Rendezvous Points

Module 6

Module Objectives

Cisco.com

- Explain the basic operation of Auto-RP.
- Explain the basic operation of PIMv2 BSR.
- Explain how to configure RP's
- How to use various IOS commands to tune and control RP operation.

Module Agenda

Cisco.com

- **Auto RP**
- **PIMv2 BSR**
- **Static RPs**
- **Tuning RP Operations**
- **Debugging RP Operation**
- **Special Cases**

Auto-RP Overview

Cisco.com

- All routers automatically learn RP address
  - No configuration necessary except on:
    - Candidate RPs
    - Mapping Agents
- Makes use of Multicast to distribute info
  - Two specially IANA assigned Groups used
    - Cisco-Announce - 224.0.1.39
    - Cisco-Discovery - 224.0.1.40
  - Typically Dense mode is used forward these groups
- Permits backup RP’s to be configured
- Can be used with Admin-Scoping

- **Auto-RP Overview**
  - Auto-RP allows all routers in the network to automatically “learn” Group-to-RP mappings.
  - There are no special configuration steps that must be taken except on the router(s) that are to function as:
    - Candidate RP’s
    - Mapping Agents
  - Multicast is used to distribute Group-to-RP mapping information via two special, IANA assigned multicast groups.
    - Cisco-Announce Group - 224.0.1.39
    - Cisco-Discovery Group - 224.0.1.40
  - Because multicast is used to distribute this information, a “Chicken and Egg” situation can occur if the above groups operate in Sparse mode. (Routers would have to know a priori what the address of the RP is before they can learn the address of the RP(s) via Auto-RP messages.) Therefore, it is recommend that these groups *always* run in Dense mode so that this information is flooded throughout the network.
  - Multiple Candidate RP’s may be defined so that in the case of an RP failure, the other Candidate RP can assume the responsibility of RP.
  - Auto-RP can be configured to support Administratively Scoped zones. (BSR cannot!) This can be important when trying to prevent high-rate group traffic from leaving a campus and consuming too much bandwidth on WAN links.

Auto-RP Fundamentals

Cisco.com

- **Candidate RPs**
  - **Multicast RP-Announcement messages**
    - **Sent to Cisco-Announce (224.0.1.39) group**
    - **Sent every rp-announce-interval (default: 60 sec)**
  - **RP-Announcements contain:**
    - **Group Range (default = 224.0.0.0/4)**
    - **Candidate's RP address**
    - **Holdtime = 3 x <rp-announce-interval>**
  - **Configured via global config command**

```
ip pim send-rp-announce <intfc> scope <ttl> [group-list acl]
```

  - **'Deny' in group-list has variable meaning**
    - **Before 12.0(1.1) Deny = "I'm not C-RP for this group-range"**
    - **After 12.0(1.1) Deny = "Force group-range to always be DM"**

- **Auto-RP Candidate RP's (C-RP's)**
  - Multicast RP-Announcement messages to the Cisco-Announce (224.0.1.39) group. These messages "announce" this router as being a Candidate for selection as RP and are sent every 60 seconds by default.
  - RP-Announce messages contain:
    - Group Range (default is all multicast groups or 224.0.0.0/4)
    - The Candidate's IP address
    - A holdtime which is used to detect when the C-RP has failed. This holdtime is 3 times the announcement interval or 3x60 = 180 seconds = 3 minutes
  - C-RP's are configured using the (rather obtuse) command:

    ip pim send-rp-announce <intfc> scope <ttl> [group-list <acl>]

    - The <intfc> specifies which IP address is used as the source address in the RP-Announce messages that are sent out all multicast interfaces on the router.
    - The <ttl> value controls the TTL of the RP-Announce message.
    - The optional 'group-list' permits a group range other than the default to be assigned.
    - This command may be configured more than once on a router so that the router will function as C-RP for multiple group ranges.
  - Note: A 'deny' in the 'group-list' access-list has a different meaning beginning with IOS release 12.0(1.1).
    - Before 12.0.(1.1): Deny means "I'm not the RP for this group range."
    - After 12.0.(1.1): Deny means "Force this group range to always work in Dense mode. Note: Only a single C-RP needs to "deny" this group range to force this to happen. In other words, the 'deny' overrides any other router's "permit" advertisement.

Auto-RP Fundamentals

Cisco.com

- **Mapping agents**
  - **Receive RP-Announcements**
    - **Stored in Group-to-RP Mapping Cache with holdtimes**
    - **Elects highest C-RP IP address as RP for group range**
  - **Multicast RP-Discovery messages**
    - **Sent to Cisco-Discovery (224.0.1.40) group**
    - **Sent every 60 seconds or when changes detected**
  - **RP-Discovery messages contain:**
    - **Elected RP's from MA's Group-to-RP Mapping Cache**
  - **Configured via global config command**

```
ip pim send-rp-discovery [<interface>] scope <ttl>
```

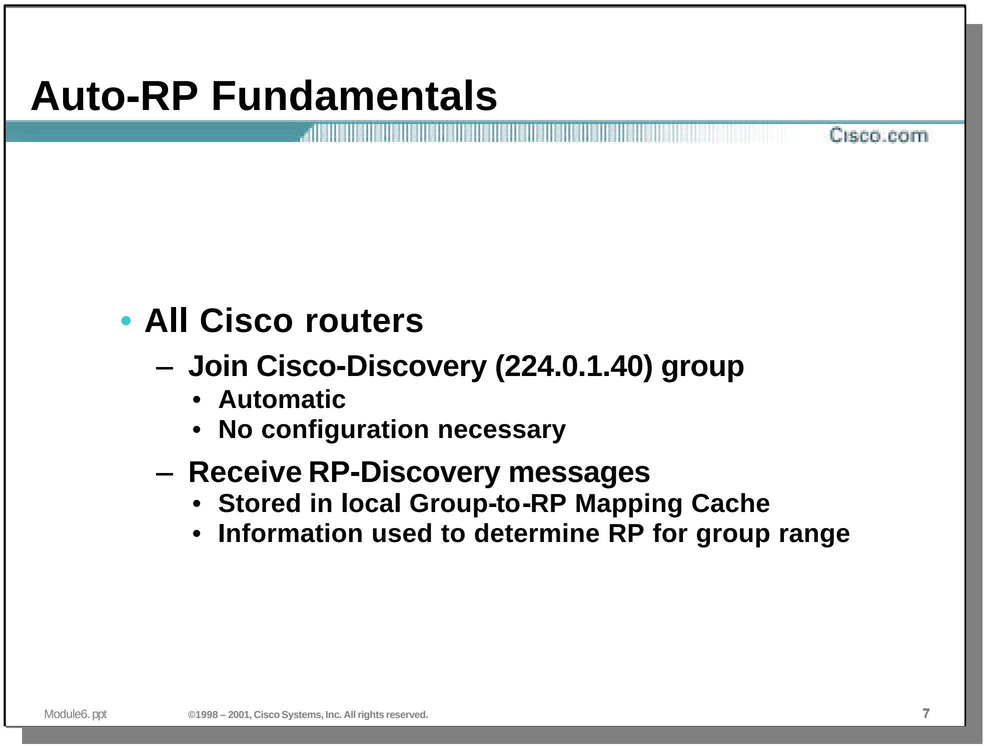
- **Source address of packets set by '<interface>' (12.0)**
  - **If not specified, source address = output interface address**
  - **Results in the appearance of multiple MA's. (one/interface)**

- **Auto-RP Mapping Agents (MA's)**
  - Mapping Agents join the RP-Announce group (224.0.1.39) in order to receive RP Announcements sent by all Candidate RP's.
  - When they receive an Announcement they:
    - Save the Announcement in the Group-to-RP mapping cache
    - Select the C-RP with the highest IP address as RP for the group range
    - The holdtimes are used to timeout an entry in the cache if a C-RP fails and is no longer sending periodic C-RP announcements.
  - Mapping Agents periodically send the elected RP's from their Group-to-RP mapping cache to all routers in the network via RP Discovery messages.
    - RP Discovery messages are multicast to the Auto-RP Discovery group 224.0.1.40.
    - They are sent every 60 seconds or when a change to the information in the Group-to-Mapping takes place.
  - MA's are configured using the (rather obtuse) command:

    ip pim send-rp-discovery[<intfc>] scope <ttl>

    - The optional <intfc> specifies which IP address is used as the source address in the RP-Discovery messages that are sent out all multicast interfaces on the router. (A Loopback interface is normally specified here.) If this interface is not specified, the source address of each multicast interface on the router is used.

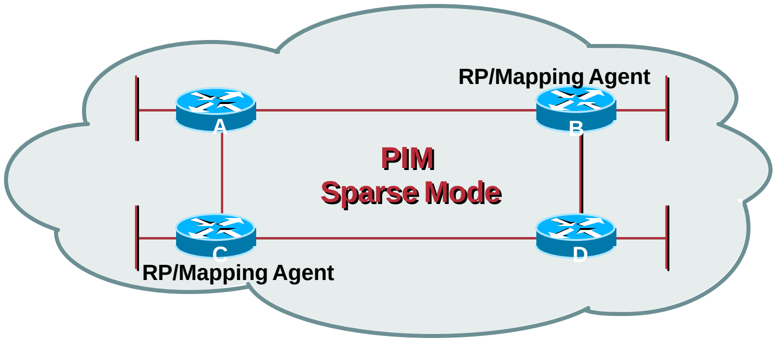
      Note: The reason that this is an optional clause is strictly to be backwards compatible with IOS releases prior to 12.0 that did not allow the interface to be specified. In practice, an interface should always be specified.
    - The <ttl> value controls the TTL of the RP-Discovery message.



- **All Cisco Routers**
  - Automatically join the Cisco-Discovery (224.0.1.40) group in order to receive Group-to-RP mapping information being multicast by the Mapping Agents in the network.
    - *No configuration is necessary!*
  - Group-to-RP mapping information contained in the RP-Discovery messages is stored in the router's local Group-to-RP mapping cache. This information is used by the router to map a Group address to the IP address of the active RP for the group.

Simple Auto-RP Configuration

Cisco.com



On each router: `ip multicast-routing`

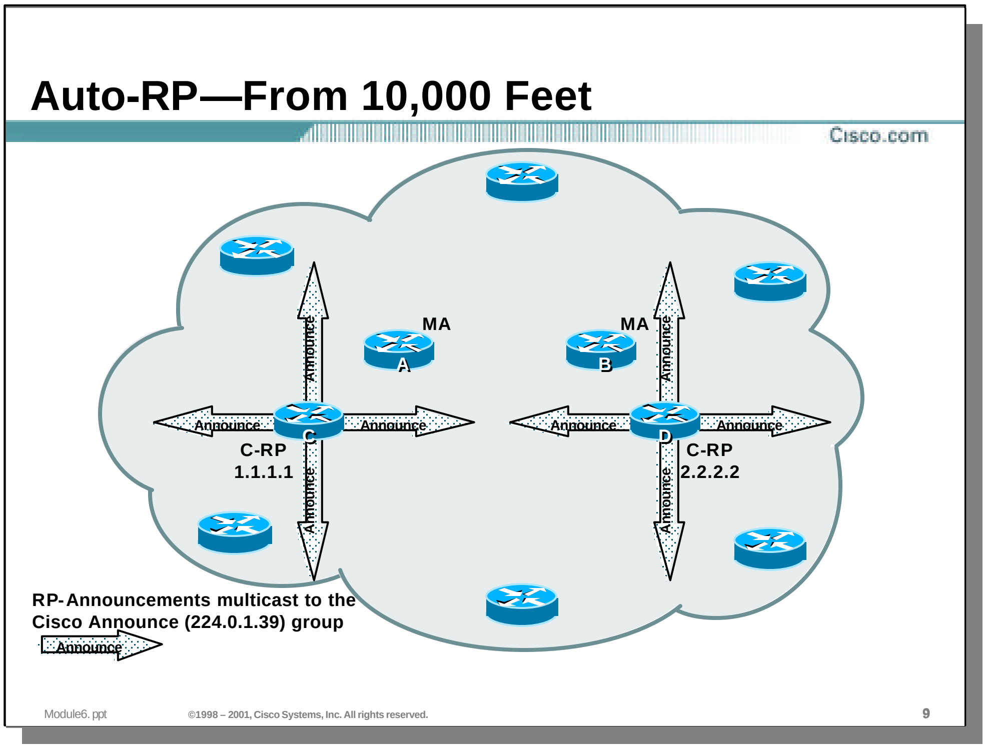
On each interface: `ip pim sparse-dense-mode`

On routers B and C:

```
ip pim send-rp-announce loopback0 scope 16
ip pim send-rp-discovery loopback0 scope 16
```


- **Example Auto-RP Configuration**
  - The above example network shows how to configure a network to run Sparse mode using Auto-RP and two Candidate-RP's/Mapping Agents.

    Note: A common practice is to combine the function of Candidate-RP and Mapping Agent on the same router. This is done more as a configuration convenience than for any operational requirement.
  - One every router in the network:
    - Configure the 'ip multicast-routing' global command to enable multicast on the router.
    - Configure the 'ip pim sparse-dense-mode' interface command on *EVERY* interface on each router. This allows the Auto-RP groups to function in Dense mode and all other groups to operate in Sparse or Dense mode depending on whether an RP has been configured for the group.
    - On the router(s) that are to function as Candidate-RP's, configure the 'ip pim send-rp-announce Loopback0 scope <ttl>' command. (Make sure the <ttl> value is sufficient to allow the message to reach all Mapping Agents in the network.)
    - On the router(s) that are to function as Mapping Agents, configure the 'ip pim send-rp-discovery Loopback0 scope <ttl>' command. (Make sure the <ttl> value is sufficient to allow the message to reach all routers in the network.)
  - No additional configuration is generally necessary. The network is now completely enabled for IP Multicast!

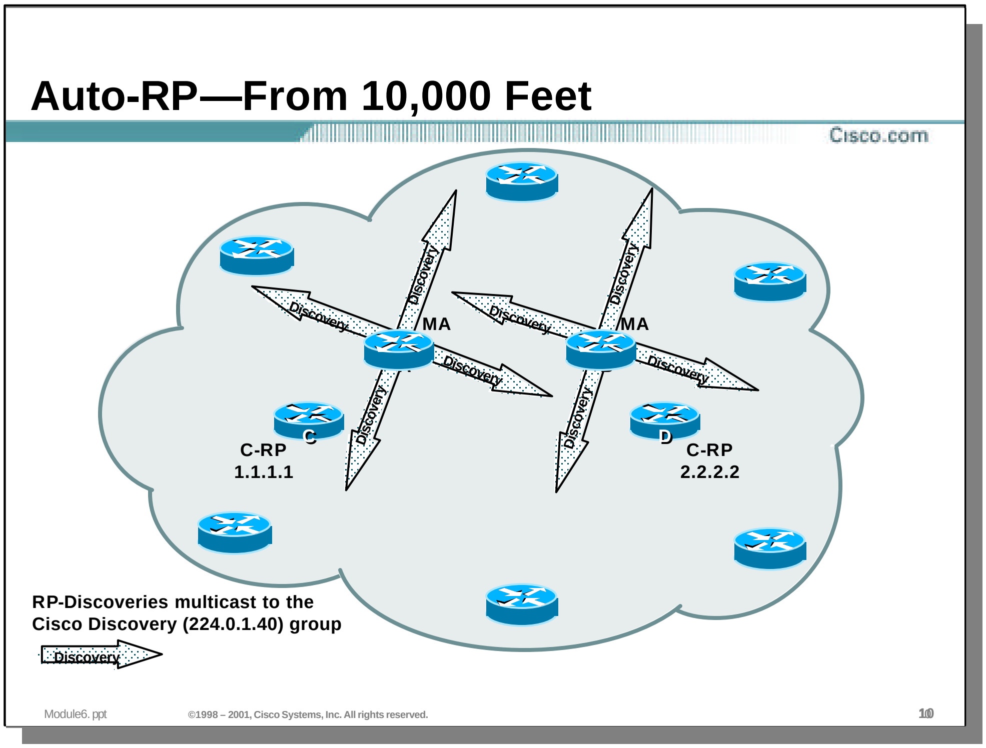


- **Auto-RP - The big picture**

  In this example, routers A and B have been configured as Mapping Agents while routers C & D have been configured as Candidate RP's.

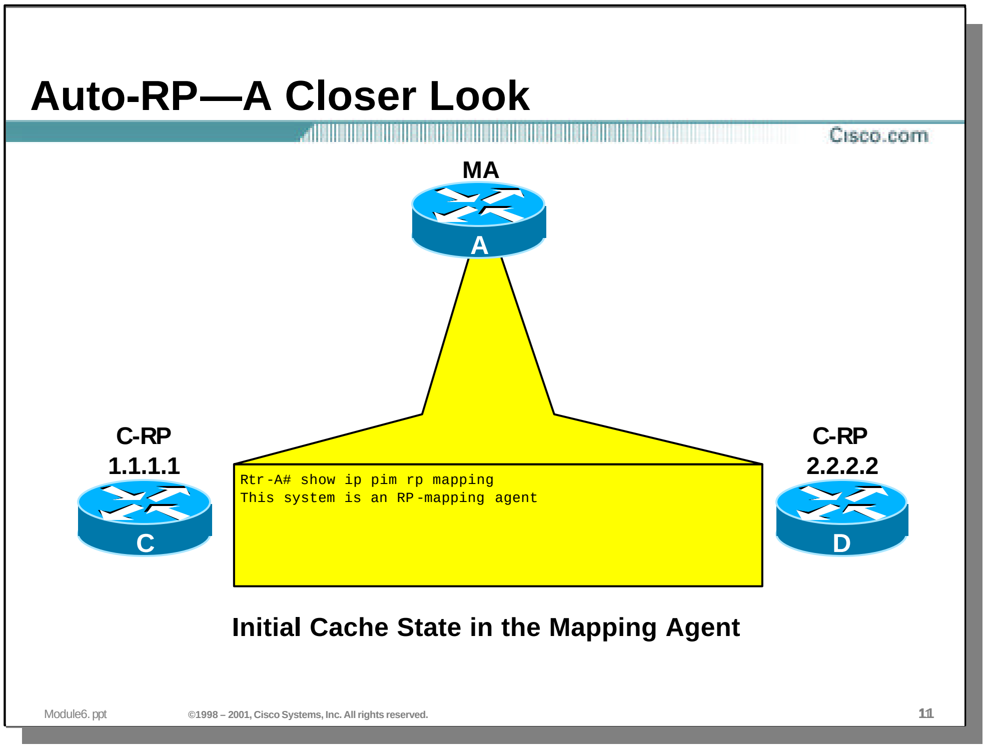
  – Step 1

  - The Candidate RP's begin multicasting their candidacy to be the RP via RP-Announce messages which are sent via the Cisco-Announce group, 224.0.1.39.



- **Auto-RP - The big picture**
  - Step 2
    - The two Mapping Agents (routers A & B) receive the RP-Announce messages from the two Candidate RP's (routers C & D).
  - Step 3
    - The C-RP with the highest IP address (in this case, router D) is stored in the Group-to-Mapping cache of the Mapping Agents.
  - Step 4
    - The Mapping Agents *both* multicast the contents of their Group-to-RP Mapping Cache to the Cisco-Discovery group, 224.0.1.40.

      Note: All Mapping Agents are transmitting this Group-to-RP Mapping information simultaneously. The originally published specification on Auto-RP implied that there was a Master-Slave relationship between Mapping Agents and that only the Master would transmit while the Slave(s) were quiet until the Master failed. This specification is in error and this is not how Auto-RP has been implemented. As long as both Mapping Agents are transmitting identical information, there is no need to add the complexity of a Master-Slave failover scheme.
  - Step 5
    - The RP Discovery messages are received via multicast by all routers in the network. The Group-to-RP mapping information contained in these messages is stored in the router's local Group-to-RP mapping cache. This information is subsequently used by the router to determine the IP address of the RP for a given group.



- **Auto-RP Up Close**

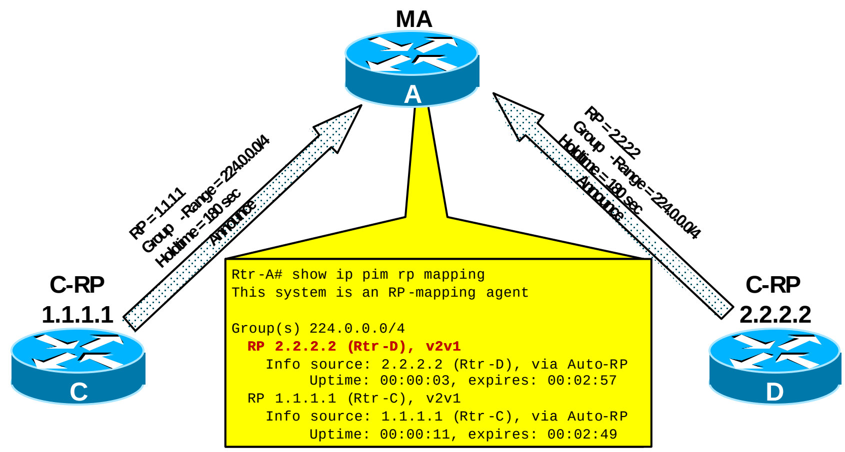
  This is the same example that was presented in the previous slides. However, in this case, we will examine the process in more detail at each step.

