Information About Cisco IOS SLB

Last Updated: April 27, 2011

To configure IOS SLB, you should understand the following concepts:

Note Some IOS SLB features are specific to one platform and are not described in this feature document. For information about those features, refer to the appropriate platform-specific documentation.

Overview

This document describes how to configure the Cisco IOS Server Load Balancing (IOS SLB) feature. For a complete description of the IOS SLB commands in this chapter, refer to the "Server Load Balancing Commands" chapter of the *Cisco IOS IP Application Services Command Reference*. To locate documentation of other commands that appear in this chapter, use the command reference master index or search online.

The SLB feature is a Cisco IOS-based solution that provides IP server load balancing. Using the IOS SLB feature:

1. The network administrator defines a *virtual* server that represents a group of *real* servers in a cluster of network servers known as a *server farm*. In this environment the clients are configured to connect to the IP address of the virtual server.
2. The virtual server IP address is configured as a loopback address, or secondary IP address, on each of the real servers.

3 When a client initiates a connection to the virtual server, the IOS SLB function chooses a real server for the connection based on a configured load-balancing algorithm.

The IOS SLB feature provides load balancing for a variety of networked devices and services, including:

- Application servers, such as Hypertext Transfer Protocol (HTTP), Telnet, File Transfer Protocol (FTP), and so on
- Firewalls
- Service nodes, such as authentication, authorization, and accounting (AAA) servers, web caches, and so on

In addition, the IOS SLB Exchange Director enables advanced load-balancing routing capabilities for the following additional service nodes:

- mobile Service Exchange Framework (mSEF) components:
    - Cisco Content Services Gateways (CSGs)

If you are running with Supervisor Engine 32 (SUP32-MSFC2A), CSG Release 3.1(3)C7(1) or later is required.

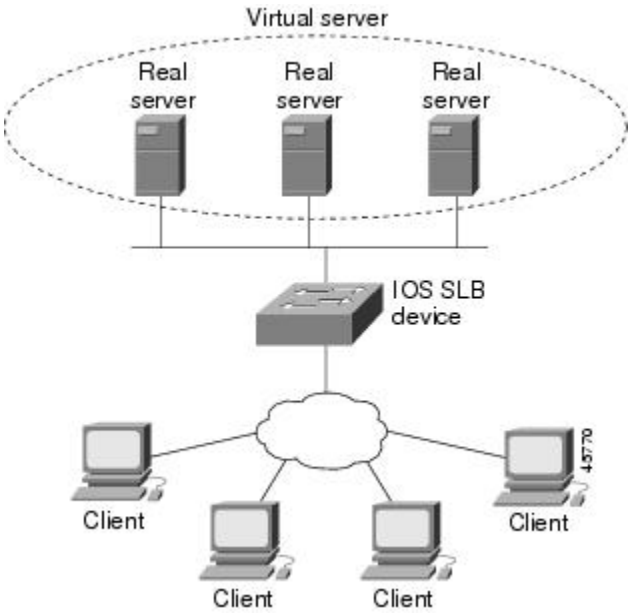
-
    - Cisco gateway general packet radio service (GPRS) support nodes (GGSNs)
    - Cisco Service Selection Gateways (SSGs)
    - Cisco Home Agents
- Other components for mobile, Public Wireless LAN (PWLAN), and Service Provider networks:
    - Wireless Application Protocol (WAP) gateways
    - Protocol optimization gateways
    - Non-Cisco GGSNs and Home Agents
    - Other RADIUS-aware flow gateways. These gateways are proxies or routing nodes that receive RADIUS Authorization and Accounting requests for users that route flows through the gateways. The Exchange Director binds the RADIUS and data flows to the same gateway, ensuring that the gateway receives a complete and consistent view of the network activity for the user.

The Exchange Director also adds the following features:

- Enhanced failover capabilities for single-chassis failover within the mSEF for Cisco Catalyst 6500 series switches and Cisco 7600 series routers. When used with Route Processor Redundancy Plus (RPR+), IOS SLB stateful backup for redundant route processors provides full IOS SLB stateful failover for these platforms.
- Flow persistence, which provides intelligent return routing of load-balanced IP flows.

The figure below illustrates a simple IOS SLB network.

Figure 1: Logical View of Cisco IOS SLB



Benefits of IOS SLB

IOS SLB shares the same software code base as Cisco IOS and has all of the software features sets of Cisco IOS software.

On Cisco Catalyst 6500 series switches, IOS SLB uses hardware acceleration to forward packets at a very high speed when running in dispatched mode.

IOS SLB assures continuous, high availability of content and applications with techniques for actively managing servers and connections in a distributed environment. By distributing user requests across a cluster of servers, IOS SLB optimizes responsiveness and system capacity, and reduces the cost of providing Internet, database, and application services for large, medium, and small-scale sites.

IOS SLB facilitates scalability, availability, and ease of maintenance as follows:

- The addition of new physical (real) servers, and the removal or failure of existing servers, can occur at any time, transparently, without affecting the availability of the virtual server.
- IOS SLB's slow start capability allows a new server to increase its load gradually, preventing failures caused by assigning many new connections to the server in a short period.
- IOS SLB supports fragmented packets and packets with IP options, buffering your servers from client or network vagaries that are beyond your control.
- IOS SLB firewall load balancing enables you to scale access to your Internet site. You can add firewalls without affecting existing connections, enabling your site to grow without impacting customers.

Using DFP enables IOS SLB to provide weights to another load-balancing system. IOS SLB can act as a DFP manager, receiving weights from host servers, and it can act as a DFP agent, sending weights to a DFP

manager. The functions are enabled independently--you can implement either one, or both, at the same time.

IOS SLB makes the administration of server applications easy. Clients know only about virtual servers; no administration is required for real server changes.

IOS SLB provides security for the real server because it never announces the real server's address to the external network. Users are familiar only with the virtual IP address. You can filter unwanted flows based on both IP address and TCP or UDP port numbers. Additionally, though it does not eliminate the need for a firewall, IOS SLB can help protect against some denial-of-service attacks.

In a branch office, IOS SLB allows balancing of multiple sites and disaster recovery in the event of full-site failure, and distributes the work of load balancing.

Cisco IOS SLB Features

Routing Features

Algorithms for Server Load Balancing

IOS SLB provides the following load-balancing algorithms:

You can specify one of these algorithms as the basis for choosing a real server for each new connection request that arrives at the virtual server.

For each algorithm, connections in the closing state are counted against the number of connections assigned to a real server. This impacts the least connections algorithm more than the other algorithms, because the least connections algorithm is influenced by the number of connections. IOS SLB adjusts the number of connections per real server, and the algorithm metrics, each time a connection is assigned.

- Weighted Round Robin Algorithm, page 5
- Weighted Least Connections Algorithm, page 5
- Route Map Algorithm, page 6

Weighted Round Robin Algorithm

The weighted round robin algorithm specifies that the real server used for a new connection to the virtual server is chosen from the server farm in a circular fashion. Each real server is assigned a weight, n, that represents its capacity to manage connections, as compared to the other real servers associated with the virtual server. That is, new connections are assigned to a given real server n times before the next real server in the server farm is chosen.

For example, assume a server farm comprised of real server ServerA with n = 3, ServerB with n = 1, and ServerC with n = 2. The first three connections to the virtual server are assigned to ServerA, the fourth connection to ServerB, and the fifth and sixth connections to ServerC.

Note To configure the IOS SLB device to use a round robin algorithm, assign a weight of n=1 to all of the servers in the server farm. GPRS load balancing *without* GTP cause code inspection enabled requires the weighted round robin algorithm. You can bind a server farm that uses weighted least connections to a virtual server providing GPRS load balancing without GTP cause code inspection enabled, but you cannot place the virtual server in service. If you try to do so, IOS SLB issues an error message. The Home Agent Director requires the weighted round robin algorithm. You can bind a server farm that uses weighted least connections to a Home Agent Director virtual server, but you cannot place the virtual server INSERVICE. If you try to do so, IOS SLB issues an error message. RADIUS load balancing requires the weighted round robin algorithm. RADIUS load balancing accelerated data plane forwarding *does not* support the weighted round robin algorithm.

Weighted Least Connections Algorithm

The weighted least connections algorithm specifies that the next real server chosen from a server farm is the server with the fewest active connections. Each real server is assigned a weight for this algorithm, also. When weights are assigned, the server with the fewest connections is based on the number of active connections on each server, and on the relative capacity of each server. The capacity of a given real server is calculated as the assigned weight of that server divided by the sum of the assigned weights of all of the real servers associated with that virtual server, or n1/(n1+n2+n3...).

For example, assume a server farm comprised of real server ServerA with n = 3, ServerB with n = 1, and ServerC with n = 2. ServerA would have a calculated capacity of 3/(3+1+2), or half of all active connections on the virtual server, ServerB one-sixth of all active connections, and ServerC one-third of all active connections. At any point in time, the next connection to the virtual server would be assigned to the real server whose number of active connections is farthest below its calculated capacity.

Note Assigning a weight of n=1 to all of the servers in the server farm configures the IOS SLB device to use a simple least-connection algorithm. GPRS load balancing *without* GTP cause code inspection enabled *does not* support the weighted least connections algorithm. GPRS load balancing *with* GTP cause code inspection enabled *does* support the weighted least connections algorithm. Access Service Network (ASN) load balancing (for Mobile Station Pre-Attachment requests), the Home Agent Director, RADIUS load balancing, and RADIUS load balancing accelerated data plane forwarding *do not* support the weighted least connections algorithm.

Route Map Algorithm

The route map algorithm is valid only with IOS SLB RADIUS load balancing accelerated data plane forwarding, also known as Turbo RADIUS load balancing. Turbo RADIUS load balancing is a high-performance solution that uses policy-based routing (PBR) route maps to manage subscriber data-plane traffic in a Cisco Content Services Gateway (CSG) environment. When Turbo RADIUS load balancing receives a RADIUS payload, it inspects the payload, extracts the framed-IP attribute, applies a route map to the IP address, and then determines which CSG is to manage the subscriber.

For more information about policy-based routing, see the “Policy-Based Routing” and “Configuring Policy-Based Routing” sections of the *Cisco IOS IP Routing Configuration Guide* .

