Deploying Inter-Domain Multicast Routing

dino@cisco.com

Nanog 15

02/01/99

Agenda

- ISP Deployment Requirements
  - MBGP
- What have been the obstacles thus far
- Can we use existing technology
  - PIM-SM
  - MSDP
- Future Technology
  - BGMP
- References/Questions

ISP Requirements

- ISPs are telling us that the unicast and multicast topologies won’t be congruent across domains
  - Due to physical/topological constraints
  - Due to policy constraints
- We need a inter-domain routing protocol that distinguishes unicast versus multicast policy

ISP Requirements

- A multicast delivery tree must only be dependent on the reachability of the source
  - Just like a unicast delivery path is dependent on the destination
- Using shared-trees will not allow one ISP to be autonomous from another ISP
  - Where to put the RP?
  - Coined the phrase “3rd party RP problem”
- Using shared-trees will not allow per source prefix policy

ISP Requirements

- Cannot use flood-and-prune technology
  - State everywhere is a non-starter
- Most important requirement is:
  - Simplicity
  - Multiple multicast routing protocols doesn't necessarily satisfy this requirement

Topologies are not Congruent

- Simply use a different routing table to find multicast sources
  - Unicast forwarding uses the unicast routing table (built by IGPs and BGP)
  - Inter-domain multicast forwarding uses the MRIB (built by MBGP)
  - PIM uses MRIB for RPF lookups
- You get all the policy tweaking features of BGP
- Required for the MIX topologies

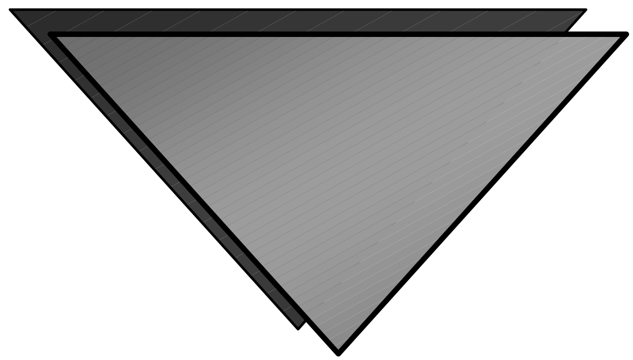
What have been the obstacles so far

- How to find group members across a large internet
- How to avoid the 3rd party RP problem if using shared-trees
- How to do scalable group-to-RP mappings if using shared-trees
- If not using shared-trees, how to find sources

How to find group members

Flood and Prune
DVMRP
PIM-DM

Broadcast Membership
MOSPF
DWRs



Rendezvous Mechanism
PIM-SM
BGMP

Can we use existing Technology

- Why not use sparse-mode PIM?
  - Where to put the root of the shared tree (the RP)
  - ISP 3rd-party RP problem
- If you did use sparse-mode PIM
  - Group-to-RP mappings would have to be distributed throughout the Internet

Avoid Shared-Trees across Domains

- If you can’t connect shared trees together easily, then don’t
- Multicast Source Discovery Protocol (MSDP)
  - Different paradigm
  - Rather than getting trees connected, get sources known to all trees

MSDP - Basic Idea

- MSDP advertises multicast sources to other routing domains
- Other domains decide if group members are active and find a way to get the data
- PIM-SM is the obvious way to get data
- “MSDP connects shared-trees together”

