CHAPTER 7

Configuring Real Servers and Server Farms

Date: 2/21/11

This chapter describes how to configure real servers and server farms on the Cisco Application Control Engine (ACE) using Cisco Application Networking Manager (ANM).

This chapter includes the following sections:

In formation About Server Load Balancing

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 can consist of an HTTP GET for a Web page or an 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.

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

ANM allows you to configure load balancing using:

- Virtual servers—See Configuring Virtual Servers, page 6-2.
- Real servers—See Configuring Real Servers, page 7-4.
- Server farms—See Configuring Server Farms, page 7-14.
- Sticky groups—See Configuring Sticky Groups, page 8-6.
- Parameter maps—See Configuring Parameter Maps, page 9-1.

For more information about SLB as configured and performed by the ACE, see:

- Configuring Virtual Servers, page 6-2
- Load-Balancing Predictors, page 7-2
- Real Servers, page 7-3
- Server Farms, page 7-4
- Configuring Health Monitoring, page 7-29
- TCL Scripts, page 7-29
- Configuring Stickiness, page 8-1

This section includes the following topics:

- Load-Balancing Predictors, page 7-2
- Real Servers, page 7-3
- Server Farms, page 7-4

Load-Balancing Predictors

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

- 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).

Note FWLB allows you to scale firewall protection by distributing traffic across multiple firewalls on a per-connection basis. All packets belonging to a particular connection must go through the same firewall. The firewall then allows or denies transmission of individual packets across its interfaces. For more information about configuring FWLB on the ACE, see the *Cisco 4700 Series Application Control Engine Appliance Server Load-Balancing Configuration Guide*.

- Hash Content— Selects the server by using a hash value based on the specified content string of the HTTP 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 Layer4—Selects the server using a Layer 4 generic protocol load-balancing method.
- 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. Cache servers perform better with the URL hash method because you can divide the contents of the caches evenly if the traffic is random enough. In a redundant configuration, the cache servers continue to work even if the active ACE switches over to the standby ACE. For information about configuring redundancy, see the "Configuring High Availability" section on page 12-1.
- Least Bandwidth—Selects the server with the least amount of network traffic or a specified sampling period. Use this type for server farms with heavy traffic, such as downloading video clips.
- Least Connections—Selects the server with the fewest number of active connections based on server weight. For the least connection 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 as determined by information from SNMP probes.
- Response—Selects the server with the lowest response time for a specific response-time measurement.
- Round Robin—Selects the next server in the list of real servers based on server weight (weighted roundrobin). Servers with a higher weight value receive a higher percentage of the connections. This is the default predictor.

Note The different 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 round-robin and least-connections predictor methods.

Related Topics

Configuring the Predictor Method for Server Farms, page 7-20

Real Servers

To provide services to clients, you configure real servers on the ACE. Real servers are dedicated physical servers that you typically configure in groups called server farms. These servers provide client services such as HTTP or XML content, website hosting, FTP file uploads or downloads, redirection for web pages that have moved to another location, and so on. You identify real servers with names and characterize them with IP addresses, connection limits, and weight values. 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. The load-balancing predictor algorithms (for example, roundrobin, least connections, and so on) determine the servers to which the ACE sends connection requests. For a listing and brief description of the load-balancing predictors, see the "Load-Balancing Predictors" section on page 7-2.

The ACE uses traffic classification maps (class maps) within policy maps to identify traffic that meets defined criteria and to apply specific actions to that traffic based on the SLB configuration.

If a primary real server fails, the ACE takes that server out of service and no longer includes it in load-balancing decisions. If you configured a backup server for the real server that failed, the ACE redirects the primary real server connections to the backup server. For information about configuring a backup server, see the "Configuring Virtual Server Layer 7 Load Balancing" section on page 6-30.

The ACE can take a real server out of service for the following reasons:

- Probe failure
- ARP timeout
- Specifying Out Of Service as the administrative state of a real server
- Specifying Inservice Standby as the administrative state of a real server

The Out Of Service and Inservice Standby selections both provide the graceful shutdown of a server.

Related Topics

- Configuring Real Servers, page 7-4

- Configuring Health Monitoring for Real Servers, page 7-30

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, having mirrored content allows several servers to share the load of increased demand during important local or international events, such as the Olympic Games. This phenomenon of a 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 listing and brief description of load-balancing predictors, see the "Load-Balancing Predictors" section on page 7-2.

Related Topics

Configuring Server Farms, page 7-14

Configuring Real Servers

Real servers are dedicated physical servers that are typically configured in groups called server farms. These servers provide services to clients, such as HTTP or XML content, streaming media (video or audio), TFTP or FTP services, and so on. When configuring real servers, you assign names to them and specify IP addresses, connection limits, and weight values.

The ACE uses traffic classification maps (class maps) within policy maps to filter specified traffic and to apply specific actions to that traffic based on the load-balancing configuration. A load-balancing predictor algorithm (such as round-robin or least connections) determines the servers to which the ACE sends connection requests. For information about configuring class maps, see the "Configuring Virtual Context Class Maps" section on page 13-6.

This section includes the following topics:

- Configuring Load Balancing on Real Servers, page 7-4
- Displaying Real Server Statistics and Status Information, page 7-7

Configuring Load Balancing on Real Servers

You can configure load balancing on real servers.

Procedure

Step 1 Choose **Config > Devices >** *context* **> Load Balancing > Real Servers**.

The Real Servers table appears.

Step 2 In the Real Servers table, click **Poll Now** to instruct ANM to poll the devices and display the current values, and click **OK** when prompted if you want to poll the devices for data now.

Step 3 Click **Add** to add a new real server, or choose a real server you want to modify and click **Edit**.

The Real Servers configuration window appears.

Step 4 In the Real Servers configuration window, configure the server using the information in Table 7-1.

Table 7-1 Real Server Attributes

Field	Description
Name	Field that allows you to either enter a unique name for this server or accept the automatically incremented value in this field. Valid entries are unquoted text strings with no spaces and a maximum of 64 characters.
Type	Type of server: • **Host**—The real server provides content and services to clients. • **Redirect**—The server redirects traffic to a new location.
State	State of the real server: • **In Service**—The real server is in service. • **Out Of Service**—The real server is out of service.
Description	Brief description for this real server. Valid entries are strings of up to 240 characters. Spaces and special characters are allowed.
IP Address	Field that appears for only real servers specified as hosts. Enter a unique IP address in dotted-decimal format (such as 192.168.11.1). The IP address cannot be an existing virtual IP address (VIP).
Fail-On-All	Field that appears only for real servers identified as host servers. By default, real servers with multiple probes configured for them have an OR logic associated with them, which means that if one of the real server probes fails, the real server fails and enters the PROBE-FAILED state. Check this checkbox to configure a real server to remain in the OPERATIONAL state unless all probes associated with it fail (AND logic). The Fail-On-All function is applicable to all probe types.
Min. Connections	Minimum number of connections to be allowed on this server before the ACE starts sending connections again after it has exceeded the Max. Connections limit. This value must be less than or equal to the Max. Connections value. By default, this value is equal to the Max. Connections value. Valid entries are from 2 to 4000000.
Max. Connections	Maximum number of active connections allowed on this server. When the number of connections exceeds this value, the ACE stops sending connections to this server until the number of connections falls below the Min. Connections value. Valid entries are from 2 to 4000000, and the default is 4000000.
Weight	Field that appears only for real servers identified as hosts. Enter the weight to be assigned to this real server in a server farm. Valid entries are from 1 to 100, and the default is 8.
Probes	Field that appears only for real servers identified as hosts. In the Probes field, choose the probes to use for health monitoring in the Available Items list, and click **Add**. The probes appear in the Selected Items list. To remove probes that you do not want to use for health monitoring, choose them in the Selected Items list, and click **Remove**. The probes appear in the Available Items list.

Table 7-1 Real Server Attributes (continued)

Field	Description
Web Host Redirection	URL string used to redirect requests to another server. This field appears only for real servers identified as redirect servers. Enter the URL and port used to redirect requests to another server. Valid entries are in the form http://host.com:port where *host* is the name of the server and *port* is the port to be used. Valid host entries are unquoted text strings with no spaces and a maximum of 255 characters. Valid port numbers are from 1 to 65535. The relocation string supports the following special characters: • %h—Inserts the hostname from the request Host header • %p—Inserts the URL path string from the request
Redirection Code	Field that appears only for real servers identified as redirect servers. Choose the appropriate redirection code: • **N/A**—Webhost redirection code is not defined. • **301**—Requested resource has been moved permanently. For future references to this resource, the client should use one of the returned URIs. • **302**—Requested resource has been found, but has been moved temporarily to another location. For future references to this resource, the client should use the request URI because the resource may be moved to other locations from time to time.
Rate Bandwidth	Bandwidth rate is the number of bytes per second and applies to the network traffic exchanged between the ACE and the real server in both directions. Specify the real server bandwidth limit in bytes per second. Valid entries are from 2 to 300000000. The default is 300000000.
Rate Connection	Connection rate is the number of connections per second received by the ACE and applies only to new connections destined to a real server. Specify the limit for connections per second. Valid entries are from 2 to 350000. The default is 350000.

Step 5 Do one of the following:

- Click **Deploy Now** to deploy this configuration on the ACE and save your entries to the running-configuration and startup-configuration files.
- Click **Cancel** to exit the procedure without saving your entries and to return to the Real Servers table.
- Click **Next** to deploy your entries and to configure another real server.

Step 6 To display statistics and status information for an existing real server, choose a real server from the Real Servers table, then click **Details**. The **show rserver** *name* **detail** CLI command output appears. See the "Displaying Real Server Statistics and Status Information" section on page 7-7 for details.

Related Topics

- Managing Real Servers, page 7-7
- Configuring Health Monitoring for Real Servers, page 7-30
- Configuring Server Farms, page 7-14

- Configuring Sticky Groups, page 8-6

Displaying Real Server Statistics and Status Information

You can display statistics and status information for a particular real server.

Procedure

Step 1 Choose **Config > Devices >** ***context*** **> Load Balancing > Real Servers**.

The Real Servers table appears.

Step 2 In the Real Servers table, choose a real server from the Real Servers table, and click **Details**.

The **show rserver** *name* **detail** CLI command output appears. For details on the displayed output fields, see either the *Cisco ACE Module Server Load-Balancing Configuration Guide* or the *Cisco ACE 4700 Series Appliance Server Load-Balancing Configuration Guide*, Chapter 2, Configuring Real Servers and Server Farms.

Step 3 Click **Update Details** to refresh the output for the **show rserver** *name* **detail** CLI command. The new information appears in a separate panel with a new timestamp; both the old and the new real server statistics and status information appear side-by-side to avoid overwriting the last updated information.

Step 4 Click **Close** to return to the Real Servers table.

Related Topics

- Configuring Real Servers, page 7-4
- Managing Real Servers, page 7-7
- Displaying Real Servers, page 7-10

Managing Real Servers

The Real Servers table (Config > Operations > Real Servers) provides the following information by default for each server:

- Server name
- IP address
- Port
- Admin State (In Service, Out Of Service, or In Service Standby)
- Operational state (See Table 7-2 for descriptions of real server operational states.)
- Number of current connections
- Current server weight
- Associated server farm
- Associated virtual servers
- Device details
- Whether the server is part of a high-availability pair

In the table, N/A indicates that either the information is not available from the database or that it is not being collected via SNMP. To identify any SNMP-related issues, select the real server's virtual context in the object selector. If there are problems with SNMP, SNMP status will appear in the upper right above the content pane.