Note RADIUS load balancing accelerated data plane forwarding requires the route map algorithm.

Bind ID Support

A bind ID allows one physical server to be bound to multiple virtual servers and report a different weight for each one. Thus, the single real server is represented as multiple instances of itself, each having a different bind ID. Dynamic Feedback Protocol (DFP) uses the bind ID to identify the instance of the real server for which a given weight is specified. Use the bind ID feature only if you are using DFP.

GPRS load balancing and the Home Agent Director do not support bind IDs.

Client-Assigned Load Balancing

Client-assigned load balancing allows you to limit access to a virtual server by specifying the list of client IP subnets that are permitted to use that virtual server. With this feature, you can assign a set of client IP subnets (such as internal subnets) connecting to a virtual IP address to one server farm or firewall farm, and assign another set of clients (such as external clients) to a different server farm or firewall farm.

GPRS load balancing and the Home Agent Director do not support client-assigned load balancing.

Connection Rate Limiting

IOS SLB enables you to specify the maximum connection rate allowed for a real server in a server farm. For more information, see the description of the **rate**command in real server configuration mode.

Content Flow Monitor Support

IOS SLB supports the Cisco Content Flow Monitor (CFM), a web-based status monitoring application within the CiscoWorks2000 product family. You can use CFM to manage Cisco server load-balancing devices. CFM runs on Windows NT and Solaris workstations, and is accessed using a web browser.

Delayed Removal of TCP Connection Context

Because of IP packet ordering anomalies, IOS SLB might "see" the end of a TCP connection (a finish [FIN] or reset [RST]) followed by other packets for the connection. This problem usually occurs if multiple paths exist for the TCP connection packets to follow. To correctly redirect the packets that arrive after the connection has ended, IOS SLB retains the TCP connection information, or context, for a specified length of time. The length of time the context is retained after the connection ends is controlled by a configurable delay timer.

Firewall Load Balancing

As its name implies, firewall load balancing:

- Enables IOS SLB to balance flows to firewalls.
- Uses a load-balancing device on each side of a group of firewalls (called a firewall farm) to ensure that the traffic for each flow travels to the same firewall, ensuring that the security policy is not compromised.

You can configure more than one firewall farm in each load-balancing device.

- Layer 3 firewalls--Have IP-addressable interfaces. Are supported by IOS SLB firewall load balancing if they are subnet-adjacent to the firewall load-balancing device and have unique MAC addresses. The device does not modify the IP addresses in the user packet. To send the packet to the chosen firewall, the device determines which interface to use and changes the Layer 2 headers accordingly. This type of routing is the standard dispatched routing used by IOS SLB.
- Layer 2 firewalls--Do not have IP addresses. Are transparent to IOS SLB firewall load balancing. IOS SLB supports Layer 2 firewalls by placing them between two IP-addressable interfaces.

Whereas many Layer 3 firewalls might exist off one Layer 3 interface on the load-balancing device (for example, one LAN), only one Layer 2 firewall can exist off each interface.

When configuring the load-balancing device, you configure a Layer 3 firewall using its IP address, and a Layer 2 firewall using the IP address of the interface of the device on the "other side" of the firewall.

To balance flows across the firewalls in a firewall farm, IOS SLB firewall load balancing performs a route lookup on each incoming flow, examining the source and destination IP addresses (and optionally the source and destination TCP or User Datagram Protocol [UDP] port numbers). Firewall load balancing applies a hash algorithm to the results of the route lookup and selects the best firewall to manage the connection request.

Note IOS SLB firewall load balancing *must* examine incoming packets and perform route lookup. With Cisco Catalyst 6500 series switches, some additional packets might need to be examined. Firewall load balancing has an impact on internal (secure) side routing performance and must be considered in the complete design.

To maximize availability and resilience in a network with multiple firewalls, configure a separate equal-weight route to each firewall, rather than one route to only one of the firewalls.

IOS SLB firewall load balancing provides the following capabilities:

- Connections initiated from either side of the firewall farm are load-balanced.

- The load is balanced among a set of firewalls--the firewall farm.
- All packets for a connection travel through the same firewall. Subsequent connections can be “sticky,” ensuring that they are assigned to the same firewall.
- Source-IP, destination-IP, and source-destination-IP sticky connections are supported.
- Probes are used to detect and recover from firewall failures.
- Redundancy is provided. Hot Standby Router Protocol (HSRP), stateless backup, and stateful backup are all supported.
- Multiple interface types and routing protocols are supported, enabling the external (Internet side) load-balancing device to act as an access router.
- Proxy firewalls are supported.

GTP IMSI Sticky Database

IOS SLB can select a GGSN for a given International Mobile Subscriber ID (IMSI), and forward all subsequent Packet Data Protocol (PDP) create requests from the same IMSI to the selected GGSN.

To enable this feature, IOS SLB uses a GTP IMSI sticky database, which maps each IMSI to its corresponding real server, in addition to its session database.

1. IOS SLB creates a sticky database object when it processes the first GTP PDP create request for a given IMSI.
2. IOS SLB removes the sticky object when it receives a notification to do so from the real server, or as a result of inactivity.
3. When the last PDP belonging to an IMSI is deleted on the GGSN, the GGSN notifies IOS SLB to remove the sticky object.

Home Agent Director

Home agents are the anchoring points for mobile nodes. They route flows for a mobile node to its current foreign agent (point of attachment).

The Home Agent Director load balances Mobile IP Registration Requests (RRQs) among a set of home agents (configured as real servers in a server farm). The Home Agent Director has the following characteristics:

- Can operate in dispatched mode or in directed server NAT mode, but not in directed client NAT mode. In dispatched mode, the home agents must be Layer 2-adjacent to the IOS SLB device.
- Does not support stateful backup. See the “Stateful Backup” section for more information.
- Delivers RRQs destined to the virtual Home Agent Director IP address to one of the real home agents, using the weighted round robin load-balancing algorithm. See the “Weighted Round Robin Algorithm” section for more information about this algorithm.
- Requires DFP in order to allocate RRQs based on capacity.

For more information about Mobile IP, home agents, and related topics, refer to the *Cisco IOS IP Mobility Configuration Guide*.

Interface Awareness

Some environments require IOS SLB to take into account the input interface when mapping packets to virtual servers, firewall farms, connections, and sessions. In IOS SLB, this function is called interface awareness. When interface awareness is configured, IOS SLB processes only traffic arriving on configured access interfaces. (An access interface is any Layer 3 interface.)

Such “sandwich” environments require IOS SLB on both sides of a farm of CSGs, SSGs, or firewalls. For example, you might want IOS SLB to perform RADIUS load balancing on one side of a farm and firewall load balancing on the other, or firewall load balancing on both sides of a firewall farm.

Maximum Connections

IOS SLB allows you to configure maximum connections for server and firewall load balancing.

- For server load balancing, you can configure a limit on the number of active connections that a real server is assigned. If the maximum number of connections is reached for a real server, IOS SLB automatically switches all further connection requests to other servers until the connection number drops below the specified limit.
- For firewall load balancing, you can configure a limit on the number of active TCP or UDP connections that a firewall farm is assigned. If the maximum number of connections is reached for the firewall farm, new connections are dropped until the connection number drops below the specified limit.

Multiple Firewall Farm Support

You can configure more than one firewall farm in each load-balancing device.

Network Address Translation

Cisco IOS Network Address Translation (NAT), RFC 1631, allows unregistered “private” IP addresses to connect to the Internet by translating them into globally registered IP addresses. As part of this functionality, Cisco IOS NAT can be configured to advertise only one address for the entire network to the outside world. This configuration provides additional security and network privacy, effectively hiding the entire internal network from the world behind that address. NAT has the dual functionality of security and address conservation, and is typically implemented in remote access environments.

- Session Redirection, page 9
- Dispatched Mode, page 10
- Directed Mode, page 10
- Server NAT, page 10
- Client NAT, page 10
- Static NAT, page 11
- Server Port Translation, page 12

Session Redirection

Session redirection NAT involves redirecting packets to real servers. IOS SLB can operate in one of two session redirection modes, dispatched mode or directed mode.

Note In both dispatched and directed modes, IOS SLB must track connections. Therefore, you must design your network so that there is no alternate network path from the real servers to the client that bypasses the load-balancing device.

Dispatched Mode

In dispatched NAT mode, the virtual server address is known to the real servers; you must configure the virtual server IP address as a loopback address, or secondary IP address, on each of the real servers. IOS SLB redirects packets to the real servers at the media access control (MAC) layer. Because the virtual server IP address is not modified in dispatched mode, the real servers must be Layer 2-adjacent to IOS SLB, or intervening routers might not be able to route to the chosen real server.

Refer to the "Configuring Virtual Interfaces" chapter of the *Cisco IOS Interface Configuration Guide* for more information about configuring the loopback address.

Note Some UDP applications cannot respond from the loopback interface. If that situation occurs, you must use directed mode.

Directed Mode

In directed NAT mode, the virtual server can be assigned an IP address that is not known to any of the real servers. IOS SLB translates packets exchanged between a client and a real server, using NAT to translate the virtual server IP address to a real server IP address.

IOS SLB supports the following types of NAT:

Note You can use both server NAT and client NAT for the same connection. IOS SLB does not support FTP or firewall load balancing in directed mode. Therefore, FTP and firewall load balancing cannot use NAT. IOS SLB supports only client NAT for TCP and UDP virtual servers. IOS SLB supports only server NAT (but not server port translation) for Encapsulation Security Payload (ESP) virtual servers or Generic Routing Encapsulation (GRE) virtual servers.

Server NAT

Server NAT involves replacing the virtual server IP address with the real server IP address (and vice versa). Server NAT provides the following benefits:

- Servers can be many hops away from the load-balancing device.
- Intervening routers can route to them without requiring tunnelling.
- Loopback and secondary interfaces are not required on the real server.
- The real server need not be Layer 2-adjacent to IOS SLB.
- The real server can initiate a connection to a virtual server on the same IOS SLB device.

Client NAT

If you use more than one load-balancing device in your network, replacing the client IP address with an IP address associated with one of the devices results in proper routing of outbound flows to the correct device. Client NAT also requires that the ephemeral client port be modified since many clients can use the same ephemeral port. Even in cases where multiple load-balancing devices are not used, client NAT can be useful to ensure that packets from load-balanced connections are not routed around the device.

