ZSCALER
The Leader in Cloud Security

ADVANCED SECURITY

How Zscaler Tackles Emerging Web Threats with High Speed, Real-Time Content Inspection in the Cloud

ABSTRACT

Leveraging a purpose built architecture capable of high-speed content inspection, the Zscaler solution inspects all web traffic in real-time. Content inspection covers not just the URL but also all headers and the full body of all requests and responses. Inspection at this level is vital to ensure security on the web today, which is dominated by dynamic, user-supplied content. Zscaler advanced security implements inspection at four levels – Knowledge of Destination, Payload, Application and Content to ensure that threats are mitigated using increasingly comprehensive scanning techniques. This is achieved without introducing noticeable latency thanks to a globally distributed architecture designed from the ground up, specifically for a Security-as-a-Service delivery model.

Attacker Evolution 3
SaaS Changes the Playing Field 3
Traditional Approaches Web Security 3
Inspection 4
Levels of Inspection 4
Knowledge of Destination 5
Knowledge of Payload 5
Knowledge of Application 6
Knowledge of Content 6
Depth of Inspection 7
Advanced Security 8
Network Effect 8
Partners 9
Malicious URLs 9
Phishing 9
Botnets 9
Vulnerabilities 10
Page Risk Index 10
History 10
A New Approach 11
Control Categories 12
Scoring 13
Functionality 13
Botnets 13
Malicious Active Content 13
Phishing 13
Communication 14
Cross Site Scripting 14
Control Access to Suspicious Destinations 14
P2P Control 14
Conclusion 14

Attacker Evolution

In nature, those that adapt to changes in their environment survive and prosper. The individuals that attack computer networks are no different. The information technology environment has evolved significantly over the past decade and attackers have adjusted their tactics along the way. Clear shifts have occurred in the attack evolution. Attacks have shifted from servers to web applications and on to web browsers. Along the way, attackers have evolved from individuals motivated by curiosity, to organized criminals seeking profit. Unfortunately, enterprises have largely failed to keep pace and continue to use dated methods to thwart attacks.

Attackers that once targeted enterprise servers have now realized that it is far easier to exploit client machines thanks to weak defenses and naive users. Buffer overflows in publicly exposed server-side services have been replaced by multi-faceted, client-side attacks leveraging social engineering, web browser vulnerabilities and trusted, yet vulnerable web applications. Web 2.0 technologies, while empowering developers to produce intuitive, user-friendly applications have also raised the bar on complexity, ensuring that vulnerable web applications are an accepted part of life on the Internet. The web browser has become a portal for attackers, allowing them to access sensitive data on desktop and mobile devices, while often permitting complete control of a machine as it is recruited into a botnet army. Enterprises must shift focus and adapt if they expect to defend against modern attacks.

SaaS Changes the Playing Field

Cloud delivered security or Security-as-a-Service (SaaS) solutions have begun to emerge in an effort to tackle the challenge of web browser security. SaaS solutions offer an inherent and critical advantage over traditional hardware or software based Secure Web Gateway (SWG) products. SaaS solutions are able to protect mobile devices just as easily as they protect assets on the Local Area Network (LAN). This is a game changing differentiator. Enterprises are becoming increasingly reliant on remote employees and 'road warriors' working both from laptops and smartphones. Attackers have recognized this shift. They know all too well that remote workers are unlikely to be protected by LAN based defenses and mobile devices therefore constitute a 'target rich' environment. SaaS vendors can inspect web traffic regardless of location but few vendors, such as Zscaler offer 'true SaaS' by requiring that no additional software run on the client device. This not only ensures that remote assets can be protected 'out of the box', but also reduces the cost and complexity associated with managing the overall solution.

Traditional Approaches Web Security

Latency is the enemy of web security. If the web browsing experience is degraded by security controls, users will not accept the solution. It cannot be avoided. Security introduces latency, as packets must be inspected in real-time. The deeper the level of inspection required, the more CPU cycles are consumed and as a result, the potential for slowing web traffic increases. Degrading throughput is a challenge for appliance vendors and it is enhanced in the multi-tenant environment introduced in SaaS based solutions. Vendors recognize this and have therefore been forced to limit the depth of content inspection in order to avoid introducing latency when inspecting web traffic. Without a high-speed, scalable infrastructure, deep inspection simply cannot be achieved. While competitors have built their web proxy solutions on top of existing technologies in order to bring solutions to market quickly, the

Zscaler infrastructure was built from the ground up with the sole purpose of creating the fastest infrastructure possible to permit deep, bi-directional inspection of web traffic.

