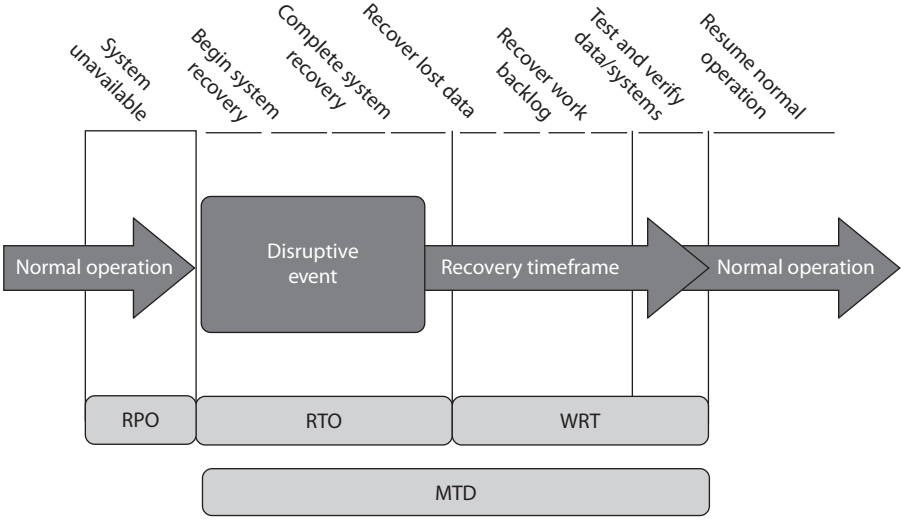


Figure 23-1 Metrics used for disaster recovery

The *recovery time objective (RTO)* is the maximum time period within which a mission-critical system must be restored to a designated service level after a disruption to avoid unacceptable consequences associated with a break in business continuity. The RTO value is smaller than the MTD value, because the MTD value represents the time after which an inability to recover significant operations will mean severe and perhaps irreparable damage to the organization's reputation or bottom line. The RTO assumes that there is a period of acceptable downtime. This means that an organization can be out of production for a certain period of time and still get back on its feet. But if the organization cannot get production up and running within the MTD window, it may be sinking too fast to properly recover.

The *work recovery time (WRT)* is the maximum amount of time available for certifying the functionality and integrity of restored systems and data so they can be put back into production. RTO usually deals with getting the infrastructure and systems back up and running, and WRT deals with ensuring business users can get back to work using them. Another way to think of WRT is as the remainder of the overall MTD value after the RTO has passed.

The *recovery point objective (RPO)* is the acceptable amount of data loss measured in time. This value represents the earliest point in time at which data must be recovered. The higher the value of data, the more funds or other resources that can be put into place to ensure a smaller amount of data is lost in the event of a disaster. Figure 23-2 illustrates the relationship and differences between the use of RPO and RTO values.

The MTD, RTO, RPO, and WRT values are critical to understand because they will be the basic foundational measures used when determining the type of recovery solutions an organization must put into place, so let's dig a bit deeper into them. As an example of RTO, let's say a company has determined that if it is unable to process product order requests for 12 hours, the financial hit will be too large for it to survive. This means that the MTD for order processing is 12 hours. To keep things simple, let's say that RTO and WRT are 6 hours each. Now, suppose that orders are processed using on-premises servers, on a site with no backup power sources, and an ice storm causes a power outage that will take days to restore. Without a plan and supporting infrastructure already in place, it would be close to impossible to migrate the servers and data to a site with power within 6 hours. The RTO (that is, the maximum time to move the servers and data) would not be met (to say nothing of the WRT) and it would likely exceed the MTD, putting the company at serious risk of collapse.

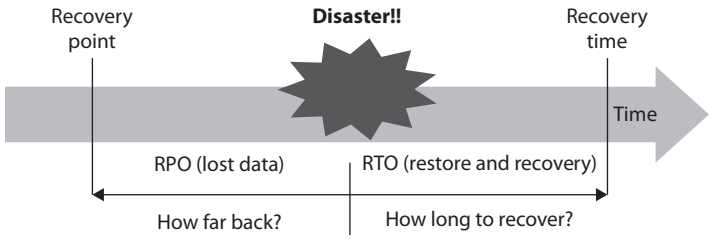


Figure 23-2 RPO and RTO measures in use

Now let's say that the same company did have a recovery site on a different power grid, and it was able to restore the order-processing services within a couple of hours, so it met the RTO requirement. But just because the systems are back online, the company still might have a critical problem. The company has to restore the data it lost during the disaster. Restoring data that is a week old does the company no good. The employees need to have access to the data that was being processed right before the disaster hit. If the company can only restore data that is a week old, then all the orders that were in some stage of being fulfilled over the last seven days could be lost. If the company makes an average of $25,000 per day in orders and all the order data was lost for the last seven days, this can result in a loss of $175,000 and a lot of unhappy customers. So just getting things up and running (meeting the RTO) is just part of the picture. Getting the necessary data in place so that business processes are up to date and relevant (RPO) is just as critical.

To take things one step further, let's say the company stood up the systems at its recovery site in two hours. It also had real-time data backup systems in place, so all of the necessary up-to-date data is restored. But no one actually tested the processes to recover data from backups, everyone is confused, and orders still cannot be processed and revenue cannot be collected. This means the company met its RTO requirement and its RPO requirement, but failed its WRT requirement, and thus failed the MTD requirement. Proper business recovery means *all* of the individual things have to happen correctly for the overall goal to be successful.

