### Other Related Presentations

#### Multicast Sessions

<table>
<thead>
<tr>
<th>Session #</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>2214</td>
<td>Introduction to IP Multicast</td>
</tr>
<tr>
<td>2215</td>
<td>PIM Multicast Routing</td>
</tr>
<tr>
<td>2216</td>
<td>Deploying IP Multicast</td>
</tr>
<tr>
<td>2217</td>
<td>Advanced IP Multicast Routing</td>
</tr>
</tbody>
</table>

#### MBGP Related Sessions

<table>
<thead>
<tr>
<th>Session #</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>2209</td>
<td>Deploying BGP</td>
</tr>
</tbody>
</table>

### Agenda

- PIM-SM review (forwarding)
- Inter-Domain Multicast
  - MBGP (routing)
  - MSDP (source discovery)
  - Topology Examples
- SSM (Source Specific Multicast)
PIM-SM Review

- Everyone here is a PIM-SM expert, right?

- So just for review...

PIM-SM Shared Tree Join

\[ (*, G) \text{ Join} \rightarrow \text{Shared Tree} \]

\[ (*, G) \text{ State created only along the Shared Tree.} \]
PIM-SM Sender Registration

(S, G) Join

Shared Tree
Source Tree
(S, G) Register
(S, G) Join

Receiver

Traffic Flow
(S, G) Register (unicast)
(RP sends a Register-Stop back to the first-hop router to stop the Register process.)

(S, G) Register-Stop (unicast)

(RP sends a Register-Stop back to the first-hop router to stop the Register process.)

(S, G) traffic begins arriving at the RP via the Source tree.

Traffic Flow
Shared Tree
Source Tree
(S, G) Register
(S, G) Register-Stop

Source

RP

(S, G) traffic begins arriving at the RP via the Source tree.

RP sends a Register-Stop back to the first-hop router to stop the Register process.

(S, G) State created only along the Source Tree.
**PIM-SM Sender Registration**

Source traffic flows natively along SPT to RP. From RP, traffic flows down the Shared Tree to Receivers.

**PIM-SM SPT Switchover**

Last-hop router joins the Source Tree. Additional (S, G) State is created along new part of the Source Tree.
Traffic begins flowing down the new branch of the Source Tree.
Additional (S, G) State is created along the Shared Tree to prune off (S, G) traffic.

(S, G) Traffic flow is now pruned off of the Shared Tree and is flowing to the Receiver via the Source Tree.
(S, G) traffic flow is no longer needed by the RP so it Prunes the flow of (S, G) traffic.

(S, G) Traffic flow is now only flowing to the Receiver via a single branch of the Source Tree.
Agenda

- PIM-SM review (forwarding)
- Inter-Domain Multicast
  - MBGP (routing)
  - MSDP (source discovery)
  - Topology Examples
- SSM (Source Specific Multicast)

Inter-Domain Multicast

- Past history
- In the future
- ISP requirements to deploy now
  - PIM-SM
  - MBGP
  - MSDP
  - Topology Examples
A long time ago, in a galaxy far, far away...

MBONE...
MBONE... vat, nv, wb, sd,...

DVMRP MBONE
- Virtual network overlaid (tunneled) on the unicast Internet infrastructure
- DVMRP MBONE uses RIP-like routing
- Flood and Prune technology
- Initially instantiated by MROUTED, and later implemented by various router vendors
- Very successful in academic circles
Past History

Problem

• DVMRP can’t scale to Internet sizes
  – Distance vector-based routing protocol
  – Periodic updates
    • Full table refresh every 60 seconds
  – Table sizes
    • Internet > 40,000 prefixes
  – Stability
    • Hold-down, count-to-infinity, etc.

In the Future?

• BGMP (Border Gateway Multicast Protocol)
  – Shared tree of domains
    • Bidirectional trees
    • Explict join-model
    • Joins sent toward root domain
  – Single root domain per group
    • Multicast group prefixes assigned by domain
    • MASC proposed as assignment method
  – Requires BGP4+ (aka MBGP)
    • Must carry group prefixes in NLRI field
    • Needed to build bidirectional trees
In the Future

BGMP Example

Domain A

Domain B

Root

224.2.2.2

Domain C

Domain D

Domain E

Domain F

Domain G

In the Future

BGMP Example

Domain A

Domain B

Root

224.2.2.2

Domain C

Domain D

Domain E

Domain F

Domain G
In the Future?

• MASC (Multicast Address Set-Claim)
  – Multicast address space is hierarchical
    • Top of hierarchy is at an Internet exchange
    • Children get address space from parent
    • Results in aggregateable multicast address space
  – Allocation has a lifetime
    • Children must renew address allocation
    • May not receive same space at renewal time
    • Parent may reclaim space at renewal time
    • Permits reallocation of space
    • Complex “garbage collection” problem

In the Future

• BGMP and MASC are a long ways off
  – Both are quite complex to implement
  – Still only in draft proposal stages
• ISP’s are deploying multicast now
  – What are their minimum requirements?
ISP Requirements to Deploy Now

- Want an explicit join protocol for efficiency
  - PIM-SM (forwarding)

- Use existing (unicast) routing model
  - Hmm

- Need interdomain source discovery
  - Separate PIM domain RP’s need to share source information
  - Hmm

ISP Requirements to Deploy Now

- Use existing (unicast) routing model
  - Need same tool-set for multicast as unicast
    - Robust set of peering and policy controls
    - Ability to separate unicast and multicast topologies
    - Use familiar configuration, operation and terminology model
  - Something like BGP but for multicast