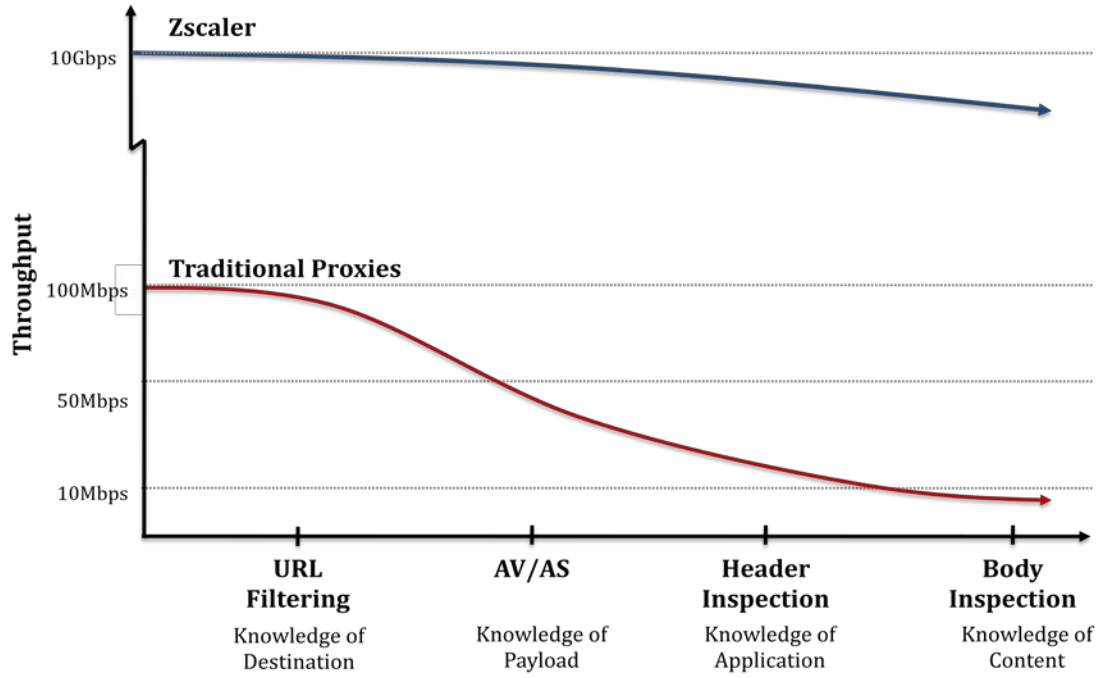

Figure 1 - Content Inspection Throughput

Inspection

Web pages must be reviewed and evaluated in order to implement security. All security solutions do this. What is not always apparent is the depth of inspection that takes place. The deeper the level of inspection performed, the greater the risk of introducing latency to web browsing. As such, most vendors limit inspection to only what can be done quickly. URLs can be quickly matched against block lists, but such an approach offers protection only against known attacks and doesn't take into account the dynamic nature of the majority of web content.

Levels of Inspection

Inspection capabilities of web security solutions can be divided into four categories as shown in Figure 2. Each subsequent level permits more comprehensive security controls, but also places greater demands on inspection engines and increases the likelihood of introducing latency. While *Knowledge of Destination* is common among security solutions, other levels are less likely to be employed.

