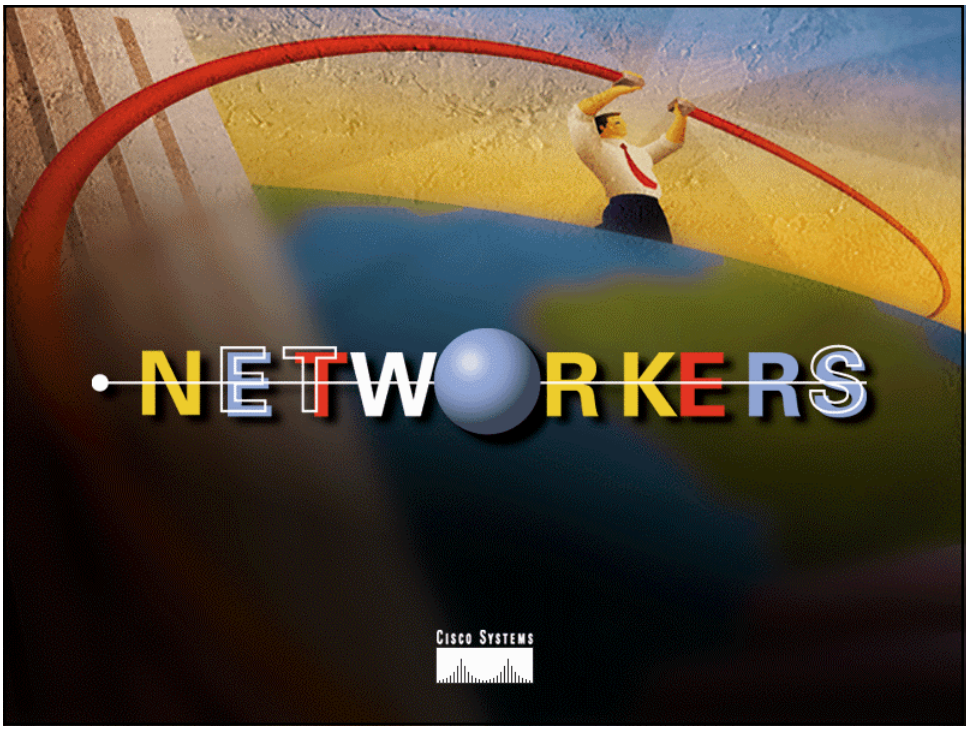

NETWORKERS
Cisco Systems


Advances in Multicast Technology
Session 320
Cisco Systems

Other Related Presentations

- Multicast sessions

Session #	Title
303	Introduction to IP Multicast
306	PIM Protocol Concepts
314	Deploying IP Multicast
320	Advances in IP Multicast

- MBGP related sessions

Session #	Title
309	Deploying BGP
317	Advanced BGP and Troubleshooting

Agenda

- Inter-Domain Multicast
  - MBGP
  - MSDP
- PGM (Pragmatic General Multicast)
- MRM (Multicast Route Monitor)
- UDLR (Unidirectional Link Routing)

Inter-Domain Multicast

- Past history
- In the future
- ISP requirements to deploy now
  - MBGP
  - MSDP

Past History

**A long time ago,
in a galaxy
far, far away…**

Past History

MBONE…

Past History

MBONE… vat, nv, wb, sd,...

Past History

- 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

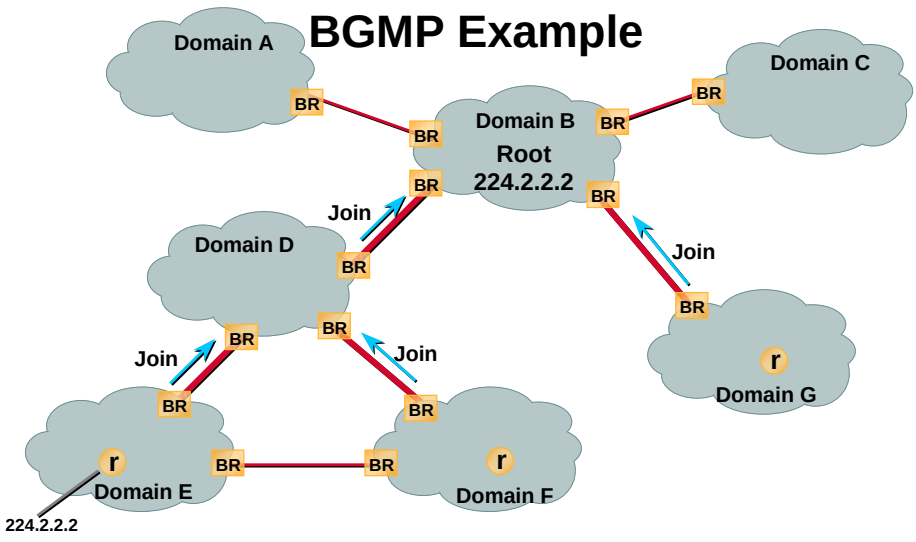
Problem

- 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.

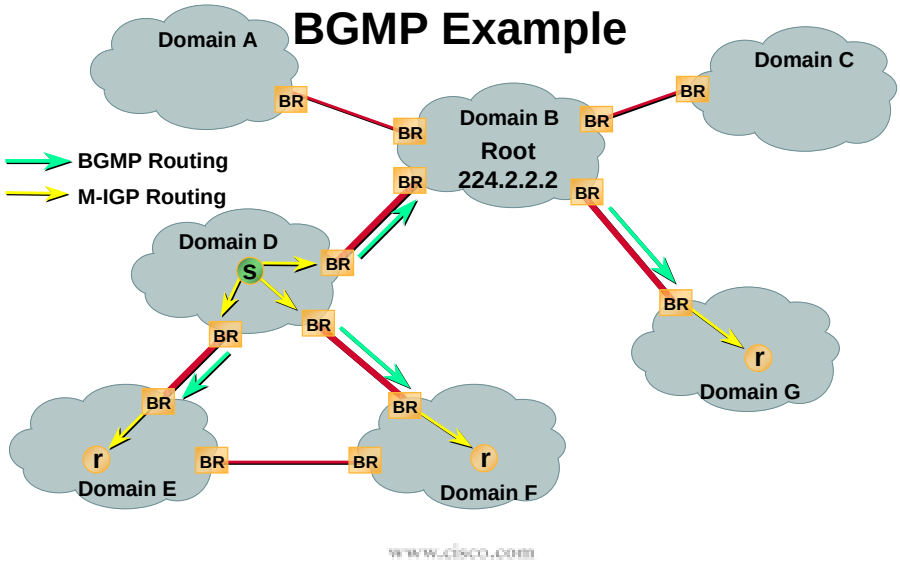
In the Future

- 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



In the Future



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 want to deploy multicast now
  - What are their minimum requirements?

