White Paper

Cisco ASA CX delivers context-aware security

Revolutionary architecture, evolutionary deployment

Cisco Security Technology Group

February, 2012

Contents

Abstract

The appification of the web and the consumerization of IT are radically changing the enterprise security landscape. The security apparatus at today's enterprises needs to be aware not only of applications and users accessing the infrastructure, but also of the device in use, the location of the user and the time-of-day before granting access. Such **context-aware** security requires a rethinking of the firewall architecture. Yet, the context-aware firewall needs to interwork with enterprises' existing security infrastructure. The Cisco® ASA CX not only provides enterprises with breakthrough security, it also protects their existing investments in security.

Big changes in the security landscape

The Internet today is seeing a major trend towards appification. Today's average employees are increasingly getting accustomed to accessing web-based services via consumer apps on mobile devices — both for personal and professional use. Such employees are demanding the ability to use their favorite apps on the corporate network. Further, consumerization of IT has also taken hold — employees expect to be able to use their personal phones, tablets and laptops at work and to do so with ease and without fear of being attacked by hackers. Finally, employees expect continued and reliable access to enterprise applications that are the backbone of their everyday workflow.

Security architects and operators are thus under sustained pressure to allow network access to the full range of applications, websites and devices. At the same time they must protect the enterprise from malicious attacks and maintain reliable access to traditional enterprise applications. The security staff today can't easily gain visibility into applications that are on the network and the devices and users that are accessing these applications. Without such visibility devising a meaningful defense is all but impossible. Even when the security staff knows the applications on the network, it lacks granular control over those applications and the ability to selectively grant users and devices access to them.

Hackers understand the constraints faced by the security staff. And they have devised ways around the typical security policies via applications that can hop ports, web-widgets that result in drive-by malicious downloads even when a user restricts surfing to reputable sites and advanced threats that use multiple weaknesses to steal confidential enterprise data.

Context-aware security needed to protect enterprises

To respond to the changing security landscape, enterprises need to enforce security based on the complete context of a situation. This context includes the identity of the user (who), the application or website that the user is trying to access (what), the origin of the access attempt (where), the time of the attempted access (when) and the properties of the device used for the access (how).

The firewall has long been the mainstay of an enterprise's defense perimeter. To fight off modern-day threats, the firewall needs to be made "context-aware." That is, it needs to extract the user and application identity, origin of the access and the type of device used for the access, and then permit or deny the access based on these attributes, in accordance with configured policy. In addition, the firewall must have the ability to detect and protect against emerging threats.

The firewall is the right place to obtain the full context of the traffic flowing through the network. The firewall already sees all the traffic crossing the trust boundary between the enterprise network and the world at large. Assessing this traffic at the firewall for conformance with corporate policies would seem logical provided the firewall had the ability to inspect the traffic for its full context. Unfortunately, most firewalls on the market today are too inflexible to easily accommodate context-awareness. Not only are these firewalls unable to extract the full context of a flow, they also lack the ability to enforce granular policies such as permitting access to Facebook but denying access to games on Facebook or permitting finance employees access to a sensitive enterprise database but denying the same to other employees.

In the instances where firewalls do have some context-aware attributes, translating the desired corporate policy to firewall rules throws up a roadblock. Current firewall rules expressed in the language of IP address, protocols and port numbers are already difficult to maintain. Manually adding more rules on top of these to account for new applications, users and mobile devices, takes significant time and effort. Even when these rules are successfully installed, the result is an expanded rule-set that only make rules maintenance more brittle and threatens the reliability of access to the traditional enterprise applications.

Clearly, the firewall needs to enable security operators to express context-aware policy in a much more business friendly language. Instead of forcing the use of IP addresses and protocol and port numbers to form convoluted rules, security operators need to have the ability to simply “block Skype” or to “allow Yahoo! Messenger but stop file transfers.” Policy expressed in such business-friendly language is faster to implement, is easier to maintain and results in substantial reduction in the complexity and fragility of security management.

A context-aware firewall goes beyond the next-generation firewalls on the market today. Most of the current next-generation firewalls are technically capable of recognizing only a portion of the context of a flow and are ineffective against emerging threats. Many do not have a user-friendly solution to manage security policies. While such firewalls may improve upon traditional firewalls in some deployments, they do not provide the comprehensive security solution that enterprises need.