Static NAT

With static NAT, address translations exist in the NAT translation table as soon as you configure static NAT commands, and they remain in the translation table until you delete the static NAT commands.

You can use static NAT to allow some users to use NAT and allow other users on the same Ethernet interface to continue with their own IP addresses. This option enables you to provide a default NAT behavior for real servers, differentiating between responses from a real server, and connection requests initiated by the real server.

For example, you can use server NAT to redirect Domain Name System (DNS) inbound request packets and outbound response packets for a real server, and static NAT to process connection requests from that real server.

Note Static NAT is not required for DNS, but we recommend it, because static NAT hides your real server IP addresses from the outside world.

IOS SLB supports the following static NAT options, configured using the **ip slb static** command:

- Static NAT with dropped connections--The real server is configured to have its packets dropped by IOS SLB, if the packets do not correspond to existing connections. This option is usually used in conjunction with the subnet mask or port number option on the **real** command in static NAT configuration mode, such that IOS SLB builds connections to the specified subnet or port, and drops all other connections from the real server.
- Static NAT with a specified address--The real server is configured to use a user-specified virtual IP address when translating addresses.
- Static NAT with per-packet server load balancing--The real server is configured such that IOS SLB is not to maintain connection state for packets originating from the real server. That is, IOS SLB is to use server NAT to redirect packets originating from the real server. Per-packet server load balancing is especially useful for DNS load balancing. IOS SLB uses DNS probes to detect failures in the per-packet server load-balancing environment.

Note Static NAT with per-packet server load balancing does not load-balance fragmented packets.

- Static NAT with sticky connections--The real server is configured such that IOS SLB is not to maintain connection state for packets originating from the real server, unless those packets match a sticky object:
  - If IOS SLB finds a matching sticky object, it builds the connection.
  - If IOS SLB does not find a matching sticky object, it forwards the packets without building the connection.

IOS SLB uses the following logic when handling a packet from a real server:

SUMMARY STEPS

1. Does the packet match a real server?
2. Does the packet match an existing connection?
3. Is the real server configured to use static NAT?
4. Is the real server configured to have its packets dropped by IOS SLB, if the packets do not correspond to existing connections?
5. Is the real server configured for per-packet server load balancing?
6. Is the real server configured to maintain connection state for sticky connections?
7. Can IOS SLB find a matching sticky object?

DETAILED STEPS

Step 1 Does the packet match a real server?

- If no, IOS SLB has no interest in the packet.
- If yes, continue.

Step 2 Does the packet match an existing connection?

- If yes, IOS SLB uses NAT to redirect the packet, in accordance with the connection control block.
- If no, continue.

Step 3 Is the real server configured to use static NAT?

- If no, IOS SLB manages the packet as usual. This functionality is also called static NAT pass-through.
- If yes, continue.

Step 4 Is the real server configured to have its packets dropped by IOS SLB, if the packets do not correspond to existing connections?

- If yes, IOS SLB drops the packet.
- If no, continue.

Step 5 Is the real server configured for per-packet server load balancing?

- If yes, IOS SLB uses NAT to redirect the packet.
- If no, continue.

Step 6 Is the real server configured to maintain connection state for sticky connections?

- If no, IOS SLB builds the connection.
- If yes, IOS SLB searches for a matching sticky object. Continue.

Step 7 Can IOS SLB find a matching sticky object?

- If no, IOS SLB drops the packet.
- If yes, IOS SLB builds the connection.

Server Port Translation

Server port translation, also known as port address translation, or PAT, is a form of server NAT that involves the translation of virtual server ports instead of virtual server IP addresses. Virtual server port

translation does not require translation of the virtual server IP address, but you can use the two types of translation together.

IOS SLB supports server port translation for TCP and UDP only.

Port-Bound Servers

Port-bound servers allow one virtual server IP address to represent one set of real servers for one service, such as HTTP, and a different set of real servers for another service, such as Telnet. When you define a virtual server, you must specify the TCP or UDP port managed by that virtual server. However, if you configure NAT on the server farm, you can also configure port-bound servers.

Packets destined for a virtual server address for a port that is not specified in the virtual server definition are not redirected.

IOS SLB supports both port-bound and non-port-bound servers, but port-bound servers are recommended.

IOS SLB firewall load balancing does not support port-bound servers.

Route Health Injection

By default, a virtual server's IP address is advertised (added to the routing table) when you bring the virtual server into service (using the **inservice** command). If there is a preferred host route to a website's virtual IP address, you can advertise that host route, but there is no guarantee that the IP address is available. However, you can use the **advertise** command to configure IOS SLB to advertise the host route only when IOS SLB has verified that the IP address is available. IOS SLB withdraws the advertisement when the IP address is no longer available. This function is known as route health injection.

Sticky Connections

You can use the optional **sticky** command to enable IOS SLB to force connections from the same client to the same load-balanced server within a server farm.

Sometimes, a client transaction can require multiple consecutive connections, which means new connections from the same client IP address or subnet must be assigned to the same real server. These connections are especially important in firewall load balancing, because the firewall might need to profile the multiple connections in order to detect certain attacks.

- IOS SLB supports source-IP sticky connections.
- Firewall load balancing supports source-IP, destination-IP, and source-destination-IP sticky connections.
- RADIUS load balancing supports calling-station-IP, framed-IP, and username sticky connections.

For firewall load balancing, the connections between the same client-server pair are assigned to the same firewall. New connections are considered to be sticky as long as the following conditions are met:

- The real server is in either OPERATIONAL or MAXCONNS_THROTTLED state.
- The sticky timer is defined on a virtual server or on a firewall farm.

This binding of new connections to the same server or firewall is continued for a user-defined period after the last sticky connection ends.

To get the client-server address sticky behavior needed for "sandwich" firewall load balancing, you must enable sticky on both sides of the firewall farm. In this configuration, client-server sticky associations are created when an initial connection is opened between a client-server address pair. After this initial connection is established, IOS SLB maintains the sticky association in the firewall load-balancing devices

on either side of the farm, and applies the sticky association to connections initiated from either the client or server IP address, by both firewall load-balancing devices.

Client subnet sticky is enabled when you specify the **sticky** command with a subnet mask. Subnet sticky is useful when the client IP address might change from one connection to the next. For example, before reaching IOS SLB, the client connections might pass through a set of NAT or proxy firewalls that have no sticky management of their own. Such a situation can result in failed client transactions if the servers do not have the logic to cope with it. In cases where such firewalls assign addresses from the same set of subnets, IOS SLB's sticky subnet mask can overcome the problems that they might cause.

Sticky connections also permit the coupling of services that are managed by more than one virtual server or firewall farm. This option allows connection requests for related services to use the same real server. For example, web server (HTTP) typically uses TCP port 80, and HTTPS uses port 443. If HTTP virtual servers and HTTPS virtual servers are coupled, connections for ports 80 and 443 from the same client IP address or subnet are assigned to the same real server.

Virtual servers that are in the same sticky group are sometimes called buddied virtual servers.

The Home Agent Director does not support sticky connections.

TCP Session Reassignment

IOS SLB tracks each TCP SYNchronize sequence number, or SYN, sent to a real server by a client attempting to open a new connection. If several consecutive SYNs are not answered, or if a SYN is replied to with an RST, the TCP session is reassigned to a new real server. The number of SYN attempts is controlled by a configurable reassign threshold.

IOS SLB firewall load balancing does not support TCP session reassignment.

Transparent Web Cache Load Balancing

IOS SLB can load-balance HTTP flows across a cluster of transparent web caches. To set up this function, configure the subnet IP addresses served by the transparent web caches, or some common subset of them, as virtual servers. Virtual servers used for transparent web cache load balancing do not answer pings on behalf of the subnet IP addresses, and they do not affect traceroute.

In some cases, such as when its cache does not contain needed pages, a web cache must initiate its own connections to the Internet. Those connections should not be load-balanced back to the same set of web caches. To address this need, IOS SLB allows you to configure **client exclude** statements, which exclude connections initiated by the web caches from the load-balancing scheme.

IOS SLB firewall load balancing does not support transparent web cache load balancing.

Security Features

Alternate IP Addresses

IOS SLB enables you to telnet to the load-balancing device using an alternate IP address. To do so, use either of the following methods:

- Use any of the interface IP addresses to telnet to the load-balancing device.
- Define a secondary IP address to telnet to the load-balancing device.

This function is similar to that provided by the LocalDirector (LD) Alias command.

Avoiding Attacks on Server Farms and Firewall Farms

IOS SLB relies on a site's firewalls to protect the site from attacks. In general, IOS SLB is no more susceptible to direct attack than is any switch or router. However, a highly secure site can take the following steps to enhance its security:

- Configure real servers on a private network to keep clients from connecting directly to them. This configuration ensures that the clients must go through IOS SLB to get to the real servers.
- Configure input access lists on the access router or on the IOS SLB device to deny flows from the outside network aimed directly at the interfaces on the IOS SLB device. That is, deny *all* direct flows from unexpected addresses.
- To protect against attackers trying to direct flows to real or nonexistent IP addresses in the firewall subnet, configure the firewalls in a private network.
- Configure firewalls to deny *all* unexpected flows targeted at the firewalls, especially flows originating from the external network.

Slow Start

To prevent an overload, slow start controls the number of new connections that are directed to a real server that has just been placed in service. In an environment that uses weighted least connections load balancing, a real server that is placed in service initially has no connections, and could therefore be assigned so many new connections that it becomes overloaded.

GPRS load balancing and the Home Agent Director do not support slow start.

SynGuard

SynGuard limits the rate of TCP start-of-connection packets (SYNchronize sequence numbers, or SYNs) managed by a virtual server to prevent a type of network problem known as a SYN flood denial-of-service attack. A user might send a large number of SYNs to a server, which could overwhelm or crash the server, denying service to other users. SynGuard prevents such an attack from bringing down IOS SLB or a real server. SynGuard monitors the number of SYNs managed by a virtual server at specific intervals and does not allow the number to exceed a configured SYN threshold. If the threshold is reached, any new SYNs are dropped.

IOS SLB firewall load balancing and the Home Agent Director do not support SynGuard.

