W H I T E P A P E R

FIREWALL LOAD BALANCING
Web Switching to Optimize Firewall Performance

- **FIREWALL LOAD BALANCING USING WEB SWITCHES**
- **IMPLEMENTATION ISSUES**
- **ADDING A DMZ**
- **FIREWALL HEALTH MONITORING**
- **PHYSICAL CONNECTION MONITORING**
- **HIGH AVAILABILITY FIREWALLS USING HOT STANDBY WEB SWITCHES**
- **WEB SWITCH ARCHITECTURE**

Alteon WebSystems, Inc.

50 Great Oaks Boulevard

San Jose, California 95119

408-360-5500

408-360-5501 fax

http://www.alteon.com

W H I T E P A P E R

WP0008B-US-6/99

1

Long a concern for many enterprises and Internet Service Providers (ISPs), network security has become a central issue, bringing firewalls into common use as a method of preventing unauthorized access to network resources.

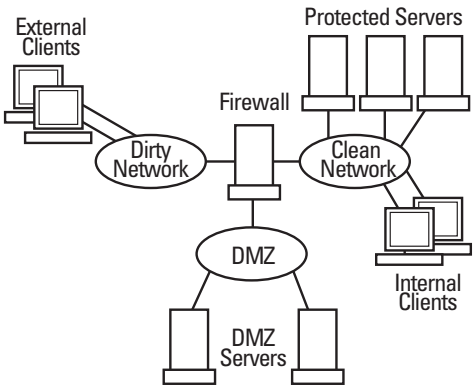
While the current generation of firewall products is very effective at preventing network intrusions, it has introduced its own problems to enterprise and ISP networks. In particular, current firewall technology limits performance and scalability and, because firewalls are often single points of failure, they can reduce network availability.

To understand why firewalls introduce these problems into networks, it is necessary to examine the typical firewall implementation. Currently, the most popular firewalls are software products installed on a server with two or three network interface cards and inserted into the data path. One network interface card is connected to the public side of the network, often to a router that connects to the Internet. This is known as the dirty side of the firewall.

The other network interface card is connected the to side of the network that attaches to resources that must be protected. This is known as the clean side of the firewall.

In some cases, an additional level of network security is implemented, more secure than the dirty side of the firewall but less secure than the clean side. This area is known as the DMZ (demilitarized zone) and is typically used for resources such as Internet Web servers where public access is required. Figure 1 shows a firewall configuration with a dirty network, a clean network and a DMZ.

FIGURE 1: Typical Firewall Configuration



Because firewalls sit on the data path, they can limit network performance and scalability. All network traffic passing between the dirty and clean networks must traverse the firewall.

Unfortunately, the processing architecture that works best for firewalls is not well suited to examining high volumes of data packets. Consequently, firewalls can slow communications by having to process every packet. Scaling the performance of firewalls can be difficult because it generally involves a forklift upgrade to a more powerful server.

Also because firewalls sit on the data path, they represent single points-of-failure that degrade network resource availability. While most firewalls can be deployed with commercially-available high-availability software in a hot-standby configuration, none of the solutions offered to date can support more than one firewall being active at a time. Consequently, users must buy and configure a second firewall and high-availability software and then watch it sit idly until a failure calls it into action.

Firewall load balancing with new Web switches solve these problems. Sophisticated Web switches allow firewalls to operate in parallel - giving users the ability to maximize firewall productivity, scale firewall performance without forklift upgrades and eliminate the firewall as a single point of failure.

FIREWALL LOAD BALANCING WITH WEB SWITCHES

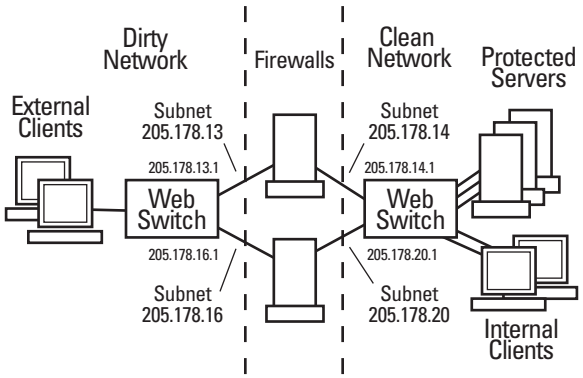
Unlike traditional packet switches, Web switches support Layer 4 and higher switching and processing functionality with the ability to maintain the state of individual TCP sessions. These devices provide excellent platforms for implementing firewall load balancing. Alteon WebSystems' Web switches - including the ACEswitch 180 PLUS, ACEswitch 180e, ACEdirector 2 and ACEdirector 3 - each provide a unique set of functionality that addresses all of the issues discussed below. For simplicity, the term "Web switch" will be used throughout the paper. However, any of these switches can implement the described functionality.

