IP Multicast Tutorial
Apricot 2013, Singapore

Greg Shepherd

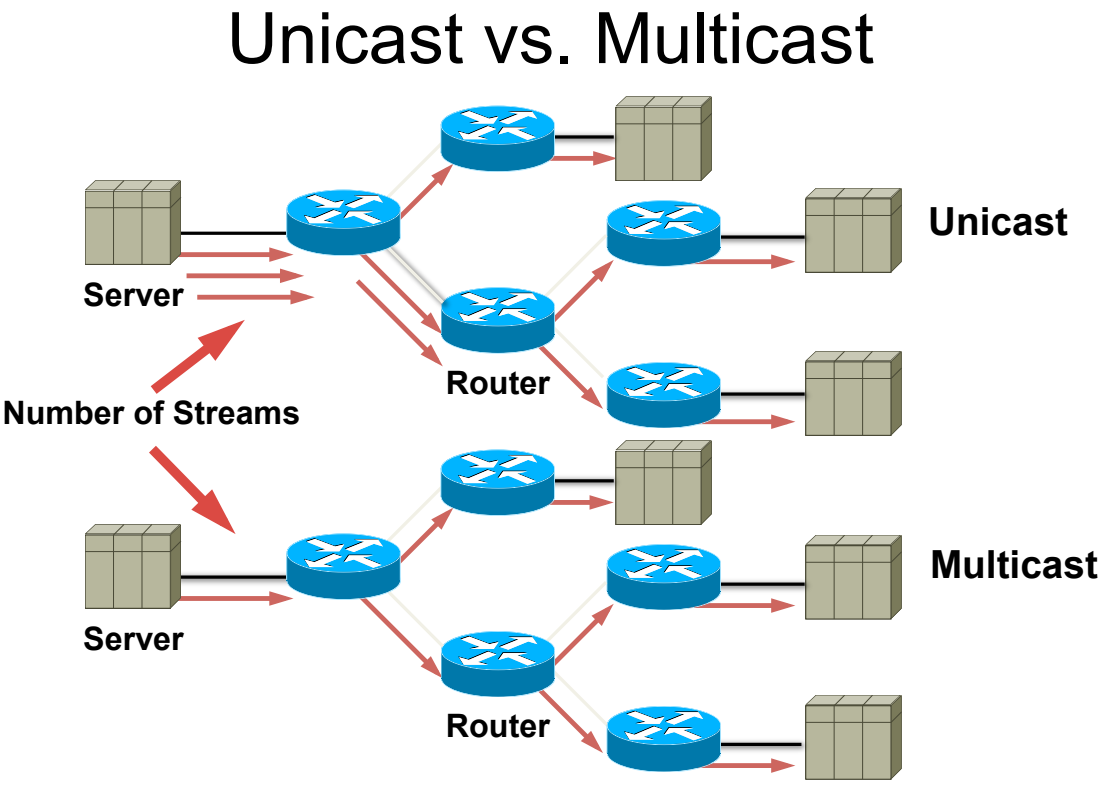
Distinguished Engineer, Cisco

Februafry 2013

Agenda

- Why Multicast?
- Multicast Fundamentals
- PIM Protocols
- RP Choices
- Multicast at Layer 2
- Interdomain IP Multicast
- Provider Services

Why Multicast?



A Brief History of Multicast

Steven Deering, 1985, Stanford University

Yeah, he was way ahead of his time and too clever for all of us.

A solution for layer2 applications in the growing layer3 campus network

-Think overlay broadcast domain

- Like VXLan ;)

Broadcast Domain

- all members receive
- all members can source
- members dynamically come and go

A Brief History of Multicast

RFC966 - 1985

```
Multi-destination delivery is useful to
several applications, including:
      - distributed, replicated databases [6,9].
      - conferencing [11].
      - distributed parallel computation,
including distributed gaming [2].
```

All inherently many-to-many applications

No mention of one-to-many services such as Video/IPTV

A Brief History of Multicast

Overlay Broadcast Domain Requirements

- Tree building and maintenance
- Network-based source discovery
- Source route information
- Overlay mechanism – tunneling

The first solution had it all

Distance Vector Multicast Routing Protocol
DVMRP, RFC1075 – 1988

A Brief History of Multicast

PIM – Protocol Independent Multicast

"Independent" of which unicast routing protocol you run
It does require that you're running one. ☺
Uses local routing table to determine route to sources

Router-to-router protocol to build and maintain distribution trees

Source discovery handled one of two ways:
1) Flood-and-prune PIM-DM, Dense Mode
2) Explicit Join w/ Rendezvous Point (RP) PIM-SM, Sparse Mode - The Current Standard

A Brief History of Multicast

PIM-SM – Protocol Independent Multicast Sparse Mode

- Tree building and maintenance
- Network-based source discovery
- ~~Source route information~~
- ~~Overlay mechanism – tunneling~~

Long, Sordid IETF history

RFC4601 – 2006 (original draft was rewritten from scratch)

Primary challenges to the final specification were in addressing Network-based source discovery.

A Brief History of Multicast

Today's dominant applications are primarily one-to-many

IPTV, Contribution video over IP, etc.

Sources are well known

SSM – Source Specific Multicast

RFC3569, RFC4608 – 2003

- Tree building and maintenance
- ~~Network-based source discovery~~
- ~~Source route information~~
- ~~Overlay mechanism – tunneling~~

Very simple and the preferred solution for one-to-many applications

Multicast Uses

- Any applications with multiple receivers
  - One-to-many or many-to-many
- Live video distribution
- Collaborative groupware
- Periodic data delivery—"push" technology
  - Stock quotes, sports scores, magazines, newspapers, adverts
- Server/Website replication
- Reducing network/resource overhead
  - More than multiple point-to-point flows
- Resource discovery
- Distributed interactive simulation (DIS)
  - War games
  - Virtual reality

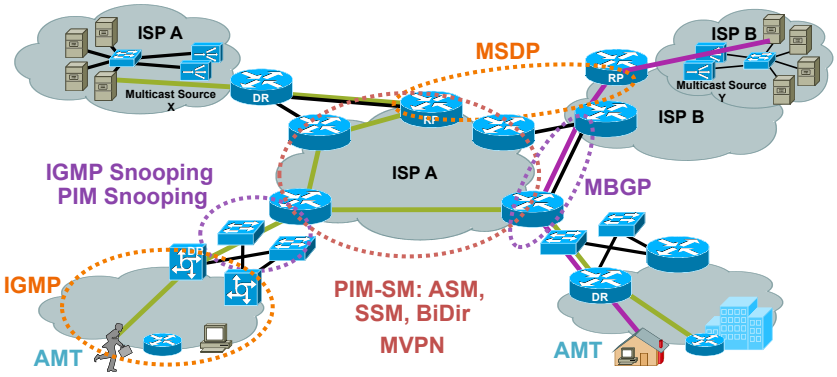
Multicast Considerations

Multicast Is UDP-Based

- **Best effort delivery:** Drops are to be expected; multicast applications should not expect reliable delivery of data and should be designed accordingly; reliable multicast is still an area for much research; expect to see more developments in this area; PGM, FEC, QoS
- **No congestion avoidance:** Lack of TCP windowing and "slow-start" mechanisms can result in network congestion; if possible, multicast applications should attempt to detect and avoid congestion conditions
- **Duplicates:** Some multicast protocol mechanisms (e.g., asserts, registers, and SPT transitions) result in the occasional generation of duplicate packets; multicast applications should be designed to expect occasional duplicate packets
- **Out of order delivery:** Some protocol mechanisms may also result in out of order delivery of packets

Multicast Fundamentals

Multicast Components



- End stations (hosts-to-routers)
  - IGMP, AMT

Campus Multicast

- Switches (Layer 2 optimization)
  - IGMP snooping PIM snooping
- Routers (multicast forwarding protocol)
  - PIM sparse mode or bidirectional PIM

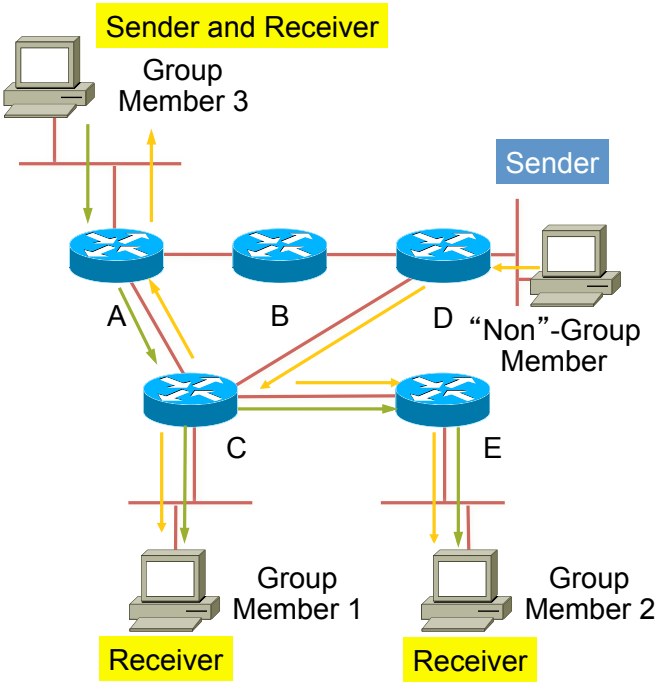
- Multicast routing across domains
  MBGP

Interdomain Multicast

- Multicast source discover
  MSDP with PIM-SM
- Source Specific Multicast
  SSM

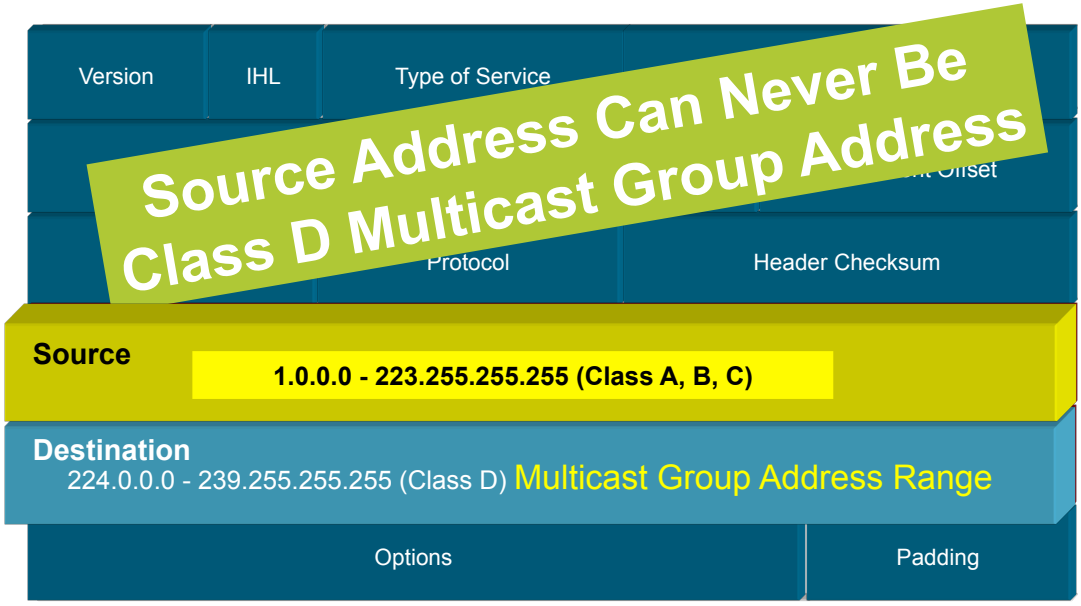
IP Multicast Group Concept

1. You must be a “member” of a group to receive its data
2. If you send to a group address, all members receive it
3. You do not have to be a member of a group to send to a group



Multicast Addressing

IPv4 Header



Multicast Addressing—224/4

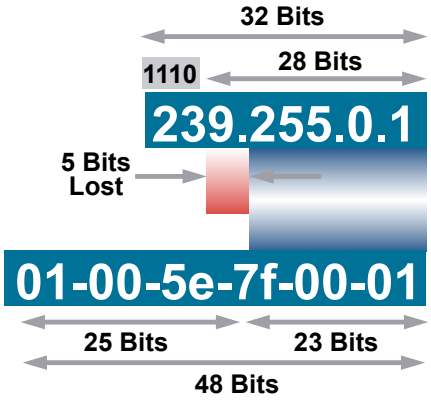
- Reserved link-local addresses
  - 224.0.0.0–224.0.0.255
  - Transmitted with TTL = 1
  - Examples
    - 224.0.0.1 All systems on this subnet
    - 224.0.0.2 All routers on this subnet
    - 224.0.0.5 OSPF routers
    - 224.0.0.13 PIMv2 routers
    - 224.0.0.22 IGMPv3
- Other reserved addresses
  - 224.0.1.0–224.0.1.255
  - Not local in scope (transmitted with TTL > 1)
  - Examples
    - 224.0.1.1 NTP (Network Time Protocol)
    - 224.0.1.32 Mtrace routers
    - 224.0.1.78 Tibco Multicast1

Multicast Addressing—224/4

- Administratively scoped addresses
  - 239.0.0.0–239.255.255.255
  - Private address space
    - Similar to RFC1918 unicast addresses
    - Not used for global Internet traffic—scoped traffic
- GLOP (honest, it's not an acronym)
  - 233.0.0.0–233.255.255.255
  - Provides /24 group prefix per ASN
- SSM (Source Specific Multicast) range
  - 232.0.0.0–232.255.255.255
  - Primarily targeted for Internet-style broadcast

Multicast Addressing

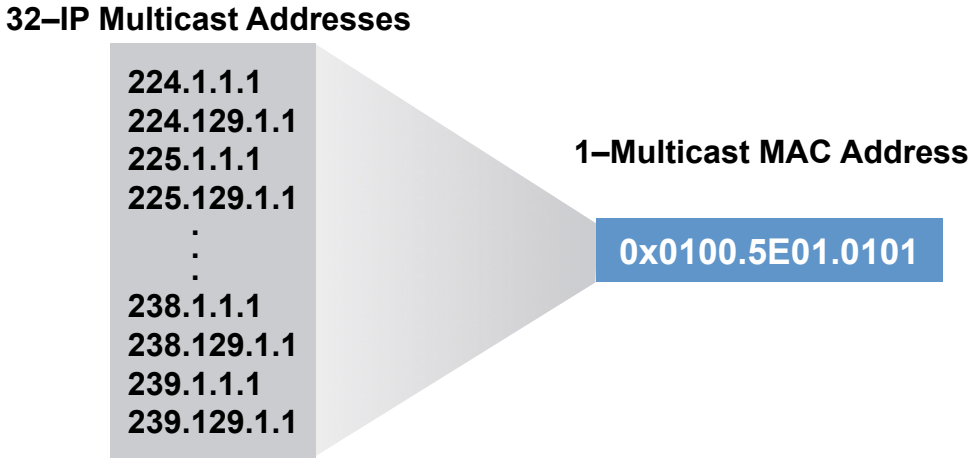
IP Multicast MAC Address Mapping



Multicast Addressing

IP Multicast MAC Address Mapping

Be Aware of the 32:1 Address Overlap



How Are Multicast Addresses Assigned?

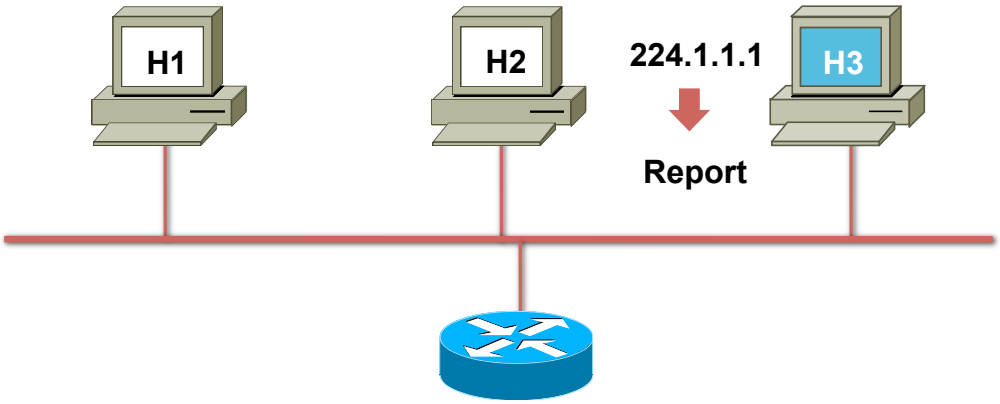
- Static global group address assignment
  - Temporary method to meet immediate needs
  - Group range: 233.0.0.0–233.255.255.255
    - Your AS number is inserted in middle two octets
    - Remaining low-order octet used for group assignment
  - Defined in RFC 2770
    - "GLOP Addressing in 233/8"
  - SSM does not require group address "ownership"
- Manual address allocation by the admin
  - Is still the most common practice

Host-Router Signaling: IGMP

- How hosts tell routers about group membership
- Routers solicit group membership from directly connected hosts
- RFC 1112 specifies version 1 of IGMP
  - Supported from Windows 95 on..
- RFC 2236 specifies version 2 of IGMP
  - Supported on latest service pack for Windows and most UNIX systems
- RFC 3376 specifies version 3 of IGMP
  - Supported in Window XP and various UNIX systems

Host-Router Signaling: IGMP

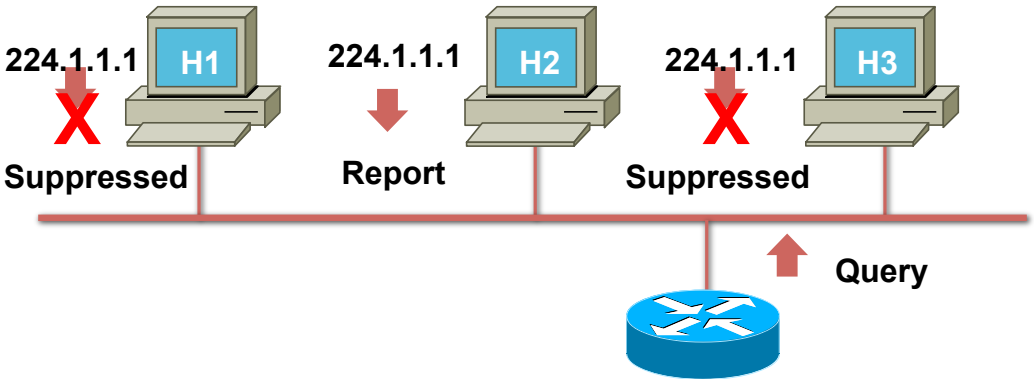
Joining a Group



- Host sends IGMP report to join group

Host-Router Signaling: IGMP

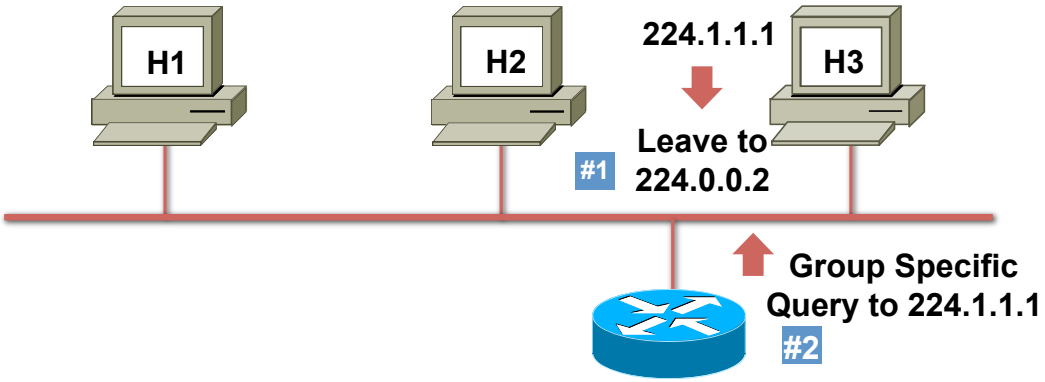
Maintaining a Group



- Router sends periodic queries to 224.0.0.1
- One member per group per subnet reports
- Other members suppress reports

Host-Router Signaling: IGMP

Leaving a Group (IGMPv2)



- Host sends leave message to 224.0.0.2
- Router sends group-specific query to 224.1.1.1
- No IGMP report is received within ~ 3 seconds
- Group 224.1.1.1 times out

Host-Router Signaling: IGMPv3

RFC 3376 – enables SSM

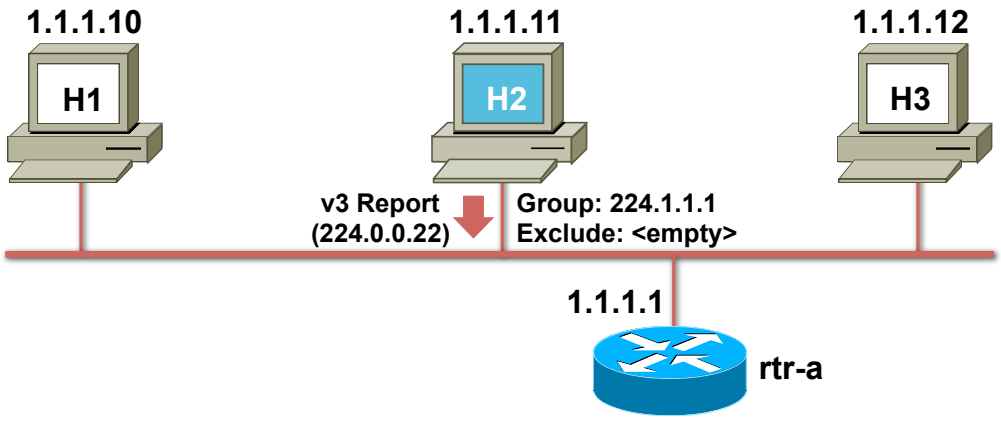
- Adds include/exclude source lists
- Enables hosts to listen only to a specified subset of the hosts sending to the group
- Requires new 'IPMulticastListen' API
- New IGMPv3 stack required in the OS
- Apps must be rewritten to use IGMPv3 include/ exclude features