Server Failure Detection and Recovery Features

- Probes, page 18

Automatic Server Failure Detection

IOS SLB automatically detects each failed Transmission Control Protocol (TCP) connection attempt to a real server, and increments a failure counter for that server. (The failure counter is not incremented if a failed TCP connection from the same client has already been counted.) If a server's failure counter exceeds a configurable failure threshold, the server is considered out of service and is removed from the list of active real servers.

For RADIUS load balancing, the IOS SLB performs automatic server failure detection when a RADIUS request is not answered by the real server.

If you have configured all-port virtual servers (that is, virtual servers that accept flows destined for all ports except GTP ports), flows can be passed to servers for which no application port exists. When the servers reject these flows, IOS SLB might fail the servers and remove them from load balancing. This situation can also occur in slow-to-respond AAA servers in RADIUS load-balancing environments. To prevent this situation, you can disable automatic server failure detection.

Note If you disable automatic server failure detection using the **no faildetect inband** command, we strongly recommend that you configure one or more probes. If you specify the **no faildetect inband** command, the **faildetect numconns**command is ignored, if specified.

Automatic Unfail

When a real server fails and is removed from the list of active servers, it is assigned no new connections for a length of time specified by a configurable retry timer. After that timer expires, the server is again eligible for new virtual server connections and IOS SLB sends the server the next qualifying connection. If the connection is successful, the failed server is placed back on the list of active real servers. If the connection is unsuccessful, the server remains out of service and the retry timer is reset. The unsuccessful connection must have experienced at least one retry, otherwise the next qualifying connection is also sent to that failed server.

Backup Server Farms

A backup server farm is a server farm that can be used when none of the real servers defined in a primary server farm is available to accept new connections. When configuring backup server farms, keep in mind the following considerations:

- A server farm can act as both primary and backup at the same time.
- The same real server cannot be defined in both primary and backup at the same time.
- Both primary and backup require the same NAT configuration (none, client, server, or both). In addition, if NAT is specified, both server farms must use the same NAT pool.

Dynamic Feedback Protocol (DFP) Agent Subsystem Support

IOS SLB supports the DFP Agent Subsystem feature, also called global load balancing, which enables client subsystems other than IOS SLB to act as DFP agents. With the DFP Agent Subsystem, you can use multiple DFP agents from different client subsystems at the same time.

For more information about the DFP Agent Subsystem, refer to the *DFP Agent Subsystem* feature document for Cisco IOS Release 12.2(18)SXD.

DFP for Cisco IOS SLB

With IOS SLB DFP support, a DFP manager in a load-balancing environment can initiate a TCP connection with a DFP agent. Thereafter, the DFP agent collects status information from one or more real host servers, converts the information to relative weights, and reports the weights to the DFP manager. The DFP manager factors in the weights when load balancing the real servers. In addition to reporting at user-defined intervals, the DFP agent sends an early report if a sudden change occurs in a real server's status.

The weights calculated by DFP override the static weights you define using the **weight**command in server farm configuration mode. If DFP is removed from the network, IOS SLB reverts to the static weights.

You can define IOS SLB as a DFP manager, as a DFP agent for another DFP manager, or as both at the same time. In such a configuration, IOS SLB sends periodic reports to the other DFP manager, which uses the information to choose the best server farm for each new connection request. IOS SLB then uses the same information to choose the best real server within the chosen server farm.

DFP also supports the use of multiple DFP agents from different client subsystems (such as IOS SLB and GPRS) at the same time.

- DFP and GPRS Load Balancing, page 17
- DFP and the Home Agent Director, page 17

DFP and GPRS Load Balancing

In GPRS load balancing, you can define IOS SLB as a DFP manager and define a DFP agent on each GGSN in the server farm. Thereafter, the DFP agent can report the weights of the GGSNs. The DFP agents calculate the weight of each GGSN based on CPU use, processor memory, and the maximum number of Packet Data Protocol (PDP) contexts (mobile sessions) that can be activated for each GGSN. As a first approximation, DFP calculates the weight as the number of existing PDP contexts divided by the maximum allowed PDP contexts:

(existing PDP contexts)/(maximum PDP contexts)

Maximum PDP contexts are specified using the **gprs maximum-pdp-context-allowed** command, which defaults to 10,000 PDP contexts. If you accept the default value, DFP might calculate a very low weight for the GGSN:

(existing PDP contexts)/10000 = Low GGSN weight

When you specify maximum PDP contexts using the **gprs maximum-pdp-context-allowed** command, keep this calculation in mind. For example, Cisco 7200 series routers acting as GGSNs are often configured with a maximum of 45,000 PDP contexts.

DFP and the Home Agent Director

When using the Home Agent Director, you can define IOS SLB as a DFP manager and define a DFP agent on each home agent in the server farm, and the DFP agent can report the weights of the home agents. The DFP agents calculate the weight of each home agent based on CPU use, processor memory, and the maximum number of bindings that can be activated for each home agent:

(maximum-number-of-bindings - current-number-of-bindings)/maximum-number-of-bindings * (cpu-use + memory-use)/32 * maximum-DFP-weight = reported-weight

The *maximum-number-of-bindings* is 235,000. The *maximum-DFP-weight* is 24.

GGSN-IOS SLB Messaging

You can enable a GGSN to notify IOS SLB when certain conditions occur. The notifications enable IOS SLB to make intelligent decisions, which in turn improves GPRS load balancing and failure detection.

The notifications sent by the GGSN use GTP with message types from the unused space (reserved for future use) and the following information elements (IEs):

- Notification type, which indicates the notification condition. For example, this could be a notification to IOS SLB to reassign the session to an alternate GGSN, when the current GGSN fails on Call Admission Control (CAC).
- Identifier of the relevant session (session key).
- Other IEs specific to the notification type. For example, for a notification to reassign, GGSN includes the create response, which it would otherwise have sent to the SGSN. This enables IOS SLB to relay this response back to SGSN when the maximum number of reassignments due to notification reach the configured limit.

GGSN-IOS SLB messaging is supported in both dispatched mode and directed modes.

INOP_REAL State for Virtual Servers

You can configure a virtual server such that, if all of the real servers that are associated with the virtual server are inactive, the following actions occur:

- The virtual server is placed in the INOP_REAL state.
- An SNMP trap is generated for the virtual server's state transition.
- The virtual server stops answering ICMP requests.

For more information, see the description of the **inservice (server farm virtual server)** command in SLB server farm virtual server configuration mode.

Probes

Probes determine the status of each real server in a server farm, or each firewall in a firewall farm. The Cisco IOS SLB feature supports DNS, HTTP, ping, TCP, custom UDP, and WSP probes:

- A DNS probe sends domain name resolve requests to real servers, and verifies the returned IP addresses.
- An HTTP probe establishes HTTP connections to real servers, sends HTTP requests to the real servers, and verifies the responses. HTTP probes are a simple way to verify connectivity for devices that are server load-balanced, and for firewalls being firewall load-balanced (even devices on the other side of a firewall).

HTTP probes also enable you to monitor applications that are server load-balanced. With frequent probes, the operation of each application is verified, not just connectivity to the application.

HTTP probes do not support HTTP over Secure Socket Layer (HTTPS). That is, you cannot send an HTTP probe to an SSL server.

- A ping probe pings real servers. Like HTTP probes, ping probes are a simple way to verify connectivity for devices and firewalls being load-balanced.
- A TCP probe establishes and removes TCP connections. Use TCP probes to detect failures on TCP port 443 (HTTPS).
- A custom UDP probe can to support a variety of applications and protocols, including:

  - RADIUS Accounting/Authorization probes
  - GTP Echo probes
  - Connectionless WSP probes
  - XML-over-UDP probes for CSG user-database load-balancing
  - Mobile IP RRQ/RRP
- A WSP probe simulates requests for wireless content and verifies the retrieved content. Use WSP probes to detect failures in the Wireless Application Protocol (WAP) stack on port 9201.

You can configure more than one probe, in any combination of supported types, for each server farm, or for each firewall in a firewall farm.

You can also flag a probe as a routed probe, with the following considerations:

- Only one instance of a routed probe per server farm can run at any given time.
- Outbound packets for a routed probe are routed directly to a specified IP address.

IOS SLB probes use the SA Agent. You might want to specify the amount of memory that the SA Agent can use, using the **rtr low-memory** command. If the amount of available free memory falls below the value specified in the **rtr low-memory** command, then the SA Agent does not allow new operations to be configured. For more details, see the description of the **rtr low-memory**command in the *Cisco IOS IP SLAs Command Reference*.

- Probes in Server Load Balancing, page 19
- Probes in Firewall Load Balancing, page 19

Probes in Server Load Balancing

Probes determine the status of each real server in a server farm. All real servers associated with all virtual servers tied to that server farm are probed.

If a real server fails for one probe, it fails for all probes. After the real server recovers, all probes must acknowledge its recovery before it is restored to service.

Note If a probe is configured for stateful backup and a failover occurs, the change in status (from backup to active) is reflected accurately in the probe in the new active IOS SLB device. However, the probe in the new backup IOS SLB device (which had been the active device before the failover) still shows its status as active.

Probes in Firewall Load Balancing

Probes detect firewall failures. All firewalls associated with the firewall farm are probed.

If a firewall fails for one probe, it is failed for all probes. After the firewall recovers, all probes must acknowledge its recovery before the probe is restored to service.

To prevent password problems, make sure you configure the HTTP probe to expect status code 401. For more details, see the description of the **expect**command.

Use the **ip http server** command to configure an HTTP server on the device. For more details, see the description of the **ip http server** command in the *Cisco IOS Configuration Fundamentals Command Reference*.

In a transparent web cache load-balancing environment, an HTTP probe uses the real IP address of the web cache, since there is no virtual IP address configured.

