Operations Area Working Group F. Baker
Internet-Draft Cisco Systems
Intended status: BCP January 21, 2012
Expires: July 24, 2012

On Firewalls in Internet Security
draft-baker-opsawg-firewalls-00

Abstract

There is an ongoing discussion regarding the place of firewalls in security. This note is intended to capture and try to make sense out of it.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on July 24, 2012.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction 3
2. Common kinds of firewalls 3
  2.1. Perimeter security: Protection from aliens and intruders 4
  2.2. Pervasive access control 5
  2.3. Intrusion Management: Contract and Reputation filters . . 5
3. Reasoning about Firewalls 7
  3.1. The End-to-End Principle 7
  3.2. Building a communication 8
  3.3. The middle way 8
4. Recommendations 9
5. IANA Considerations 10
6. Security Considerations 10
7. Acknowledgements 10
8. References 11
  8.1. Normative References 11
  8.2. Informative References 11
Author's Address 11

1. Introduction

There is an ongoing discussion regarding the place of firewalls in security. This note is intended to capture and try to make sense out of it.

The IETF has a long and fractured discussion on security. Many early RFCs simply didn't address the topic - and said as much. When the IESG started complaining about that, it was told that there was no market interest in the topic that was measurable in money spent. Those who *were* interested in the topic set forth frameworks, rules, and procedures without necessarily explaining how they would be useful in deployment, and dismissed questions as "from those who don't understand." In many cases, as a result, deployments have been underwhelming in both quantity and quality, and the Internet is noted for its problems with security. What is clear is that people need to think clearly about security, their own and that of others. What is not clear is how to do so in a coherent and scalable manner.

Prophylactic perimeter security in the form of firewalls, and the proper use of them, have been a fractious sub-topic in this area. One could compare them to the human skin. The service that the skin performs for the rest of the body is to keep common crud out, and as a result prevent much damage and infection that could otherwise occur. The body supplies prophylactic perimeter security for itself and then presumes that the security perimeter has been breached; real defenses against attacks on the body include powerful systems that detect changes (anomalies) counterproductive to human health, and recognizable attack syndromes such as common or recently-seen diseases. One might well ask, in view of those superior defenses, whether there is any value in the skin at all; the value is easily stated, however. It is not in preventing the need for the stronger solutions, but in making their expensive invocation less needful and more focused.

This note will address common kinds of firewalls and the claims made for them. It will suggest a line of reasoning about the use of firewalls. It will attempt to end the bickering on the topic, which is, for the most part, of little value in illuminating the discussion.

2. Common kinds of firewalls

There are at least three common kinds of firewalls:

- Context or Zone-based firewalls, that protect systems within a perimeter from systems outside it,

- Pervasive routing-based measures, which protect intermingled systems from each other by enforcing role-based policies, and
- Systems that analyze application behavior and trigger on events that are unusual, match a signature, or involve an untrusted peer.

2.1. Perimeter security: Protection from aliens and intruders

As discussed in [RFC6092], the most common kind of firewall is used at the perimeter of a network. Perimeter security assumes two things: that applications and equipment inside the perimeter are under the control of the local administration and are therefore probably doing reasonable things, and that applications and equipment outside the perimeter are unknown. It may make simple permission rules, such as that external web clients are permitted to access a specific web server or that SMTP peers are permitted to access internal SMTP MTAs. Apart from those rules, a session may be initiated from inside the perimeter, and responses from outside will be allowed through the firewall, but sessions may never be initiated from outside.

In addition, perimeter firewalls often perform some level of testing, either as application proxies or through deep packet inspection, to verify that the protocol claimed to be being passed is in fact the protocol being passed.

The existence and definition of zone-based perimeter defenses is arguably a side-effect of the deployment of Network Address Translation [RFC2993]; applications frequently make the mistake of coupling application identities to network layer addresses, and in so doing make two other coupling assumptions: that an address useful to and understood by one application is useful to and understood by another, and that addresses are unlikely to change within a time frame useful to the application. Network Address Translation forces the translator to interpret packet payloads and change addresses where used by applications. If the transport or application headers are not understood by the translator, this has the effect of damaging or preventing communication. Detection of such issues can be sold as a security feature, although it is really a side-effect of a failure.