- Solution: Multiprotocol BGP (MBGP)
MBGP—Multiprotocol BGP

- MBGP overview
- MBGP capability negotiation
- MBGP NLRI exchange
- MBGP-DVMRP redistribution
- Unicast to Multicast NLRI Translation

MBGP Overview

- MBGP: Multiprotocol BGP (aka multicast BGP in multicast networks)
  - Defined in RFC 2283 (extensions to BGP)
  - Can carry different types of routes
    - Unicast
    - Multicast
  - Both routes carried in same BGP session
  - Does not propagate multicast state info
  - Same path selection and validation rules
    - AS-Path, LocalPref, MED, ...
**MBGP Overview**

- New multiprotocol attributes
  - MP_REACH_NLRI
  - MP_UNREACH_NLRI
- MP_REACH_NLRI and MP_UNREACH_NLRI
  - Address Family Information (AFI) = 1 (IPv4)
    - Sub-AFI = 1 (NLRI is used for unicast)
    - Sub-AFI = 2 (NLRI is used for multicast RPF check)
    - Sub-AFI = 3 (NLRI is used for both unicast and multicast RPF check)

---

**The BGP UPDATE Message**

- Unfeasible Routes Length (2 Octets)
- Withdrawn Routes (Variable)
- Total path Attribute Length (2 Octets)
- Path Attributes (Variable)
- Network Layer Reachability Information (Variable)

- A BGP update is used to advertise a single feasible route to a peer, or to withdraw multiple unfeasible routes
- Each update message contains attributes, like origin, AS-Path, Next-Hop, .... MP_REACH_NLRI
### MP_REACH_NLRI Attribute

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
<th>RFC 1700</th>
<th>May be Zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address Family Identifier (2 Octets)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subsequent Address Family Identifier (1 Octet)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length of the Next-Hop Address (1 Octet)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Network Address of Next-Hop (Variable)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of SNPAs (1 Octet)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length of first SNPA (1 Octet)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First SNPA (Variable)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length of second SNPA (1 Octet)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second SNPA (Variable)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length of last SNPA (1 Octet)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Last SNPA (Variable)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Network Layer Reachability Information (Variable)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### MP_UNREACH_NLRI Attribute

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
<th>Length (I Octet)</th>
<th>Prefix (Variable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address Family Identifier (2 Octets)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subsequent Address Family Identifier (1 Octet)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Withdrawn Routes (Variable)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MBGP Overview

- Separate BGP tables maintained
  - Unicast Routing Information Base (RIB)
  - Multicast Routing Information Base (MRIB)
- Unicast RIB (U-RIB)
  - Contains unicast prefixes for unicast forwarding
  - Populated with BGP unicast NLRI
    - AFI = 1, Sub-AFI = 1 or 3
- Multicast RIB (M-RIB)
  - Contains unicast prefixes for RPF checking
  - Populated with BGP multicast NLRI
    - AFI = 1, Sub-AFI = 2 or 3

MBGP allows different unicast and multicast topologies and different policies
  - Same IP address may have different signification
    - Unicast routing information
    - Multicast RPF information
  - For same IPv4 address two different NLRI with different next-hops
  - Can use existing or new BGP peering topology for multicast
MBGP Overview

• What is in the Cisco IOS® implementation?
  – All the familiar BGP configuration knobs
  – Carries multicast routes in MP_REACH_NLRI
  – NLRI capability negotiation
  – Redistribution between MBGP and IGP (includes DVMRP)
  – Unicast to Multicast NLRI Translation
    • Used for older non-MBGP stub sites

MBGP—Capability Negotiation

• BGP routers establish BGP sessions through the OPEN message
• OPEN message contains optional parameters
• BGP session is terminated if OPEN parameters are not recognised
• New parameter: CAPABILITIES
  • Multiprotocol extension
  • Multiple routes for same destination
New keyword on neighbor command

```
neighbor <foo> remote-as <asn> nlri multicast unicast
```

- Configures router to negotiate either or both NLRI
- If neighbor configures both or subset, common NRLI is used in both directions
- If there is no match, notification is sent and peering doesn’t come up

If neighbor doesn’t include the capability parameters in open, Cisco backs off and reopens with no capability parameters

- Peering comes up in unicast-only mode
- Hidden command

```
neighbor <foo> dont-capability-negotiate
```
MBGP—Capability Negotiation

BGP Session for Unicast and Multicast NLRI

AS 123

192.168.100.0/24

AS 321

192.168.100.2 open active, local address 192.168.100.1
BGP: 192.168.100.2 sending OPEN, version 4
BGP: 192.168.100.2 rcv OPEN w/option parameter type: 2, len: 6
BGP: 192.168.100.2 OPEN has CAPABILITY code: 1, length 4
BGP: 192.168.100.2 OPEN has MP_EXT CAP for afi/safi: 1/1
BGP: 192.168.100.2 went from Active to OpenSent
BGP: 192.168.100.2 went from OpenSent to OpenConfirm
BGP: 192.168.100.2 went from OpenConfirm to Established

MBGP NLRI Exchange

- BGP/MBGP configuration allows you to:
  - Define which NLRI type are exchanged (unicast, multicast, both)
  - Set NLRI type through route-maps (redistribution)
  - Define policies through standard BGP attributes (for unicast and/or multicast NLRI)
- No redistribution allowed between MBGP and BGP tables
  - NLRI type can be set with set nlri route-map command
MBGP NLRI Exchange