Protocol Support Features

Protocol Support

IOS SLB supports the following protocols:

- Access Service Network (ASN)
- Domain Name System (DNS)
- Encapsulation Security Payload (ESP)
- File Transfer Protocol (FTP)
- Generic Routing Encapsulation (GRE)
- GPRS Tunneling Protocol v0 (GTP v0)
- GPRS Tunneling Protocol v1 (GTP v1)
- GPRS Tunneling Protocol v2 (GTP v2)
- Hypertext Transfer Protocol (HTTP)
- Hypertext Transfer Protocol over Secure Socket Layer (HTTPS)
- Internet Message Access Protocol (IMAP)
- Internet Key Exchange (IKE, was ISAKMP)
- IP in IP Encapsulation (IPinIP)
- Mapping of Airline Traffic over IP, Type A (MATIP-A)
- Network News Transport Protocol (NNTP)
- Post Office Protocol, version 2 (POP2)
- Post Office Protocol, version 3 (POP3)
- RealAudio/RealVideo through RTSP
- Remote Authentication Dial-In User Service (RADIUS)
- Simple Mail Transport Protocol (SMTP)
- Telnet
- Transmission Control Protocol (TCP) and standard TCP protocols
- User Datagram Protocol (UDP) and standard UDP protocols
- X.25 over TCP (XOT)
- Wireless Application Protocol (WAP), including:
    - Connectionless Secure WSP
    - Connectionless WSP
    - Connection-Oriented Secure WSP
    - Connection-Oriented WSP

AAA Load Balancing

IOS SLB provides RADIUS load-balancing capabilities for RADIUS authentication, authorization, and accounting (AAA) servers.

IOS SLB provides the following RADIUS load-balancing functions:

- Balances RADIUS requests among available RADIUS servers and proxy servers.
- Routes RADIUS request retransmissions (such as retransmissions of unanswered requests) to the same RADIUS server or proxy server as the original request.
- Provides session-based automatic failure detection.
- Supports both stateless backup and stateful backup.

In addition, IOS SLB can load-balance devices that proxy the RADIUS Authorization and Accounting flows in both traditional and mobile wireless networks. For more information, see the "RADIUS Load Balancing" section.

Audio and Video Load Balancing

IOS SLB can balance RealAudio and RealVideo streams through Real-Time Streaming Protocol (RTSP), for servers running RealNetworks applications.

VPN Server Load Balancing

IOS SLB can balance Virtual Private Network (VPN) flows, including the following flows:

- IP Security (IPSec) flows. An IPSec flow consists of a UDP control session and an ESP tunnel.
- Point-to-Point Tunneling Protocol (PPTP) flows. A PPTP flow consists of a TCP control session and a GRE tunnel.

Redundancy Features

An IOS SLB device can represent one point of failure, and the servers can lose their connections to the backbone, if either of the following occurs:

- The IOS SLB device fails.
- A link from a switch to the distribution-layer switch becomes disconnected.

To reduce that risk, IOS SLB supports the following redundancy enhancements, based on HSRP:

- Stateless Backup, page 21
- Stateful Backup, page 21
- Active Standby, page 22

Stateless Backup

Stateless backup provides high network availability by routing IP flows from hosts on Ethernet networks without relying on the availability of one Layer 3 switch. Stateless backup is particularly useful for hosts that do not support a router discovery protocol (such as the Intermediate System-to-Intermediate System [IS-IS] Interdomain Routing Protocol [IDRP]) and do not have the functionality to shift to a new Layer 3 switch when their selected Layer 3 switch reloads or loses power.

Stateful Backup

Stateful backup enables IOS SLB to incrementally backup its load-balancing decisions, or "keep state," between primary and backup switches. The backup switch keeps its virtual servers in a dormant state until HSRP detects failover; then the backup (now primary) switch begins advertising virtual addresses and processing flows. You can use HSRP to configure a timer for failure detection.

Stateful backup provides IOS SLB with a one-to-one stateful or idle backup scheme. This means that only one instance of IOS SLB is handling client or server flows at a given time, and that there is at most one backup platform for each active IOS SLB switch.

The Home Agent Director do not support stateful backup.

Note If a probe is configured for stateful backup and a failover occurs, the change in status (from backup to active) is reflected accurately in the probe in the new active IOS SLB device. However, the probe in the new backup IOS SLB device (which had been the active device before the failover) still shows its status as active.

Active Standby

Active standby enables two IOS SLBs to load-balance the same virtual IP address while at the same time acting as backups for each other. If a site has only one virtual IP address to load-balance, an access router is used to direct a subset of the flows to each IOS SLB using policy-based routing.

IOS SLB firewall load balancing supports active standby. That is, you can configure two pairs of firewall load balancing devices (one pair on each side of the firewalls), with each device in each pair handling traffic and backing up its partner.

Exchange Director Features

IOS SLB supports the Exchange Director for the mobile Service Exchange Framework (mSEF) for CiscoCisco 7600 series routers. The Exchange Director provides the following features:

- ASN Load Balancing, page 22
- GPRS Load Balancing, page 23
- Dual-Stack Support for GTP Load Balancing, page 25
- Home Agent Director, page 25
- KeepAlive Application Protocol (KAL-AP) Agent Support, page 25
- RADIUS Load Balancing, page 27
- RADIUS Load Balancing Accelerated Data Plane Forwarding, page 28
- WAP Load Balancing, page 29
- Stateful Backup of Redundant Route Processors, page 29
- Flow Persistence, page 29

ASN Load Balancing

IOS SLB can provide load balancing across a set of Access Service Network (ASN) gateways. The gateway server farm appears to the base station as one ASN gateway.

When a Mobile Subscriber Station (MSS) wants to enter the network, the base station sends a Mobile Station Pre-Attachment request to the virtual IP address of the IOS SLB. IOS SLB selects an ASN gateway and forwards the request to that gateway. The gateway responds directly to the base station with a Mobile Station Pre-Attachment response. If configured to do so, the base station then returns a Mobile Station Pre-Attachment ACK to IOS SLB, which forwards the ACK to the selected gateway. Thereafter, all subsequent transactions flow between the base station and the gateway.

If sticky connections are enabled for the ASN gateways, IOS SLB makes a load-balancing decision once for a subscriber and then forwards all subsequent requests from the same subscriber to the same Cisco Broadband Wireless Gateway (BWG). The sticky information is replicated to the standby IOS SLB.

IOS SLB populates the sticky database with Mobile Stations IDs (MSIDs), with one sticky entry for each MSS. The sticky database enables IOS SLB to perform persistent session tracking of the real server selected for the MSID. The first packet sent to a virtual IP address from an MSS creates the session object and the sticky object. Subsequent packets from the MSS use the MSID to find the real server in the sticky database, if the session lookup fails. All packets that belong to a given MSS are load-balanced to same BWG as long as the sticky object exists.

Redundancy support has been provided by replicating the sticky MSID entries to the backup IOS SLB. Redundancy works in both the intra-chassis (stateful switchover) and inter-chassis (HSRP) environments. Sessions need not be replicated to the standby IOS SLB.

GPRS Load Balancing

GPRS is the packet network infrastructure based on the European Telecommunications Standards Institute (ETSI) Global System for Mobile Communication (GSM) phase 2+ standards for transferring packet data from the GSM mobile user to the packet data network (PDN). The Cisco gateway GPRS support node (GGSN) interfaces with the serving GPRS support node (SGSN) using the GTP, which in turn uses UDP for transport. IOS SLB provides GPRS load balancing and increased reliability and availability for the GGSN.

When configuring the network shared by IOS SLB and the GGSNs, keep the following considerations in mind:

- Specify static routes (using **ip route** commands) and real server IP addresses (using **real** commands) such that the Layer 2 information is correct and unambiguous.
- Choose subnets carefully, using one of the following methods:
  - Do not overlap virtual template address subnets.
  - Specify next hop addresses to real servers, not to interfaces on those servers.
- IOS SLB assigns all PDP context creates from a specific IMSI to the same GGSN.
- IOS SLB supports GTP Version 0 (GTP v0), Version 1 (GTP v1), and Version 2 (GTP v2). Support for GTP enables IOS SLB to become “GTP aware,” extending IOS SLB knowledge into Layer 5.
- GPRS load balancing maps enable IOS SLB to categorize and route user traffic based on access point names (APNs).

IOS SLB supports two types of GPRS load balancing:

- GPRS Load Balancing without GTP Cause Code Inspection, page 23
- GPRS Load Balancing with GTP Cause Code Inspection, page 24

GPRS Load Balancing without GTP Cause Code Inspection

GPRS load balancing *without* GTP cause code inspection enabled is recommended for Cisco GGSNs. It has the following characteristics:

- Can operate in dispatched mode or in directed server NAT mode, but not in directed client NAT mode. In dispatched mode, the GGSNs must be Layer 2-adjacent to the IOS SLB device.
- Supports stateful backup only if sticky connections are enabled. See the “Stateful Backup” section for more information.

- Delivers tunnel creation messages destined to the virtual GGSN IP address to one of the real GGSNs, using the weighted round robin load-balancing algorithm. See the “Weighted Round Robin Algorithm” section for more information about this algorithm.
- Requires DFP in order to account for secondary PDP contexts in GTP v1 and GTP v2.

GPRS Load Balancing with GTP Cause Code Inspection

GPRS load balancing *with* GTP cause code inspection enabled allows IOS SLB to monitor all PDP context signaling flows to and from GGSN server farms. This enables IOS SLB to monitor GTP failure cause codes, detecting system-level problems in both Cisco and non-Cisco GGSNs.

The table below lists the PDP create response cause codes and the corresponding actions taken by IOS SLB.

Table 1: PDP Create Response Cause Codes and Corresponding IOS SLB Actions

Cause Code	IOS SLB Action
Request Accepted	Establish session
No Resource Available	Fail current real, reassign session, drop the response
All dynamic addresses are occupied	Fail current real, reassign session, drop the response
No memory is available	Fail current real, reassign session, drop the response
System Failure	Fail current real, reassign session, drop the response
Missing or Unknown APN	Forward the response
Unknown PDP Address or PDP type	Forward the response
User Authentication Failed	Forward the response
Semantic error in TFT operation	Forward the response
Syntactic error in TFT operation	Forward the response
Semantic error in packet filter	Forward the response
Syntactic error in packet filter	Forward the response
Mandatory IE incorrect	Forward the response
Mandatory IE missing	Forward the response
Optional IE incorrect	Forward the response
Invalid message format	Forward the response
Version not supported	Forward the response

GPRS load balancing *with* GTP cause code inspection enabled has the following characteristics:

- Must operate in directed server NAT mode.
- Supports stateful backup. See the “Stateful Backup” section for more information.