FIREWALL LOAD BALANCING IMPLEMENTATION

To implement firewall load balancing, two Web switches are required - one on the clean side of the firewalls and one on the dirty side. An example configuration is shown in Figure 2. In this example, filters that redirect all incoming IP traffic are configured on the ports on the dirty-side (left-hand) Web switch that connects to the dirty network and on the ports on the clean-side (right-hand) Web switch that connects to the clean network.

The redirection filters are configured to load balance the filtered traffic across a group of real servers represented by IP addresses. For the dirty side Web switch in Figure 2, the real servers are the IP addresses 205.178.14.1 and 205.178.20.1, which are IP interfaces on the clean side Web switch. Conversely, for the clean side Web switch, the real servers are 205.178.13.1 and 205.178.16.1, which are IP interfaces on the dirty side Web switch.

FIGURE 2: Typical Firewall Configuration



For each port on an Web switch that is connected to a firewall, a static route is configured to the port on the partner Web switch that connects to the same firewall. For example, in Figure 2, the port on the dirty-side Web switch that attaches to the upper firewall is configured with a static route that points to IP interface 205.178.14.1 on the clean-side Web switch while the port that attaches to the lower firewall has a static route to IP interface 205.178.20.1. Similarly, the port on the clean-side Web switch that attaches to the upper firewall is configured with a static route that points to IP interface 205.178.13.1 on the dirty-side Web switch while the port that attaches to the lower firewall has a static route pointing to IP interface 205.178.16.1.

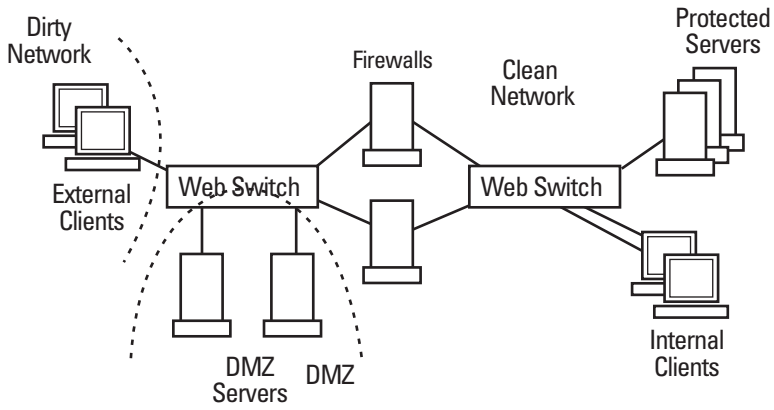
Since the real servers are IP interfaces on the partner Web switch and static routes are defined to these interfaces, one traversing the upper firewall and the other traversing the lower firewall, redirecting traffic along one or the other of these static routes has the effect of balancing the traffic across the firewalls. And because Web switches intelligently maintain state information about the traffic flowing through them, they ensure that all traffic between specific IP source/destination address pairs flows through the same firewall. This, in turn, ensures that sessions established by the firewalls are maintained for their duration.

ADDING A DMZ

With firewall load balancing, a DMZ can easily be configured. A major advantage of implementing a DMZ in conjunction with firewall load balancing is that traffic filtering - needed to determine what traffic goes to the DMZ and what must pass through the firewall - is performed on the Web switches. Doing this offloads firewalls from having to perform this function thereby significantly increasing firewall performance.

The DMZ servers can be connected to the Web switch on the dirty side of the firewall. Figure 3 shows a typical firewall load balancing configuration with a DMZ.

FIGURE 3: Typical Firewall Load Balancing Topology with DMZ



The DMZ servers can be attached to the Web switch directly or through an intermediate hub or switch. The Web switch is configured with filters to permit or deny access to the DMZ servers. In this manner, two levels of security are implemented; one that restricts access through the use of filters configured on the Web switch and another that restricts access through the use of stateful inspection performed by the firewalls.

FIREWALL HEALTH MONITORING

To maintain high availability, the Web switches monitor firewall health and direct packets only to healthy firewalls. The Web switches monitor the health of the firewalls by pinging each configured interface on its partner Web switch through each firewall on a regular basis.