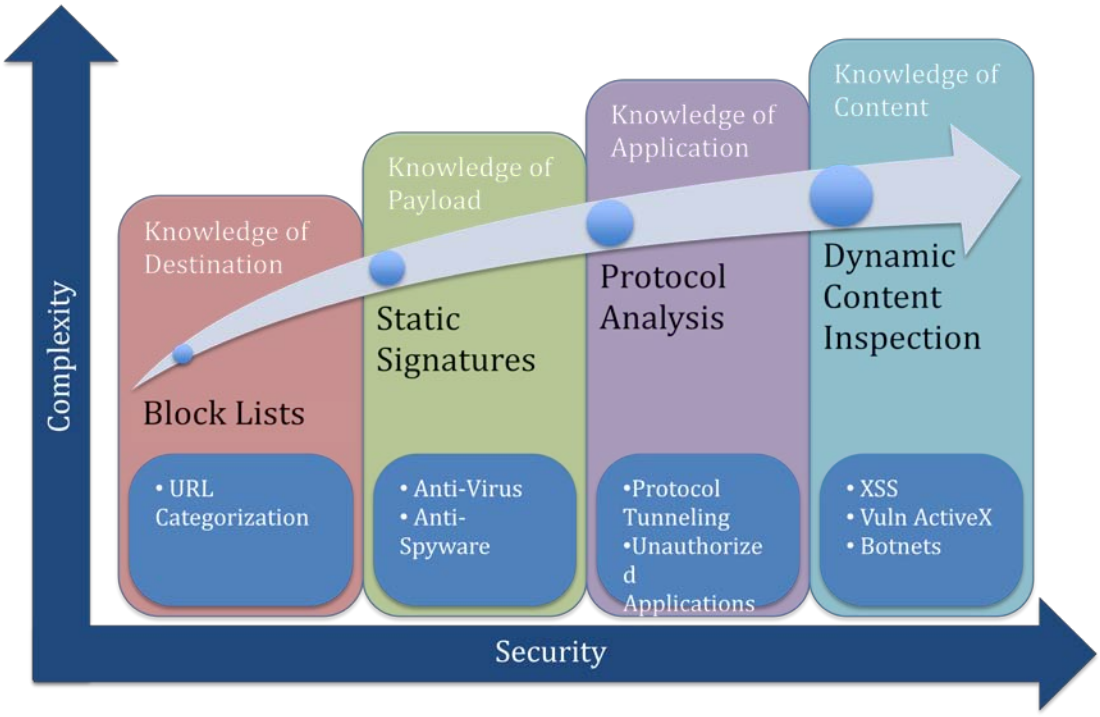

Figure 2 - Levels of Inspection

Knowledge of Destination

All web security solutions implement security controls through *Knowledge of Destination*. If it is possible to know ahead of time that web content violates a security policy, then blocking access to it can be implemented simply by inspecting the URL of the request. Block lists (aka black lists) follow this approach by archiving URLs associated with known malicious content. When an end user surfs to a page, the request is checked against the block list and should a malicious site be identified, the request is blocked. This is a simple, efficient way to implement security, but *Knowledge of Destination* alone provides insufficient protection as entities deploy Web 2.0 based sites.

Static content is no longer the norm. In a Web 2.0 world, content is dynamic in nature and user supplied content is not only accepted, but encouraged. Two individuals can request the same web page but receive markedly different results. Perhaps the web application customizes content based on user preferences or identifies the web browser version and adjusts accordingly. Maybe the site includes external content such as advertising space or news feeds. As such, a static block list simply cannot definitively identify malicious content. Additionally, thanks to millions of users providing the bulk of content, especially on popular social networking sites, it is simply not possible to keep up with new content purely by reviewing/updating block lists. Content must be reviewed dynamically, in real-time in order to implement security.

Knowledge of Payload

The web includes a significant volume of binary content in the form of pictures, video, audio, documents, executable files, etc. that could be malicious in nature. It is therefore important for such content to be evaluated by anti-virus (AV) and anti-spyware (AS) engines. While AV/AS has long been an accepted practice on the desktop, it is still not employed as an in-line solution in the majority of enterprises. One is not a replacement for the other. Rather, in combination, host and network based

AV/AS solutions are complimentary and represent an important component of a defense-in-depth approach to web security.

Host based AV must receive regular signature updates and should that process be delayed or blocked altogether for a variety of reasons, a machine can become infected. In-line or network AV/AS will provide protection in situations where host based solutions are either not up to date or have been disabled. The latter is a common problem when the machine has become infected. Malicious code often seeks to disable AV/AS to ensure that subsequent signature updates do not detect and quarantine the initial infection. Employing network based AV/AS ensures that a single point of failure does not exist when detecting malicious binaries.

Implementing in-line AV/AS in a web security solution is challenging. Once again, due to the real-time nature of the web, latency can quickly be introduced when running AV/AS engines in-line. This is especially true for SWG appliances that provide in-line AV/AS using an entirely separate third party appliance which communicates with the SWG via Internet Content Adaptation Protocol (ICAP). ICAP is a message transmission protocol commonly used to permit standardized communication between devices but such a setup is inefficient and creates unacceptable latency for most enterprises. Zscaler has taken a different approach by implementing AV/AS capabilities directly within the gateway responsible for content inspection. Moreover, files are scanned in phases to further streamline the process. If for example, the first 1MB of a 100MB file is found to be malicious, the file will be blocked immediately, without requiring the full download to complete. Additionally, hashing algorithms are continually employed to quickly identify content that has previously been scanned. The result is in-line AV/AS at unprecedented speed, resulting in transparent protection for end users.

Knowledge of Application

Today, virtually all protocols are tunneled through HTTP/HTTPS. This is true because ports 80 and 443 always represent a path to the Internet. As such, modern applications requiring external communication typically employ intelligent techniques to probe the network looking for ways to communicate with remote servers. They will generally begin by trying an efficient, proprietary protocol on a high numbered port and then slowly regress to using more common protocols, eventually settling on tunneling traffic through HTTP/HTTPS if needed. For this reason, traditional firewalls are of limited value. Applications are no longer defined by the ports that they communicate on. Identifying applications requires inspecting traffic to distinguish unique identifiers, typically within the URI or headers of the request/response.

Malicious code will often tunnel communications in order to bypass firewalls restricting outbound traffic. Additionally, employees will leverage applications with tunneling capabilities to bypass security controls on the LAN. Anonymizers are often leveraged to bypass URL filtering controls, while peer-to-peer (P2P) clients may be used to transfer files or conduct bandwidth intensive tasks such as sharing media files. Unauthorized communications such as IRC may also be tunneled through web protocols. Identifying and blocking such traffic is not a trivial challenge, as it requires researching the applications in question to understand their traffic patterns in order to implement appropriate pattern matching technologies. Moreover, research must be ongoing as application upgrades often change the application behavior used for detection. Zscaler provides controls to manage popular anonymizer, P2P and IRC applications with HTTP/HTTPS tunneling capabilities. Blocking is based on traffic inspection, not simple block lists which can be easily bypassed.