ISP Requirements to Deploy Now

- Want an explicit join protocol for efficiency
  - ☑ PIM-SM
- Use existing (unicast) operation model
  - Hmmm
- Will not share RP with competitors
  - Results in Third-party Resource Dependency
  - Hmmm
- Want flexibility regarding RP placement
  - Hmmm

ISP Requirements to Deploy Now

- Use existing (unicast) operation 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—Multiprotocol BGP

- MBGP overview
- MBGP capability negotiation
- MBGP NLRI exchange
- MBGP-DVMRP redistribution
- BGP-to-MBGP redistribution

MBGP Overview

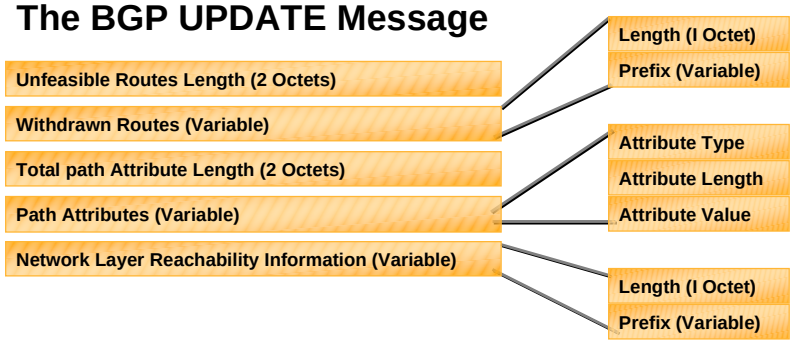
- MBGP: Multiprotocol BGP
  (aka multicast BGP in multicast networks)
  - Defined in RFC 2283 (extensions to BGP)
  - Can carry different types of routes
    - Unicast
    - Multicast
  - Both routes carried in same BGP session
  - Does not propagate multicast state info
  - Same path selection and validation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

MBGP Overview

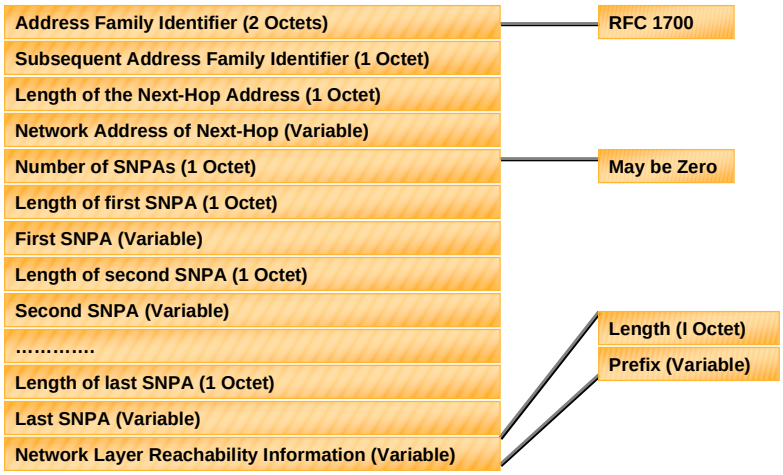
The BGP UPDATE Message



- 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

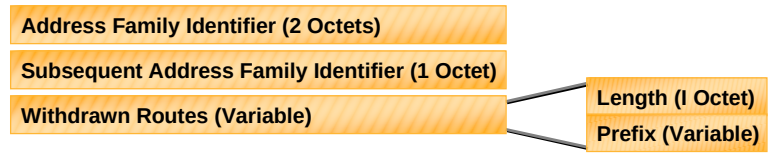
MBGP Overview

MP_REACH_NLRI Attribute



MBGP Overview

MP_UNREACH_NLRI Attribute



MBGP Overview

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

MBGP Overview

- 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 DVMRP
  - Redistribution of BGP stubs into MBGP