EXAM TIP An RTO is the amount of time it takes to recover from a disaster, and an RPO is the acceptable amount of data, measured in time, that can be lost from that same event.

The actual MTD, RTO, and RPO values are derived during the *business impact analysis (BIA)*, the purpose of which is to be able to apply criticality values to specific business functions, resources, and data types. A simplistic example is shown in Table 23-1. The company must have data restoration capabilities in place to ensure that mission-critical data is never older than one minute. The company cannot rely on something as slow as backup tape restoration, but must have a high-availability data replication solution in place. The RTO value for mission-critical data processing is two minutes or less. This means that the technology that carries out the processing functionality for this type of data cannot be down for more than two minutes. The company probably needs failover technology in place that will shift the load once it notices that a server goes offline.

Data Type	RPO	RTO
Mission critical	Continuous to 1 minute	Instantaneous to 2 minutes
Business critical	5 minutes	10 minutes
Business	3 hours	8 hours

Table 23-1 RPO and RTO Value Relationships

What Is the Difference Between Preventive Measures and Recovery Strategies?

Preventive mechanisms are put into place not only to try to reduce the possibility that the organization will experience a disaster, but also, if a disaster does hit, to lessen the amount of damage that will take place. Although the organization cannot stop a tornado from coming, for example, it could choose to move its facility from Tornado Alley to an area less prone to these weather events. As another example, the organization cannot stop a car from plowing into and taking out a transformer that it relies on for power, but it can have a separate power feed from a different transformer in case this happens.

Recovery strategies are processes designed to rescue the company after a disaster takes place. These processes integrate mechanisms such as establishing alternate sites for facilities, implementing emergency response procedures, and possibly activating the preventive mechanisms that have already been implemented.

In this same scenario, data that is classified as "Business" can be up to three hours old when the production environment comes back online, so a less frequent data replication process is acceptable. Because the RTO for business data is eight hours, the company can choose to have hot-swappable hard drives available instead of having to pay for the more complicated and expensive failover technology.

The DR team has to figure out what the company needs to do to actually recover the processes and services it has identified as being so important to the organization overall. In its business continuity and recovery strategy, the team closely examines the critical, agreed-upon business functions, and then evaluates the numerous recovery and backup alternatives that might be used to recover critical business operations. It is important to choose the right tactics and technologies for the recovery of each critical business process and service in order to assure that the set MTD values are met.

So what does the DR team need to accomplish? The team needs to actually define the recovery processes, which are sets of predefined activities that will be implemented and carried out in response to a disaster. More importantly, these processes must be constantly reevaluated and updated as necessary to ensure that the organization meets or exceeds the MTDs. It all starts with understanding the business processes that would have to be recovered in the aftermath of a disaster. Armed with that knowledge, the DR team can make good decisions about data backup, recovery, and processing sites, as well as overall services availability, all of which we explore in the next sections.

Business Process Recovery

A *business process* is a set of interrelated steps linked through specific decision activities to accomplish a specific task. Business processes have starting and ending points and are repeatable. The processes should encapsulate the knowledge about services, resources, and operations provided by an organization. For example, when a customer requests

to buy a book via a company's e-commerce site, the company's order fulfillment system must follow a business process such as this:

1. Validate that the book is available.
2. Validate where the book is located and how long it would take to ship it to the destination.
3. Provide the customer with the price and delivery date.
4. Verify the customer's credit card information.
5. Validate and process the credit card order.
6. Send the order to the book inventory location.
7. Send a receipt and tracking number to the customer.
8. Restock inventory.
9. Send the order to accounting.

The DR team needs to understand these different steps of the organization's most critical processes. The data is usually presented as a workflow document that contains the roles and resources needed for each process. The DR team must understand the following about critical business processes:

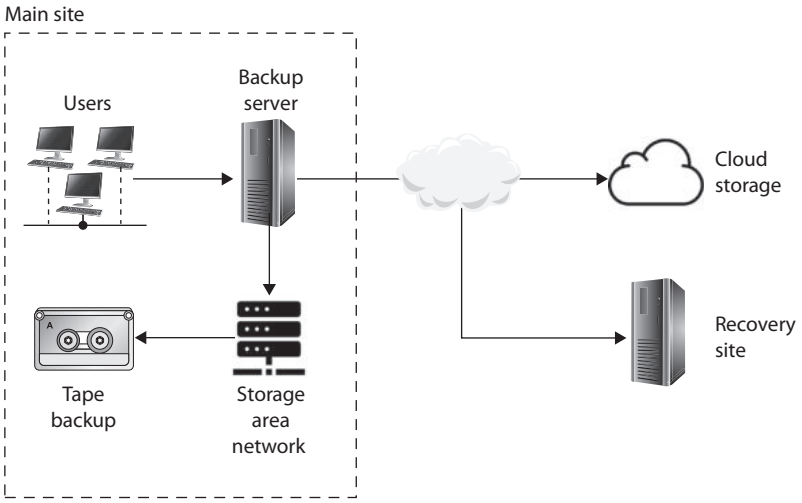
- Required roles
- Required resources
- Input and output mechanisms
- Workflow steps
- Required time for completion
- Interfaces with other processes

This will allow the team to identify threats and the controls to ensure the least amount of process interruption.