This section includes the following topics:

Activating Real Servers

You can activate a real server.

Procedure

Step 1 Choose **Config > Operations > Real Servers**.

The Real Servers table appears.

Step 2 From the Real Servers table, choose the servers that you want to activate, and click **Activate**.

The Activate Server window appears.

Step 3 In the Reason field of the Activate Server window, enter a reason for this action.

You might enter a trouble ticket, an order ticket, or a user message.

Note Do not enter a password in this field.

Step 4 Do one of the following:

- Click **OK** to activate the server and to return to the Real Servers table. The server appears in the table with the status Inservice.
- Click **Cancel** to exit this procedure without activating the server and to return to the Real Servers table.

Related Topics

- Managing Real Servers, page 7-7
- Suspending Real Servers, page 7-9
- Displaying Real Servers, page 7-10

Suspending Real Servers

You can suspend a real server.

Procedure

Step 1 Choose **Config > Operations > Real Servers**.

The Real Servers table appears.

Step 2 In the Real Servers table, choose the server that you want to suspend, and click **Suspend**.

The Suspend Real Servers window appears.

Step 3 In the Reason field of the Suspend Real Servers window, enter the reason for this action.

You might enter a trouble ticket, an order ticket, or a user message.

Note Do not enter a password in this field.

Step 4 From the Suspend Real Servers Type drop-down list, choose one of the following:

- **Graceful**
- **Suspend**
- **Suspend and Clear Connections** (clears the existing connections to this server as part of the shutdown process)

Note Graceful suspend and suspend options vary by device type. For the commands deployed by the device type when these options are selected, see the "CLI Commands Sent from the Real Server Table" section on page 7-12.

For the CSS and CSM, when you perform a graceful suspend operation, ANM saves the last known non-zero service (or real server) weight, which sets the weight to zero. ANM references the saved weight when performing an Activate operation. If the current weight is zero, and a non-zero weight has been saved for that service (or real server), the Activate operation also sets the weight to the saved value.

Step 5 Do one of the following:

- Click **Deploy Now** to suspend the server and to return to the Real Servers table. The server appears in the table with the status Out Of Service.
- Click **Cancel** to exit this procedure without suspending the server and to return to the Real Servers table.

Related Topics

- Managing Real Servers, page 7-7
- Activating Real Servers, page 7-8
- Displaying Real Servers, page 7-10

Modifying Real Servers

You can modify server weight and connection limits for real servers.

Procedure

Step 1 Choose **Config > Operations > Real Servers**.

The Real Servers table appears.

Step 2 In the Real Servers table, choose the servers whose configuration you want to modify, and click **Change Weight** below the table to the right of Activate and Suspend.

The Change Weight Real Servers window appears.

Step 3 In the Change Weight Real Servers window, enter the following information for the selected server:

- Reason for change such as trouble ticket, order ticket or user message.

  Note Do not enter a password in this field.

- Weight (For allowable ranges for each device type, see Table 7-4).

Step 4 Do one of the following:

- Click **Deploy Now** to accept your entries and to return to the Real Servers table. The server appears in the table with the updated information.
- Click **Cancel** to exit this procedure without saving your entries and to return to the Real Servers table.

Related Topics

- Managing Real Servers, page 7-7
- Activating Real Servers, page 7-8
- Displaying Real Servers, page 7-10

Displaying Real Servers

To display all real servers, choose **Config > Operations > Real Servers**.

The Real Servers table displays the following information by default:

- Server name
- IP address
- Port
- Admin State (In Service, Out Of Service, or In Service Standby)
- Operational state (See Table 7-2 for descriptions of real server operational states.)
- Number of current connections
- Current server weight
- Associated server farm

- Associated virtual servers
- Device details
- Whether the server is part of a high availability pair

In the table, N/A indicates that either the information is not available from the database or that it is not being collected via SNMP. To identify any SNMP-related issues, select the real server's virtual context in the object selector. If there are problems with SNMP, SNMP status will appear in the upper right above the content pane.

Table 7-2 Real Server Operational States

State	Description
Failed	Server has failed and will not be retried for the amount of time specified by its retry timer.
Inband probe failed	Server has failed the inband Health Probe agent.
Inservice	Server is in use as a destination for server load balancing client connections.
Inservice standby	Server is the backup real server, which remains inactive unless the primary real server fails.
Operation wait	Server is ready to become operational but is waiting for the associated redirect virtual server to be in service.
Out of service	Server is not in use by a server load balancer as a destination for client connections.
Probe failed	Server load-balancing probe to this server has failed. No new connections will be assigned to this server until a probe to this server succeeds.
Probe testing	Server has received a test probe from the server load balancer.
Ready to test	Server has failed and its retry timer has expired; test connections will begin flowing to it soon.
Return code failed	Server has been disabled because it returned an HTTP code that matched a configured value.
Test wait	Server is ready to be tested. This state is applicable only when the server is used for HTTP redirect load balancing.
Testing	Server has failed and has been given another test connection. The success of this connection is not known.
Throttle: DFP	DFP has lowered the weight of the server to throttle level; no new connections will be assigned to the server until DFP raises its weight.
Throttle: max clients	Server has reached its maximum number of allowed clients.
Throttle: max connections	Server has reached its maximum number of connections and is no longer being given connections.
Unknown	State of the server is not known.

Related Topics

- Displaying Real Server Statistics and Status Information, page 7-7
- Activating Real Servers, page 7-8
- Suspending Real Servers, page 7-9
- Modifying Real Servers, page 7-10

CLI Commands Sent from the Real Server Table

Table 7-3 displays the CLI commands dispatched to the device for a given Real Servers table option, and is sorted by device.

Table 7-3 CLI Commands Deployed from the Real Servers Table

Command	Sample CLI Sent
ACE Modules and Appliances	
Real Server Activation	`serverfarm host sf1` `rserver rs1 80` `inservice`
Real Server Graceful Suspend	`serverfarm host sf1` `rserver rs1 80` `inservice standby`
Real Server Suspend	`serverfarm host sf1` `rserver rs1 80` `no inservice`
Real Server Suspend and Clear Connections	`serverfarm host sf1` `rserver rs1 80` `no inservice` `clear conn rserver rs1 80 serverfarm sf1`
Real Server Change Weight	`serverfarm host sf1` `rserver rs1 80` `weight 2`
CSMs	
Real Server Activation	`serverfarm host sf1` `real 10.10.10.10 80` `inservice`
Real Server Graceful Suspend	`serverfarm host sf1` `real 10.10.10.10 80` `weight 0`
Real Server Suspend	`serverfarm host sf1` `real 10.10.10.10 80` `no inservice`
Real Server Suspend and Clear Connections	`serverfarm host sf1` `real 10.10.10.10 80` `no inservice` `clear module contentSwitchingModule 3 connections real 10.10.10.10`

Table 7-3 CLI Commands Deployed from the Real Servers Table (continued)

Command	Sample CLI Sent
Real Server Change Weight	`serverfarm host sf1` `  rserver 10.10.10.10 80` `  weight 2`
CSM Named Real Commands Sent	
Real Server Activation	`serverfarm host sf1` `  real name rs1 80` `  inservice`
Real Server Graceful Suspend	`serverfarm host sf1` `  real name rs1 80` `  weight 0`
Real Server Suspend	`serverfarm host sf1` `  real name rs1 80` `  no inservice`
Real Server Suspend and Clear Connections	`serverfarm host sf1` `  real name rs1 80` `  no inservice` `clear module contentSwitchingModule 3 connections real 10.10.10.10`
Real Server Change Weight	`serverfarm host sf1` `  real name rs1 80` `  weight 2`
CSS Devices	
Real Server Activation	`service myReal7` `  active`
Real Server Graceful Suspend	`service myReal7` `  weight 0`
Real Server Suspend	`service myReal7` `  suspend`
Real Server Suspend and Clear Connections	`service myReal7` `  suspend`
Real Server Change Weight	`service myReal7` `  weight 2`

Server Weight Ranges

Table 7-4 displays the allowable server weight ranges by device type.

Table 7-4 Real Servers Table Server Weight Ranges

Device Type	Valid Weight Configurations
ACE Appliances and Modules	1 to 100
CSMs	0 to 100
CSS Devices	0 to 10

Configuring Server Farms

You can configure load balancing using server farms, which are groups of networked real servers that contain the same content and that typically reside in the same physical location in a data center. Websites often include groups of servers configured in a server farm. Load-balancing software distributes client requests for content or services among the real servers based on the configured policy and traffic classification, server availability and load, and other factors. If one server goes down, another server can take its place and continue to provide the same content to the clients who requested it.

This section includes the following topics:

Configuring Load Balancing Using Server Farms

Procedure

Step 1 Choose **Config > Devices >** ***context*** **> Load Balancing > Server Farms**.

The Server Farms table appears.

Step 2 In the Server Farms table, click **Poll Now** to instruct ANM to poll the devices and display the current values, and click **OK** when prompted if you want to poll the devices for data now.

Step 3 Click **Add** to add a new server farm, or choose an existing server farm and click **Edit**.

The Server Farms configuration window appears.

Step 4 In the Server Farms configuration window, configure the server farm using the information in Table 7-5.

Table 7-5 Server Farm Attributes

Field	Description
Name	Unique name for this server farm or accept the automatically incremented value in this field. Valid entries are unquoted text strings with no spaces and a maximum of 64 characters.
Type	Type of server farm as follows: • **Host**—Server farm consists of real servers that provide content and services to clients. • **Redirect**—Server farm consists only of real servers that redirect client requests to alternate locations specified in the real server configuration. (See "Configuring Real Servers" section on page 7-4.)
Description	Brief description for this server farm. Valid entries are unquoted alphanumeric text strings with no spaces and a maximum of 240 characters.
Fail Action	Action that the ACE is to take with respect to connections if any real server in the server farm fails: • **N/A**—The ACE is to take no action if any server in the server farm fails. • **Purge**—The ACE is to remove connections to a real server if that real server in the server farm fails. The ACE sends a reset command to both the client and the server that failed. • **Reassign**—The ACE is to reassign the existing server connections to the backup real server (if configured) if the real server fails after you enter this command. If a backup real server has not been configured for the failing server, this selection leaves the existing connections untouched in the failing real server.