- Tracks the number of open PDP contexts for each GGSN, which enables GGSN server farms to use the weighted least connections (**leastconns**) algorithm for GPRS load balancing. See the "Weighted Least Connections Algorithm" section for more information about this algorithm.
- Enables IOS SLB to deny access to a virtual GGSN if the carrier code of the requesting International Mobile Subscriber ID (IMSI) does not match a specified value.
- Enables IOS SLB to account for secondary PDP contexts even without DFP.

Dual-Stack Support for GTP Load Balancing

IPv6 support enables IOS SLB to manage IPv6 addresses for GTP load balancing, for all versions of GTP (v0, v1, v2).

Dual-stack support enables IOS SLB to manage dual-stack implementations for GTP load balancing. A dual stack implementation is one that uses both IPv4 and IPv6 addresses.

When configuring dual-stack support for GTP load balancing, keep the following considerations in mind:

- The real server must be configured as a dual-stack real server, with the IPv4 and IPv6 addresses, using the **real** command in SLB server farm configuration mode.
- The virtual server must be configured as a dual-stack virtual server, with the virtual IPv4 and IPv6 addresses and the optional IPv6 prefix, using the **virtual** command in SLB virtual server configuration mode.
- To specify the primary IPv6 server farm and optional backup IPv6 server farm, use the **serverfarm** command in SLB virtual server configuration mode.
- The **client** command in SLB virtual server configuration mode is not supported.
- The gateway must be configured with the IPv4 and IPv6 addresses of the virtual server.
- The interface between IOS SLB and the gateway must be configured with dual-stack addresses.
- All HSRP instances (both IPv4 and IPv6) for the client-facing interface must have the same HSRP state.

Home Agent Director

The Home Agent Director load balances Mobile IP Registration Requests (RRQs) among a set of home agents (configured as real servers in a server farm). Home agents are the anchoring points for mobile nodes. Home agents route flows for a mobile node to its current foreign agent (point of attachment).

The Home Agent Director has the following characteristics:

- Can operate in dispatched mode or in directed server NAT mode, but not in directed client NAT mode. In dispatched mode, the home agents must be Layer 2-adjacent to the IOS SLB device.
- Does not support stateful backup. See the "Stateful Backup" section for more information.
- Delivers RRQs destined to the virtual Home Agent Director IP address to one of the real home agents, using the weighted round robin load-balancing algorithm. See the "Weighted Round Robin Algorithm" section for more information about this algorithm.
- Requires DFP in order to allocate RRQs based on capacity.

For more information about Mobile IP, home agents, and related topics, refer to the *Cisco IOS IP Configuration Guide*, Release 12.2.

KeepAlive Application Protocol (KAL-AP) Agent Support

Support for the KeepAlive Application Protocol (KAL-AP) agent support enables IOS SLB to perform load balancing in a global server load balancing (GSLB) environment. KAL-AP provides load information

along with its keepalive response message to the KAL-AP manager or GSLB device, such as the Global Site Selector (GSS), and helps the GSLB device load-balance client requests to the least-loaded IOS SLB devices.

When configuring KAL-AP agent support for IOS SLB, keep the following considerations in mind:

- KAL-AP agent support automatically detects the Virtual Private Network (VPN) routing and forwarding (VRF) ID of an incoming request packet, and uses the same VRF ID when sending a response.
- A client that uses DNS caching might contact IOS SLB directly, instead of sending requests through the GSS. Therefore, configure the DNS setting in the client to avoid such a situation.

KAL-AP calculates the load value in one of two ways: relatively or absolutely. (IOS SLB CPU/memory load might affect the final KAL-AP load value.)

- Relative KAL-AP Load Value, page 26
- Absolute KAL-AP Load Value, page 26

Relative KAL-AP Load Value

If the **farm-weight**command is not configured in server farm configuration mode, or if DFP is not enabled for the IOS SLB, KAL-AP calculates a relative load value, using the following formula:

KAL-AP Load = **256** - (number-*of-active-real-servers* * **256**/*number-of-inservice-real-servers*)

For example, if a site is provisioned with two real servers, and both real servers are inservice but only one is currently active, the resulting KAL-AP load value for that site is:

KAL-AP Load = **256** - (**1*** **256**/**2**) = **256** - **128** = **128**

Absolute KAL-AP Load Value

If the **farm-weight**command is configured in server farm configuration mode, and DFP is enabled for the IOS SLB, KAL-AP calculates an absolute load value, using the following formula:

KAL-AP Load = **256** - (*sum-of-max-dfp-weights-of-real-servers* * **256**/*farm-weight*)

Note The maximum DFP weight for a real server is configured using the **gprs dfp max-weight**command in global configuration mode. However, the actual maximum DFP weight reported to KAL-AP is proportional to the load on the GGSN. For example, if a GGSN is configured with a maximum DFP weight of 100, but the GGSN is 50 percent loaded, it reports a maximum DFP weight of 50 to KAL-AP. If the DFP connection to the real server is down, KAL-AP uses the setting of the **weight** command in SLB real server configuration mode. If no **weight** command is configured for the real server, KAL-AP uses the default weight of 8.

For example, consider a site with the following settings:

- A server farm configured with a farm weight of 200.
- GGSN-1 configured with a maximum DFP weight of 100, 0 percent loaded (so it reports a DFP weight of 100).
- GGSN-2 configured with a maximum DFP weight of 100, 50 percent loaded (so it reports a DFP weight of only 50).

The resulting KAL-AP load value for that site is:

KAL-AP Load = **256** - [(**100** + **50**)* **256**/**200**] = **256** - **192** = **64**

For best results, configure a **farm-weight** that is equal to the sum of the maximum DFP weights for the real servers in the server farm. For example, if there are three real servers in a server farm, configured with maximum DFP weights of 100, 50, and 50, then configure a **farm-weight** of 200 (that is, 100 + 50 + 50). If a real server is added to or removed from the server farm, you must adjust the **farm-weight** accordingly.

RADIUS Load Balancing

IOS SLB provides RADIUS load-balancing capabilities for RADIUS servers. In addition, IOS SLB can load-balance devices that proxy the RADIUS Authorization and Accounting flows in both traditional and mobile wireless networks, if desired. IOS SLB does this by correlating data flows to the same proxy that processed the RADIUS for that subscriber flow.

IOS SLB provides RADIUS load balancing in mobile wireless networks that use service gateways, such as the Cisco Service Selection Gateway (SSG) or the Cisco Content Services Gateway (CSG). The following mobile wireless networks are supported:

- GPRS networks. In a GPRS mobile wireless network, the RADIUS client is typically a GGSN.
- Simple IP CDMA2000 networks. CDMA2000 is a third-generation (3-G) version of Code Division Multiple Access (CDMA). In a simple IP CDMA2000 mobile wireless network, the RADIUS client is a Packet Data Service Node (PDSN).
- Mobile IP CDMA2000 networks. In a Mobile IP CDMA2000 mobile wireless network, both the Home Agent (HA) and the PDSN/Foreign Agent (PDSN/FA) are RADIUS clients.

IOS SLB provides the following RADIUS load-balancing functions:

- Balances RADIUS requests among available RADIUS servers and proxy servers.
- Routes RADIUS request retransmissions (such as retransmissions of unanswered requests) to the same RADIUS server or proxy server as the original request.
- Routes all of a subscriber's RADIUS flows, as well as all non-RADIUS data flows for the same subscriber, to the same service gateway.
- Supports multiple service gateway server farms (for example, one farm of SSGs and another of CSGs). IOS SLB examines the input interface in a packet to route it to the correct service gateway server farm.
- Supports multiple WAP gateway server farms behind a RADIUS load balancing virtual server, using RADIUS calling station IDs and usernames to select specific server farms. This enhancement enables RADIUS load balancing on both the control plane and the data plane. RADIUS load balancing on the control plane enables RADIUS messages to be load-balanced to AAA servers for subscriber authorization, authentication and accounting. RADIUS load balancing on the data plane enables data flows for a subscriber to maintain a consistent network path to the destination network device. In addition, the RADIUS virtual server can acknowledge RADIUS accounting messages and build or delete sticky objects, rather than having to forward the messages to the specified server.
- Can route data packets to a real server in the CSG farm, then to a real server in the SSG farm.
- Routes RADIUS Accounting-Request messages from a RADIUS client to the service gateway that processed the RADIUS Access-Request message for the subscriber. The service gateway can then clean up the host entry it has created for the subscriber.
- Uses the weighted round robin load-balancing algorithm. See the "Weighted Round Robin Algorithm" section for more information about this algorithm.
- Facilitates SSG single sign-on through the RADIUS protocol.
- Provides session-based automatic failure detection.
- Supports both stateless backup and stateful backup.

To perform RADIUS load balancing, IOS SLB uses the following RADIUS sticky databases:

- The IOS SLB RADIUS framed-IP sticky database associates each subscriber's IP address with a specific service gateway. In a GPRS mobile wireless network, IOS SLB uses the RADIUS framed-IP sticky database to route packets correctly.

Note Subscriber IP addresses are assigned by service gateways or by RADIUS clients. If subscriber IP addresses are assigned from disjoint per-service gateway pools (so that the next-hop service gateway can be chosen based on the source IP address), IOS SLB can use policy routing to route subscriber flows.

- The IOS SLB RADIUS calling-station-ID sticky database associates each subscriber's calling station ID with a specific service gateway.
- The IOS SLB RADIUS username sticky database associates each subscriber's username with a specific service gateway.
- RADIUS load balancing maps enable IOS SLB to categorize and route user traffic based on RADIUS calling station IDs and usernames. RADIUS load balancing maps is mutually exclusive with Turbo RADIUS load balancing and RADIUS load balancing accounting local acknowledgement.
- RADIUS load balancing accounting local acknowledgement:
  - Enables IOS SLB to respond to RADIUS accounting packets with an ACK response while maintaining sticky objects for those sessions.
  - Is mutually exclusive with RADIUS load balancing maps and Turbo RADIUS load balancing.
- In a CDMA2000 mobile wireless network, to route packets correctly, IOS SLB requires both the RADIUS framed-IP sticky database and either the RADIUS username sticky database or the RADIUS calling-station-ID sticky database.
- The IOS SLB RADIUS International Mobile Subscriber ID (IMSI) sticky database maps the IMSI address for each user to the corresponding gateway. This enables IOS SLB to forward all subsequent flows for the same user to the same gateway.

