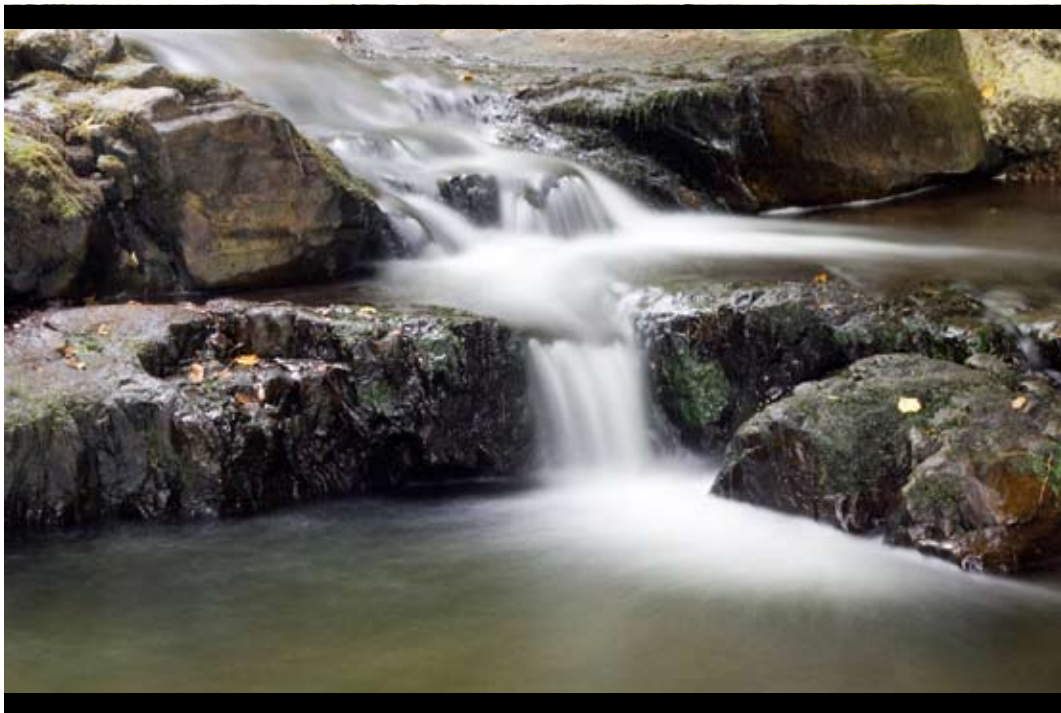


Scaling the Secure Web Gateway



Introduction

In this paper, the term content security refers primarily to protections such as URL filtering and malware protection that are applied to outbound sessions initiated by end users as opposed to protections applied to web or other n-tier servers on incoming requests.

The history of each new wave of network security deployments has been a progressive expansion to “appliance sprawl” and content security is the latest casualty.

In the face of tremendous scaling pressures due to traffic growth combined with increasing connectivity requirements and adding new features to combat new threats, security architects have had little choice but to build large server and appliance farms glued together with load balancers, switches and ineffective clustering techniques. In content security, the specific requirements since 2005 have grown from simple URL filtering to malware protection, as well as ancillary functions like SSL decryption and proxy/caches. Just in the last year, even newer protections have been added, including reputation services and botnet protection. In the near future, Data Leak Protection (DLP) will start being added as well.

Crossbeam solves this scaling problem at an architectural level, using the advanced intelligence in its X-Series operating system (XOS™ software) to enable the deployment of the whole set of content security services in an extremely efficient way and give back control of the infrastructure to the network security team. The result is better scaling, lower cost, and an infrastructure that supports customer business objectives.

Appliance Sprawl Comes to Content Security

The first instances of appliance sprawl occurred with firewalls. Early firewalls ran on servers with limited processing power. As soon as traffic increased beyond the capability of a single server, load balancers were applied and surrounded multiple firewall servers to provide scaling. Subsequently, similar strategies were deployed for intrusion detection and prevention systems.

Content security is now experiencing even greater pressure. Because malware is now present on 5%-10% of legitimate web sites and growing at over 500% per year (!), security teams are being forced to apply inline malware protection on all web requests as opposed to relying purely on URL filtering which traditionally has been deployed in offline mode. In addition, bot-controlled PCs are increasingly performing

the stealth transfer of personally identifiable information outside the enterprise back to hacker sites, a process known as “phoning home.” This, again, requires inline deep packet inspection of traffic flows. Finally, traffic is once more growing at a significant rate, with many Crossbeam customers reporting doubling of traffic each year.