Revolutionary architecture enables context-awareness

While context-awareness inside firewalls is necessary to protect enterprises against modern-day threats, bringing the full set of required capabilities to market has proven impossible until now: Cisco has been able to develop a fully context-aware firewall by carefully following a set of key principles.

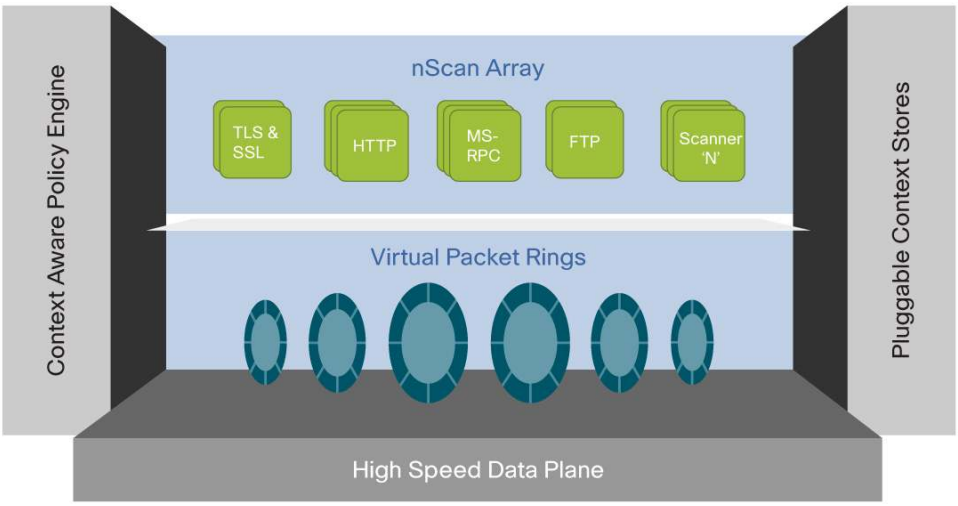
First, the firewall needs to be **built from the ground up** to detect, propagate, and act on the full context of a flow. Second, the firewall implementation needs to be **complete** — it needs to implement all of the features that enterprises need, not just the ones that protect against recent threats. Third, the architecture needs to be **flexible** to accommodate mechanisms to fight future threats, while requiring minimal maintenance downtime. Fourth, it needs the ability to use both **local and global context** — the ability to use information from both within and outside an enterprise’s network. Fifth, the firewall needs to deliver context-awareness while maintaining **high performance**.

The Cisco ASA CX is built using a new firewall architecture that can perform context-aware inspection and make context-aware decisions without compromising on any of the principles outlined above.

Re-imagining the firewall

The ASA CX re-imagines the firewall using four architectural constructs that work in concert (Figure 1). The first is a **Virtual Packet Ring** that enables packets to flow from the hardware interfaces into main memory and back out at high speeds. The second is an **nScan Array** that inspects multiple flows in parallel to obtain the full context of each flow and enforce the policy decision associated with that context. The third is a **Context-aware Policy Engine** that the nScan Array uses to make policy decisions. The fourth is a set of **Pluggable Context Stores** that consist of dynamically update-able databases that the nScan Array consults to make enforcement decisions. Together, these architectural constructs make it possible for the ASA CX to adhere to the principles outlined above.

Figure 1. The ASA CX Architecture



The ASA CX's Virtual Packet Rings are implemented using shared memory. Each flow is mapped into a Virtual Packet Ring such that no subsequent data copies are needed as the flow is inspected by the nScan Array. With Virtual Packet Rings, data packets enter and exit without being copied multiple times. The avoidance of copies in turn enhances the packet throughput of the system. It also makes the system's performance more predictable and reduces the latency experienced by the end user who initiated the flow.

The nScan Array is a set of engines that inspect incoming flows by accessing packets maintained by the Virtual Packet Rings. Multiple inspection engines work in parallel on a flow to extract the full context of the flow and make a permit/deny decision on the flow. For example an incoming flow may be subjected to an HTTP inspection in parallel with an IPS inspection. A flow is allowed through the firewall only when both the HTTP and IPS inspectors determine that the flow is permissible. The inspection engines in the nScan Array can be dynamically updated without requiring an upgrade of the entire system or forcing downtime. This enables the ASA CX to defend against new exploits with minimal time lag from the exploit's discovery, and with minimal operator intervention.