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

MBGP—Capability Negotiation

- 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

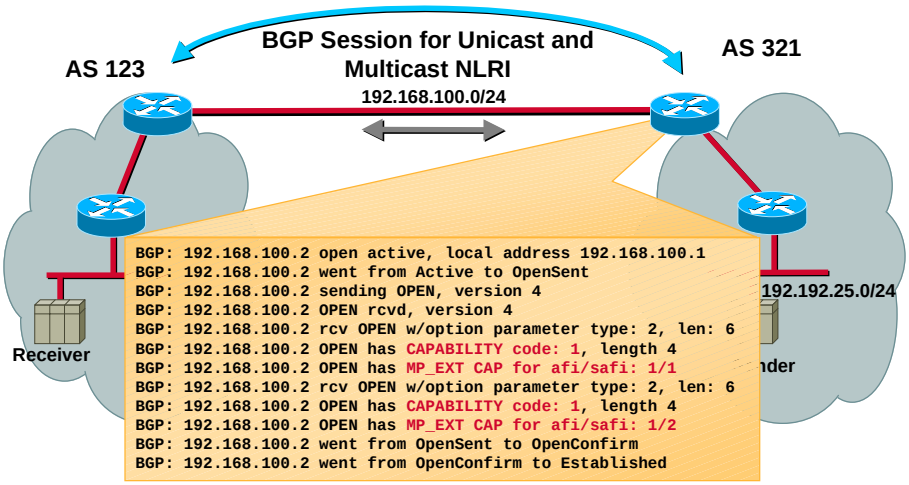
MBGP—Capability Negotiation

- 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

Congruent Topologies



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

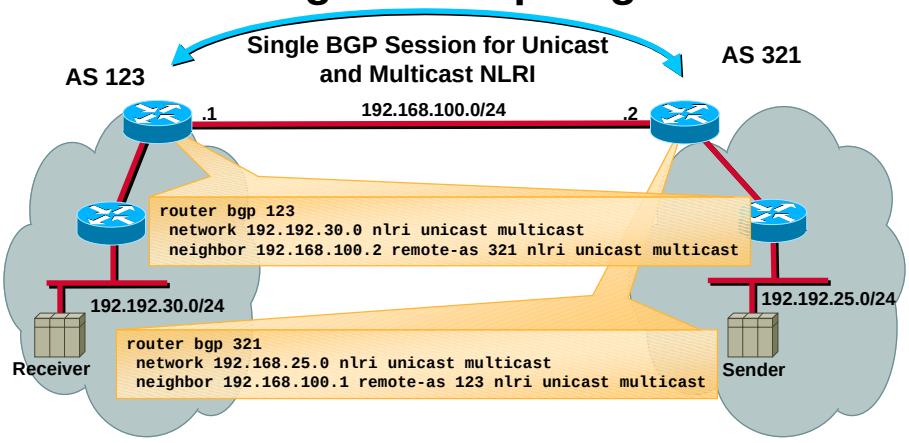
MBGP NLRI Exchange

- 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]
```

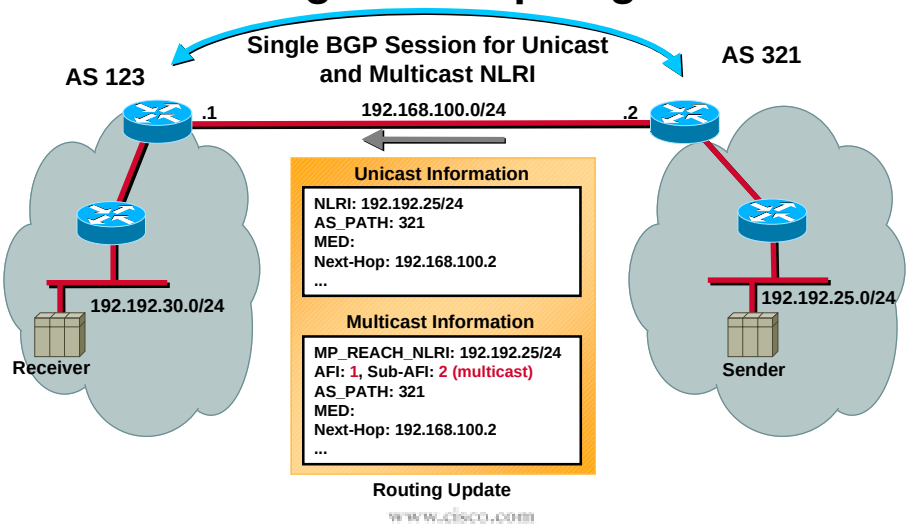
MBGP—NLRI Information

Congruent Topologies



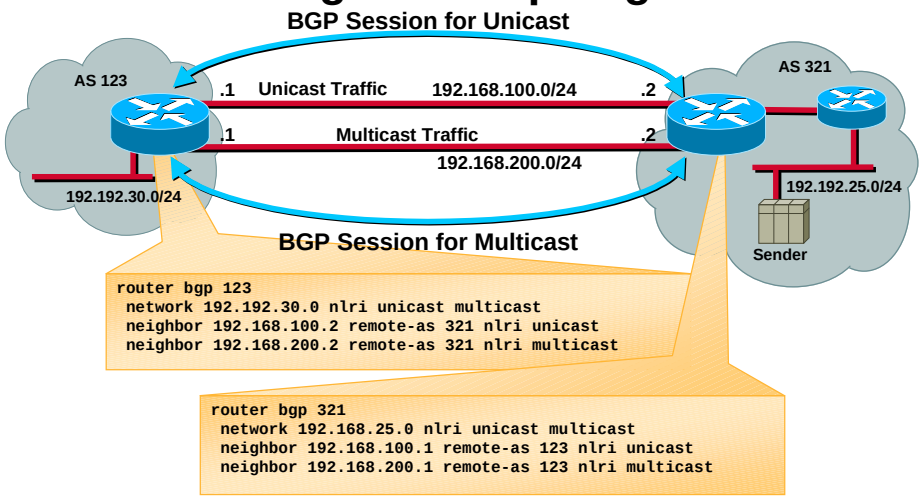
MBGP—NLRI Information

Congruent Topologies



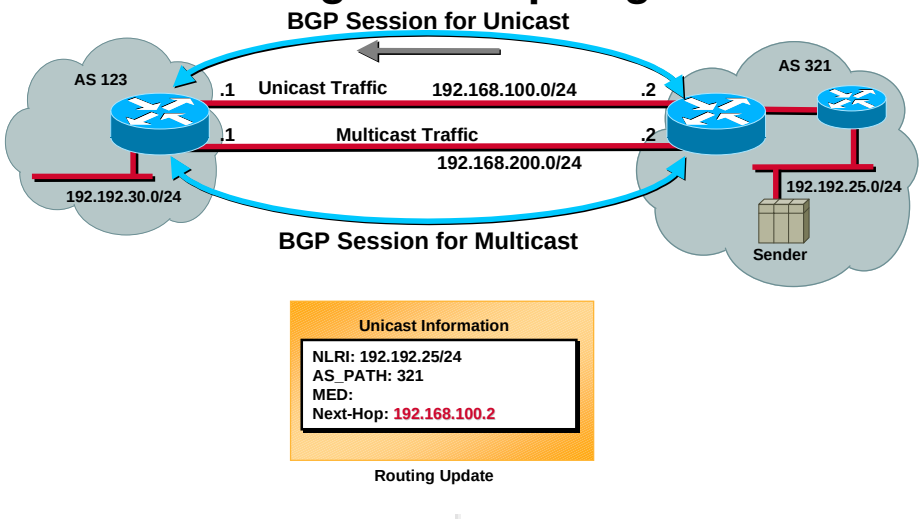
MBGP—NLRI Information

Incongruent Topologies



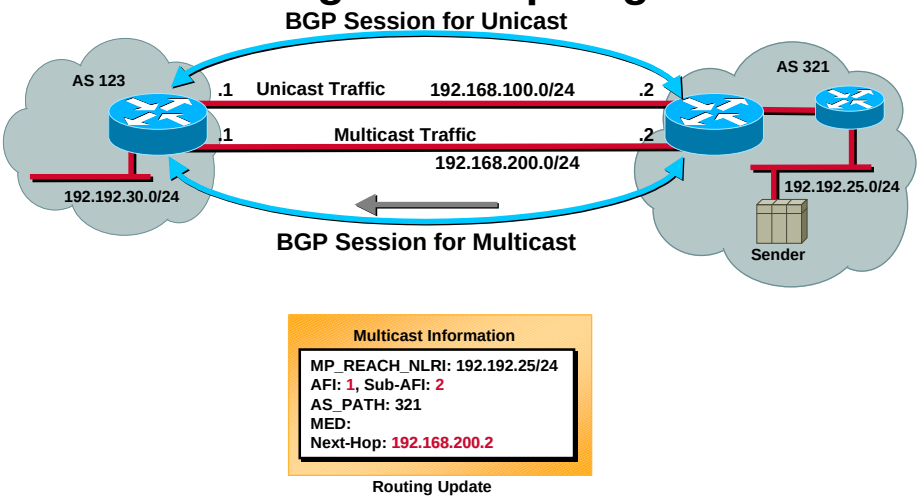
MBGP—NLRI Information

Incongruent Topologies



MBGP—NLRI Information

Incongruent Topologies



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

DVMRP <-> MBGP Redistribution

- You can also put routes in the MRIB that are currently in the DVMRP routing table

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

- You can do your typical set operations
- Used when connecting DVMRP access points into the MBGP backbone
- Used at strategic interconnect points with the old DVMRP MBONE

DVMRP <-> MBGP Redistribution

- MBGP routes can be sent into DVMRP
  - However, we recommend tail sites using DVMRP access to accept DVMRP default route

```
interface tunnel0
ip dvmrp metric 1 route-map <map> mbgp
```

- Can use typical match operations

BGP to MBGP Redistribution

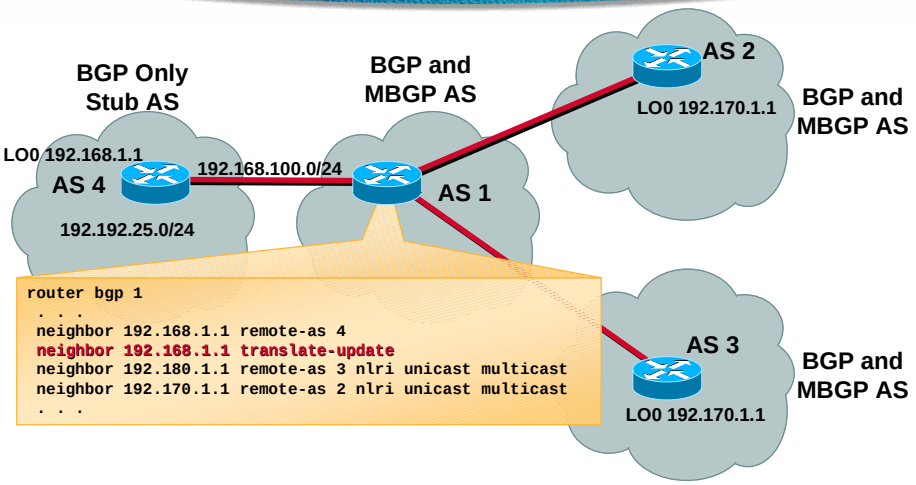
- BGP stubs that don't have MBGP support need to get their routes into the MBGP backbone
- They get external routes via MBGP default or static default
- Use command