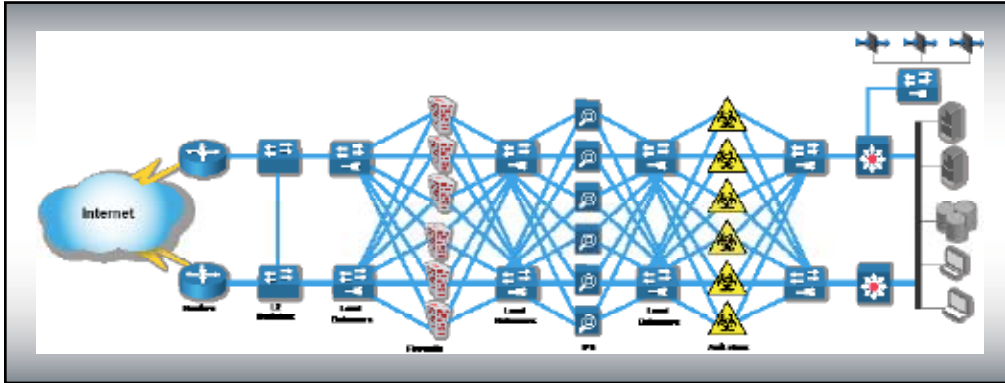


Figure 1: Appliance Sprawl Begins

This progression has happened so quickly that, much like the earlier firewall scaling strategies, various layers of servers running URL filtering, malware protection, and even proxy/caches have multiplied quickly with similar load balancing and clustering strategies. Unlike firewalls, however, the problem is more severe because packet payloads must be inspected. In firewalls, the target of security processing is primarily packet headers which are highly structured and smaller and thus less CPU intensive. Each step further into the packet causes much higher CPU utilization, as more software is required to deal with increasingly unstructured data after the packet header – as is the case with content inspection.

The combination, therefore, of deep packet inspection pressures along with the requirement to run inline and handle rapidly growing traffic has caused security teams to multiply the number of boxes performing each specific function. An example of this is shown in Figure 1: Appliance Sprawl Begins, and represents an actual customer prior to deploying Crossbeam equipment. Note that this particular architecture does not include newer content security functions such as reputation services, data leak prevention, and P2P protections. Each, in turn, requires its own layer of appliances or servers and generates successively worse waves of equipment growth.

These inline appliance complexes have also turned out to be inflexible as new services are added. This is because the order of security processing matters to the efficiency of the architecture. For example, in a classic combination of proxy/cache, URL filtering and AV scanning, it is important that the caching function be applied first in order to determine whether a forbidden web site has already been flagged. If not, then the URL filtering function can make the next determination to see whether the request should be forwarded. Finally, if the request is permitted, only then might the AV function have to inspect the HTTP transaction to

determine whether a downloaded file needs to be scanned. The problem is how to insert the next function in the architecture. Imagine that data leak prevention needs to be added with similar sets of multiple boxes in order to scale. Where is this function inserted? Once this issue has been decided, the existing infrastructure must then be designed for implementation. And this cycle repeats itself with each new service.

It should be noted that the complexity of these layered architectures becomes worse as various security vendors migrate their applications to Layer 2 mode (see Figure 2: Layer 2 Chains) and requires significant design time simply to understand packet flow between Layer 2 and Layer 3 functions.

A secondary problem has also surfaced recently for content security infrastructures: the growth of segmented transaction zones. Because of problems such as worm outbreaks, organizations have increasingly adopted more aggressive segmentation of their networks, with “choke points” between boundaries. In addition, these zones often contain their own specific content that requires protection such that the concepts of “outside” and “inside” the perimeter lose their meanings. What really matters is the particular policy to be applied as traffic traverses the boundaries between the transaction zones. Organizations that only maintain content protections on Internet-facing boundaries risk missing significant internally vulnerable web assets.

The specific problem raised by segmentation is that it may require duplicate infrastructures between zones and not just at the Internet boundary. In addition, the zone transitions will typically require different combinations of protections depending on the assets being protected and the type of traffic crossing the boundaries. For example, a particular zone might need malware protection for requests to internal servers but no URL filtering in this case (since the servers are internal and not part of global URL filter databases), whereas another zone needs both. But the choice then becomes building separate security layers for specific zone traversals, resulting once again in a multiplication of boxes. Worse, when faced with this kind of multiplication, the security team may simply decide that cost and complexity are too high and the protection is foregone altogether.