Host-Router Signaling: IGMPv3

New Membership Report Address

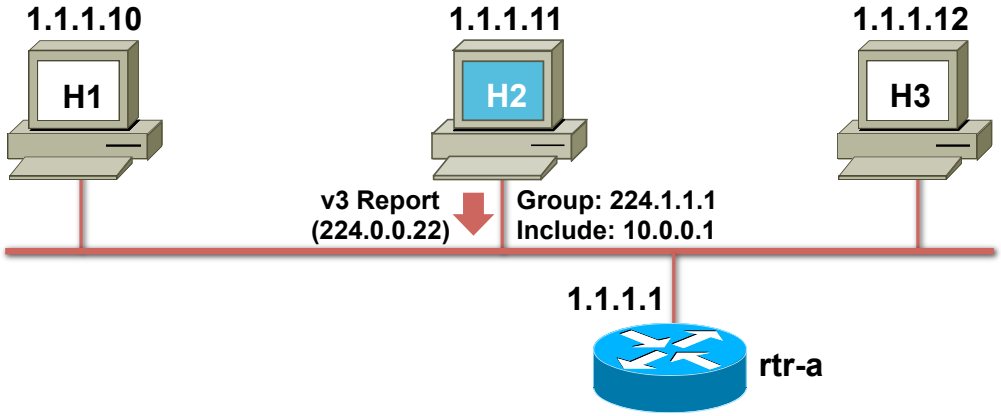
- 224.0.0.22 (IGMPv3 routers)
  - All IGMPv3 hosts send reports to this address
    - Instead of the target group address as in IGMPv1/v2
  - All IGMPv3 routers listen to this address
  - Hosts do not listen or respond to this address
- No report suppression
  - All hosts on wire respond to queries
    - Host's complete IGMP state sent in single response
  - Response interval may be tuned over broad range
    - Useful when large numbers of hosts reside on subnet

IGMPv3—Joining a Group



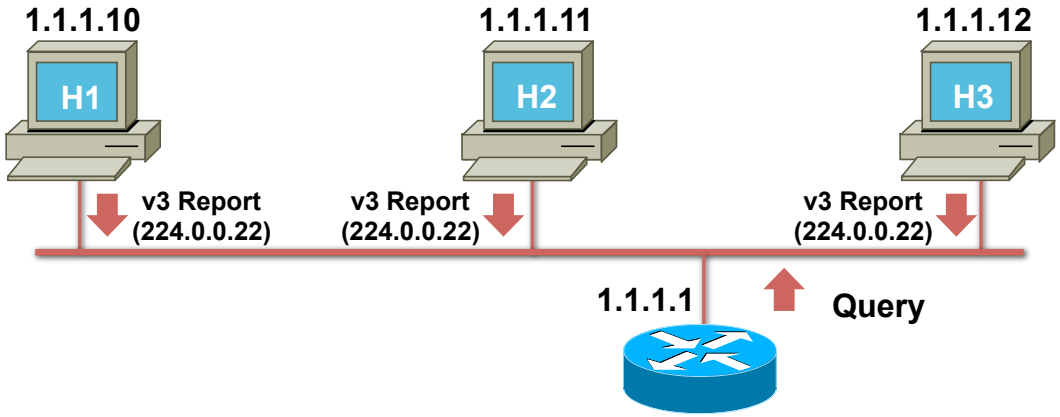
- Joining member sends IGMPv3 report to 224.0.0.22 immediately upon joining

IGMPv3—Joining Specific Source(s)



- IGMPv3 report contains desired source(s) in the include list
- Only "Included" source(s) are joined

IGMPv3—Maintaining State



- Router sends periodic queries
- All IGMPv3 members respond
- Reports contain multiple group state records

Multicast L3 Forwarding

Multicast Routing is Backwards from Unicast Routing

- Unicast routing is concerned about where the packet is going
- Multicast routing is concerned about where the packet came from
  - Initially

Unicast vs. Multicast Forwarding

Unicast Forwarding

- Destination IP address directly indicates where to forward packet
- Forwarding is hop-by-hop
  - Unicast routing table determines interface and next-hop router to forward packet

Unicast vs. Multicast Forwarding

Multicast Forwarding

- Destination IP address (group) doesn't directly indicate where to forward packet
- Forwarding is Outging Interface List dependent (OIF)
  - Receivers must first be "connected" to the tree before traffic begins to flow
    - Connection messages (PIM joins) follow unicast routing table toward multicast source
    - Build multicast distribution trees that determine where to forward packets
    - Distribution trees rebuilt dynamically in case of network topology changes
    - Each router in the path maintains an OIF list per tree state

Reverse Path Forwarding (RPF)

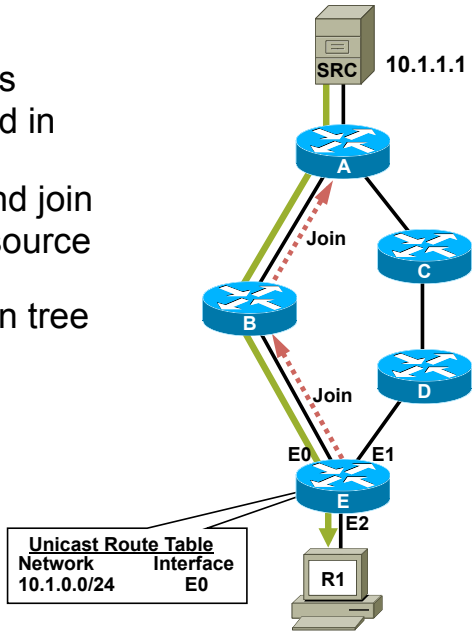
The RPF Calculation

- The multicast source address is checked against the unicast routing table
- This determines the interface and upstream router in the direction of the source to which PIM joins are sent
- This interface becomes the "Incoming" or RPF interface
  - A router forwards a multicast datagram only if received on the RPF interface

Reverse Path Forwarding (RPF)

RPF Calculation

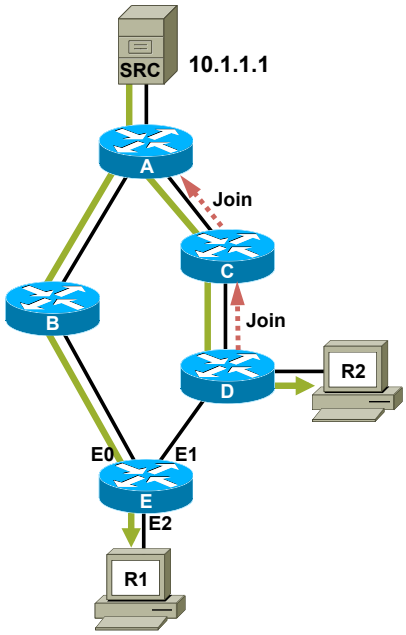
- Based on source address
- Best path to source found in unicast route table
- Determines where to send join
- Joins continue towards source to build multicast tree
- Multicast data flows down tree



Reverse Path Forwarding (RPF)

RPF Calculation

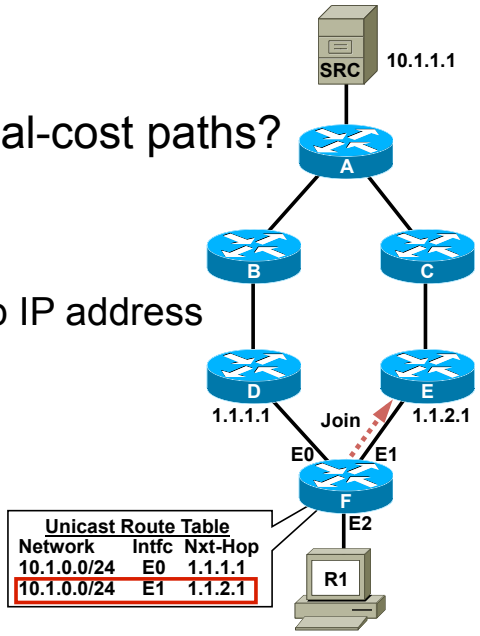
- Based on source address
- Best path to source found in unicast route table
- Determines where to send join
- Joins continue towards source to build multicast tree
- Multicast data flows down tree
- Repeat for other receivers



Reverse Path Forwarding (RPF)

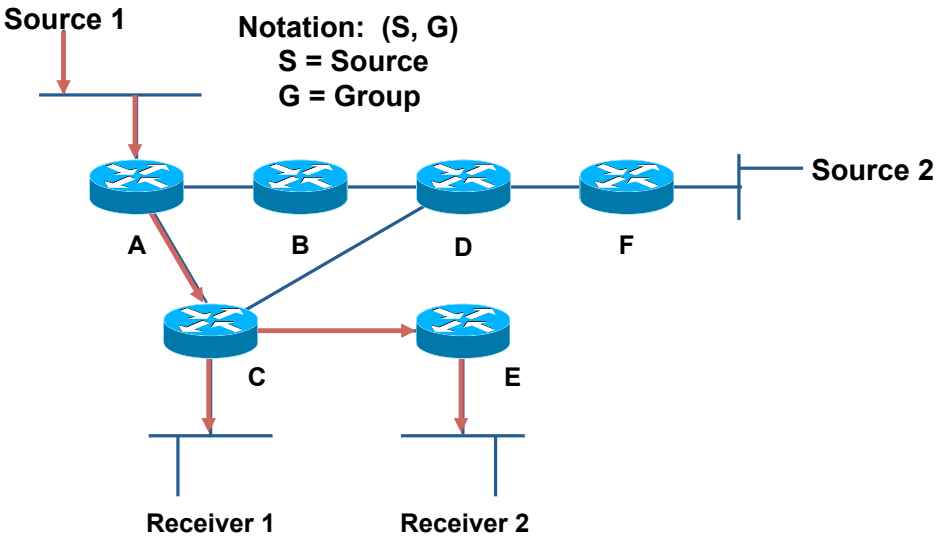
RPF Calculation

- What if we have equal-cost paths?
  - We can't use both
- Tie-breaker
  - Use highest next-hop IP address



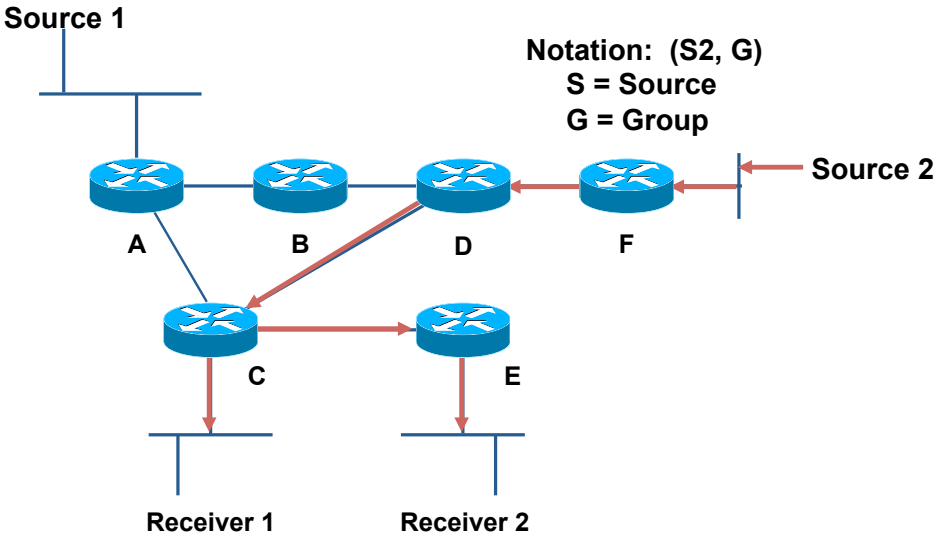
Multicast Distribution Trees

Shortest Path or Source Distribution Tree



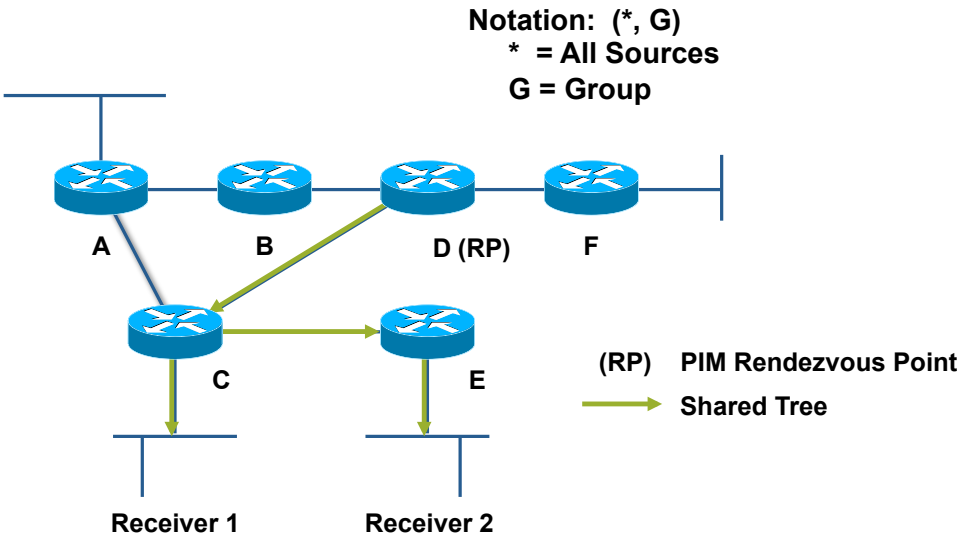
Multicast Distribution Trees

Shortest Path or Source Distribution Tree

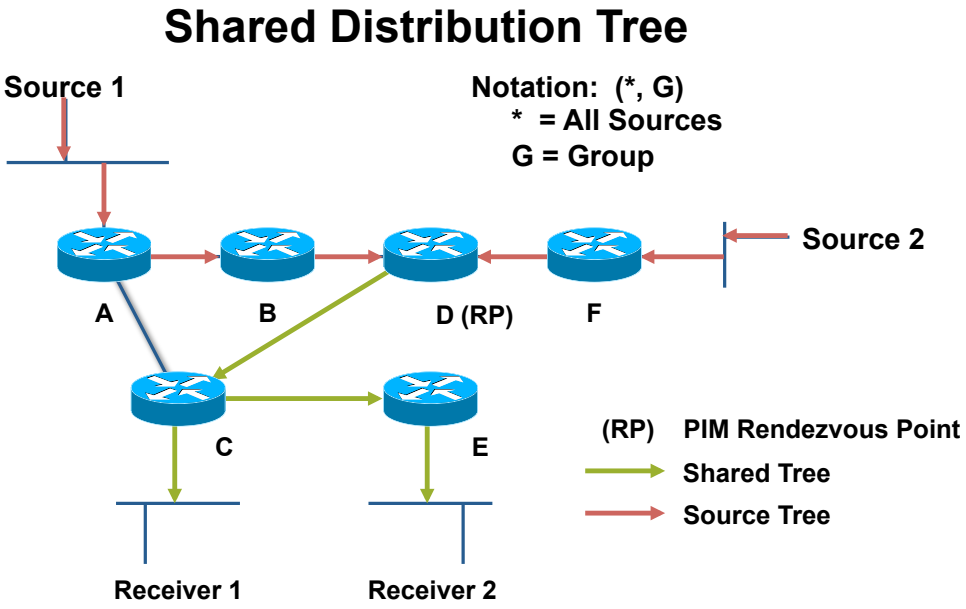


Multicast Distribution Trees

Shared Distribution Tree



Multicast Distribution Trees



Multicast Distribution Trees

Characteristics of Distribution Trees

- Source or shortest path trees
  - Uses more memory O (S x G) but you get optimal paths from source to all receivers; minimizes delay
- Shared trees
  - Uses less memory O(G) but you may get suboptimal paths
    from source to all receivers; may introduce extra delay

Multicast Tree Creation

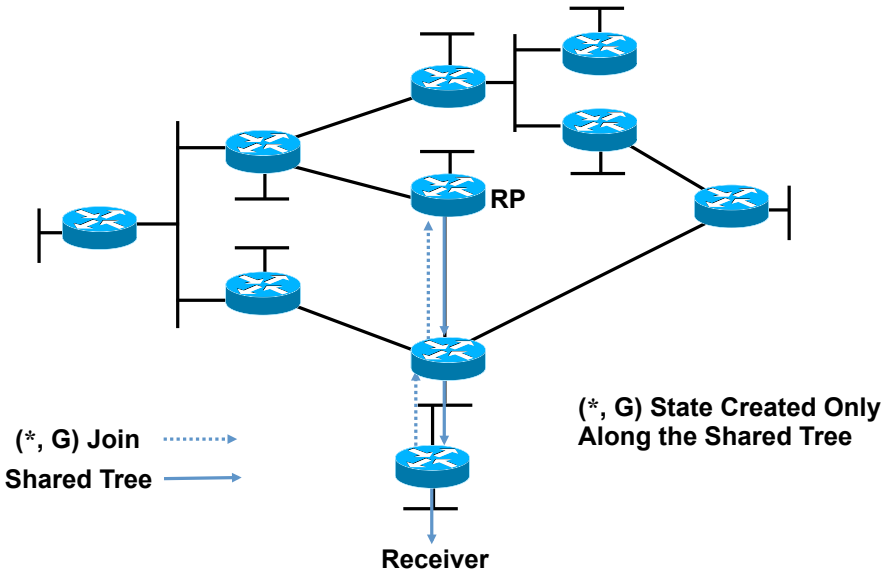
- PIM join/prune control messages
  - Used to create/remove distribution trees
- Shortest path trees
  - PIM control messages are sent toward the source
- Shared trees
  - PIM control messages are sent toward RP

PIM Protocol Variants

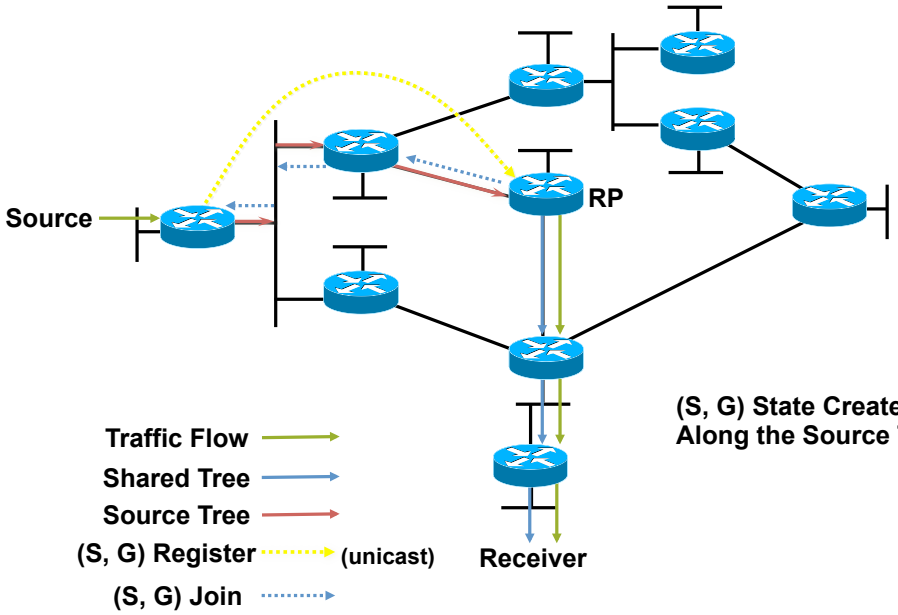
Major Deployed PIM Variants

- PIM-SM
  - ASM
    - Any Source Multicast/RP/SPT/shared tree
  - SSM
    - Source Specific Multicast, no RP, SPT only
  - BiDir
    - Bidirectional PIM, no SPT, shared tree only

PIM-SM Shared Tree Join



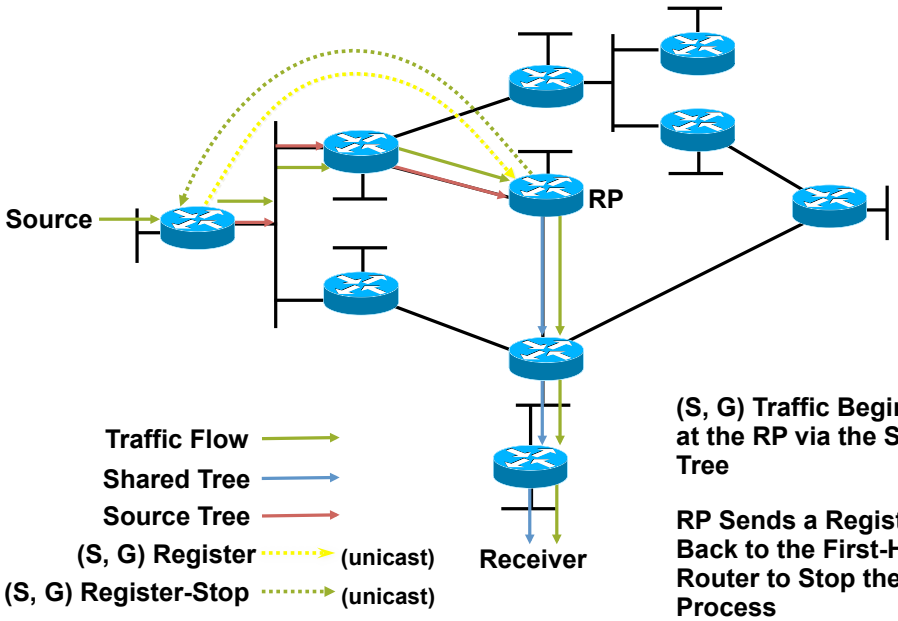
PIM-SM Sender Registration



Traffic Flow
Shared Tree
Source Tree
(S, G) Register (unicast)
(S, G) Join

(S, G) State Created Only Along the Source Tree

PIM-SM Sender Registration

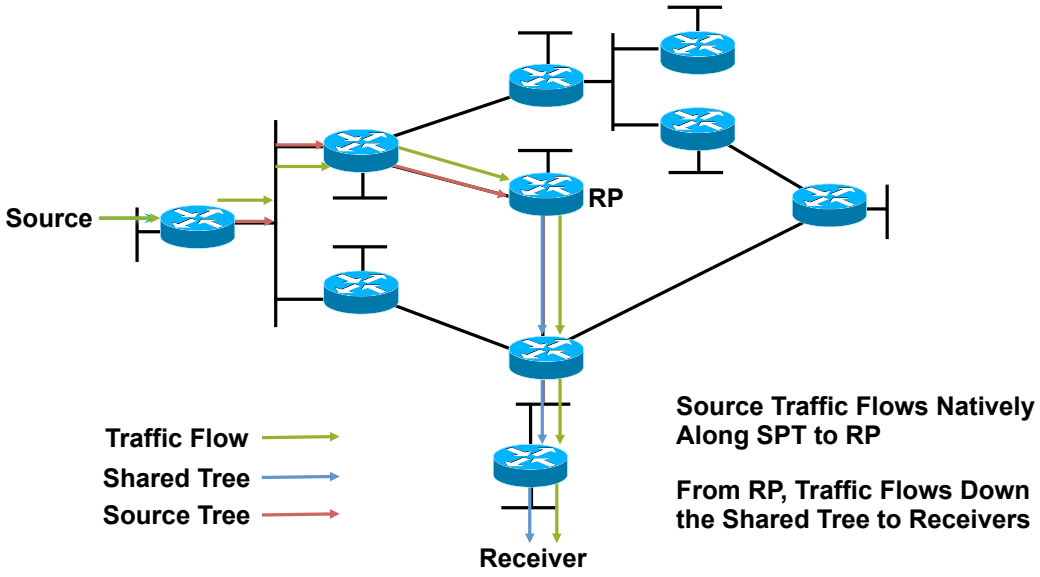


Traffic Flow
Shared Tree
Source Tree
(S, G) Register (unicast)
(S, G) Register-Stop (unicast)