Table 7-5 Server Farm Attributes (continued)

Field	Description
Failaction Reassign Across Vlans	Option that is available for the ACE module A2(3.0) and later releases only. This field appears only when the Fail Action is set to Reassign. Check the check box to specify that the ACE reassigns the existing server connections to the backup real server on a different VLAN interface (commonly referred to as a bypass VLAN) if the real server fails. If a backup real server has not been configured for the failing server, this option has no effect and leaves the existing connections untouched in the failing real server. Note the following configuration requirements and restrictions when you enable this option: • Enable the Transparent option (see the next Field) to instruct the ACE not to use NAT to translate the ACE VIP address to the server IP address. The Failaction Reassign Across Vlans option is intended for use in stateful firewall load balancing (FWLB) on your ACE, where the destination IP address for the connection coming in to the ACE is for the end-point real server, and the ACE reassigns the connection so that it is transmitted through a different next hop. • Enable the MAC Sticky option on all server-side interfaces to ensure that packets that are going to and coming from the same server in a flow will traverse the same firewalls or stateful devices (see the "Configuring VLAN Interfaces" section on page 11-5). • Configure the Predictor Hash Address option after you add the serverfarm (see the "Configuring the Predictor Method for Server Farms" section on page 7-20). • You must configure identical policies on the primary interface and the backup-server interface. The backup interface must have the same feature configurations as the primary interface. • If you configure a policy on the backup-server interface that is different from the policies on the primary-server interface, that policy will be effective only for new connections. The reassigned connection will always have only the primary-server interface policies. • Interface-specific features (for example, NAT, application protocol inspection, outbound ACLs, or SYN cookie) are not supported. • You cannot reassign connections to the failed real server after it comes back up. This restriction also applies to same-VLAN backup servers. • Real servers must be directly connected to the ACE. This requirement also applies to same-VLAN backup server. • You must disable sequence number randomization on the firewall (see the "Configuring Connection Parameter Maps" section on page 9-3). • Probe configurations should be similar on both ACEs and the interval values should be low. For example, if you configure a high interval value on ACE-1 and a low interval value on ACE-2, the reassigned connections may become stuck because of the probe configuration mismatch. ACE-2 with the low interval value will detect the primary server failure first and will reassign all its incoming connections to the backup-server interface VLAN. ACE-1 with the high interval value may not detect the failure before the primary server comes back up and will still point to the primary server. To minimize packet loss, we recommend the following probe parameter values on both ACEs: Interval: 2, Faildetect: 2, Passdetect interval: 2, and Passdetect count: 5.

Table 7-5 Server Farm Attributes (continued)

Field	Description
Transparent	Field that appears only for host server farms. Specify whether network address translation from the VIP address to the server IP is to occur. Check the check box to indicate that network address translation from the VIP address to the server IP address is to occur. Uncheck the check box to indicate that network address translation from the VIP address to the server IP address is not to occur.
Fail-On-All	Field that appears only for host server farms. By default, real servers that you configure in a server farm inherit the probes that you configure directly on that server farm. When you configure multiple probes on a server farm, the real servers in the server farm use an OR logic with respect to the probes, which means that if one of the probes configured on the server farm fails, all the real servers in that server farm fail and enter the PROBE-FAILED state. With AND logic, if one server farm probe fails, the real servers in the server farm remain in the operational state. If all the probes associated with the server farm fail, then all the real servers in that server farm fail and enter the PROBE-FAILED state. Check this check box to configure the real servers in a server farm to use AND logic with respect to multiple server farm probes. The Fail-On-All function is applicable to all probe types.
Partial-Threshold Percentage	Field that appears only for host server farms. Enter the minimum percentage of real servers in the primary server farm that must remain active for the server farm to stay up. If the percentage of active real servers falls below this threshold, the ACE takes the server farm out of service. Valid entries are from 0 to 99. The default is 0.
Back Inservice	Field that appears only for host server farms. Enter the percentage of real servers in the primary server farm that must be active again for the ACE to place the server farm back into service. Valid entries are from 0 to 99. The value in this field should be larger than the value in the Partial Threshold Percentage field. The default is 0.
Probes	Field that appears only for host server farms. In the Available Items list, choose the probes to use for health monitoring, and click **Add**. The selected probes appear in the Selected Items list. To remove probes that you do not want to use for health monitoring, select them in the Selected Items list, and click **Remove**. The selected probes appear in the Available Items list.

Step 5 Do one of the following:

- Click **Deploy Now** to deploy this configuration on the ACE and save your entries to the running-configuration and startup-configuration files.

  The window refreshes with additional configuration options:

  - To add real servers to the server farm, see the "Adding Real Servers to a Server Farm" section on page 7-18.
  - To specify a predictor method for the server farm, see the "Configuring the Predictor Method for Server Farms" section on page 7-20.
  - To configure return code checking, see the "Configuring Server Farm HTTP Return Error-Code Checking" section on page 7-25.

- Click **Cancel** to exit the procedure without saving your entries and to return to the Server Farms table.

- Click **Next** to deploy your entries and to configure another server farm.

Step 6 (Optional) To display statistics and status information for an existing server farm, choose a server farm from the Server Farms table, and click **Details**.

The **show serverfarm** *name* **detail** CLI command output appears. See the "Displaying Server Farm Statistics and Status Information" section on page 7-28 for details.

Related Topics

- Configuring Health Monitoring for Real Servers, page 7-30
- Configuring Real Servers, page 7-4
- Configuring Sticky Groups, page 8-6
- Configuring the Predictor Method for Server Farms, page 7-20
- Configuring Server Farm HTTP Return Error-Code Checking, page 7-25

Adding Real Servers to a Server Farm

You can add real servers to a server farm. After adding a server farm (see the "Configuring Server Farms" section on page 7-14), you can associate real servers with it and configure predictors and retcode maps. The options for these attributes appear after you have successfully added a new server farm.

Assumptions

This topic assumes the following:

- A server farm has been added to ANM (see the "Configuring Server Farms" section on page 7-14).
- At least one real server exists.

Procedure

Step 1 Choose **Config > Devices >** *context* **> Load Balancing > Server Farms**.

The Server Farms table appears.

Step 2 In the Server Farms table, choose the server farm that you want to associate with real servers.

The Real Servers table appears.

Step 3 In the Real Servers table, click **Add** to add a new entry, or select an existing server and click **Edit** to modify it.

The Real Servers configuration pane appears.

Step 4 In the Real Servers configuration pane, configure the real server using the information in Table 7-6.

Table 7-6 Real Server Configuration Attributes

Field	Description
Name	Server that you want to associate with the server farm.
Port	Port number to be used for server port address translation (PAT). Valid entries are from 1 to 65535.
Backup Server Name	Server that is to act as the backup server for the server farm. Leave this field blank to indicate that there is no designated backup server for the server farm.

Table 7-6 Real Server Configuration Attributes (continued)

Field	Description
Backup Server Port	Server port number. If you select a backup server, enter the backup server port number. Valid entries are from 1 to 65535.
Fail-On-All	Field that appears only for real servers identified as host servers. By default, real servers with multiple probes configured for them have an OR logic associated with them. This means that if one of the real server probes fails, the real server fails and enters the PROBE-FAILED state. Check this checkbox to configure a real server to remain in the OPERATIONAL state unless all probes associated with it fail (AND logic). The Fail-On-All function is applicable to all probe types.
State	State of this server as follows: • **In Servic**e—The server is in service. • **In Service Standby**—The server is a backup server and remains inactive unless the primary server fails. If the primary server fails, the backup server becomes active and starts accepting connections. • **Out Of Service**—The server is out of service.
Min. Connections	Minimum number of connections that the number of connections must fall below before the ACE resumes sending connections to the server after it has exceeded the number in the Max. Connections field. The number in this field must be less than or equal to the number in the Max. Connections field. For ACE appliances, valid entries are from 2 to 4294967295. For ACE modules, valid entries are from 2 to 4000000.
Max. Connections	Maximum number of active connections that can be sent to the server. When the number of connections exceeds this number, the ACE stops sending connections to the server until the number of connections falls below the number specified in the Min. Connections field. For ACE appliances, valid entries are from 2 to 4294967295. For ACE modules, valid entries are from 2 to 4000000.
Weight	Weight to assign to the server. Valid entries are from 1 to 100. The default is 8.
Probes	Probes to apply to the server. Choose the probes in the Available Items list that you want to apply to this server, and click **Add**. The selected probes appear in the Selected Items list. To remove probes that you do not want to use, choose the probes in the Selected Items list, and click **Remove**. The selected probes appear in the Available Items list.
Rate Bandwidth	Bandwidth rate, which is the number of bytes per second and applies to the network traffic exchanged between the ACE and the real server in both directions. Specify the bandwidth limit in bytes per second. Valid entries are from 2 to 300000000. The default is 300000000.
Rate Connection	Connection rate, which is the number of connections per second received by the ACE and applies only to new connections destined to a real server. Specify the limit for connections per second. Valid entries are from 2 to 350000. The default is 350000.

Step 5 When you finish configuring this server for this server farm, do one of the following:

- Click **Deploy Now** to deploy this configuration on the ACE and save your entries to the running-configuration and startup-configuration files. You return to the Real Servers table.
- Click **Cancel** to exit this procedure without saving your entries and to return to the Real Servers table.
- Click **Next** to deploy your entries and to add another real server for this server farm.

Related Topics

- Configuring Health Monitoring for Real Servers, page 7-30
- Configuring Real Servers, page 7-4
- Configuring Sticky Groups, page 8-6
- Configuring the Predictor Method for Server Farms, page 7-20
- Configuring Server Farm HTTP Return Error-Code Checking, page 7-25

Configuring the Predictor Method for Server Farms

You can configure the predictor method for a server farm. The predictor method specifies how the ACE is to select a server in the server farm when it receives a client request for a service. After adding a server farm (see the "Configuring Server Farms" section on page 7-14), you can associate real servers with it and configure the predictor method and retcode maps. The options for these attributes appear after you have successfully added a new server farm.

Note You can configure only one predictor method per server farm.

Assumptions

This topic assumes the following:

- A server farm has been added to ANM (see the "Configuring Server Farms" section on page 7-14.)
- At least one real server exists.

Procedure

Step 1 Choose **Config > Devices >** *context* **> Load Balancing > Server Farms**.

The Server Farms table appears.

Step 2 In the Server Farms table, choose the server farm that you want to configure the predictor method for, and click the **Predictor** tab.

The Predictor configuration pane appears.

Step 3 In the Type field of the Predictor configuration pane, choose the method that the ACE is to use to select a server in this server farm when it receives a client request (see Table 7-7).

Step 4 Enter the required information for the selected predictor method (see Table 7-7).

Table 7-7 Predictor Method Attributes

Predictor Method	Description / Action
Hash Address	Server selection method that uses a hash value based on the source or destination IP address. To configure the hash address predictor method, in the Mask Type field, indicate whether server selection is based on source IP address or the destination IP address as follows: • **N/A**—This option is not defined. • **Destination**—The server is selected based on the destination IP address. • **Source**—The server is selected based on the source IP address. In the IP Netmask field, choose the subnet mask to apply to the address. If none is specified, the default is 255.255.255.255.
Hash Content	Server selection method that uses a hash value based on the specified content string of the HTTP packet body. **a.** In the Begin Pattern field, enter the beginning pattern of the content string and the pattern string to match before hashing. If you do not specify a beginning pattern, the ACE starts parsing the HTTP body immediate following the offset byte. You cannot configure different beginning and ending patterns for different server farms that are part of the same traffic classification. Valid entries are unquoted text strings with no spaces and a maximum of 255 alphanumeric characters. The ACE supports regular expressions for matching string expressions. Table 13-34 lists the supported characters that you can use for matching string expressions. **b.** In the End Pattern field, enter the pattern that marks the end of hashing. If you do not specify either a length or an end pattern, the ACE continues to parse the data until it reaches the end of the field or the end of the packet, or until it reaches the maximum body parse length. You cannot configure different beginning and ending patterns for different server farms that are part of the same traffic classification. Valid entries are unquoted text strings with no spaces and a maximum of 255 alphanumeric characters. The ACE supports regular expressions for matching string expressions. Table 13-34 lists the supported characters that you can use for matching string expressions. **c.** In the Length (Bytes) field, enter the length in bytes of the portion of the content (starting with the byte after the offset value) that the ACE uses for sticking the client to the server. Valid entries are from 1 to 1000 bytes. The offset and length can vary from 0 to 1000 bytes. If the payload is longer than the offset but shorter than the offset plus the length of the payload, the ACE sticks the connection based on that portion of the payload starting with the byte after the offset value and ending with the byte specified by the offset plus the length. The total of the offset and the length cannot exceed 1000. **Note** You cannot specify both the length and the end-pattern options for a Hash Content predictor. **d.** In the HTTP Content Offset (Bytes) field, enter the portion of the content that the ACE uses to stick the client on a particular server by indicating the bytes to ignore starting with the first byte of the payload. Valid entries are integers from 0 to 999 bytes. The default is 0, which indicates that the ACE does not exclude any portion of the content.
Hash Cookie	Server selection method that uses a hash value based on the cookie name. In the Cookie Name field, enter a cookie name in the form of an unquoted text string with no spaces and a maximum of 64 characters.