  – Step 1

  - At time zero, the Group-to-RP mapping caches in the Mapping Agents are empty since no RP-Announcements have been received.
  - The output of the ‘show ip pim rp mapping’ command shows that router A is a Mapping Agent and that the Group-to-RP mapping cache is empty.

Auto-RP—A Closer Look

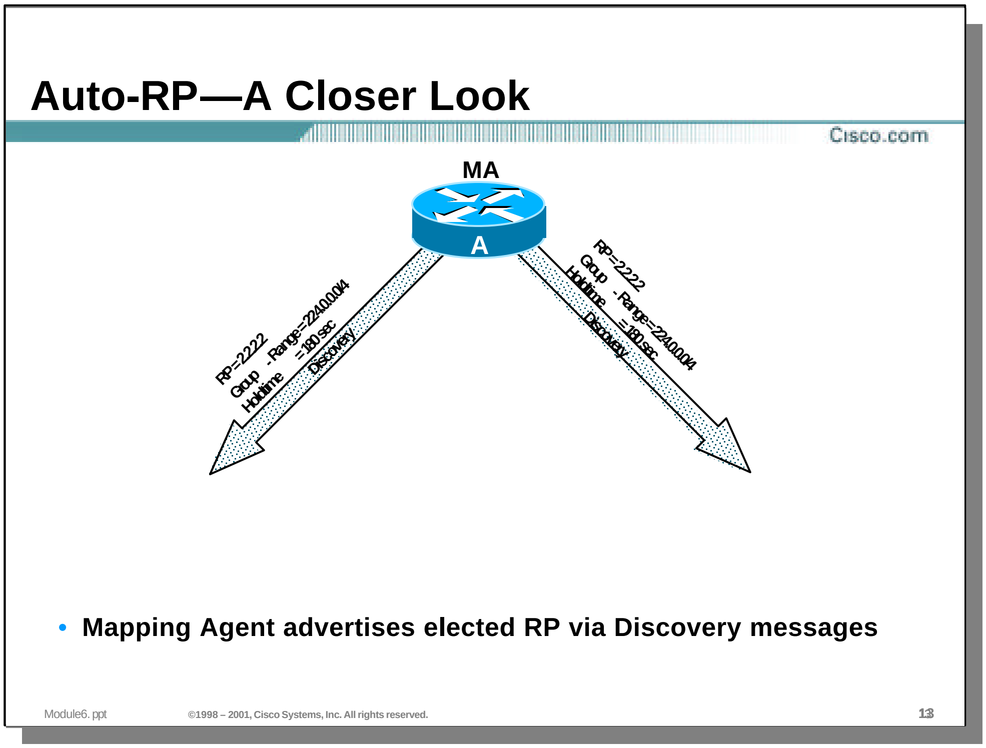
Cisco.com



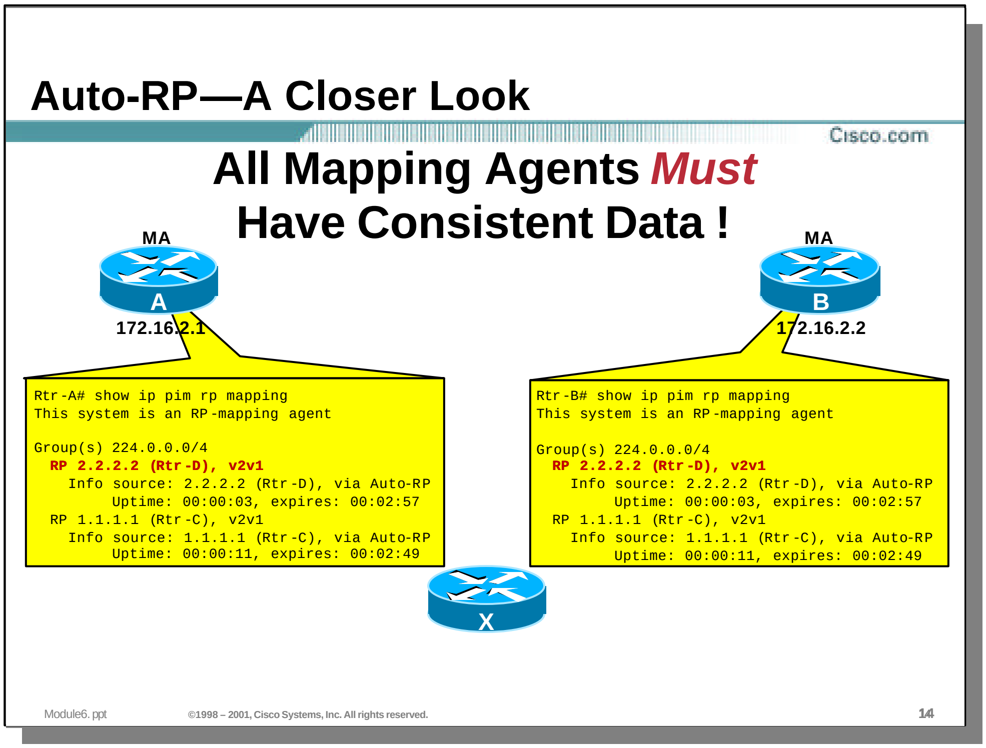
- **C-RP Information is Stored in MA's Group-to-RP Mapping Cache**
- **Mapping Agent elects highest IP Address as RP**

Module6. ppt 12

- **Auto-RP Up Close**
  - Step 2
    - Routers C and D begin sending their RP Announce messages advertising themselves as a candidate to be RP for all multicast groups. (Note the group range, the IP address of the C-RP and the holdtime in the message.)
  - Step 3
    - The Mapping Agent (router A) receives these RP Announcements and stores this information in its Group-to-RP mapping cache.
    - The output of the 'show ip pim rp mapping' command on the Mapping Agent (router A) now shows both router C and D as candidates for group range 224.0.0.0/4 (i.e. all multicast groups with the exception of the Auto-RP groups).
    - The Mapping Agent then elects the C-RP with the highest IP address as the active RP for the group range.



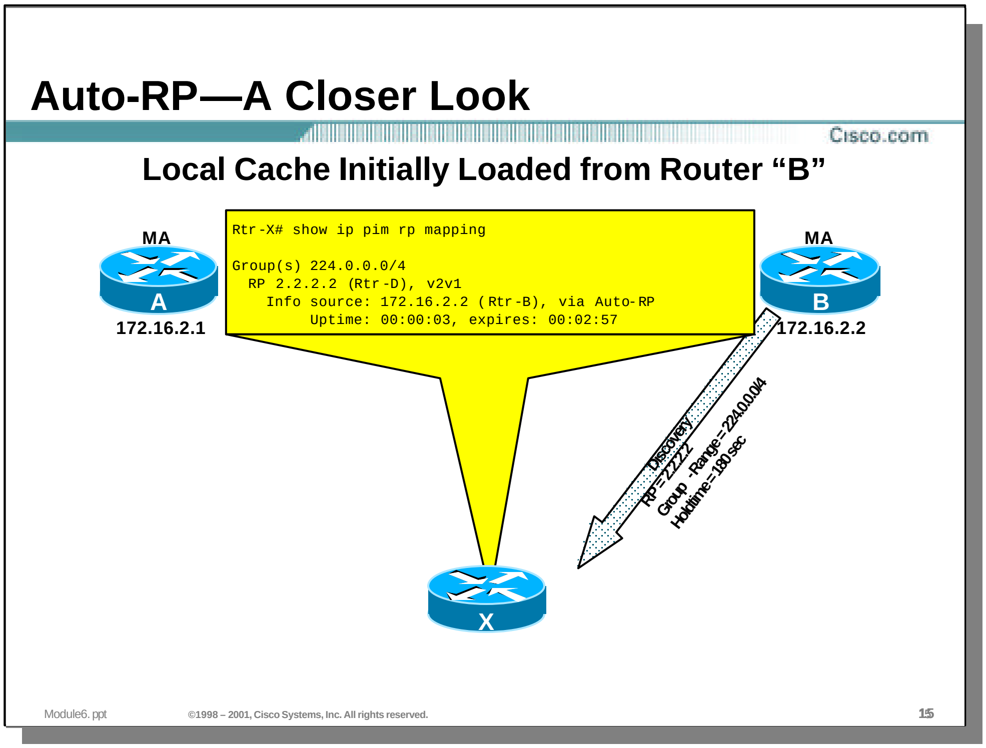
- **Auto-RP Up Close**
  - Step 4
    - The Mapping Agent begins advertising the results of the RP election to the rest of the network via Auto-RP Discovery messages.



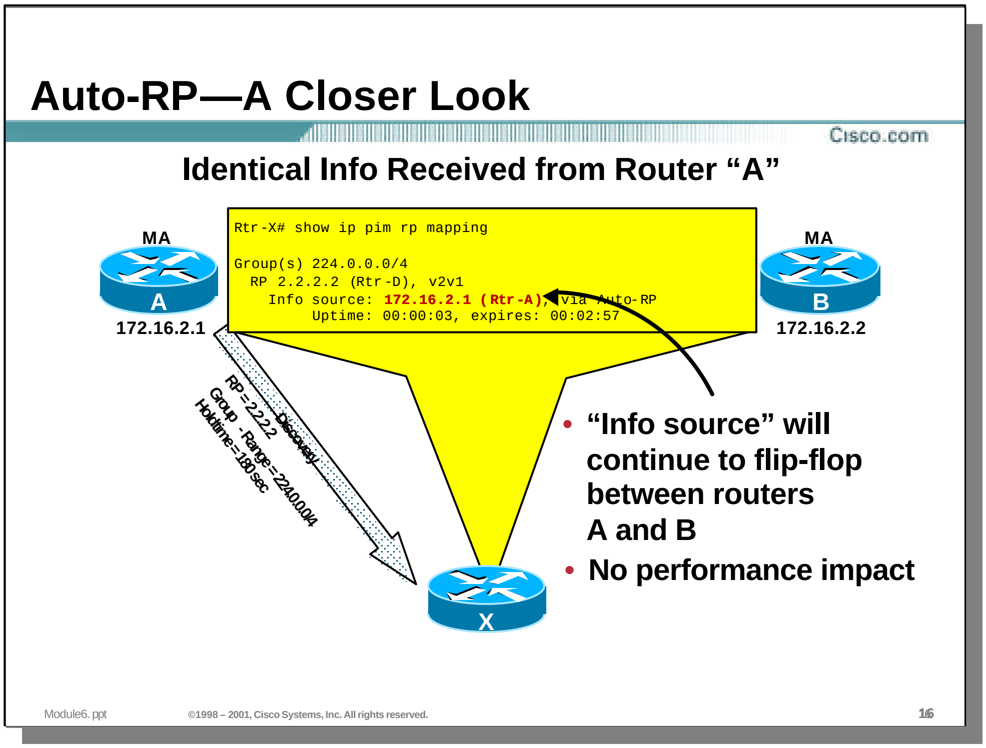
- **Auto-RP Up Close**

  It is critical that all Mapping Agents in the PIM-SM domain have identical information in their Group-to-RP mapping caches. Note that in our example network, they do.

  If the information in the mapping caches are *not* identical, it can cause the routers in the network to flip-flop between two different RPs.



- **Auto-RP Up Close**
  - Step 6
    - Assume that router B is the first MA to send its RP Discovery message containing its Group-to-RP mapping cache contents.
  - Step 7
    - The routers in the network (router X in this example) all receive this RP Discovery message and install the information in their local Group-to-RP mapping cache.
    - The output of the 'show ip pim rp mapping' command shows that router D is currently selected as the RP for group range 224.0.0.0/4 (i.e. all multicast groups with the exception of the Auto-RP groups) and that this information was most recently received from router B.



- **Auto-RP Up Close**
  - Step 8
    - Next, router A sends an RP Discovery message containing its Group-to-RP mapping cache contents.
  - Step 9
    - The routers in the network (router X in this example) all receive this RP Discovery message and update the information in their local Group-to-RP mapping cache. Since both Mapping Agents are sending identical information, the only thing that will change in the local Group-to-RP mapping cache is the "source" of the information.
    - The output of the 'show ip pim rp mapping' command shows that router D is still selected as the RP for group range 224.0.0.0/4 (i.e. all multicast groups with the exception of the Auto-RP groups). However, the data reflects that this information was most recently received from router A.
    - The flip-flop of the information source in the routers' local Group-to-RP mapping cache has little or no performance impact on the router.

Module Agenda

Cisco.com

- Auto RP
- **PIMv2 BSR**
- Static RPs
- Tuning RP Operations
- Debugging RP Operation
- Special Cases

PIMv2 BSR Overview

Cisco.com

- A single Bootstrap Router (BSR) is elected
  - Multiple Candidate BSR's (C-BSR) can be configured
    - Provides backup in case currently elected BSR fails
  - C-RP's send C-RP announcements to the BSR
    - C-RP announcements are sent via unicast
    - BSR stores ALL C-RP announcements in the "RP-set"
  - BSR periodically sends BSR messages to all routers
    - BSR Messages contain entire RP-set and IP address of BSR
    - Messages are flooded hop-by-hop throughout the network away from the BSR
  - All routers select the RP from the RP-set
    - All routers use the same selection algorithm; select same RP
- BSR cannot be used with Admin-Scoping

Module6. ppt 18

- **BSR Overview**
  - Bootstrap Router (BSR)
    - A single router is elected as the BSR from a collection of Candidate BSR's.
    - If the current BSR fails, a new election is triggered.
    - The election mechanism is pre-emptive based on C-BSR priority.
  - Candidate RP's (C-RP's)
    - Send C-RP announcements directly to the BSR via unicast. (Note: C-RP's learn the IP address of the BSR via periodic BSR messages.)
    - The BSR stores the complete collection of all received C-RP announcements in a database called the "RP-set".
  - The BSR periodically sends out BSR messages to all routers in the network to let them know the BSR is still alive.
  - BSR messages are flooded hop-by-hop throughout the network.
    - Multicast to the "All-PIM Routers" group (224.0.0.13) with a TTL of 1.
  - BSR messages also contain:
    - The complete "RP-set" consisting of *all* C-RP announcements.
    - The IP Address of the BSR so that C-RP's know where to send their announcements.
  - All routers receive the BSR messages being flooded throughout the network.
    - Select the active RP for each group range using a common hash algorithm that is run against the RP-set. This results in all routers in the network selecting the same RP for a given group-range.
  - ***BSR cannot be used with Admin-Scoping!***
    - Admin scoping was not considered when BSR was designed. One problem is that C-RP announcements that are unicast to the BSR cross multicast boundaries. There are several other problems as well.

PIMv2 BSR Fundamentals

Cisco.com

- **Candidate RPs**
  - **Unicast PIMv2 C-RP messages to BSR**
    - **Learns IP address of BSR from BSR messages**
    - **Sent every rp-announce-interval (default: 60 sec)**
  - **C-RP messages contain:**
    - **Group Range (default = 224.0.0.0/4)**
    - **Candidate's RP address**
    - **Holdtime = 3 x <rp-announce-interval>**
  - **Configured via global config command**

```
ip pim rp-candidate <intfc> [group-list acl]
```


- **BSR Candidate RP's (C-RP)**
  - C-RP Messages
    - Sent periodically (default: 60sec) directly to the BSR via unicast.
    - Messages contain the Group-range, C-RP address and a holdtime.
    - The IP address of the current BSR is learned from the periodic BSR messages that are received by all routers in the network.
  - C-RP's are configured using the following command:

```
ip pim rp-candidate <intfc> [group-list <acl>]
```

    - The <intfc> parameter dictates the IP Address that is advertised in the C-RP message. In most cases, a Loopback interface is used.
    - The optional 'group-list' access-list can be used to specify a group-range other than the default of 224.0.0.0/4 (i.e. all multicast groups)
    - This command may be configured more than once on a router so that the router will function as C-RP for multiple group ranges.

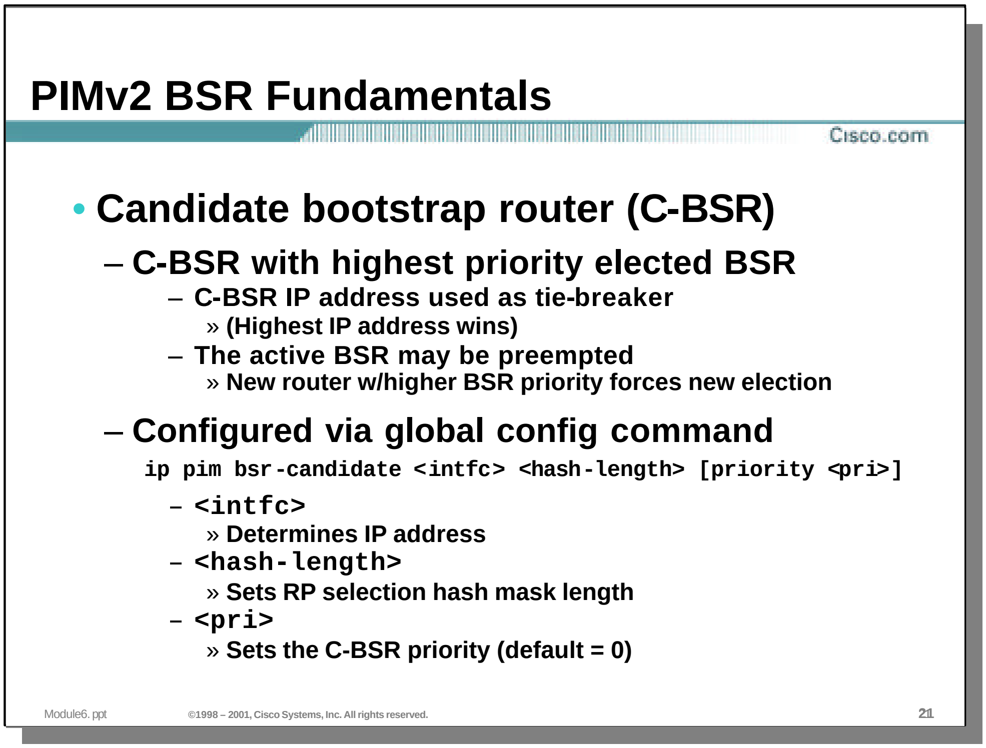
PIMv2 BSR Fundamentals

Cisco.com

- **Bootstrap router (BSR)**
  - **Receive C-RP messages**
    - **Accepts and stores ALL C-RP messages**
    - **Stored in Group-to-RP Mapping Cache w/holdtimes**
  - **Originates BSR messages**
    - **Multicast to All-PIM-Routers (224.0.0.13) group**
      - **(Sent with a TTL = 1)**
    - **Sent out all interfaces. Propagate hop-by-hop**
    - **Sent every 60 seconds or when changes detected**
  - **BSR messages contain:**
    - **Contents of BSR's Group-to-RP Mapping Cache**
    - **IP Address of active BSR**

Module6. ppt 20

- **Bootstrap Router**
  - The primary purpose of the Bootstrap router is to collect all C-RP announcements in to a database called the RP-set and to periodically send the RP-set out to all other routers in the network inside of BSR messages.
  - BSR Messages
    - Sent periodically (default: 60 secs) by the BSR out all multicast interfaces.
    - BSR messages are multicast to the All-PIM-Routers (224.0.0.13) group with a TTL of 1. These messages are received by all PIM neighbors who retransmit them (again with a TTL of 1) out all interfaces except the one in which the messages was received. (An RPF check is done to insure the BSR message came in on the correct interface in the direction of the BSR.)
    - BSR messages contain the RP-set and the IP address of the currently active BSR. (This is how C-RP's know where to unicast their C-RP messages.



- **Candidate Bootstrap Routers (C-BSRs)**
  - C-BSR's participate in the BSR election mechanism.
    - The C-BSR with the highest priority is elected as the BSR.
    - The highest IP address of the C-BSR's is used as a tie-breaker.
    - The election mechanism is preemptive. If a new C-BSR with a higher priority comes up, it triggers a new election.
  - C-BSR's are configured using the following command:
    - ip pim bsr-candidate <intfc> <hash-length> [priority <pri>]
    - The <intfc> parameter is used to specify the BSR's IP address which is forwarded in BSR messages. (This is where C-RP's will send their messages if the C-BSR is elected as BSR.)
    - The <hash-length> parameter specifies the number of bits in the hash. This can be used to control RP load balancing across a group range where different RP's are selected for different groups within a group range whose size is defined by the hash-length in bits.
    - The optional <pri> value permits the C-BSR to be configured with a priority other than the default of zero.

PIMv2 BSR Fundamentals

Cisco.com

- **BSR election mechanism**
  - **C-BSR's:**
    - **Begin in Candidate-BSR state**
      - **BSR-Timeout timer started (150 seconds)**
      - **If higher priority (preferred) BSR message received**
        - » **Restart timer and forward BSR message**
        - » **Copy info to local Group-to-RP mapping cache**
        - » **Otherwise, discard BSR message**
      - **If timer expires, transition to Elected-BSR state**
    - **While in Elected-BSR state**
      - **Periodically originate own BSR messages**
        - » **Include local Group-to-RP mapping cache in msg**
      - **Return to Candidate-BSR state if preferred BSR message (higher priority) received**

Module6. ppt 22

- **BSR Election**

  How and when routers in the network forward BSR messages plays a key role in the BSR election mechanism. The algorithm used to decide whether to process and forward an incoming BSR message depends on whether the router is a Candidate BSR or not.