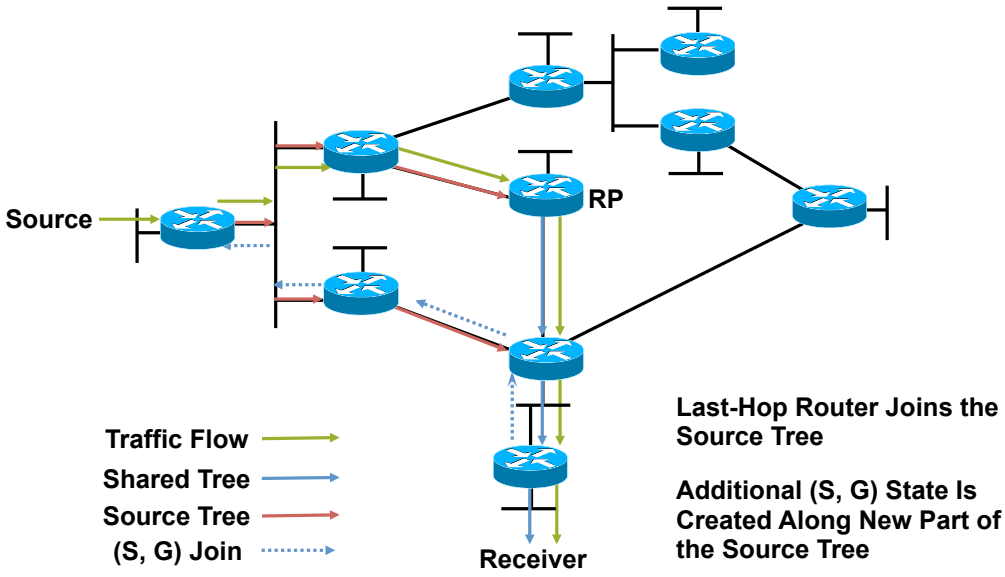
(S, G) Traffic Begins Arriving at the RP via the Source Tree

RP Sends a Register-Stop Back to the First-Hop Router to Stop the Register Process

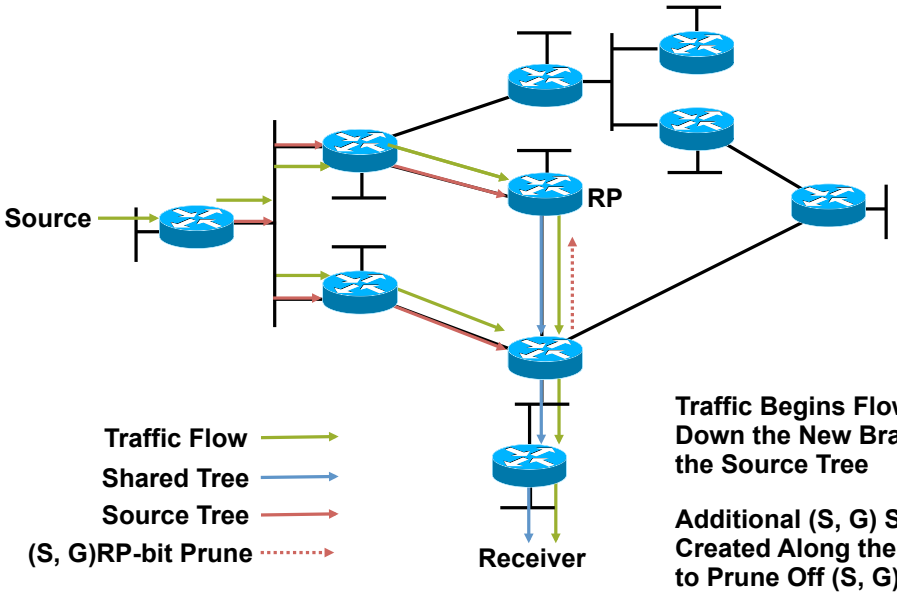
PIM-SM Sender Registration



PIM-SM SPT Switchover



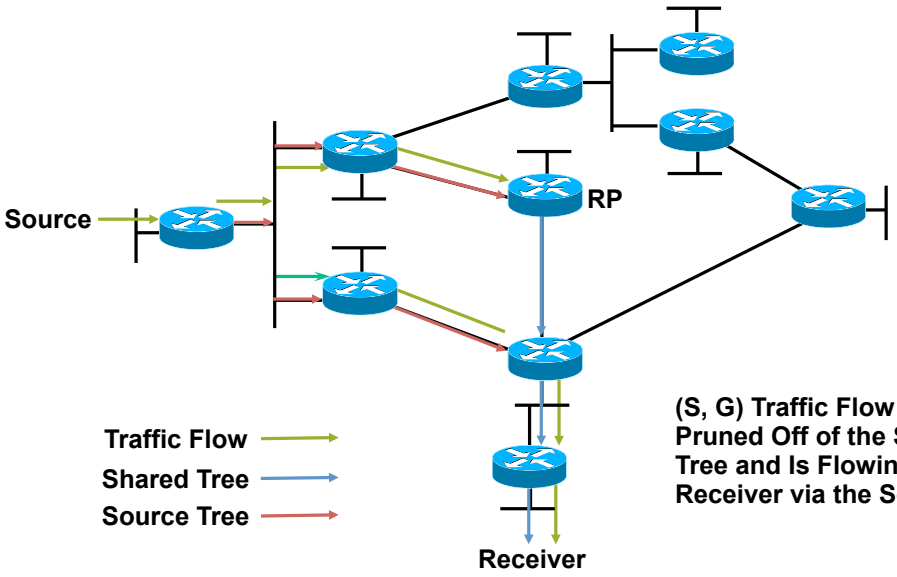
PIM-SM SPT Switchover



Traffic Begins Flowing Down the New Branch of the Source Tree

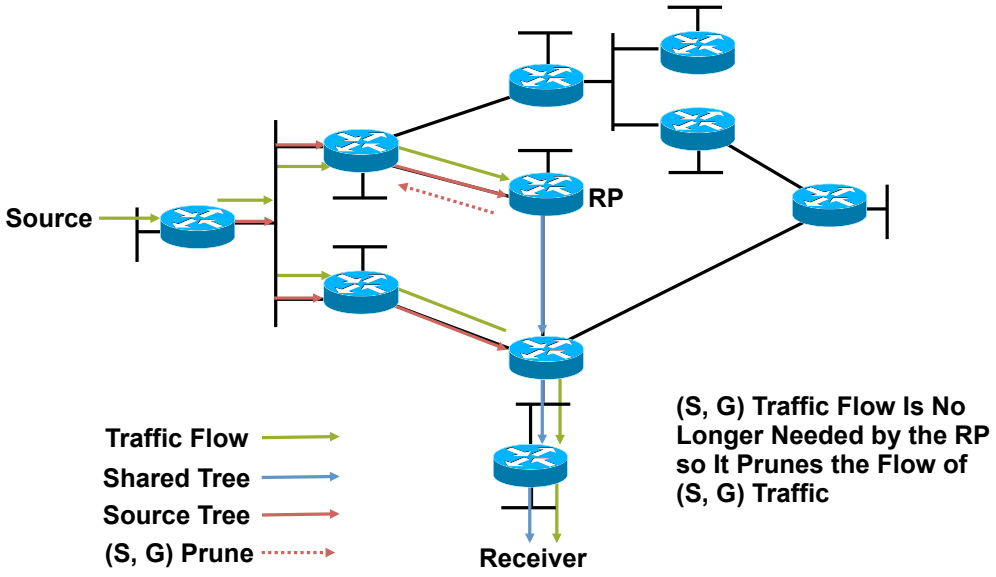
Additional (S, G) State Is Created Along the Shared Tree to Prune Off (S, G) Traffic

PIM-SM SPT Switchover

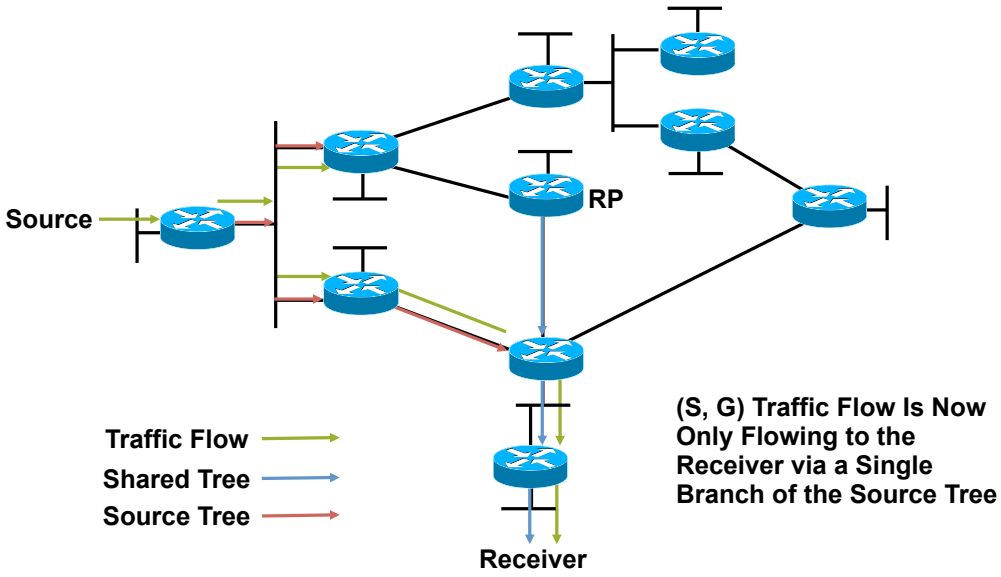


(S, G) Traffic Flow Is Now Pruned Off of the Shared Tree and Is Flowing to the Receiver via the Source Tree

PIM-SM SPT Switchover



PIM-SM SPT Switchover



“The default behavior of PIM-SM is that routers with directly connected members will join the shortest path tree as soon as they detect a new multicast source.”

PIM Frequently Forgotten Fact

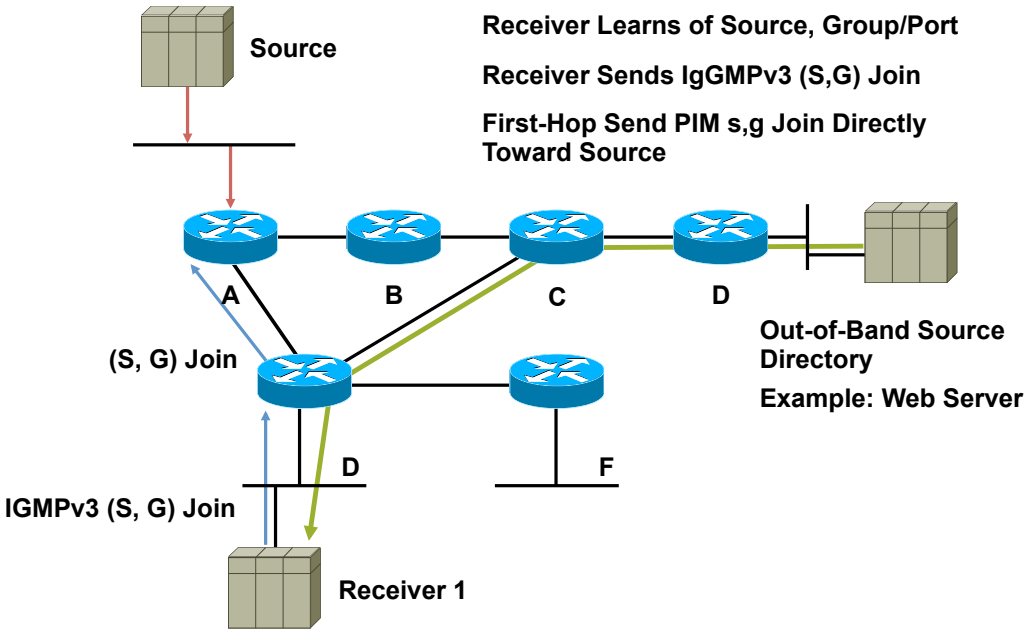
PIM-SM—Evaluation

- Effective for sparse or “dense” distribution of multicast receivers
- Advantages
  - Traffic only sent down “joined” branches
  - Can switch to optimal source-trees for high traffic sources dynamically
    - Sounds clever but it actually switches for all sources by default
  - Unicast routing protocol-independent
  - Basis for interdomain, multicast routing
    - When used with MBGP, MSDP and/or SSM

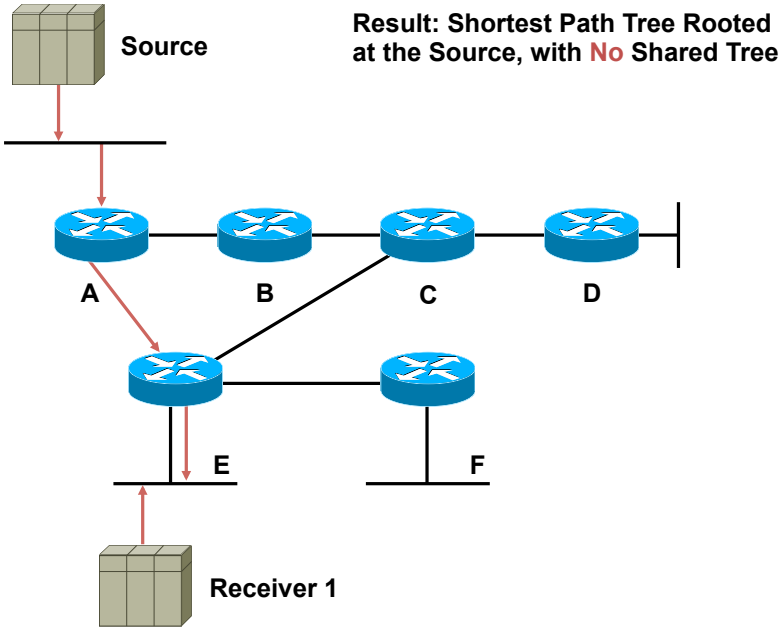
Source Specific Multicast

- Assume a one-to-any multicast model
  - Example: video/audio broadcasts, stock market data
- Why does ASM need a shared tree?
  - So that hosts and first hop routers can learn who the active source is for the source discovery
- What if this was already known?
  - Hosts could use IGMPv3 to signal exactly which (S, G) SPT to join
  - The shared tree and RP wouldn't be necessary
  - Different sources could share the same group address and not interfere with each other
- Result: Source Specific Multicast (SSM)
- RFC 3569: An Overview of Source Specific Multicast (SSM)

PIM Source Specific Mode



PIM Source Specific Mode



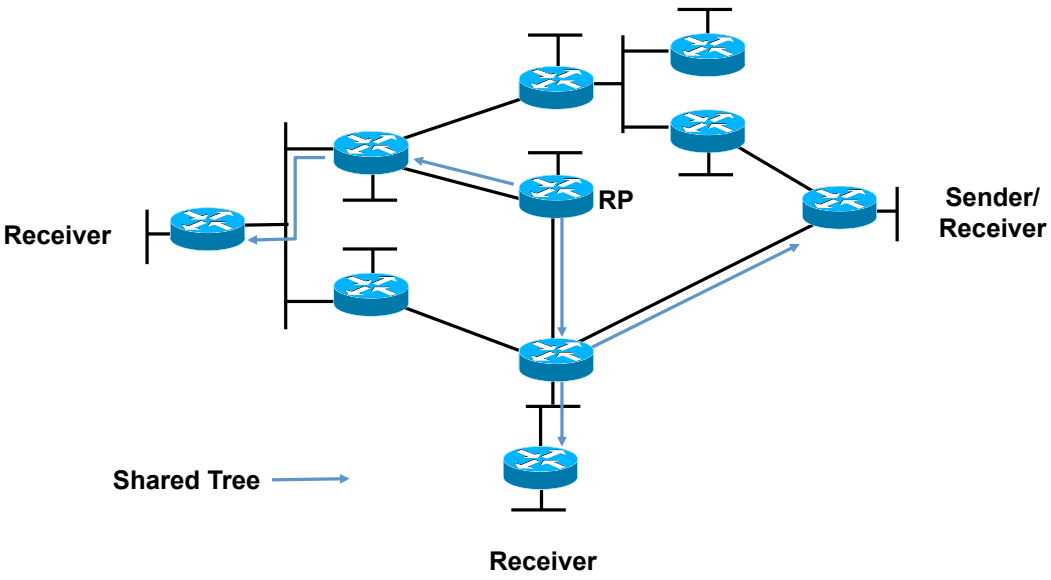
SSM—Evaluation

- Ideal for applications with one source sending to many receivers
- Uses a simplified subset of the PIM-SM protocol
  - Simpler network operation
- Solves multicast address allocation problems
  - Flows differentiated by both source and group
    - Not just by group
  - Content providers can use same group ranges
    - Since each (S,G) flow is unique
- Helps prevent certain DoS attacks
  - "Bogus" source traffic
    - Can't consume network bandwidth
    - Not received by host application

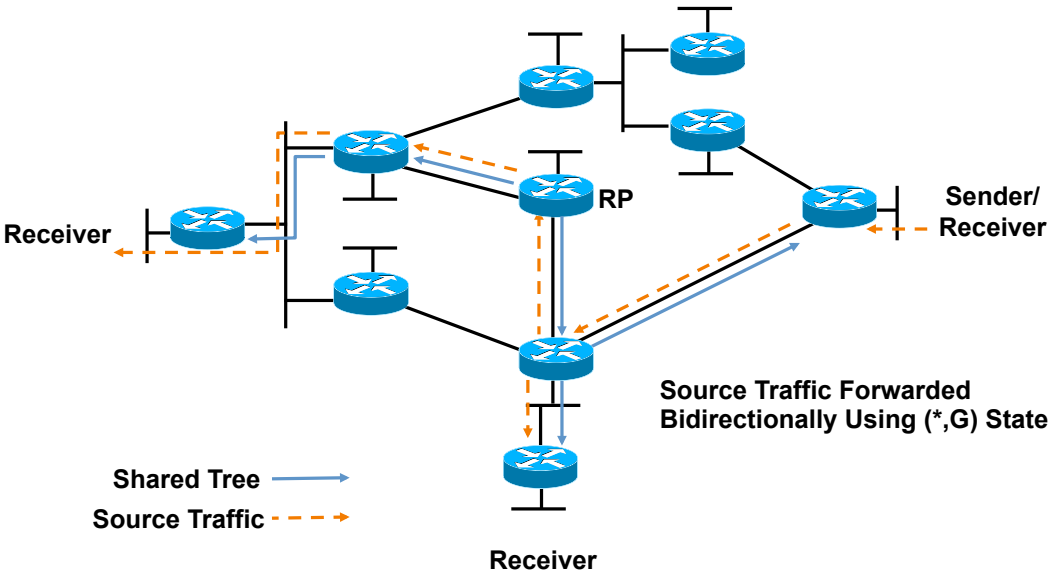
Many-to-Many State Problem

- Creates huge amounts of (S,G) state
  - State maintenance workloads skyrocket
    - High OIL fan-out makes the problem worse
  - Router performance begins to suffer
- Using shared trees only
  - Provides some (S, G) state reduction
    - Results in (S, G) state only along SPT to RP
    - Frequently still too much (S, G) state
    - Need a solution that only uses (*, G) state

Bidirectional PIM—Overview



Bidirectional PIM—Overview



Bidir PIM—Evaluation

- Ideal for many to many applications
- Drastically reduces network mroute state
  - Eliminates **all** (S,G) state in the network
    - SPTs between sources to RP eliminated
    - Source traffic flows both up and down shared tree
  - Allows many-to-any applications to scale
    - Permits virtually an unlimited number of sources

RP Choices

PIM-SM ASM RP Requirements

- Group to RP mapping
  - Consistent in all routers within the PIM domain
- RP redundancy requirements
  - Eliminate any single point of failure

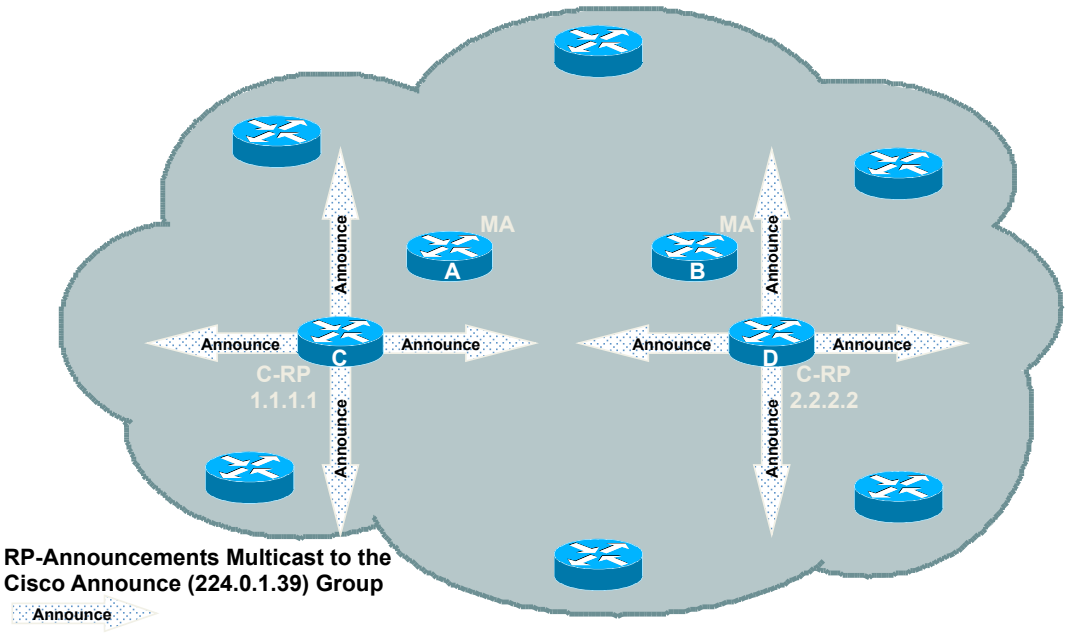
How Does the Network Know About the RP?