The inspection engines in the nScan Array use the Context-aware Policy Engine to make policy decisions and take enforcement actions. The policy engine is a highly optimized data-structure that matches the attributes of a flow to the appropriate policy and its associated action. The inspection engines also use data in the Pluggable Context Stores to arrive at inspection decisions. Consider the case where a security operator has configured a policy to prevent all but a specific set of users from accessing a corporate database. In this case, when a new flow comes up for evaluation, an inspection engines first looks up the policy engine to determine the appropriate action. In doing so, the engine recognizes the need to extract the user identity associated with the traffic flow. Thus it accesses the user ID context store, obtains the user ID associated with the flow and tests this user ID against the users permitted access to the database. If the extracted user ID matches one of the IDs configured in the policy engine, the flow is permitted. If not, the flow is denied access.

There are multiple context stores in the system, one each for the different types of entities that are needed by the inspection engines. Thus there are context stores for user IDs, devices, URLs, and so on. Conceptually, context stores fall into two categories: those that are generated locally (e.g., user IDs and devices) and those that are obtained from an external source (e.g., domain and URL databases). Regardless of the type, each store is dynamically refreshed so that the inspection engines have the most up-to-date information while making their inspection decisions.

As soon as an inspection engine finishes working on a flow, it moves off to examine the next flow under consideration. Depending on the traffic pattern and the configured policies, the ASA CX system may determine that it needs additional inspection engines of a particular type (e.g., HTTP) and fewer inspection engines of another type (e.g., FTP). The system dynamically adjusts the nScan Array to contain the optimal number and type of inspection engines to keep traffic flowing smoothly.

The ASA CX hardware architecture, which consists of several CPU cores (based on hardware configuration), gives the nScan Array extensive parallel inspection capacity. Coupled with dynamic optimization of inspection engines in response to traffic patterns, and the efficient shuffling of traffic between network ports by the Virtual Packet Rings, this inspection capacity enables ASA CX to deliver high-performance context-aware security. At the same time, dynamically updatable inspection engines and Pluggable Context Stores give the ASA CX immense flexibility to combat new threats as they arise, with minimal intervention from the security operator.

While the ASA CX has been designed for flexibility and performance, the architecture seamlessly supports the complete set of security features that enterprises have come to expect from Cisco’s firewalls. This includes inspection engines developed to counter threats to specific applications, and capabilities such as remote access and network address translation.

Granular control of applications

The ASA CX implements traffic inspection in two stages. At the first stage, an inspection engine classifies traffic at a coarse level. In many cases, the coarse classification is all that is needed and the engine makes an enforcement decision without further investment of computing power. However, in some cases the coarse classification engine determines that to enforce a particular security policy it needs to inspect the traffic more deeply. In these cases, the traffic is guided to a deep classification engine that carries out the more granular inspection. The two-level inspection mechanism provides the security staff with immense flexibility without compromising on performance.

Consider the case where the security staff wants to permit use of Yahoo! Messenger for collaboration between employees, but also wants to prevent file transfers via the Messenger to protect the enterprise from data leakage. Enforcing such a policy requires the deep inspection engine, as the inspection process needs to look beyond port and protocol. The ASA CX's deep inspection engine first identifies and permits the Yahoo! Messenger flow as it comes in. Realizing that a file transfer may be initiated at any point in the midst of a chat session, the engine continues to monitor the flow. At some later point in time if a user initiates a file transfer, the inspection engine recognizes a new sub-flow, flags it, and prevents the transfer from completing.

Similar visibility and granular control is available for more than a thousand applications, in the ASA CX. However, deep inspection is not unnecessarily triggered for traffic for which coarse classification is sufficient (FTP from a permitted network, for example).

Finally, note that the ASA CX monitors application flows on all IP ports, not just a specific list of ports. As a result it can effectively detect and control non-web applications such as Yahoo! Messenger and Skype, as well as applications that hop ports.

Comprehensive user identification

Firewalls can implement user identification in two broad ways. One is passive user identification based on scraping Active Directory (AD) agent logs and associating IP address with users for a short amount of time. Any traffic to or from a known IP address is thus attributed to the user associated with that IP address. The other is active user authentication where a firewall uses NT LAN Manager (NTLM) or Kerberos to process a user's identity.