While this can have useful side-effects, such as preventing the passage of attack traffic that masquerades as some well-known protocol, it also has the nasty side-effect of making innovation difficult. For example, One of the issues in the deployment of Explicit Congestion Notification [RFC3168], for example, has been that common firewalls often test unused bits and require them to be set to zero to close covert channels. A similar problem has slowed the deployment of SCTP [RFC4960], in that a firewall will often not

permit a protocol it doesn't know even if a user behind it opens the session. When a new protocol or feature is defined, the firewall needs to stop applying that rule, and that can be difficult to make happen.

2.2. Pervasive access control

Another access control model, often called "Role-based", tries to control traffic in flight regardless of the perimeter. Given a rule that equipment located in a given routing domain or with a specific characteristic (such as "student dorms") should not be able to access equipment in another domain or with a specific characteristic (such as "academic records"), it might prevent routing from announcing the second route in the domain of the first, or it might tag individual packets ("I'm from the student dorm") and filter on those tags at enforcement points throughout network. Such rules can be applied to individuals are well as equipment; in that case, the host needs to tag the traffic, or there must be a reliable correlation between equipment and its user.

One common use of this model is in data centers, in which physical or virtual machines from one tenant (which is not necessarily an "owner" as much as it is a context in which the system is used) might be co-resident with physical or virtual machines from another. Inter-tenant attacks, espionage, and fraud are prevented by enforcing a rule that traffic from systems used by any given tenant is only delivered to other systems used by the same tenant. This might, of course have nuances; under stated circumstances, identified systems or identified users might be able to cross such a boundary.

The major impediment in deployment is complexity. The administration has the option to assign policies for individuals on the basis of their current location (e.g. as the cross-product of people, equipment, and topology), meaning that policies can multiply wildly. The administrator that applies a complex role-based access policy is probably most justly condemned to live in the world he or she has created.

2.3. Intrusion Management: Contract and Reputation filters

The model proposed in Advanced Security for IPv6 CPE [I-D.vyncke-advanced-ipv6-security] could be compared to purchasing an anti-virus software package for one's computer. The proposal is to install a set of filters, perhaps automatically updated, that identify "bad stuff" and make it inaccessible, while not impeding anything else.

It depends on four basic features:

- A frequently-updated signature-based Intrusion Prevention System which inspects a pre-defined set of protocols at all layers (from layer-3 to layer-7) and uses a vast set of heuristics to detect attacks within one or several flow. Upon detection, the flow is terminated and an event is logged for further optional auditing.

- A centralized reputation database that scores prefixes for degree of trust. This is unlikely to be on addresses per se, as Privacy Addresses change regularly and frequently.

- Local correlation of attack-related information, and

- Global correlation of attacks seen, in a reputation database

The proposal doesn't mention anomaly-based intrusion detection, which could be used to detect day-zero attacks and new applications or attacks. This would be an obvious extension.

The comparison to anti-virus software is real; anti-virus software uses similar algorithms, but on API calls or on data exchanged rather than on network traffic, and for identified threats is often effective.

The proposal also has weaknesses:

- People don't generally maintain anti-virus packages very well, letting contracts expire,

- Reputation databases have a bad reputation for distributing information which is incorrect or out of date,

- Anomaly-based analysis identifies changes but is often ineffective in determining whether new application or application behaviors are pernicious (false positives). Someone therefore has to actively decide - a workload the average homeowner might have little patience for, and

- Signature-based analysis applies to attacks that have been previously identified, and must be updated as new attacks develop. As a result, in a world in which new attacks literally arise daily, the administrative workload and be intense, and reflexive responses like accepting https certificates that are out of date or the download and installation of unsigned software on the assumption that the site admin is behind are themselves vectors for attack.

Security has to be maintained to be useful, because attacks are maintained.

3. Reasoning about Firewalls

3.1. The End-to-End Principle

One common complaint about firewalls in general is that they violate the End-to-End Principle [Saltzer]. The End-to-End Principle is often incorrectly stated as requiring that "application specific functions ought to reside in the end hosts of a network rather than in intermediary nodes, provided they can be implemented 'completely and correctly' in the end hosts" or that "there should be no state in the network." What it actually says is heavily nuanced, and is a line of reasoning applicable when considering any two communication layers.