- Static configuration
  - Manually on every router in the PIM domain
- AutoRP
  - Originally a Cisco® solution
  - Facilitated PIM-SM early transition
- BSR
  - draft-ietf-pim-sm-bsr

Static RPs

- Hard-configured RP address
  - When used, must be configured on every router
  - All routers must have the same RP address
  - RP failover not possible
    - Exception: if anycast RPs are used
- 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

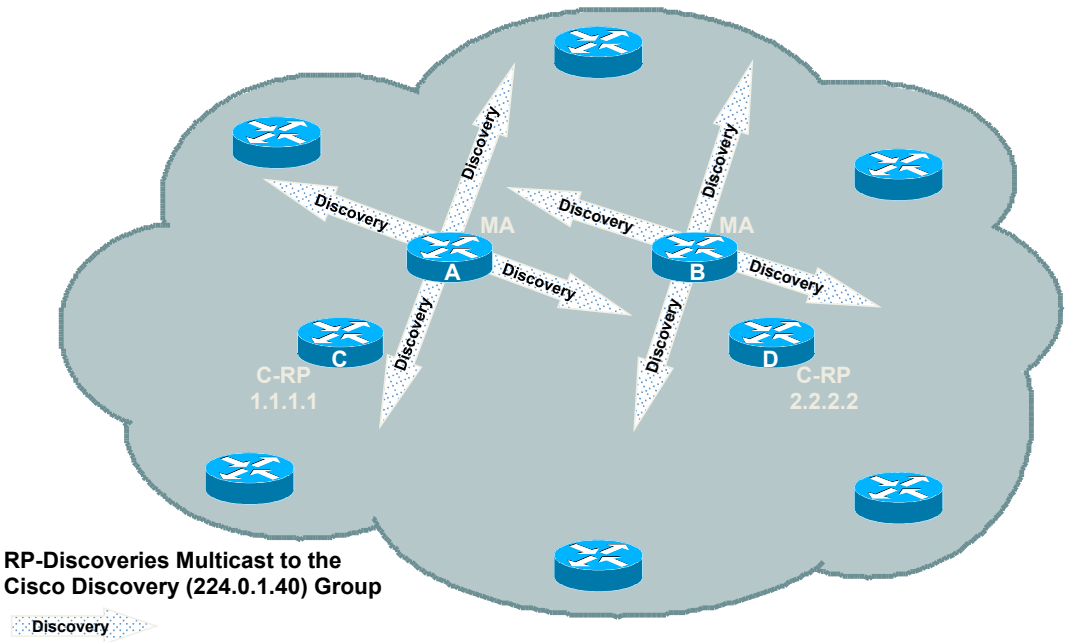
Auto-RP—From 10,000 Feet



RP-Announcements Multicast to the Cisco Announce (224.0.1.39) Group

Announce

Auto-RP—From 10,000 Feet

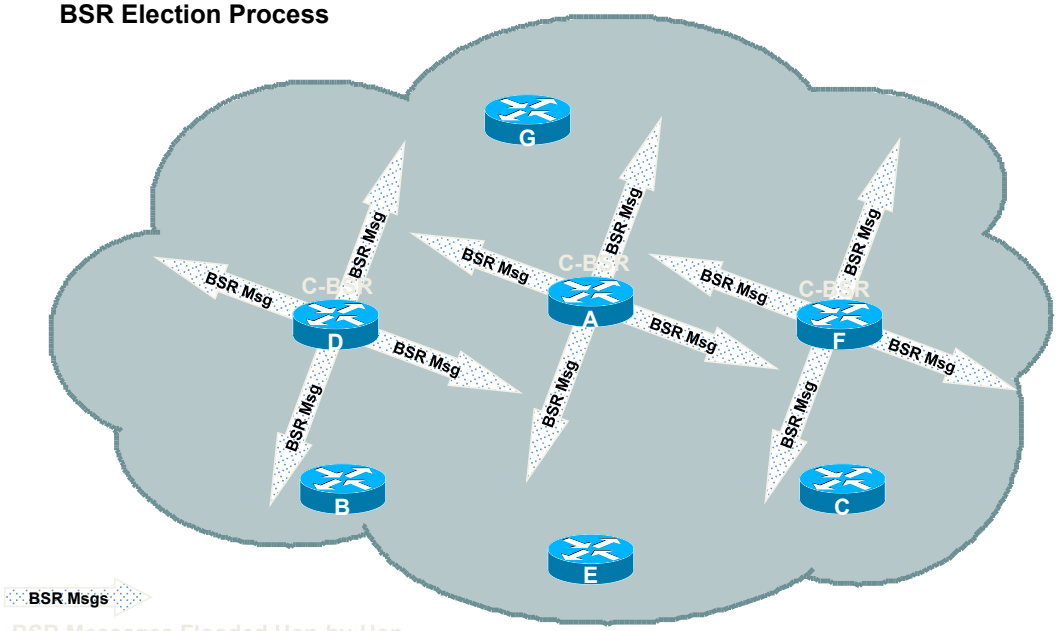


RP-Discoveries Multicast to the Cisco Discovery (224.0.1.40) Group

Discovery

BSR—From 10,000 Feet

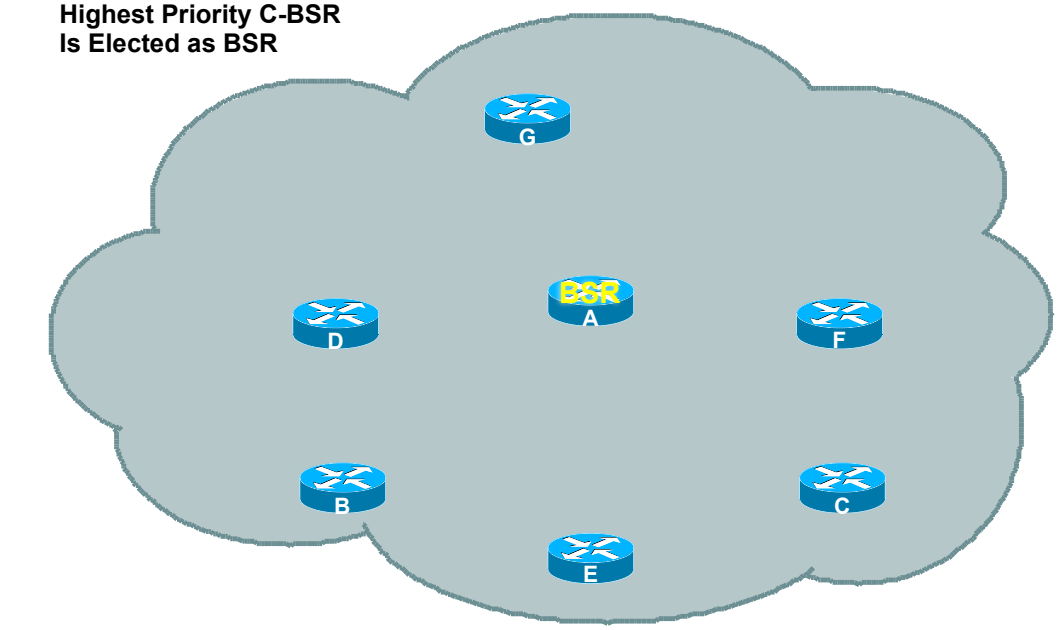
BSR Election Process



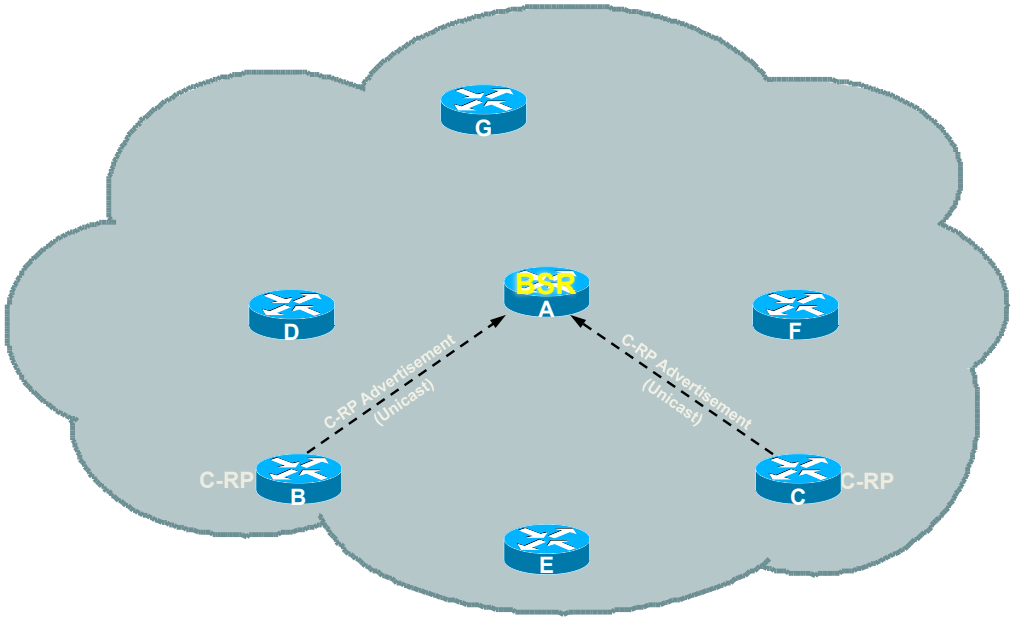
BSR Messages Flooded Hop-by-Hop

BSR—From 10,000 Feet

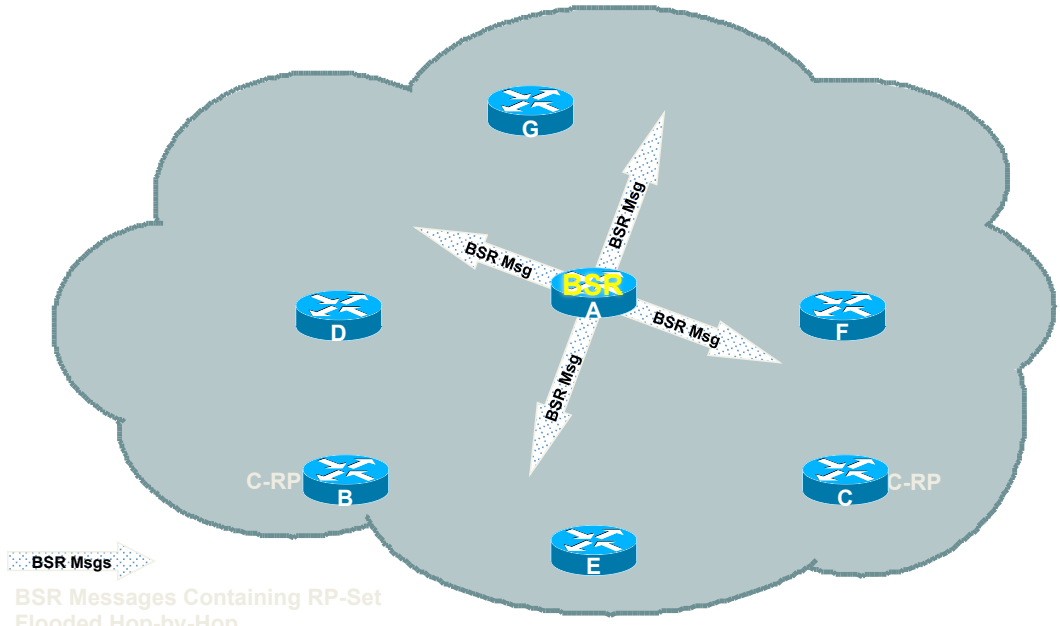
Highest Priority C-BSR Is Elected as BSR



BSR—From 10,000 Feet



BSR—From 10,000 Feet

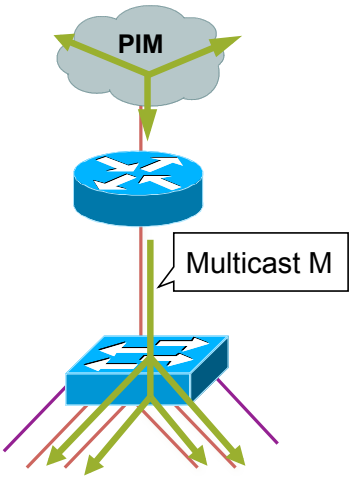


Multicast at Layer 2

L2 Multicast Frame Switching

Problem: Layer 2 Flooding of Multicast Frames

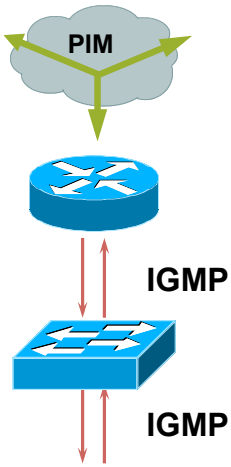
- Typical L2 switches treat multicast traffic as unknown or broadcast and must “flood” the frame to every port
- Static entries can sometimes be set to specify which ports should receive which group(s) of multicast traffic
- Dynamic configuration of these entries would cut down on user administration



L2 Multicast Frame Switching

IGMPv1–v2 Snooping

- Switches become “IGMP”-aware
- IGMP packets intercepted by the NMP or by special hardware ASICs
  - Requires special hardware to maintain throughput
- Switch must examine contents of IGMP messages to determine which ports want what traffic
  - IGMP membership reports
  - IGMP leave messages
- Impact on low-end, Layer 2 switches
  - Must process all Layer 2 multicast packets
  - Admin load increases with multicast traffic load
  - Generally results in switch **meltdown**



L2 Multicast Frame Switching

Impact of IGMPv3 on IGMP Snooping

- IGMPv3 reports sent to separate group (224.0.0.22)
  - Switches listen to just this group
  - Only IGMP traffic—no data traffic
  - Substantially reduces load on switch CPU
  - Permits low-end switches to implement IGMPv3 snooping
- No report suppression in IGMPv3
  - Enables individual member tracking
- IGMPv3 supports source-specific includes/ excludes

Summary—Frame Switches

IGMP Snooping

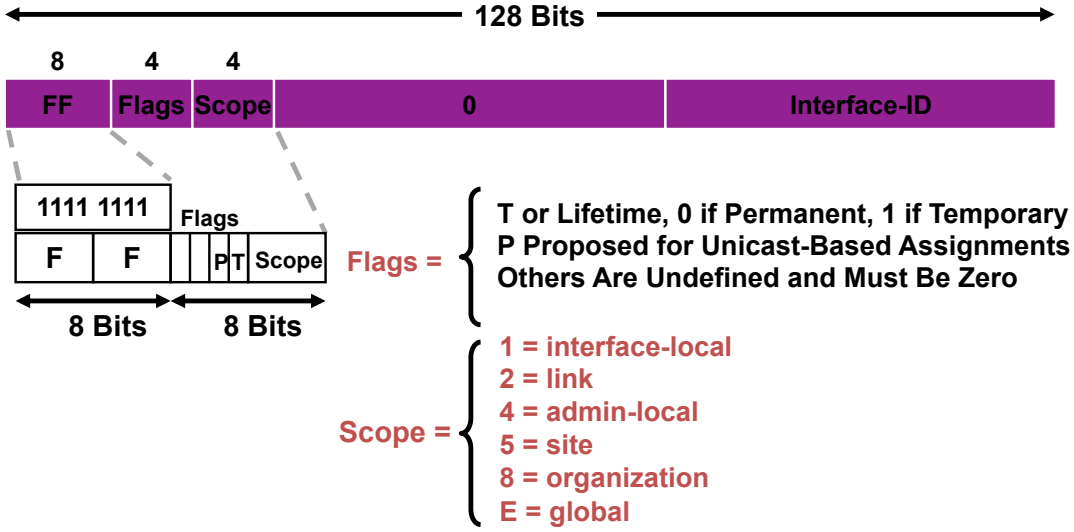
- Switches with Layer 3-aware hardware/ ASICs
- High-throughput performance maintained
- Increases cost of switches
- Switches without Layer 3-aware hardware/ ASICs
- Suffer serious performance degradation or even **meltdown!**
- Shouldn't be a problem when IGMPv3 is implemented

IPv6 Multicast

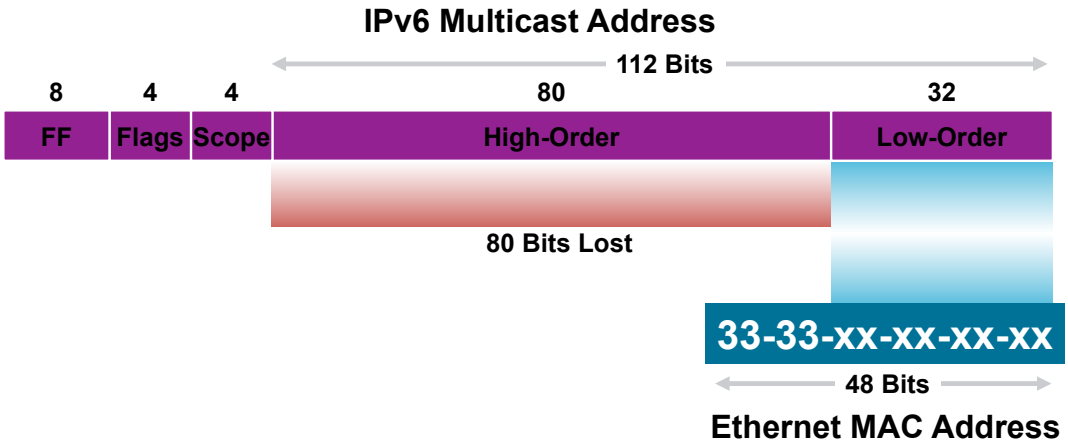
IPv4 vs. IPv6 Multicast

IP Service	IPv4 Solution	IPv6 Solution
Address Range	32-Bit, Class D	128-Bit (112-Bit Group)
Routing	Protocol-Independent All IGPs and GBP4+	Protocol-Independent All IGPs and BGP4+ with v6 Mcast SAFI
Forwarding	PIM-DM, PIM-SM: ASM, SSM, BiDir	PIM-SM: ASM, SSM, BiDir
Group Management	IGMPv1, v2, v3	MLDv1, v2
Domain Control	Boundary/Border	Scope Identifier
Interdomain Source Discovery	MSDP Across Independent PIM Domains	Single RP Within Globally Shared Domains

IPv6 Multicast Addresses (RFC3513)



IPv6 Layer 2 Multicast Addressing Mapping



Unicast-Based Multicast Addresses

- RFC 3306—unicast-based multicast addresses
  - Similar to IPv4 GLOP addressing
  - Solves IPv6 global address allocation problem
  - Flags = 00PT
    - P = 1, T = 1 → Unicast-based multicast address
- Example
  - Content provider's unicast prefix
    - 1234:5678:9::/48
  - Multicast address
    - FF3x:0030:1234:5678:0009::0001

IP Routing for Multicast

- RPF-based on reachability to v6 source same as with v4 multicast
- RPF still protocol-independent
  - Static routes, mroutes
  - Unicast RIB: BGP, ISIS, OSPF, EIGRP, RIP, etc.
  - Multiprotocol BGP (mBGP)
    - Support for v6 mcast subaddress family
    - Provide translate function for nonsupporting peers

IPv6 Multicast Forwarding

- PIM-Sparse Mode (PIM-SM)
  - RFC4601
- PIM Source Specific Mode (SSM)
  - RFC3569 SSM overview (v6 SSM needs MLDv2)
  - Unicast, prefix-based multicast addresses ff30::/12
  - SSM range is ff3X::/96
- PIM Bi-Directional Mode (BiDir)
  - draft-ietf-pim-bidir-09.txt