Passive user authentication has the advantage that it works with all applications that are in use — since such authentication does not interact directly with an application. However, it leaves an authentication hole in the cases where an IP address is reassigned from one device to another. In such cases, passive user authentication cannot correctly bind an IP address to a user's identity.

Active user authentication performs true authentication via a protocol such as NTLM or Kerberos, from within an application, making it more precise than passive authentication. Unfortunately, only certain applications, such as web browsers and Microsoft Outlook, implement NTLM or Kerberos. As a result, active user authentication may not provide adequate coverage by itself.

Several firewalls implement only one of these user identification mechanisms. In comparison, the ASA CX makes use of both passive and active mechanisms to determine a user's identity, forming as complete a picture of a traffic flow's context as possible. The ASA CX also has the ability to profile and exempt non-interactive devices such as printers from authentication, preventing false alarms from these devices.

In networks that deploy the Cisco Identity Services Engine (ISE) and implement Cisco TrustSec® technology, the ASA CX can do even better. Network endpoints implementing TrustSec, tag user flows based on user identity. ISE associates policy actions with tags and distributes them to enforcement points such as Cisco switches and firewalls. The ASA CX interoperates with ISE - it receives the policy-tag associations from ISE and populates its local context store with these associations. Subsequently, when the ASA CX sees incoming flows with TrustSec tags, it knows the appropriate enforcement action for them. This high-fidelity visibility into and granular control over user transactions goes well beyond the ability of any other firewall on the market.

Unprecedented device- and location-based control

Cisco AnyConnect™ is the most widely deployed remote access VPN solution in the market today, with more than 100 million seats deployed. AnyConnect is the preferred choice for most large enterprises implementing VPNs, and is available across a broad spectrum of devices, including PCs, iPhones, and iPads. When AnyConnect is used on an end-user device and the other end of the VPN connection resides on the ASA CX, the ASA CX obtains access to the device's OS-type, OS version and ownership (corporate vs private); the type of security software installed on the device (e.g., antivirus scanner); and the status of the security software on the device (e.g., clean, needs update). This device information significantly enhances the context available to the ASA CX. Using this context, the ASA CX can enforce policies such as permitting access to email from a private laptop but denying access to sensitive enterprise applications, or, for a corporate-owned laptop, allowing access to email as well as enterprise applications.

In the future, AnyConnect will work even when the end-user device is directly on the corporate network — in other words when the device is not connected via a VPN. In this situation, AnyConnect will directly transmit parameters such as the status of the security software on the device, to the ASA CX to take into account while making enforcement decisions.

Most competing firewalls do not have a robust integration with a widely deployed endpoint solution such as AnyConnect. As a result, these firewalls are blind to the device used for access, and are unable to obtain the full context of the access. Thus, these firewalls make sub-optimal enforcement decisions, irrespective of whether they permit or deny the access.

Further, the ASA CX can interoperate with ISE to enforce device-based policies based on TrustSec tags. The mechanism at work is similar to the user-ID-based enforcement discussed earlier. The only change is that tags are associated with devices rather than with users; consequently, the local device store is updated and used, rather than the local user store. Consider the case where an enterprise using TrustSec wants to implement a policy that denies mobile devices on a guest network access to enterprise data. Any flow originating at a device in the guest network is assigned a tag that tells Cisco switches and firewalls that the device has limited permissions. Subsequently, if the ASA CX comes across a flow destined for a corporate database but with a "guest device" tag, it knows to deny the flow.

Finally, the ASA CX can enforce location-based control using location information derived from the AnyConnect headend. Consider the case where the security staff wants to allow access to sensitive information on the corporate intranet to personnel on the road. However, based on a history of break-ins, the staff only wants to allow such access if the travelling personnel are in the United States. The ASA CX can be configured to implement such a policy. When a new VPN connection comes into the ASA CX, an inspection engine examines the originating IP address of the connection. Using the Pluggable Context Store, the engine determines that the connection originates outside the U.S. Thus, the ASA CX denies the flow, safeguarding access to the intranet in accordance with the configured policy.

