

Configuring Load Balancing on the CSS 11500

Document ID: 28862

Introduction

Before You Begin

Conventions

Prerequisites

Components Used

Services

Round Robin

Weighted Round Robin

Least Connections/Bytes

ArrowPoint Content Aware (ACA)

Other Methods

Configure

Network Diagram

Configurations

Testing

Related Information

Introduction

Cisco CSS 11500 series content services switches offer multiple services to load balance services within a content rule.

Before You Begin

Conventions

For more information on document conventions, see the Cisco Technical Tips Conventions.

Prerequisites

There are no specific prerequisites for this document.

Components Used

This document is not restricted to specific software and hardware versions.

Services

Services that can help you load balance include:

- Round Robin
- Weighted Round Robin
- Least Connections/Bytes
- ArrowPoint Content Aware (ACA)
- Other Methods

Round Robin

This service distributes Layers 3–5 requests in rotation. The connections are prone to falling into a black hole if requests overload the server.

Weighted Round Robin

Weighted Round Robin (WRR) behaves like the Round Robin algorithm. However, using WRR, you can manually weight servers to get picked more often.

Least Connections/Bytes

With this service, a CSS 11500 correlates the server load and (to?) the number of active connections. A CSS 11500 cannot recognize real server performance differences.

ArrowPoint Content Aware (ACA)

The CSS 11500 uses the ArrowPoint Content Aware (ACA) service to gather response time data for every flow for building statistical averages/variances for every service and content rule. The best servers are used, while the slow servers are pruned from the eligible list. This service also can manage persistent connections for e-commerce based on IP addresses (range), Secure Socket Layer (SSL) application IDs, and cookies.

Other Methods

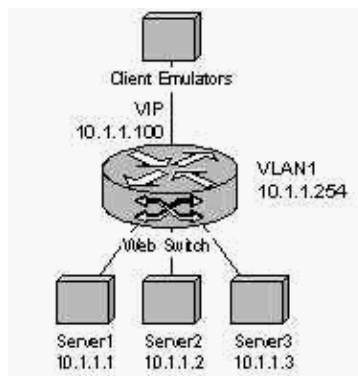
Other balancing methods are **urlhash**, **domainhash**, **url**, **domain**, **srcip**, and **destip**. However, these methods are not covered in this document. For more information about these balancing methods, refer to the CSS Command Reference.

Configure

In the following example, two identical Hypertext Transfer Protocol (HTTP) servers are connected to a CSS 11500. Different algorithms are used for evaluation. This example uses load balanced HTML content with the wildcard symbol `/*.html`. The Virtual IP (VIP) address 10.1.1.101 uses Network Address Translation (NTA) to reach the servers' IP addresses. The ACA dynamic load-balancing algorithm is used. ACA uses measured metrics to determine the best server within the rule to use.

Note: To find additional information on the commands used in this document, use the Command Lookup Tool (registered customers only).

Network Diagram



Configurations

Device 1

```
!Generated MAY  5 15:50:40
!Active version: ap0310027

configure
!***** GLOBAL *****

username admin des-password
  ip route 0.0.0.0 0.0.0.0 10.1.1.101

!***** CIRCUIT *****

circuit VLAN1
  ip address 10.1.1.254 255.255.255.0

!***** SERVICE *****

service Server1
  ip address 10.1.1.1
  keepalive type http
  keepalive uri "/"
  active

service Server2
  ip address 10.1.1.2
  keepalive type http
  keepalive uri "/"
  active

service Server3
  ip address 10.1.1.3
  keepalive type http
  keepalive uri "/"
  weight 5

!--- Makes the server get hit more often.
!--- The default weight is 1.

  active

!***** OWNER *****

owner foo.com
  content L3_LeastConnections
    vip address 10.1.1.100
    add service Server1
    add service Server2
    add service Server3
    balance leastconn

!--- Balance based on least connections content L3_RoundRobin.

  active
  content L3_RoundRobin
    vip address 10.1.1.100
    VIP address 10.1.1.100
    add service Server1
    add service Server2
    add service Server3

!--- The default is round robin.
```

```

    active

content L5_ACA
    port 80
    protocol tcp
    VIP address 10.1.1.100
    add service Server1
    add service Server2
    add service Server3
    balance aca

!--- Used to dynamically balance server.

    url "/*.html"

!--- Use this rule only with HTML documents.

    active
content L5_WeightedRR
    port 80
    protocol tcp
    VIP address 10.1.1.100
    add service Server1
    add service Server2
    add service Server3
    balance weightedrr

!--- Use the weight information found in the service.

    url "/*.gif"

!--- Only use this rule for GIF documents.

    Use the weight info found in the service
    active

```

Testing

Use the following steps to verify or troubleshoot your configuration of CSS load balancing.

Note: Certain **show** commands are supported by the Output Interpreter Tool (registered customers only), which allows you to view an analysis of **show** command output.

1. Verify that all servers are up by using the **show service summary** command.
2. Activate the L3_Least Connections rule.
3. Start the client emulators.
4. Issue the show summary command to see the hit counts by service. The last server does not get hit as often if the first servers are fast enough to handle the connections.
5. Activate the L3_Round Robin rule.
6. Start the client emulators. All servers will be hit equally.
7. Start the client emulators and have them request 1.gif and 2.gif. The switch recognizes that the requested file ends in the .gif file extension and applies the L5_WRR rule. Secondly, the weighting on the third server is five times that of the other servers so 5x more .gif files are served from.
8. Repeat the test with the HTML document. Server3 again will receive the most hits. ACA uses a combination of dynamic learned response time information and the load factors on the server with manual tuning that use parameters such as weight.

Note: The three servers need substantial traffic for you to see the merits of using the ACA service.

9. The tests were repeated using clients. Different content (HTML, GIF, JPEG) was requested. In this

example, Round Robin treated every JPEG equally because no other rule matched it. Five times as many GIFs were served by Server3. ACA determined that while Server3 was the preferred server, it was overloaded, so it redistributed requests over the other servers dynamically. To determine a preferred server, use the **show service summary** command and look at the results in the Connections and Load columns.

Service Name	State	Conn	Weight	Avg Load	State Transitions
Server1	Alive	22	1	40	0
Server2	Alive	25	1	9	0
Server3	Alive	68	5	76	0

The **show summary** command output shows that the ACA rule hit the servers, based on their loads. Round Robin hit the servers equally. WRR hit Server3 the most because of the manual weighting. The Least Connections results show that the first two servers handled most of the load.

Global Bypass Counters:

No Rule Bypass Count: 0

Acl Bypass Count: 0

Owner	Content Rules	State	Services	Service Hits
foo.com	L5_ACA	Active	Server1	520
			Server2	608
			Server3	854
	L3_RoundRobin	Active	Server1	665
			Server2	665
			Server3	665
	L5_WeightedRR	Active	Server1	278
			Server2	277
			Server3	1387
L3_LeastConnecti	Suspended	Server1	665	
		Server2	650	
		Server3	201	

Note: The counters are per owner and per rule. To clear all counters, issue the **zero all** command at the config-owner[foo.com])# prompt. To clear counters for a rule, enter into the configuration mode for the rule and then issue the **zero all** command.

Related Information

- [Cisco CSS 11500 Series Product Support Page](#)
 - [End-of-Sale CSS Models](#)
 - [Content Delivery Products](#)
 - [Technical Support – Cisco Systems](#)
-

All contents are Copyright © 2006–2007 Cisco Systems, Inc. All rights reserved. Important Notices and Privacy Statement.

Updated: Aug 02, 2007

Document ID: 28862