Table 7-7 Predictor Method Attributes (continued)

Predictor Method	Description / Action
Hash Header	Server selection method that uses a hash value based on the header name. In the Header Name field, choose the HTTP header to be used for server selection as follows: • To specify an HTTP header that is not one of the standard HTTP headers, click the first radio button and enter the HTTP header name in the Header Name field. Valid entries are unquoted text strings with no spaces and a maximum of 64 characters. • To specify one of the standard HTTP headers, click the second radio button, and then choose one of the HTTP headers from the list.
Hash Layer4	Layer 4 generic protocol load-balancing method. Use this predictor to load balance packets from protocols that are not explicitly supported by the ACE. **a.** In the Begin Pattern field, enter the beginning pattern of the Layer 4 payload and the pattern string to match before hashing. If you do not specify a beginning pattern, the ACE starts parsing the HTTP body immediate following the offset byte. You cannot configure different beginning and ending patterns for different server farms that are part of the same traffic classification. Valid entries are unquoted text strings with no spaces and a maximum of 255 alphanumeric characters. The ACE supports regular expressions for matching string expressions. Table 13-34 lists the supported characters that you can use for matching string expressions. **b.** In the End Pattern field, enter the pattern that marks the end of hashing. If you do not specify either a length or an end pattern, the ACE continues to parse the data until it reaches the end of the field or the end of the packet, or until it reaches the maximum body parse length. You cannot configure different beginning and ending patterns for different server farms that are part of the same traffic classification. Valid entries are unquoted text strings with no spaces and a maximum of 255 alphanumeric characters. The ACE supports regular expressions for matching string expressions. Table 13-34 lists the supported characters that you can use for matching string expressions. **c.** In the Length (Bytes) field, enter the length in bytes of the portion of the payload (starting with the byte after the offset value) that the ACE uses for sticking the client to the server. Valid entries are from 1 to 1000 bytes. The offset and length can vary from 0 to 1000 bytes. If the payload is longer than the offset but shorter than the offset plus the length of the payload, the ACE sticks the connection based on that portion of the payload starting with the byte after the offset value and ending with the byte specified by the offset plus the length. The total of the offset and the length cannot exceed 1000. **Note** You cannot specify both the length and end-pattern options for a Hash Layer 4 predictor. **d.** In the HTTP Content Offset (Bytes) field, enter the portion of the content that the ACE uses to stick the client on a particular server by indicating the bytes to ignore starting with the first byte of the payload. Valid entries are from 0 to 999 bytes. The default is 0, which indicates that the ACE does not exclude any portion of the content.

Table 7-7 Predictor Method Attributes (continued)

Predictor Method	Description / Action
Hash URL	Server selection method that uses a hash value based on the URL. Use this method to load balance firewalls. Enter values in one or both of the pattern fields as follows: • In the URL Begin Pattern field, enter the beginning pattern of the URL and the pattern string to parse. • In the URL End Pattern field, enter the ending pattern of the URL and the pattern string to parse. Valid entries for these fields are unquoted text strings with no spaces and a maximum of 255 alphanumeric characters for each pattern that you configure.
Least Bandwidth	Server with the least amount of network traffic over a specified sampling period. **a.** In the Assess Time (Seconds) field, enter the number of seconds for which the ACE is to collect traffic information. Valid entries are from 1 to 10 seconds. **b.** In the Least Bandwidth Samples field, enter the number of samples over which you want to weight and average the results of the probe query to calculate the final load value. Valid entries are 1, 2, 4, 8, and 16 (values from 1 to 16 that are also a power of 2).
Least Connections	Server with the fewest number of connections. In the Slow Start Duration (Seconds) field, enter the slow-start value to be applied to this predictor method. Valid entries are from 1 to 65535, where 1 is the slowest ramp-up value. The slow-start mechanism is used to avoid sending a high rate of new connections to servers that you have just put into service.

Table 7-7 Predictor Method Attributes (continued)

Predictor Method	Description / Action
Least Loaded	Least loaded server based on information from SNMP probes. **a.** In the SNMP Probe Name field, choose the name of the SNMP probe to use. **b.** In the Auto Adjust field, configure the autoadjust feature to assign a maximum load value of 16000 to that server to prevent it from being flooded with new incoming connections. The ACE periodically adjusts this load value based on feedback from the server's SNMP probe and other configured options. Options include the following: – **N/A**—Indicates that this option is not defined. – **Average**—Instructs the ACE to apply the average load of the server farm to a real server whose load reaches zero. The average load is the running average of the load values across all real servers in the server farm. – **Off**—Overrides the default behavior of the ACE of setting the load value for a server with a load of zero to 16000. When you configure this parameter, the ACE sends all new connections to the server that has a load of zero until the next load update arrives from the SNMP probe for this server. There may be times when you want the ACE to send all new connections to a real server whose load is zero. **c.** In the Weight Connection field, check the check box to instruct the ACE to use the current connection count in the final load calculation for a real server. When you configure this option, the ACE includes the current connection count in the total load calculation for each real server in a server farm. Uncheck the check box to reset the behavior of the ACE to the default of excluding the current connection count from the load calculation. To instruct the ACE to select the server with the lowest load, use the predictor least-loaded command in server farm host or redirect configuration mode. With this predictor, the ACE uses SNMP probes to query the real servers for load parameter values (for example, CPU utilization or memory utilization). This predictor is considered adaptive because the ACE continuously provides feedback to the load-balancing algorithm based on the behavior of the real server. To use this predictor, you must associate an SNMP probe with it. The ACE queries user-specified OIDs periodically based on a configurable time interval. The ACE uses the retrieved SNMP load value to determine the server with the lowest load. The syntax of this predictor command is as follows: **predictor least-loaded probe** *name* The *name* argument specifies the identifier of the existing SNMP probe that you want the ACE to use to query the server. Enter an unquoted text string with no spaces and a maximum of 64 alphanumeric characters. For example, to configure the ACE to select the real server with the lowest load based on feedback from an SNMP probe called PROBE_SNMP, enter: `host1/Admin(config)# serverfarm SF1` `host1/Admin(config-sfarm-host)# predictor leas` `t-loaded probe PROBE_SNMP` `host1/Admin(config-sfarm-host-predictor)#` To reset the predictor method to the default of round-robin, enter: `host1/Admin(config-sfarm-host)# no predictor`

Table 7-7 Predictor Method Attributes (continued)

Predictor Method	Description / Action
Response	Server selection method based on the lowest response time for a requested response-time measurement. a. In the Response Type field, select the type of measurement to use as follows: – **App-Req-To-Resp**—The response time from when the ACE sends an HTTP request to a server to the time that the ACE receives a response from the server for that request. – **Syn-To-Close**—The response time from when the ACE sends a TCP SYN to a server to the time that the ACE receives a CLOSE from the server. – **Syn-To-Synack**—The response time from when the ACE sends a TCP SYN to a server to the time that the ACE receives a SYN-ACK from the server. b. In the Response Samples field, enter the number of samples over which you want to average the results of the response-time measurement. Valid entries are 1, 2, 4, 8, and 16 (values from 1 to 16 that are also a power of 2). c. In the Weight Connection field, check the check box to instruct the ACE to use the current connection count in the final load calculation for a real server. When you configure this option, the ACE includes the current connection count in the total load calculation for each real server in a server farm. Uncheck the check box to reset the behavior of the ACE to the default of excluding the current connection count from the load calculation.
Round Robin	Server selection method in which The ACE selects the next server in the list of servers based on server weight. This method is the default predictor.

Step 5 Click **Deploy Now** to deploy this configuration on the ACE and save your entries to the running-configuration and startup-configuration files.

Related Topics

- Configuring Health Monitoring for Real Servers, page 7-30
- Configuring Real Servers, page 7-4
- Configuring Sticky Groups, page 8-6
- Adding Real Servers to a Server Farm, page 7-18
- Configuring Server Farm HTTP Return Error-Code Checking, page 7-25

Configuring Server Farm HTTP Return Error-Code Checking

Note This feature is available only for server farms configured as hosts. It is not available for server farms configured with the type Redirect.

You can configure HTTP return error-code checking (retcode map) for a server farm. After adding a server farm (see the "Configuring Server Farms" section on page 7-14), you can associate real servers with it and configure the predictor method and retcode maps. These options appear after you have successfully added a server farm.

Assumption

A host type server farm has been added to ANM (sSee the "Configuring Server Farms" section on page 7-14).

Procedure

Step 1 Choose **Config > Devices >** ***context*** **> Load Balancing > Server Farms**.

The Server Farms table appears.

Step 2 In the Server Farms table, choose the server farm that you want to configure for return error-code checking, and click the **Retcode Map** tab.

The Retcode Map table appears.

Step 3 In the Retcode Map table, click **Add** to add a new entry to the table.

The Retcode Map configuration pane appears.

Note You cannot modify an entry in the Retcode Map table. Instead, delete the existing entry, then add a new one.

Step 4 In the Lowest Retcode field of the Retcode Map configuration pane, enter the minimum value for an HTTP return error code.

Valid entries are from 100 to 599. This number must be less than or equal to the number in the Highest Retcode field.

Step 5 In the Highest Retcode field, enter the maximum number for an HTTP return error code.

Valid entries are from 100 to 599. This number must be greater than or equal to the number in the Lowest Retcode field.

Step 6 In the Type field, specify the action to be taken and related options using the information in Table 7-8.

Note For ACE appliances, the only available option is Count.

Table 7-8 Return-Code Type Configuration Options

Option	Description
Count	Total number of return codes received for each return code number that you specify.
Log	Syslog error message generated when the number of events reaches a specified threshold. a. In the Threshold field, enter the number of events that the ACE is to receive before generating a syslog error message. Valid entries are from 1 to 4294967295. b. In the Reset (Seconds) field, enter the time interval in seconds for which the ACE checks for the return code. Valid entries are from 1 to 4294967295 seconds.

Table 7-8 Return-Code Type Configuration Options (continued)

Option	Description
Remove	The ACE generates a syslog error message when the number of events reaches a specified threshold and then removes the server from service. **a.** In the Threshold field, enter the number of events that the ACE is to receive before generating a syslog error message and removing the server from service. Valid entries are from 1 to 4294967295. **b.** In the Reset (Seconds) field, enter the time interval in seconds for which the ACE checks for the return code. Valid entries are from 1 to 4294967295 seconds. **c.** In the Resume Service (Seconds) field, enter the number of seconds that the ACE waits before it resumes service for the real server automatically after taking the real server out of service. Valid entries are 30 to 3600 seconds. The default is 300 seconds.