Knowledge of Content

A typical web request leads to dozens of responses from multiple web servers resulting in hundreds of kilobytes of data. Mashups and user-supplied content ensure that much of the content received has not

been vetted in any way to ensure that it is not malicious in nature. Additionally, legitimate sites are regularly compromised, and serve as a catalyst for attackers who are then able to leverage the site to attack the many trusting users visiting it each day. For these reasons, all web content received must be considered to be untrusted regardless of the source.

Deep content inspection, which permits high-speed pattern matching of all content regardless of location, without introducing latency, is a significant but necessary challenge. Threats identified via *Knowledge of Content* simply cannot be identified ahead of time for the simple reason that they don't exist ahead of time. Dynamic content requires real-time inspection. Deep inspection at this level covers not just the URL, but all headers and the full body of all requests and responses. This could not be achieved without a global architecture of proxy technologies designed specifically for this purpose. Zscaler is able to achieve this level of inspection in a SaaS solution because the system was designed from the ground up with *Knowledge of Content* as the goal.

Depth of Inspection

Consider a typical request to a social networking site such as Facebook. A single request to a page can result in dozens of responses from both facebook.com and third party sites. Content on the site is dynamic in nature and continually changing as it is provided by millions of users contributing to the site. Security controls relying solely on the *Knowledge of Destination* alone will be minimally effective in such a situation.

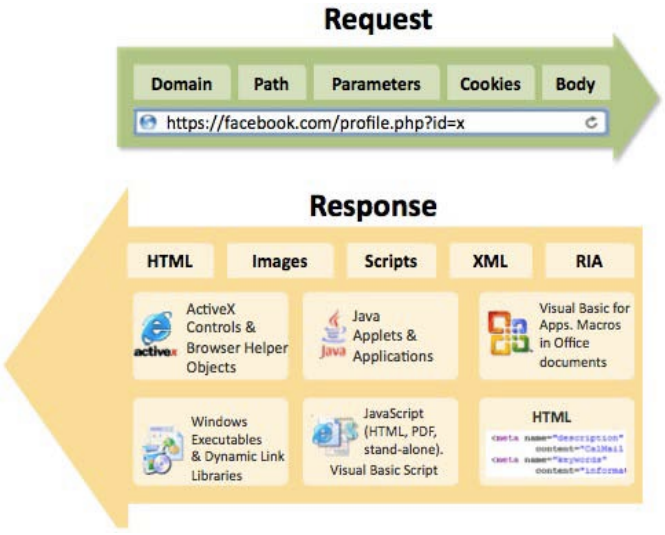

Figure 3 - Web Request/Response

Let's break down a typical web request/response scenario as identified in Figure 3. The request is not limited to simply the URL but also potentially includes cookies and body content. All components of the request can impact the response, which is returned by the web server. Much of web content today is dynamically generated based on information submitted in the request. The page being requested is determined by the domain and path alone. However the content displayed on that page can be influenced by the parameters submitted either in the URI (GET request) or body (POST request). Cookies may also contain information about the user, which can lead to customization. Even components such as the browser being used, plugins installed or geographic location that the request originated from can influence the content returned. Many security solutions focus only on blocking at a site (domain) or perhaps page (path) level, however, as can be seen, two users can request the same web page on a site and receive completely different content based on a variety of factors. Attacks can also exist in virtually any component of the request. Cross-site scripting (XSS) attacks for example are likely to exist in parameters of the request but could also be found in the body, headers or even cookie contents. Requests can also involve the upload of content (*Knowledge of Payload*) or result from an application tunneling an alternate protocol through HTTP/HTTPS traffic (*Knowledge of Application*). Restricting inspection to only the domain and path of a request will miss entire classes of attacks.

Once a request has been made and the initial page content is returned, it typically results in a variety of subsequent requests and redirects before the full 'page' is completely rendered. The full content returned includes not just HTML, but client-side scripts (e.g. JavaScript), data (e.g. XML), binary content (e.g. images/video), Rich Internet Applications (e.g. Flash and Silverlight), etc. As discussed, the content returned is dynamic and there is no guaranteed that two requests to the same page will result in the same content being returned. Response content must therefore be independently evaluated for malicious content in real time – a significant challenge when latency cannot be introduced. Downloaded files must be scanned by in-line AV/AS engines (*Knowledge of Payload*) and the full content body of all responses must be inspected (*Knowledge of Content*), all without introducing noticeable latency. Malicious content such as JavaScript targeting a known web browser vulnerability or a vulnerable ActiveX control being exploited could exist deep within the body of a dynamically generated response. If a solution is not capable of deep inspection, such attacks will never be identified.