For example, in Figure 2, the clean-side switch will ping the dirty-side switch on the 205.178.13.1 interface (via the upper firewall) and on the 205.178.16.1 interface (via the lower firewall).

If an Web switch interface fails to respond to a user-specified number of pings, it (and, by implication, the associated firewall), is placed in a Server Failed state. At this time, the partner Web switch stops sending traffic to that interface, distributing it across the remaining, healthy Web switch interfaces and firewalls.

When an Web switch's interface is in the Server Failed state, its partner Web switch continues to send pings to it at a user-configurable rate. After the first successful ping, the interface (and its associated firewall) is brought back into service.

PHYSICAL CONNECTION MONITORING

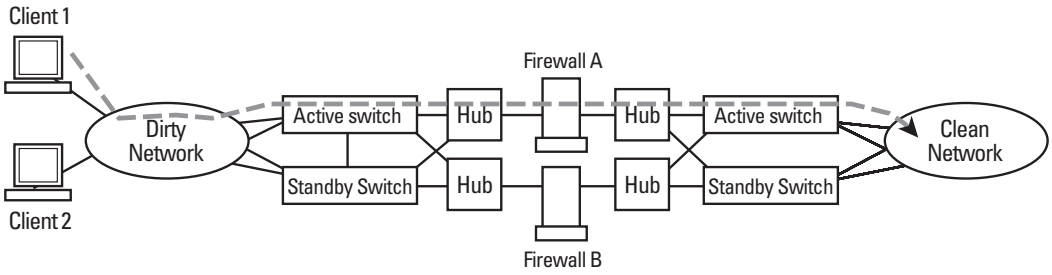
Web switches also monitor physical link status of switch ports connected to firewalls. If the physical link to a firewall goes down, that firewall is placed immediately in the Server Failed state. When an Web switch detects that a failed physical link to a firewall has been restored, it brings the firewall back into action.

HIGH AVAILABILITY FIREWALLS USING HOT STANDBY WEB SWITCHES

The firewall health monitoring techniques described in the previous section are one way that Web switches assure high application availability. Even higher levels of application availability can be achieved by using hot standby Web switches with firewall load balancing.

With firewall load balancing, Web switches can be used in pairs, with one active and the other in hot standby mode, to build network topologies with no system-wide single point-of-failure. This means that the Web switches are not single points-of-failure and that their use does not force a single point-of-failure at some other point in the network.

FIGURE 4: Hot Standby Web Switch Configuration



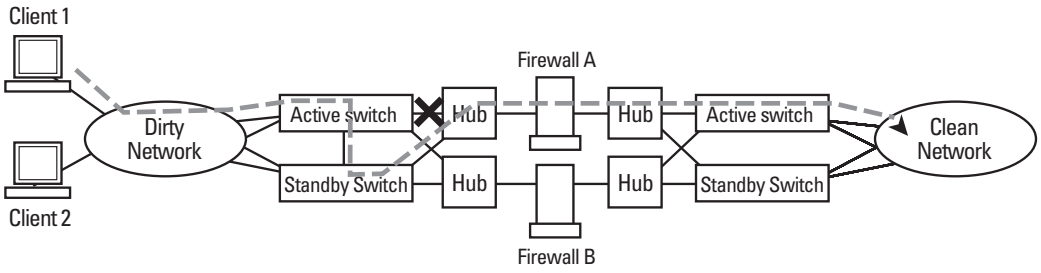
An example of a hot standby Web switch configuration is shown in Figure 4. Note that the topology supports use of redundant Web switches as well as redundant network devices connected to them.

In a hot standby configuration, no traffic flows through the standby Web switch. For example, all traffic from client 1 flows through the active Web switch on the dirty side of the firewalls, through one of the firewalls (firewall A in this example) and through the active Web switch on the clean side, as shown by the arrows.

In this example, four hubs are used, two connected to the Web switches on the dirty side of the firewall and two connected to the Web switches on the clean side. Other options are available. For example, multiple NICs with special failover capabilities or NICs with multiple network interfaces may be in each firewall.

When implementing a hot standby configuration, one of the Web switches in each pair is designated as the active switch, the other as the standby switch. A direct link, known as the failover link, is configured between the active Web switch and the standby Web switch. The failover link is used to send keep-alive messages between the active and standby Web switches. Data traffic crosses the failover link only in the event that at a port on the active Web switch fails.