RP Mapping Mechanisms for IPv6

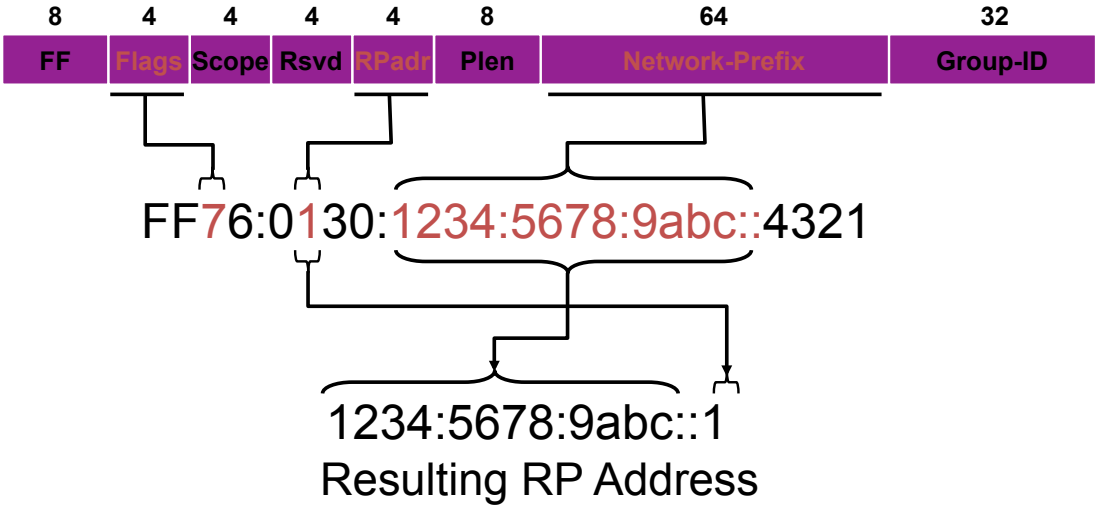
- Static RP assignment
- BSR
- Auto-RP—no current plans
- Embedded RP

Embedded RP Addressing—RFC3956

- Proposed new multicast address type
  - Uses unicast-based multicast addresses (RFC 3306)
- RP address is embedded in multicast address
- Flag bits = 0RPT
  - R = 1, P = 1, T = 1 → Embedded RP address
- Network-Prefix::RPadr = RP address
- For each unicast prefix you own, you now also own:
  - 16 RPs for each of the 16 multicast scopes (256 total) with 2^32 multicast groups assigned to each RP (2^40 total)

Embedded RP Addressing—Example

Multicast Address with Embedded RP Address



Multicast Listener Discover—MLD

- MLD is equivalent to IGMP in IPv4
- MLD messages are transported over ICMPv6
- Version number confusion
  - MLDv1 corresponds to IGMPv2
    - RFC 2710
  - MLDv2 corresponds to IGMPv3, needed for SSM
    - RFC 3810
- MLD snooping
  - draft-ietf-magma-snoop-12.txt

Interdomain IP Multicast

MBGP Overview

MBGP: Multiprotocol BGP

- Defined in RFC 2858 (extensions to BGP)
- Can carry different types of routes
  - Unicast
  - Multicast
- Both routes carried in same BGP session
- Does not propagate multicast state info
  - That's PIM's job
- Same path selection and validation rules
  - AS-Path, LocalPref, MED…

MBGP Overview

- Separate BGP tables maintained
  - Unicast prefixes for unicast forwarding
  - Unicast prefixes for multicast RPF checking
- AFI = 1, Sub-AFI = 1
  - Contains unicast prefixes for unicast forwarding
  - Populated with BGP unicast NLRI
- AFI = 1, Sub-AFI = 2
  - Contains unicast prefixes for RPF checking
  - Populated with BGP multicast NLRI

MBGP Overview

MBGP Allows Divergent Paths and Policies

- Same IP address holds dual significance
  - Unicast routing information
  - Multicast RPF information
- For same IPv4 address two different NLRI with different next-hops
- Can therefore support both congruent and incongruent topologies

MBGP—Summary

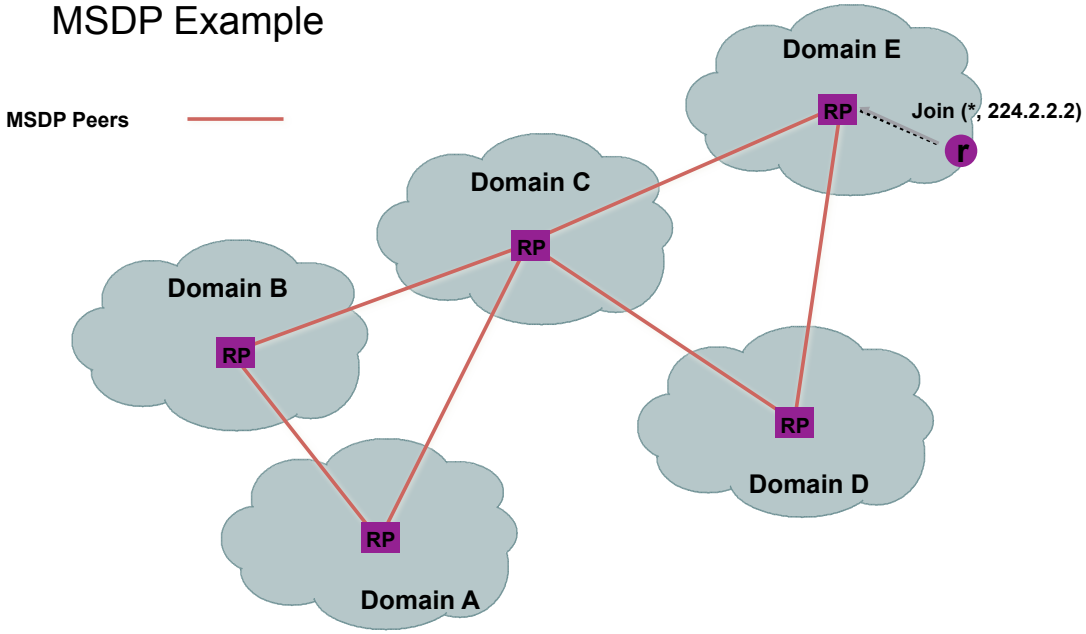
- Solves part of inter-domain problem
  - Can exchange unicast prefixes for multicast RPF checks
  - Uses standard BGP configuration knobs
  - Permits separate unicast and multicast topologies if desired
- Still must use PIM to:
  - Build distribution trees
  - Actually forward multicast traffic
  - PIM-SM recommended

MSDP

- RFC 3618
- ASM only
  - RPs knows about all sources in their domain
    - Sources cause a "PIM Register" to the RP
    - Tell RPs in other domains of it's sources
      - Via MSDP SA (Source Active) messages
  - RPs 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 required for interdomain ASM source discovery

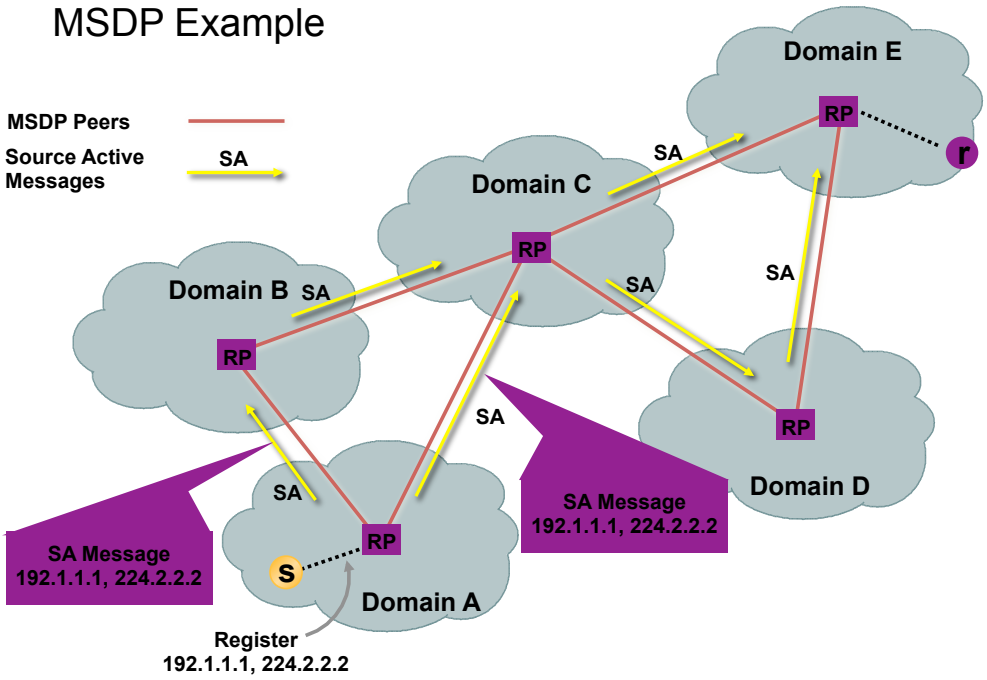
MSDP Overview

MSDP Example



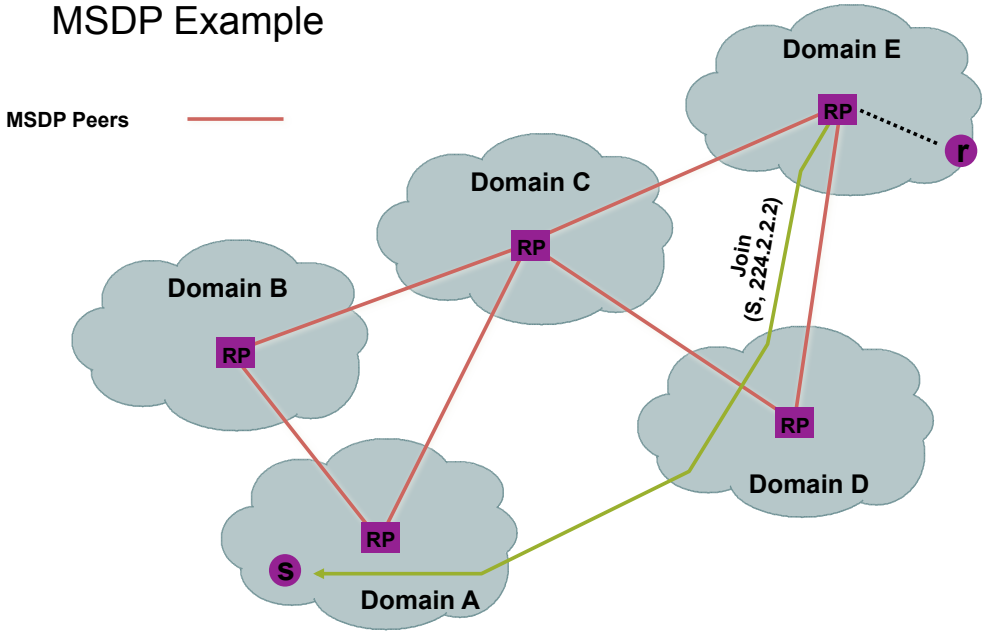
MSDP Overview

MSDP Example



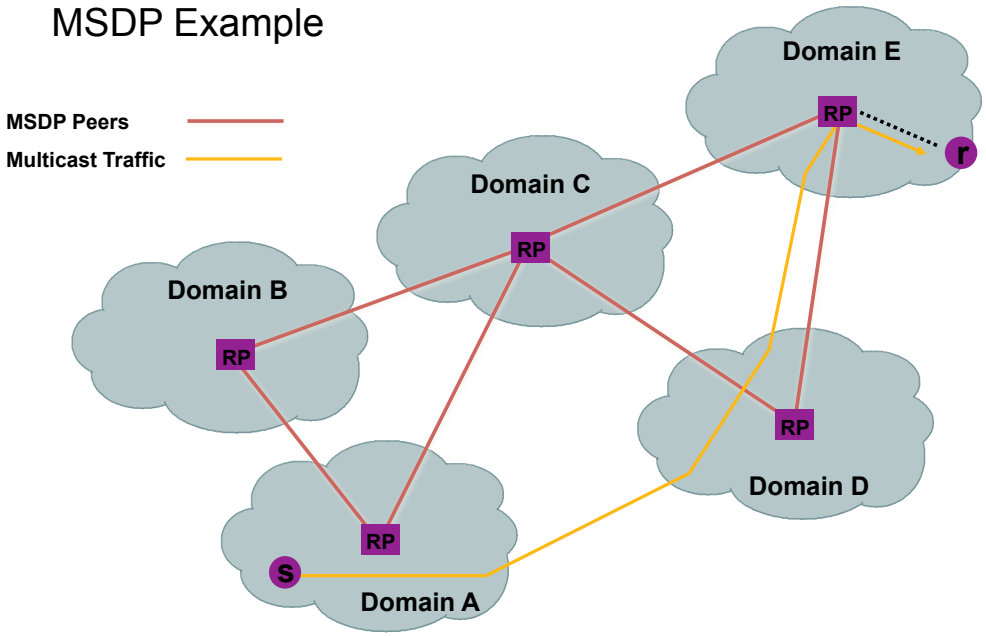
MSDP Overview

MSDP Example



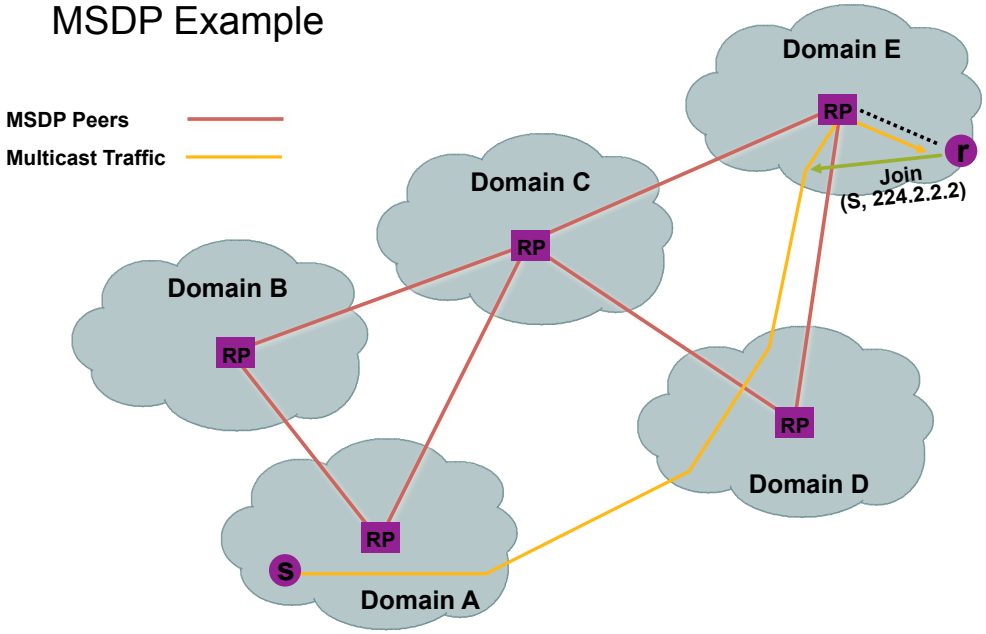
MSDP Overview

MSDP Example



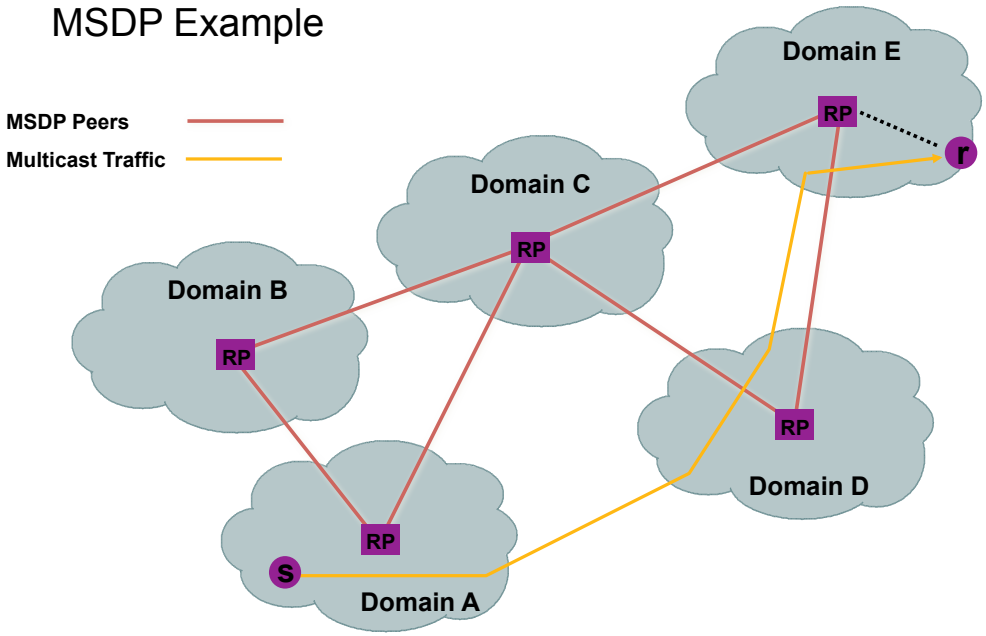
MSDP Overview

MSDP Example

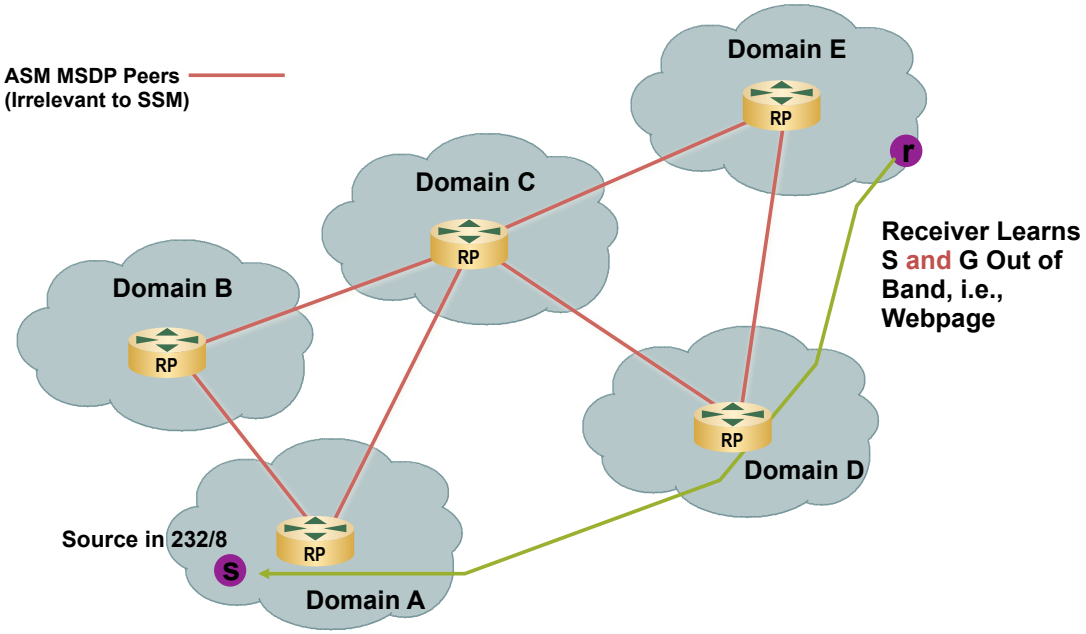


MSDP Overview

MSDP Example



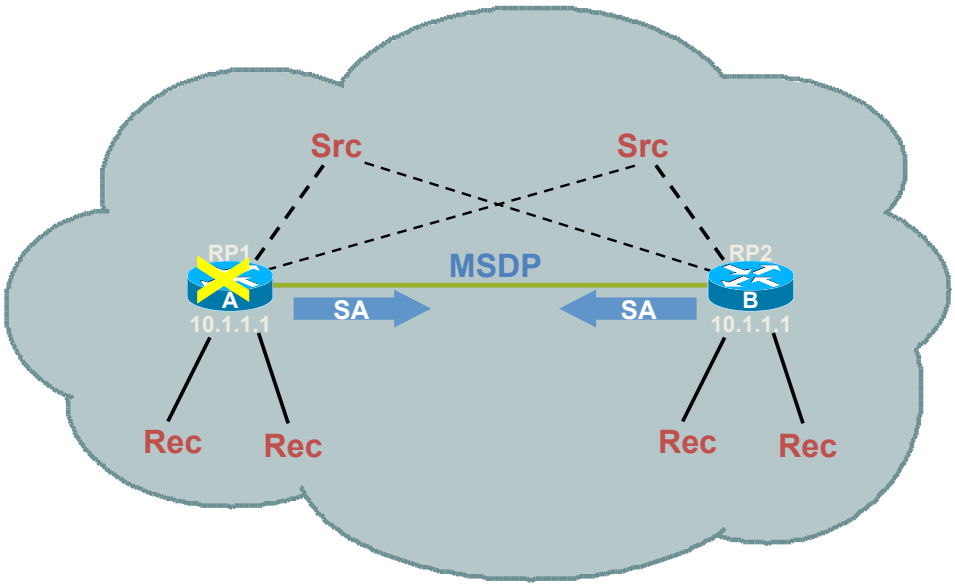
MSDP wrt SSM—Unnecessary



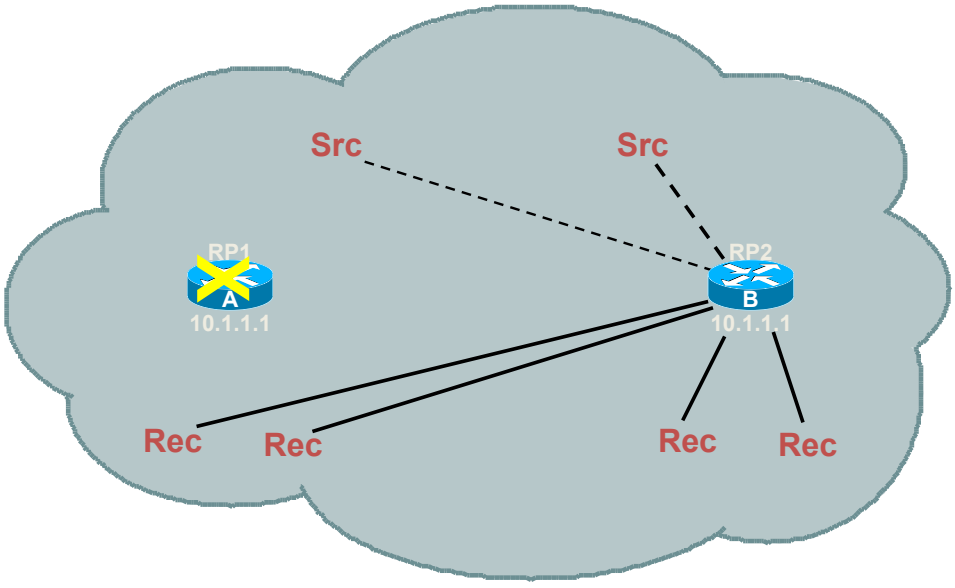
Anycast RP—Overview

- Redundant RP technique for ASM which uses MSDP for RP synchronization
- Uses single defined RP address
  - Two or more routers have same RP address
    - RP address defined as a loopback interface
    - Loopback address advertised as a host route
  - Senders and receivers join/register with closest RP
    - Closest RP determined from the unicast routing table
    - Because RP is statically defined
- MSDP session(s) run between all RPs
  - Informs RPs of sources in other parts of network
  - RPs join SPT to active sources as necessary