- **BSR Message forwarding by C-BSR routers**

  C-BSR's operate in one of two states, Candidate-BSR or Elected-BSR. Initially, a C-BSR comes up in C-BSR state.

- **C-BSR State**
  - A BSR-Timeout timer started with a period of 150 seconds. If this timer expires, the router transitions to the **Elected-BSR** State.
  - If a BSR message is received with higher priority than the C-BSR's priority, then the BSR (whose address is in the BSR message) is considered to be "preferred" and the BSR message is processed as follows:
    - The BSR-Timeout timer is reset.
    - The BSR message is forwarded out all other interfaces.
    - The RP-set in the BSR message is copied into the local Group-to-RP mapping cache.
  - If a BSR message is received with a priority less than the C-BSR's priority, the BSR message is simply discarded.

- **Elected BSR State**
  - The router has been elected as the BSR and periodically originates BSR messages containing the current RP-set.
  - If a BSR message is received from another router with a higher priority, forward the BSR message and transition back to **C-BSR** state; otherwise discard the BSR message.

PIMv2 BSR Fundamentals

Cisco.com

- **BSR election mechanism**
  - **Non C-BSRs (i.e., all other routers):**
    - **Start in Accept-Any state**
      - **Accepts first BSR message received**
      - **Saves BSR info and forwards BSR message**
      - **Transitions to Accept-Preferred state**
    - **While in Accept-Preferred state**
      - **Starts BSR-Timeout timer**
      - **Only accept and forward preferred BSR messages**
        - » **(i.e., BSR messages with priority > current BSR priority)**
      - **Otherwise, discard BSR message**
      - **Return to Accept-Any state if timer expires**

Module6. ppt 23

- **BSR Message forwarding by Non C-BSRs**
  - Non C-BSR routers operate in two states, **Accept-Any** state and **Accept-Preferred** state. When a non C-BSR router boots up, it starts in the **Accept-Any** state.
- **Accept-Any State**
  - Accept the first BSR message received and process it as follows:
    - Copy the RP-Set into the local Group-to-RP mapping cache.
    - Save the current BSR priority and BSR address in the BSR message.
    - Transition to the **Accept-Preferred** state.
- **Accept-Preferred State**
  - Start the BSR-Timeout timer with a period of 150 seconds. If this timer expires, transition back to the **Accept-Any** state.
  - Accept only BSR messages that are "preferred". (A "preferred" BSR message is one with a priority greater than or equal to the current BSR priority.) The accepted BSR message is then processed as follows:
    - The BSR-Timeout timer is reset.
    - The BSR message is forwarded out all other interfaces.
    - The RP-set in the BSR message is copied into the local Group-to-RP mapping cache.
    - Save the current BSR priority and BSR address in the BSR message.
  - If a BSR message is received with a priority less than the C-BSR's priority, the BSR message is simply discarded. (Remember, the IP address of the BSR is used to break any ties with the winner being the C-BSR with the highest IP address.)

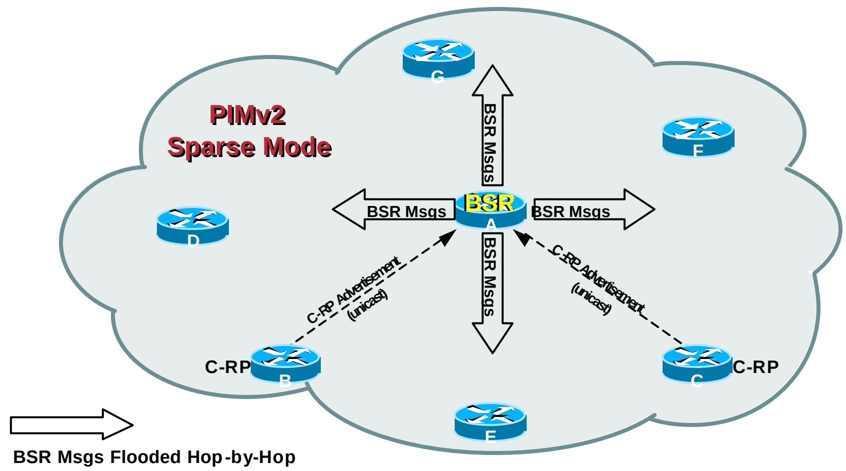
PIMv2 BSR Fundamentals

Cisco.com

- **All PIMv2 routers**
  - **Receive BSR messages**
    - **Stored in local Group-to-RP Mapping Cache**
    - **Information used to determine active BSR address**
  - **Selects RP using Hash algorithm**
    - **Selected from local Group-to-RP Mapping Cache**
    - **All routers select same RP using same algorithm**
    - **Permits RP-load balancing across group range**

- **All PIMv2 Routers**
  - Accept BSR messages based on the rules described in the previous pages. When a BSR message is accepted:
    - The RP-Set in the BSR message is stored in the local Group-to-RP mapping cache.
    - The BSR message is forwarded out all other interfaces (except the one in which it was received) on the router.
  - Selects RP using a Hash Algorithm
    - The RP for a group is selected from the set of C-RP's (stored in the Group-to-RP mapping cache) that have advertised their candidacy for a matching group-range.
    - The same hashing algorithm is used by all routers to select the RP from the set of C-RP's in the RP-set. Since all routers run the same algorithm on the same RP-set (received from the BSR), all routers will select the same RP for a given group.
    - The hashing algorithm permits multiple C-RP's to load balance the duties of RP across a range of groups. Only one C-RP will be selected as RP for any *single* group in the group range. However, the hash algorithm may select other C-RP's as RP for another group *within the group range.*
    - For example, given a BSR hash length of 30 bits being used on IPv4 group addresses, this results in a remainder of 2 bits of an IPv4 address or 4 group addresses that a C-RP will serve as RP. In this scenario, if C-RP routers A and B both advertise their candidacy for group-range 224.1.1.0/24 and the hash algorithm selects router A as RP for 224.1.1.0, the hash length of 28 bits will also cause router A to be selected as RP for groups 224.1.1.1, 224.1.1.2 and 224.1.1.3 (i.e. a contiguous group range of of 4 addresses.) If the hash algorithm selects router B as RP for group 224.1.1.4, it will also select router B for groups 224.1.1.5, 224.1.1.6 and 224.1.1.7.

Basic PIMv2 BSR



- **BSR Example**
  - Step 1
    - Candidate RP's unicast their C-RP messages to the previously elected BSR. (The C-RP's learned the IP address of the BSR from the BSR messages that are being flooded throughout the network.)
  - Step 2
    - The BSR receives and stores *ALL* C-RP information in a database called the RP-Set (which is stored in the Group-to-RP mapping cache on Cisco routers).
  - Step 3
    - The BSR periodically sends BSR messages containing the RP-set out all of its interfaces. These BSR messages are forwarded hop-by-hop away from the BSR by all routers in the network. The RP-set is used by all routers in the network to calculate the RP for a group using a common hash algorithm.

Module Agenda

- Auto RP
- PIMv2 BSR
- **Static RPs**
- Tuning RP Operations
- Debugging RP Operation
- Special Cases

Static RP's

Cisco.com

- **Hard-coded RP address**
  - **When used, must be configured on every router**
  - **All routers must have the same RP address**
  - **RP fail-over not possible**
    - **Exception: If Anycast RPs are used. (See MSDP module.)**
- **Command**

```
ip pim rp-address <address> [group-list <acl>] [override]
```

  - **Optional group list specifies group range**
    - **Default: Range = 224.0.0.0/4** ***(Includes Auto-RP Groups!!!!)***
  - **Override keyword "overrides" Auto-RP information**
    - **Default: Auto-RP learned info takes precedence**

Module6. ppt 27

- **Hard-code RP Addresses**
  - Requires every router in the network to be manually configured with the IP address of a *single* RP.
  - If this RP fails, there is no way for routers to fail-over to a standby RP.
    - The exception to this rule is if "Anycast-RP's" are in use. This requires MSDP to be running between each RP in the network.
- **Command**

```
ip pim rp-address <address> [group-list <acl>] [override]
```

  - The 'group-list' allows a group range to be specified.
    - The default is ALL multicast groups or 224.0.0.0/4
    - DANGER, WILL ROBINSON!!!

      The default range includes the Auto-RP groups (224.0.1.39 and 224.0.1.40) which will cause this router to attempt to operate these groups in Sparse mode. This is normally not desirable and can often lead to problems where some routers in the network are trying to run these groups in Dense mode (which is the normal method) while others are trying to use Sparse mode. This will result in some routers in the network being starved of Auto-RP information. This in turn, can result in members of some groups to not receive multicast traffic.
  - The 'override' keyword permits the statically defined RP address to take precedence over Auto-RP learned Group-to-RP mapping information.
    - The default is that Auto-RP learned information has precedence.

Module Agenda

Cisco.com

- Static RPs
- Auto RP
- PIMv2 BSR
- **Tuning RP Operations**
- Debugging RP Operation
- Special Cases

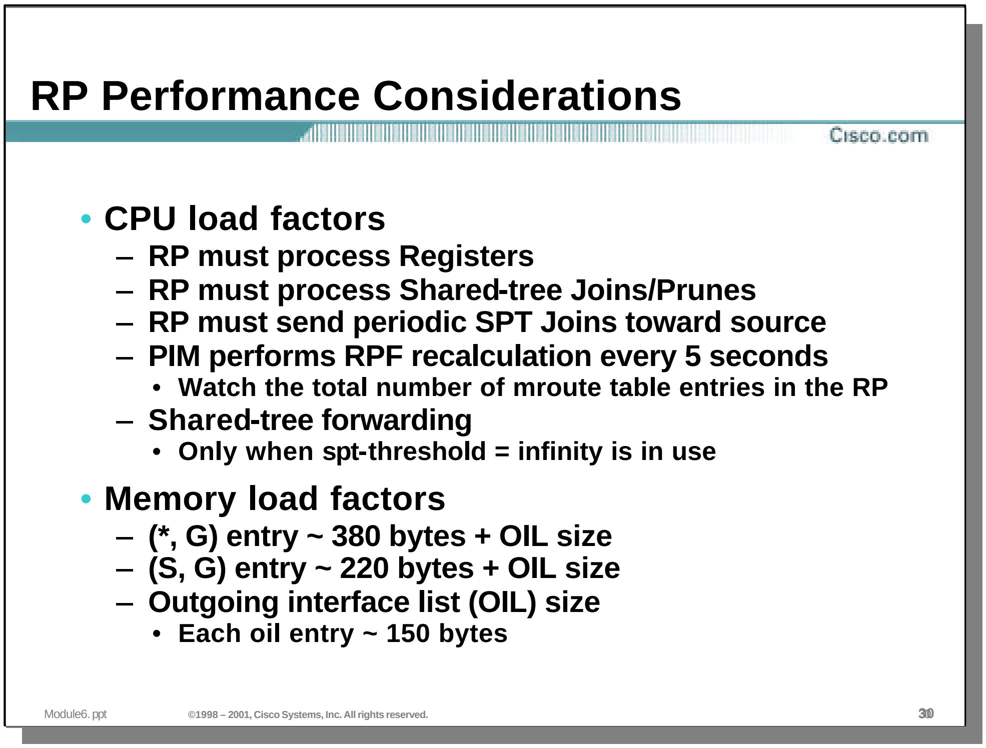
RP Placement

Cisco.com

- **Q: “Where do I put the RP?”**
  - **A: “Generally speaking, it’s not critical”**
- **SPT’s are normally used by default**
  - **RP is a place for source and receivers to meet**
  - **Traffic does not normally flow through the RP**
  - **RP is therefore not a bottleneck**
- **Exception: SPT-Threshold = Infinity**
  - **Traffic stays on the shared tree**
  - **RP could become a bottleneck**

Module6. ppt 29

- **PIM MFAQ (Most Frequently Asked Question)**
  - Q: Where should I put the RP?
  - A: Generally speaking, it is not critical.
    - The default behavior of PIM-SM is to switch to the Shortest-Path Tree (aka the Source Tree) and bypass the RP as soon as a new source is detected. This means that in most cases, multicast traffic does not flow through the RP. Therefore, the RP does not become a point of congestion.
    - The default behavior can be overridden in Cisco routers by setting the SPT-Threshold to “Infinity”. This prevents the Cisco router from joining the SPT and keeps all group traffic flowing down the Shared Tree. In this case, the RP *could* become a bottleneck.



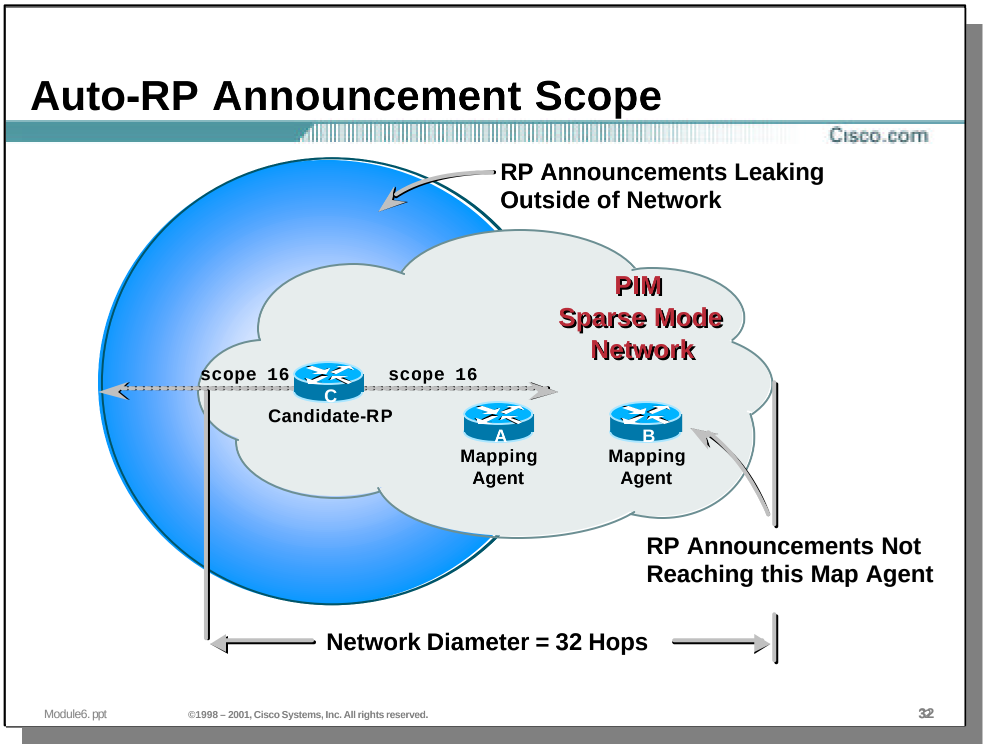
- **RP Performance Considerations**
  - CPU Load Factors
    - The RP will receive all Register messages for any new sources in the network. Although processing of Register messages is done at Process Level, the impact on the router is usually small since the RP will immediately send back a Register-Stop message.
    - The RP will receive and must process all Shared Tree Join/Prune messages from downstream routers on the Shared Tree. Downstream routers continue to send periodic (once a minute) Join/Prune messages up the Shared Tree. The number of these Join/Prune messages is generally quite small and therefore has little impact on the RP.
    - The RP must send periodic (once a minute) SPT Joins toward all sources for which it has members active on a branch of the Shared Tree.
    - In order to detect a network topology change, *ALL* PIM routers perform an RPF recalculation on every (*, G) and (S, G) entry in the mroute table every 5 seconds. The impact of this will grow as the total number of entries in the mroute table increase and as the number of entries in the unicast routing table increase. (The later is due to the fact that each RPF calculation requires the route to the source to be looked up in the unicast routing table. If this table is quite large, as would be the case for poorly aggregated address space, the lookup can take more effort than when the number of entries is kept small.) Except for the following load factor, this is the most significant CPU load factor.
    - Any traffic that does have to flow through the RP requires it to replicate the packets out all outgoing interfaces.
  - Memory Factors
    - The amount of memory consumed by PIM is primarily a function of the size of the mroute table. (See the numbers in the slide for details.)

Dealing with Overloaded RP’s

Cisco.com

- **Increase CPU horsepower**
- **Increase memory**
- **Use SPTs if not already**
- **Split RP load across several RPs!**

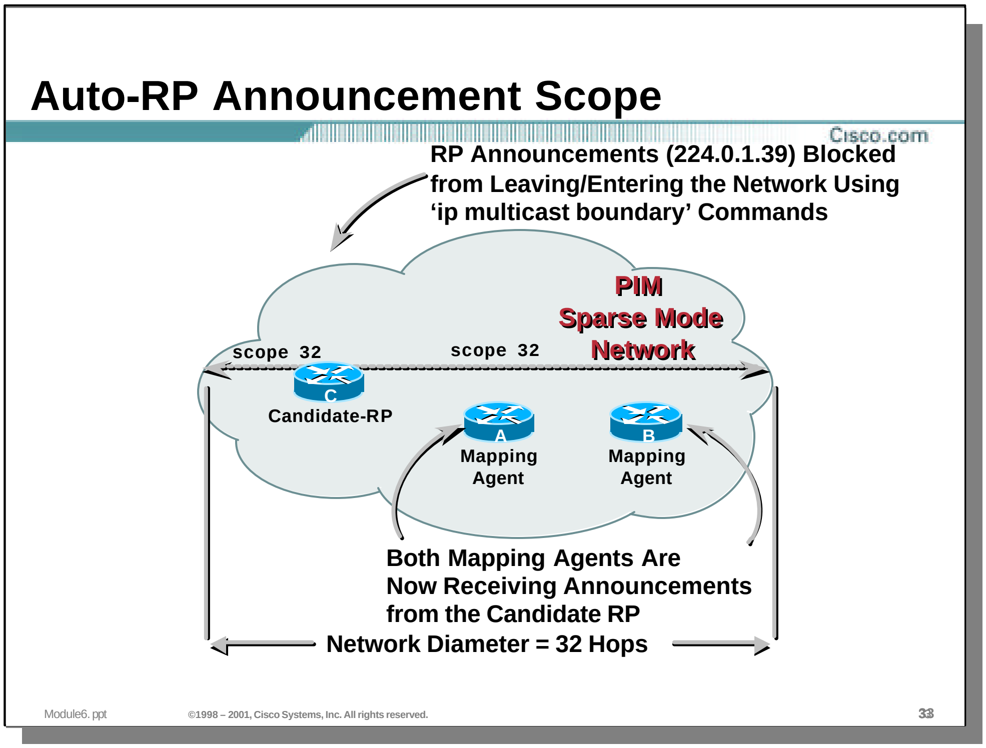
- **Dealing with overloaded RP’s**
  - If possible, increase the CPU horsepower of the RP. In some cases, this can be accomplished by changing the RP or RSP card in the router.
  - If the multicast traffic in the network results in an *extremely large* number of mroute table entries, it may be necessary to increase the amount of memory in the router. This scenario is not likely to occur except in the cases where a low-end router with a minimal amount of memory is used in a network with a large number of multicast sources.
  - If the RP is overloaded due to the multicast packet replication and forwarding demands, insure that Shortest Path Trees are in use by making sure all routers in the network have an SPT-Threshold set to the default value of zero.
  - If all else fails, split the RP load across several RPs by assigning different group ranges to different RPs. (The “Anycast-RP” technique can be used in conjunction with MSDP to allow more than on RP to be active for a *single* group and to share the load.)



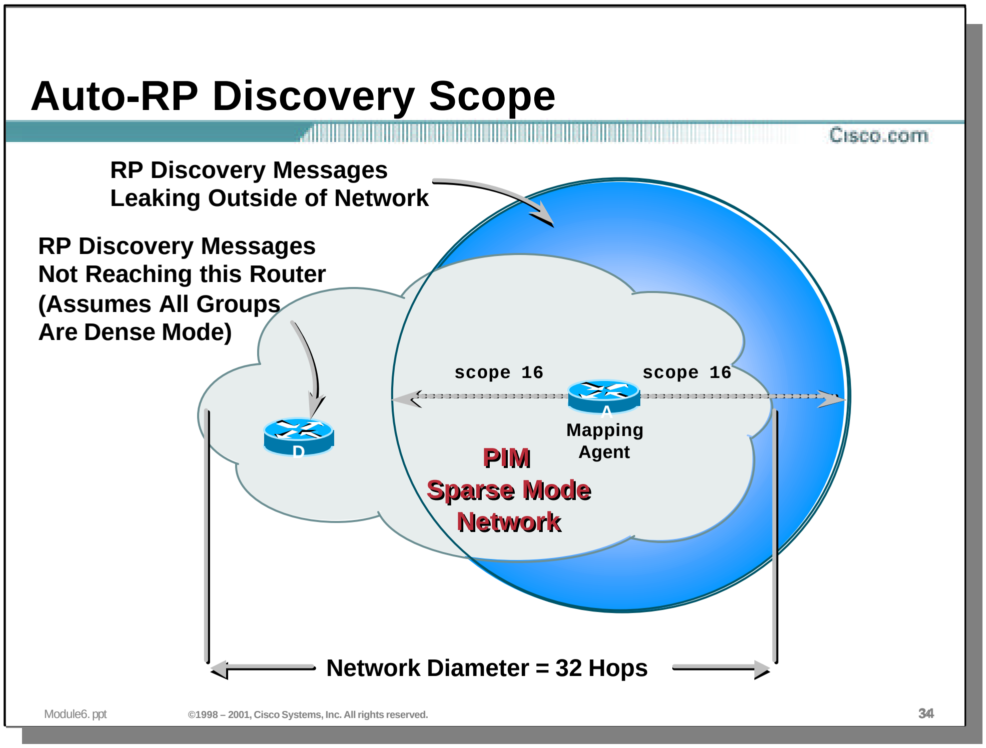
- **Auto-RP Announcement Scope**
  - Care must be taken in the selection of the TTL scope of RP Announcement messages that are sent by C-RPs to insure that the messages reach all Mapping Agents in the network.
- **Example**

  In the diagram above, an arbitrary scope of 16 was used in the ‘ip pim send-rp-announce’ command on the C-RP router. However, the maximum diameter of the network is greater than 16 hops and in this case one Mapping Agent is further away than 16 hops. As a result, this Mapping Agent does not receive the RP Announcement messages from the C-RP. This can cause the two Mapping Agents to have different information in their Group-to-RP mapping caches. If this occurs, each Mapping Agent will advertise a different router as the RP for a group which will have disastrous results.