RADIUS Load Balancing Accelerated Data Plane Forwarding

RADIUS load balancing accelerated data plane forwarding, also known as Turbo RADIUS load balancing, is a high-performance solution that uses basic policy-based routing (PBR) route maps to manage subscriber data-plane traffic in a Cisco Content Services Gateway (CSG) environment.

When Turbo RADIUS load balancing receives a RADIUS payload, it takes the following actions:

1. Inspects the payload.
2. Extracts the framed-IP attribute.
3. Applies a route map to the IP address.
4. Determines which CSG is to manage the subscriber.

If vendor-specific attribute (VSA) correlation is configured, and if the Cisco VSA is buffered, then the Cisco VSA is injected into the RADIUS Accounting-Start packet.

Turbo RADIUS load balancing does not require VSA correlation, but it does require a server farm configured with **predictor route-map** on the accounting virtual server.

Note When you specify the **predictor route-map** command in SLB server farm configuration mode, no further commands in SLB server farm configuration mode or real server configuration mode are allowed.

For more information about policy-based routing, see the “Policy-Based Routing” and “Configuring Policy-Based Routing” sections of the *Cisco IOS IP Routing Configuration Guide* .

In a mobile Service Exchange Framework (mSEF) environment, Turbo RADIUS load balancing does not require firewall load balancing on the network side of the CSG cluster. (Standard RADIUS load balancing does require firewall load balancing on the network side of the cluster.)

Turbo RADIUS load balancing:

- Supports simple IP access control lists (ACLs) and match and set next-hop pairs.
- Is mutually exclusive with RADIUS load balancing maps and RADIUS load balancing accounting local acknowledgement.
- Is mutually exclusive with the optional ACL logging facility. In order to use Turbo RADIUS load balancing, you must first disable the logging facility. For more information, see the description of the **access-list (IP standard)** command in the *Cisco IOS Security Command Reference* , Cisco IOS 12.4.

WAP Load Balancing

You can use IOS SLB to load-balance wireless session protocol (WSP) sessions among a group of WAP gateways or servers on an IP bearer network. WAP runs on top of UDP on a set of well known ports, with each port indicating a different WAP mode:

- Connectionless WSP mode (IP/UDP [9200]/WSP). In connectionless WSP mode, WSP is a simple 1-request/1-response protocol in which one server-bound packet results in a server response of one or more packets.
- Connection-oriented WSP mode (IP/UDP [9201]/WTP/WSP). In connection-oriented WSP mode, WTP manages retransmissions of WDP events, and WSP operates using a defined session bring-up/tear-down sequence. IOS SLB uses a WAP-aware finite state machine (FSM), driven by events in WSP sessions, to reassign sessions. This FSM operates only on port 9201, where the WSP sessions are not encrypted and WTP manages retransmissions.
- Connectionless secure WSP mode (IP/UDP [9202]/WTLS/WSP). This mode functions the same as connectionless WSP mode, but with security provided by WTLS.
- Connection-oriented secure WSP mode (IP/UDP [9203]/WTLS/WTP/WSP). This mode functions the same as connection-oriented WSP mode, but with security provided by WTLS.

IOS SLB uses WSP probes to detect failures in the WAP stack on port 9201.

Stateful Backup of Redundant Route Processors

When used with RPR+, IOS SLB supports the stateful backup of redundant route processors for mSEF for the CiscoCisco 7600 routers. This enables you to deploy Cisco Multiprocessor WAN Application Modules (MWAMs) in the same chassis as IOS SLB, while maintaining high availability of load-balancing assignments.

Flow Persistence

Flow persistence provides intelligent return routing of load-balanced IP flows to the appropriate node, without the need for coordinated hash mechanisms on both sides of the load-balanced data path, and without using Network Address Translation (NAT) or proxies to change client or server IP addresses.

Restrictions for Cisco IOS SLB

General Restrictions

- *Does not support load balancing of flows between clients and real servers that are on the same local-area network (LAN) or virtual LAN (VLAN). The packets being load-balanced cannot enter and leave the load-balancing device on the same interface.*
- You cannot configure IOS SLB from different user sessions at the same time.
- Do not configure an IOS SLB virtual IP address on the same subnet as any real server IP address, unless all server farms that include the real server IP address are configured using the **nat server** command.
- Operates in a standalone mode and currently does not operate as a MultiNode Load Balancing (MNLB) Services Manager. Does not support IOS SLB and MNLB configured with the same virtual IP address, even if they are for different services. The presence of IOS SLB does not preclude the use of the existing MNLB Forwarding Agent with an external Services Manager (such as the LocalDirector) in an MNLB environment. (MNLB is sometimes called Cisco Application Services Architecture, or CASA.)
- Does not support coordinating server load-balancing statistics among different IOS SLB instances for backup capability.
- Supports FTP and firewall load balancing only in dispatched mode.
- Does not support Dynamic Host Configuration Protocol (DHCP) load balancing.
- Does not support Internet Protocol version 6 (IPv6).
- When operating in dispatched mode, real servers must be Layer 2-adjacent, tag-switched, or through a GRE tunnel.

When operating in directed mode with server NAT, real servers need not be Layer 2-adjacent to IOS SLB. This function allows for more flexible network design, because servers can be placed several Layer 3 hops away from the IOS SLB switch.

- When operating in directed mode as a member of a multicast group, IOS SLB can receive multicast flows but cannot send multicast flows. This is not a restriction when operating in dispatched mode.
- Supports client Network Address Translation (NAT) and server port translation for TCP and UDP virtual servers only.
- When balancing streams to a virtual IP address that is the same as one of the IOS interface IP addresses (loopback, Ethernet, and so on), IOS SLB treats all UDP packets to that address as traceroute packets and replies with “host unreachable” ICMP packets. This occurs even if the IOS listens to the target UDP port. To avoid this issue, configure the virtual server as a network (address/31), not as a host (address/32).
- Do not use the virtual IP address configured in the IOS SLB virtual server for UDP-based router management applications such as SNMP. Doing so can result in high CPU usage. (This is not a problem for a UDP virtual server that is configured with destination port number 0.)
- The DFP agent requires a delay between hello messages of at least 3 seconds. Therefore, if your DFP manager provides a timeout specification, you must set the timeout to at least 3 seconds.
- When both IOS SLB and the Web Cache Communication Protocol (WCCP) are configured on a Cisco Catalyst 6500 series switch, and WCCP Input Redirection is configured with IOS SLB, Layer 2 WCCP forwarding must be used between the router and the cache. In this case, WCCP and IOS SLB both run in hardware and are processed in the correct order. If Generic Routing Encapsulation (GRE) forwarding is used, then IOS SLB takes precedence over WCCP and there is no redirection, because GRE forwarding is done on the MSFC. Note that the WCCP forwarding method, either Layer 2 or GRE, is configured on the cache engine and not on the switch.

- Do not configure IOS SLB and a Cisco Service Selection Gateway (SSG) on the same device.
- For “sandwich” configurations (that is, those that require an IOS SLB on both sides of a farm of CSGs, SSGs, or firewalls), if a flow is to be directed through two IOS SLB instances (virtual servers or firewall farms), the IOS SLB instances must reside in different Virtual Private Network (VPN) routing and forwardings (VRFs).
- If you do not configure an access interface using the **access** command in server farm, virtual server, or firewall farm configuration mode, IOS SLB installs the wildcards for the server farm, virtual server, or firewall farm in all of the available interfaces of the device, including the VRF interfaces. If IOS SLB is not required on the VRF interfaces, use the **access** command to limit wildcards to the specified interfaces only.
- VRF-aware IOS SLB does not operate “between” VRFs. That is, the server farm interface and the client traffic interface must use the same VRFs.

Static NAT Restrictions

- Does not work with client NAT server farms. That is, if a real server is using a virtual IP address for server NAT, and a server farm is associated with that same virtual IP address, then you cannot configure the server farm to use client NAT.
- Requires that each real server be associated with only one virtual server, to ensure that IOS SLB can create connections correctly.
- Requires a 0-port virtual server.
- Does not support virtual service FTP.
- Static NAT with per-packet server load balancing does not load-balance fragmented packets.

Backup Server Farm Restrictions

- Does not support defining the same real server in both primary and backup server farms.
- Requires the same NAT configuration (none, client, server, or both) for both primary and backup server farms. In addition, if NAT is specified, both server farms must use the same NAT pool.
- Does not support HTTP redirect load balancing. If a primary server farm specifies a redirect virtual server, you cannot define that primary as a backup, nor can you define a backup for that primary.

Firewall Load Balancing Restrictions

- Is not limited to one firewall farm in each load-balancing device.
- Requires that each firewall must have its own unique MAC address and must be Layer 2-adjacent to each device. The firewalls can be connected to individual interfaces on the device, or they can all share a VLAN and connect using one interface.
- Requires an Ethernet interface between each firewall load-balancing device and each firewall.
- On each firewall load-balancing device, IOS SLB requires that each Layer 2 firewall be connected to one Layer 3 (IP) interface.
- Flows with a destination IP address on the same subnet as the configured firewall IP addresses are not load-balanced. (Such flows could be a firewall console session or other flows on the firewall LAN.)
- Does not support the following IOS SLB functions:
    - NAT
    - Port-bound servers
    - SynGuard
    - TCP session reassignment
    - Transparent web cache load balancing

Restrictions for GPRS Load Balancing Without GPRS Tunneling Protocol (GTP) Cause Code Inspection Enabled

- If a real server is defined in two or more server farms, each server farm must be associated with a different virtual server.
- Operates in either dispatched or directed server NAT mode only.
- Supports stateful backup only if sticky connections are enabled.
- Cannot load-balance network-initiated PDP context requests.
- Does not support the following IOS SLB functions:
  - Bind IDs (A bind ID allows one physical server to be bound to multiple virtual servers and report a different weight for each.)
  - Client-assigned load balancing
  - Slow start
  - Weighted least connections load-balancing algorithm

Restrictions for GPRS Load Balancing With GTP Cause Code Inspection Enabled

- If a real server is defined in two or more server farms, each server farm must be associated with a different virtual server.
- Operates in directed server NAT mode only.
- Cannot load-balance network-initiated PDP context requests.
- Requires inbound and outbound signaling to flow through IOS SLB.
- Requires either the SGSN or the GGSN to echo its peer.
- Does not support the following IOS SLB functions:
  - Bind IDs
  - Client-assigned load balancing
  - Slow start

GTP v2 Restrictions

- Does not support client NAT.
- IOS SLB can balance GTP v2 control packets for packet data network gateways (PGWs) and for serving gateways (SGWs). If a PGW load-balancing device and an SGW load-balancing device are configured in the same Supervisor Engine, you must configure a separate virtual server for each device.
- IOS SLB checks for and processes only the following GTP v2 messages.:
  - GTP_CREATE_SESSION_REQ
  - GTP_ECHO_REQ
  - GTP_SLB_NOTIFICATION

All other messages are dropped.