Zero-day malware protection

Cisco has the world's largest threat analysis system in Cisco Security Intelligence Operations (SIO). The SIO network contains more than 700,000 global sensors that see more than 5 billion web requests per day as well as 35% of the global email traffic. In addition, SIO receives threat telemetry from AnyConnect endpoints. SIO continuously collates all of the information that it receives, categorizing networks, domains and applications and assigning them reputation scores. These scores are centrally computed, and quickly distributed to Cisco equipment configured to receive them. ASA CXs receive these scores, update their global context stores, and begin recognizing emerging threats and commence enforcement actions against these threats.

Consider the case where a section of popular news site contains an HTTP GET request that maliciously redirects the browser to a host containing malware. The unsuspecting user visiting the website, may inadvertently trigger the GET request and download the malware on her computer.

ASA CX can seamlessly protect against such "drive-by" downloads, without any intervention from the user or the owner of the news website. Using the reputation information from SIO, the ASA CX "knows" that the news website itself is reputable. However, the ASA CX does not ascribe the website's reputation to the host in the malicious GET request. When the GET request results in an attempt by the host to download the malware onto the user's computer, the ASA CX realizes that the host in question has a low reputation. Thus it prevents the download from proceeding, thereby protecting the user's computer.

Intuitive management

As Cisco re-imagined the firewall with the ASA CX, it has re-thought firewall management as well with Cisco Prime™ Security Manager (PRSM), which utilizes Web 2.0 technologies to simplify everyday use and provides a consistent experience, regardless of whether the security staff is managing one firewall or several.

Ease of expressing business policies

PRSM is optimized to translate business policies easily into the appropriate firewall configuration. Let's assume that the security staff wanted to configure access to Yahoo! Messenger while preventing file transfers using the Messenger. With a traditional management application, this configuration would require coordinated creation of multiple rules based on IP addresses, protocols and port numbers, and these rules would have to be constantly modified to take into account any port-hopping implemented by the Messenger.

PRSM replaces this tedious and error-prone configuration with a few mouse clicks. The security operator navigates to the applications pane, searches for Yahoo! Messenger by name and checks the box for disabling file transfer. The operator does not need to configure IP addresses, protocols or port numbers. Instead, the policy configuration that the operator desires is expressed in everyday business language — i.e. "allow Yahoo! Messenger but disable file transfers." The translation of the business policy into low-level rules for IP addresses, protocols and ports is completely handled by PRSM.

Comprehensive reporting

For context-aware security to work, the security staff must be able to generate reports and conduct detailed analysis of specific events. PRSM includes a comprehensive and flexible reporting package to facilitate such analysis.

The ASA CX generates events, based on configuration, in a standard format. These events can be stored on-device or shipped to the off-box management application for aggregation and long-term archival. These events are also continuously summarized into reports, enabling security operators to access the reports on the fly. Further, events can be transformed from the PRSM standard format to any other format using a converter. Thus, security operators can continue to use their favorite third party tools to analyze the events and reports. Finally, the reports support a drill-down capability such that security operators can get detailed information on events for debugging, troubleshooting, and the like.

Consistent on-device, off-box experience

Cisco PRSM is available in two variants. The first is a web-based on-device version that is integrated with the ASA CX. The second is an off-box version that is typically used in situations where a network contains multiple ASA CXs. The on-device version is identical to the off-box version except for the latter's ability to manage multiple firewalls. Thus, from a security operator's point of view, the experience of managing the ASA CX is consistent irrespective of the management application variant — on-device or off-box — that the operator chooses.

Expressive UI

Cisco PRSM contains several constructs designed to provide security operators with a flexible yet efficient device management interface.

First, a search bar is available throughout the UI. An operator can conduct a free-form search for a policy, object or event from any screen and then quickly access the item of interest. In addition, operators can attach user-data to any policy or object to facilitate efficient searching at a later time. Second, all UI panels are equipped with auto-completion and auto-suggestion shortcuts allowing operators to save time while managing the ASA CX. Third, the policy language is extremely flexible, allowing operators to construct succinct policies even for complex situations. For example, an operator can express policies such as "allow access to all users in the engineering group, except users A and B." Most competing firewall management applications lack such expressiveness, forcing operators to construct artificially elongated rules.