```
neighbor <foo> translate-update [nlri unicast multicast]
```

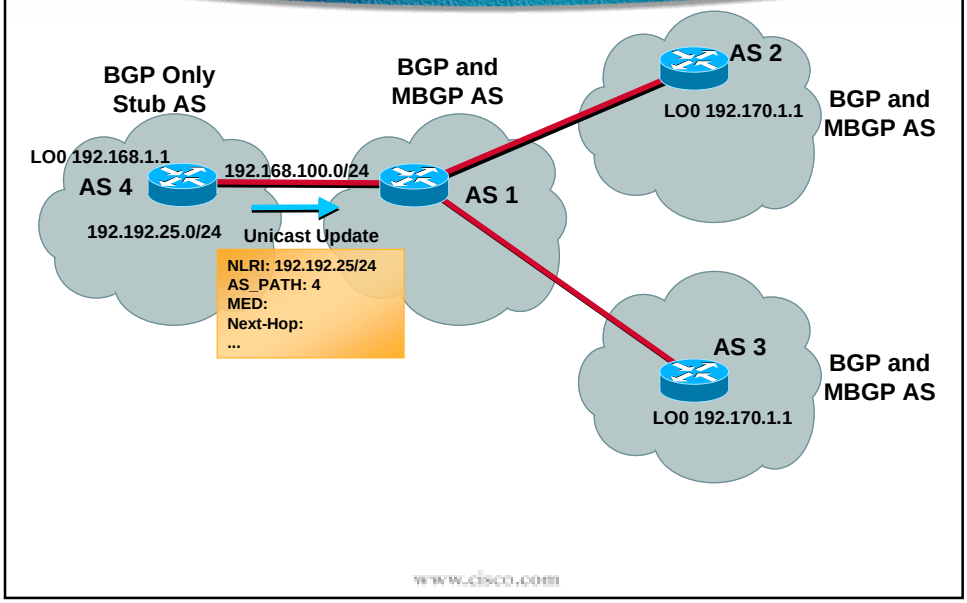
BGP to MBGP Redistribution

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

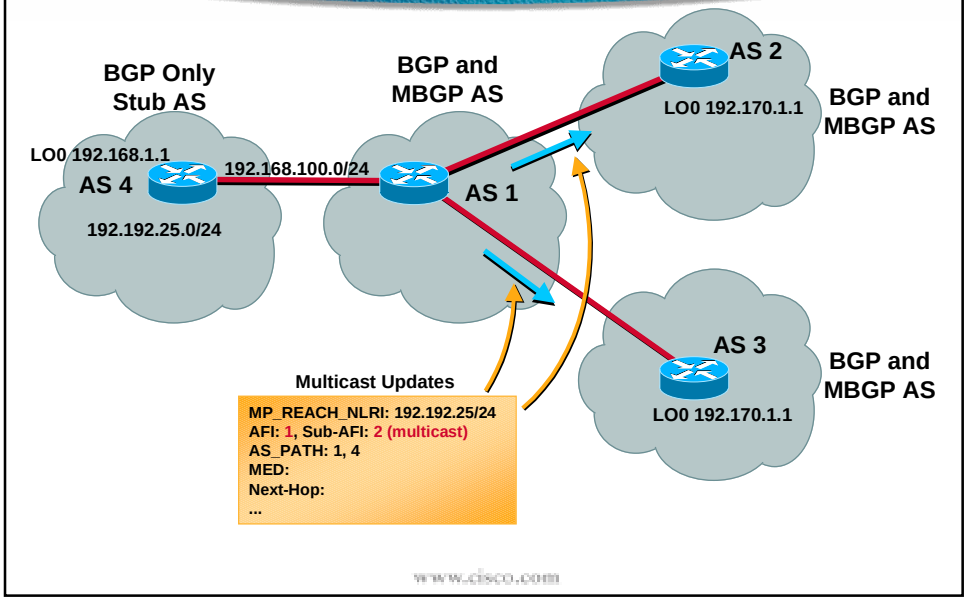
BGP to MBGP Redistribution



BGP to MBGP Redistribution



BGP to MBGP Redistribution



MBGP—Summary

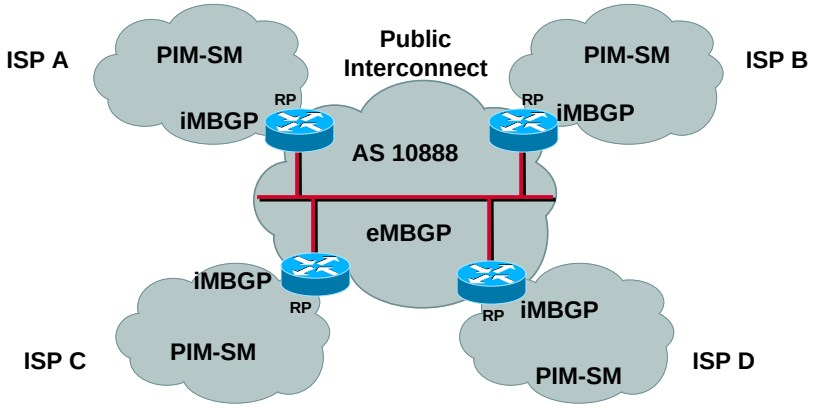
- Solves part of inter-domain problem
  - Can exchange multicast routing information
  - Uses standard BGP configuration knobs
  - Permits separate unicast and multicast topologies if desired
- Still must use PIM to:
  - Build distribution trees
  - Actually forward multicast traffic
  - PIM-SM recommended
    - But there's still a problem using PIM-SM here… (more on that later)

ISP Requirements to Deploy Now

- Interim solution: MBGP + PIM-SM
  - Environment
    - ISPs run MBGP and PIM-SM (internally)
    - ISPs multicast peer at a public interconnect
  - Deployment
    - Each ISP puts their own administered RP attached to the interconnect
    - That RP as well as all border routers run MBGP
    - The interconnect runs dense-mode PIM

ISP Requirements to Deploy Now

Interim Solution: MBGP + PIM-SM



ISP Requirements to Deploy Now

- Interim solution: MBGP + PIM-SM
  - Too restrictive regarding RP placement
    - Need multiple interconnect points between ISP's
  - Using multiple interconnect points
    - Fine if all ISP RP's at same interconnect
    - Can degenerate into large PIM-DM cloud
  - Back to the "requirements list"

ISP Requirements to Deploy Now

- Want an explicit join protocol for efficiency
  - ☑ PIM-SM
- Use existing (unicast) operation model
  - ☑ MBGP
- Will not share RP with competitors
  - Results in third-party resource dependency
  - Hmmm
- Want flexibility regarding RP placement
  - 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
      - Hmmm. 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—logical RP

MSDP Concept

- Simple but elegant
  - Abandon inter-domain shared trees; just use inter-domain source trees
  - Reduces to 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