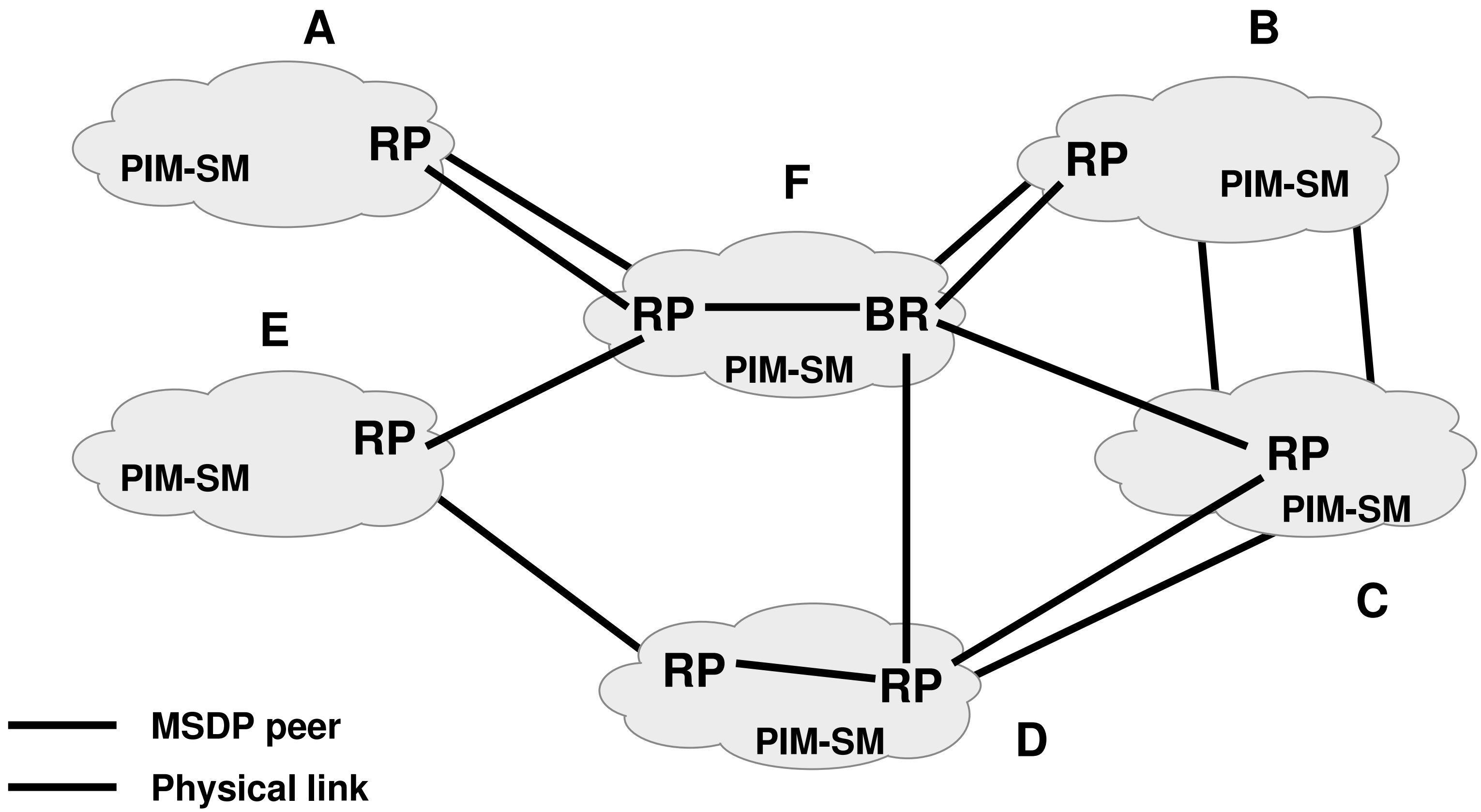
MSDP - Basic Architectural Model

- Use intra-domain PIM-SM shared-trees for rendezvousing local sources and local receivers
- Use inter-domain PIM-SM source-trees for receiving data from external sources for local receivers

MSDP - Protocol Details

- MSDP typically runs in an RP
- MSDP can run in any router that wants to aggregate MSDP sessions
- MSDP atypically can run in a dense-mode border router

Connecting Domains Together with MSDP



MSDP - Elements of Operation

- Receivers in a domain join the shared-tree
- The RP is known only to routers in the domain
- When a source goes active in a domain, it's packets get to the RP in that domain
- The RP sends a Source-Active (SA) message identifying the source and group it sends to

MSDP - Elements of Operation (continued)

- The RP's address is also in the SA message to accommodate "peer-RPF" like flooding
- Each MSDP peer receives SA message and forwards away from the originating RP
- Use MBGP or BGP to determine what peer is "away from the originating RP"