Advanced Security

Zscaler is the only SaaS based web security solution capable of achieving deep, real-time inspection at all levels – *Knowledge of Destination, Payload, Application and Content*. With high-speed gateways deployed around the world and geoIP technology, web traffic is always routed to the gateway in the closest geographic proximity. This eliminates needless latency caused by inefficient routing. Once traffic reaches the Zscaler gateway, *Single Scan, Multi-Action* scanning running on a purpose built system allows gateway engines to efficiently scan content in a single pass to implement security at all levels. This architecture permits security controls that can inspect content bi-directionally, at any level, whether or not it is encrypted.

Network Effect

Beyond the ability to implement security in real time, a SaaS architecture permits unique abilities to identify previously unknown threats and leverage this knowledge to protect all clients. This is known as the network effect. The knowledge that can be obtained from the system grows exponentially as users are added to the system. Zscaler taps into this potential by implementing off-line processes to further inspect content using methods that simply could not be performed in real-time due to the time involved to perform the depth of analysis necessary.

Off-line analysis includes data mining to identify anomalous traffic patterns. When attacks are identified, signatures are created to ensure that further attacks are identified, blocked and reported in real-time. In this way, knowledge gained from a single targeted attack can be leveraged to protect all users across all companies. This is the power of the network effect.

Binary files are continuously re-analyzed using multiple AV/AS engines from a multitude of vendors. Such analysis would not be possible in real time due to the latency that such a process would introduce but can be implemented via offline processing where latency is not a concern.

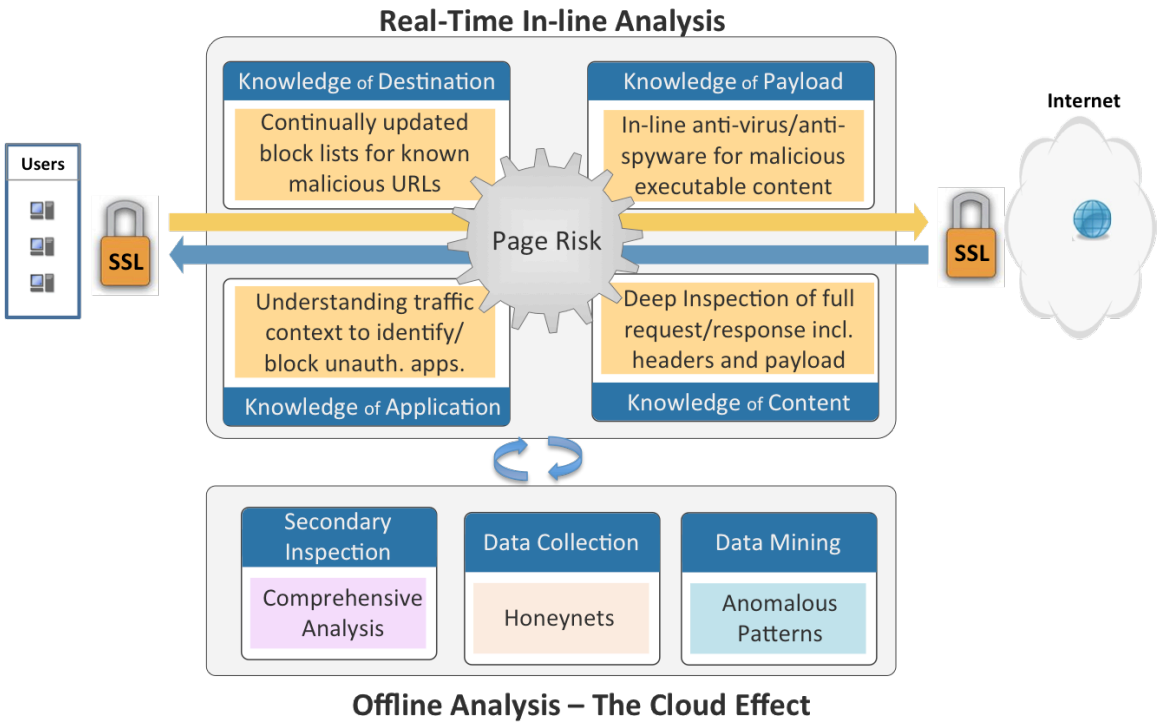

Figure 4 - Real-Time and Offline Analysis

Partners

Sharing information with trusted partners is essential to continually learn of emerging threats. Zscaler constantly evaluates partner data feeds to identify those that will improve threat knowledge and enhance client protections. Partner data feeds are integrated in the following four separate domains.

Malicious URLs

On a daily basis, thousands of pages are identified that are known to be hosting malicious code. The code is designed to compromise web browsers accessing such pages by way of known or unknown vulnerabilities. When malicious URLs are identified, block lists leveraged by global Zscaler gateways can be instantaneously updated, ensuring that users are transparently protected.

Phishing

Phishing has become a lucrative industry for attackers. Setting up sites designed to social engineer victims into sharing confidential data such as credit card and social security numbers is trivial to do. The sites can be quickly assembled and disappear shortly after they first emerge. Receiving multiple feeds identifying such sites and quickly disseminating the information to gateways where it can be leveraged is a critical component to protect against phishing attacks.