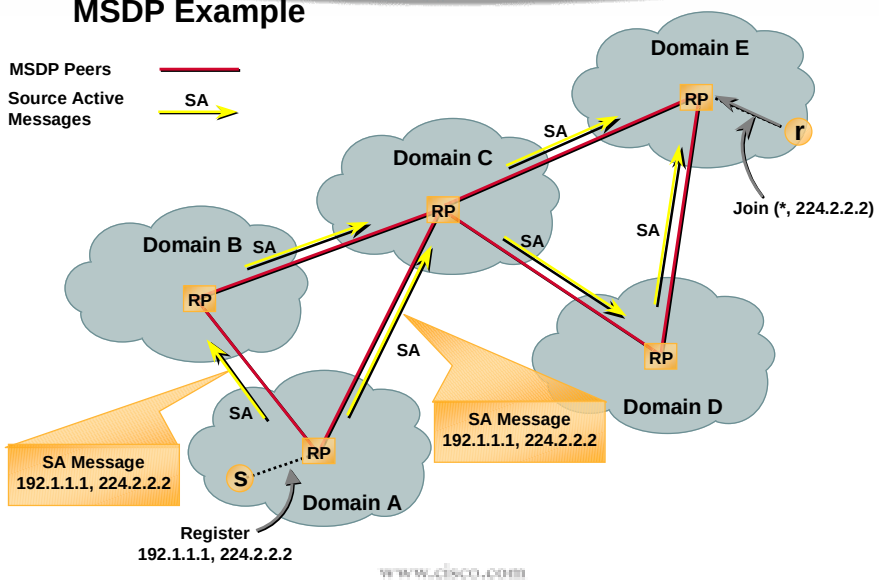
      Only necessary if there are receivers for the group

MSDP Design Points

- MSDP peers talk via TCP connections
  - UDP encapsulation option
- Source Active (SA) messages
  - Peer-RPF forwarded to prevent loops
    - RPF check on AS-PATH back to the peer RP
    - If successful, flood SA message to other peers
    - Stub sites accept all SA messages
      - Since they have only one exit (e.g., default peer)
  - MSDP speaker may cache SA messages
    - Reduces join latency

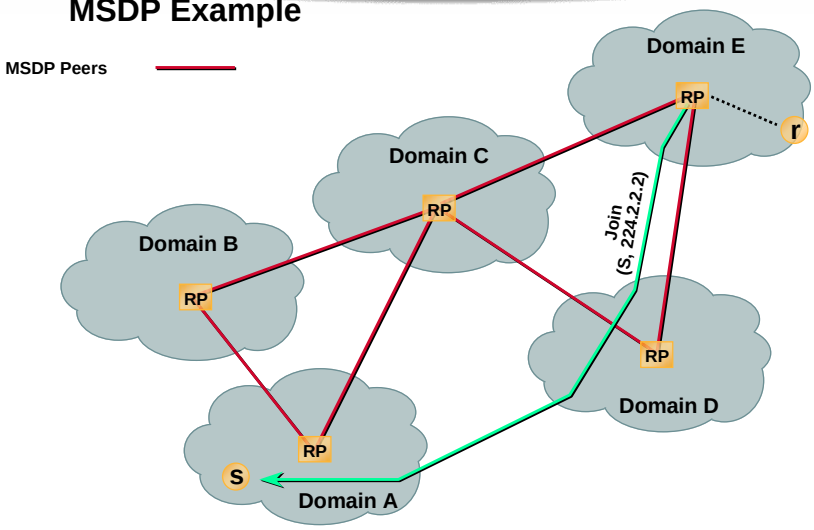
MSDP Overview

MSDP Example



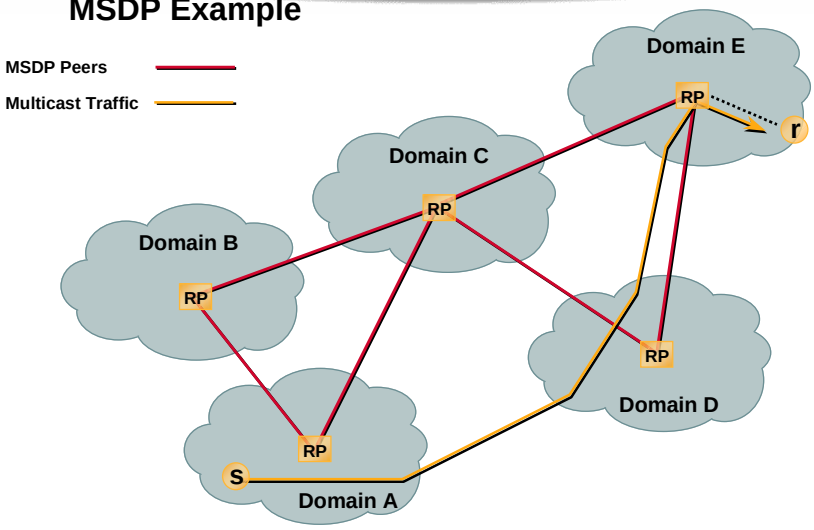
MSDP Overview

MSDP Example



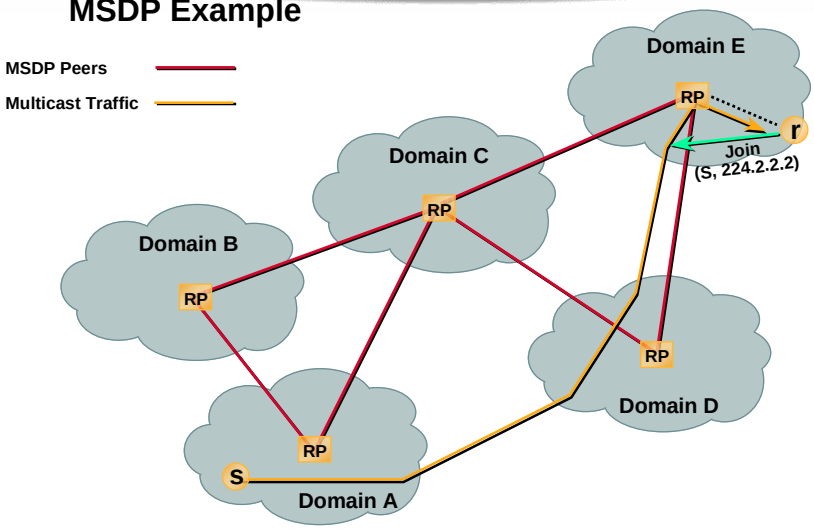
MSDP Overview

MSDP Example



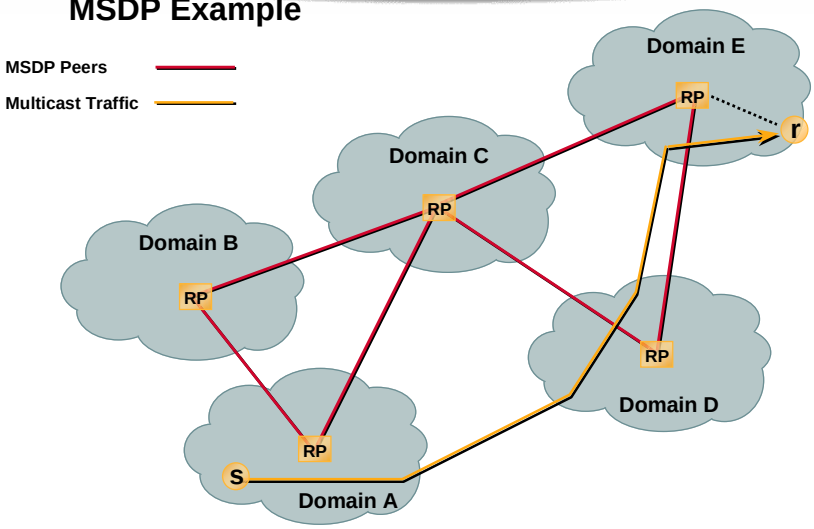
MSDP Overview

MSDP Example



MSDP Overview

MSDP Example



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 (Cont.)

- **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
- Use existing (unicast) operation model
  - ☑ MBGP
- Will not share RP with competitors
  - ☑ MSDP
- Want flexibility regarding RP placement
  - ☑ MSDP

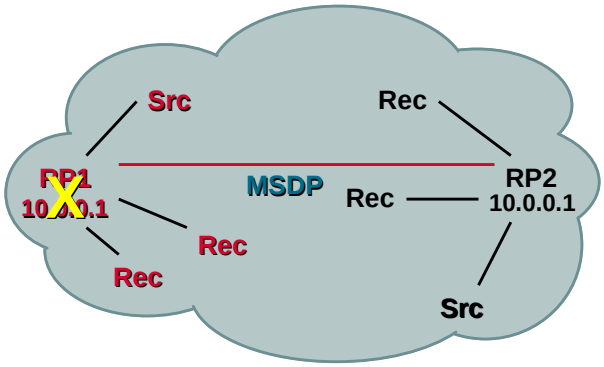
MSDP Application—Logical RP

- draft-ietf-mboned-logical-rp-00.txt
- 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

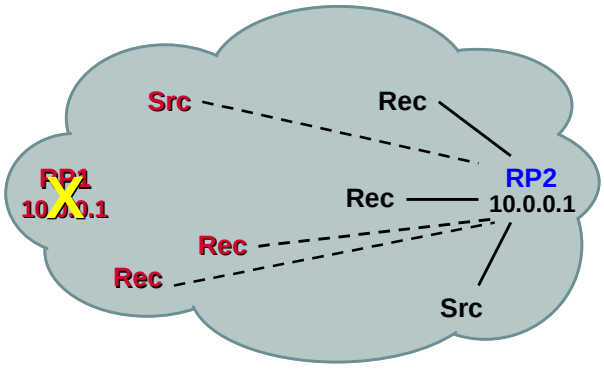
MSDP Application—Logical RP

- 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

Logical RP—Convergence



Logical RP—Convergence



Agenda

- Inter-Domain Multicast
  - MBGP
  - MSDP
- PGM (Pragmatic General Multicast)
- MRM (Multicast Route Monitor)
- UDLR (Unidirectional Link Routing)

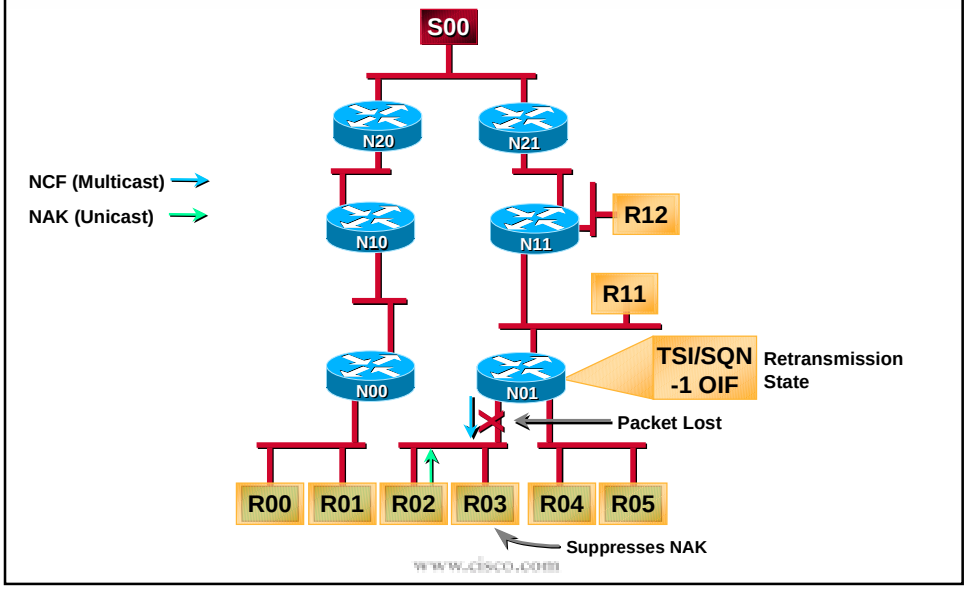
Pragmatic General Multicast (PGM)

- IETF draft
  - draft-speakman-pgm-spec-02.txt
- Routers assist the retransmit process
  - NAK suppression mechanism
  - Retransmission constraint mechanism
  - Maintain NAK/retransmission state only
- Important point:
  - Routers don't do the retransmitting

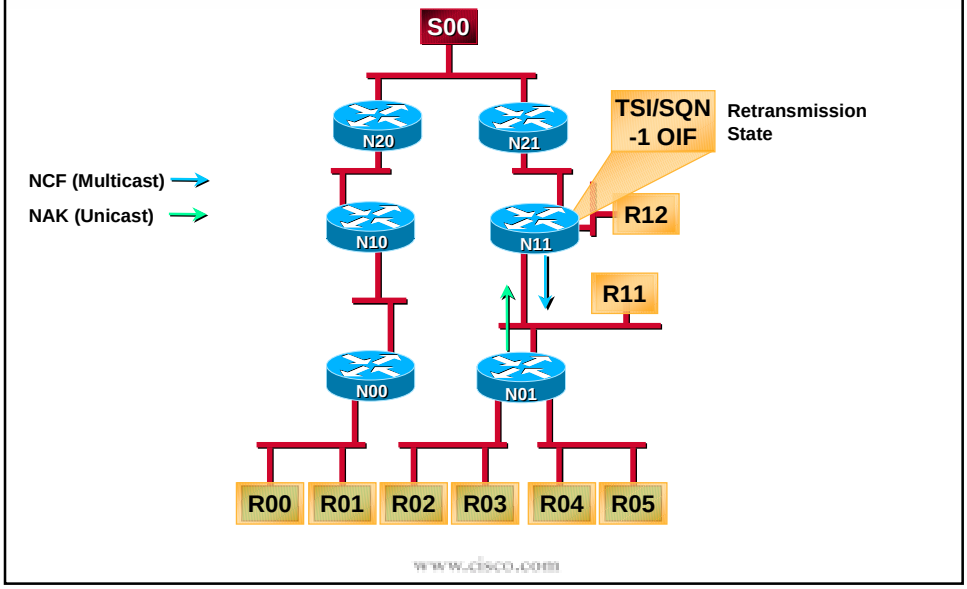
PGM—Pragmatic General Multicast

- Source multicasts packets (ODATA)