  Notice also, that the C-RP is fewer than 16 hops way from the edge of the network. This can result in RP Announcement messages leaking into adjacent networks and causing Auto-RP problems in those networks.



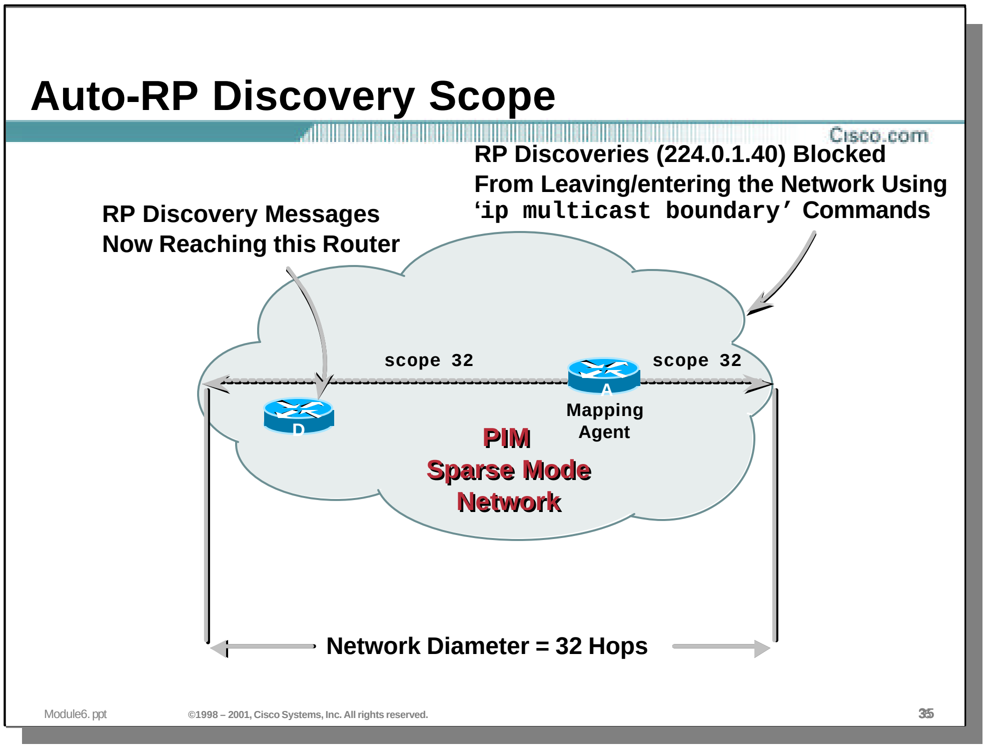
- **Auto-RP Announcement Scope**
  - The best way to avoid the problems on the preceding page is to use a sufficiently large enough scope so that the RP Announcement messages reach all Mapping Agents in the network.
- **Example**
  - In the above diagram, the maximum network diameter is 32. Therefore by setting the scope to 32 or greater, we are assured that the RP Announcements will reach both Mapping Agents shown in the example network.
  - In order to prevent RP Announcement messages from leaking into adjacent networks, a multicast boundary is defined for the Cisco-Announce (224.0.1.39) multicast group on all border routers in the network. This not only stops RP Announcement messages from leaking out, *it more importantly, stops any from leaking in from adjacent networks.*



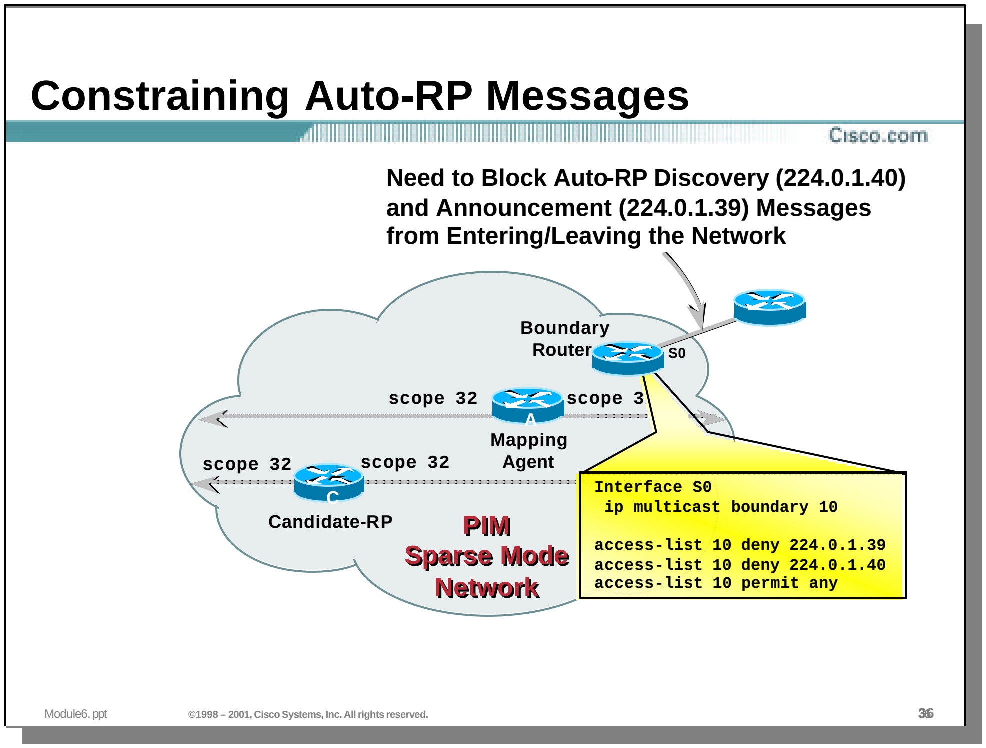
- **Auto-RP Discovery Scope**
  - Care must be taken in the selection of the TTL scope of RP Discovery messages that are sent by Mapping Agents to insure that the messages reach all routers in the network.
- **Example**

  In the diagram above, an arbitrary scope of 16 was used in the 'ip pim send-rp-discovery' command on the Mapping Agent. However, the maximum diameter of the network is greater than 16 hops and in this case, at least one router is further away than 16 hops. As a result, this router does not receive the RP Discovery messages from the MA. This can result in the router having no Group-to-RP mapping information. If this occurs, the router will attempt operate in Dense mode for all multicast groups while other routers in the network are working in Sparse mode.

  Notice also, that the MA is fewer than 16 hops way from the edge of the network. This can result in RP Discovery messages leaking into adjacent networks and causing Auto-RP problems in those networks.



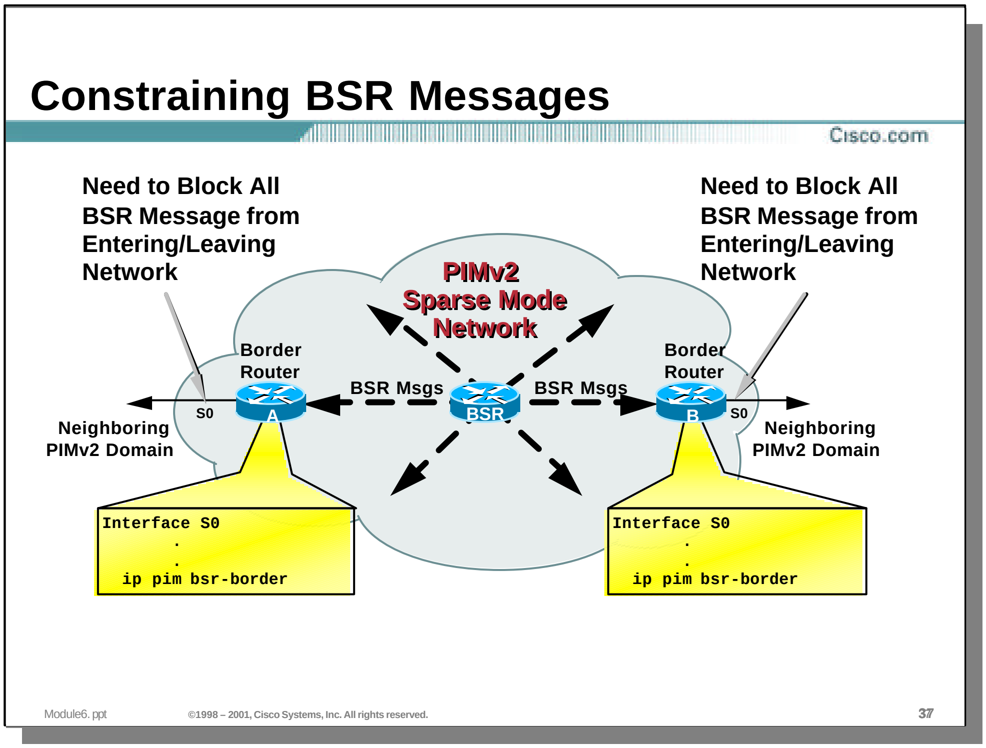
- **Auto-RP Discovery Scope**
  - The best way to avoid the problems on the preceding page is to use a sufficiently large enough scope so that the RP Discovery messages reach all routers in the network.
- **Example**
  - In the above diagram, the maximum network diameter is 32. Therefore by setting the scope to 32 or greater, we are assured that the RP Discovery messages will reach the farthest router in the network.
  - In order to prevent RP Discovery messages from leaking into adjacent networks, a multicast boundary is defined for the Cisco-Discovery (224.0.1.40) multicast group on all border routers in the network. This not only stops RP Discovery messages from leaking out, *it more importantly, stops any from leaking in from adjacent networks.*



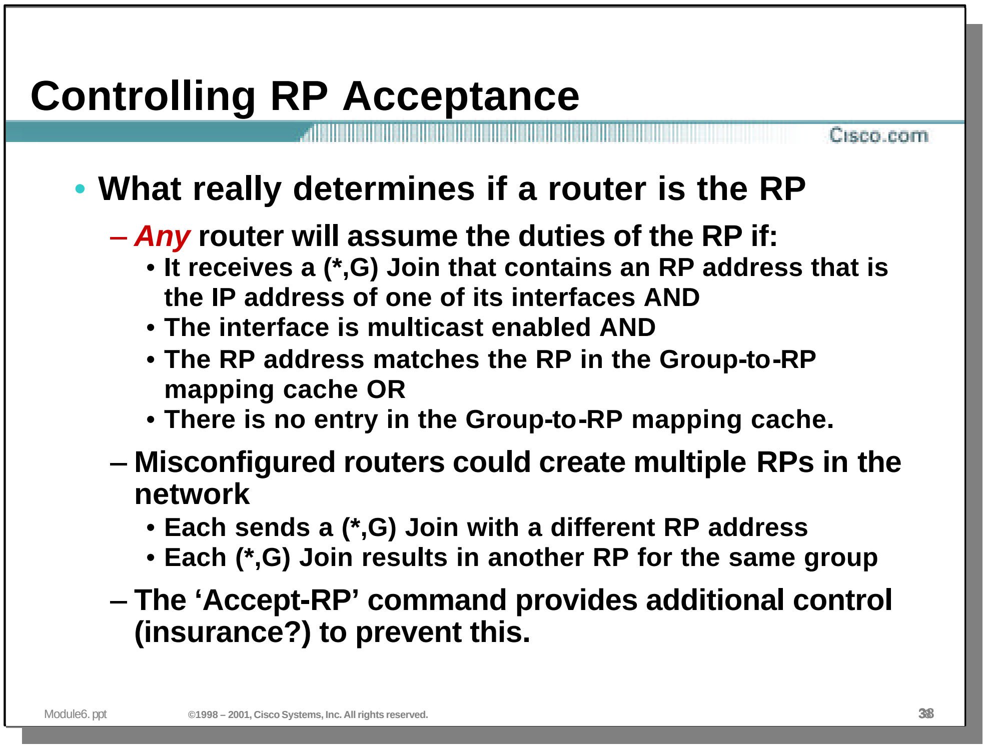
- **Constraining Auto-RP Messages**
  - This example shows how to configure the multicast boundary on a border router so that Auto-RP messages do not leak into or out of the network.
    - On the border interface (in this case, Serial0) the ‘ip multicast boundary’ command is used.
    - The access list associated with the ‘ip multicast boundary’ command is as follows:

```
access-list 10 deny 224.0.1.39
access-list 10 deny 224.0.1.40
access-list 10 permit any
```

The above access list stops the flow of multicast traffic for the two Auto-RP groups (224.0.1.39 and 224.0.1.40) while allowing all other multicast traffic to enter or exit via interface Serial 0.



- **Constraining BSR Messages**
  - Like Auto-RP, allowing BSR messages to leak into or out of a network can cause problems both in the local network and in adjacent networks.
  - In order to block BSR messages from entering or exiting on a given interface, the 'ip pim bsr-border' interface command can be used.



- **By default, a router operates as an RP for a group if:**
  - It receives a (*, G) Join containing one of its addresses as the RP or
  - It receives a (S, G) Register message.
- **Basic sanity check**
  - Routers will perform a rudimentary sanity check to see if it actually should be the RP for group "G" by searching the Group-to-RP mapping cache for an entry for group "G". If an entry is found and the RP for the group is ***not*** this router, then the router will discard the (*, G) Join or (S,G) Register and will ***not*** become the RP. Otherwise, it will assume that it ***is*** the RP for group "G" and assume duties of the RP.
- **Extended sanity checking**
  - In order to provide additional control and sanity checking over which router should be accepted as the RP, the IOS command '**ip pim accept-rp**' was created.

Controlling RP Acceptance

Cisco.com

- **Accept-RP Command**
  - **Global configuration commands**

```
ip pim accept-rp <rp-address> [<acl>]
ip pim accept-rp Auto-rp [<acl>]
ip pim accept-rp 0.0.0.0 [<acl>]
```

  - **Multiple commands accepted**
    - **Command list sorted in order shown above**
    - **Only one Auto-RP and one 0.0.0.0 (wildcard) accepted**
    - **Omitting ACL implies 224.0.0.0/4 group range**
  - **Search Rules**
    - **Top down search**
    - **Stop on RP address match—Apply ACL and exit**
    - **Exception: Auto-RP denies RP/Group**
      - **Apply 0.0.0.0 (wildcard) entry (if it exists)**

Module6. ppt 39

- **Command format**

  ip pim accept-rp <rp-address> [<acl>]
  ip pim accept-rp Auto-rp [<acl>]
  ip pim accept-rp 0.0.0.0 [<acl>]

  – The option <acl> is used to specify which groups are valid using standard permit and deny clauses.

    • Omitting the <acl> assumes a 'permit 224.0.0.0 15.255.255.255'.

  – If more than one of the above commands is configured, they are sorted in the order shown above.

  – The "Auto-rp" entry matches any RP address learned via Auto-RP.

  (Note: This form has implied "deny" clauses for the Auto-RP groups, 224.0.1.39 and 224.0.1.40, that cannot be overridden in the optional <acl>. This helps prevent the Auto-RP groups from accidentally switching to Sparse mode.)

  – The "0.0.0.0" (wildcard) matches any RP address.

  – While multiple 'ip pim accept-rp <rp-address>' commands may be configured, only a single "Auto-rp" and a single "0.0.0.0" (wildcard) command is accepted.

- **Search Rules**

  – The list of configured commands is searched from top down and stops at the first entry that matches the RP address.

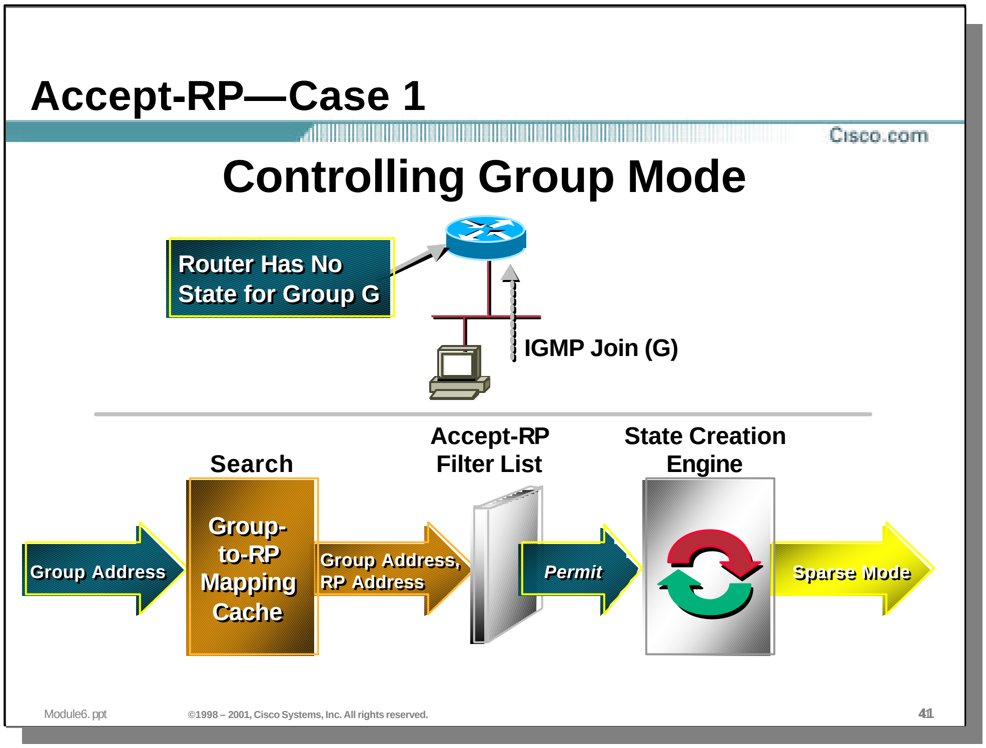
  – The <acl> is applies and the RP is either permitted or denied.

  – Exception: If an "Auto-RP" entry "denies" an RP and a "0.0.0.0" entry exists, the 0.0.0.0 entry is also tried.

Controlling RP Acceptance

Cisco.com

- **Accept-RP Command Usage**
  - **Case 1 — Controlling Group mode**
  - **Case 2 — Accepting (*, G) joins**
  - **Case 3 — Accepting PIM registers at the RP**

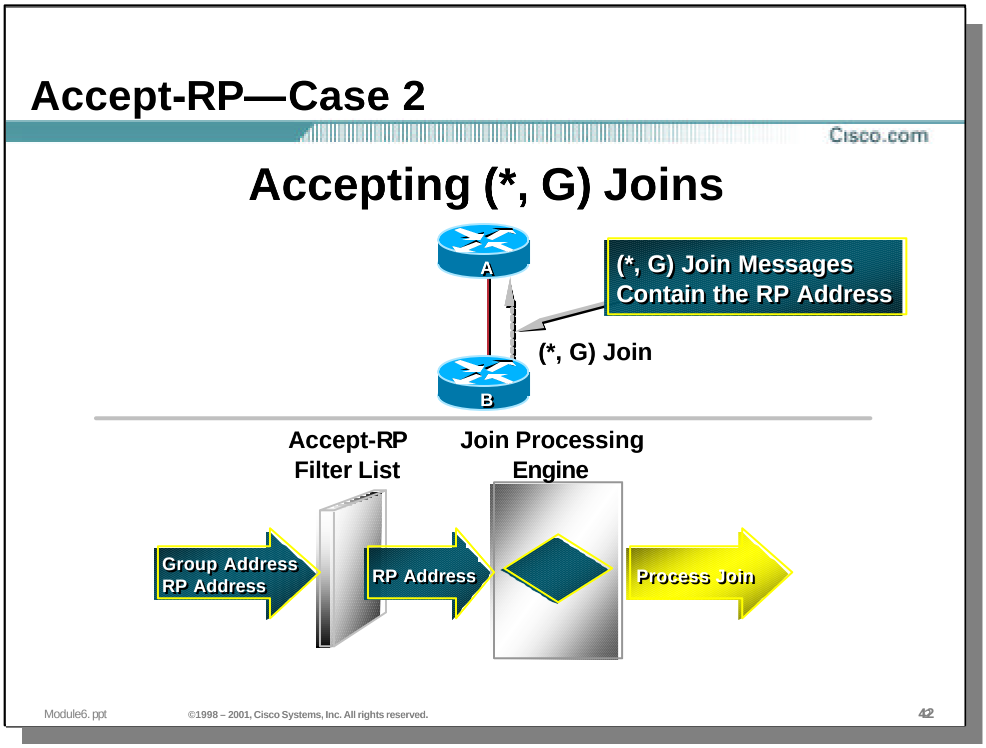


- **Case 1 - Controlling Group Mode**

  If a router has no (*, G) state when an IGMP Membership Report is received for group "G", the router will apply the configured Accept-RP rules to determine if there is a valid RP for this group or not. If there isn't, the (*,G) entry is created and set as a Dense mode group. If there is a valid RP, the (*,G) entry is set in Sparse mode.