Data Backup

Data has become one of the most critical assets to nearly all organizations. It may include financial spreadsheets, blueprints on new products, customer information, product inventory, trade secrets, and more. In Chapter 2, we stepped through risk analysis procedures and, in Chapter 5, data classification. The DR team should not be responsible for setting up and maintaining the organization's data classification procedures, but the team should recognize that the organization is at risk if it does not have these procedures in place. This should be seen as a vulnerability that is reported to management. Management would need to establish another group of individuals who would identify the organization's data, define a loss criterion, and establish the classification structure and processes.

The DR team's responsibility is to provide solutions to protect this data and identify ways to restore it after a disaster. Data usually changes more often than hardware and software, so these backup or archival procedures must happen on a continual basis. The data backup process must make sense and be reasonable and effective. If data in the files changes several times a day, backup procedures should happen a few times a day or nightly to ensure all the changes are captured and kept. If data is changed once a month, backing up data every night is a waste of time and resources. Backing up a file and its corresponding changes is usually more desirable than having multiple copies of that one file. Online backup technologies usually record the changes to a file in a transaction log, which is separate from the original file.



The IT operations team should include a backup administrator, who is responsible for defining which data gets backed up and how often. These backups can be full, differential, or incremental, and are usually used in some type of combination with each other. Most files are not altered every day, so, to save time and resources, it is best to devise a backup plan that does not continually back up data that has not been modified. So, how do we know which data has changed and needs to be backed up without having to look at every file's modification date? This is accomplished by setting an *archive bit* to 1 if a file has been modified. The backup software reviews this bit when making its determination of whether the file gets backed up and, if so, clears the bit when it's done.

The first step is to do a *full backup*, which is just what it sounds like—all data is backed up and saved to some type of storage media. During a full backup, the archive bit is cleared, which means that it is set to 0. An organization can choose to do full backups only, in which case the restoration process is just one step, but the backup and restore processes could take a long time.

Most organizations choose to combine a full backup with a differential or incremental backup. A *differential process* backs up the files that have been modified since the *last full backup*. When the data needs to be restored, the full backup is laid down first, and then

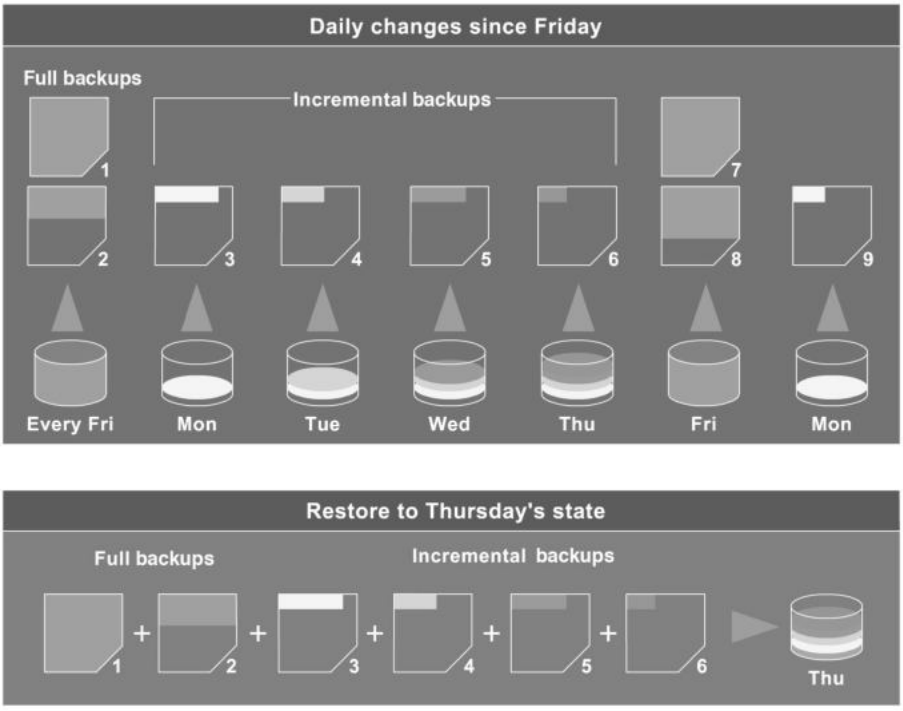


Figure 23-3 Backup software steps

the most recent differential backup is put down on top of it. The differential process does not change the archive bit value.

An *incremental process* backs up all the files that have changed since the *last full or incremental backup* and sets the archive bit to 0. When the data needs to be restored, the full backup data is laid down, and then each incremental backup is laid down on top of it in the proper order (see Figure 23-3). If an organization experienced a disaster and it used the incremental process, it would first need to restore the full backup on its hard drives and lay down every incremental backup that was carried out before the disaster took place (and after the last full backup). So, if the full backup was done six months ago and the operations department carried out an incremental backup each month, the backup administrator would restore the full backup and start with the older incremental backups taken since the full backup and restore each one of them until they were all restored.

Which backup process is best? If an organization wants the backup and restoration processes to be simple, it can carry out just full backups—but this may require a lot of hard drive space and time. Although using differential and incremental backup processes is more complex, it requires fewer resources and less time. A differential backup takes more time in the backing-up phase than an incremental backup, but it also takes less time

to restore than an incremental backup because carrying out restoration of a differential backup happens in two steps, whereas in an incremental backup, every incremental backup must be restored in the correct sequence.

Whatever the organization chooses, it is important to not mix differential and incremental backups. This overlap could cause files to be missed, since the incremental backup changes the archive bit and the differential backup does not.