  - Identified by Transport Session Id (TSI)
  - Sequenced by Sequence Number (SQ)
- Receivers detect drops via TSI/SQ
  - Waits random delay before sending NAK
  - NAK's are unicast to upstream PGM router
- Routers send NAK Confirmations (NCF)
  - NCF's are multicast back to receivers
  - Other receivers suppress NAK's upon hearing NCF

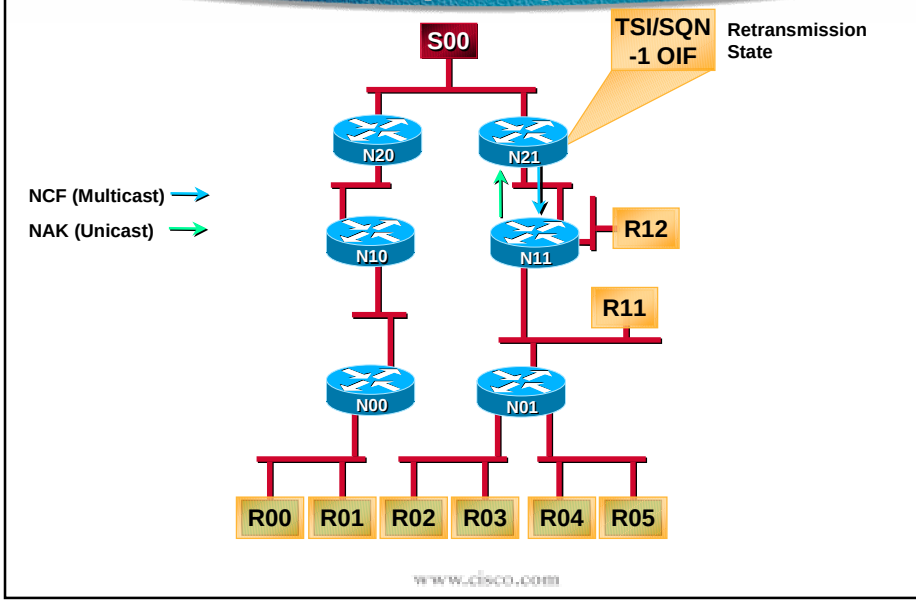
Pragmatic General Multicast (PGM)



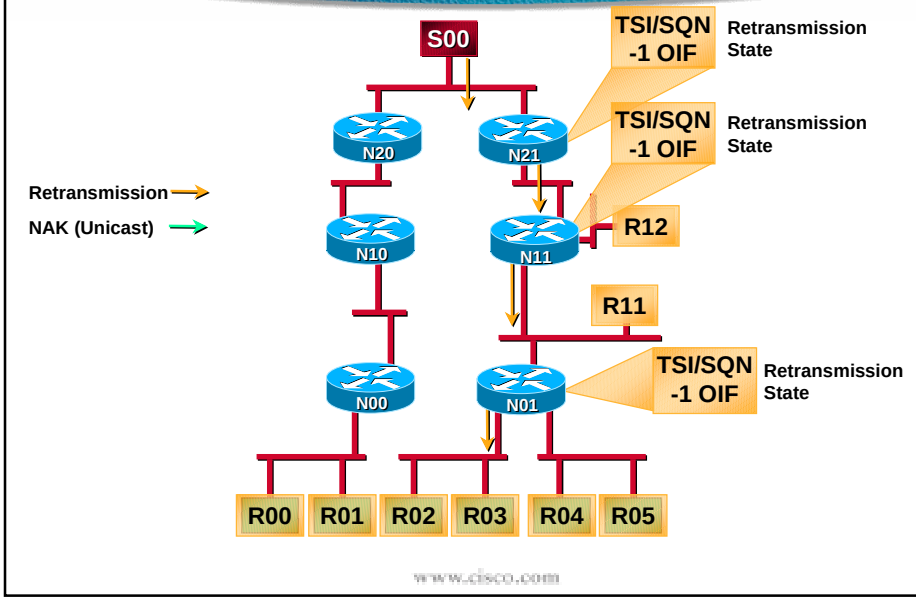
Pragmatic General Multicast (PGM)



Pragmatic General Multicast (PGM)



Pragmatic General Multicast (PGM)



Agenda

- Inter-Domain Multicast
  - MBGP
  - MSDP
- PGM (Pragmatic General Multicast)
- MRM (Multicast Route Monitor)
- UDLR (Unidirectional Link Routing)

Multicast Route Monitor (MRM)

- What/who is MRM designed for
- Detection of faults in multicast routing
- Isolation of faults
- Reliability
- Security

MRM—Motivation and Goals

- Detection and alarm of network problems in close to real-time
- Good coverage of faults
- Good extensibility
- Low overhead (scale)

MRM—Partitioning the Tasks

- Fault detection:
  - Identification
  - Classification
  - May involve a large number of systems
- Fault isolation:
  - Find system/LAN/region with trouble
  - Ideally involving as few systems as possible

MRM—Fault Detection

- Topological disconnectivity (physical/logical)
- Blackholes in forwarding path
- Excessive/persistent packet losses
- Excessive duplicates

MRM—Fault Isolation

- mtrace
- rsh, snmp based tools, etc
- traceroute
- A combination of tools

MRM—Protocol Components

- MRM monitor
  - Activates diagnostic tests, collects, summarizes, presents diagnostic output
- Test Sender (TS)
  - A system that originates traffic for testing purposes
- Test Receiver (TR)
  - Collects data and reports to the MRM monitor

MRM—Protocol Messages

- Beacon messages
  - Originated by the MRM monitor
  - Addressed to a well-known multicast address
  - Signals the liveness of the MRM monitor
  - Medium to carry periodically refreshed requests
- MRM monitor requests
  - Source specification requests