> [Saltzer] "presents a design principle that helps guide placement of functions among the modules of a distributed computer system. The principle, called the end-to-end argument, suggests that functions placed at low levels of a system may be redundant or of little value when compared with the cost of providing them at that low level."

In other words, the End-to-End Argument is not a prohibition against lower layer retries of transmissions, which can be important in certain LAN technologies, nor of the maintenance of state, nor of consistent policies imposed for security reasons. It is, however, a plea for simplicity. Any behavior of a lower communication layer, whether found in the same system as the higher layer (and especially application) functionality or in a different one, that from the perspective of a higher layer introduces inconsistency, complexity, or coupling extracts a cost. That cost may be in user satisfaction, difficulty of management or fault diagnosis, difficulty of future innovation, reduced performance, or other forms. Such costs need to be clearly and honestly weighed against the benefits expected, and used only if the benefit outweighs the cost.

From that perspective, introduction of a policy that prevents communication under an understood set of circumstances, whether it is to prevent access to pornographic sites or prevents traffic that can be characterized as an attack, does not fail the end to end argument; there are any number of possible sites on the network that are inaccessible at any given time, and the presence of such a policy is easily explained and understood.

What does fail the end-to-end argument is behavior that is intermittent, difficult to explain, or unpredictable. If I can sometimes reach a site and not at other times, or reach it using this host or application but not another, I wonder why that is true, and may not even know where to look for the issue.

3.2. Building a communication

Any communication requires at least three components:

- a sender, someone or some thing that sends a message,
- a receiver, someone or some thing that receives the message, and
- a channel, which is a medium by which the message is communicated.

In the Internet, the IP network is the channel; it may traverse something as simple as a directly connected cable or as complex as a sequence of ISPs, but it is the means of communication. In normal communications, a sender sends a message via the channel to the receiver, who is willing to receive and operate on it. In contrast, attacks are a form of harassment. A receiver exists, but is unwilling to receive the message, has no application to operate on it, or is by policy unwilling to. Attacks on infrastructure occur when message volume overwhelms infrastructure or uses infrastructure but has no obvious receiver.

By that line of reasoning, a firewall primarily protects infrastructure, by preventing traffic that would attack it from it. The best prophylactic might use a procedure for the dissemination of Flow Specification Rules [RFC5575] to drop traffic sent by an unauthorized or inappropriate sender or which has no host or application willing to receive it as close as possible to the sender.

In other words, as discussed in Section 1, a firewall compares to the human skin, and has as its primary purpose the prophylactic defense of a network. By extension, the firewall also protects a set of hosts and applications, and the bandwidth that serves them, as part of a strategy of defense in depth. A firewall is not itself a security strategy; the analogy to the skin would say that a body protected only by the skin has an immune system deficiency and cannot be expected to long survive. That said, every security solution has a set of vulnerabilities; the vulnerabilities of a layered defense is the intersection of the vulnerabilities of the various layers (e.g., a successful attack has to thread each layer of defense).

3.3. The middle way

There is therefore no one way to prevent attacks; as noted in Section 2, there are different kinds of firewalls, and they address different views of the network. A zone-based firewall (Section 2.1) views the network as containing zones of trust, and deems applications inside its zone of protection to be trustworthy. A role-based firewall (Section 2.2) identifies parties on the basis of

membership in groups, and prevents unauthorized communication between groups. A reputation, anomaly, or signature-based intrusion management system depends on active administration, and permits known applications to communicate while excluding unknown or known-evil applications. In each case, the host or application is its own final bastion of defense, but preventing a host from accepting incoming traffic (so-called "host firewalls") does not defend infrastructure. Each type of prophylactic has a purpose, and none of them is a complete prophylactic defense.