  – Step 1
    - The Group-to-RP mapping cache is searched for the group address in the IGMP Join message. If an entry is not found, then the group is created in Dense mode.

  – Step 2
    - If a matching entry ***is*** found in the Group-to-RP mapping cache, the Group and RP addresses are run through the Accept-RP filters. If a "permit" is returned, then the group is created in Sparse mode; otherwise the group is created in Dense mode.



- **Case 2 - Accepting (*, G) Joins**

  If a router receives (*, G) Join it will apply the configured Accept-RP rules to determine if the RP address contained in the (*, G) Join is valid or not.

  – Step 1 (not shown)

  - The Group-to-RP mapping cache is searched for the group address in the (*, G) Join message. If an entry is found and it is a “negative” entry indicating that the group has been forced to always be in Dense mode, then the (*, G) Join is not accepted and an error message is generated.

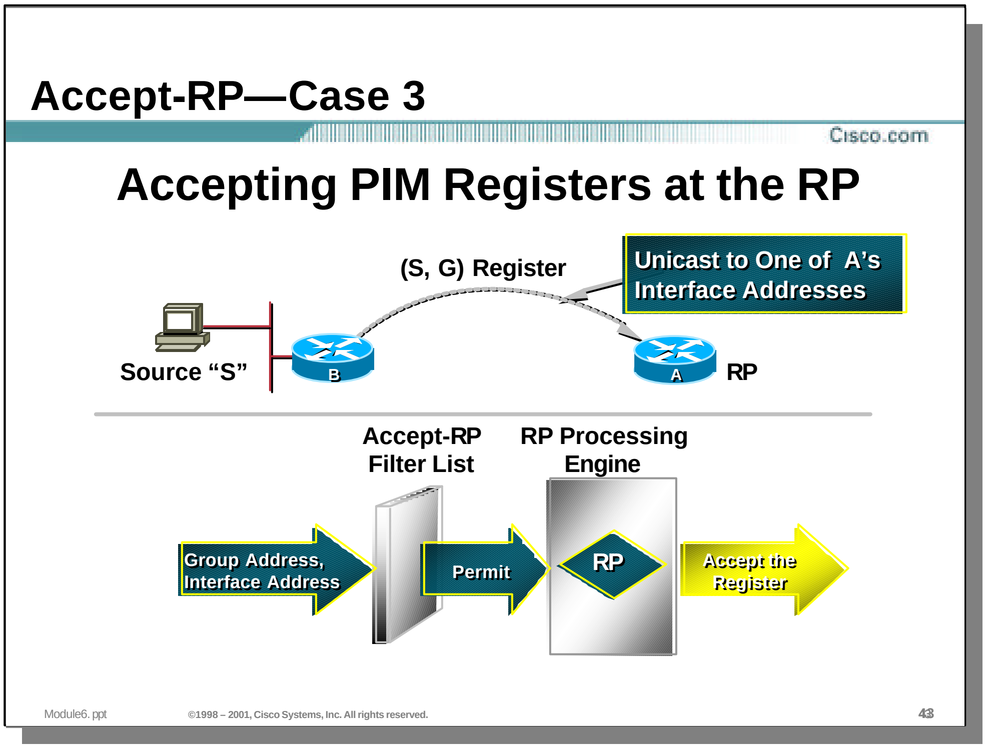
  – Step 2

  - The Group and RP addresses (contained in the (*, G) Join) are run through the Accept-RP filters. If a “permit” is returned, then the (*, G) Join is processed normally; otherwise the (*, G) Join is dropped and an error message is generated.

- **Example**

  – When Auto-RP is in use, it is normally the case that the two Auto-RP groups, 224.0.1.39 and 224.0.1.40 should be operating in Dense mode. However, if a downstream router is misconfigured with a static RP address, it will send (*, G) Joins for these Auto-RP groups. The routers that receive these (*, G) Joins will create a (*,G) entry ***in Sparse mode*** for these Auto-RP groups. This can result in portions of the network trying to operate these groups in Dense mode while other parts of the network are operating in Sparse mode. This will generally cause the Auto-RP mechanisms to fail. The following Accept-RP command will cause a router to reject any (*,G) Joins for the Auto-RP groups and prevent these Joins from propagating.

```
ip pim accept-rp Auto-rp
```



- **Case 3 - Accepting PIM Register messages**

  If a router receives PIM Register message, it will apply the configured Accept-RP rules to determine if the router is permitted to be the RP or not.

  – Step 1 (not shown)

  - The Group-to-RP mapping cache is searched for the group address in the (*, G) Join message. If an entry is found and it is a "negative" entry indicating that the group has been forced to always be in Dense mode, then the PIM Register is not accepted, a Register-Stop is sent back to the first-hop router and an error message is generated.

  – Step 2

  - The Group (contained in the multicast packet encapsulated in the Register message) and RP addresses (the destination IP address in the Register message) are run through the Accept-RP filters. If a "permit" is returned, then the PIM Register is processed normally; otherwise a Register-Stop is sent back to the first-hop router and an error message is generated.

Filtering C-RP Announcements

Cisco.com

- **Use on Mapping Agents to filter out bogus C-RP's**
  - **Some protection from RP-Spoofing denial-of-service attacks**
  - **Multiple commands may be configured as needed**
- **Global command**

```
ip pim rp-announce-filter rp-list <acl> [group-list <acl>]
```

  - **rp-list <acl>**
    - » **Specifies from which routers C-RP Announcements are accepted.**
  - **group-list <acl>**
    - » **Specifies which groups in the C-RP Announcement are accepted.**
    - » **If not specified, defaults to deny all groups**

- **Filtering RP Announcements**

  Network Administrators may wish to configure Mapping Agents so that they will only accept C-RP Announcements from well-known routers in the network. This will prevent C-RP Announcements from bogus routers from being accepted and potentially being selected as the RP.

- **Global Command**

```
ip pim rp-announce-filter rp-list <acl> [group-list <acl>]
```

– The rp-list <acl> specifies the IP address(es) from which C-RP announcements will be accepted.

– The option group-list <acl> specifies the group range(s) that are acceptable for the routers in the rp-list. If not specified, the default group-list <acl> is `deny all`

– Multiple instances of this command may be configured.

Controlling Source Registration

Cisco.com

- **New global command - IOS 12.0(6)**

```
ip pim accept-register [list <acl>] | [route-map <map>]
```

  - **Used on RP to filter incoming Register messages**
  - **Filter on Source address alone (Simple ACL)**
  - **Filter on (S, G) pair (Extended ACL)**
  - **May use route-map to specify what to filter**
    - **Filter by AS-PATH if (m)BGP is in use.**
- **Prevents unwanted sources from sending**
  - **First hop router blocks traffic from reaching net**
  - **Note: Source can still send traffic on local wire**

- **Controlling Source Registration**

  In some cases, it may be desirable to control which hosts in the network can actually source traffic to a group. While there is currently no way to prevent a bogus source from transmitting traffic on its local segment, we can prevent it from being registered to the RP. This will, in most cases, prevent this traffic from going past the first-hop router and reaching other hosts in the network.

  A new IOS command, ‘ip pim accept-register’ was introduced which when configured on an RP, controls which (S, G) Register messages will be accepted and which will be rejected.

- **Global Command (IOS 12.0(6) or later)**

  ip pim accept-register [list <acl>] | [route-map <map>]

  – If the “list <acl>” is specified, the <acl> can either be a simple access list to control which hosts may send to any groups or an extended access list that specifies both source and group address combinations that are permitted or denied from sending.

  – If the “route-map <map>” is specified, then only matching (S, G) traffic will be accepted. (Note: This permits other matching criteria to be considered such as AS-PATH.)

Module Agenda

Cisco.com

- Static RPs
- Auto RP
- PIMv2 BSR
- Tuning RP Operations
- Debugging RP Operation
- Special Cases

Module6. ppt 46

Debugging Auto-RP Operation

Cisco.com

- **Understand the Auto-RP mechanisms**
  - *This is the fundamental debugging tool for problems with Auto-RP!!!*
- **Verify Group-to-RP Mapping Caches**
  - **First on the Mapping Agents**
    - **Other routers will learn Group-to-RP mapping info from these routers**
      - **If not correct, use debug commands to see what's wrong**
    - **Make sure all MA's have consistent Group-to-RP information**
      - **If not, watch for TTL Scoping problems**
  - **Then on other routers**
    - **If info doesn't match MA, there is a problem distributing the information**
    - **Use show and debug commands to find where the break is**

Module6. ppt 47

- **Debugging Auto-RP**
  - First and foremost, you must understand the fundamental mechanisms behind Auto-RP in order to debug problems!
  - Verify Group-to-RP Mapping Caches on Mapping Agents
    - Because other routers in the network will learn the Group-to-RP mapping information from the Mapping Agents, it is important that this information is correct on the Mapping Agents. If the information is not correct, verify that the C-RP's are configured correctly and that C-RP Announcements are being received properly by the Mapping Agent.
    - If multiple Mapping Agents are in use, make sure that their Group-to-RP mapping information is identical. If not, the routers in the network will oscillate between the different RP's selected by the Mapping Agents. Again, make sure all Mapping Agents are properly receiving Auto-RP Announcements from all C-RP's in the network. *Watch out for TTL scoping problems!*
  - Verify Group-to-RP Mapping Caches on all other routers
    - Group-to-RP mapping information should match the MA's. If not, verify that the router is properly receiving Auto-RP Discovery messages from the Mapping Agents. *Again, watch out for TTL scoping problems!*

Debugging Auto-RP Operation

Cisco.com

- **Insure Auto-RP group state is correct**
  - **Should normally be in Dense mode**
  - **Watch out for mixed DM and SM conditions**
    - **Can occur when Static RP's are also defined**
      - **Always 'deny' Auto-RP groups on Static RP configurations**
    - **Use 'Accept-RP' filters on all routers as insurance**
  - **Watch out for DM problems in NBMA networks**
    - **(See Module 7 for details)**

- **Common Problem - Incorrect Auto-RP Group mode**

  The two Auto-RP groups, 224.0.1.39 and 224.0.1.40 are normally run in Dense mode so that this information is flooded throughout the network. Only in very rare situations is it desirable to run these two groups in Sparse mode because this creates a "chicken-and-the-egg" problem. (How do you join the RP for the Auto-RP groups if you don't know the RP address?)

  Therefore, the following situations should be verified:

  – Insure that the Auto-RP groups are operating in Dense mode on all routers in the network.

  – Mixed DM/SM situations can arise when static RP addresses have been configured on some routers in the network. To avoid this:

  - Always specify an <acl> on any 'ip pim rp-address' commands that denies groups 224.0.1.39 and 224.0.1.40.
  - Configure Accept-RP filters on all routers in the network that "deny" groups 224.0.1.39 and 224.0.1.40.

    (Note: The 'ip pim accept-rp auto-rp' command has an implied set of "deny" clauses for these two groups to prevent them from switching into Sparse mode.)

Debugging BSR Operation

Cisco.com

- **Understand the BSR mechanisms**
  - ***This is the fundamental debugging tool for problems with BSR!!!***
- **Verify Group-to-RP Mapping Caches**
  - **First on the BSR**
    - **Other routers will learn Group-to-RP mapping info from this router**
      - **If not correct, use debug commands to see what's wrong**
  - **Then on other routers**
    - **If info doesn't match BSR, there is a problem distributing the information**
    - **Use show and debug commands to find where the break is**

Module6. ppt 49

- **Debugging BSR Operation**
  - First and foremost, you must understand the fundamental mechanisms behind BSR in order to debug problems!
  - Verify Group-to-RP Mapping Caches on the elected BSR.
    - Because other routers in the network will learn the Group-to-RP mapping information from the elected BSR, it is important that this information is correct on this router. If the information is not correct, verify that the PIMv2 C-RP's are configured correctly and that C-RP Announcements are being received properly by the BSR.
  - Verify Group-to-RP Mapping Caches on all other routers
    - Group-to-RP mapping information should match the BSR. If not, verify that the router is properly receiving BSR messages.

Module Agenda

Cisco.com

- Static RPs
- Auto RP
- PIMv2 BSR
- Tuning RP Operations
- Debugging RP Operation
- Special Cases

RP on a Stick

Cisco.com

- Triggering conditions on the RP:
  - A (*,G) entry (i.e. Shared Tree) exists with a *single* outgoing interface on the RP.
  - And an (S,G) entry (i.e a Source) exists on the same interface with a *Null* outgoing interface list.
- Results in special “state” at the RP
  - Frequently misunderstood and rarely seen
    - Default behavior is to join SPT which avoids this
  - Mishandled in versions of IOS prior to 12.0
  - Requires the Proxy Join Timer to resolve
  - Need to understand concept of “atomic joins”

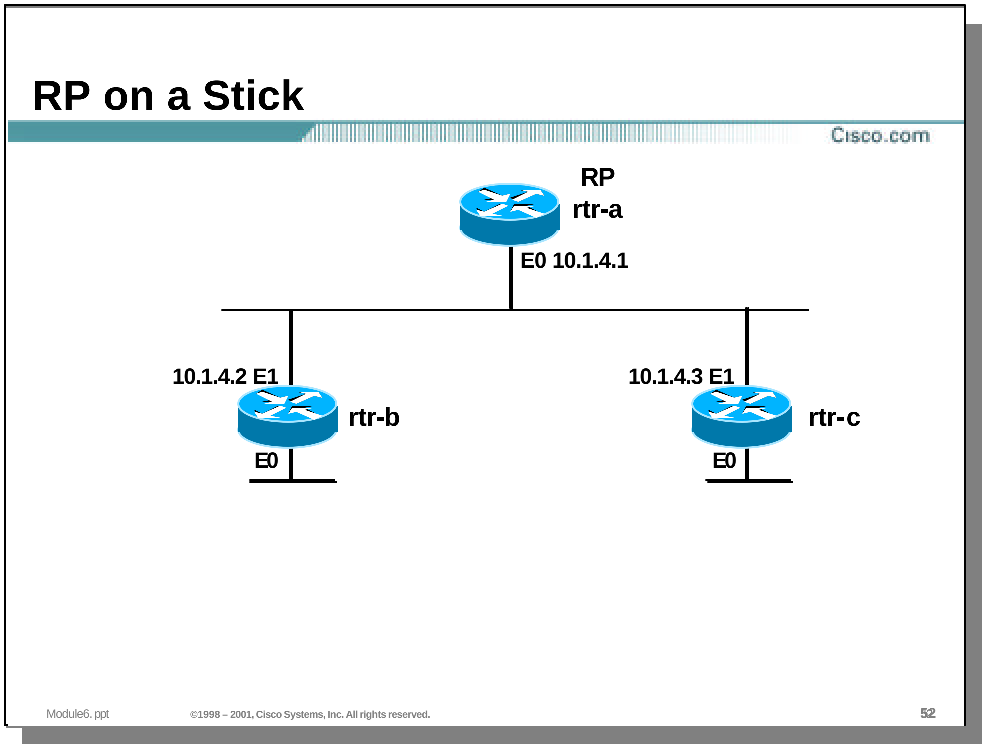
Module6. ppt 51

- **RP on a Stick**

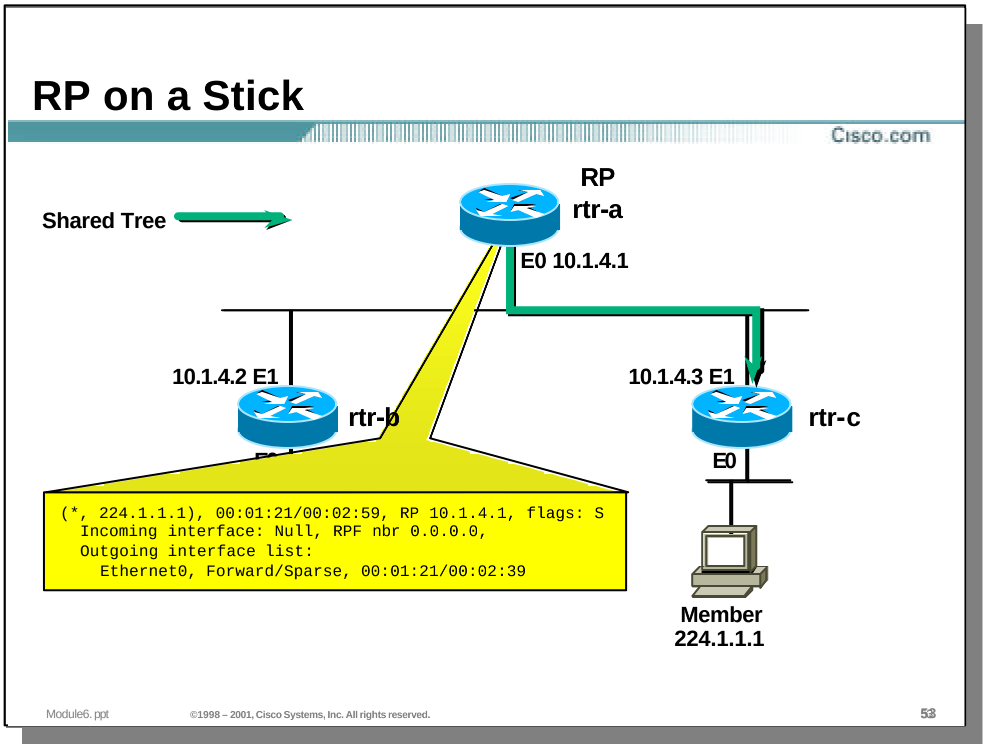
  This is a special situation that occurs under the following conditions:

  - All branches of the Shared Tree are out a single interface on the RP (i.e. there is only a single interface in the (*, G) OIL at the RP.)
  - All sources for the group are out the same interface. (This would result in (S, G) entries with Null OIL’s since the incoming interface can never appear in the OIL of an entry.)

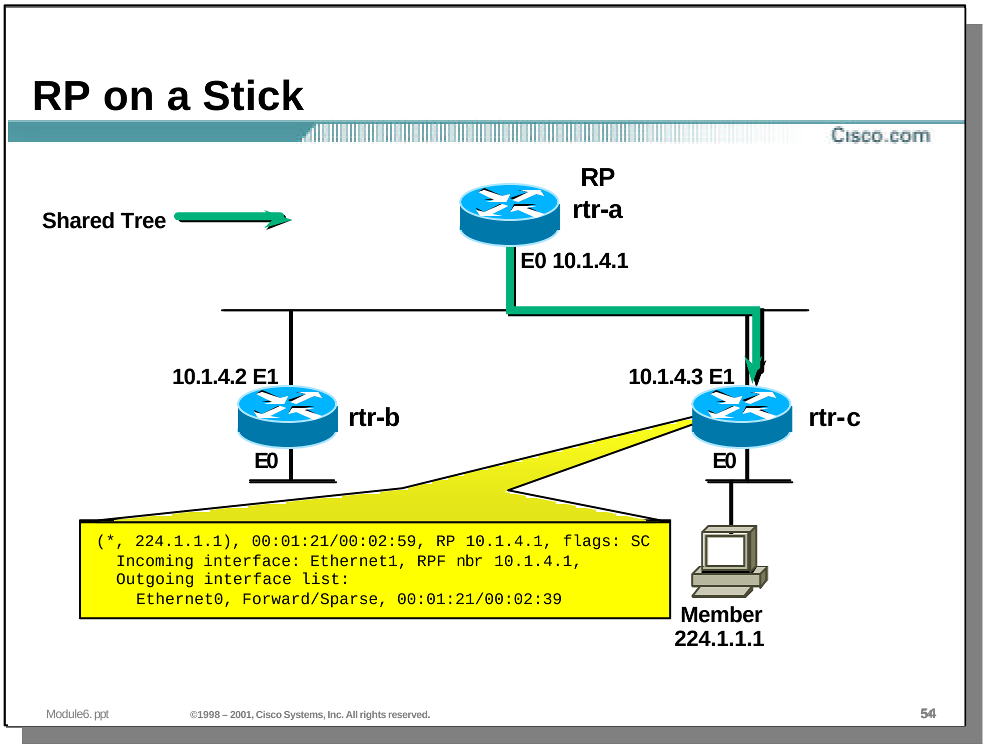
- **Unusually State results in this condition**
  - Special PIM rules had to be created that were not in the original PIMv2 specification in order to avoid situations where :
    - (S, G) traffic flows were incorrectly pruned.
    - (S, G) traffic continued to flow to the RP only to be dropped.
    - (S, G) state would get *stuck* in the RP and the First-Hop router even when the source has long since stopped sending.
  - Problem was solved in IOS 12.0 by:
    - Special “Proxy Join” Timer and
    - Introduction of “Atomic” and “Non-Atomic” (*, G) Joins



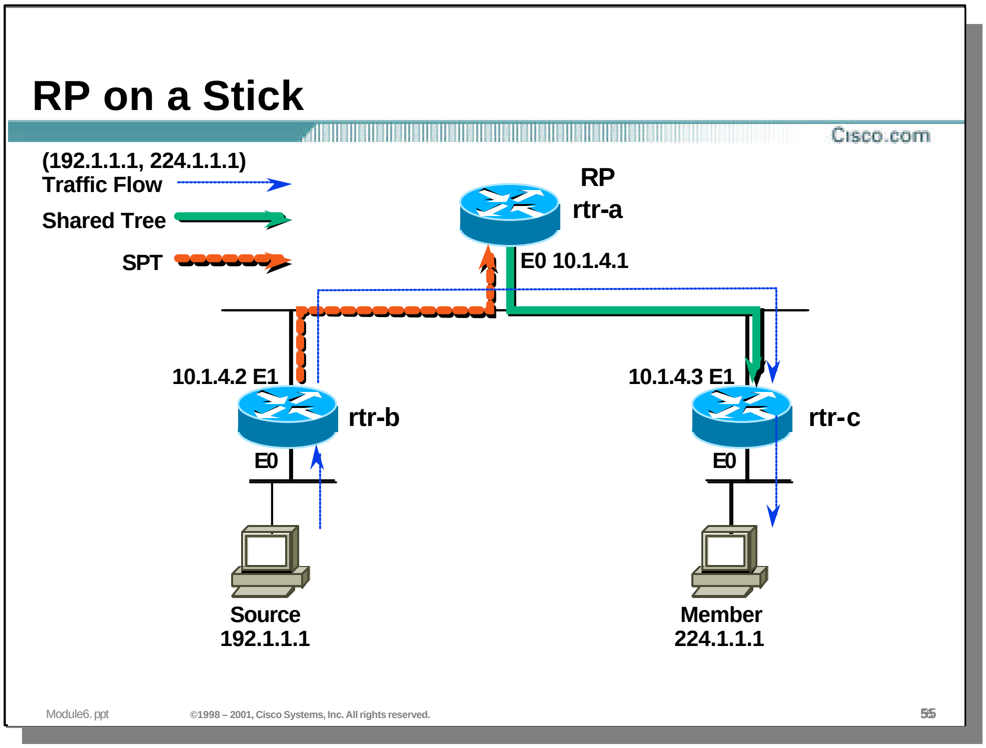
- **RP-on-a-Stick Example**
  - Consider that above network topology where both "rtr-b" and "rtr-c" share a common Ethernet segment with the RP.