  - Statistics collection requests
- Statistics reports

MRM—Reliability

- Positive acknowledgement for unicast request messages, or single packet reports
- Periodic refresh of requests for multicast addressed requests
- Use TCP for large reports
- Critical TSs and TRs send low-frequency periodic liveness reports to the MRM monitor

MRM—Security

- All MRM messages carry monotonically increasing sequence numbers
- Use MD5 as the standard authentication mechanism

MRM—More Information

- ftp://ftpeng.cisco.com/ipmulticast/mrm
- mrm.guide
- Images

Agenda

- Inter-Domain Multicast
  - MBGP
  - MSDP
- PGM (Pragmatic General Multicast)
- MRM (Multicast Route Monitor)
- UDLR (Unidirectional Link Routing)

UDLR

- Applicable environments
- The problem
- Cisco solutions
  - UDLR-Tunnels
  - IGMP-UDLR

Applicable Environments

- Satellite systems
- ADSL connections
  - Where bandwidths are asymmetric
- Cable systems
  - Where bandwidths and link-type are asymmetric
- ATM partially meshed SVCs

The Fundamental Problem

- Both unicast and multicast routing protocols forward data on interfaces in which they have received routing control information
- The model can only work on bi-directional links

The Problem (In More Detail)

- Unicast routing
  - If I received an update on interface serial0 for prefix P, then I will forward data for destinations that match prefix P out serial0 (distance vector)
- Multicast routing
  - If I receive a Join on interface serial0 for group G, then I will forward data for traffic destined for group G out serial0 (sparse-mode)

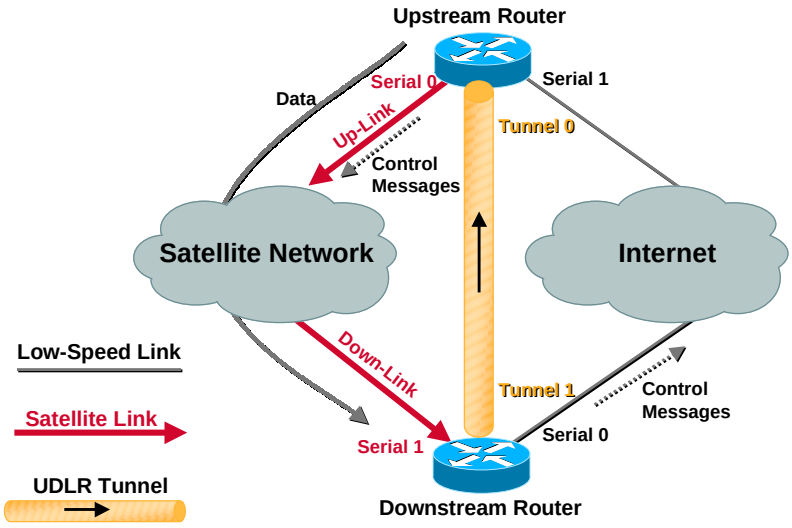
Cisco Solutions

- UDLR-Tunnels for unicast and multicast routing
- IGMP-UDLR for large-scale multicast routing

UDLR-Tunnels

- Extend GRE tunnels to be configured as one-way
- Associate the one-way tunnel with a one-way interface (which goes in the opposite direction)
- ULPs don’t see tunnel as an interface
- Mapping performed at the link-layer so real one-way interface looks bi-directional

UDLR-Tunnels



UDLR-Tunnels

- **How to configure (upstream router)**

