



CHAPTER 1

Overview

This chapter describes server load balancing (SLB) as implemented in the Cisco 4700 Series Application Control Engine (ACE) appliance. It contains the following major sections:

- [Server Load-Balancing Overview](#)
- [Load-Balancing Predictors](#)
- [Real Servers and Server Farms](#)
- [Configuring Traffic Classifications and Policies](#)
- [Connection Limits and Rate Limiting](#)
- [Operating the ACE Strictly as a Load Balancer](#)
- [Where to Go Next](#)

Server Load-Balancing Overview

Server load balancing (SLB) is the process of deciding to which server a load-balancing device should send a client request for service. For example, a client request may consist of a HyperText Transport Protocol (HTTP) GET for a web page or a File Transfer Protocol (FTP) GET to download a file. The job of the load balancer is to select the server that can successfully fulfill the client request and do so in the shortest amount of time without overloading either the server or the server farm as a whole.

The ACE supports the load balancing of the following protocols:

- Generic protocols

- HTTP
- Remote Authentication Dial-In User Service (RADIUS)
- Reliable Datagram Protocol (RDP)
- Real-Time Streaming Protocol (RTSP)
- Session Initiation Protocol (SIP)

Depending on the load-balancing algorithm—or *predictor*—that you configure, the ACE performs a series of checks and calculations to determine which server can best service each client request. The ACE bases server selection on several factors including the source or destination address, cookies, URLs, HTTP headers, or the server with the fewest connections with respect to load.

Load-Balancing Predictors

The ACE uses the following predictors to select the best server to fulfill a client request:

- Application response—Selects the server with the lowest average response time for the specified response-time measurement based on the current connection count and server weight (if configured).
- Hash address—Selects the server using a hash value based on either the source or destination IP address or both. Use these predictors for firewall load balancing (FWLB). For more information about FWLB, see [Chapter 6, Configuring Firewall Load Balancing](#).
- Hash content—Selects the server using a hash value based on a content string in the Trusted Third Parties (TTP) packet body.
- Hash cookie—Selects the server using a hash value based on a cookie name.
- Hash header—Selects the server using a hash value based on the HTTP header name.
- Hash URL—Selects the server using a hash value based on the requested URL. You can specify a beginning pattern and an ending pattern to match in the URL. Use this predictor method to load balance cache servers.
- Least bandwidth—Selects the server that processed the least amount of network traffic based on the average bandwidth that the server used over a specified number of samples.

- Least connections—Selects the server with the fewest number of active connections based on server weight. For the least-connections predictor, you can configure a slow-start mechanism to avoid sending a high rate of new connections to servers that you have just put into service.
- Least loaded—Selects the server with the lowest load based on information obtained from Simple Network Management Protocol (SNMP) probes. To use this predictor, you must associate an SNMP probe with it.
- Round-robin—Selects the next server in the list of real servers based on the server weight (weighted round-robin). Servers with a higher weight value receive a higher percentage of the connections. This is the default predictor.

**Note**

The hash predictor methods do not recognize the weight value that you configure for real servers. The ACE uses the weight that you assign to real servers only in the least-connections, application-response, and round-robin predictor methods.

For more information about load-balancing predictors, see [Chapter 2, Configuring Real Servers and Server Farms](#).

Real Servers and Server Farms

This section briefly describes real servers and server farms and how they are implemented on the ACE. It contains the following topics:

- [Real Servers](#)
- [Server Farms](#)
- [Health Monitoring](#)

Real Servers

To provide services to clients, you configure *real servers* (the actual physical servers) on the ACE. Real servers provide client services such as HTTP or XML content, hosting websites, FTP file uploads or downloads, redirection for web pages that have moved to another location, and so on. The ACE also allows you to configure backup servers in case a server is taken out of service for any reason.

After you create and name a real server on the ACE, you can configure several parameters, including connection limits, health probes, and weight. You can assign a weight to each real server based on its relative importance to other servers in the server farm. The ACE uses the server weight value for the weighted round-robin and the least-connections load-balancing predictors. For a listing and brief description of the ACE predictors, see the “[Load-Balancing Predictors](#)” section. For more detailed information about the ACE load-balancing predictors and server farms, see [Chapter 2, Configuring Real Servers and Server Farms](#).