Botnets

The growth of botnets is one of the greatest threats facing enterprises today. For this reason, Zscaler works with partners to identify known botnet command and control (C&C) servers. With such information, it is possible to continually monitor outbound requests to identify those destined for C&C servers, indicating the presence of infected machines on a given network.

Vulnerabilities

Vulnerabilities are continuously being discovered in applications. Zscaler not only monitors public sources to ensure that signature based protections are deployed for applicable client side vulnerabilities but also participates in a variety of private and commercial programs. In doing so, Zscaler gains access to vulnerability details ahead of the general public, enabling the deployment of signatures so that customers can be protected from attack simply by surfing the web via Zscaler's global gateways.

Page Risk Index

Many times, a definitive rule exists for blocking malicious content. Perhaps, an anti-virus signature has returned a positive response or a user is attempting to access a URL, which has been previously black listed. In such cases, blocking is a straightforward, binary decision. Access to the requested content is either blocked or allowed, based on a pre-defined security policy. However, new threats emerge every day, for which signatures have not yet been written. As such, the concept of reputation plays an important role in providing comprehensive security to end-users.

History

IP reputation has become standard functionality for email security vendors. The idea being that if spam email has previously been identified from a particular source IP address, that same address has an increased likelihood of delivering spam going forward. The more spam detected, the higher the likelihood that subsequent email messages will also be spam. This concept worked well for email security as an IP address is a reasonable and consistent identifier for an email server. Web security vendors have attempted to adapt this same concept. An IP address is not however a strong identifier for sources of malicious content on the web as a single web server may host content from multiple sources. As such, vendors have attempted to translate the concept of IP reputation to that of domain reputation. Vendors calculate a reputation score for a given domain based on a variety of static variables such as the results of periodic security scans. While this approach can provide insight into the security reputation of a given site, it is of limited effectiveness, especially in an environment driven by dynamic, user-supplied content where reputation scores are continually changing.