The net effect of this appliance sprawl has been to increase capital and operational costs including license renewals, maintenance on each box, troubleshooting, and staff increases. The pressure on the security team is heightened by the imperative to “do more with less” just at a time when they have only been able to “do more with more.”

Layer 2 Chains

The requirement to perform content scanning inline has created other difficult challenges from a deployment perspective. Chief among these is the requirement to run in Layer 2 or “transparent” mode with fail-open capabilities. While security vendors generally apply bridging strategies to meet the requirement, the result, from a deployment perspective, is a chain of Layer 2 devices that are difficult to troubleshoot. If a packet does not get through to its destination, how do you know which device stopped it?

Figure 2: Layer 2 Chains

The Crossbeam Secure Web Gateway

Crossbeam offers an entirely new and much more flexible way to meet the requirements of content security infrastructures. Using the intelligence and power of the Crossbeam X-Series Next Generation Security Platform, customers can scale content security performance without having to add boxes, add new features without having to redesign their architecture, provide “five-nines” of high availability and reduce the number of operators required to manage the equipment. The solution also enables a rational and simple migration that does not require everything to be changed at once. Customers typically replace from twenty to fifty devices with just a couple of Crossbeam chassis, depending on the complexity and performance demands of the network.

The Crossbeam Secure Web Gateway (SWG) consists of an X-Series Next Generation Security Platform running customer-preferred combinations of the following best-of-breed security applications: URL filtering, malware scanning, anti-virus, P2P protection, bot-net protection, proxy/cache, and reputation services. In Figure 3: X80 with URL filtering and malware scanning enabled, an X80 is shown running URL filtering and malware protection (shaded here to show the particular capacity allocation in this example).

Other functions, such as SSL decryption and data leak prevention, are also easily added. Additionally, the order and specific functions to be applied may be determined entirely by the customer. The security functions themselves are provided by industry leading best-of-breed vendors such as Trend Micro and Websense.

The simplicity of the solution and its flexibility derive from an architecture that is designed and purpose-built to handle the requirements of scaling security services in large networks. This architecture comprises an intelligent, distributed, and virtualized system software layer that orchestrates the functions of three tightly integrated blade types: network processing modules (NPM), application processing modules (APM), and control processing modules (CPM). In Figure 3: X80 with URL Filtering and Malware Scanning Enabled, the NPMs are the four blades on the left, the APMs are the blades in the middle with no interfaces, and the CPMs are the two blades on the right. With just two of each type of blade, the entire system is completely redundant in itself and there is no single point of failure. Scaling is simply a function of adding blades and the X80, shown here, drives up to 40Gbps of wirespeed throughput.



Figure 3: X80 with URL Filtering and Malware Scanning Enabled

The system is managed either via a graphical user interface, via command line interface, or in a distributed fashion by a network manager called SecureShore™ Network Management System that can manage up to 1000 separate Crossbeam devices.

The theory of operation of the X-Series is important for the understanding of the SWG. Overall, traffic enters the system via the interfaces on the NPM. The NPM then distributes the traffic to the APMs running the security software, making sure that the correct sequence of services is applied and in the most efficient manner. Management is done via the CPM.

The specifics are as follows:

- The administrator decides which security applications/services will be run and then assigns those services to groups of APMs. These groups are called Virtual Application Processing groups (VAP groups) and they act as a single entity no matter how many blades are in the group. In Figure 3, malware scanning and URL filtering each consist of 3-blade VAP groups.
- The administrator creates policies that determine how traffic entering the NPMs is to be serviced by the security applications running on the APMs. Security services can be applied sequentially and/or in parallel (as in the case of monitoring applications such as behavior anomaly detection systems). Once the NPM determines the order of applications it also determines the fastest processor within each VAP group in order to achieve maximum performance. It then distributes the traffic to the APMs across a telco-class switching fabric that is dedicated to the data plane. Each NPM instantiates a separate switched set of links to the APMs. Built on a combination of 30G switches, 10G network processors, and 16-core security processors, each NPM can deliver 10Gbps of wirespeed throughput with a total of 40Gbps for a fully loaded chassis. The system software that operates the NPM also supports easy insertion of the chassis into any large network with support for extended routing functions, VLAN trunks, link aggregation, and various failover protocols.
- The system as a whole is monitored by the CPM across the backplane on an out-of-band switching fabric with dedicated 1Gbps management lanes to each blade. The CPM works with each of the other blade types to ensure highest performance and availability of the system. For example, blade failure causes instant rerouting of traffic to other blades within the VAP group with no service loss and capacity restoration takes approximately 30 seconds as standby blades are brought into service with no human intervention. Thus, in the face of failure, with zero-touch by humans, the system sustains no service disruption and complete capacity restoration.