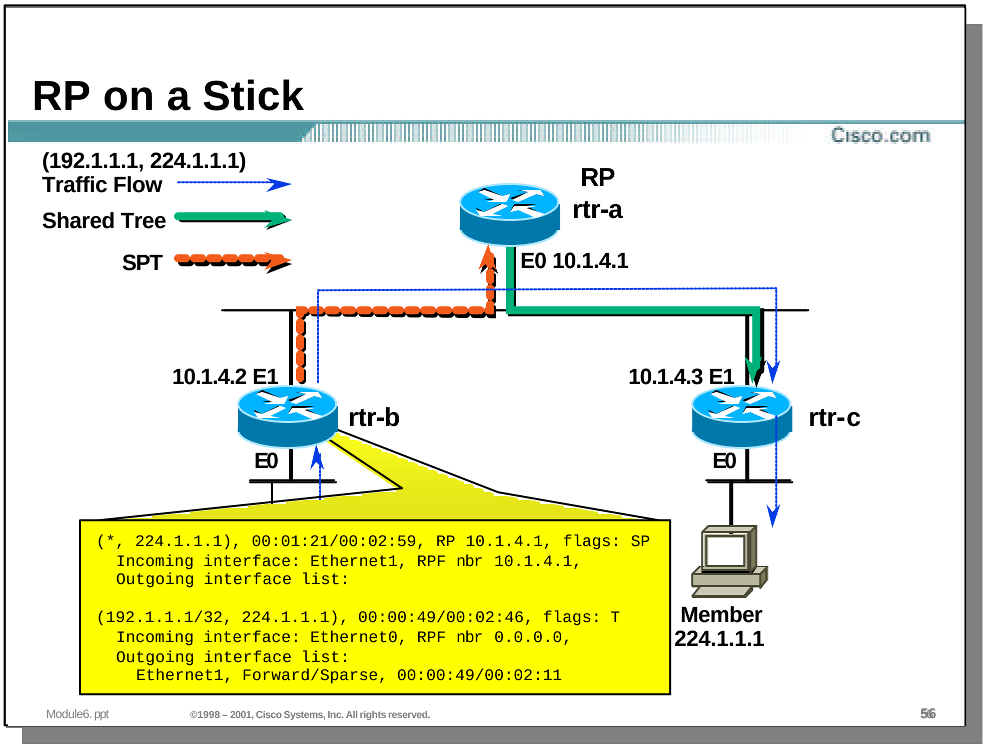
- **RP-on-a-Stick Example**
  - When a host behind “rtr-c” joins group 224.1.1.1, a branch of the Shared Tree is created (shown by the solid arrow) which results in the following state on the RP:
    - (*, 224.1.1.1), 00:01:21/00:02:59, RP 10.1.4.1, flags: S
    - Incoming interface: Null, RPF nbr 0.0.0.0,
    - Outgoing interface list:
    - Ethernet0, Forward/Sparse, 00:01:21/00:02:39



- **RP-on-a-Stick Example**
  - This also results in the following state on "rtr-c":
    - (*, 224.1.1.1), 00:01:21/00:02:59, RP 10.1.4.1, flags: SC
    - Incoming interface: Ethernet1, RPF nbr 10.1.4.1,
    - Outgoing interface list:
    - Ethernet0, Forward/Sparse, 00:01:21/00:02:39



- **RP-on-a-Stick Example**
  - Now assume that source 192.1.1.1 behind "rtr-b" begins sending to group 224.1.1.1. After the normal Register process has completed, a branch of the SPT (shown by the heave dashed arrow) is built from "rtr-b" to the RP. This allows traffic to flow to the members as shown by the thin dashed arrows.

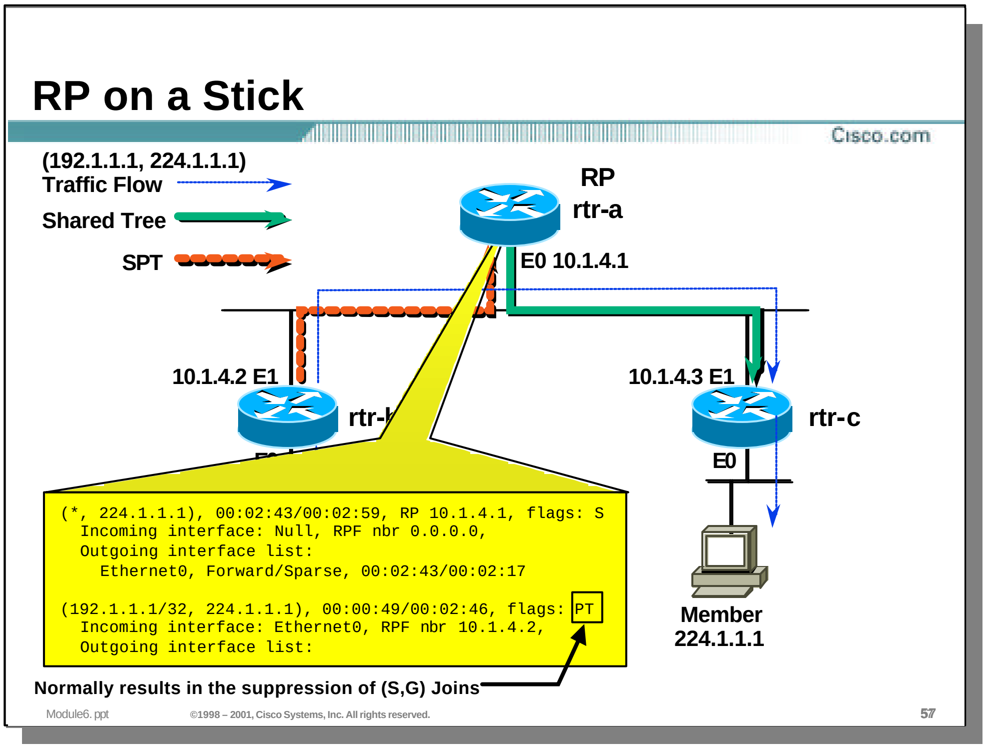


- **RP-on-a-Stick Example**

  The creation of the SPT results in the following state on "rtr-b":

```
(*, 224.1.1.1), 00:01:21/00:02:59, RP 10.1.4.1, flags: SP
  Incoming interface: Ethernet1, RPF nbr 10.1.4.1,
  Outgoing interface list:

(192.1.1.1/32, 224.1.1.1), 00:00:49/00:02:46, flags: T
  Incoming interface: Ethernet0, RPF nbr 0.0.0.0,
  Outgoing interface list:
    Ethernet1, Forward/Sparse, 00:00:49/00:02:11
```



- **RP-on-a-Stick Example**

  The creation of the SPT also results in the following state on the RP:

```
(*, 224.1.1.1), 00:02:43/00:02:59, RP 10.1.4.1, flags: S
  Incoming interface: Null, RPF nbr 0.0.0.0,
  Outgoing interface list:
    Ethernet0, Forward/Sparse, 00:02:43/00:02:17

(192.1.1.1/32, 224.1.1.1), 00:00:49/00:02:46, flags: PT
  Incoming interface: Ethernet0, RPF nbr 10.1.4.2,
  Outgoing interface list:
```

  – Notice that the OIL of the (S, G) entry is Null which, in turn, results in the “P” flag being set. Normally, this would cause the RP to send an (S, G) Prune toward the source to shut off the flow of (S, G) traffic. However in this case, that would starve the host behind “rtr-c” of the desired group traffic. Obviously, something else must be done to prevent this.

RP on a Stick

Cisco.com

- **Solution requires three new concepts**
  - **Atomic & Non-Atomic Joins**
  - **Proxy Join Timer/Flag**
  - **Header-only Registers**

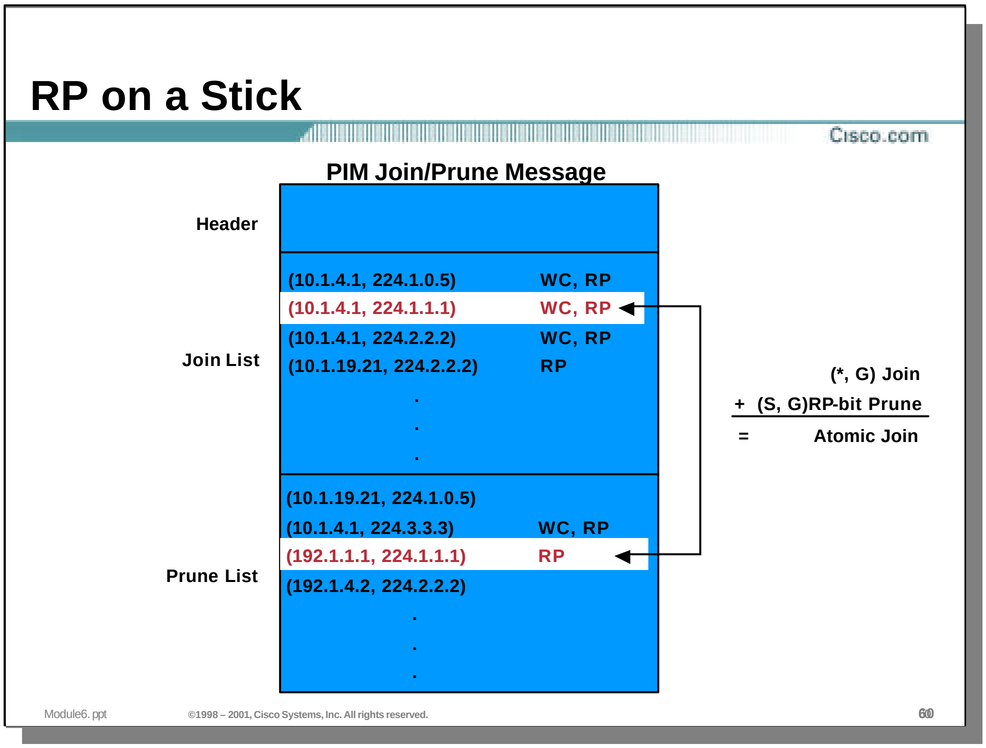
- **RP-on-a-Stick Solution**
  - In order to deal of this special situation, several new concepts were added to the 12.0 implementation of PIM. These are:
    - Atomic vs. Non-Atomic (*, G) Joins
    - The Proxy Join Timer (and its flag) on (S, G) entries
    - Header-only Registers (aka Data-less Registers)
  - Each of the above are discussed in the following pages

RP on a Stick

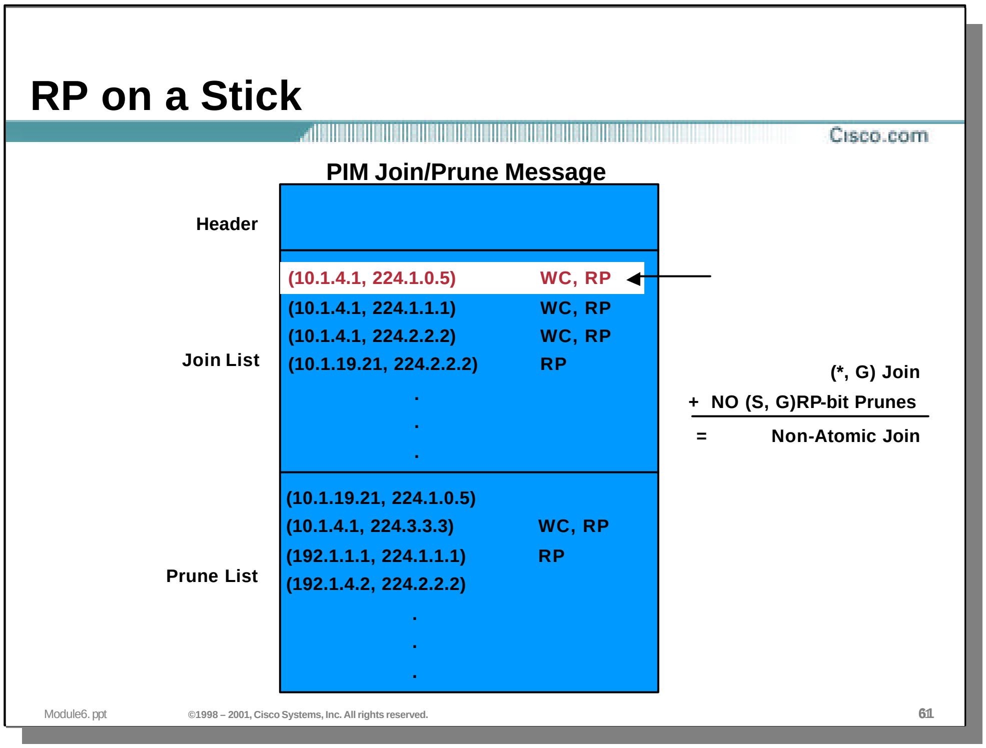
- **Non-Atomic Joins**
  - **Contains (*, G) Join for group “G” only**
    - **This is the type of (*, G) Join we are familiar with**
- **Atomic Joins**
  - **Contains (*, G) Join for group “G” followed by**
  - **(S, G)RP-bit Prunes for all sources in group “G”**
    - **Used to prune unnecessary (S, G) traffic from the Shared Tree after switchover to the SPT.**
  - **All in the same PIM Join/Prune message**

Module6. ppt 59

- **Non-Atomic Joins**
  - This is a PIM Join/Prune message that contains only a (*, G) Join for group “G” in the Join list without any associated (S, G)RP-bit Prunes for group “G” in the Prune list.
    - This is the typical (*, G) Join that has been described in most of the examples in Module 5, “PIM-SM”.
- **Atomic Joins**
  - This is a PIM Join/Prune message that contains a (*, G) Join for group “G” in the Join list ***AND*** a complete list of all (S, G)RP-bit Prunes for group “G” in the Prune list.
    - Remember, these (S, G)RP-bit Prunes are used to Prune specific (S, G) traffic off of ***the Shared Tree*** after a router has joined the SPT directly toward the source.



- **Example: Atomic (*, G) Join**
  - In this example, a (*, G) entry for group 224.1.1.1 in the Join list of the PIM Join/Prune message. (The WC (wildcard) and RP (RP-Tree) bits tell us that this entry is a (*, G) Join to RP 10.1.4.1)
  - In addition, there is an (S, G) entry for group 224.1.1.1 (192.1.1.1, 224.1.1.1) with the RP-bit ***set*** in the Prune list. (I.e. an (S, G)RP-bit Prune exists for group 224.1.1.1).
  - Because both a (*, G) Join along with one or more (S, G)RP-bit Prunes exist in this Join/Prune message for group 224.1.1.1, it is said to contain an *Atomic (*, G) Join for group 224.1.1.1*.



- **Example: Non-Atomic (*, G) Join**
  - Also in this example, is a (*, G) entry for group 224.1.0.5 in the Join list of the PIM Join/Prune message. (The WC (wildcard) and RP (RP-Tree) bits tell us that this entry is a (*, G) Join to RP 10.1.4.1)
  - In addition, there are no (S, G) entries for group 224.1.0.5 ***with the RP-bit set*** in the Prune list. (I.e. there are no (S, G)RP-bit Prunes for group 224.1.0.5).
  - Because only a (*, G) Join exists in this Join/Prune message for group 224.0.1.5 without any corresponding (S, G)RP-bit Prunes, it is said to contain an *Non-Atomic (*, G) Join for group 224.0.1.5*.

RP on a Stick

Cisco.com

- **Proxy Join Timer**
  - **Used on (S, G) entries only**
  - **Started on the RP by the initial (S, G) Register**
    - **If (S, G) OIL is Null AND (*, G) OIL is Non-Null**
  - **Started/Restarted by receipt of a Non-Atomic Join**
    - **The “X” flag indicates when it is running**
    - **Times out in 2 minutes**
  - **Controls the sending of (S, G) Joins and Prunes down the SPT**
  - **When Proxy Join Timer is Running (X flag set)**
    - **Send (S, G) Joins down SPT**
    - **Suppress sending any (S, G) Prunes down SPT**

Module6. ppt 62

- **Proxy Join Timer**

  The Proxy Join Timer only exists on (S, G) entries in the Mroute table. Its purpose is to attract (S, G) traffic to the router even when the OIL of the (S, G) entry is Null. This maintains the flow of (S, G) traffic in cases such as the RP-on-a-Stick.

- **Proxy Join Timer Rules**
  - The Proxy Join Timer is started when the RP receives the first (S, G) Register message when a source goes active if:
    - The OIL is null in resulting (S, G) entry AND the OIL is non-Null in the (*, G) entry. (This is the RP-on-a-Stick condition.)
  - The Proxy Join Timer is started whenever the router receives a Non-Atomic (*,G) Join and an (S, G) entry already exists.
    - This timer runs for 2 minutes unless restarted by the receipt of another Non-Atomic (*, G) Join.
    - When this timer is running on an (S, G) entry, the “X” flag will be displayed in the flags field of the entry.
  - When the Proxy Join Timer is running, the router will:
    - Send periodic (S, G) Joins toward the source *even though the OIL is Null.*
    - Suppress the sending of (S, G) Prunes toward the source *even though the OIL is Null.*

RP on a Stick

Cisco.com

- **Header-only Registers**
  - **Used to keep (S, G) state alive in the RP**
  - **Sent every 2 minutes by First-hop router**
    - **As long as source is still active**
    - **Continues sending until a Register-Stop is received**
  - **Register Messages contains null (S,G) data packet**
    - **Processed by the RP**
    - **Resets (S, G) entry timer at the RP**
    - **RP doesn't send Null packet down Shared Tree**

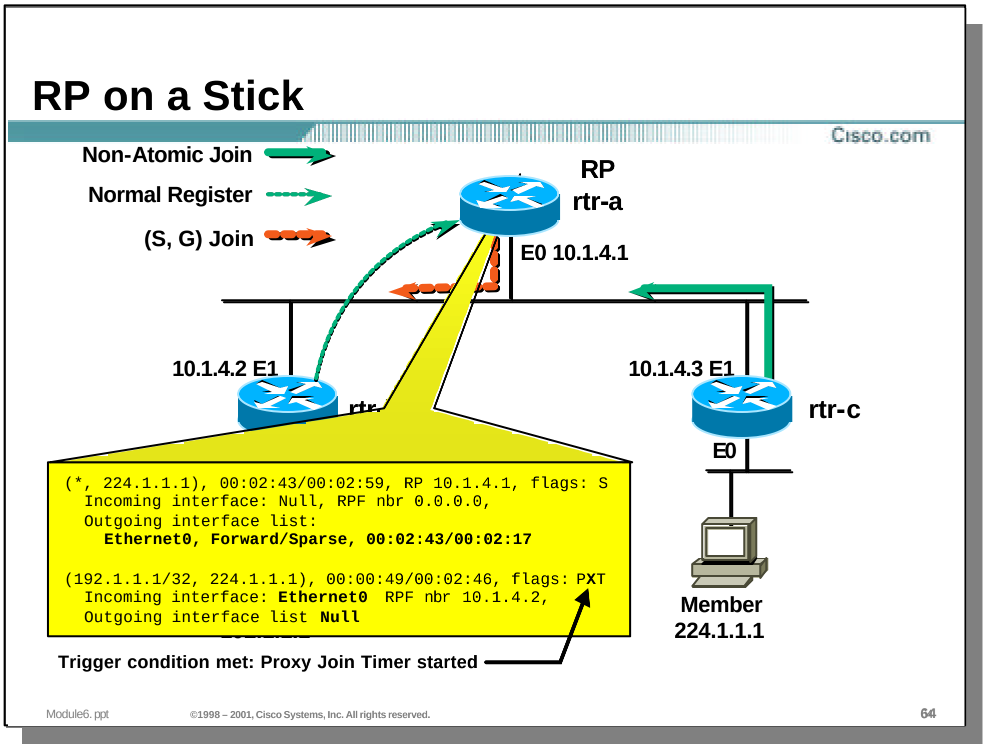
Module6. ppt 63

- **Header-only (Data-less) Registers**

  Normally, the Expire timer of an (S, G) entry is reset to 3 minutes every time the router forwards a packet associated with that entry. However, in the RP-on-a-Stick case, the (S, G) entry has a Null OIL and is therefore not forwarding any packets. This would normally result in the (S, G) entry timing out at the RP. This can not be allowed to happen as it is possible that another member somewhere in the network could join the Shared Tree via another interface. If the (S, G) entry was allowed to timeout, the RP would not be able to trigger the "Batch-Join" to rejoin the SPT when this new member joined. (Because there wouldn't be any (S, G) entry to tell the RP of the active source.)