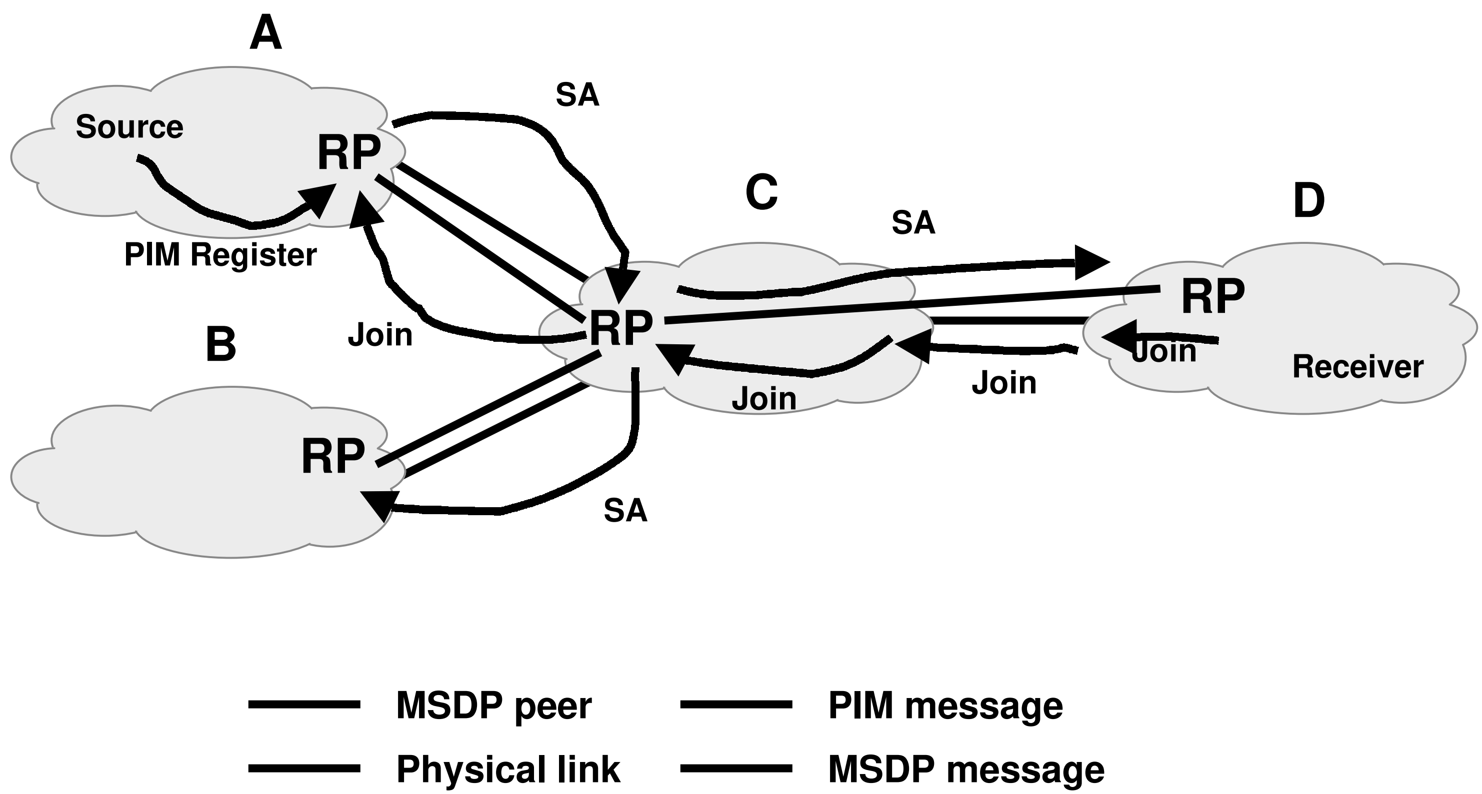
MSDP - Elements of Operation (continued)

- Each MSDP speaking RP will examine SA message to see if any local members are joined to the group
- If so, the RP joins to source described in SA message
- Otherwise, the SA message is ignored (Flood-and-Join model)

MSDP - More Details How it Works

- Source goes active in PIM-SM domain
- It's packets get PIM registered to domain's RP
- RP sends SA message to it's MSDP peers
- Other MSDP peers forward to their peers using the AS-path associated with the originating RP's prefix
- If an RP in another domain has receivers, it sends a (S,G) Join using an inter-domain route to source S