```
interface tunnel0
tunnel udlr receive-only serial0
```

- **How to configure (downstream router)**

```
interface tunnel1
tunnel udlr send-only serial1
```

UDLR-Tunnels

- Features
  - All IP unicast routing protocols supported
  - IS-IS (via CLNS) is supported
  - All IP multicast routing protocols supported
  - HDLC keepalives
  - PPP Link Quality Monitoring (LQM)

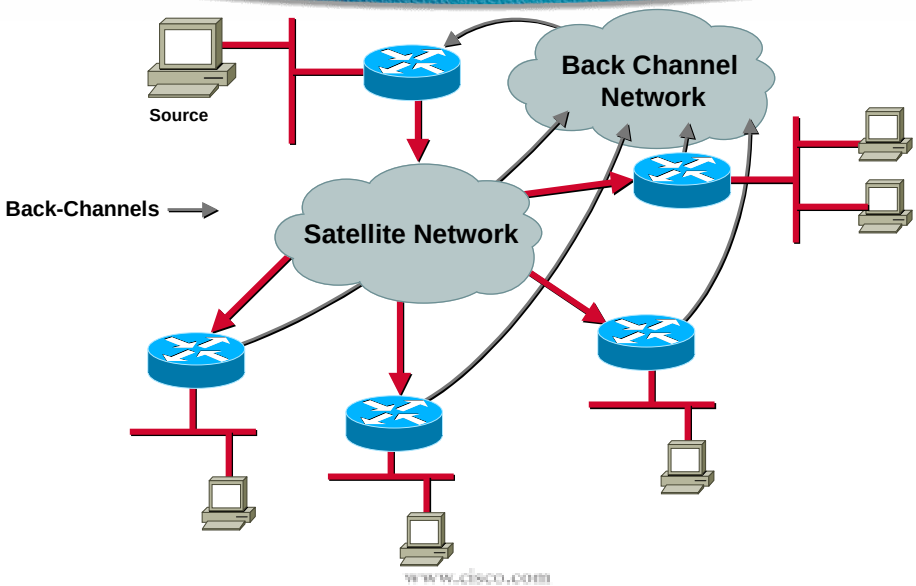
UDLR-Tunnels

- Caution!
- This is not a general purpose scalable solution for UDLR routing
- You have to limit the number of tunnels that fan-into the upstream router
- Useful for small transit clouds

IGMP-UDLR

- Used for large scale multicast routing over widespread unidirectional links
- Design goals
  - Eliminate static multicast routes and static group membership
  - Reduce the number of control messages sent
  - Built-in fault tolerance

IGMP-UDLR Environment



IGMP-UDLR—Basic Idea

- Downstream routers listen for IGMP queries
- They select a querier
- Host sends IGMP report to join group
- Downstream router forwards IGMP report to querier
- Querier (upstream router) populates olist for data forwarding
- Querier echos IGMP report back out one-way link to suppress other downstream reports

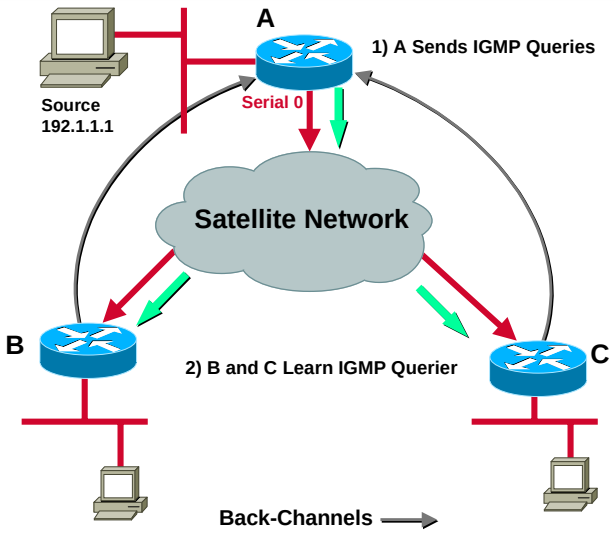
IGMP-UDLR—Basic Idea (Cont.)

- Other downstream routers remember reporter for group and monitor it's reporting status for the group
- When the reporter goes down or leaves the group, a new reporter forwards IGMP reports
- Leaves work the same way

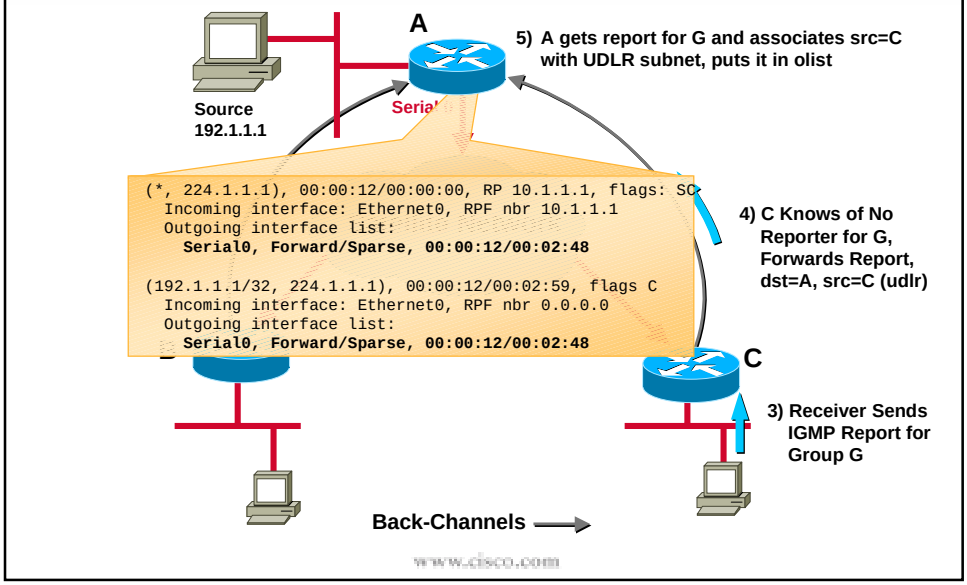
IGMP-UDLR Scalability

- Groups are dynamic so only joined group traffic traverses UDLR link
- Report suppression allows one report per group per UDLR link (irrespective of the number of members and member subnets)

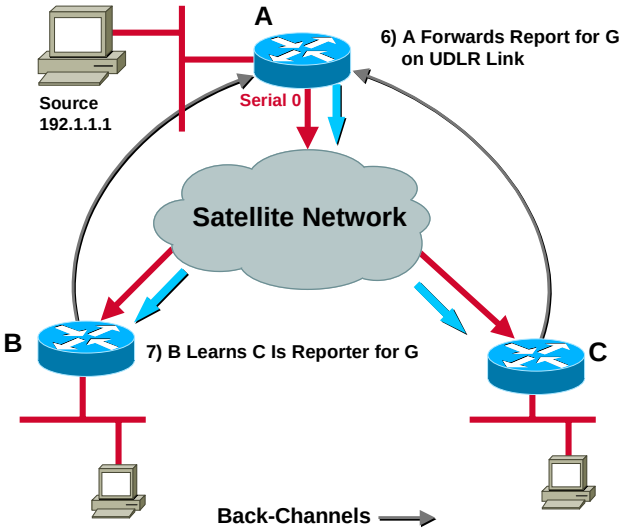
IGMP-UDLR Details for Joining



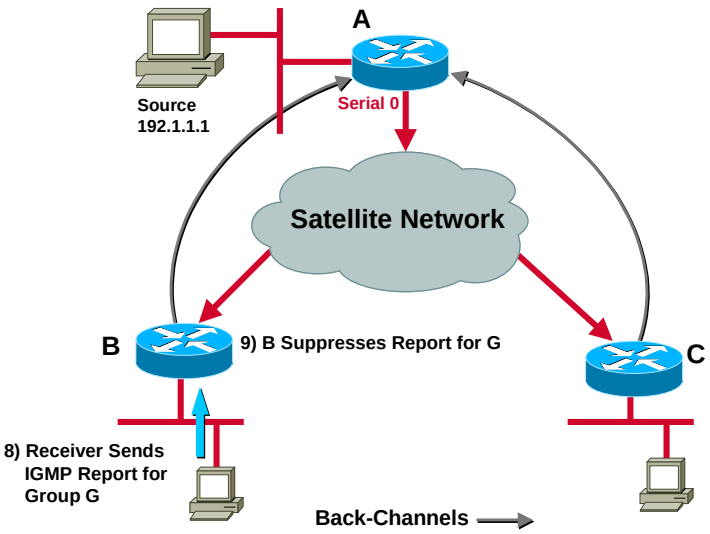
IGMP-UDLR Details for Joining



IGMP-UDLR Details for Joining



IGMP-UDLR Details for Joining



UDLR Documentation

- URLs
  - ftp://ftpeng.cisco.com/dino/udlr
  - udlr-tunnel-commands
  - udlr-commands (IGMP-UDLR)
  - udlr.txt
- Mailing lists
  - udlr-beta@cisco.com

Documentation and Contact Info

- **EFT/Beta Site Web Page:**
  - **ftp://ftpeng.cisco.com/ipmulticast.html**
- **EFT/Beta Mailing List:**
  - **multicast-support@cisco.com**
- **Customer Support Mailing List:**
  - **cs-ipmulticast@cisco.com**

If All Else Fails—RTFB[1]

[1] Read this fine book

Cisco Systems

Empowering the Internet Generation℠