Step 7 Do one of the following:

- Click **Deploy Now** to deploy this configuration on the ACE and save your entries to the running-configuration and startup-configuration files.
- Click **Cancel** to exit this procedure without saving your entries and to return to the Retcode Map table.
- Click **Next** to deploy your entries and to add another retcode map.

Related Topics

- Information About Virtual Contexts, page 5-2
- Configuring Virtual Context Class Maps, page 13-6
- Configuring Virtual Context Policy Maps, page 13-31
- Configuring Real Servers, page 7-4
- Configuring Sticky Groups, page 8-6

Displaying All Server Farms

You can display all server farms associated with a virtual context.

Procedure

Step 1 Choose **Config > Devices**.

The Virtual Contexts table appears.

Step 2 In the Virtual Contexts table, choose the virtual context with the server farms you want to display, and click **Load Balancing > Server Farms**.

The Server Farms table appears with the following information:

- Server farm name
- Server farm type (either host or redirect)
- Description

- Number of real servers associated with the server farm
- Number of predictor methods for the server farm
- Number of entries in the HTTP retcode map table

You can click on any of the entries in the last three columns to view specific information about those entries.

Related Topics

- Displaying Server Farm Statistics and Status Information, page 7-28
- Configuring Server Farms, page 7-14
- Adding Real Servers to a Server Farm, page 7-18
- Configuring the Predictor Method for Server Farms, page 7-20
- Configuring Server Farm HTTP Return Error-Code Checking, page 7-25

Displaying Server Farm Statistics and Status Information

You can display statistics and status information for a particular server farm.

Procedure

Step 1 Choose **Config > Devices >** ***context*** **> Load Balancing > Server Farms**.

The Server Farms table appears.

Step 2 In the Server Farms table, choose a server farm from the Server Farms table, and click **Details**.

The **show serverfarm** *name* **detail** CLI command output appears. For details about the displayed output fields, see the *Cisco ACE Module Server Load-Balancing Configuration Guide* or the *Cisco ACE 4700 Series Appliance Server Load-Balancing Configuration Guide*, Chapter 2, Configuring Real Servers and Server Farms.

Step 3 Click **Update Details** to refresh the output for the **show serverfarm** *name* **detail** CLI command.

The new information appears in a separate panel with a new timestamp; both the old and the new server farm statistics and status information appear side-by-side to avoid overwriting the last updated information.

Step 4 Click **Close** to return to the Server Farms table.

Related Topics

- Displaying All Server Farms, page 7-27
- Configuring Server Farms, page 7-14
- Adding Real Servers to a Server Farm, page 7-18
- Configuring the Predictor Method for Server Farms, page 7-20
- Configuring Server Farm HTTP Return Error-Code Checking, page 7-25

Configuring 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 (see the "TCL Scripts" section on page 7-29).

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, it can make reliable load-balancing decisions.

Health monitoring on the ACE tracks the state of a server by sending out probes. Also referred to as out-of-band health monitoring, the ACE verifies the server response or checks for any network problems that can prevent a client to reach a server. Based on the server response, the ACE can place the server in or out of service, and can make reliable load-balancing decisions.

The ACE identifies the health of a server in the following categories:

- Passed—The server returns a valid response.
- Failed—The server fails to provide a valid response to the ACE is unable to reach a server for a specified number of retries.

By configuring the ACE for health monitoring, the ACE sends active probes periodically to determine the server state.

The ACE supports 4000 unique probe configurations which includes ICMP, TCP, HTTP, and other predefined health probes. The ACE also allows the opening of 1000 sockets simultaneously.

This section includes the following topics:

TCL Scripts

The ACE supports several specific types of health probes (for example HTTP, TCP, or ICMP health probes) when you need to use a diverse set of applications and health probes to administer your network. The basic health probe types supported in the current ACE software release may not support the specific probing behavior that your network requires. To support a more flexible health-probing functionality, the ACE allows you to upload and execute Toolkit Command Language (TCL) scripts on the ACE.

The TCL interpreter code in the ACE is based on Release 8.44 of the standard TCL distribution. You can create a script to configure health probes. Script probes operate similar to other health probes available in the ACE software. As part of a script probe, the ACE executes the script periodically, and the exit code

that is returned by the executing script indicates the relative health and availability of specific real servers. For information on health probes, see the "Configuring Health Monitoring for Real Servers" section on page 7-30.

For your convenience, the following sample scripts for the ACE are available to support the TCL feature and are supported by Cisco TAC:

- CHECKPORT_STD_SCRIPT
- ECHO_PROBE_SCRIPT
- FINGER_PROBE_SCRIPT
- FTP_PROBE_SCRIPT
- HTTP_PROBE_SCRIPT
- HTTPCONTENT_PROBE
- HTTPHEADER_PROBE
- HTTPPROXY_PROBE
- IMAP_PROBE
- LDAP_PROBE
- MAIL_PROBE
- POP3_PROBE
- PROBENOTICE_PROBE
- RTSP_PROBE
- SSL_PROBE_SCRIPT
- TFTP_PROBE

These scripts are located in the probe: directory and are accessible in both the Admin and user contexts. Note that the script files in the probe: directory are read-only, so you cannot copy or modify them. However, you can copy files from the probe: directory. For more information, see either the *Cisco Application Control Engine Module Administration Guide* or the *Cisco 4700 Series Application Control Engine Appliance Administration Guide*.

To load a script into memory on the ACE and enable it for use, use the script file command. For detailed information on uploading and executing TCL scripts on the ACE, see either the *Cisco ACE Module Server Load-Balancing Configuration Guide* or the *Cisco ACE 4700 Series Appliance Server Load-Balancing Configuration Guide*.

Configuring Health Monitoring for Real Servers

You can establish monitoring of real servers to determine their viability in load-balancing decisions. To check the health and availability of a real server, the ACE periodically sends a probe to the real server. Depending on the server response, the ACE determines whether or not to include the server in its load-balancing decision.

Procedure

Step 1 Choose **Config > Devices >** ***context*** **> Load Balancing > Health Monitoring**.

The Health Monitoring table appears.

Step 2 In the Health Monitoring table, click **Add** to add a new health monitoring probe, or choose an existing entry and click **Edit** to modify it.

The Health Monitoring window appears.

Step 3 In the Name field of the Health Monitoring window, enter a name that identifies the probe and that associates the probe with the real server.

Valid entries are text strings with a maximum of 64 characters.

Step 4 In the Type field, choose the type of probe that you want to use.

The probe type determines what the probe sends to the real server. See Table 7-9 for the types of probes and their descriptions.

Table 7-9 Probe Types

Probe Type	Description
DNS	Sends a request to a DNS server giving it a configured domain. To determine if the server is up, the ACE must receive the configured IP address for that domain.
ECHO-TCP	Sends a string to the server and compares the response with the original string. If the response string matches the original, the server is marked as passed. If not, the ACE retries as configured before the server is marked as failed.
ECHO-UDP	Sends a string to the server and compares the response with the original string. If the response string matches the original, the server is marked as passed. If not, the ACE retries as configured before the server is marked as failed.
FINGER	Sends a probe to the server to verify that a defined username is a username on the server.
FTP	Initiates an FTP session. By default, this probe is for an anonymous login with the option of configuring a user ID and password. The ACE performs an FTP GET or LS to determine the outcome of the problem. This probe supports only active connections.
HTTP	Sets up a TCP connection and issues an HTTP request. Any valid HTTP response causes the probe to mark the real server as passed.
HTTPS	Similar to an HTTP probe, but this probe uses SSL to generate encrypted data.
ICMP	Sends an ICMP request and listens for a response. If the server returns a response, the ACE marks the real server as passed. If there is no response and times out, or an ICMP standard error occurs, such as DESTINATION_UNREACHABLE, the ACE marks the real server as failed.
IMAP	Initiates an IMAP session, using a configured user ID and password. Then, the probe attempts to retrieve e-mail from the server and validates the result of the probe based on the return codes received from the server.
POP	Initiates a POP session, using a configured user ID and password. Then, the probe attempts to retrieve e-mail from the server and validates the result of the probe based on the return codes received from the server.
RADIUS	Connects to a RADIUS server and logs into it to determine if the server is up.
RTSP	Establishes a TCP connection and sends a request packet to the server. The ACE compares the response with the configured response code to determine whether the probe succeeded.
Scripted	Executes probes from a configured script to perform health probing. This method allows you to author specific scripts with features not present in standard probes. For ACE appliances, the script probe file name must first be established on the device.

Table 7-9 Probe Types (continued)

Probe Type	Description
SIP-TCP	Establishes a TCP connection and sends an OPTIONS request packet to the user agent on the server. The ACE compares the response with the configured response code or expected string, or both, to determine whether the probe has succeeded. If you do not configure an expected status code, any response from the server is marked as failed.
SIP-UDP	Establishes a UDP connection and sends an OPTIONS request packet to the user agent on the server. The ACE compares the response with the configured response code or expected string, or both, to determine whether the probe has succeeded. If you do not configure an expected status code, any response from the server is marked as failed.
SMTP	Initiates an SMTP session by logging into the server.
SNMP	Establishes a UDP connection and sends a maximum of eight SMNP OID queries to probe the server. The ACE weighs and averages the load information that is retrieved and uses it as input to the least-loaded algorithm for load-balancing decisions. If the retrieved value is within the configured threshold, the server is marked as passed. If the threshold is exceeded, the server is marked as failed.
TCP	Initiates a TCP handshake and expects a response. By default, a successful response causes the probe to mark the server as passed. The probe then sends a FIN to end the session. If the response is not valid, or if there is no response, the probe marks the real server as failed.
TELNET	Establishes a connection to the real server and verifies that a greeting from the application was received.
UDP	Sends a UDP packet to a real server. The probe marks the server as failed only if an ICMP Port Unreachable messages is returned.

Step 5 Enter health monitoring general attributes (see Table 7-10).

Note Click **More Settings** to access the additional general attributes for the selected probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-10 Health Monitoring General Attributes

Field	Action
Description	Description for this probe. Valid entries are unquoted alphanumeric text strings with no spaces and a maximum of 240 characters.
Probe Interval (Seconds)	Number of seconds that the ACE is to wait before sending another probe to a server marked as passed. Valid entries are from 2 to 65535. The default is 15.
Pass Detect Interval (Seconds)	Number of seconds that the ACE is to wait before sending another probe to a server marked as failed. Valid entries are from 2 to 65535. The default is 60.
Fail Detect	Consecutive number of times that an ACE must detect that probes have failed to contact a server before marking the server as failed. Valid entries are from 1 to 65535. The default is 3.
More Settings	
Pass Detect Count	Number of successful probe responses from the server before the server is marked as passed. Valid entries are from 1 to 65535. The default is 3.
Receive Timeout (Seconds)	Number of seconds the ACE is to wait for a response from a server that has been probed before marking the server as failed. Valid entries are from 1 to 65535. The default is 10.

Table 7-10 Health Monitoring General Attributes (continued)

Field	Action
Destination IP Address[1]	Preferred destination IP address. By default, the probe uses the IP address from the real or virtual server configuration for the destination IP address. To override the destination address that the probe uses, enter the preferred destination IP address in this field using dotted-decimal notation, such as 192.168.11.1.
Is Routed [2]	Check box that indicates that the destination IP address is routed according to the ACE internal routing table. Uncheck the check box to indicate that the destination IP address is not routed according to the ACE internal routing table.