- IOS SLB supports the following GTP_SLB notification messages:
  - GTP_SLB_NOTIF_REASSIGN_REAL
  - GTP_SLB_NOTIF_PDP_DELETION.
  - GTP_SLB_NOTIF_PDP_STATUS

VPN Server Load Balancing Restrictions

- Does not support Internet Control Message Protocol (ICMP) and wildcard (0-protocol) virtual servers.

RADIUS Load Balancing Accelerated Data Plane Forwarding Restrictions

- Requires the route map algorithm.
- Requires redundant CSGs for best results.
- Requires static provisioning of load distribution by subscriber address range.
- Supports only simple IP access control lists (ACLs).
- When VSA correlation is used, IOS SLB maintains the correlation information only in the active RADIUS load-balancing device, not in the backup RADIUS load-balancing device. The backup RADIUS load-balancing device does not receive VSA correlation information from the active RADIUS load-balancing device.
- All Accounting-Request and Access-Accept messages must include the RADIUS-assigned Framed-ip-address attribute. The source IP address for each subscriber flow must also match the value of the Framed-ip-address attribute in the Access-Accept message.
- RADIUS accounting must be enabled on the RADIUS client, which is typically a Network Access Server (NAS).
- When you specify the **predictor route-map** command in SLB server farm configuration mode, no further commands in SLB server farm configuration mode or real server configuration mode are allowed.

VSA Correlation Restrictions

- VSA correlation might result in a degradation of performance.
- IOS SLB maintains the correlation information only in the active RADIUS load-balancing device, not in the backup RADIUS load-balancing device. The backup RADIUS load-balancing device does not receive VSA correlation information from the active RADIUS load-balancing device.
- The Cisco VSA is injected into the RADIUS Accounting-Start packet. The Cisco VSA is not injected into any other RADIUS messages or packets, such as interim RADIUS Accounting On or Off messages or RADIUS Accounting-Stop packets.
- You cannot configure **radius inject acct** commands and **radius inject auth** commands on the same virtual server.

RADIUS Load Balancing for GPRS Restrictions

- Requires the weighted round robin algorithm.
- Does not support fragmented RADIUS packets.
- All Accounting-Request and Access-Accept messages must include the RADIUS-assigned Framed-ip-address attribute. The source IP address for each subscriber flow must also match the value of the Framed-ip-address attribute in the Access-Accept message.
- RADIUS accounting must be enabled on the RADIUS client, which is typically a Network Access Server (NAS).

RADIUS Load Balancing for CDMA2000 Restrictions

- Requires the weighted round robin algorithm.
- Does not support fragmented RADIUS packets.
- All subscribers on the mobile network must be assigned a unique IP address (that is, no overlapping IP addresses) which can be routed in the mobile wireless network.

- Each User-Name attribute must correspond to one subscriber, or at most to a very small number of subscribers. Otherwise, one SSG might be burdened with an unexpectedly large load.
- For simple IP networks, the following additional restrictions apply:
  - The PDSN must include the User-Name attribute in all RADIUS Access-Request and Accounting-Start packets. The value of the User-Name attribute for a subscriber must be the same in all the packets (except for Cisco PDSNs that provide MSID-based access).
  - The PDSN must include the Framed-ip-address attribute and the NAS-ip-address in all RADIUS Accounting-Start and Accounting-Stop packets. The value of the Framed-ip-address attribute must equal the source IP address in subscriber data packets routed by RADIUS load balancing for SSG service.
  - The PDSN must include the NAS-ip-address in all Accounting-Requests. For BSC/PCF hand-offs, the Accounting-Stop must include the 3GPP2-Session-Continue VSA with a value of **1**, to prevent the destruction of RADIUS load balancing sticky database objects for the subscriber.
- For Mobile IP networks, the following additional restrictions apply:
  - For a subscriber session, the PDSN and HA must send the RADIUS Access-Request and Accounting-Start packets with the User-Name attribute. The value of the User-Name attribute in all PDSN and HA RADIUS packets must be the same for the session.
  - For a subscriber session, the PDSN and HA must send RADIUS Accounting-Request packets with a Framed-ip-address attribute equal to the source IP address in subscriber data packets routed by RADIUS load balancing for SSG service. All RADIUS Accounting-Requests sent by the PDSN and HA must also include the NAS-ip-address attribute.
  - The PDSN must include the 3GPP2-Correlation-Identifier attribute in all Accounting-Requests.

Home Agent Director Restrictions

- A Registration Request (RRQ) must include the network access identifier (NAI) to be load-balanced.
- An RRQ must include a home agent IP address of either 0.0.0.0 or 255.255.255.255 to be load-balanced.
- For fast switching, the NAI in the RRQ cannot be more than 96 bytes deep in the packet. If the NAI is deeper than 96 bytes, IOS SLB manages the packet at the process level.
- Operates in either dispatched or directed server NAT mode only.
- Does not support the following IOS SLB functions:
  - Bind IDs
  - Client-assigned load balancing
  - Slow start
  - Stateful backup
  - Sticky connections
  - Weighted least connections load-balancing algorithm

Restrictions for HTTP Probes

- HTTP probes do not support HTTP over Secure Socket Layer (HTTPS). That is, you cannot send an HTTP probe to an SSL server.

Restrictions for UDP Probes

- UDP probes do not support fragmented Response packets.
- UDP probes do not support hosts that require a particular source port value in probe packets. UDP probes select an ephemeral port for each probe.

- Protocols and applications that have Message Digest Algorithm Version 5 (MD5) checksums generated from payload must be captured by a “sniffer” to obtain a correct checksum.
- For Cisco IOS Multiprotocol Label Switching (MPLS):
  - Clients can connect to IOS SLB through the MPLS cloud in a Supervisor Engine 720 environment.
  - The MPLS client interface must be configured with Tunnel Engineering. No other MPLS configuration is supported.
  - The MPLS client interface must receive packets as IP packets.
  - The MPLS client interface must be behind a Penultimate Hop Popping (PHP) router.
- For Cisco Catalyst 6500 series switches and Cisco 7600 series routers:
  - Supports Native Cisco IOS only (c6sup images). Native Cisco IOS requires the MSFC and the Policy Feature Card (PFC). When running redundant MSFCs in the same Catalyst 6500 switch, stateful backup between the two MSFCs is not supported, but stateless backup between the two MSFCs is supported.

The term MSFC refers to an MSFC1, MSFC2, or MSFC3, except when specifically differentiated.

The term PFC refers to a PFC1, PFC2, or PFC3, except when specifically differentiated.

- ◦ Requires that the Multilayer Switching (MLS) flow mode operate in full-flow mode or in interface full-flow mode. IOS SLB automatically sets the flow mode for its own use. For more information about how to set the MLS flow, refer to the *Catalyst 6000 Family Cisco IOS Software Configuration Guide* .
  - When operating in dispatched mode, real servers must be Layer 2-adjacent to IOS SLB (that is, not beyond an additional router), with hardware data packet acceleration performed by the PFC. All real servers in the same server farm must be on the same VLAN. The loopback address must be configured in the real servers.
  - Requires that all real servers in a firewall farm be on the same VLAN. Real servers in different firewall farms can be on different VLANs.
  - Provides no hardware data packet acceleration in directed mode. (Hardware data packet acceleration is performed by the PFC, and in directed mode the packets are managed by the MSFC, not the PFC.)
  - For the Cisco Supervisor Engine 2, “sandwich” configurations that require firewall load balancing are not supported, because such configurations require VRF. VRF is not supported for the Supervisor Engine 2.

ASN Release 6 Load Balancing Restrictions

- Operates in either dispatched or directed server NAT mode only. In directed mode, IOS SLB changes the destination IP address of the Mobile Station Pre-Attachment request to that of the selected Access Service Network (ASN) gateway real server.
- Requires DFP
- Does not support the following features:
  - Client NAT
  - Weighted least connections algorithm (for Mobile Station Pre-Attachment requests)
- When the base station is configured to send Mobile Station Pre-Attachment ACKnowledgement, or ACK, packets directly to an ASN gateway, bypassing IOS SLB, you must ensure that the session can time out without failing the real server. To do so, configure the **no faildetect inband** command real server configuration mode.
- For stateful backup and sticky connections:

- ASN sticky connections are supported only on the Cisco Broadband Wireless Gateway (BWG) Release 2.0 or later.
- If you are running ASN on a Cisco BWG, we recommend that you configure the **gw port** command in virtual server configuration mode.
- Do not use port number 2231 as the communication port between the Cisco BWG and the IOS SLB that is providing load balancing for the ASN.
- If you are not running ASN on a Cisco BWG, you must use the **sticky**command in virtual server configuration mod for sticky object deletion, since delete and network address identifier (NAI) update notifications on communication ports are not expected.
- To enable the Cisco BWG to send notifications for ASN to IOS SLB, configure the **wimax agw slb port** command in global configuration mode on the Cisco BWG.

Note Cisco BWG commands are documented in the *Cisco Broadband Wireless Gateway Command Reference* .

- To enable the Cisco BWG to send NAI-update notifications to IOS SLB when an MSID is registered, configure the **wimax agw slb notify nai-updates** command in global configuration mode on the Cisco BWG.
- To enable the Cisco BWG to send delete notifications to IOS SLB when an MSID is unregistered or deleted, configure the **wimax agw slb notify session-deletion** command in global configuration mode on the Cisco BWG.When you configure IOS SLB firewall load balancing, the load-balancing devices use route lookup to recognize flows destined for the firewalls. To enable route lookup, you must configure each device with the IP address of each firewall that will route flows to that device.