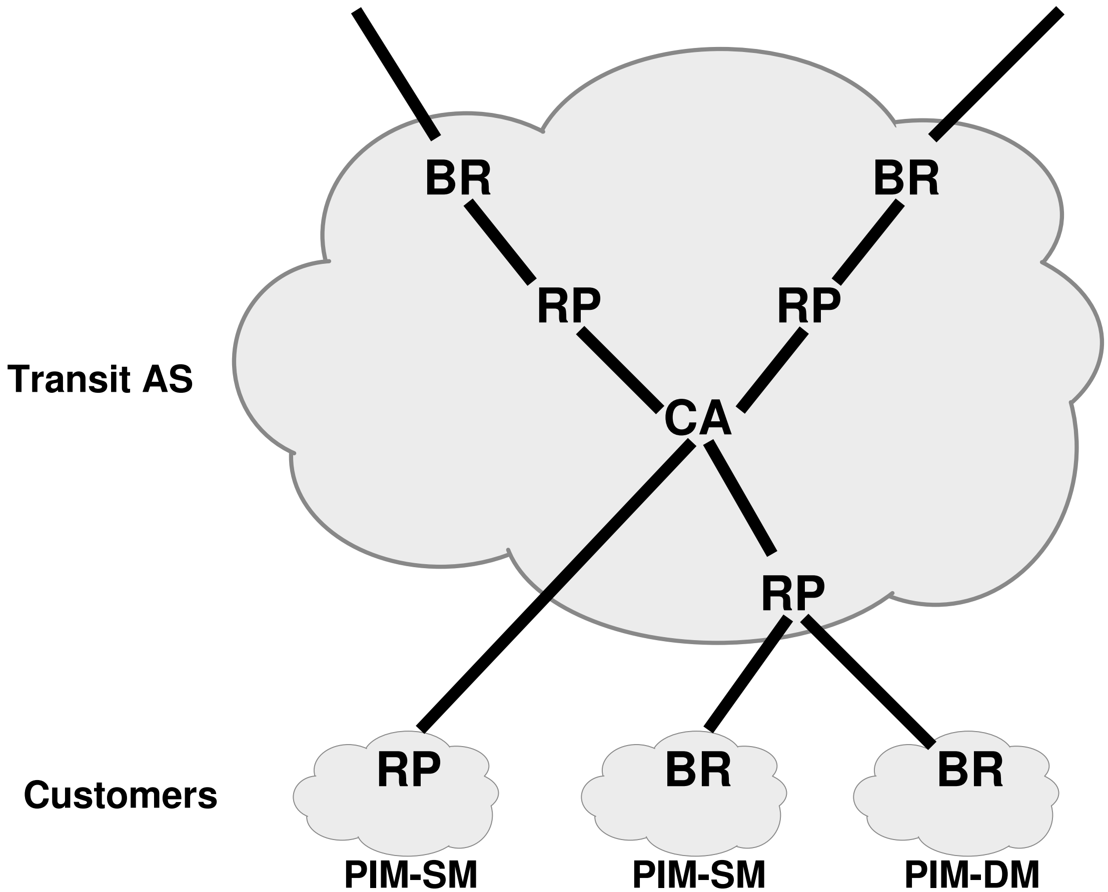
How MSDP works with PIM-SM



MSDP - Observations

- State is not created from SA messages
- Only forwarding state is maintained
- Periodic SA transmission required to inform distant RPs if source is still active (not for lost messages)
- TCP is used for MSDP sessions so there is no message loss

A Possible ISP/Customer Scenario



MSDP - Observations

- What if a receiver joins right after an SA message is received by the local RP? Join Latency?
  - MSDP SA caching peers can be configured
  - Local RPs can send SA-Requests and get immediate response back for all active sources for a given group just joined

MSDP - Observations

- What about bursty sources?
  - When the RP gets data from local sources
  - Does it queue the packet or drop it while it waits for joiners?
- SA messages encapsulate data for the first `n` packets of a data flow (when an (S,G) entry is created in the local RP)

MSDP - Conclusions

- There is no shared tree across domains
  - No 3rd-party RP problem
- Group-to-RP mappings not distributed across the Internet

Avoid Shared-Trees all together

- With PIM, shared-trees are used to bootstrap source-trees
- If hosts could tell routers about (S,G) pairs to join, PIM routers could build source-trees without the necessity for shared-trees
  - But IGMPv3 is not in product form
  - No directory mechanism is really available
- Could use DNS?
  - But need DNS with Dynamic Updates

Future Technology

- Does sending control messages for active multicast sources scale?
  - In the short-term, yes
  - In the long-term, depends on the success of multicast routing in the Internet
- Long-term plan is to migrate MSDP to BGMP when the problem approaches (the Internet way)

Border Gateway Multicast Protocol (BGMP)

- Designed specifically for inter-domain multicast routing
- Like BGP, runs over TCP and sends Join/Prune messages on MBR boundaries
- MBRs get connected together via the MIGP deployed in the domain
- BGMP builds bi-directional shared tree of domains per group

Border Gateway Multicast Protocol (BGMP)

- But didn't you say that shared-trees across domains are problematic?
  – The root of the shared-tree of domains is the "root domain"
  – Not a router in the middle of some ISP's topology
  – Typically a customer will be joined to the group or sourcing to the group
- BGMP has option to build source-trees
  – For MIGPs that support source-trees only

Border Gateway Multicast Protocol (BGMP)

- But didn't we have the same problem with PIM, we now have "group-to-root domain" mappings?
  - Yes, they do have to be distributed across the Internet
  - BGMP/MASC designers believe the information can be aggregated
  - BGMP is currently dependent on these group-to-root domain mappings - more specifically *domain group allocation algorithms*

Border Gateway Multicast Protocol (BGMP) - Observations

- Uses rendezvous mechanism so we don’t have to advertise group membership or active sources
- Supports bursty sources without knowing ahead of time the burstiness of the source
- We need more deployment experience how bi-directional shared-trees can actually work
  - Are they robust during topology changes

BGMP - Bottom Line

- Work in progress
- We need your feedback!

References/Questions

- PIM-SM version 2
  - **`RFC 2362`**
- MBGP (Multiprotocol Extensions for BGP-4)
  - **`RFC 2283`**
- MSDP
  - **`ID draft-farinacci-msdp-00.txt`**
- BGMP
  - **`ID draft-ietf-idmr-gum-04.txt`**