Anycast RP—Overview



Anycast RP—Overview



Multicast Provider Services

MVPNs

Multicast VPN (MVPN)

- Allows an ISP to provide its MPLS VPN customers the ability to transport their multicast traffic across MPLS packet-based core networks
- Requires IPmc enabled in the core
- MPLS may still be used to support unicast
- A scalable architecture solution for MPLS networks based on native multicast deployment in the core

Multicast VPN (MVPN)

- Uses draft-rosen-vpn-mcast encapsulation and signaling to build MVPN Multicast VPN (MVPN)'
  - GRE encapsulation
  - PIM inside PIM
- Not universally deployed
  - Not all VPN providers offer MVPN services

Terminology

Multicast VRF (MVRF): A VRF that supports both unicast and multicast forwarding tables

- Per VRF multicast routing and forwarding
- PIM/IGMP/MSDP and other multicast protocols operate in the context of the VRF
- RPF check using unicast routing information in the same VRF
- Special configuration not required to create or enable an MVRF

Multicast Distribution Tree (MDT): Used to carry multicast C-packets among PE routers in a common Multicast Domain (set of VRFs that can send Multicast packets to each other)

MDTs

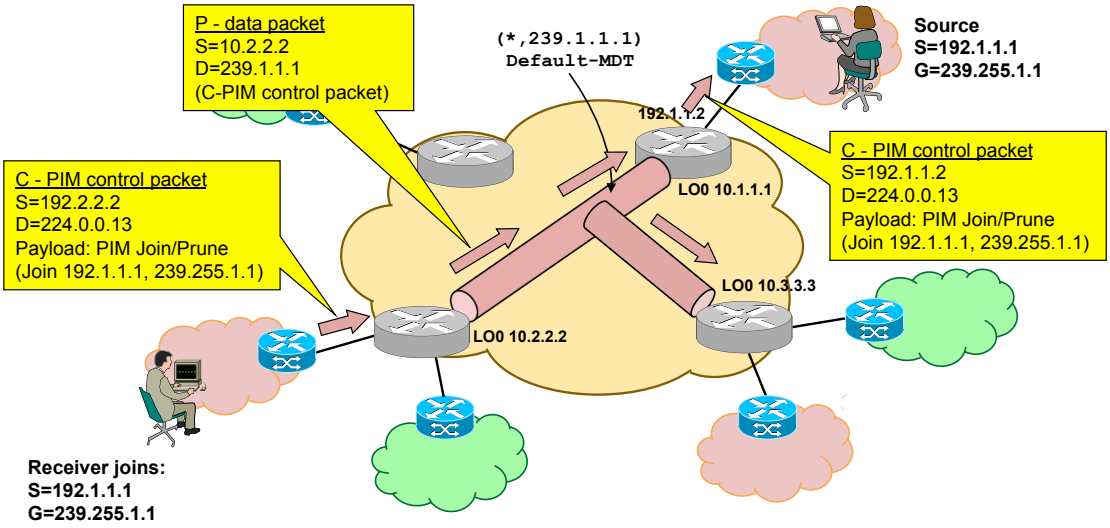
Default MDT Groups

- Configured for every MVRF if MPLS or IP core network present
- Used for PIM control traffic, low bandwidth sources, and flooding of dense-mode traffic

Data MDT Groups

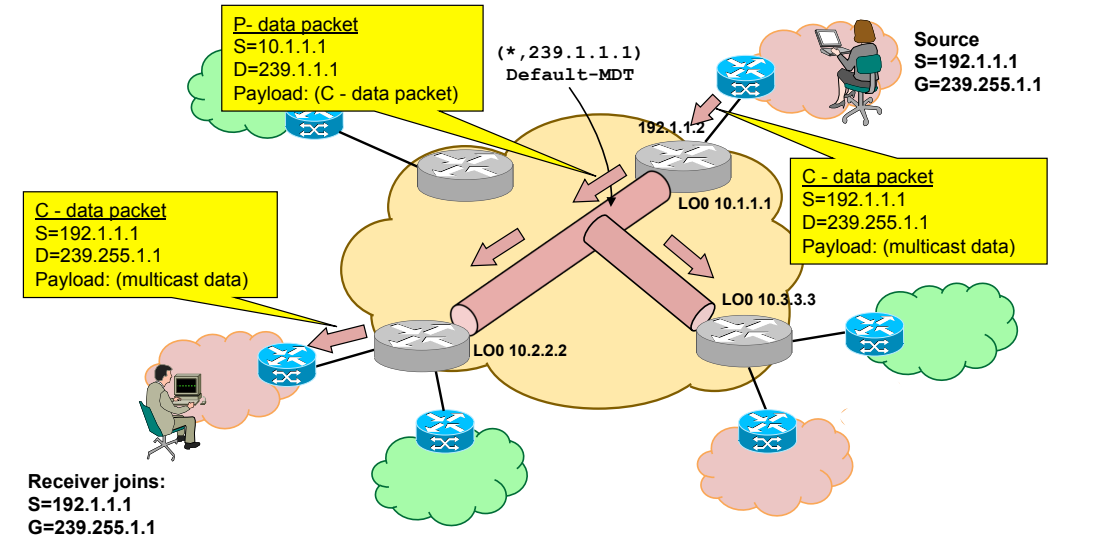
- Optionally configured
- Used for high bandwidth sources to reduce replication to uninterested PEs

Default MDT (MI-PMSI)



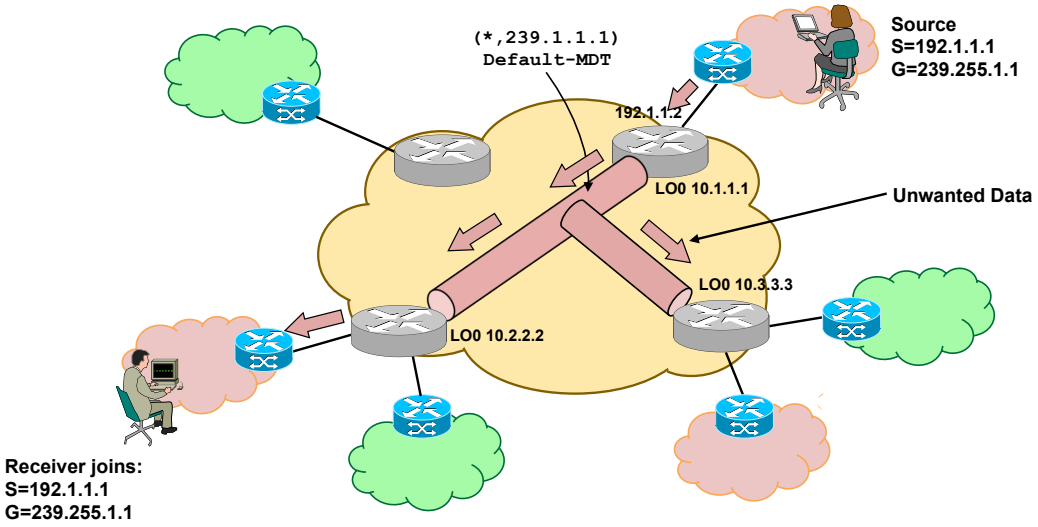
PIM Control Flow Traffic

Default MDT (MI-PMSI)



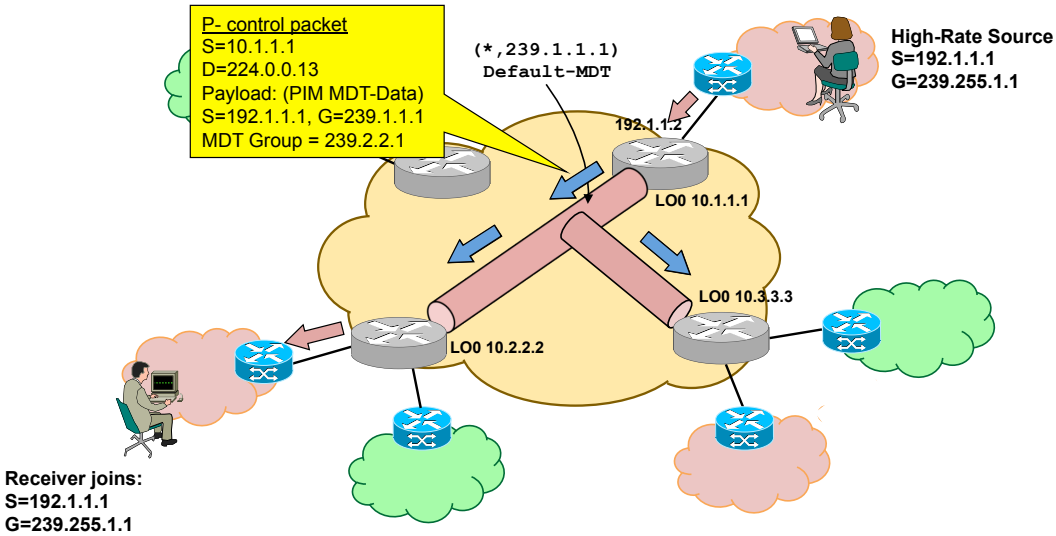
Multicast Data Traffic Flow

Default MDT (MI-PMSI)



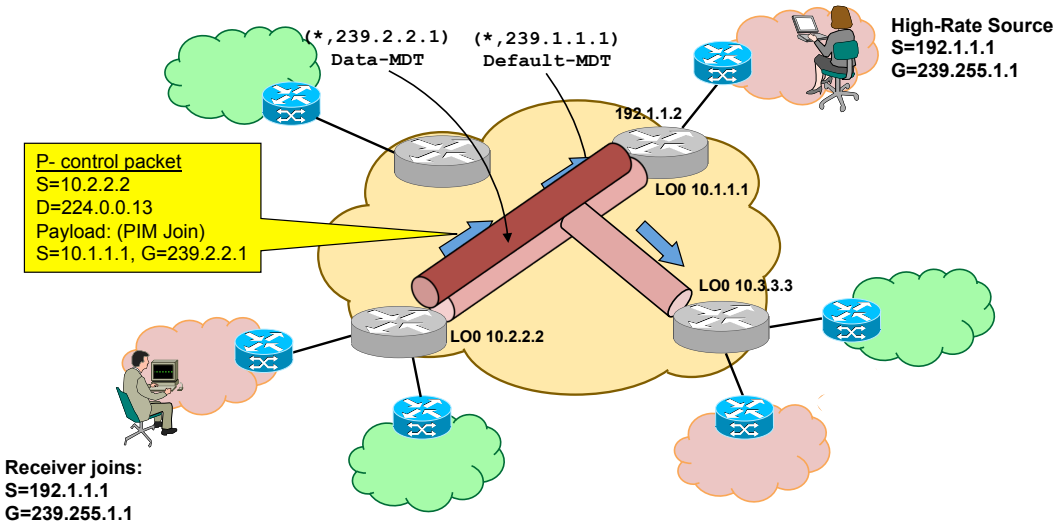
Advantage: Reduces multicast state in the P routers in the core
Disadvantage: May result in wasted bandwidth

Data MDT (S-PMSI)



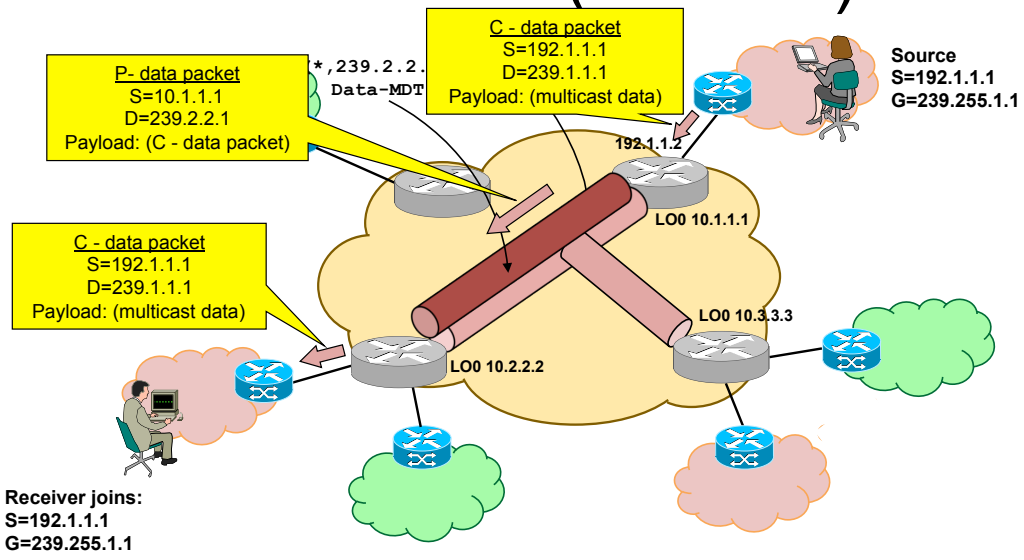
PE router signals switch to Data MDT using new group 229.2.2.1

Data MDT (S-PMSI)



Data-MDT is built using group 239.2.2.1

Data MDT (S-PMSI)



Data only goes to PE routers that have receivers

Label Switched Multicast

What is Label Switched Multicast ?

- IP multicast packets are transported using MPLS encapsulation.
- MPLS encoding for LSM documented in rfc5332.
- Unicast and Multicast share the same label space.
- MPLS protocols RSVP-TE and LDP are modified to support P2MP and MP2MP LSPs.

LSM Protocols

For BUILDING LSP's:

- Multicast LDP (MLDP)
  - Extensions to LDP
  - Support both P2MP and MP2MP LSP
  - draft-ietf-mpls-ldp-p2mp-08
- RSVP-TE P2MP
  - Extensions to unicast RSVP-TE
  - RFC4875

LSM Protocols

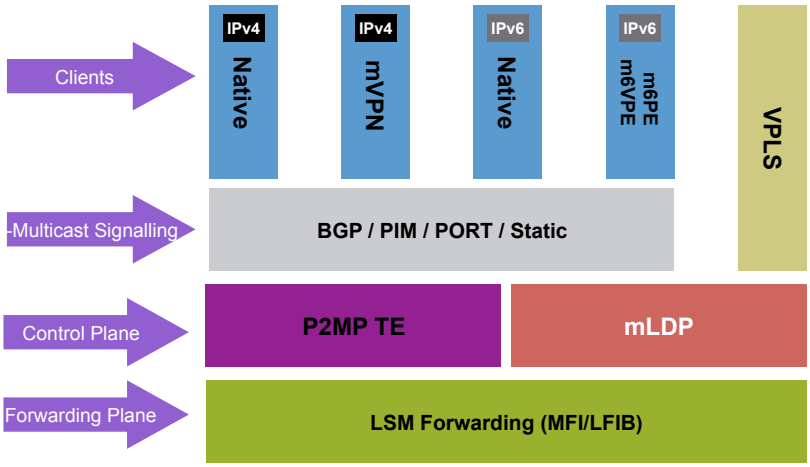
For ASSIGNING FLOWS to LSPs:

- BGP
  - draft-ietf-l3vpn-2547bis-mcast-bgp-08
  - Also describes Auto-Discovery
- PIM
  - draft-ietf-l3vpn-2547bis-mcast-10
- MLDP In-band signaling
- Static

LSM Services

- LSM architecture supports a range of services or “clients”
- Clients use combination of multicast signalling and control plane
- All LSM traffic is forwarded using MFI or LFIB mechanisms

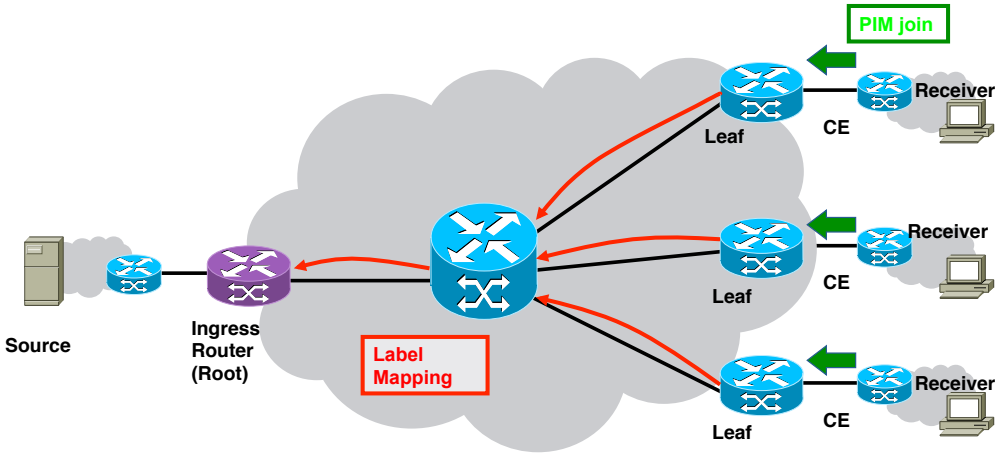
  Shares the same forwarding plane as unicast MPLS



mLDP overview

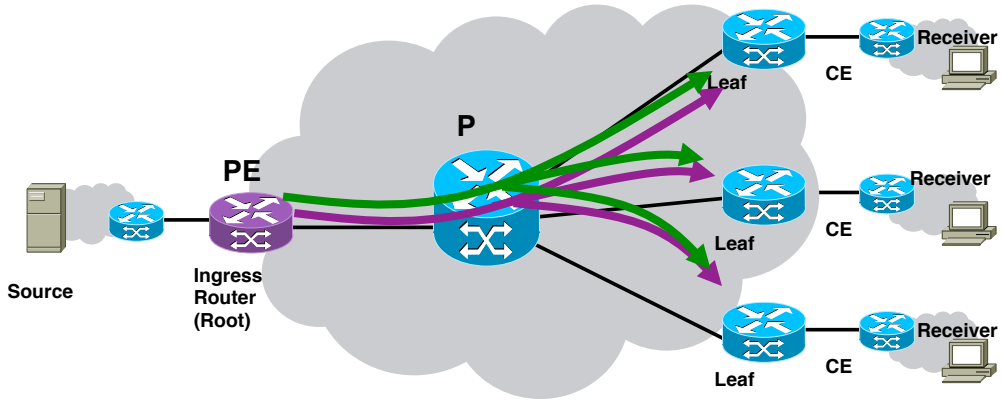
- LSPs are build from the receiver to the root, which is a scaling advantage
- Supports P2MP and MP2MP LSPs
- Supports FRR via RSVP TE unicast backup path
- No periodic signaling, reliable using TCP
- Control plane is P2MP or MP2MP
- Data plane is P2MP

mLDP example, signalling



- The egress (leafs) receives a PIM Join.
- The leaf send a mLDP label mapping to the ingress PE (via the core).
- The ingress PE received one update due to receiver driven logic.
- The core router received 3 update messages

mLDP example, state

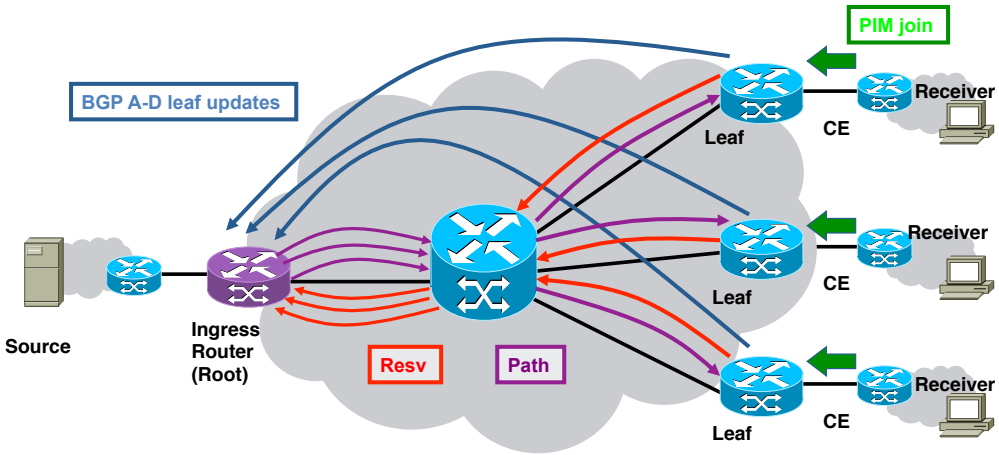


- Control Plane: **1 P2MP LSP**
- Forwarding Plane: **1 P2MP LSP**
- **P: 1 P2MP FEC (independent of the number of leaves), 4 control msg**
- **PE: 1 P2MP FEC (independent of the number of leaves), 1 control msg**
- **When a leaf wants to leave, msg is only sent to the next branch point, not all the way to ingress PE;**

RSVP-TE overview

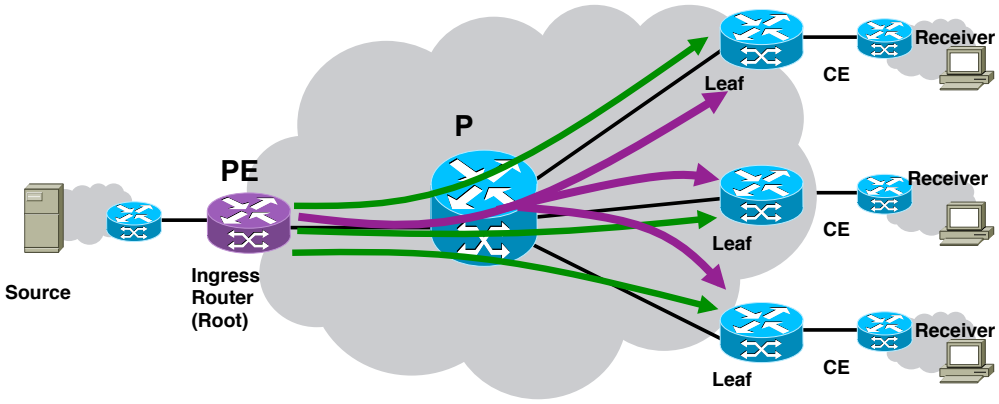
- LSPs are build from the head-end to the tail-end
- Supports only P2MP LSPs
- Supports traffic engineering
  - – Bandwidth reservation
  - – Explicit routing
  - – Fast ReRoute
- Signaling is periodic
- P2P technology at control plane
  - – Inherits P2P scaling limitations
- P2MP at the data plane
  - –Scalability advantage in forwarding plane

RSVP-TE example, signalling



- The egress (leafs) receives a PIM Join.
- The Leafs sends a BGP A-D leaf to notify the ingress PE.
- The ingress sends RSVP-TE Path messages to the leaves.
- The leaves respond with RSVP-TE Resv messages.
- The core router received 6 updates.