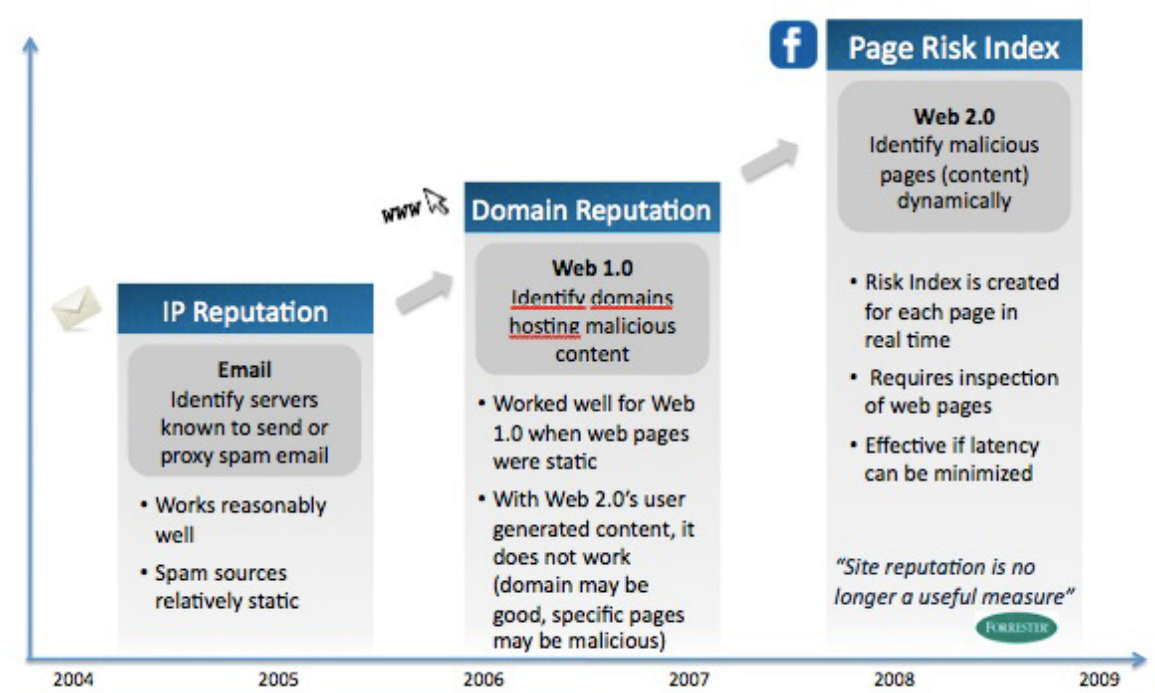

Figure 5 - Evolution of Reputation Scores

A New Approach

The size and growth rate of the Internet, combined with a trend toward increasing volumes of dynamic, user-supplied content, ensures that static measures of web reputation alone will never be adequate. Requests will always be made for which no definitive data is available to determine if the request represents a security risk. Content isn't static so how can a static reputation score expect to succeed? Two users can request the same content at the same time and receive markedly different results. Why? A variety of factors could play a role. The response could be customized based on individual user preferences, different browser versions or the geographic location where the request originated. Perhaps the page contains random content such as a banner ad. In short, static reputation scores simply cannot provide an accurate assessment of dynamic content. For that reason, dynamically calculated variables are required to automatically assess the risk for any given web request/response scenario. No individual metric will provide an accurate risk score for all scenarios. Rather, it is necessary to leverage a blending of a variety of risk indicators and apply appropriate weighting to each variable to achieve a comprehensive risk index or score. An 'in the cloud' security model is an ideal environment to calculate such a risk score as it ensures that all requests and responses pass through a central processing node and therefore allow for complete inspection of all data to and from a given resource.

Calculating a dynamic risk score is a challenging proposition. The score must be calculated 'on the fly' for every individual request. Numerous variables must individually by calculated and then combined into an overall risk score. The calculated score must then be compared to a predefined risk threshold in order to make a block/allow decision, and this must occur without adding latency to the browsing experience of each and every user. Thanks to high performance cloud-based architecture developed by Zscaler, we have been able to implement the first ever 100% dynamically calculated reputation score for web content. Known as Page Risk Index (PRI), Zscaler's patent-pending approach to web reputation is a comprehensive weighting of a variety of risk indicators in two separate and distinct control categories – domain analysis and content analysis. A PRI score is calculated for each and every web request

(request), which is then compared to previously established thresholds in order to determine how the request should be handled.

Control Categories

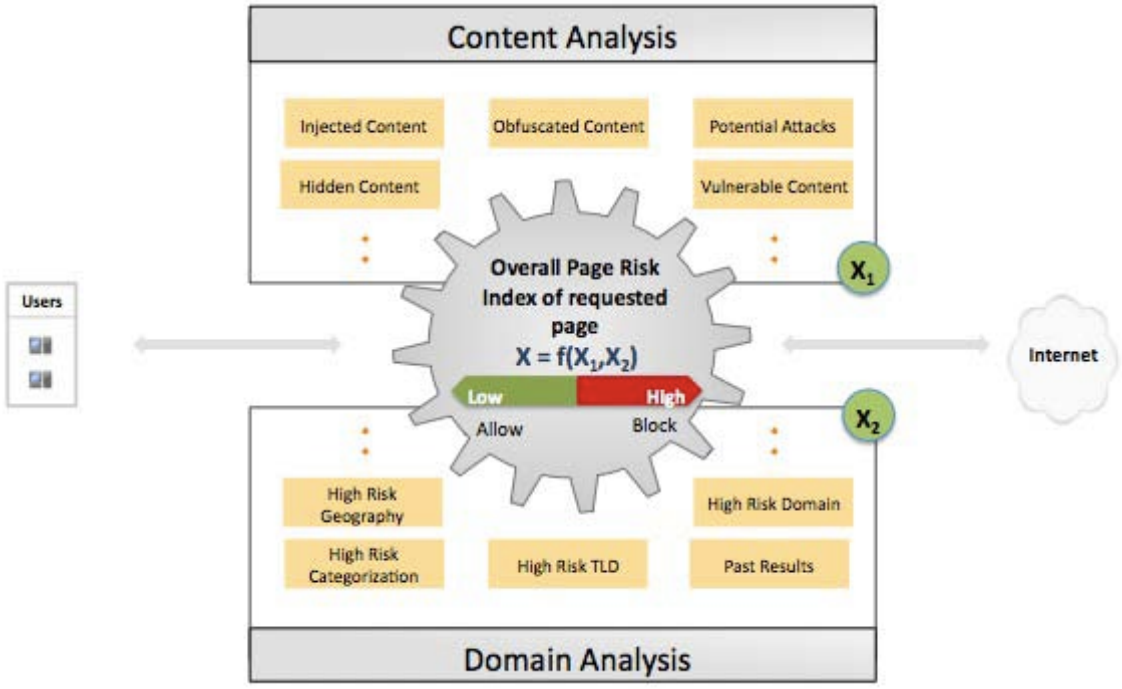

Figure 6 - Web Risk Index Calculation

The various risk indicators will be drawn from two separate control categories. Each category is defined below along with a summary of some of the key risk categories:

1. **Domain Analysis** – A weighted risk score is calculated for the domain hosting the content requested.
   a. **Geography** – Using geoIP technology, the geographic source of the content is determined. Based on a statistical analysis of the geographic distribution of past malicious content, a risk score is assigned.
   b. **Categorization** – Certain content categorizations such as pornography, shareware, etc. have a higher statistical likelihood of hosting malicious content.
   c. **TLD** – Given the availability and cost of certain top level domains (TLDs), analysis shows that some are more likely than others to host malicious content.
   d. **Domain** – A variety of partner data feeds are inspected to determine if the domain has historically been a source of malicious content.
   e. **Past Results** – Historical results are taken into considered when calculating the overall domain risk score.
2. **Content Analysis** – A weighted risk score is calculated by inspecting all content returned for a given request. The inspection is done in real-time and all content is inspected.
   a. **Injected Content** – Attackers commonly inject malicious content into otherwise legitimate websites. Page content is inspected to identify code injected into a web page, designed to directly initiate a browser attack or redirect the browser to an alternate page hosting malicious content.
   b. **Hidden Content** – HTML code such as zero-pixel IFRAMES/images designed to pull content from a third party domain without providing a visual indicator.