PRSM's UI is optimized to the security operator's everyday workflows. For example, in response to a complex ticket, an operator can create multiple changes in the UI, preview them, and only then commit the changes to take effect. Similarly, the operator can track the history of changes to a policy and the objects used by a policy. Such detail allows operators to collaborate on defining policies that are touched by multiple people, or to decommission objects that are no longer in use.

Model-based design

At the core of Cisco PRSM is a model-based design. All configurable objects are modeled with a schema that describes the object in its entirety, including the types of configurations that can be applied to any element of the object. Data associated with the schema is stored in a configuration database (Config DB) on the ASA CX. Together, the Config DB and the schema completely define every instance of a configurable object on the ASA CX.

The UI template for the management application is auto-generated from this schema, as is the code template to wire the schema to the inspection engines in the nScan Array. Generating code in this manner from the schema not only ensures consistent implementation for the configurable objects and minimizes software errors in the management application.

REST API

Cisco PRSM is built using a Representational State Transfer Application Programming Interface (REST API). This API abstracts the various configurable objects in the Config DB and allows the management application to apply a consistent set of operations to those objects. The REST API is also available to the security staff, should it desire to write scripts or develop a custom management application for the ASA CX.

Evolutionary deployment model

Even though context-aware firewalls are needed for complete protection against security threats, by no means do they render stateful Layer 3/Layer 4 (L3/L4) firewalls obsolete. L3/L4 firewalls remain useful in a wide variety of applications, including in the data center, where traditional five-tuple-based rules are both effective and sufficient. Nor is wholesale replacement of existing L3/L4 firewalls by context-aware ones desirable — given the hardware and process investment enterprises have already made.

Cisco protects existing hardware investments via a context-aware (CX) module that can be added on to an existing Cisco L3/L4 firewall, converting it to an ASA CX. Once this module is installed and configured, some portion of the traffic — that which needs full context-aware inspection — is routed through it. For the rest of the traffic, inspection based on IP addresses, ports, and protocols is sufficient and such traffic continues to be serviced by the L3/L4 firewall as before. As a result, enterprises interested in adding context-awareness to their security infrastructure do not need to replace their existing hardware with new firewalls. Nor do they need to add another device to their network — since the CX module simply adds on to an existing firewall device. The CX module provides all the additional visibility and control that the security staff needs to protect the enterprise.

Further, the ASA CX protects enterprises' investment in existing processes and firewall rules. Since existing Cisco L3/L4 firewalls are not ripped out of the network, the rules configured on those firewalls require continued maintenance. However, because Cisco PRSM is capable of reading, translating and managing the configuration for existing L3/L4 firewalls, security operators are spared from having to deal with two independent firewall management applications.

Finally, for enterprises that do not deploy Cisco firewalls, the ASA CX removes the need to make a choice between firewalls that "bolt on" some context-aware capabilities to an outdated architecture and next-gen firewalls that do not implement all the features available on stateful L3/L4 firewalls. The former typically implement an extensive feature list that enterprises rely on today — such as remote access and network address translation — but have limited context-awareness. The latter provide a broader, even if incomplete, set of context-aware capabilities, but miss important features that the security staff has come to expect. By moving to the ASA CX, such enterprises can maintain their existing use of traditional firewall features while gaining visibility into and control over applications, users and devices on their networks.

Thus, the ASA CX can be deployed in an evolutionary manner, interworking with existing hardware and processes whenever possible. At the same time, it enables enterprises to benefit from the revolutionary architecture that brings full context-aware protection to their networks.

Completing the Cisco SecureX framework

With SecureX, Cisco laid out a framework to protect the enterprise in the face of sweeping changes such as appification of the web, and cosumerization of IT. The SecureX framework embeds security into the network by intelligently and dynamically combining local context from AnyConnect endpoints and TrustSec tags, global reputation information from SIO and context-aware policy enforcement on a firewall. The ASA CX completes the SecureX framework by enabling context-awareness in the firewall.

Please see http://www.cisco.com/go/asa for more information on the ASA CX, including details on feature availability.

References

The Future of Network Security: Cisco's SecureX Architecture, 2011, Cisco Systems

Cisco TrustSec: Context-Aware Secure Access for Borderless Networks, 2011, Cisco Systems

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

Printed in USA C11-700240-00 02/12