RSVP-TE example, state



- Control Plane: **3 P2P sub-LSPs from the ingress to the leaves**
- Data Plane: The 3 sub-LSP are merged into **one P2MP for replication**
- **P:** one state for each individual leaf, **total 3** in example**; 12 path/resv msg**
- **Ingress PE: 3 LSPs, 6 path/resv msg**
- **When a leaf want to leave, control-msg is sent all the way to ingress PE to remove the LSP;**

Applications of LSM

- IPTV / Internet multicast transport
  - draft-ietf-mpls-mldp-in-band-signaling-02
  - 1-1 mapping between IP multicast flow and LSP
  - Forwarding uses the global table (non-VPN)
- VPLS
  - draft-ietf-pwe3-p2mp-pw-00
  - Use MLDP to create Pseudowires
- Carriers Carrier service
  - draft-wijnands-mpls-mldp-csc-01
  - A provider offering services to another provider

Applications of LSM (cont)

- MVPN (Rosen Model)
  - RFC6037
  - Using MLDP MP2MP for the default MDT (MI-PMSI).
  - Using MLDP or RSVP-TE P2MP for the data MDT (MS-PMSI).
  - Same as GRE model, just the encapsulation changed.
- MVPN (Dynamic partitioned MDT)
  - draft-rosen-l3vpn-mvpn-mspmsi-05.
  - Dynamic model of above.
  - Using MLDP MP2MP for the dynamic MDT.

LSM Status

LSM Protocols	Distinct Properties
MLDP draft-ietf-mpls-ldp-p2mp-08	▪ Dynamic Tree Building suitable for broad set Multicast Applications ▪ FRR as optional capability ▪ Receiver-driven dynamic tree building approach
P2MP RSVP-TE RFC-4875	▪ Deterministic bandwidth guarantees over entire tree (calculation overhead limits this to static tree scenarios) ▪ Headend-defined trees ▪ FRR inherent in tree setup ▪ Useful for small but significant subset of Multicast Applications: Broadcast TV where bandwidth restrictions exist

LSM – decision points

- MLDP and RSVP are both useful tree building protocols for transporting multicast over MPLS.
- It depends on the application and the scalability/ feature requirements which protocol is preferred.
- Aggregation is useful to limit the number of LSPs that are created. Too much aggregation causes flooding.
- There are different options to assign multicast flows to LSP's, PIM, BGP, MLDP in-band signaling and static.
- For general purpose MVPN we recommend MLDP for tree building and PIM for assigning flows to the LSP.

Internet IP Multicast?

Internet IP Multicast

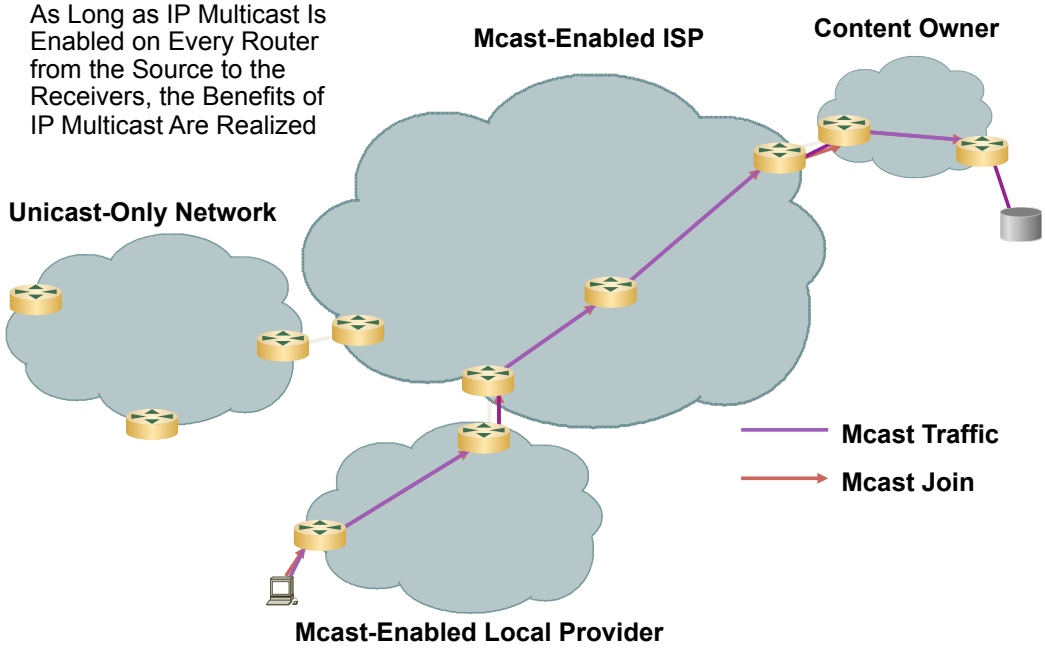
- We can build multicast distribution trees.
  - PIM
- We can RPF on interdomain sources
  - MBGP
- We no longer need (or want) network-based source discovery
  - SSM
- So interdomain IP Multicast is in every home, right?

Internet IP Multicast

- What worked?
- What didn't work?
- What's being done to fix it?

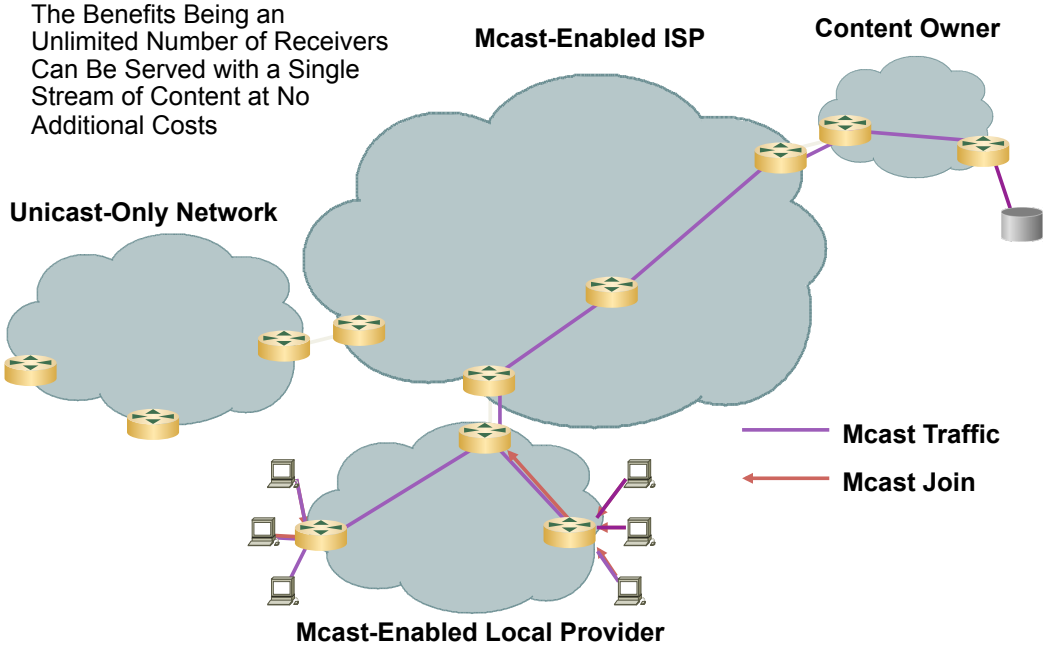
What Worked?

As Long as IP Multicast Is Enabled on Every Router from the Source to the Receivers, the Benefits of IP Multicast Are Realized

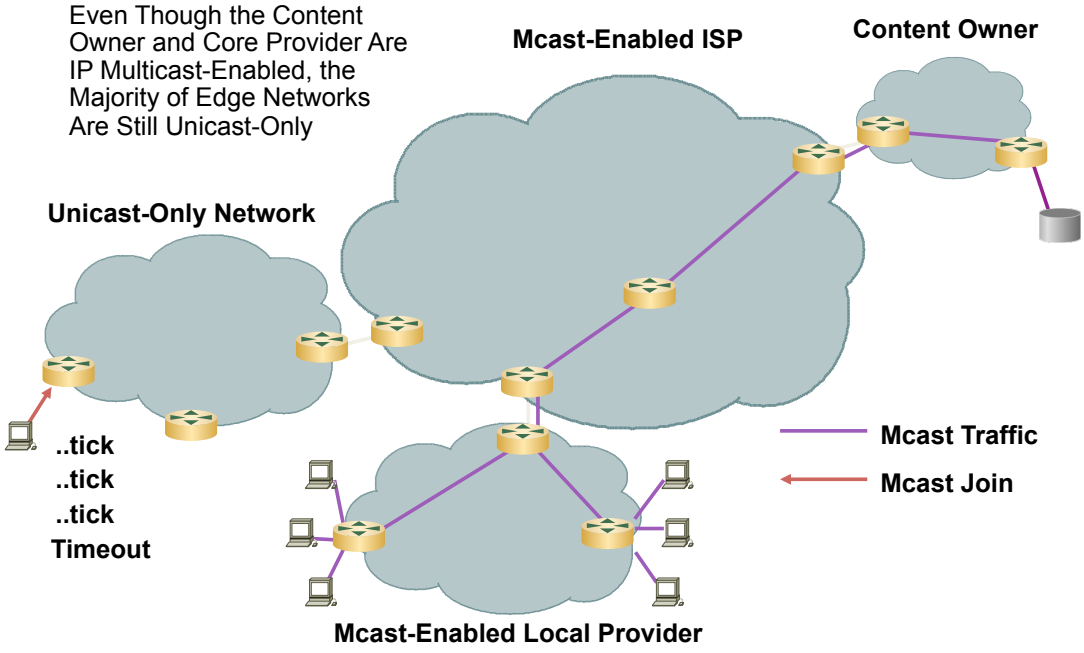


What Worked?

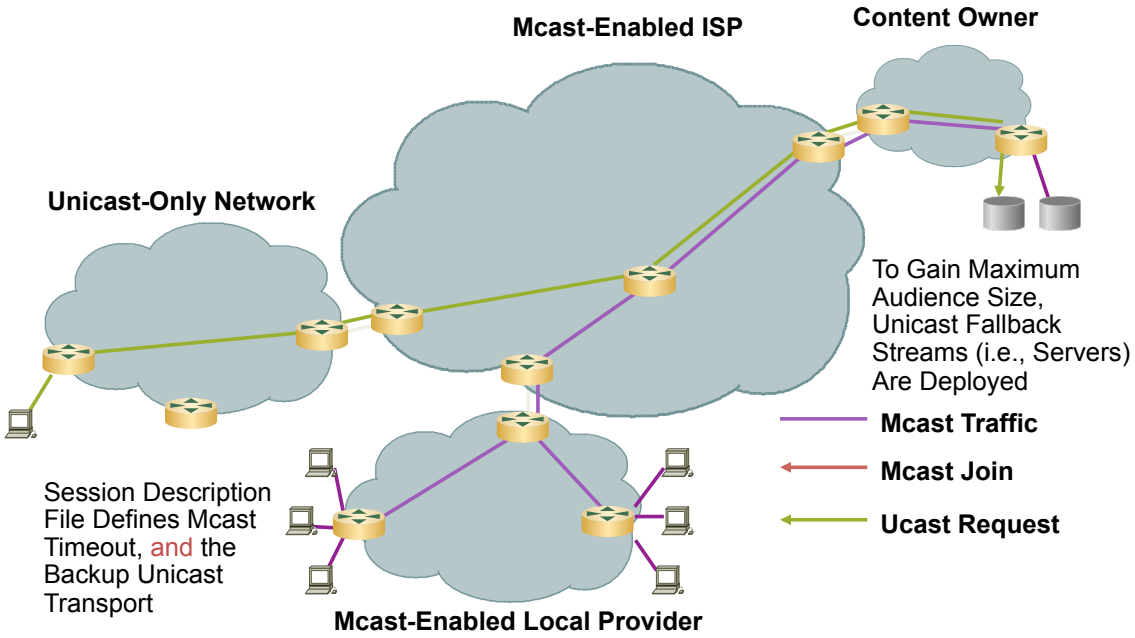
The Benefits Being an Unlimited Number of Receivers Can Be Served with a Single Stream of Content at No Additional Costs



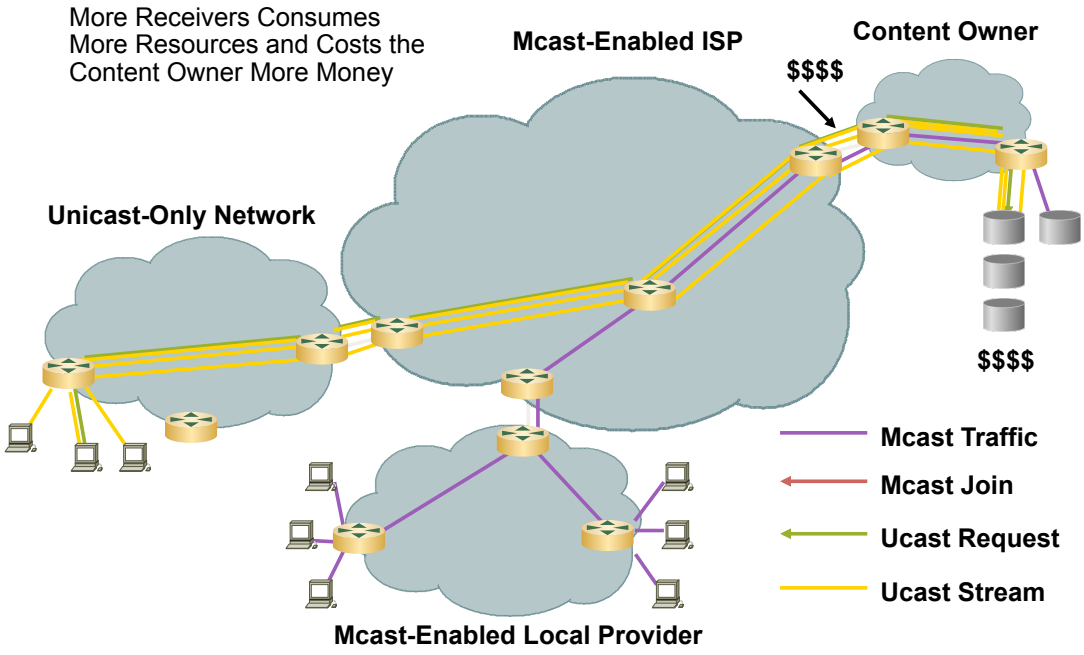
What Didn't?



What Didn't?



What Didn't?



What's Wrong?

- Without an overlay mechanism, Multicast in the Internet is an all or nothing solution
  - Each receiver must be on an IP multicast-enabled path
  - Many core networks have IP multicast-enabled, but few edge networks accept multicast transit traffic
- Even Mcast-aware content owners are forced to provide unicast streams to gain audience size
- Unicast cannot scale dynamically for live content
  - Splitters/caches just distribute the problem
    - Still has a cost per user

AMT—Automatic Multicast Tunneling

- Automatic IP multicast without explicit tunnels

  http://www.ietf.org/internet-drafts/draft-ietf-mboned-auto-multicast-X.txt

- Allow multicast content distribution to extend to unicast-only connected receivers

  Bring the flat scaling properties of multicast to the Internet

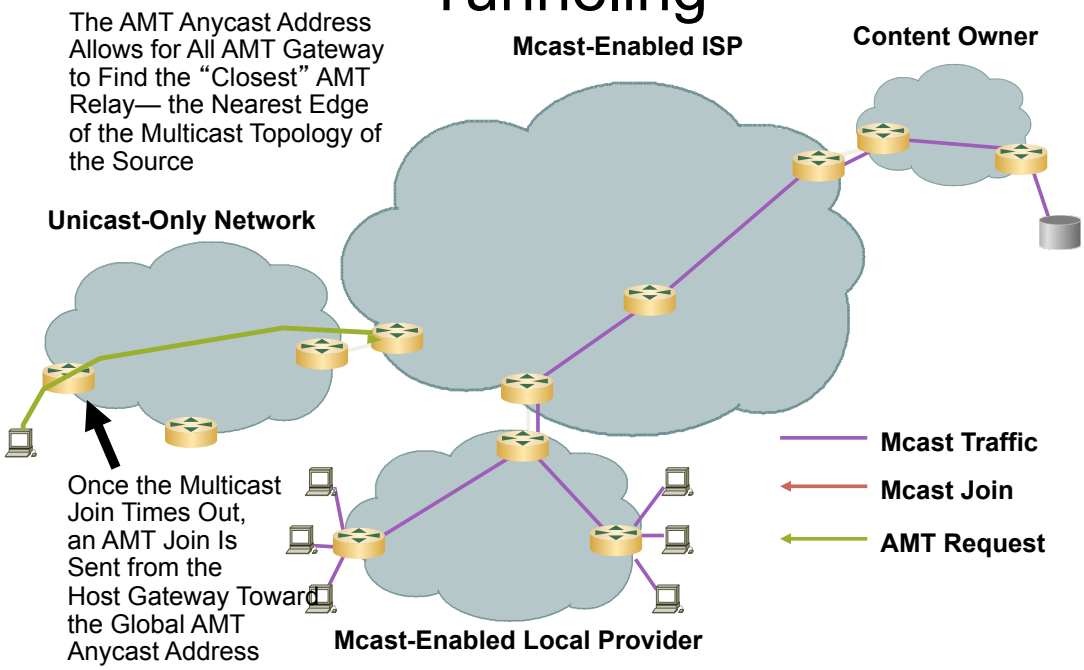
- Provide the benefits of multicast wherever multicast is deployed

  Let the networks which have deployed multicast benefit from their deployment

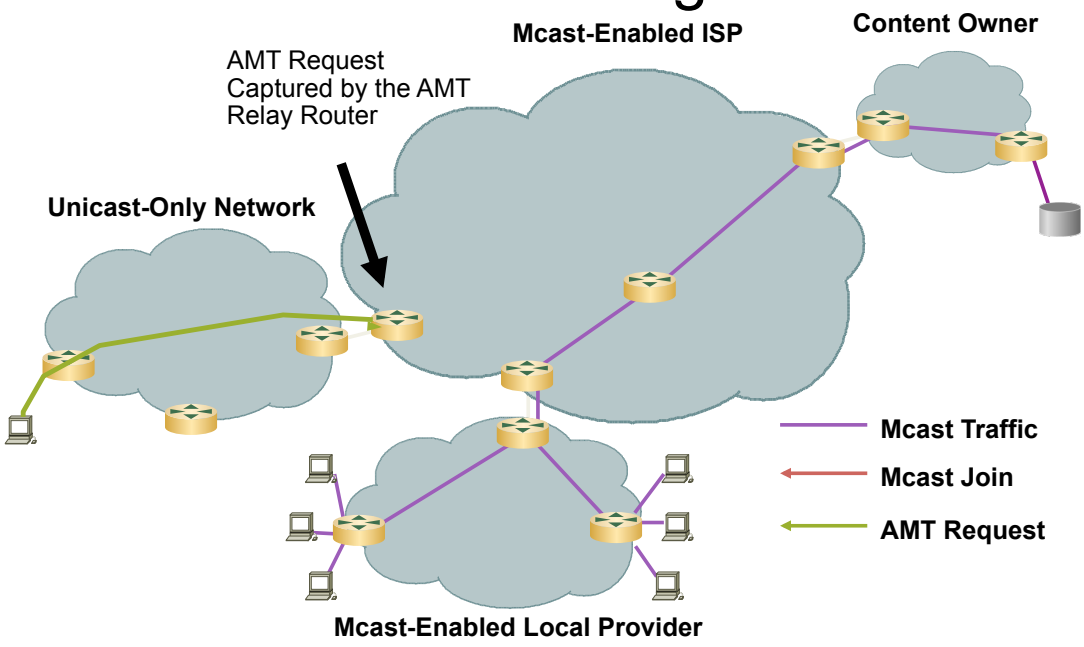
- Work seamlessly with existing applications