Critical data should be backed up and stored onsite *and* offsite. The onsite backups should be easily accessible for routine uses and should provide a quick restore process so operations can return to normal. However, onsite backups are not enough to provide real protection. The data should also be held in an offsite facility in case of disasters. One decision the CISO needs to make is where the offsite location should be in reference to the main facility. The closer the offsite backup storage site is, the easier it is to access, but this can put the backup copies in danger if a large-scale disaster manages to take out the organization's main facility and the backup facility. It may be wiser to choose a backup facility farther away, which makes accessibility harder but reduces the risk. Some organizations choose to have more than one backup facility: one that is close and one that is farther away.

Backup Storage Strategies

A backup strategy must take into account that failure can take place at any step of the process, so if there is a problem during the backup or restoration process that could corrupt the data, there should be a graceful way of backing out or reconstructing the data

Restoring Data from Backups: A Cautionary Tale

Can we actually restore data from backups? Backing up data is a wonderful thing in life, but making sure it can be properly restored is even better. Many organizations have developed a false sense of security based on the fact that they have a very organized and effective process of backing up their data. That sense of security can disappear in seconds when an organization realizes in a time of crisis that its restore processes do not work. For example, one company had paid an offsite backup facility to use a courier to collect its weekly backup tapes and transport them to the offsite facility for safekeeping. What the company did not realize was that this courier used the subway and many times set the tapes on the ground while waiting for the subway train. A subway has many large engines that create their own magnetic field. This can have the same effect on media as large magnets, meaning that the data can be erased or corrupted. The company never tested its restore processes and eventually experienced a disaster. Much to its surprise, it found out that three years of data were corrupted and unusable.

Many other stories and experiences like this are out there. Don't let your organization end up as an anecdote in someone else's book because it failed to verify that its backups could be restored.

from the beginning. The procedures for backing up and restoring data should be easily accessible and comprehensible even to operators or administrators who are not intimately familiar with a specific system. In an emergency situation, the same person who always does the backing up and restoring may not be around, or outsourced consultants may need to be temporarily hired to meet the restoration time constraints.

There are four commonly used backup strategies that you should be aware of:

- **Direct-attached storage** The backup storage is directly connected to the device being backed up, typically over a USB cable. This is better than nothing, but is not really well suited for centralized management. Worse yet, many ransomware attacks look for these attached storage devices and encrypt them too.
- **Network-attached storage (NAS)** The backup storage is connected to the device over the LAN and is usually a storage area network (SAN) managed by a backup server. This approach is usually centrally managed and allows IT administrators to enforce data backup policies. The main drawback is that, if a disaster takes out the site, the data may be lost or otherwise be rendered inaccessible.
- **Cloud storage** Many organizations use cloud storage as either the primary or secondary repository of backup data. If this is done on a virtual private cloud, it has the advantage of providing offsite storage so that, even if the organization's site is destroyed by a disaster, the data is available for recovery. Obviously, WAN connectivity must be reliable and fast enough to support this strategy if it is to be effective.
- **Offline media** As ransomware becomes more sophisticated, we are seeing more instances of attackers going after NAS and cloud storage. If your data is critical enough that you have to decrease the risk of it being lost as close to zero as you can, you may want to consider offline media such as tape backups, optical discs, or even external drives that are disconnected after each backup (and potentially removed offsite). This is the slowest and most expensive approach, but is also the most resistant to attacks.

Electronic vaulting and remote journaling are other solutions that organizations should be aware of. *Electronic vaulting* makes copies of files as they are modified and periodically transmits them to an offsite backup site. The transmission does not happen in real time, but is carried out in batches. So, an organization can choose to have all files that have been changed sent to the backup facility every hour, day, week, or month. The information can be stored in an offsite facility and retrieved from that facility in a short amount of time.

This form of backup takes place in many financial institutions, so when a bank teller accepts a deposit or withdrawal, the change to the customer's account is made locally to that branch's database and to the remote site that maintains the backup copies of all customer records.

Electronic vaulting is a method of transferring bulk information to offsite facilities for backup purposes. *Remote journaling* is another method of transmitting data offsite, but this usually only includes moving the journal or transaction logs to the offsite facility, not the actual files. These logs contain the deltas (changes) that have taken place to the individual files. Continuing with the bank example, if and when data is corrupted and needs to be restored, the bank can retrieve these logs, which are used to rebuild the lost data. Journaling is efficient for database recovery, where only the reapplication of a series of changes to individual records is required to resynchronize the database.

EXAM TIP Remote journaling takes place in real time and transmits only the file deltas. Electronic vaulting takes place in batches and moves the entire file that has been updated.

An organization may need to keep different versions of software and files, especially in a software development environment. The object and source code should be backed up along with libraries, patches, and fixes. The offsite facility should mirror the onsite facility, meaning it does not make sense to keep all of this data at the onsite facility and only the source code at the offsite facility. Each site should have a full set of the most current and updated information and files.

Another software backup technology is *tape vaulting*. Many organizations back up their data to tapes that are then manually transferred to an offsite facility by a courier or an employee. This manual process can be error-prone, so some organizations use *electronic tape vaulting*, in which the data is sent over a serial line to a backup tape system at the offsite facility. The company that maintains the offsite facility maintains the systems and changes out tapes when necessary. Data can be quickly backed up and retrieved when necessary. This technology improves recovery speed, reduces errors, and allows backups to be run more frequently.