- MRIB is populated by:
  - Receiving AFI/SAFI 1/2 MP_REACH_NLRI from neighbors
  - Configured/stored locally by:

    network <foo> <foo-mask> [nlri multicast unicast]
    redistribute <unicast> route-map <map>
    aggregate-address <foo> <foo-mask> [nlri multicast unicast]
    neighbor <foo> default-originate [nlri multicast unicast]

- Note: Syntax changes in 12.07T/12.1 forward

MBGP — NLRI Information

Congruent Topologies

BGP Session for Unicast and Multicast NLRI

```
router bgp 321
  neighbor 192.168.100.1 remote-as 123 nlri unicast multicast
  network 192.192.25.0 255.255.255.0 nlri unicast multicast
  no auto-summary
```

Sender

192.168.100.1/24

Receiver

192.168.10.0/24

AS 123

AS 321
MBGP — NLRI Information

Congruent Topologies

Routing Update

Unicast Information
NLRI: 192.192.25/24
AS_PATH: 321
MED:
Next-Hop: 192.168.100.2
...

Multicast Information
MP_REACH_NLRI: 192.192.25/24
AFI: 1, Sub-AFI: 2 (multicast)
AS_PATH: 321
MED:
Next-Hop: 192.168.100.2
...

MBGP MBGP — NLRI Information

Incongruent Topologies

router bgp 321
... network 192.192.25.0 niri unicast multicast neighbor 192.168.100.1 remote-as 123 niri unicast neighbor 192.168.200.1 remote-as 123 niri multicast
MBGP — NLRI Information

Incongruent Topologies

```
AS 123
  .1 Unicast Traffic 192.168.100.0/24 .2
  .1
  .1

192.168.200.0/24

AS 321
  192.192.25.0/24

Multicast Traffic .2

Sender

Unicast Information
NLRI: 192.192.25/24
AS_PATH: 321
MED:
Next-Hop: 192.168.100.2

Routing Update

Multicast Information
MP_REACH_NLRI: 192.192.25/24
AFI: 1, Sub-AFI: 2
AS_PATH: 321
MED:
Next-Hop: 192.168.200.2

Routing Update
```
MBGP — NLRI Information

Incongruent Topologies

Unicast RIB

<table>
<thead>
<tr>
<th>Network</th>
<th>Next-Hop</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>192.192.25.0/24</td>
<td>192.168.100.2</td>
<td>321</td>
</tr>
</tbody>
</table>

Multicast RIB

<table>
<thead>
<tr>
<th>Network</th>
<th>Next-Hop</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>192.192.25.0/24</td>
<td>192.168.200.2</td>
<td>321</td>
</tr>
</tbody>
</table>

New ‘nlri’ keyword used to control RIB population. (e.g. ‘network’ command)

- Unicast RIB only

routerr bgp 100
network 160.10.1.0 255.255.255.0 nlri unicast
network 160.10.3.0 255.255.255.0 nlri unicast
no auto-summary
New ‘nlri’ keyword used to control RIB population. (e.g. ‘network’ command)
- Unicast RIB only
- Multicast RIB only
Populating the MRIB

- Just to restate

  ```
  network <foo> <foo-mask> [nlri multicast unicast]
  redistribute <unicast> route-map <map>
  aggregate-address <foo> <foo-mask> [nlri multicast unicast]
  neighbor <foo> default-originate [nlri multicast unicast]
  ```

DVMRP <-> MBGP Redistribution

- DVMRP can be distributed as any other IGP

  ```
  router bgp <asn>
  redistribute dvmrp route-map <map>
  ```

- You can do your typical set operations

- Used to connect legacy sites to the current native multicast Internet.

- Currently used at the AMES MIX to connect with the old DVMRP MBONE
Unicast-Multicast NLRI Translation

- Strictly transitional—not recommended
- BGP stubs that don’t have MBGP support need to get their routes into the MBGP backbone
- Stub gets external routes via existing BGP peering
- Use hidden command
  neighbor <foo> translate-update [nlri unicast multicast]

Unicast-Multicast NLRI Translation

- BGP Update received by translating router is translated into an MP_REACH_NLRI attribute
  - As if the neighbor sent AFI 1/SAFI 2 routes
Unicast-Multicast NLRI Translation

Arriving Unicast update intercepted by ‘translate-update’ Front-end
A translated Multicast update is created & passed to the IN Process
Original Unicast update is passed on to the IN Process
Both updates processed normally by the IN Process

router bgp 1
   neighbor 192.168.1.1 remote-as 4
   neighbor 192.168.1.1 translate-update nrlr unicast multicast
   neighbor 192.180.1.1 remote-as 3 nrlr unicast multicast
   neighbor 192.170.1.1 remote-as 2 nrlr unicast multicast

AS 2
AS 4
AS 1
LO0 192.168.1.1
192.168.100.0/24
192.168.25.0/24

LO0 192.170.1.1

Unicast-Multicast NLRI Translation

BGP Only Stub AS

LO0 192.168.1.1
192.168.100.0/24
192.192.25.0/24

AS 4

Unicast Update

NLR: 192.192.25/24
AS_PATH: 4
MED:
Next-Hop:
...

MBGP AS

AS 1

AS 2

AS 3

Unicast-Multicast NLRI Translation

MP_REACH_NLRI: 192.192.25/24
AFI: 1, Sub-AFI: 2 (multicast)
AS_PATH: 1, 4
MED:
Next-Hop:
...