  No OS kernel changes

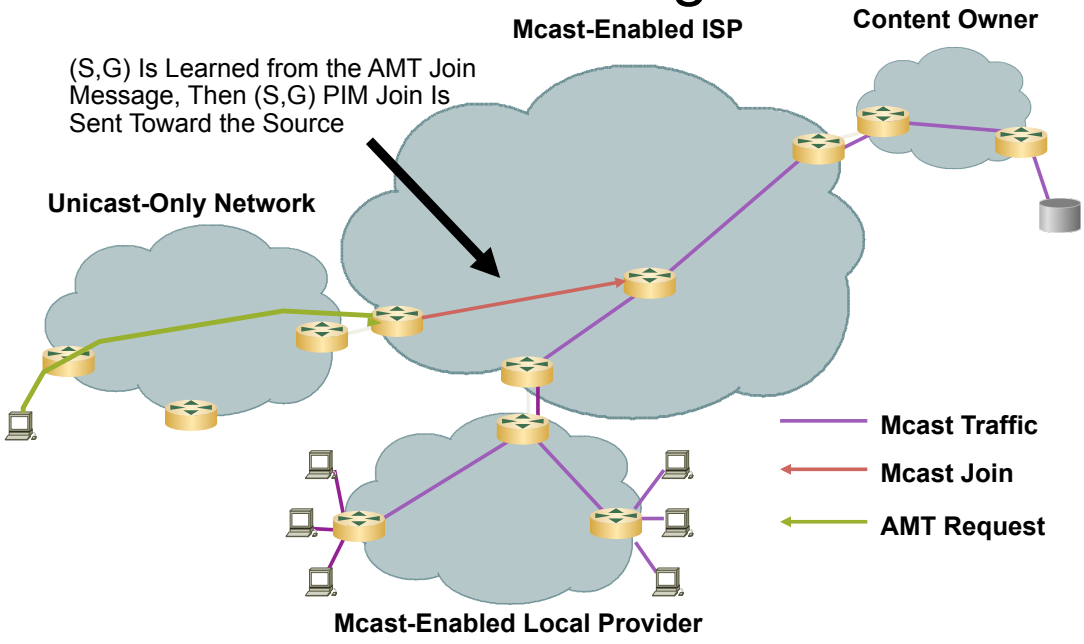
AMT—Automatic Multicast Tunneling



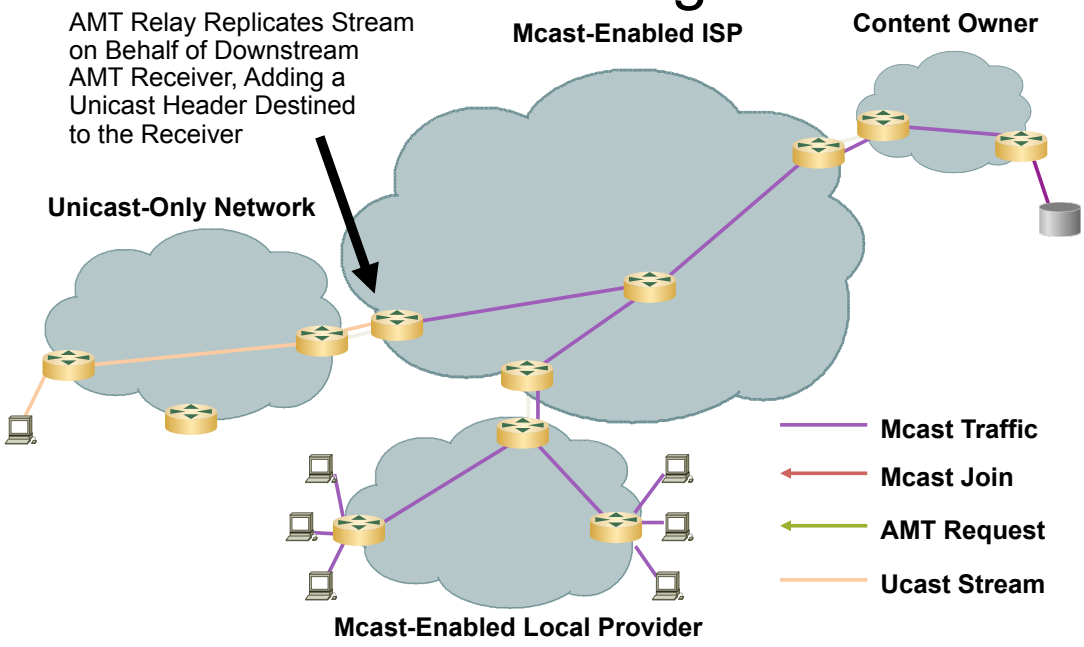
AMT—Automatic Multicast Tunneling



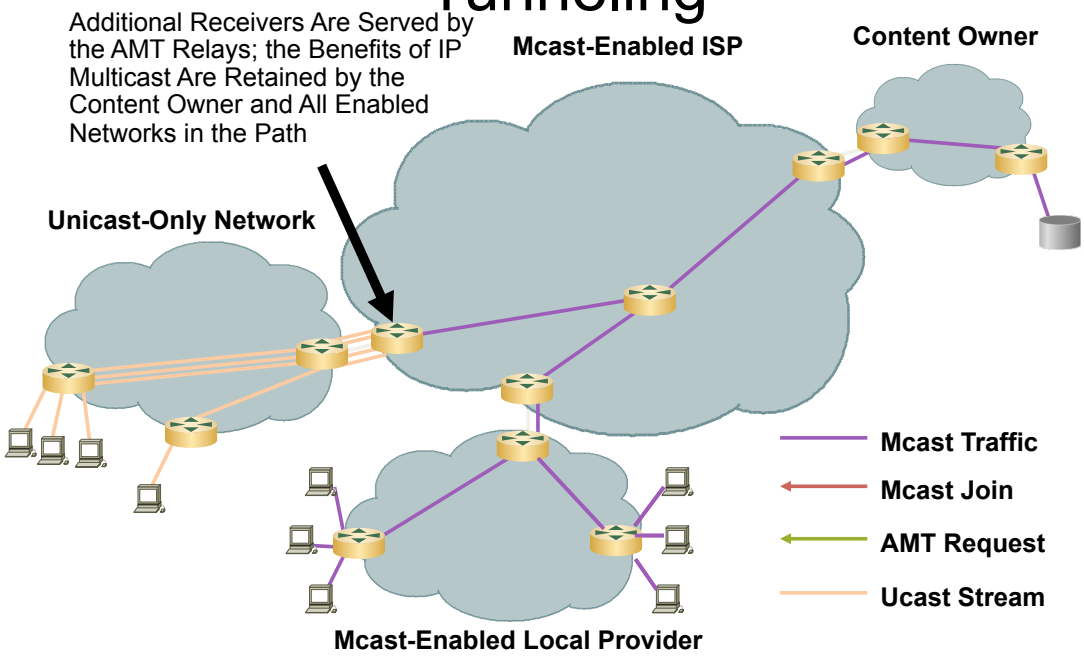
AMT—Automatic Multicast Tunneling



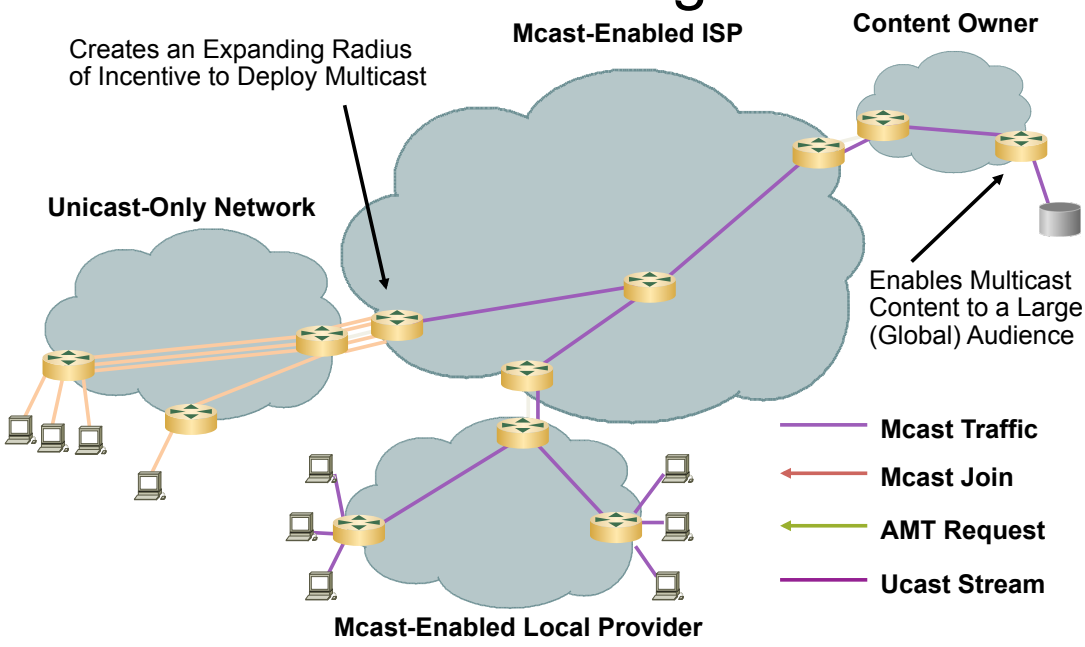
AMT—Automatic Multicast Tunneling



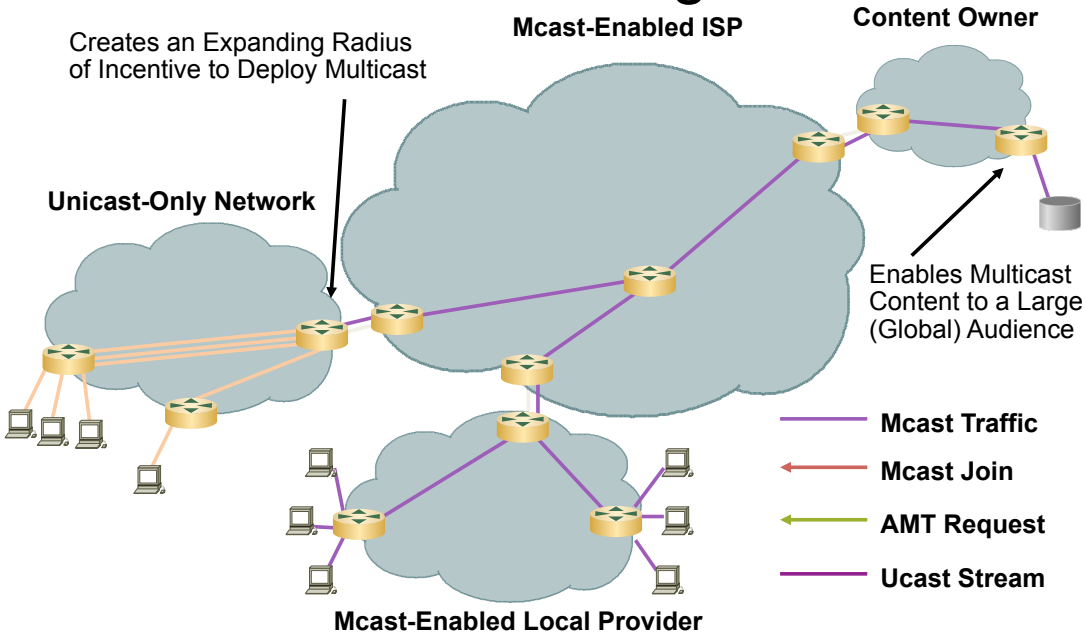
AMT—Automatic Multicast Tunneling



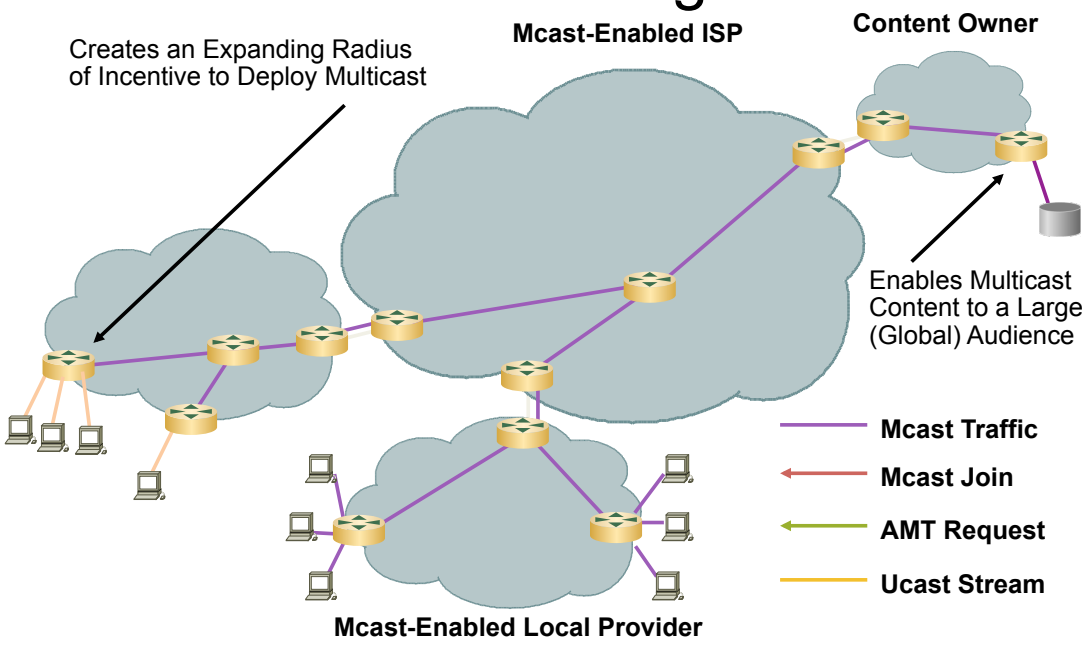
AMT—Automatic Multicast Tunneling



AMT—Automatic Multicast Tunneling



AMT—Automatic Multicast Tunneling



AT&T
AMT Multicast Trials

Outline

- AT&T Trial Activity – AMT Multicast
  - Motivation
  - Overview of Trial
  - Technical Learnings
- Opportunities for Further Discussion
  - Improved standardization & development

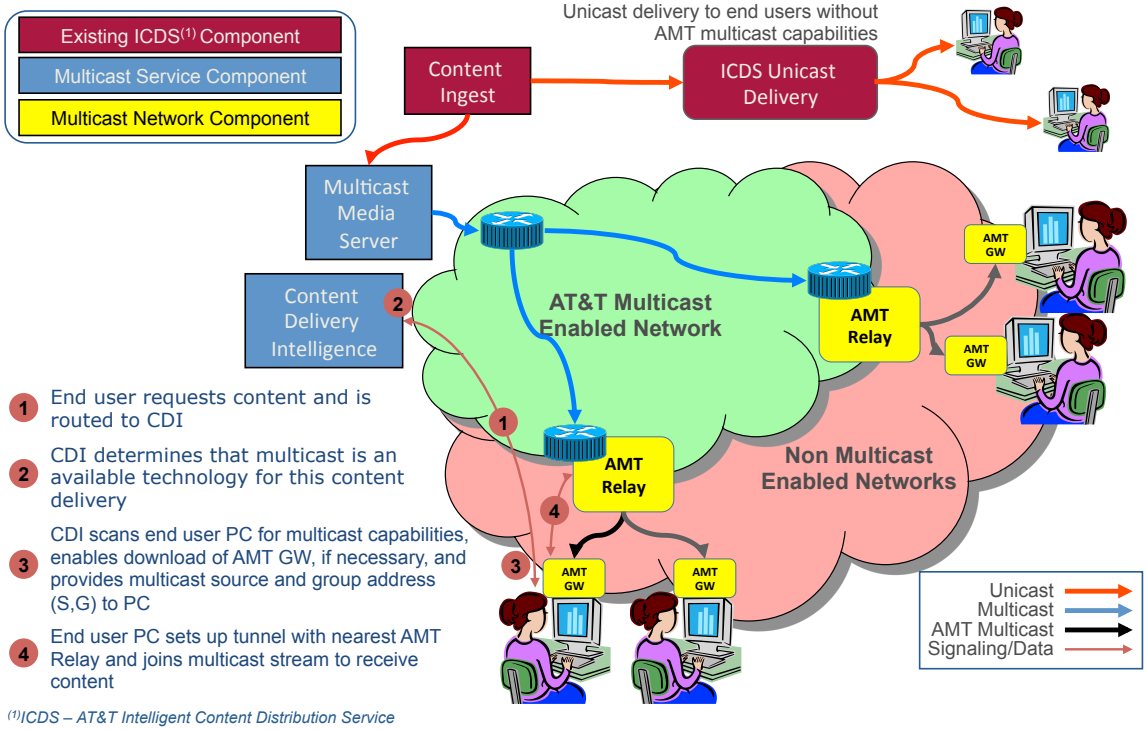
AT&T AMT Multicast – Motivation

- Multicast will play a critical role in cost-effective delivery of content for both network and content providers
- However, the Internet is not currently universally multicast capable – especially broadband access networks and home network equipment such as RGs (Residential Gateways)
- As an interim solution, use AMT (Automatic IP Multicast without Explicit Tunnels)[1] to tunnel through non-multicast-capable networks

AT&T AMT Multicast Trials
Overview

- Use PIM-SSM (Protocol Independent Multicast – Source Specific Multicast) over AT&T multicast-capable CBB (Core Backbone) network
- AT&T-developed AMT Relay and GW (Gateway)
- End-to-end Service delivery perspective *Integrated with AT&T CDN (Content Delivery Network)*
  - Content-request handling/routing logic
  - Coexistent with unicast
  - Service Logic interacting with PC Client
  - Seamless failover
  - Service Assurance
  - Accounting/Reporting
  - AAA (future)

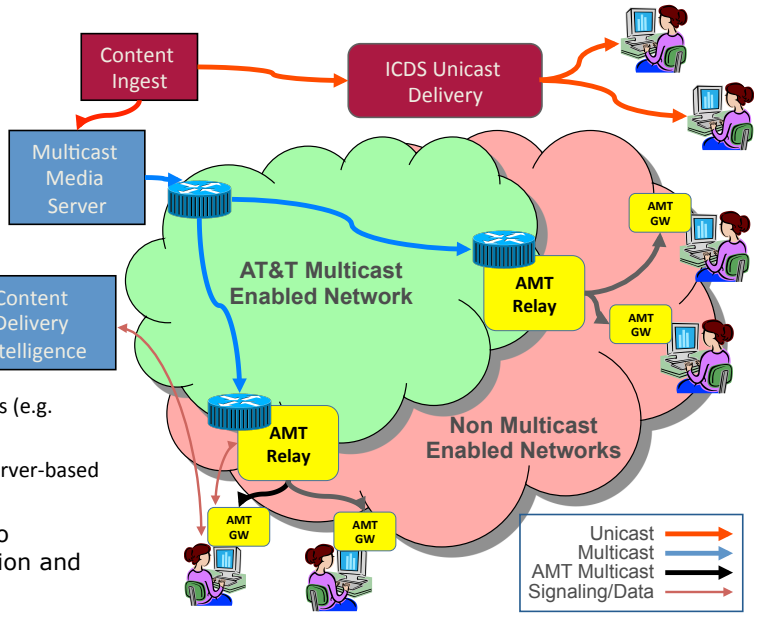
AT&T AMT Multicast Trial Architecture



(1)ICDS – AT&T Intelligent Content Distribution Service

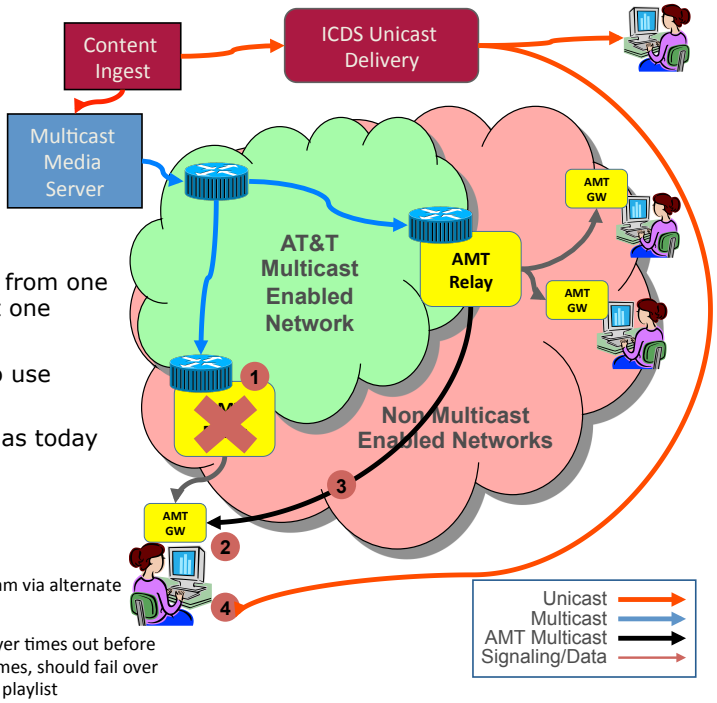
AT&T AMT Multicast Trials
What has worked well

- ✓ Quality of AMT multicast perceived as good as unicast
- ✓ Pushing GW client, as needed
- ✓ Pushing appropriate stream to user
- ✓ Anycast Routing for AMT Relays
- ✓ Accounting information uploaded from PC
- ✓ Handling home network environments (e.g. multiple PCs behind NAT, WiFi)
- ✓ Performance & scalability of Relay (Server-based for Trial)
- ✓ Tuning MTU on Media Server to accommodate AMT encapsulation and avoid fragmentation



AT&T AMT Multicast Trials
Failure Recovery Model

- ✓ AMT Relay advertises anycast route for Discovery. Once GW discovers Relay, uses its unicast address.
- ✓ AMT GWs can detect Relay failure and "Rediscover" new Relay
- ✓ Media Player should allow failover from one stream to another on playlist (last one unicast)
- ✓ Multicast media servers could also use anycast for their source address
- ✓ Unicast media server redundancy as today



Example:

1. AMT Relay fails mid-stream
2. AMT GW detects failure and Rediscovers alternate Relay
3. Rejoins stream via alternate Relay
4. If Media Player times out before stream resumes, should fail over to unicast in playlist

Now You Know…

- Why multicast?
- Multicast fundamentals
- PIM protocols
- RP choices
- Multicast at Layer 2
- Interdomain IP multicast
- Provider Services
- Enough to be dangerous

Q & A