1. The Dest IP Address field is not applicable to the Scripted probe type.
2. The Is Routed field is not applicable to the RTSP, Scripted, SIP-TCP, and SIP-UDP probe types.

Table 7-11 lists the default port numbers for each probe type.

Table 7-11 Default Port Numbers for Probe Types

Probe Type	Default Port Number
DNS	53
Echo	7
Finger	79
FTP	21
HTTP	80
HTTPS	443
ICMP	Not applicable
IMAP	143
POP3	110
RADIUS	1812
RTSP	554
Scripted	1
SIP (both TCP and UDP)	5060
SMTP	25
SNMP	161
Telnet	23
TCP	80
UDP	53

Step 6 Enter the attributes for the specific probe type selected as follows:

- For DNS probes, see Table 7-12.
- For Echo-TCP probes, see Table 7-13.
- For Echo-UDP probes, see Table 7-14.
- For Finger probes, see Table 7-15.
- For FTP probes, see Table 7-16.

- For HTTP probes, see Table 7-18.
- For HTTPS probes, see Table 7-18.
- There are no specific attributes for ICMP probes.
- For IMAP probes, see Table 7-19.
- For POP probes, see Table 7-20.
- For RADIUS probes, see Table 7-21.
- For RTSP probes, see Table 7-22.
- For Scripted probes, see Table 7-23.
- For SIP-TCP probes, see Table 7-24.
- For SIP-UDP probes, see Table 7-25.
- For SMTP probes, see Table 7-26.
- For SNMP probes, see Table 7-27.
- For TCP probes, see Table 7-28.
- For Telnet probes, see Table 7-29.
- For UDP probes, see Table 7-30.

Step 7 Do one of the following:

- Click **Deploy Now** to deploy this configuration on the ACE and save your entries to the running-configuration and startup-configuration files.
- Click **Cancel** to exit this procedure without saving your entries and to return to the Health Monitoring table.
- Click **Next** to deploy your entries and to configure another probe.

Step 8 (Optional) To display statistics and status information for a particular probe, choose the probe from the Health Monitoring table, and click **Details**.

The **show probe** *name* **detail** CLI command output appears. See the "Displaying Health Monitoring Statistics and Status Information" section on page 7-53 for details.

Related Topics

- Configuring DNS Probe Expect Addresses, page 7-49
- Configuring Headers for HTTP and HTTPS Probes, page 7-50
- Configuring Health Monitoring Expect Status, page 7-51
- Displaying Health Monitoring Statistics and Status Information, page 7-53
- Configuring Real Servers, page 7-4
- Configuring Server Farms, page 7-14
- Configuring Sticky Groups, page 8-6

Configuring Probe Attributes

You can configure health monitoring probe-specific attributes.

This section includes the following topics:

Refer to the following topics for additional configuration options for health-monitoring probes:

DNS Probe Attributes

Table 7-12 lists the DNS probe attributes.

Note Click **More Settings** to access the additional attributes for the DNS probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-12 DNS Probe Attributes

Field	Action
Domain Name	Domain name that the probe is to send to the DNS server. Valid entries are unquoted text strings with a maximum of 255 characters.
More Settings	
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.

To configure expect addresses for DNS probes, see the "Configuring DNS Probe Expect Addresses" section on page 7-49.

Echo-TCP Probe Attributes

Table 7-13 lists the Echo-TCP probe attributes.

Note Click **More Settings** to access the additional attributes for the Echo-TCP probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-13 Echo-TCP Probe Attributes

Field	Action
Send Data	ASCII data that the probe is to send to the server. Valid entries are unquoted text strings with no spaces and a maximum of 255 characters.
More Settings	
Is Connection	Check box that indicates that the connection parameters are configured. Uncheck the check box to indicate that connection parameters are not configured.
Open Timeout (Seconds)	Number of seconds to wait when opening a connection with a real server. Valid entries are integers from 1 to 65535. The default is 10.

Echo-UDP Probe Attributes

Table 7-14 lists the Echo-UDP probe attributes.

Note Click **More Settings** to access the additional attributes for the Echo-UDP probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-14 Echo-UDP Probe Attributes

Field	Action
Send Data	ASCII data that the probe is to send to the server. Valid entries are unquoted text strings with no spaces and a maximum of 255 characters.
More Settings	
Port	Number that the probe is to use. By default, the probe uses the port number based on its type.

Finger Probe Attributes

Table 7-15 lists the Finger probe attributes.

Note Click **More Settings** to access the additional attributes for the Finger probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-15 Finger Probe Attributes

Field	Action
Send Data	ASCII data that the probe is to send to the server. Valid entries are unquoted text strings with no spaces and a maximum of 255 characters.
More Settings	
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
Is Connection	Check box that indicates that the connection parameters are configured. Uncheck the check box to indicate that connection parameters are not configured.
Open Timeout (Seconds)	Number of seconds to wait when opening a connection with a real server. Valid entries are from 1 to 65535. The default is 10.

FTP Probe Attributes

Table 7-16 lists the FTP probe attributes.

Note Click **More Settings** to access the additional attributes for the FTP probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-16 FTP Probe Attributes

Field	Action
More Settings	
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
Is Connection	Check box that indicates that the connection parameters are configured. Uncheck the check box to indicate that connection parameters are not configured.
Open Timeout (Seconds)	Number of seconds to wait when opening a connection with a real server. Valid entries are integers from 1 to 65535. The default is 10.

To configure probe expect statuses for FTP probes, see the "Configuring Health Monitoring Expect Status" section on page 7-51.

HTTP Probe Attributes

Table 7-17 lists the HTTP probe attributes.

Note Click **More Settings** to access the additional attributes for the HTTP probe type. By default, ANM hides the probe attributes with default values and the probe attributes which are not commonly used.

Table 7-17 HTTP Probe Attributes

Field	Action
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
Request Method Type	Type of HTTP request method that is to be used for this probe. Choose one of the following: • **N/A**—This option is not defined. • **Get**—The HTTP request method is a GET with a URL of "/". This request method directs the server to get the page, and the ACE calculates a hash value for the content of the page. If the page content information changes, the hash value no longer matches the original hash value and the ACE assumes the service is down. This is the default request method. • **Head**—The server is to only get the header for the page. Using this method can prevent the ACE from assuming that the service is down due to changed content and therefore changed hash values.
Request HTTP URL	Field that appears if you chose Head or Get in the Request Method Type field. Enter the URL path on the remote server. Valid entries are strings of up to 255 characters specifying the URL path. The default path is "/".
More Settings	
Is Connection	Check box to indicate that connection parameters are configured. Uncheck the check box to indicate that connection parameters are not configured.
Open Timeout (Seconds)	Number of seconds to wait when opening a connection with a real server. Valid entries are from 1 to 65535. The default is 10.
User Name	User identifier to be used for authentication on the real server. Valid entries are unquoted text strings with a maximum of 64 characters.
Password	Password to be used for authentication on the real server. Valid entries are unquoted text strings with a maximum of 64 characters. Reenter the password in the Confirm field.
Expect Regular Expression	Expected response data from the probe destination. Valid entries are text strings (quotes allowed) with a maximum of 255 characters.
Expect Regex Offset	Number of characters into the received message or buffer where the ACE is to begin looking for the string specified in the Expect Regular Expression field. Valid entries are from 1 to 4000.
Hash	Check box that indicates that the ACE is to use an MD5 hash for an HTTP GET probe. Uncheck the check box to indicate that the ACE should not use an MD5 hash for an HTTP GET probe.
Hash String	Field that appears if the Hash check box is selected. Enter the 32-bit hash value that the ACE is to compare with the hash that is generated from the HTTP page sent by the server. If you do not provide this value, the ACE generates a value the first time it queries the server, stores this value, and matches this value with other responses from the server. A successful comparison causes the probe to maintain an Alive state. Enter the MD5 hash value as a quoted or unquoted hexadecimal string with 16 characters.

To configure probe headers and expect statuses for HTTP probes, see:

- Configuring Headers for HTTP and HTTPS Probes, page 7-50
- Configuring Health Monitoring Expect Status, page 7-51

HTTPS Probe Attributes

Table 7-18 lists the HTTPS probe attributes.

Note Click **More Settings** to access the additional attributes for the HTTPS probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-18 HTTPS Probe Attributes

Field	Action
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
Request Method Type	Type of HTTP request method that is to be used for this probe. Choose one of the following: • **N/A**—This option is not defined. • **Get**—The HTTP request method is a GET with a URL of "/". This request method directs the server to get the page, and the ACE calculates a hash value for the content of the page. If the page content information changes, the hash value no longer matches the original hash value and the ACE assumes the service is down. This is the default request method. • **Head**—The server is to only get the header for the page. Using this method can prevent the ACE from assuming that the service is down due to changed content and as a result changed hash values.
Request HTTP URL	Field that appears if you chose Head or Get in the Request Method Type field. Enter the URL path on the remote server. Valid entries are strings of up to 255 characters specifying the URL path. The default path is "/".

Table 7-18 HTTPS Probe Attributes (continued)

Field	Action
Cipher	Choose the cipher suite to be used with this HTTPS probe: • RSA_ANY—The HTTPS probe accepts all RSA-configured cipher suites and that no specific suite is configured. This is the default action. • RSA_EXPORT1024_WITH_DES_CBC_SHA • RSA_EXPORT1024_WITH_RC4_56_MD5 • RSA_EXPORT1024_WITH_RC4_56_SHA • RSA_EXPORT_WITH_DES40_CBC_SHA • RSA_EXPORT_WITH_RC4_40_MD5 • RSA_WITH_3DES_EDE_CBC_SHA • RSA_WITH_AES_128_CBC_SHA • RSA_WITH_AES_256_CBC_SHA • RSA_WITH_DES_CBC_SHA • RSA_WITH_RC4_128_MD5 • RSA_WITH_RC4_128_SHA
SSL Version	Version of SSL or TLS to be used in ClientHello messages sent to the server as follows: • **All**—The probe is to use all SSL versions. • **SSLv3**—The probe is to use SSL version 3. • **TLSv1**—The probe is to use TLS version 1. By default, the probe sends ClientHello messages with an SSL version 3 header and a TLS version 1 message.
More Settings	
Is Connection	Check box that indicates that the connection parameters are configured. Uncheck the check box to indicate that connection parameters are not configured.
Open Timeout (Seconds)	Number of seconds to wait when opening a connection with a real server. Valid entries are from 1 to 65535. The default is 10.
User Name	User identifier to be used for authentication on the real server. Valid entries are unquoted text strings with a maximum of 64 characters.
Password	Password to be used for authentication on the real server. Valid entries are unquoted text strings with a maximum of 64 characters. Reenter the password in the Confirm field.
Expect Regular Expression	Expected response data from the probe destination. Valid entries are text strings (quotes allowed) with a maximum of 255 characters.
Expect Regex Offset	Number of characters into the received message or buffer where the ACE is to begin looking for the string specified in the Expect Regular Expression field. Value entries are from 1 to 4000.