AS 2

LO0 192.170.1.1

MBGP AS

Lo0 192.168.100.0/24

AS 1

MBGP AS

192.168.100.0/24

192.192.25.0/24

AS 2

MBGP AS

AS 3

MBGP AS

AS 3

Lo0 192.170.1.1
**Unicast-Multicast NLRI Translation**

- **BGP Only Stub AS**
  - 192.168.1.1
  - 192.168.100.0/24
  - 192.192.25.0/24

- **MBGP AS**
  - 192.192.25/24
  - AS_PATH: 1, 4
  - MED:
  - Next-Hop:

- **AS 2**
  - LO0 192.170.1.1

- **AS 3**
  - LO0 192.170.1.1

**MBGP—Summary**

- **Solves part of inter-domain problem**
  - Can exchange multicast routing information
  - Uses standard BGP configuration knobs
  - Permits separate unicast and multicast topologies if desired

- **Still must use PIM to:**
  - Build distribution trees
  - Actually forward multicast traffic
  - PIM-SM recommended
    - But there’s still a problem using PIM-SM here… (more on that later)
ISP Requirements to Deploy Now

- Want an explicit join protocol for efficiency
  - √ PIM-SM (forwarding)
- Use existing (unicast) operation model
  - √ MBGP (routing)
- Need interdomain source discovery
  - Separate PIM domain RP’s need to share source information
  - Hmmm

ISP Requirements to Deploy Now

- Will not share RP with competitors
  - Firm requirement
  - Third-party resource dependency
    - “If my customers are multicasting on group G whose RP is in my competitor’s network and that RP goes down, my customers lose connectivity.”
- Want flexibility re: RP placement
  - May need to place RP(s) someplace other than a single interconnect point
ISP Requirements to Deploy Now

- Must interconnect PIM-SM domains
  - Inter-domain rendezvous mechanism?
    - Requires dynamic DNS (or something similar)
    - Still results in third-party RP problem
  - Interconnect using shared trees
    - That’s BGMP! Can’t wait
  - Interconnect using source trees
    - Need a way to discover all multicast sources
      - Hmmmm. Interesting idea!

- Solution: MSDP
  - Multicast Source Discovery Protocol

MSDP—Multicast Source Discovery Protocol

- MSDP concepts
- MSDP design points
- MSDP example
- Cisco MSDP implementation
- MSDP configuration
- MSDP application—Anycast RP
MSDP Concept

• Simple but elegant
  – Utilize inter-domain source trees
  – Reduces problem to locating active sources
  – RP or receiver last-hop can join inter-domain source tree

MSDP Concepts

• Works with PIM-SM only
  – RP’s knows about all sources in a domain
    • Sources cause a “PIM Register” to the RP
    • Can tell RP’s in other domains of its sources
      – Via MSDP SA (Source Active) messages
  – RP’s know about receivers in a domain
    • Receivers cause a “(*, G) Join” to the RP
    • RP can join the source tree in the peer domain
      – Via normal PIM (S, G) joins
MSDP Overview

MSDP Example

MSDP Peers
Source Active Messages

SA Message 192.1.1.1, 224.2.2.2

Register 192.1.1.1, 224.2.2.2

MSDP Example

MSDP Peers

Domain A
Domain B
Domain C
Domain D
Domain E

Join (*, 224.2.2.2)

Register (S, 224.2.2.2)
MSDP Overview

MSDP Example

MSDP Peers

Multicast Traffic

Domain A

Domain B

Domain C

Domain D

Domain E

RP

Join (S, 224,2,2,2)

MSDP Peers

Multicast Traffic
MSDP Overview

MSDP Example

MSDP Design Points

- MSDP peers talk via TCP connections
- Source Active (SA) messages
  - Peer-RPF forwarded to prevent loops
    - RPF check on AS-PATH back to the peer RP
      - other rules from draft apply
    - If successful, flood SA message to other peers
    - Stub sites can be configured to accept all SA messages (similar to static default for routing)
      - Since they have only one exit (e.g., default peer)
  - MSDP speaker may cache SA messages
    - Reduces join latency
MSDP Peers

- MSDP establishes a neighbor relationship between MSDP peers
  - Peers connect using TCP port 639
  - Peers send keepalives every 60 secs (fixed)
  - Peer connection reset after 75 seconds if no MSDP packets or keepalives are received
- MSDP peers must run BGP!
  - May be an MBGP peer, a BGP peer or both
  - Exception: BGP is unnecessary when peering with only a single MSDP peer.

MSDP SA Messages

- MSDP SA Message Contents
  - One or more messages (in TLV format)
    - Keepalives
    - Source Active (SA) Messages
    - Source Active Request (SA-Req) Messages
    - Source Active Response (SA-Resp) Message
  - Source Active (SA) Messages
    - Used to advertise active Sources in a domain
    - Can also carry initial multicast packet from source
      - Hack for Bursty Sources (ala SDR)
    - SA Message Contents:
      - IP Address of Originator (usually an RP)
      - Number of (S, G)'s pairs being advertised
      - List of active (S, G)'s in the domain
      - Encapsulated Multicast packet
MSDP SA Messages

- MSDP Message Contents (cont.)
  - SA Request (SA-Req) Messages
    - Used to request a list of active sources for a group
    - Sent to an MSDP SA Cache Server
    - Reduces Join Latency to active sources
    - SA Request Messages contain:
      - Requested Group Address
  - SA Response (SA-Resp) Messages
    - Sent in response to an SA Request message
    - SA Response Messages contain:
      - IP Address of Originator (usually an RP)
      - Number of (S, G)’s pairs being advertised
      - List of active (S, G)’s in the domain
  - Keepalive messages
    - Used to keep MSDP peer connection up