Server Farms

Typically, in data centers, servers are organized into related groups called *server farms*. Servers within server farms often contain identical content (referred to as mirrored content) so that if one server becomes inoperative, another server can take its place immediately. Also, mirrored content allows several servers to share the load of increased demand during important local or international events, for example, the Olympic Games. This sudden large demand for content is called a *flash crowd*.

After you create and name a server farm, you can add existing real servers to it and configure other server-farm parameters, such as the load-balancing predictor, server weight, backup server, health probe, and so on. For a description of the ACE predictors, see the “[Load-Balancing Predictors](#)” section. For more detailed information about the ACE load-balancing predictors and server farms, see [Chapter 2, Configuring Real Servers and Server Farms](#).

Health Monitoring

You can instruct the ACE to check the health of servers and server farms by configuring health probes (sometimes referred to as *keepalives*). After you create a probe, you assign it to a real server or a server farm. A probe can be one of many types, including TCP, ICMP, Telnet, HTTP, and so on. You can also configure scripted probes using the TCL scripting language.

The ACE sends out probes periodically to determine the status of a server, verifies the server response, and checks for other network problems that may prevent a client from reaching a server. Based on the server response, the ACE can place the server in or out of service, and, based on the status of the servers in the server farm, can make reliable load-balancing decisions. For more information about out-of-band health monitoring, see [Chapter 4, Configuring Health Monitoring](#).

Configuring Traffic Classifications and Policies

The ACE uses several configuration elements to classify (filter) interesting traffic and then to perform various actions on that traffic before making the load-balancing decision. These filtering elements and subsequent actions form the basis of a traffic policy for SLB. This section contains the following topics:

- [Filtering Traffic with ACLs](#)
- [Classifying Layer 3 and Layer 4 Traffic](#)
- [Classifying Layer 7 Traffic](#)
- [Configuring a Parameter Map](#)
- [Creating Traffic Policies](#)
- [Applying Traffic Policies to an Interface Using a Service Policy](#)

Filtering Traffic with ACLs

To permit or deny traffic to or from a specific IP address or an entire network, you can configure an access control list (ACL). ACLs provide a measure of security for the ACE and the data center by allowing access only to traffic that you explicitly authorize on a specific interface or on all interfaces. An ACL consists of a series of permit or deny entries with special criteria for the source address, destination address, protocol, port, and so on. All ACLs contain an implicit deny statement, so you must include an explicit permit entry to allow traffic to and through the ACE. For more information about ACLs, see the *Cisco 4700 Series Application Control Engine Appliance Security Configuration Guide*.

Classifying Layer 3 and Layer 4 Traffic

To classify Layer 3 and Layer 4 network traffic, you configure class maps and specify match criteria according to your application requirements. When a traffic flow matches certain match criteria, the ACE applies the actions specified in the policy map with which the class map is associated. A policy map acts on traffic ingressing the interface to which the policy map is applied through a service policy (globally to all VLAN interfaces in a context or to a single VLAN interface).

Class maps that operate at Layer 3 and Layer 4 for SLB typically use virtual IP (VIP) addresses as matching criteria. For details about Layer 3 and Layer 4 class maps for SLB, see [Chapter 3, Configuring Traffic Policies for Server Load Balancing](#).

Classifying Layer 7 Traffic

In addition to Layer 3 and Layer 4 class maps, you can also configure Layer 7 class maps for advanced load-balancing matching criteria, such as HTTP cookie, header, and URL settings. After you configure a Layer 7 class map, you associate it with a Layer 7 policy map. You cannot apply a Layer 7 policy map to an interface directly (see the “[Creating Traffic Policies](#)” section). For details about Layer 7 class maps for SLB, see [Chapter 3, Configuring Traffic Policies for Server Load Balancing](#).