Data repositories commonly have replication capabilities, so that when changes take place to one repository (i.e., database) they are replicated to all the other repositories within the organization. The replication can take place over telecommunication links, which allow offsite repositories to be continuously updated. If the primary repository goes down or is corrupted, the replication flow can be reversed, and the offsite repository updates and restores the primary repository. Replication can be asynchronous or synchronous. *Asynchronous replication* means the primary and secondary data volumes are out of sync. Synchronization may take place in seconds, hours, or days, depending upon the technology in place. With *synchronous replication*, the primary and secondary repositories are always in sync, which provides true real-time duplication. Figure 23-4 shows how offsite replication can take place.

The DR team must balance the cost to recover against the cost of the disruption. The balancing point becomes the recovery time objective. Figure 23-5 illustrates the relationship between the cost of various recovery technologies and the provided recovery times.

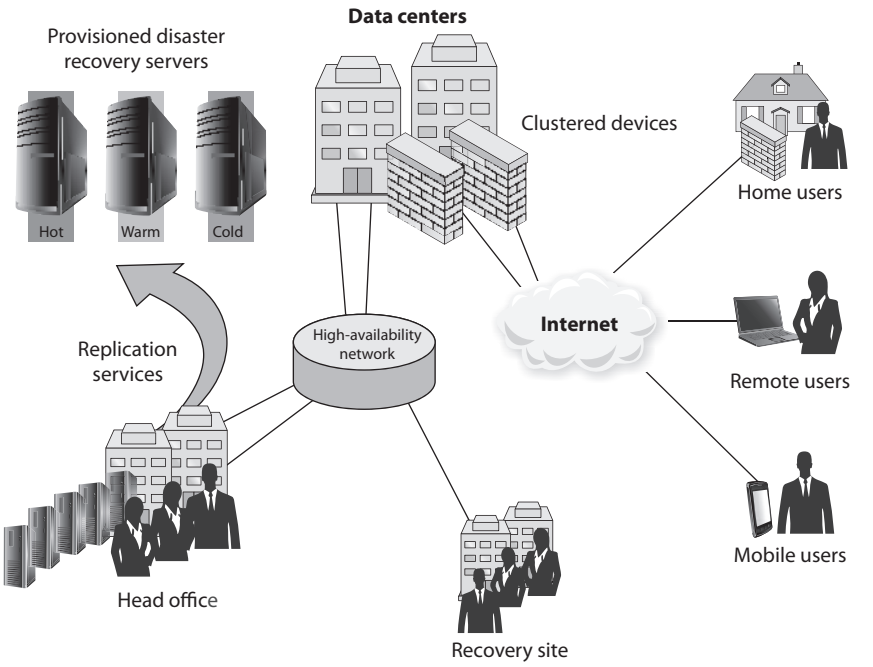


Figure 23-4 Offsite data replication for data recovery purposes

Choosing a Software Backup Facility

An organization needs to address several issues and ask specific questions when it is deciding upon a storage facility for its backup materials. The following list identifies just some of the issues that an organization needs to consider before committing to a specific vendor for this service:

- Can the media be accessed in the necessary timeframe?
- Is the facility closed on weekends and holidays, and does it only operate during specific hours of the day?
- Are the facility's access control mechanisms tied to an alarm and/or the police station?
- Does the facility have the capability to protect the media from a variety of threats?
- What is the availability of a bonded transport service?
- Are there any geographical environmental hazards such as floods, earthquakes, tornadoes, and so on that might affect the facility?
- Does the facility have a fire detection and suppression system?
- Does the facility provide temperature and humidity monitoring and control?
- What type of physical, administrative, and logical access controls are used?

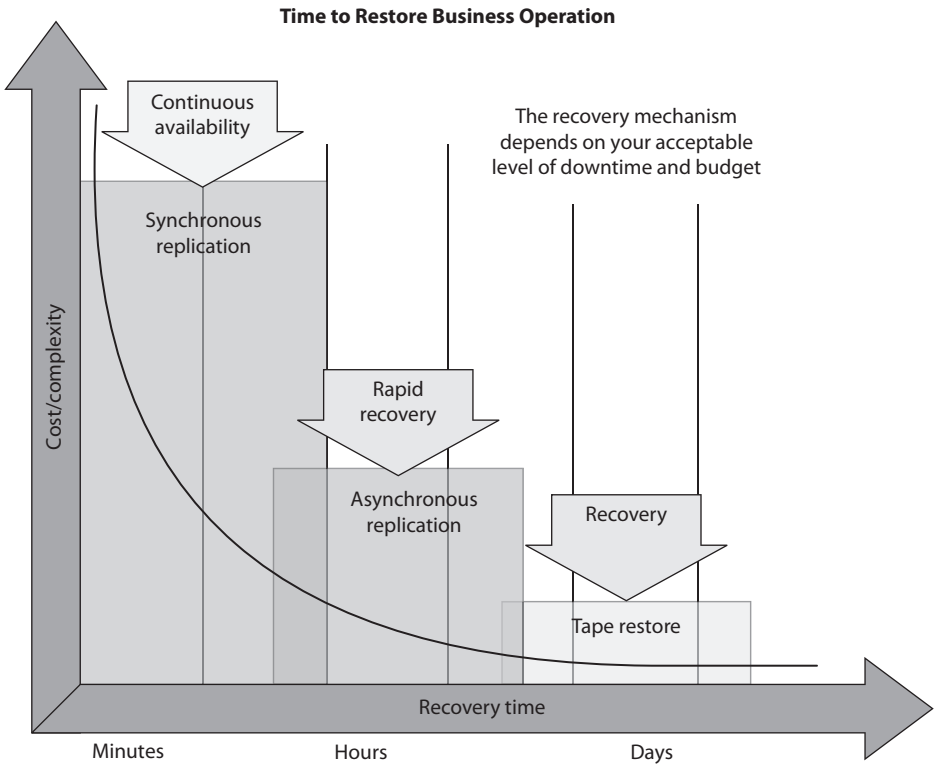


Figure 23-5 The criticality of data recovery will dictate the recovery solution.