Receiving SA Messages

- Skip RPF Check and process SA if:
  - Sending MSDP peer = only MSDP peer
    (i.e. the ‘default-peer’ or only ‘msdp-peer’ configured.)
  - Sending MSDP peer = Mesh-Group peer
- RPF Check the received SA message.
  - Lookup best BGP path to RP in SA message
    - Search MBGP RIB first then BGP RIB
      - If path to RP not found, RPF Check Fails; ignore SA message
  - Sending MSDP Peer = BGP peer?
    - Yes: Is best path to RP via this BGP peer?
      - If yes, RPF Check Succeeds; process SA message
    - No: Is next AS in best path to RP = AS of MSDP peer?
      - If yes, RPF Check Succeeds; process SA message
Receiving SA Messages

• Detailed RPF Check rules
  – Case 1: Sending MSDP Peer = iBGP peer
    • Is best path to RP via this BGP peer?
      – Translation: Is the address of the iBGP peer that advertised the best path to the RP = address of the sending MSDP peer?
  – Case 2: Sending MSDP Peer = eBGP peer
    • Is best path to RP via this BGP peer?
      – Translation: Is the AS of eBGP peer = next AS in the best path to the RP?
  – Case 3: Sending MSDP Peer != BGP peer
    • Is the next AS in best path to RP = AS of the sending MSDP peer?
      – Translation: None needed.

MSDP/iMBGP mesh-peering

MSDP SA

MSDP/iMBGP mesh-peering

RPF Success!
RPF Check Example

RPF rule when MSDP == internal (m)BGP peer

BGP Table router B

Network Next Hop Path
*> 172.16.0.2/32 172.16.3.1 i

Who is the iBGP peer advertising this route, in our example 172.16.3.1

MSDP Peers router B

MSDP Peer AS State
172.16.3.1 1 UP
172.16.4.1 1 UP

Is the MSDP == BGP peer

RPF Failure!

RPF Check Example

RPF rule when MSDP == external (m)BGP peer

BGP Neighbours router B

Neighbor AS Up/Down
172.16.3.1 1 06:06:08
172.16.4.1 3 06:09:06

BGP Table

Network Next Hop Path
* 172.16.0.2/32 172.16.3.1 i
172.16.0.2/32 172.16.4.1 3 1 1

Who is the BGP peer advertising this route

Is the MSDP AS == Last AS in RP route

RPF Success!
RPF Check Example

**RPF rule when MSDP == external (m)BGP peer**

MSDP Peers router B

<table>
<thead>
<tr>
<th>MSDP Peer</th>
<th>AS</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>172.16.3.1</td>
<td>1</td>
<td>UP</td>
</tr>
<tr>
<td>172.16.4.1</td>
<td>3</td>
<td>UP</td>
</tr>
</tbody>
</table>

BGP Neighbors router B

<table>
<thead>
<tr>
<th>Neighbor</th>
<th>AS</th>
<th>Up/Down</th>
</tr>
</thead>
<tbody>
<tr>
<td>172.16.3.1</td>
<td>1</td>
<td>06:06:08</td>
</tr>
<tr>
<td>172.16.4.1</td>
<td>3</td>
<td>06:09:06</td>
</tr>
</tbody>
</table>

BGP Table

<table>
<thead>
<tr>
<th>Network</th>
<th>Next Hop</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>*&gt; 172.16.0.2/32</td>
<td>172.16.3.1</td>
<td>1 i</td>
</tr>
<tr>
<td>172.16.0.2/32</td>
<td>172.16.4.1</td>
<td>3 i</td>
</tr>
</tbody>
</table>

**Is the MSDP AS == Last AS in RP route**

RPF Failure!

Who is the BGP peer advertising this route

RPF Check Example

**RPF rule when MSDP != (m)BGP peer**

MSDP Peers router B

<table>
<thead>
<tr>
<th>MSDP Peer</th>
<th>AS</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>172.16.3.1</td>
<td>1</td>
<td>UP</td>
</tr>
<tr>
<td>172.16.4.1</td>
<td>3</td>
<td>UP</td>
</tr>
</tbody>
</table>

BGP Table

<table>
<thead>
<tr>
<th>Network</th>
<th>Next Hop</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>*&gt; 172.16.0.2/32</td>
<td>172.16.3.1</td>
<td>1 i</td>
</tr>
<tr>
<td>172.16.0.2/32</td>
<td>172.16.4.1</td>
<td>3 i</td>
</tr>
<tr>
<td>*&gt; 172.16.4.0/24</td>
<td>172.16.3.1</td>
<td>1 i</td>
</tr>
<tr>
<td>*&gt; 172.16.3.0/24</td>
<td>172.16.4.1</td>
<td>3 i</td>
</tr>
</tbody>
</table>

**Is the MSDP AS == Last AS in RP route**

RPF Success!
**Cisco MSDP Implementation**

- Cisco implementation current with ID:
  - draft-ietf-msdp-spec-02.txt
- Multiple peer support
  - Peer with BGP, MBGP, or static peers
- SA caching (off by default)
- Sending and receiving SA-requests
- Sending and receiving SA-responses
Cisco MSDP Implementation

- SA input and output filtering
- SA-request input filtering
- Default peer support
  - So a tail site can MSDP with a backbone provider without requiring the two to BGP peer
- Triggered join support when creating an (S,G) learned by MSDP
- Mesh groups
  - Reduces RPF-flooding of SA messages between fully meshed MSDP peers