  To prevent this from happening, the behavior of the First-Hop DR was changed in IOS 12.0 so that (S, G) Header-only (aka Data-less) Registers would be sent periodically (every 2 minutes) to the RP. These Header-only Registers cause the RP to reset the Expire timer in the (S, G) entry thereby preventing it from timing out.

- **Contents of Header-only Registers**
  - Header-only Registers contain a specially formatted null or "data-less" (S, G) packet.
  - These "null" (S, G) packets are not forwarded down the Shared Tree by the RP.

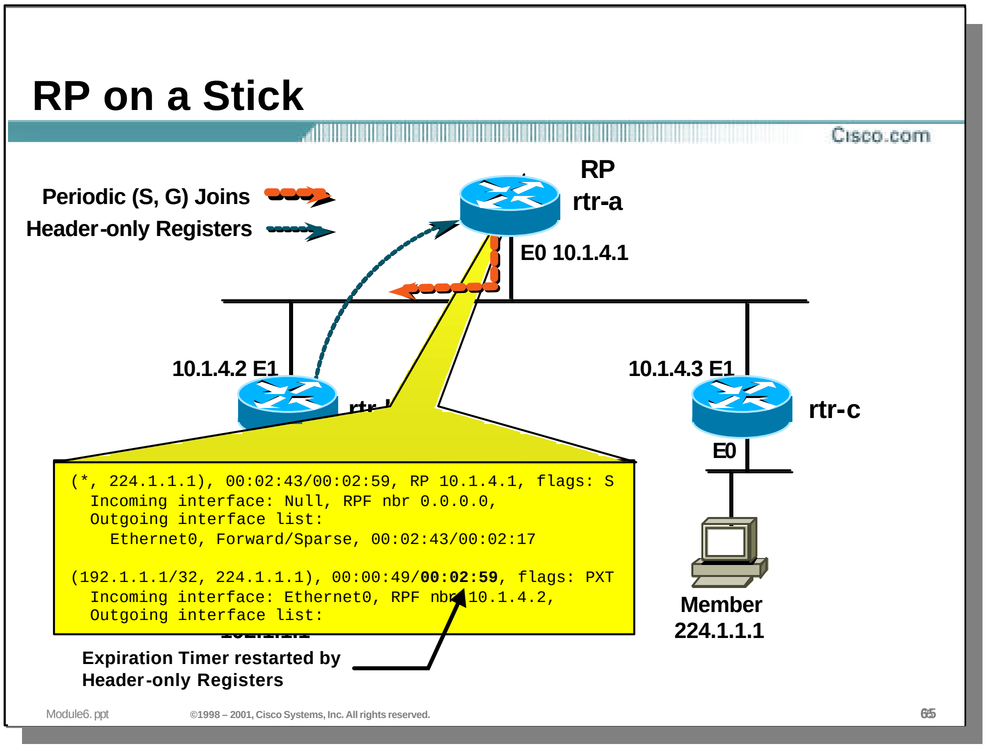


- **RP-on-a-Stick Example**
  - In this example, "rtr-c" is sending Non-Atomic (*, G) Joins to the RP to keep on the Shared Tree. (Note that "rtr-c" has not joined the SPT at this point. This could be due to the SPT-Threshold being set to *Infinity*.)
  - The RP is now running version 12.0 or later of IOS. Therefore, when the Non-Atomic (*, G) Join for group 224.1.1.1 is received, the RP starts the Proxy Join Timer in all (S, G) entries for group 224.1.1.1. This results in the following state in the RP:

```
(*, 224.1.1.1), 00:02:43/00:02:59, RP 10.1.4.1, flags: S
  Incoming interface: Null, RPF nbr 0.0.0.0,
  Outgoing interface list:
    Ethernet0, Forward/Sparse, 00:02:43/00:02:17

(192.1.1.1/32, 224.1.1.1), 00:00:49/00:02:46, flags: PXT
  Incoming interface: Ethernet0, RPF nbr 10.1.4.2,
  Outgoing interface list:
```

  - Notice the "X" flag is set in the above example. This causes the RP to continue sending (S, G) Joins toward the source (even though the OIL is Null) which, in turn, will keep the traffic flowing to the member across the common Ethernet segment.



- **RP-on-a-Stick Example**
  - The First-hop router (rtr-b) is also running version 12.0 or later of IOS and it will therefore send periodic Header-only (S, G) Register messages to the RP.
  - When RP receives these Header-only (S, G) Registers, (roughly every 2 minutes), it resets the Expire timer in the corresponding (S, G) entry in the Mroute table. This results in the following state in the RP:

```
(*, 224.1.1.1), 00:02:43/00:02:59, RP 10.1.4.1, flags: S
  Incoming interface: Null, RPF nbr 0.0.0.0,
  Outgoing interface list:
    Ethernet0, Forward/Sparse, 00:02:43/00:02:17

(192.1.1.1/32, 224.1.1.1), 00:00:49/00:02:59, flags: PXT
  Incoming interface: Ethernet0, RPF nbr 10.1.4.2,
  Outgoing interface list:
```

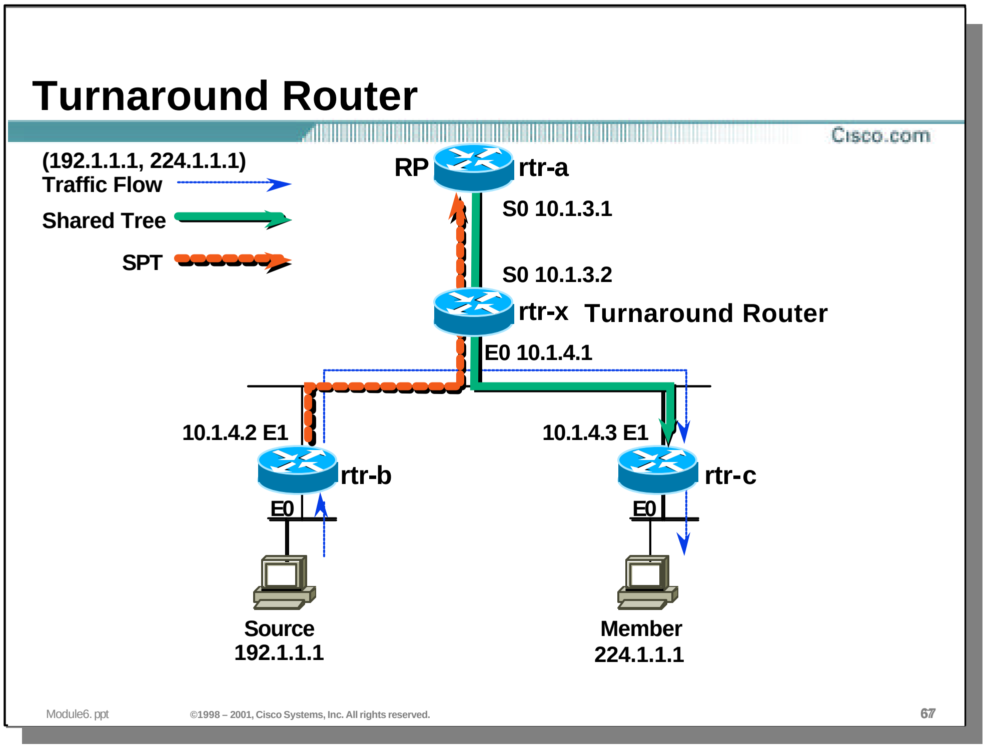
(Notice the Expire timer in the (S, G) entry has been reset.)

Turnaround Router

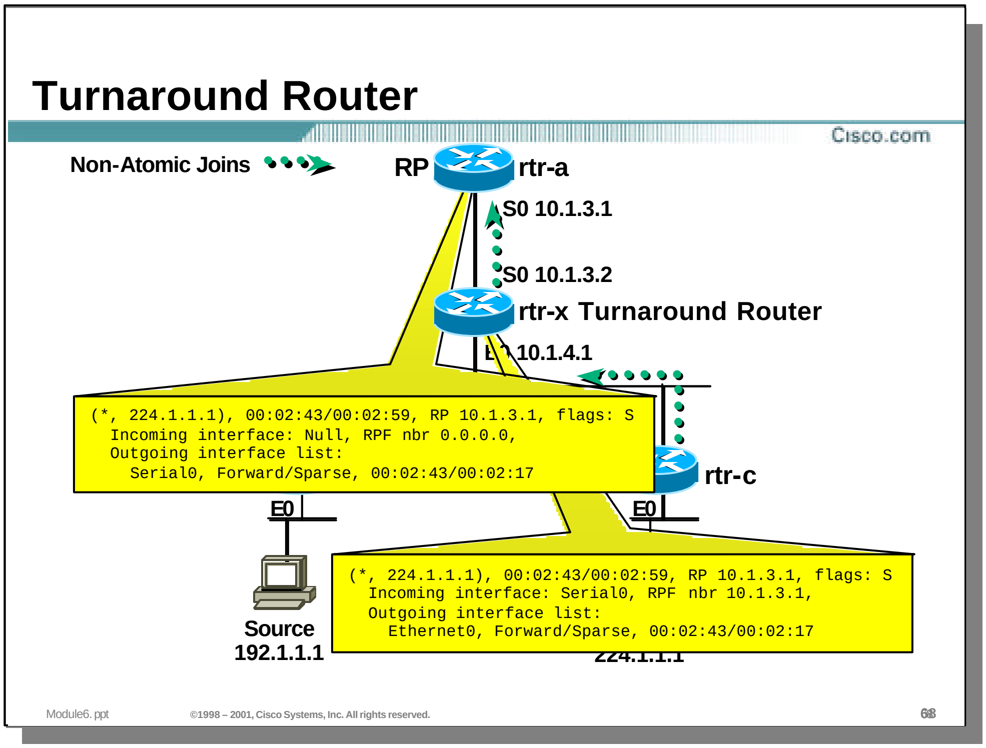
- **Extension of the RP-on-a-Stick Problem**
  - **Occurs when the SPT and the Shared Tree share a single common path**
  - **Want to avoid pulling traffic to the RP unnecessarily**
- **Special “state” in the Turnaround Router**
  - **Uses special techniques to resolve**
    - **Proxy Join Timer**
    - **Atomic and Non-Atomic Joins**
    - **Header-only Registers**

- **Turnaround Router**
  - As it turns out, the RP-on-a-Stick problem is actually a special case of another problem referred to the *Turnaround Router* problem. This situation occurs whenever :
    - There is only a single branch of the Shared Tree and
    - the Shared Tree and a SPT share a common path to the RP.
  - We want to have the (S, G) traffic flow along the SPT toward the RP and “turnaround” at the appropriate router in the network and flow back down the Shared Tree *without* sending unnecessary traffic to the RP.
- **Turnaround Router Solution**
  - Once again, the new concepts of
    - Proxy Join Timer
    - Atomic and Non-Atomic Joins
    - Header-only Registers

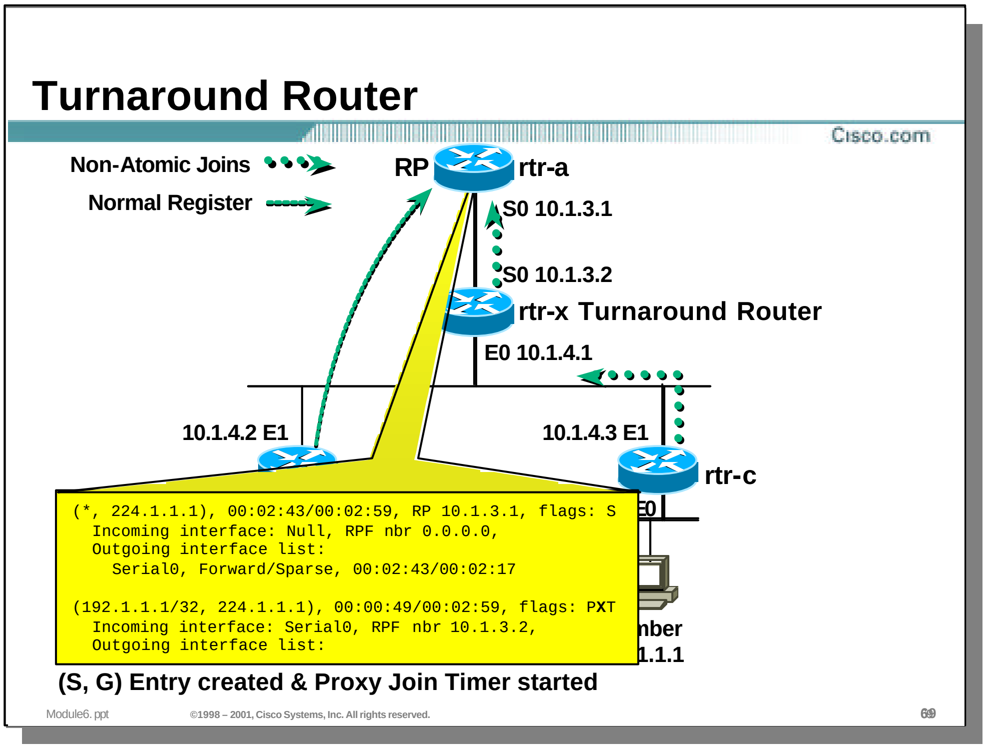
    permit the routers to solve this problem.



- **Turnaround Router Example**
  - In this example, we once again have a single branch of the 224.1.1.1 Shared Tree at the RP.
  - The SPT for source (192.1.1.1, 224.1.1.1) merges with the single branch of the 224.1.1.1 Shared Tree at “rtr-x”. This router is referred to as the *Turnaround Router* because it is here that we want the (S, G) traffic to *turnaround* and flow back down the Shared Tree to the members of group 224.1.1.1.
  - Additionally, we do *not* want the (S, G) traffic flow to all the way to the RP as it is unnecessary traffic *because there is only the single branch of the Shared Tree.* In cases where the number of hops between the Turnaround Router and the RP is large or where the links along this path are congested, the flow of traffic to the RP would simply waste precious network resources.
  - Instead, we want the traffic to only flow as shown by the thin dashed arrows in the drawing above.



- **Turnaround Router — Step-by-Step**
  - Step 1
    - The host connected to "rtr-c" joins group 224.1.1.1. This causes "rtr-c" to create (*, G) state and sends a Non-Atomic (*, G) Join toward the RP.
  - Step 2
    - The Turnaround Router (rtr-x) receives this Non-Atomic (*, G) Join and it too creates (*, G) state and sends a Non-Atomic (*, G) Join to the RP.
  - Step 3
    - The RP receives the (*, G) Join and creates (*, G) state with only **Serial0** in the OIL.



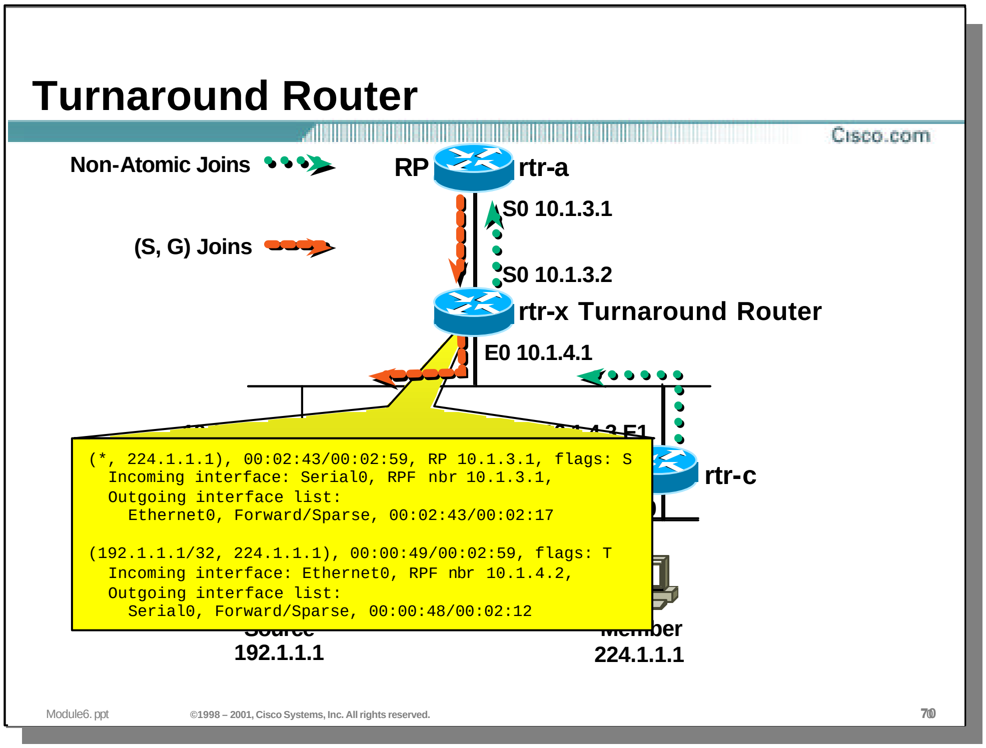
- **Turnaround Router — Step-by-Step**
  - Step 4
    - The source, 192.1.1.1, begins sending to group 224.1.1.1. This causes the first-hop router (rtr-b) to send an (S, G) Register to the RP.
  - Step 5
    - The RP processes the Register message and creates an (S, G) entry. Because the OIL of the newly created (S, G) entry is Null and the OIL of the (*, G) entry is non-Null, the RP starts the Proxy Timer in the (S, G) entry and sends an (S, G) Join toward the source.

  At this point, the state in the RP is as follows:

```
(*, 224.1.1.1), 00:02:43/00:02:59, RP 10.1.3.1, flags: S
  Incoming interface: Null, RPF nbr 0.0.0.0,
  Outgoing interface list:
    Serial0, Forward/Sparse, 00:02:43/00:02:17

(192.1.1.1/32, 224.1.1.1), 00:00:49/00:02:59, flags: PXT
  Incoming interface: Serial0, RPF nbr 10.1.3.2,
  Outgoing interface list:
```

  Notice that the Proxy Join Timer is running (note the “X” flag in the (S,G) entry.)
    - While the Proxy Join Timer is running, the RP will continue to send periodic (S, G) Joins toward the source.
    - The Proxy Join Timer will be restarted each time the RP receives another Non-Atomic Join from “rtr-x”.



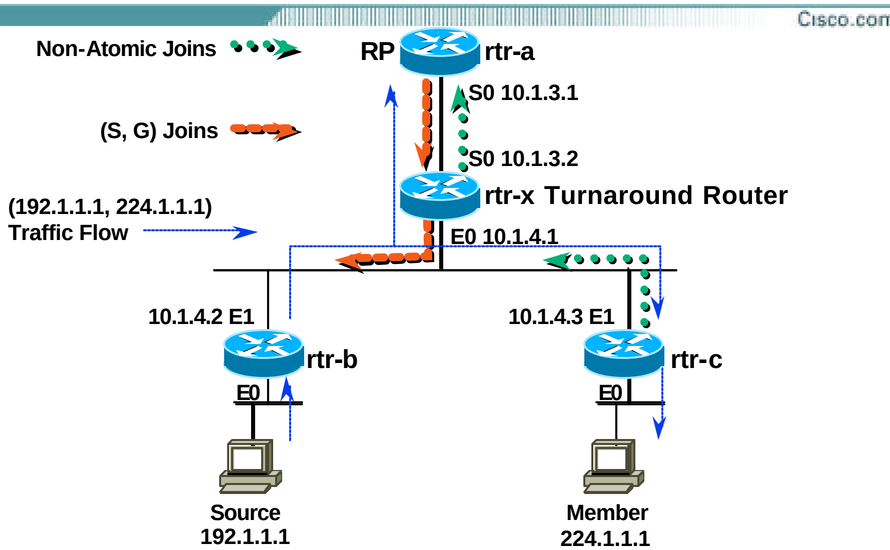
- **Turnaround Router — Step-by-Step**
  - Step 6
    - The (S, G) Join travels hop-by-hop building the SPT from the source to the RP.