The questions and issues that need to be addressed will vary depending on the type of organization, its needs, and the requirements of a backup facility.

Documentation

Documentation seems to be a dreaded task to most people, who will find many other tasks to take on to ensure they are not the ones stuck with documenting processes and procedures. However, without proper documentation, even an organization that does a terrific job of backing up data to an offsite facility will be scrambling to figure which backups it needs when a disaster hits.

Restoration of files can be challenging but restoring a whole environment that was swept away in a flood can be overwhelming, if not impossible. Procedures need to be documented because when they are actually needed, it will most likely be a chaotic and frantic atmosphere with a demanding time schedule. The documentation may need to include information on how to install images, configure operating systems and servers, and properly install utilities and proprietary software. Other documentation could include a calling tree, which outlines who should be contacted, in what order, and

Storing Business Continuity and Disaster Recovery Plans

Once the business continuity and disaster recovery plans are completed, where should they be stored? Should the organization have only one copy and keep it safely in a file cabinet next to Bob so that he feels safe? Nope. There should be two or three copies of these plans. One copy may be at the primary location, but the other copies should be at other locations in case the primary facility is destroyed. This reduces the risk of not having access to the plans when needed.

These plans should not be stored in a file cabinet, but rather in a fire-resistant safe. When they are stored offsite, they need to be stored in a way that provides just as much protection as the primary site would provide.

who is responsible for doing the calling. The documentation must also contain contact information for specific vendors, emergency agencies, offsite facilities, and any other entity that may need to be contacted in a time of need.

Most network environments evolve over time. Software is installed on top of other software, configurations are altered over the years to properly work in a unique environment, and service packs and patches are routinely installed to fix issues and update software. To expect one person or a group of people to go through all these steps during a crisis and end up with an environment that looks and behaves exactly like the original environment and in which all components work together seamlessly may be a lofty dream.

So, the dreaded task of documentation may be the saving grace one day. It is an essential piece of business, and therefore an essential piece in disaster recovery and business continuity. It is, therefore, important to make one or more roles responsible for proper documentation. As with all the items addressed in this chapter, simply saying "All documentation will be kept up to date and properly protected" is the easy part—saying and doing are two different things. Once the DR team identifies tasks that must be done, the tasks must be assigned to individuals, and those individuals have to be accountable. If these steps are not taken, the organization may have wasted a lot of time and resources defining these tasks, and still be in grave danger if a disaster occurs.

Human Resources

One of the resources commonly left out of the DR equation is people. An organization may restore its networks and critical systems and get business functions up and running, only to realize it doesn't know the answer to the question, "Who will take it from here?" The area of human resources is a critical component to any recovery and continuity process, and it needs to be fully thought out and integrated into the plan.

What happens if we have to move to an offsite facility that is 250 miles away? We cannot expect people to drive back and forth from home to work. Should we pay for temporary housing for the necessary employees? Do we have to pay their moving costs? Do we need to hire new employees in the area of the offsite facility? If so, what skill set

do we need from them? These are all important questions for the organization's senior leaders to answer.

If a large disaster takes place that affects not only the organization's facility but also surrounding areas, including housing, employees will be more worried about their families than their organization. Some organizations assume that employees will be ready and available to help them get back into production, when in fact they may need to be at home because they have responsibilities to their families.

Regrettably, some employees may be killed or severely injured in the disaster, and the organization should have plans in place to replace employees quickly through a temporary employment agency or a job recruiter. This is an extremely unfortunate scenario to contemplate, but it is part of reality. The team that considers all threats and is responsible for identifying practical solutions needs to think through all of these issues.

Organizations should already have *executive succession planning* in place. This means that if someone in a senior executive position retires, leaves the organization, or is killed, the organization has predetermined steps to carry out to ensure a smooth transition to that executive's replacement. The loss of a senior executive could tear a hole in the organization's fabric, creating a leadership vacuum that must be filled quickly with the right individual. The line-of-succession plan defines who would step in and assume responsibility for this role. Many organizations have "deputy" roles. For example, an organization may have a deputy CIO, deputy CFO, and deputy CEO ready to take over the necessary tasks if the CIO, CFO, or CEO becomes unavailable.

Often, larger organizations also have a policy indicating that two or more of the senior staff cannot be exposed to a particular risk at the same time. For example, the CEO and president cannot travel on the same plane. If the plane were to crash and both individuals were killed, then the company could face a leadership crisis. This is why you don't see the president of the United States and the vice president together too often. It is not because they don't like each other and thus keep their distance from each other. It is because there is a policy indicating that to protect the United States, its top leaders cannot be under the same risk at the same time.