Each type of defense, however, can be assisted by enabling an application running in a host to inform the network of what it is willing to receive. As noted in Section 2.1, a zone-based firewall, generally denies all incoming sessions and permits responses to sessions initiated outbound from the zone, but can in some cases be configured to also permit specific classes of incoming session requests, such as WWW or SMTP to an appropriate server. A simple way to enable a zone-based firewall to prevent attacks on infrastructure (traffic to an un instantiated address or to an application that is off) while not impeding traffic that has a willing host and application would be for the application to inform the firewall of that willingness to receive. The Port Control Protocol [I-D.ietf-pcp-base], or PCP, is an example of a protocol designed for that purpose.

4. Recommendations

A general recommendation for the IETF: the IETF should not seek to standardize something that is not being requested by consumers or industry.

Zone-based firewalls, when used, SHOULD exclude all session initiation from outside the zone regardless of attributes such as the use of IPsec. They SHOULD also facilitate the use of a protocol such as PCP by hosts to identify traffic (IPsec AH, IPsec ESP, transports in general, or transports using specified destination port ranges) that they are willing to receive, and interpret that into rules permitting specified traffic to those specific systems. Being fully automated and easily understood, such firewalls are appropriate for networks with passive administration.

Role-based firewalls can be implemented using routing technology. For example, if Alice should not be able to send a message to Bob, Alice might not be able to obtain Bob's address from DNS, Alice's routing system might not have a route to Bob, or Bob's routing system might not have a route to Alice. Role-based firewalls can also be implemented using filtering technology; Alice, Alice's router, Bob's

router, or Bob may have a filter that prevents communication between them. While there can be issues in specific cases, a routing implementation is generally more scalable and more easily managed.

Reputation, anomaly, or signature-based intrusion management is generally proprietary; a service maintains the list of exclusions, which must be updated as new kinds of attacks are developed. Implementations SHOULD be designed for frequent and scalable updating.

As further discussed in Section 2.1, firewalls of any type SHOULD NOT attempt to perform the kind of deep packet inspection and surgery that is common with Network Address Translators [RFC2993]. There is marginal value in detecting the spoofing of applications by attack traffic, but the side-effects of preventing protocol improvement and application innovation are destructive and unnecessary.

Apart from routing protocols and infrastructure protocols intended to manage network configuration and use of addresses such as DNS or DHCP, applications MUST NOT expect a peer to be able to interpret network layer addresses carried in their payload. Network layer addresses carried for documentation purposes, such as in an SMTP envelope or a syslog message, have other value and don't violate this recommendation.

5. IANA Considerations

This memo asks the IANA for no new parameters.

Note to RFC Editor: This section will have served its purpose if it correctly tells IANA that no new assignments or registries are required, or if those assignments or registries are created during the RFC publication process. From the author"s perspective, it may therefore be removed upon publication as an RFC at the RFC Editor"s discretion.

6. Security Considerations

This note reasons about security considerations. It introduces no new ones.

7. Acknowledgements

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

8.2. Informative References

[I-D.ietf-pcp-base]
Wing, D., Cheshire, S., Boucadair, M., Penno, R., and P. Selkirk, "Port Control Protocol (PCP)", draft-ietf-pcp-base-22 (work in progress), January 2012.

[I-D.vyncke-advanced-ipv6-security]
Vyncke, E., Yourtchenko, A., and M. Townsley, "Advanced Security for IPv6 CPE", draft-vyncke-advanced-ipv6-security-03 (work in progress), October 2011.

[RFC2993] Hain, T., "Architectural Implications of NAT", RFC 2993, November 2000.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

[RFC4960] Stewart, R., "Stream Control Transmission Protocol", RFC 4960, September 2007.

[RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J., and D. McPherson, "Dissemination of Flow Specification Rules", RFC 5575, August 2009.

[RFC6092] Woodyatt, J., "Recommended Simple Security Capabilities in Customer Premises Equipment (CPE) for Providing Residential IPv6 Internet Service", RFC 6092, January 2011.

[Saltzer] Saltzer, JH., Reed, DP., and DD. Clark, "End-to-end arguments in system design", ACM Transactions on Computer Systems (TOCS) v.2 n.4, p277-288, Nov 1984.

Author's Address

Fred Baker
Cisco Systems
Santa Barbara, California 93117
USA

Email: fred@cisco.com