c. **Obfuscated Content** – Attackers will commonly obfuscate malicious content such as JavaScript in an effort to hide the true purpose of code or complicate debugging efforts.
d. **Vulnerable Content** – Attackers may include content designed to trigger known web browser vulnerabilities.
e. **Potential Attacks** – Content inspection may reveal potential attack vectors.

Scoring

Figure 7 - Page Risk Index User Interface

All variables within both the page and domain risk index categories are appropriately weighted. A total PRI score is then calculated with a value between 0 and 100. Enterprises can control the acceptable risk level based on their own risk tolerance. When a dynamic PRI score is calculated which exceeds the pre-defined threshold set via the Zscaler administrative UI, the content will be blocked and logged. Administrators will then be able to review blocked requests via the reporting capabilities within the Secure module.

PRI scores would not need to be calculated during a request for which a definitive rule was in place, which allowed or disallowed the request outright. Having a separate block list rule, which prohibited any traffic to Site X, would be an example of a situation for which a PRI calculation would not be required. Any request to Site X would be denied and there would not therefore be any reason to perform a PRI calculation. The PRI score would instead be calculated in those situations where a rule was not in place to definitively determine if a given request should be allowed or denied. The PRI score, when compared to a predefined and customizable risk threshold would then be used to determine if the request is permitted.

Functionality

Leveraging the bi-directional deep inspection capabilities previously discussed, Zscaler has deployed protections against a variety of attacks. A 'defense-in-depth' approach is applied by *Knowledge of Destination, Payload, Application and Content*.

Botnets

Monitoring the destination of outbound traffic is used to identify direct interactions with known C&C servers. Botnet research is conducted to then additionally deploy signatures that can identify unique patterns within traffic for various botnets. In doing so, newly deployed botnet C&C servers can continually be identified and blocked accordingly.

Malicious Active Content

Sites hosting malicious content are continually identified through partner feeds, internal data mining and content-based inspection. Known vulnerable ActiveX controls are also blocked when identified within response bodies. With the prevalence of vulnerable ActiveX controls installed on typical Windows systems, this has become one of the most popular attack vectors leading to compromised PCs.

Phishing

Block lists provide an efficient approach to identifying known phishing sites. A variety of block lists are combined from partner feeds and internal data mining efforts, ensuring broad coverage. Such lists are continually updated and provide effective protection against known phishing sites. Sites are however

continually emerging and block lists alone are not sufficient. Heuristic detection methods are therefore additionally employed. Real-time response body inspection allows for the recognition of traits commonly found on phishing sites, thereby identifying previous unknown phishing sites.

Communication

Both infected machines and end users seeking to bypass other security controls employ unauthorized communications. Zscaler has researched common IRC clients and anonymizers to deploy detections based not on block lists but on unique patterns within the communication protocols. This ensures that traffic is blocked regardless of destination.

Cross Site Scripting

Cross-site scripting attacks (XSS) are by far the most prevalent web application vulnerability found on otherwise legitimate websites. XSS can be leveraged by attackers to control scripting languages such a JavaScript, which are then executed within browsers visiting vulnerable sites. Zscaler employs two separate approaches to identify XSS vulnerabilities. The first monitors requests and identifies the presence of active script when such content in not appropriate. The second patent pending approach injects a unique Zscaler cookie into communication with any domain. When content from the cookie is identified in any subsequent request, attempts to steal the cookie's content—the typical goal of XSS attacks—can be thwarted.

Control Access to Suspicious Destinations

Employing geoIP based technologies, Zscaler is able to empower administrators to block content from any geographic location.

P2P Control

Once again, by researching popular applications in the P2P categories of file sharing, anonymizers and VoIP applications, Zscaler is able to identify and block alternate protocols being tunneled through HTTP/HTTPS transactions.

Conclusion

With a globally deployed, high-speed architecture, Zscaler has implemented a platform capable of the deep content inspection necessary for robust security in a SaaS solution. The Zscaler platform permits all levels of inspection including *Knowledge of Destination, Payload, Application and Content*. With the ability to conduct bi-directional, real-time inspection, emerging threats can be addressed without the need to deploy and manage software internally. Protections are maintained, managed and updated continuously without any necessary intervention on the part of those responsible for administering the service within the enterprise. Policies can be updated and reports reviewed through an intuitive web portal which permits for uniform protection of all enterprise clients regardless of location.