MSDP Configuration

- Configure peers
  ip msdp peer <ip-address> [connect-source <i/f>]
- Configure default peer
  ip msdp default-peer <ip-address> [prefix-list acl]
- SA caching
  ip msdp cache-sa-state [list <acl>]
- Mesh groups
  ip msdp mesh-group <name> <ip-address>
MSDP Configuration

- Filtering
  - Can filter SA in/out, groups, with acls or route-maps
- TTL Scoping
  - `ip msdp ttl-threshold <ip-address> <ttl>`
- For more configuration commands see: ftp://ftpeng.cisco.com/ipmulticast/msdp-commands

ISP Requirements to Deploy Now

- Want an explicit join protocol for efficiency
  - ✓ PIM-SM (forwarding)
- Use existing (unicast) operation model
  - ✓ MBGP (routing)
- Need interdomain source discovery
  - ✓ MSDP (source discovery)
Within a domain, deploy more than one RP for the same group range
Give each RP the same IP address assignment
Sources and receivers use closest RP
May be used intra-domain (enterprise) to provide redundancy and RP load sharing

Sources from one RP are known to other RPs using MSDP
When an RP goes down, sources and receivers are taken to new RP via unicast routing
– Fast convergence
Anycast RP—Convergence

![Anycast RP—Convergence Diagram]

1. **RP1** (10.0.0.1)
2. **MSDP**
3. **RP2** (10.0.0.1)
4. **Src**
5. **Rec**
Which protocol...

- ...is used to propagate forwarding state?
  PIM-SM - Protocol Independent Multicast-Sparse Mode

- ...maintains routing information for interdomain RPF checking?
  MBGP - Multiprotocol Border Gateway Protocol

- ...exchanges active source information between RPs?
  MSDP - Multicast Source Discovery Protocol

Putting it all together...

- PIM-SM, MBGP, MSDP
- Public multicast peering (MIX)
- Transit topology examples
- Address allocation
  - GLOP draft-ietf-mboned-static-allocation-00.txt
ISP Requirements at the MIX

• Current solution: MBGP + PIM-SM + MSDP
  – Environment
    • ISPs run iMBGP and PIM-SM (internally)
    • ISPs multicast peer at a public interconnect
  – Deployment
    • Border routers run eMBGP
    • The interfaces on interconnect run PIM-SM
    • RPs’ MSDP peering is fully meshed
    • All peers set a common distance for eMBGP

ISP Requirements at the MIX

Peering Solution: MBGP + PIM-SM + MSDP

Redistribute old MBONE DVMRP routes into MBGP
Multicast Transit Design Objectives

- **PIM Border Constraints**
  - Confine registers within domain
  - Confine local groups
  - Confine RP announcements
  - Control SA advertisements via MSDP

- **Border RPF check**
  - RPF check against unicast routes to multicast sources

- **MSDP RPF check**
  - RPF check toward RP in received SAs

Recommended MSDP SA Filter

```plaintext
! domain-local applications
access-list 111 deny ip any host 224.0.2.2 !
access-list 111 deny ip any host 224.0.1.3 ! Rwhod
access-list 111 deny ip any host 224.0.1.24 ! Microsoft-ds
access-list 111 deny ip any host 224.0.1.22 ! SVRLOC
access-list 111 deny ip any host 224.0.1.2 ! SGI-Dogfight
access-list 111 deny ip any host 224.0.1.35 ! SVRLOC-DA
access-list 111 deny ip any host 224.0.1.60 ! hp-device-disc
! auto-rp groups
access-list 111 deny ip any host 224.0.1.39
access-list 111 deny ip any host 224.0.1.40
! scoped groups
access-list 111 deny ip any 239.0.0.0 0.255.255.255
! loopback, private addresses (RFC 1918)
access-list 111 deny ip 10.0.0.0 0.255.255.255 any
access-list 111 deny ip 127.0.0.0 0.255.255.255 any
access-list 111 deny ip 172.16.0.0 0.15.255.255 any
access-list 111 deny ip 192.168.0.0 0.0.255.255 any
access-list 111 permit ip any any
```

Multicast Boundary Filter

Minimum Recommended

!deny auto-rp groups
access-list 1 deny 224.0.1.40
access-list 1 deny 224.0.1.39
!deny scoped groups
access-list 1 deny 239.0.0.0 0.255.255.255
!permit the rest of 224/4
access-list 1 permit 224.0.0.0 15.255.255.255

Single-Homed, ISP RP, Non-MBGP

PIM Border Constraints

Tail-site Customer

Transit AS109

ip pim send-rp-announce Loopback0 scope 255
ip pim send-rp-discovery Loopback0 scope 255
int pos0/0
ip pim sparse-dense-mode
Checking PIM Border (RP mapping)

Tail-site Customer

192.168.100.0/24

Receiver

Transit AS109

pos0/0 1.1.1.1

pos0/0 1.1.1.2

RP

RP pos0/0 1.1.1.1 RP pos0/0 1.1.1.2

192.168.100.10

Tail-site Customer

Transit-tail

show ip pim rp mapping

PIM Group-to-RP Mappings

This system is an RP (Auto-RP)
This system is an RP-mapping agent

Group(s) 224.0.0.0/4

RP 3.3.3.7 (loopback.transit.net), v2vl

Info source: 1.1.1.2 (tail.transit.net), via Auto-RP

Uptime: 21:57:41, expires: 00:02:08

Transit-tail

show ip pim rp mapping

PIM Group-to-RP Mappings

This system is an RP (Auto-RP)
This system is an RP-mapping agent

Group(s) 224.0.0.0/4

RP 3.3.3.7 (loopback.transit.net), v2vl

Info source: 3.3.3.7 (loopback.transit.net), via Auto-RP

Uptime: 22:08:47, expires: 00:02:14
**Single-Homed, ISP RP, Non-MBGPSingle-Homed, ISP RP, Non-MBGP**