Table 7-18 HTTPS Probe Attributes (continued)

Field	Action
Hash	Check box that indicates that the ACE is to use an MD5 hash for an HTTP GET probe. Uncheck this check box to indicate that the ACE is not to use an MD5 hash for an HTTP GET probe.
Hash String	Field that appears if the Hash check box is selected. Enter the 32-bit hash value that the ACE is to compare with the hash that is generated from the HTTP page sent by the server. If you do not provide this value, the ACE generates a value the first time it queries the server, stores this value, and matches this value with other responses from the server. A successful comparison causes the probe to maintain an Alive state. Enter the MD5 hash value as a quoted or unquoted hexadecimal string with 16 characters.

To configure probe headers and expect statuses for HTTPS probes, see:

- Configuring Headers for HTTP and HTTPS Probes, page 7-50
- Configuring Health Monitoring Expect Status, page 7-51

IMAP Probe Attributes

Table 7-19 lists the IMAPprobe attributes.

Note Click **More Settings** to access the additional attributes for the IMAP probe type. By default, ANM hides the probe attributes with default values and the probe attributes are not commonly used.

Table 7-19 IMAP Probe Attributes

Field	Action
User Name	User identifier to be used for authentication on the real server. Valid entries are unquoted text strings with a maximum of 64 characters.
Password	Password to be used for authentication on the real server. Valid entries are unquoted text strings with a maximum of 64 characters. Reenter the password in the Confirm field.
Mailbox Name	User mailbox name from which to retrieve e-mail for this IMAP probe. Valid entries are unquoted text strings with a maximum of 64 characters.
Request Command	Request method command for this probe. Valid entries are text strings with a maximum of 32 characters and no spaces.
More Settings	
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
Is Connection	Check box that indicates that the connection parameters are configured. Uncheck the check box to indicate that connection parameters are not configured.
Open Timeout (Seconds)	Number of seconds to wait when opening a connection with a real server. Valid entries are from 1 to 65535. The default is 10.

POP Probe Attributes

Table 7-20 lists the POP probe attributes.

Note Click **More Settings** to access the additional attributes for the POP probe type. By default, ANM hides the probe attributes with default values and the probe attributes which are not commonly used.

Table 7-20 POP Probe Attributes

Field	Action
User Name	User identifier to be used for authentication on the real server. Valid entries are unquoted text strings with a maximum of 64 characters.
Password	Password to be used for authentication on the real server. Valid entries are unquoted text strings with a maximum of 64 characters. Reenter the password in the Confirm field.
Request Command	Request method command for this probe. Valid entries are text strings with a maximum of 32 characters and no spaces.

Table 7-20 POP Probe Attributes (continued)

Field	Action
More Settings	
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
Is Connection	Check box that indicates that the connection parameters are configured. Uncheck the check box to indicate that connection parameters are not configured.
Open Timeout	Number of seconds to wait when opening a connection with a real server. Valid entries are from 1 to 65535. The default is 10.

RADIUS Probe Attributes

Table 7-21 lists the RADIUS probe attributes.

Note Click **More Settings** to access the additional attributes for the RADIUS probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-21 RADIUS Probe Attributes

Field	Action
User Secret	Shared secret to be used to allow probe access to the RADIUS server. Valid entries are case-sensitive strings with no spaces and a maximum of 64 characters.
User Name	User identifier to be used for authentication on the real server. Valid entries are unquoted text strings with a maximum of 64 characters.
Password	Password to be used for authentication on the real server. Valid entries are unquoted text strings with a maximum of 64 characters. Reenter the password in the Confirm field.
More Settings	
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
NAS IP Address	IP address of the Network Access Server (NAS) in dotted-decimal format, such as 192.168.11.1.

RTSP Probe Attributes

Table 7-22 lists the RTSP probe attributes.

Note Click **More Settings** to access the additional attributes for the RTSP probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-22 RTSP Probe Attributes

Field	Action
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
RTSP Require Header Value	Require header for the probe.

Table 7-22 RTSP Probe Attributes (continued)

Field	Action
RTSP Proxy Require Header Value	Proxy-Require header for the probe.
RTSP Request Method Type	Request method type: • **N/A**—No request method is selected. • **Describe**—Probe is to use the Describe request type.
More Settings	
Is Connection	Check box that indicates that the connection parameters are configured. Uncheck the check box to indicate that connection parameters are not configured.
Open Timeout (Seconds)	Number of seconds to wait when opening a connection with a real server. Valid entries are from 1 to 65535. The default is 10.

To configure probe expect statuses for RTSP probes, see the "Configuring Health Monitoring Expect Status" section on page 7-51.

Scripted Probe Attributes

Table 7-23 lists the HTTP probe attributes.

Note Click **More Settings** to access the additional attributes for the Scripted probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-23 Scripted Probe Attributes

Field	Action
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
Script Name	Local name that you want to assign to this file on the ACE. This file can reside in the disk0: directory or the probe: directory (if the probe: directory exists). **Note** The script file must first be established on the ACE device and the name must be entered exactly as is appears on the device. See your ACE documentation for more details. Valid entries are unquoted text strings with no spaces and a maximum of 255 characters.
Script Arguments	Valid arguments, which are unquoted text strings with no spaces; separate multiple arguments with a space. The field limit is 255 characters.
More Settings	
Script Needs To Be Copied From Remote Location?	Check box that indicates that the file needs to be copied from a remote server. Uncheck this check box to indicate that the script resides locally.

Table 7-23 Scripted Probe Attributes (continued)

Field	Action
Protocol	Field that appears if the script is to be copied from a remote server. Choose the protocol to be used for copying the script: • **FTP**—The script is to be copied using FTP. • **TFTP**—The script is to be copied using TFTP.
User Name	Field that appears if FTP is selected in the Protocol field. Enter the name of the user account on the remote server.
Password	Field that appears if FTP is selected in the Protocol field. Enter the password for the user account on the remote server. Reenter the password in the Confirm field.
Source File Name	Field appears if the script is to be copied from a remote server. Enter the host IP address, path, and filename of the file on the remote server in the format *host-ip/path/filename* where: • *host-ip* represents the IP address of the remote server. • *path* represents the directory path of the file on the remote server. • *filename* represents the filename of the file on the remote server. For example, your entry might be 192.168.11.2/usr/bin/my-script.ext.

SIP-TCP Probe Attributes

Table 7-24 lists the SIP-TCP probe attributes.

Note Click **More Settings** to access the additional attributes for the SIP-TCP probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-24 SIP-TCP Probe Attributes

Field	Action
More Settings	
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
Is Connection	Check box that indicates that the connection parameters are configured. Uncheck the check box to indicate that connection parameters are not configured.
Open Timeout (Seconds)	Number of seconds to wait when opening a connection with a real server. Valid entries are from 1 to 65535. The default is 10.

Table 7-24 SIP-TCP Probe Attributes (continued)

Field	Action
Expect Regular Expression	Expected response data from the probe destination. Valid entries are text strings with a maximum of 255 characters. This field accepts both single and double quotes. Double quotes are considered delimiters so they don't appear on the device. Single quotes will appear on the device.
Expect Regex Offset	Number of characters into the received message or buffer where the ACE is to begin looking for the string specified in the Expect Regular Expression field. Value entries are from 1 to 4000.

To configure probe expect statuses for SIP-TCP probes, see the "Configuring Health Monitoring Expect Status" section on page 7-51.

SIP-UDP Probe Attributes

Table 7-25 lists the SIP-UDP probe attributes.

Note Click **More Settings** to access the additional attributes for the SIP-UDP probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-25 SIP-UDP Probe Attributes

Field	Action
More Settings	
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
Expect Regular Expression	Expected response data from the probe destination. Valid entries are text strings with a maximum of 255 characters. This field accepts both single and double quotes. Double quotes are considered delimiters so they don't appear on the device. Single quotes will appear on the device.
Expect Regex Offset	Number of characters into the received message or buffer where the ACE is to begin looking for the string specified in the Expect Regular Expression field. Value entries are from 1 to 4000.

To configure probe expect statuses for SIP-UDP probes, see the "Configuring Health Monitoring Expect Status" section on page 7-51.

SMTP Probe Attributes

Table 7-26 lists the SMTP probe attributes.

Note Click **More Settings** to access the additional attributes for the SMTP probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-26 SMTP Probe Attributes

Field	Action
More Settings	
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
Is Connection	Check box that indicates that the connection parameters are configured. Clear the check box to indicate that connection parameters are not configured.
Open Timeout (Seconds)	Number of seconds to wait when opening a connection with a real server. Valid entries are from 1 to 65535, and the default value is 10.

To configure probe expect statuses for SMTP probes, see the "Configuring Health Monitoring Expect Status" section on page 7-51.

SNMP Probe Attributes

Table 7-27 lists the SNMP probe attributes.

Note Click **More Settings** to access the additional attributes for the SNMP probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-27 SNMP Probe Attributes

Field	Action
SNMP Community	SNMP community string. Valid entries are unquoted text strings with no spaces and a maximum of 255 characters.
More Settings	
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
SNMP Version	SNMP version for the probe: • **N/A**—No version is selected. • **SNMPv1**—This probe is to use SNMP version 1. • **SNMPv2c**—This probe is to use SNMP version 2c.

To configure the SNMP OID for SNMP probes, see the "Configuring an OID for SNMP Probes" section on page 7-52.

TCP Probe Attributes

Table 7-28 lists the TCP probe attributes.

Note Click **More Settings** to access the additional attributes for the TCP probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-28 TCP Probe Attributes

Field	Action
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
Send Data	ASCII data that the probe is to send to the server. Valid entries are unquoted text strings with no spaces and a maximum of 255 characters.
More Settings	
Is Connection	Check box that indicates that the connection parameters are configured. Uncheck the check box to indicate that connection parameters are not configured.
Open Timeout (Seconds)	Number of seconds to wait when opening a connection with a real server. Valid entries are from 1 to 65535. The default is 10.
Expect Regular Expression	Expected response data from the probe destination. Valid entries are text strings (quotes allowed) with a maximum of 255 characters.
Expect Regex Offset	Number of characters into the received message or buffer where the ACE is to begin looking for the string specified in the Expect Regular Expression field. Value entries are from 1 to 4000.

Telnet Probe Attributes

Table 7-29 lists the Telnet probe attributes.

Note Click **More Settings** to access the additional attributes for the Telnet probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-29 Telnet Probe Attributes

Field	Action
More Settings	
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
Is Connection	Check box that indicates that the connection parameters are configured. Clear the check box to indicate that connection parameters are not configured.
Open Timeout (Seconds)	Enter the number of seconds to wait when opening a connection with a real server. Valid entries are from 1 to 65535. The default is 10.

UDP Probe Attributes

Table 7-30 lists the UDP probe attributes.

Note Click **More Settings** to access the additional attributes for the UDP probe type. By default, ANM hides the probe attributes with default values and the probe attributes that are not commonly used.

Table 7-30 UDP Probe Attributes

Field	Action
Port	Port number that the probe is to use. By default, the probe uses the port number based on its type.
Send Data	ASCII data that the probe is to send to the server. Valid entries are unquoted text strings with no spaces and a maximum of 255 characters.
More Settings	
Expect Regular Expression	Expected response data from the probe destination. Valid entries are text strings (quotes allowed) with a maximum of 255 characters.
Expect Regex Offset	Number of characters into the received message or buffer where the ACE is to begin looking for the string specified in the Expect Regular Expression field. Value entries are from 1 to 4000.