Configuring a Parameter Map

A parameter map combines related HTTP or RTSP actions for use in a Layer 3 and Layer 4 policy map. Parameter maps provide a means of performing actions on traffic that ingresses an ACE interface based on certain criteria, such as HTTP header and cookie settings, server connection reuse, the action to take when an HTTP header, cookie, or URL exceeds a configured maximum length, and so on. After you configure a parameter map, you associate it with a Layer 3 and Layer 4 policy map. For details about configuring an HTTP or RTSP load-balancing parameter map, see [Chapter 3, Configuring Traffic Policies for Server Load Balancing](#).

Creating Traffic Policies

The ACE uses policy maps to combine class maps and parameter maps into traffic policies and to perform certain configured actions on the traffic that matches the specified criteria in the policies. Policy maps operate at Layer 3 and Layer 4, as well as Layer 7. Because the ACE considers a Layer 7 policy map a child policy, you must associate each Layer 7 policy map with a Layer 3 and Layer 4 policy map before you can apply the traffic policy to an interface using a service policy. For more information about configuring SLB traffic policies, see [Chapter 3, Configuring Traffic Policies for Server Load Balancing](#).

Applying Traffic Policies to an Interface Using a Service Policy

To apply a traffic policy to one or more interfaces, you use a service policy. You can use a service policy on an individual interface or globally on all interfaces in a context in the input direction only. When you use a service policy on an interface, you apply and activate a Layer 3 and Layer 4 policy map with all its class-map, parameter-map, and Layer 7 policy-map associations and match criteria. For more information about using a service policy to apply a traffic policy to an interface, see [Chapter 3, Configuring Traffic Policies for Server Load Balancing](#).

Connection Limits and Rate Limiting

To help protect system resources, the ACE allows you to limit the following items:

- Maximum number of connections
- Connection rate (connections per second applied to new connections destined to a real server)
- Bandwidth rate (bytes per second applied to network traffic between the ACE and a real server in both directions)

For more information, see [Chapter 2, Configuring Real Servers and Server Farms](#) and the *Cisco 4700 Series Application Control Engine Appliance Security Configuration Guide*.

Operating the ACE Strictly as a Load Balancer

You can operate your ACE strictly as an SLB device. If you want to use SLB only, you must configure certain parameters and disable some of the ACE security features. By default, the ACE performs TCP/IP normalization checks and ICMP security checks on traffic that enters the ACE interfaces. Using the following configuration will also allow asymmetric routing as required by your network application.

The major configuration items are as follows:

- Configuring a global permit-all ACL and applying it to all interfaces in a context to open all ports

- Disabling TCP/IP normalization
- Disabling ICMP security checks
- Configuring SLB

To operate the ACE for SLB only, perform the following steps:

- Step 1** Configure a global permit-all ACL and apply it to all interfaces in a context. This action will open all ports in the current context.

```
host1/Admin(config)# access-list ACL1 extended permit ip any any
host1/Admin(config)# access-group input ACL1
```

- Step 2** Disable the default TCP/IP normalization checks on each interface where you want to configure a VIP using a load-balancing service policy. For details about TCP normalization, see the *Cisco 4700 Series Application Control Engine Appliance Security Configuration Guide*.

```
host1/Admin(config)# interface vlan 100
host1/Admin(config-if)# no normalization
```



Caution

Disabling TCP normalization may expose your ACE and your data center to potential security risks. TCP normalization helps protect the ACE and the data center from attackers by enforcing strict security policies that are designed to examine traffic for malformed or malicious segments.

- Step 3** Disable the default ICMP security checks on each interface where you want to configure a VIP using a load-balancing service policy. For details about the ICMP security checks, see the *Cisco 4700 Series Application Control Engine Appliance Security Configuration Guide*.

```
host1/Admin(config-if)# no icmp-guard
```



Caution

Disabling the ACE ICMP security checks may expose your ACE and your data center to potential security risks. In addition, after you enter the **no icmp-guard** command, the ACE no longer performs Network Address Translations (NATs) on the ICMP header and payload in error packets, which potentially can reveal real host IP addresses to attackers.

Step 4 Configure SLB. For details, see the remaining chapters in this guide.

Where to Go Next

To start configuring SLB on your ACE, proceed to [Chapter 2, Configuring Real Servers and Server Farms](#).