### Border RPF Check

**Tail-site Customer**

- `ip route 0.0.0.0 0.0.0.0 1.1.1.2`

**Transit AS109**

- `ip route 192.168.100.0 255.255.255.0 1.1.1.1`
- `router bgp 109`
  
  ... network 192.168.100.0 nlri unicast multicast

### MSDP RPF Check

**Tail-site Customer**

- `no RP / no MSDP`

**Transit AS109**

- `no downstream RP`
- `no downstream MSDP peering`
PIM Border Constraints

Tail-site Customer

192.168.100.0/24

Receiver

Transit AS109

pos0/0 1.1.1.1

pos0/0 1.1.1.2

int pos0/0

ip pim sparse-mode

ip pim bsr-border

ip multicast boundary 1

ip msdp sa-filter out 1.1.1.2 111

ip msdp sa-filter in 1.1.1.2 111

Note: Access-list 111 = Recommended SA Filter
Border RPF Check

Tail-site Customer

Receiver

Transit AS109

RP

ip route 0.0.0.0 0.0.0.0 1.1.1.2

ip route 192.168.100.0 255.255.255.0 1.1.1.1

router bgp 109

... network 192.168.100.0 nlri unicast multicast

MSDP RPF Check

Tail-site Customer

Receiver

Transit AS109

RP

ip msdp peer 1.1.1.1 connect-source loopback 0

ip msdp peer 1.1.1.2 connect-source loopback 0
Single-Homed, Customer RP, MBGP

PIM Border Constraints

Tail-site Customer

192.168.100.0/24

Receiver

pos0/0 1.1.1.1

int pos0/0
ip pim sparse-mode
ip pim bsr-border
ip multicast boundary 1
ip msdp sa-filter out 1.1.1.2 111
ip msdp sa-filter in 1.1.1.2 111

Transit AS109

RP

pos0/0 1.1.1.2

Note: Access-list 111 = Recommended SA Filter
### Border RPF Check

**Tail-site Customer**

```
router bgp 100
network 192.168.100.0 nlri unicast multicast
neighbor 1.1.1.2 remote-as 109 nlri unicast multicast
neighbor 1.1.1.2 update-source pos0/0
```

**Transit AS109**

```
router bgp 109
neighbor 1.1.1.1 remote-as 100 nlri unicast multicast
neighbor 1.1.1.1 update-source pos0/0
```

### MSDP RPF Check

**Tail-site Customer**

```
ip msdp peer 1.1.1.1 connect-source loopback 0
```

**Transit AS109**

```
ip msdp peer 1.1.1.2 connect-source loopback 0
```
PIM Border Constraints

Customer AS100

pos0/0 1.1.1.1
pos1/0 1.1.2.1

Transit AS109

pos0/0 1.1.1.2

Transit AS110

pos0/0 1.1.2.2

int pos0/0
ip pim sparse-mode
ip pim bsr-border
ip multicast boundary 1

int pos1/0
ip msdp sa-filter out 1.1.1.2 111
ip msdp sa-filter in 1.1.1.2 111

Dual-Homed, Customer RP, MBGP
Incongruent Multicast—Unicast

Receiver

192.168.100.0/24

Dual-Homed, Customer RP, MBGP
Incongruent Multicast—Unicast

RP

RP

192.168.100.10

ip multicast boundary 1

ip msdp sa-filter out 1.1.1.1 111
ip msdp sa-filter in 1.1.1.1 111
Hey, this site knows no multicast so there is no PIM to constrain.

```
router bgp 100
 network 192.168.100.0 nlri unicast multicast
 neighbor 1.1.1.1 remote-as 100 nlri unicast multicast
 neighbor 1.1.1.1 update-source pos 0/0
 neighbor 1.1.2.2 remote-as 110
 neighbor 1.1.2.2 update-source pos 1/0
```
**Dual-Homed, Customer RP, MBGP Incongruent Multicast—Unicast**

**Border RPF Check**

Customer AS100

```
router bgp 109
neighbor 1.1.1.1 remote-as 100 nlri unicast multicast
neighbor 1.1.1.1 update-source pos 0/0
```

Transit AS109

```
router bgp 110
neighbor 1.1.2.1 remote-as 100
neighbor 1.1.2.1 update-source pos 0/0
```
Dual-Homed, Customer RP, MBGP
Incongruent Multicast—Unicast

MSDP RPF Check

Customer AS100
pos0/0 1.1.1.1
pos1/0 1.1.2.1

RP
ip msdp peer 1.1.1.2 connect-source loopback 0
ip msdp peer 1.1.2.1 connect-source loopback 0