Configuring DNS Probe Expect Addresses

You can specify the IP address that the ACE expects to receive in response to a DNS request. When a DNS probe sends a domain name resolve request to the server, it verifies the returned IP address by matching the received IP address with the configured addresses.

Assumption

A DNS probe has been configured. See the "Configuring Health Monitoring for Real Servers" section on page 7-30 for more information.

Procedure

Step 1 Choose **Config > Devices >** ***context*** **> Load Balancing > Health Monitoring**.

The Health Monitoring table appears.

Step 2 In the Health Monitoring table, choose the DNS probe that you want to configure with an expected IP address.

The Expect Addresses table appears.

Step 3 In the Expect Addresses table, click **Add** to add an entry to the Expect Addresses table.

The Expect Address configuration pane appears.

Note You cannot modify an entry in the Expect Addresses table. Instead, delete the existing entry, then add a new one.

Step 4 In the IP Address field of the Expect Address configuration pane, enter the IP address that the ACE is to expect as a server response to a DNS request.

Valid entries are unique IP addresses in dotted-decimal notation, such as 192.168.11.1.

Step 5 Do one of the following:

- Click **Deploy Now** to deploy this configuration on the ACE and save your entries to the running-configuration and startup-configuration files.
- Click **Cancel** to exit this procedure without saving your entry and to return to the Expect Addresses table.

- Click **Next** to deploy your entry and to add another IP Address to the Expect Addresses table.

Related Topics

- Configuring Health Monitoring for Real Servers, page 7-30
- DNS Probe Attributes, page 7-35
- Displaying Health Monitoring Statistics and Status Information, page 7-53

Configuring Headers for HTTP and HTTPS Probes

You can specify header fields for HTTP and HTTPS probes.

Assumption

An HTTP or HTTPS probe has been configured. See the "Configuring Health Monitoring for Real Servers" section on page 7-30 for more information.

Procedure

Step 1 Choose **Config > Devices >** ***context*** **> Load Balancing > Health Monitoring**.

The Health Monitoring table appears.

Step 2 In the Health Monitoring table, choose the HTTP or HTTPS probe that you want to configure with a header.

The Probe Headers table appears.

Step 3 In the Probe Headers table, click **Add** to add an entry, or choose an existing entry and click **Edit** to modify it.

The Probe Headers configuration pane appears.

Step 4 In the Header Name field of the Probe Headers configuration pane, choose the HTTP header the probe is to use.

Step 5 In the Header Value field, enter the string to assign to the header field.

Valid entries are text strings with a maximum of 255 characters. If the string includes spaces, enclose the string with quotes.

Step 6 Do one of the following:

- Click **Deploy Now** to deploy this configuration on the ACE and save your entries to the running-configuration and startup-configuration files.
- Click **Cancel** to exit this procedure without saving your entry and to return to the Probe Headers table.
- Click **Next** to deploy your entry and to add another header entry to the Probe Headers table.

Related Topics

- Configuring Health Monitoring for Real Servers, page 7-30
- HTTP Probe Attributes, page 7-37

- HTTPS Probe Attributes, page 7-39
- Displaying Health Monitoring Statistics and Status Information, page 7-53

Configuring Health Monitoring Expect Status

You can configure a single or range of code responses that the ACE expects from the probe destination. When the ACE receives a response from the server, it expects a status code to mark a server as passed. By default, there are no status codes configured on the ACE. If you do not configure a status code, any response code from the server is marked as failed.

Expect status codes can be configured for FTP, HTTP, HTTPS, RTSP, SIP-TCP, SIP-UDP, and SMTP probes.

Assumption

An FTP, HTTP, HTTPS, RTSP, SIP-TCP, SIP-UDP or SMTP probe has been configured. See the "Configuring Health Monitoring for Real Servers" section on page 7-30 for more information.

Procedure

Step 1 Choose **Config > Devices >** ***context*** **> Load Balancing > Health Monitoring**.

The Health Monitoring table appears.

Step 2 In the Health Monitoring table, choose the probe that you want to configure for expect status codes, and click the **Expect Status** tab.

The Expect Status table appears.

Step 3 In the Expect Status table, click **Add** to add an entry, or select an existing entry and click **Edit** to modify it.

The Expect Status configuration pane appears.

Step 4 In the Expect Status configuration pane, configure a single expect status code as follows:

a. In the Min. Expect Status Code field, enter the expect status code for this probe. Valid entries are from 0 to 999.

b. In the Max. Expect Status code, enter the same expect status code that you entered in the Min Expect Status Code field.

Step 5 In the Expect Status configuration pane, configure a range of expect status codes as follows:

a. In the Min. Expect Status Code, enter the lower limit of the range of status codes. Valid entries are from 0 to 999.

b. In the Max. Expect Status Code, enter the upper limit of a range of status codes. Valid entries are from 0 to 999. The value in this field must be greater than or equal to the value in the Min Expect Status Code field.

Step 6 Do one of the following:

- Click **Deploy Now** to deploy this configuration on the ACE and save your entries to the running-configuration and startup-configuration files.
- Click **Cancel** to exit this procedure without saving your entries and to return to the Expect Status table.

- Click **Next** to deploy your entries and to add another expect status code to the Expect Status table.

Related Topics

- Configuring Health Monitoring for Real Servers, page 7-30
- FTP Probe Attributes, page 7-37
- HTTP Probe Attributes, page 7-37
- SMTP Probe Attributes, page 7-46
- Displaying Health Monitoring Statistics and Status Information, page 7-53

Configuring an OID for SNMP Probes

You can configure OID queries to probe the server. When the ACE sends a probe with an SNMP OID query, the ACE uses the retrieved value as input to the least-loaded algorithm for load-balancing decisions. Least-loaded load balancing bases the server selection on the server with the lowest load value. If the retrieved value is within the configured threshold, the server is marked as passed. If the threshold is exceeded, the server is marked as failed.

The ACE allows a maximum of eight OID queries to probe the server.

Assumption

An SNMP probe has been configured. See the "Configuring Health Monitoring for Real Servers" section on page 7-30 for more information.

Procedure

Step 1 Choose **Config > Devices >** ***context*** **> Load Balancing > Health Monitoring**.

The Health Monitoring table appears.

Step 2 In the Health Monitoring table, choose the SNMP probe for which you want to specify an OID.

The SNMP OID for Server Load Query table appears.

Step 3 In the SNMP OID for Server Load Query table, click **Add** to add an entry, or choose an existing entry and click **Edit** to modify it.

The SNMP OID configuration pane appears.

Step 4 In the SNMP OID field of the SNMP OID configuration pane, enter the OID that the probe is to use to query the server for a value.

Valid entries are unquoted strings with a maximum of 255 alphanumeric characters in dotted-decimal notation, such as .1.3.6.1.4.2021.10.1.3.1. The OID string is based on the server type.

Step 5 In the Max. Absolute Server Load Value field, enter the OID value in the form of an integer and to indicate that the retrieved OID value is an absolute value instead of a percent.

Valid entries are from 1 to 4294967295.

When the ACE sends a probe with an SNMP OID query, the ACE uses the retrieved value as input to the least-loaded algorithm for load-balancing decisions. By default, the ACE assumes that the retrieved OID value is a percentile value. Use this option to specify that the retrieved OID value is an absolute value.

Step 6 In the Server Load Threshold Value field, specify the threshold at which the server is to be taken out of service as follows:

- When the OID value is based on a percent, valid entries are integers from 1 to 100.
- When the OID is based on an absolute value, valid entries are from 1 to the value specified in the Maximum Absolute Server Load Value field.

Step 7 In the Server Load Weighting field, enter the weight to assign to this OID for the SNMP probe.

Valid entries are from 0 to 16000.

Step 8 Do one of the following:

- Click **Deploy Now** to deploy this configuration on the ACE and save your entries to the running-configuration and startup-configuration files.
- Click **Cancel** to exit this procedure without saving your entries and to return to the SNMP OID table.
- Click **Next** to deploy your entries and to add another item to the SNMP OID table.

Related Topics

- Configuring Health Monitoring for Real Servers, page 7-30
- SNMP Probe Attributes, page 7-47
- Displaying Health Monitoring Statistics and Status Information, page 7-53

Displaying Health Monitoring Statistics and Status Information

You can display statistics and status information for a particular probe.

Procedure

Step 1 Choose **Config > Devices >** ***context*** **> Load Balancing > Health Monitoring**.

The Health Monitoring table appears.

Step 2 In the Health Monitoring table, choose a probe from the Health Monitoring table, and click **Details**.

The **show probe** *name* **detail** CLI command output appears. For details on the displayed output fields, see the *Cisco ACE Module Server Load-Balancing Configuration Guide* or the *Cisco ACE 4700 Series Appliance Server Load-Balancing Configuration Guide*, Chapter 4, Configuring Health Monitoring.

Note For a DNS probe, the detailed probe results always identify a default DNS domain of www.Cisco.com.

Step 3 Click **Update Details** to refresh the output for the **show probe** *name* **detail** CLI command.

Step 4 Click **Close** to return to the Health Monitoring table.

Related Topics

- Configuring Health Monitoring for Real Servers, page 7-30

Configuring Secure KAL-AP

You can configure a secure keepalive-appliance protocol (KAL-AP) associated with a virtual context. A KAL-AP on the ACE enables communication between the ACE and a Global Site Selector (GSS), which sends KAL-AP requests to report the server states and loads for global-server load-balancing (GSLB) decisions. The ACE uses KAL-AP through a UDP connection to calculate weights and provide information for server availability to the KAL-AP device. The ACE acts as a server and listens for KAL-AP requests. When KAL-AP is initialized on the ACE, the ACE listens on the standard 5002 port for any KAL-AP requests. You cannot configure any other port.

The ACE supports secure KAL-AP for MD5 encryption of data between it and the GSS. For encryption, you must configure a shared secret as a key for authentication between the GSS and the ACE context.

Assumptions

This topic assumes the following:

- You have created a virtual context that specifies the Keepalive Appliance Protocol over UDP.
- You have enabled KAL-AP on the ACE by configuring a management class map and policy map, and apply it to the appropriate interface.

Procedure

Step 1 Choose **Config > Devices >** ***context*** **> Load Balancing > Secure KAL-AP**.

The Secure KAL-AP table appears.

Step 2 In the Secure KAL-AP table, click **Add** to configure secure KAL-AP for MD5 encryption of data.

The Secure KAL-AP configuration window appears.

Step 3 In the IP Address field of the Secure KAL-AP configuration window, enable secure KAL-AP by configuring the VIP address for the GSS.

Enter the IP address in dotted-decimal notation (for example, 192.168.11.1).

Step 4 In the Hash Key field, enter the MD5 encryption method shared secret between the KAL-AP device and the ACE.

Enter the shared secret as a case-sensitive string with no spaces and a maximum of 31 alphanumeric characters. The ACE supports the following special characters in a shared secret:

, . / = + - ^ @ ! % ~ # $ * ( )

Step 5 Do one of the following:

- Click **Deploy Now** to deploy this configuration on the ACE and save your entries to the running-configuration and startup-configuration files. The ACE validates the secure KAL-AP configuration and deploys it.
- Click **Cancel** to exit this procedure without accepting your entries and to return to the Secure KAL-AP table.
- Click **Next** to accept your entries.

Related Topics

- Creating Virtual Contexts, page 5-2
- Setting Match Conditions for Layer 3/Layer 4 Management Traffic Class Maps, page 13-12