Recovery Site Strategies

Disruptions, in BCP terms, are of three main types: nondisasters, disasters, and catastrophes. A *nondisaster* is a disruption in service that has significant but limited impact on the conduct of business processes at a facility. The solution could include hardware, software, or file restoration. A *disaster* is an event that causes the entire facility to be unusable for a day or longer. This usually requires the use of an alternate processing facility and restoration of software and data from offsite copies. The alternate site must be available to the organization until its main facility is repaired and usable. A *catastrophe* is a major disruption that destroys the facility altogether. This requires both a short-term solution, which would be an offsite facility, and a long-term solution, which may require rebuilding the original facility. Disasters and catastrophes are rare compared to nondisasters, thank goodness.

When dealing with disasters and catastrophes, an organization has three basic options: select a dedicated site that the organization owns and operates itself; lease a commercial facility, such as a "hot site" that contains all the equipment and data needed to quickly

restore operations; or enter into a formal agreement with another facility, such as a service bureau, to restore its operations. When choosing the right solution for its needs, the organization evaluates each alternative's ability to support its operations, to do it within an acceptable timeframe, and to have a reasonable cost.

An important consideration with third parties is their reliability, both in normal times and during an emergency. Their reliability can depend on considerations such as their track record, the extent and location of their supply inventory, and their access to supply and communication channels. Organizations should closely query the management of the alternative facility about such things as the following:

- How long will it take to recover from a certain type of incident to a certain level of operations?
- Will it give priority to restoring the operations of one organization over another after a disaster?
- What are its costs for performing various functions?
- What are its specifications for IT and security functions? Is the workspace big enough for the required number of employees?

To recover from a disaster that prevents or degrades use of the primary site temporarily or permanently, an organization must have an offsite backup facility available. Generally, an organization establishes contracts with third-party vendors to provide such services. The client pays a monthly fee to retain the right to use the facility in a time of need, and then incurs an activation fee when the facility actually has to be used. In addition, a daily or hourly fee is imposed for the duration of the stay. This is why service agreements for backup facilities should be considered a short-term solution, not a long-term solution.

It is important to note that most recovery site contracts do not promise to house the organization in need at a specific location, but rather promise to provide what has been contracted for somewhere within the organization's locale. On, and subsequent to, September 11, 2001, many organizations with Manhattan offices were surprised when they were redirected by their backup site vendor not to sites located in New Jersey (which were already full), but rather to sites located in Boston, Chicago, or Atlanta. This adds yet another level of complexity to the recovery process, specifically the logistics of transporting people and equipment to unplanned locations.

An organization can choose from three main types of leased or rented offsite recovery facilities:

- **Hot site** A facility that is fully configured and ready to operate within a few hours. All the necessary equipment is already installed and configured. In many cases, the remote data backup services are included, so the RPO can be down to an hour or even less. These sites are a good choice for an organization with a very small MTD. Of course, the organization should conduct regular tests (annually, at least) to ensure the site is functioning in the necessary state of readiness.

  The hot site is, by far, the most expensive of the three types of offsite facilities. The organization has to pay for redundant hardware and software, in addition

to the expenses of the site itself. Organizations that use hot sites as part of their recovery strategy tend to limit them to mission-critical systems only.

- **Warm site** A facility that is usually partially configured with some equipment, such as HVAC, and foundational infrastructure components, but does not include all the hardware needed to restore mission-critical business functions. Staging a facility with duplicate hardware and computers configured for immediate operation is extremely expensive, so a warm site provides a less expensive alternate. These sites typically do not have data replicated to them, so backups would have to be delivered and restored onto the warm site systems after a disaster.

  The warm site is the most widely used model. It is less expensive than a hot site, and can be up and running within a reasonably acceptable time period. It may be a better choice for organizations that depend on proprietary and unusual hardware and software, because they will bring their own hardware and software with them to the site after the disaster hits. Drawbacks, however, are that much of the equipment has to be procured, delivered to, and configured at the warm site after the fact, and testing will be more difficult. Thus, an organization may not be certain that it will in fact be able to return to an operating state within its RTO.
- **Cold site** A facility that supplies the basic environment, electrical wiring, HVAC, plumbing, and flooring but none of the equipment or additional services. A cold site is essentially an empty data center. It may take weeks to get the site activated and ready for work. The cold site could have equipment racks and dark fiber (fiber that does not have the circuit engaged) and maybe even desks. However, it would require the receipt of equipment from the client, since it does not provide any.

  The cold site is the least expensive option, but takes the most time and effort to actually get up and functioning right after a disaster, as the systems and software must be delivered, set up, and configured. Cold sites are often used as backups for call centers, manufacturing plants, and other services that can be moved lock, stock, and barrel in one shot.

After a catastrophic loss of the primary facility, some organizations will start their recovery in a hot or warm site, and transfer some operations over to a cold site after the latter has had time to set up.

It is important to understand that the different site types listed here are provided by service bureaus. A *service bureau* is a company that has additional space and capacity to provide applications and services such as call centers. An organization pays a monthly subscription fee to a service bureau for this space and service. The fee can be paid for contingencies such as disasters and emergencies. You should evaluate the ability of a service bureau to provide services just as you would evaluate divisions within your own organization, particularly on matters such as its ability to alter or scale its software and hardware configurations or to expand its operations to meet the needs of a contingency.

NOTE Related to a service bureau is a *contingency supplier*; its purpose is to supply services and materials temporarily to an organization that is experiencing an emergency. For example, a contingency supplier might provide raw materials such as heating fuel or backup telecommunication services. In considering contingency suppliers, the BCP team should think through considerations such as the level of services and materials a supplier can provide, how quickly a supplier can ramp up to supply them, and whether the supplier shares similar communication paths and supply chains as the affected organization.