  At this point, the state in the Turnaround Router (rtr-x) is as follows:

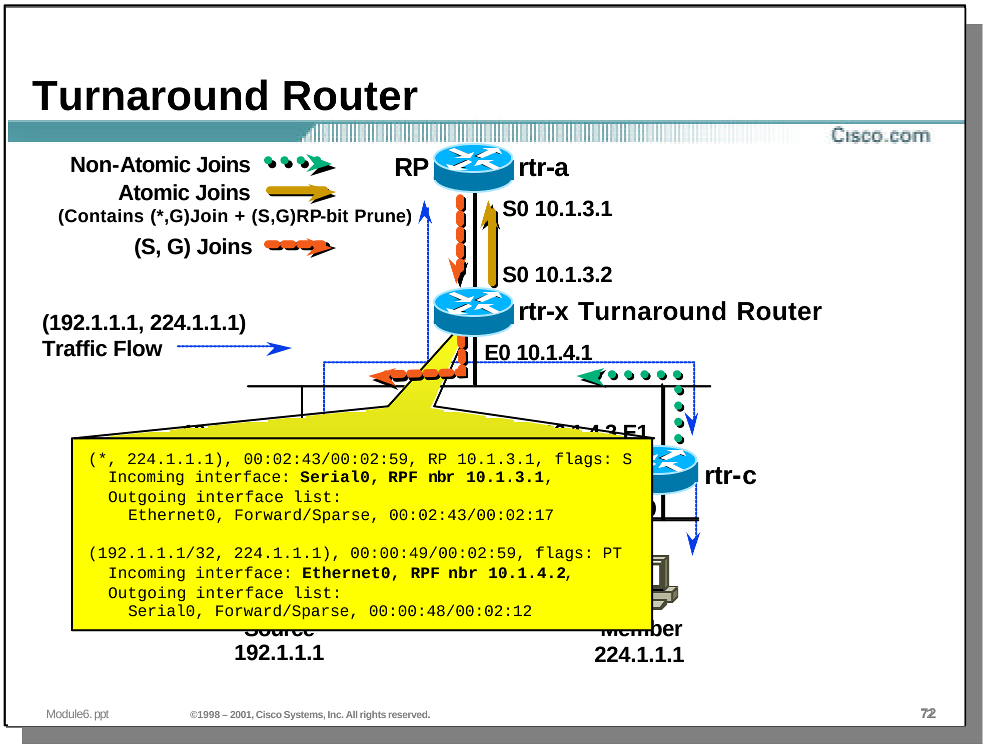
```
(*, 224.1.1.1), 00:02:43/00:02:59, RP 10.1.3.1, flags: S
  Incoming interface: Serial0, RPF nbr 10.1.3.1,
  Outgoing interface list:
    Ethernet0, Forward/Sparse, 00:02:43/00:02:17

(192.1.1.1/32, 224.1.1.1), 00:00:49/00:02:59, flags: T
  Incoming interface: Ethernet0, RPF nbr 10.1.4.2,
  Outgoing interface list:
    Serial0, Forward/Sparse, 00:00:48/00:02:12
```

Turnaround Router



- **Turnaround Router — Step-by-Step**
  - Once “rtr-b” receives the (S, G) Join, traffic begins to flow as shown above.

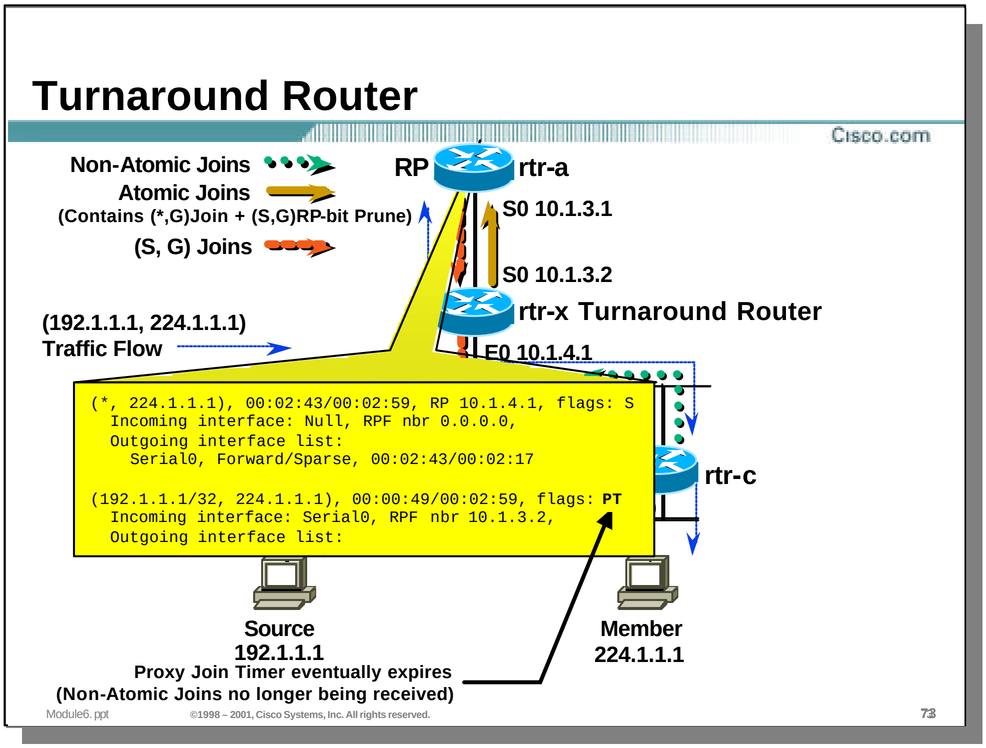


- **Turnaround Router — Step-by-Step**

  Step 7

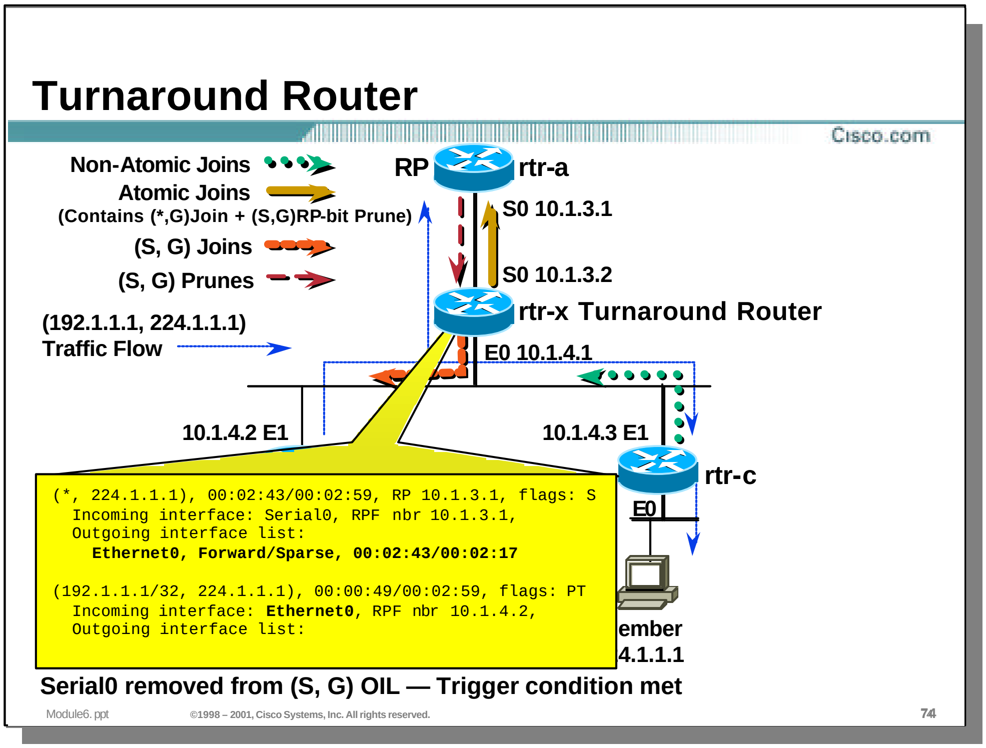
  - Router “rtr-x” detects that the paths of the SPT and the Shared Tree diverge at this point. As a result, “rtr-x” begins sending periodic (S,G)RP-bit Prunes up the Shared Tree in the same Join/Prune message with the periodic (*, G) Joins. In other words, *it begins sending Atomic Joins to the RP instead of Non-Atomic Joins!*

    (Note: Router “rtr-x” knows that the SPT and Shared Tree paths have diverged at this point because the RPF information (Incoming Interface and/or RPF neighbor) of the (S, G) entry is different than the (*,G) entry.)

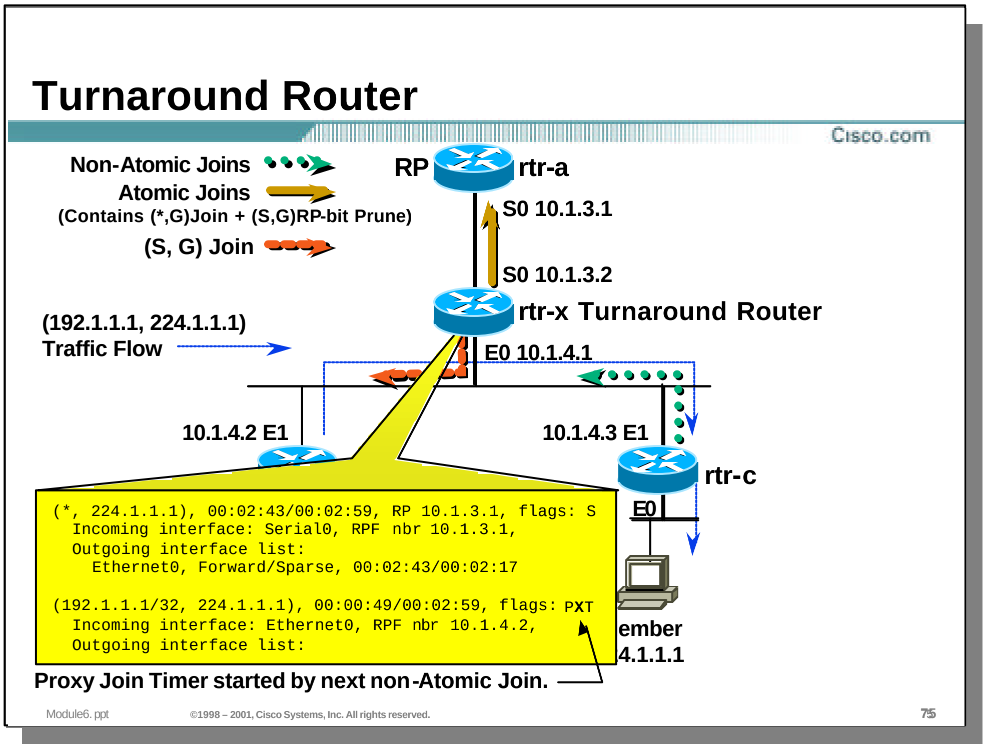


- **Turnaround Router — Step-by-Step**
  - Because the RP is no longer receiving Non-Atomic Joins, the Proxy Join Timer for the (S, G) entry is no longer being restarted and it eventually times out. This is indicated by the “X” flag being clear in the (S, G) entry shown below:

    - (*, 224.1.1.1), 00:02:43/00:02:59, RP 10.1.4.1, flags: S
    - Incoming interface: Null, RPF nbr 0.0.0.0,
    - Outgoing interface list:
    - Serial0, Forward/Sparse, 00:02:43/00:02:17
    -
    - (192.1.1.1/32, 224.1.1.1), 00:00:49/00:02:59, flags: PT
    - Incoming interface: Serial0, RPF nbr 10.1.3.2,
    - Outgoing interface list:



- **Turnaround Router — Step-by-Step**
  - Step 8
    - Now that the Proxy Join Timer is no longer running, the RP resumes its normal behavior and sends an (S, G) Prunes toward the source in response to the arrival of (S, G) packets.
  - Step 9
    - When “rtr-x” receives the (S, G) Prune, it removes **Serial0** from its outgoing interface list. This results in the Turnaround trigger condition in rtr-x.

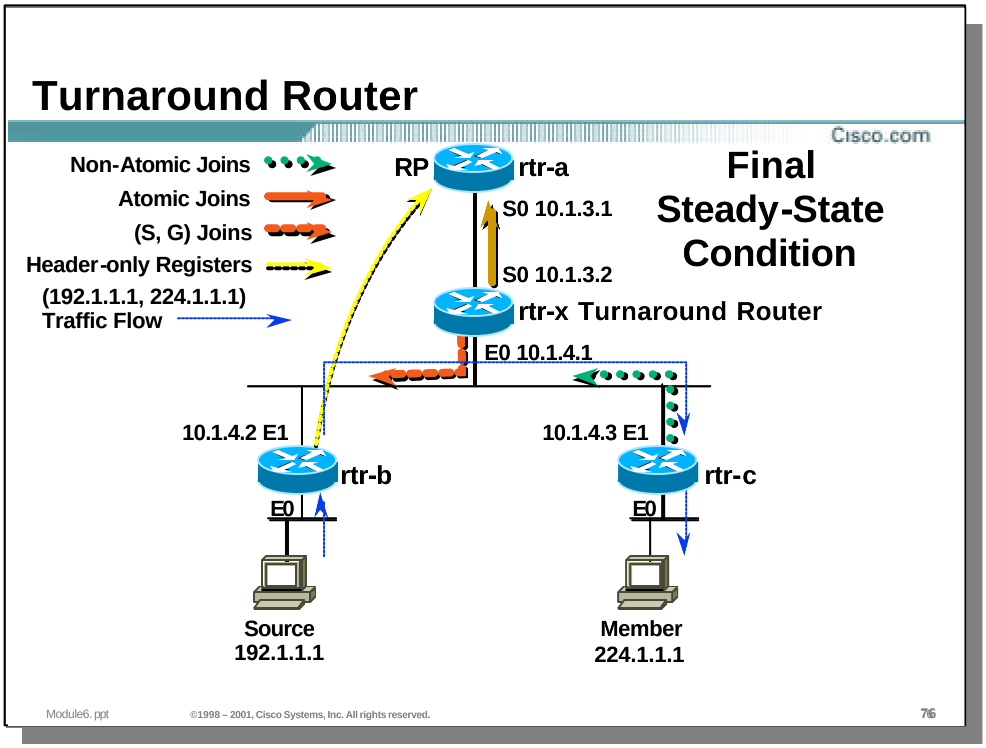


- **Turnaround Router — Step-by-Step**

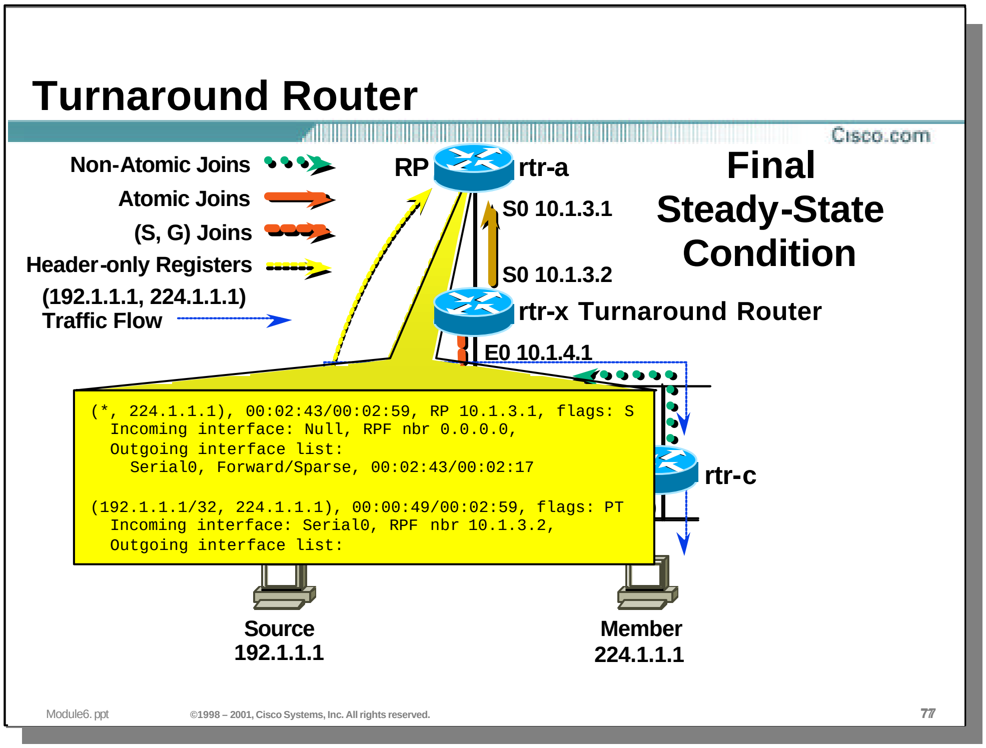
  As a result of Serial0 being removed from the (S, G) OIL, the flow of traffic to the RP is shutoff.

  – Step 10

  - Non-Atomic Joins arriving at “rtr-x” now start the Proxy Join Timer. (Note the “X” flag in the (S, G) entry.) This causes the Turnaround Router (rtr-x) to suppress sending (S, G) Prunes and instead, send (S, G) Joins toward the source. This keeps the traffic flowing as shown.



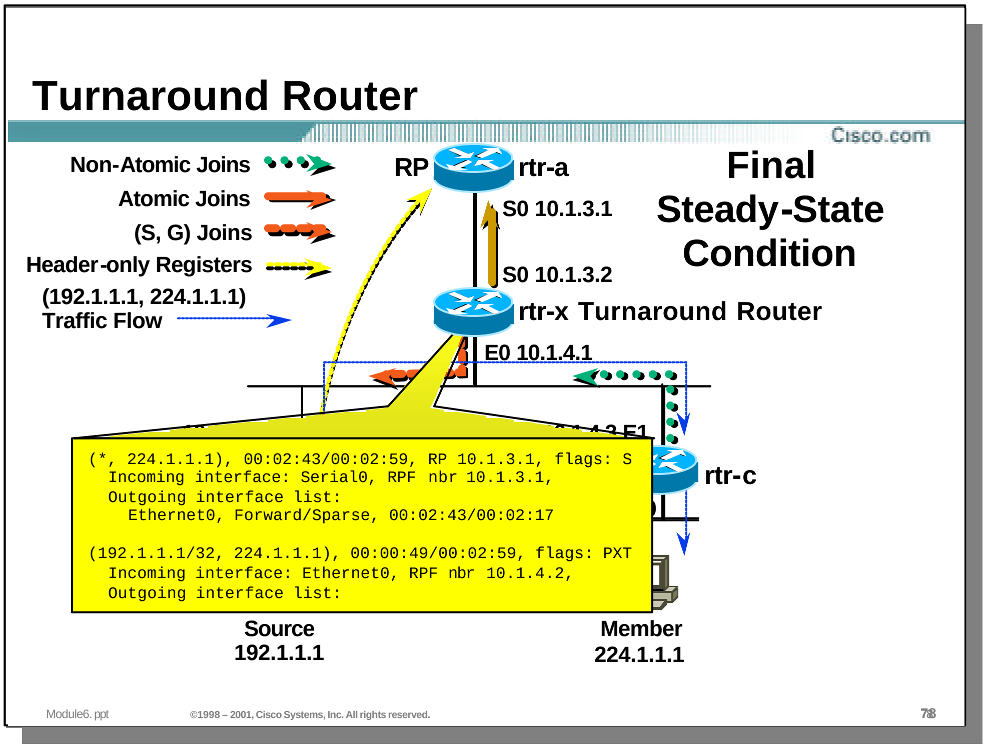
- **Turnaround Router**
  - Step 11
    - Finally, Header-only Registers sent by the First-hop router (rtr-b) continue to reset the Expire timer in the (S, G) entry at the RP. This prevents the (S, G) entry from timing out and being deleted at the RP.



- **Turnaround Router**
  - As a result of the Header-only Registers, the state in the RP will be as follows as long as the source and member remain active:

```
(*, 224.1.1.1), 00:02:43/00:02:59, RP 10.1.3.1, flags: S
  Incoming interface: Null, RPF nbr 0.0.0.0,
  Outgoing interface list:
    Serial0, Forward/Sparse, 00:02:43/00:02:17

(192.1.1.1/32, 224.1.1.1), 00:00:49/00:02:59, flags: PT
  Incoming interface: Serial0, RPF nbr 10.1.3.2,
  Outgoing interface list:
```



- **Turnaround Router**
  - As a result of the Non-Atomic Joins, the state in the Turnaround router will be as follows as long as the source and member remain active :

```
(*, 224.1.1.1), 00:02:43/00:02:59, RP 10.1.3.1, flags: S
  Incoming interface: Serial0, RPF nbr 10.1.3.1,
  Outgoing interface list:
    Ethernet0, Forward/Sparse, 00:02:43/00:02:17

(192.1.1.1/32, 224.1.1.1), 00:00:49/00:02:59, flags: PXT
  Incoming interface: Ethernet0, RPF nbr 10.1.4.2,
  Outgoing interface list:
```

CISCO SYSTEMS

EMPOWERING THE INTERNET GENERATION℠