Transit AS109
pos0/0 1.1.1.2
pos0/0 1.1.2.2

192.168.100.0/24

Customer AS100
pos0/0 1.1.1.1
pos1/0 1.1.2.1
RP

Receiver

ip msdp peer 1.1.1.1 connect-source loopback 0

Transit AS110
pos1/0 1.1.2.1
pos0/0 1.1.2.2

Dual-Homed, Customer RP, MBGP
Congruent Multicast—Unicast

PIM Border Constraints

Customer AS100
pos0/0 1.1.1.1
pos1/0 1.1.2.1
RP

int pos0/0
ip pim sparse-mode
ip pim bsr-border
ip multicast boundary 1

int pos1/0
ip pim sparse-mode
ip pim bsr-border
ip multicast boundary 1
ip msdp sa-filter out 1.1.1.2 111
ip msdp sa-filter in 1.1.1.2 111
ip msdp sa-filter out 1.1.2.2 111
ip msdp sa-filter in 1.1.2.2 111
Dual-Homed, Customer RP, MBGP Congruent Multicast—Unicast

PIM Border Constraints

Customer AS100

RP

pos0/0 1.1.1.1

pos1/0 1.1.2.1

192.168.100.0/24

Receiver

Transit AS109

Unicast & Multicast Transit

pos0/0 1.1.1.2

Transit AS110

Unicast & Multicast Transit

pos0/0 1.1.2.2

Dual-Homed, Customer RP, MBGP Congruent Multicast—Unicast

PIM Border Constraints

Customer AS100

RP

pos0/0 1.1.1.1

pos1/0 1.1.2.1

192.168.100.0/24

Receiver

Transit AS109

Unicast & Multicast Transit

pos0/0 1.1.1.2

Transit AS110

Unicast & Multicast Transit

pos0/0 1.1.2.2

int pos0/0
ip pim sparse-mode
ip pim bsr-border
ip multicast boundary 1
ip msdp sa-filter out 1.1.1.1 111
ip msdp sa-filter in 1.1.1.1 111

int pos0/0
ip pim sparse-mode
ip pim bsr-border
ip multicast boundary 1
ip msdp sa-filter out 1.1.2.1 111
ip msdp sa-filter in 1.1.2.1 111
Dual-Homed, Customer RP, MBGP Congruent Multicast—Unicast

Border RPF Check

Customer AS100

pos0/0 1.1.1.1

pos1/0 1.1.2.1

RP

192.168.100.0/24

Transit AS109

pos0/0 1.1.1.2

pos1/0 1.1.2.1

Transit AS110

pos0/0 1.1.2.2

router bgp 100

network 192.168.100.0 nleri unicast multicast
neighbor 1.1.1.2 remote-as 100 nleri unicast multicast
neighbor 1.1.1.2 update-source pos/0
neighbor 1.1.2.2 remote-as 110 nleri unicast multicast
neighbor 1.1.2.2 update-source pos/1/0

router bgp 109

neighbor 1.1.1.1 remote-as 100 nleri unicast multicast
neighbor 1.1.1.1 update-source pos 0/0
**Dual-Homed, Customer RP, MBGP Congruent Multicast—Unicast**

**Border RPF Check**

Customer AS100

- RP pos0/0 1.1.1.1
- pos1/0 1.1.2.1

Transit AS110

- pos0/0 1.1.2.2

Transit AS109

- pos0/0 1.1.1.2

MSDP RPF Check

Customer AS100

- RP pos0/0 1.1.1.1
- pos1/0 1.1.2.1

Transit AS110

- pos0/0 1.1.2.2

Transit AS109

- pos0/0 1.1.1.2

**Configuration**

```
router bgp 110
neighbor 1.1.2.1 remote-as 100 nlri unicast multicast
neighbor 1.1.2.1 update-source pos0/0
```

```
ip msdp peer 1.1.1.1 connect-source loopback 0
ip msdp peer 1.1.2.1 connect-source loopback 0
```

```
ip msdp peer 1.1.2.1 connect-source loopback 0
```
GLOP—Static Allocation of 233/8

- Temporary allocation of 233/8
  - rfc2770
- Statically assigned by mapping AS number into middle octets.
  - http://gigapop.uoregon.edu/glop/index.html
- Provides each AS with /24 addresses to use while waiting another solution

Agenda

- PIM-SM review
- Inter-Domain Multicast
  - MBGP
  - MSDP
  - Topology Examples
- SSM (Source Specific Multicast)
Source Specific Multicast

- Simplify solution for well-known sources, particularly in cases where there is a single source sending to a given group.
  - Allow immediate use of shortest forwarding path to a specific source, without need to create shared tree.
  - Eliminate dependence on MSDP for finding sources.
  - Simplify address allocation for global, single source groups when combined with elimination of shared trees (232/8).

PIM Source Specific Mode

Receiver learns of source, group/port
Receiver sends IGMPv3 (S,G) Join
First-hop sends PIM (S,G) Join directly toward Source

(S, G) Join

Out-of-band source directory, example: web server
**PIM Source Specific Mode**

Result: Shortest path tree rooted at the source, with no shared tree.

Out-of-band source directory, example: web server

**Objective of SSM Multicast**

Optimize multicast forwarding when sources are few and specifically known to receiver in advance of joining.
Where Is SSM?

- Framework
draft-holbrook-ssm-00.txt
- BCP proposal
draft-shepherd-232-00.txt
- Supported in IOS 12
- Waiting for IGMPv3 on hosts

Documentation and Contact Info

- EFT/Beta Site Web Page:
  ftp://ftpeng.cisco.com/ipmulticast.html
- TAC Support Mailing List:
tac@cisco.com
- Customer Support Mailing List:
cs-ipmulticast@cisco.com
If All Else Fails—RTFB

Advanced IP Multicast Routing
Session 2217
Please Complete Your Evaluation Form

Session 2217