Given this flexible allocation of resources and security applications, it then becomes very simple to create a Secure Web Gateway with the correct functions applied in the desired order. The order can be changed, capacity re-allocated, or functions added and removed at will. The flexible allocation of capacity is critical here because malware scanning may take up more processing horsepower than URL filtering. A typical deployment for 25,000 users might have three URL filtering blades, and four malware scanning blades. Alternatively, for customers who still see high cache hit rates, a configuration with three proxy/cache blades, two URL filtering blades, and three malware protection blades might suffice. In the case where SSL decryption is required, an optional blade dedicated to this purpose may also be configured. Because APM performance is so high, a single blade is capable of decrypting 3DES encrypted traffic at 500Mbps or AES at 2Gbps. Alternatively, on-board APM SSL accelerators are also available for applications that include native SSL decryption capabilities. The important point is that the administrator can decide which traffic should be decrypted and then sequence the decrypted traffic through multiple separate security services without having to decrypt/re-encrypt multiple times.

Thus, in a single chassis, administrators of large network security architectures can migrate to a total content security solution that:

- Handles both non-encrypted and encrypted traffic
- Provides proxy caching for initial traffic reduction
- Delivers URL filtering for new requests
- Delivers malware scanning on requests that are allowed by the URL filtering solution
- Delivers reputation and anti-x (virus, spam, phishing, and spyware) services for mail traffic
- Delivers bot-net prevention

The system can deliver one or all of these functions depending on the requirements of the network and the organizations being protected.

Summary – The Dramatic Benefits of the New Approach

The Crossbeam SWG solution is “next generation” because it completely breaks the old mold of rigid appliance farms built on the philosophy of “see a threat, buy a box.” Instead, as the very definition of a content gateway expands and changes each year, the architecture can anticipate future risks and easily absorb these changes while continuing to deliver high performance and high availability.

The flexibility and intelligence of the Crossbeam solution helps customers consolidate anywhere from 20 to 50 existing separate appliances into one highly available, scalable system. When compared to competing solutions, customers can experience savings of \$500,000 to \$1M in the first year alone simply by eliminating the capital and operational expenses associated with appliance sprawl. The performance gains of a system that can scale to throughput of 40Gbps also ensure a long product life with no fork-lift upgrades across multiple depreciation cycles.

Security teams who invest in the Crossbeam SWG report dramatic gains in productivity and security team effectiveness. Indeed, some customers have not added any network security staff in four years because of the simplicity and effectiveness of the solution. With best-of-breed security functionality from industry leaders as a foundational component of the SWG, the defensive posture of the solution is unequalled. The ultra-high performance and reliability of the Crossbeam platform ensure that network security teams deliver a safe and great end-user experience without sacrificing security or availability and deliver new capabilities in faster time with fewer devices, less cost, and less staff.

About Crossbeam Systems

Crossbeam Systems, Inc. transforms the way enterprises, service providers and government agencies architect and deliver security services. The basis of Crossbeam's solution is its Next Generation Security Platform, a highly scalable hardware platform that facilitates the consolidation, virtualization and simplification of security services delivery, while preserving the customers' choice of best-of-breed security applications. Crossbeam offers the only security platform that delivers unparalleled network performance, scalability, adaptability and resiliency. Customers choose Crossbeam to intelligently manage risk, accelerate and maintain compliance, and protect their businesses from evolving threats. Crossbeam is headquartered in Boxborough, Mass., and has offices in Europe and Asia Pacific. More information is available at: www.crossbeam.com



Corporate Headquarters

Crossbeam Systems, Inc.
80 Central Street
Boxborough, MA 01719
Tel: +1 (978) 318 7500
Fax: +1 (978) 287 4210

www.crossbeam.com