FIGURE 5: Single Link Failure Reconfiguration



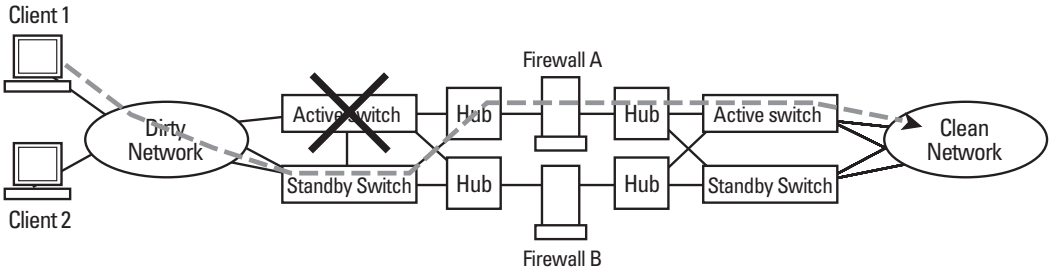
If the active Web switch detects a link failure, it communicates that information to the standby Web switch via the failover link. If the corresponding port on the standby Web switch is healthy, that port is activated.

For example, Figure 5 shows the case where the link from the active Web switch on the dirty side of firewalls to firewall A has failed. In this case, the port on the standby Web switch connecting firewall A becomes active. All traffic between client 1 and firewall A now passes through the active Web switch crosses the failover link, goes through the standby Web switch and traverses the newly activated link.

In circumstances where an entire active Web switch fails, standby Web switch will become active, as shown Figure 6. Traffic flows from client 1 to firewall A, as shown.

As noted earlier, the failover link is used to send keep-alive messages between the active and standby Web switches. If an Web switch fails to receive keep-alive messages from its counterpart, it may be an indication that the failover link has failed or that its counterpart has failed. The switch must distinguish between these two cases.

FIGURE 6: Web Switch Failure Reconfiguration



It is important to determine if the failover link has failed but both Web switches are healthy in order to avoid the split brain problem - where the standby Web switch attempts to become active while the active Web switch is still active. This situation could disrupt communications.

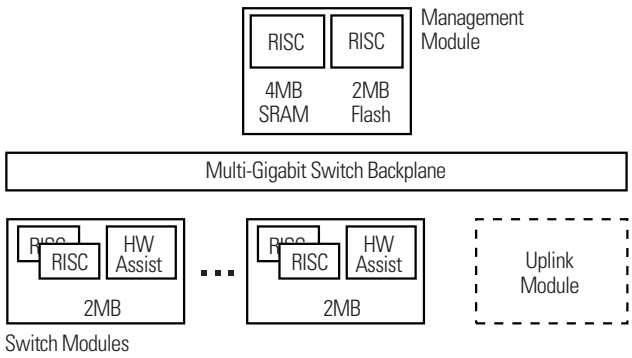
On the other hand, if the standby Web switch stops receiving keep-alive messages because the original active Web switch actually has failed, it must determine this so it can become active. Distinguishing between the two cases is accomplished by combining physical and data link layer health checks used on all Web switch ports with specialized messaging.

WEB SWITCH ARCHITECTURE

Examining thousands or tens of thousands of packets per second, determining which must be sent using existing connections to the firewalls, which must be directed to the firewalls using new connections and which should be redirected to the DMZ requires vast amounts of processing capacity and memory. Additional, background processing is also necessary to perform tasks such as checking the health of the firewalls, exchanging keep-alive messages when hot-standby switches are used and collecting and reporting statistics for network management.

The Web switch's distributed processing architecture is ideally designed for processor-intensive packet examination and manipulation. Each switch port integrates a switching ASIC that comprises a hardware-assisted forwarding engine and dual, 90-MHz RISC processors. Two additional RISC processors support switch-wide management functions. See Figure 7.

FIGURE 7: Web Switch Distributed Processing Architecture



On each switch port, the processor in the switching ASIC handles packet examination and forwarding. Background tasks such as firewall health checking, keep-alive message exchange, and statistics gathering and reporting are handled by the central management processors. With this architecture, processing tasks for each session are distributed to different processors for parallel operations, increasing overall performance.

SUMMARY

Alteon WebSystems' Web switching products, the ACEswitch 180 PLUS, the ACEswitch 180e, the ACEdirector 2 and the ACEdirector 3, offer high-performance firewall load balancing. Firewall load balancing offers significant benefits by improving firewall performance, scalability and availability. By coupling Alteon WebSystems' firewall load balancing with firewalls from leading vendors, users can build complete load balanced firewall solutions.