Most organizations use warm sites, which have some devices such as networking equipment, some computers and data storage, but very little else. These organizations usually cannot afford a hot site, and the extra downtime would not be considered detrimental. A warm site can provide a longer-term solution than a hot site. Organizations that decide to go with a cold site must be able to be out of operation for a week or two. The cold site usually includes power, raised flooring, climate control, and wiring.

The following provides a quick overview of the differences between offsite facilities.

Hot site advantages:

- Ready within hours or even minutes for operation
- Highly available
- Usually used for short-term solutions, but available for longer stays
- Recovery testing is easy

Hot site disadvantages:

- Very expensive
- Limited systems

Tertiary Sites

An organization may recognize the danger of the primary recovery site not being available when needed. This could be the case if the service provider assumes that not every customer will attempt to occupy the site at the same time, and then a major regional disaster affects more organizations than anticipated. It could also happen if a disaster affects the recovery site itself (e.g., fire, flood). Mitigating this risk could require a *tertiary site*, a backup recovery site just in case the primary is unavailable. The tertiary site is sometimes referred to as a "backup to the backup." This is basically plan B if plan A does not work out. Obviously, this is a very expensive proposition, so its costs should be balanced with the risks it is intended to mitigate.

Warm and cold site advantages:

- Less expensive
- Available for longer timeframes because of the reduced costs
- Practical for proprietary hardware or software use

Warm and cold site disadvantages:

- Limited ability to perform recovery testing
- Resources for operations not immediately available

Reciprocal Agreements

Another approach to alternate offsite facilities is to establish a *reciprocal agreement* with another organization, usually one in a similar field or that has similar technological infrastructure. This means that organization A agrees to allow organization B to use its facilities if organization B is hit by a disaster, and vice versa. This is a cheaper way to go than the other offsite choices, but it is not always the best choice. Most environments are maxed out pertaining to the use of facility space, resources, and computing capability. To allow another organization to come in and work out of the same shop could prove to be detrimental to both organizations. Whether it can assist the other organization while tending effectively to its own business is an open question. The stress of two organizations working in the same environment could cause tremendous levels of tension. If it did work out, it would only provide a short-term solution. Configuration management could be a nightmare. Does the other organization upgrade to new technology and retire old systems and software? If not, one organization's systems may become incompatible with those of the other.

If your organization allows another organization to move into its facility and work from there, you may have a solid feeling about your friend, the CEO, but what about all of her employees, whom you do not know? The mixing of operations could introduce many security issues. Now you have a new subset of people who may need to have privileged and direct access to your resources in the shared environment. Close attention needs to be paid when assigning these other people access rights and permissions to your critical assets and resources, if they need access at all. Careful testing is recommended to see if one organization or the other can handle the extra loads.

Offsite Location

When choosing a backup facility, it should be far enough away from the original site so that one disaster does not take out both locations. In other words, it is not logical to have the backup site only a few miles away if the organization is concerned about tornado damage, because the backup site could also be affected or destroyed. There is a rule of thumb that suggests that alternate facilities should be, at a bare minimum, at least 5 miles away from the primary site, while 15 miles is recommended for most low-to-medium critical environments, and 50 to 200 miles is recommended for critical operations, to give maximum protection in cases of regional disasters.

Reciprocal agreements have been known to work well in specific businesses, such as newspaper printing. These businesses require very specific technology and equipment that is not available through any subscription service. These agreements follow a "you scratch my back and I'll scratch yours" mentality. For most other organizations, reciprocal agreements are generally, at best, a secondary option for disaster protection. The other issue to consider is that these agreements are usually not enforceable because they're not written in legally binding terms. This means that although organization A said organization B could use its facility when needed, when the need arises, organization A may not have a legal obligation to fulfill this promise. However, there are still many organizations who do opt for this solution either because of the appeal of low cost or, as noted earlier, because it may be the only viable solution in some cases.

Organizations that have a reciprocal agreement need to address the following important issues before a disaster hits:

- How long will the facility be available to the organization in need?
- How much assistance will the staff supply in integrating the two environments and ongoing support?
- How quickly can the organization in need move into the facility?
- What are the issues pertaining to interoperability?
- How many of the resources will be available to the organization in need?
- How will differences and conflicts be addressed?
- How does change control and configuration management take place?
- How often can exercising and testing take place?
- How can critical assets of both organizations be properly protected?

A variation on a reciprocal agreement is a consortium, or *mutual aid agreement*. In this case, more than two organizations agree to help one another in case of an emergency. Adding multiple organizations to the mix, as you might imagine, can make things even more complicated. The same concerns that apply with reciprocal agreements apply here, but even more so. Organizations entering into such agreements need to formally and legally document their mutual responsibilities in advance. Interested parties, including the legal and IT departments, should carefully scrutinize such accords before the organization signs onto them.

Redundant Sites

Some organizations choose to have a *redundant site*, or mirrored site, meaning one site is equipped and configured exactly like the primary site, which serves as a redundant environment. The business-processing capabilities between the two sites can be completely synchronized. A redundant site is owned by the organization and mirrors the original production environment. A redundant site has clear advantages: it has full availability, is ready to go at a moment's notice, and is under the organization's complete control. This is, however, one of the most expensive backup facility options, because a full