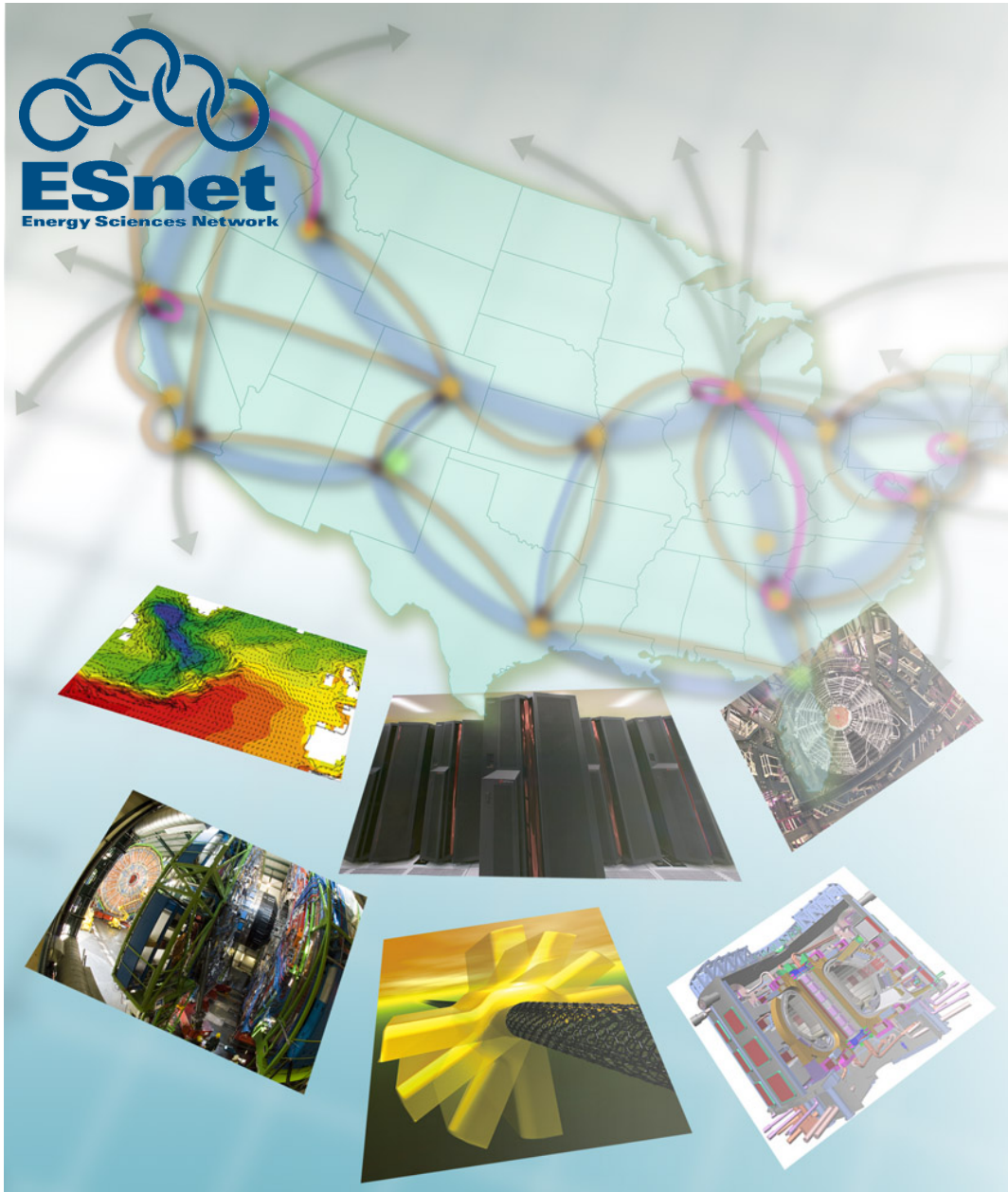


Inter-domain ASM Multicast Networking

Michael P. O'Connor
moc@es.net

Energy Sciences Network
Lawrence Berkeley National Laboratory

August 13, 2007

Networking for the Future of Science

BERKELEY LAB

Office of Science
U.S. DEPARTMENT OF ENERGY

Introduction

Multicast is a network application. Unlike other distribution methods, multicast communications affect the network routing state in the routers they pass through. This state manipulation must work flawlessly not only in your network but through your Internet Service Provider and all the way to the endpoint you're communicating with.

Many reliable multicast protocol implementations exist. All the major routing equipment manufacturers support them.

Any source multicast (ASM) is required to support the many to many conferencing model required by access grid conferencing. This talk focuses exclusively on the ASM multicast model.

Types of Data Delivery

- **Unicast:** Data is delivered to one specific recipient. One-to-one delivery.
- **Broadcast:** Data is delivered to all hosts. One to all delivery.
- **Multicast:** Data is delivered to all hosts that have expressed interest. One-to-many delivery.
- **Anycast:** Data is delivered to the nearest host of those sharing a single address. One to any delivery.

What is multicast?

It’s a network application:

Multicast distribution provides an **efficient** method for delivering traffic that can be characterized as “one-to-many” or “many-to-many”.

Multicast enabled networks are responsible for replicating data and delivering it only to listeners who have tuned in to the session.

Routers in the network build a distribution tree where the sender is the root and each network with at least one interested listener is a leaf. When a new listener tunes in, the network must build a branch from the new leaf toward the root. When a leaf no longer contains listeners, the branch must be pruned. When there are no longer any senders, the distribution tree must be torn down.

The local network support staff is almost always the only group with the knowledge and access privileges required to configure a multicast enabled network.

Addressing

Multicast group addresses are defined in the IPv4 “class D” address range 224.0.0.0 to 239.255.255.255 or using prefix notation 224.0.0.0/4.

Multicast sources transmit packets with a multicast group destination address. The source address is set to the unicast address of the sender.

Source addresses are Unicast
Group addresses are from the Class D multicast range

(S,G) notation is used to define routing state for a particular Source Group pair in a network router.

Special Addresses

A few brief examples:

224.0.0.0/24 Link local multicast addresses
224.2.0.0/16 Session Announcement Protocol (SAP)
232.0.0.0/8 Source Specific Multicast range
233.0.0.0/8 GLOP space
239.0.0.0/8 administratively scoped multicast range

For detailed description of reserved multicast group space:
http://www.iana.org/assignments/multicast-addresses

GLOP space

0 - 7	8 - 23	24 - 31
233	16 bit AS	Local bits

If you have an AS number you have a /24 in GLOP space. You should use your GLOP space for AG virtual venues at your site.

Example:

AG Test room 233.2.171.39 is in the Argonne National Lab GLOP space.

AS 683 = 2 * 256 + 171

GLOP calculator

http://www.shepfarm.com/multicast/glop.html

GLOP is not an acronym or abbreviation; for some odd reason it was selected as the name for this clever mechanism.

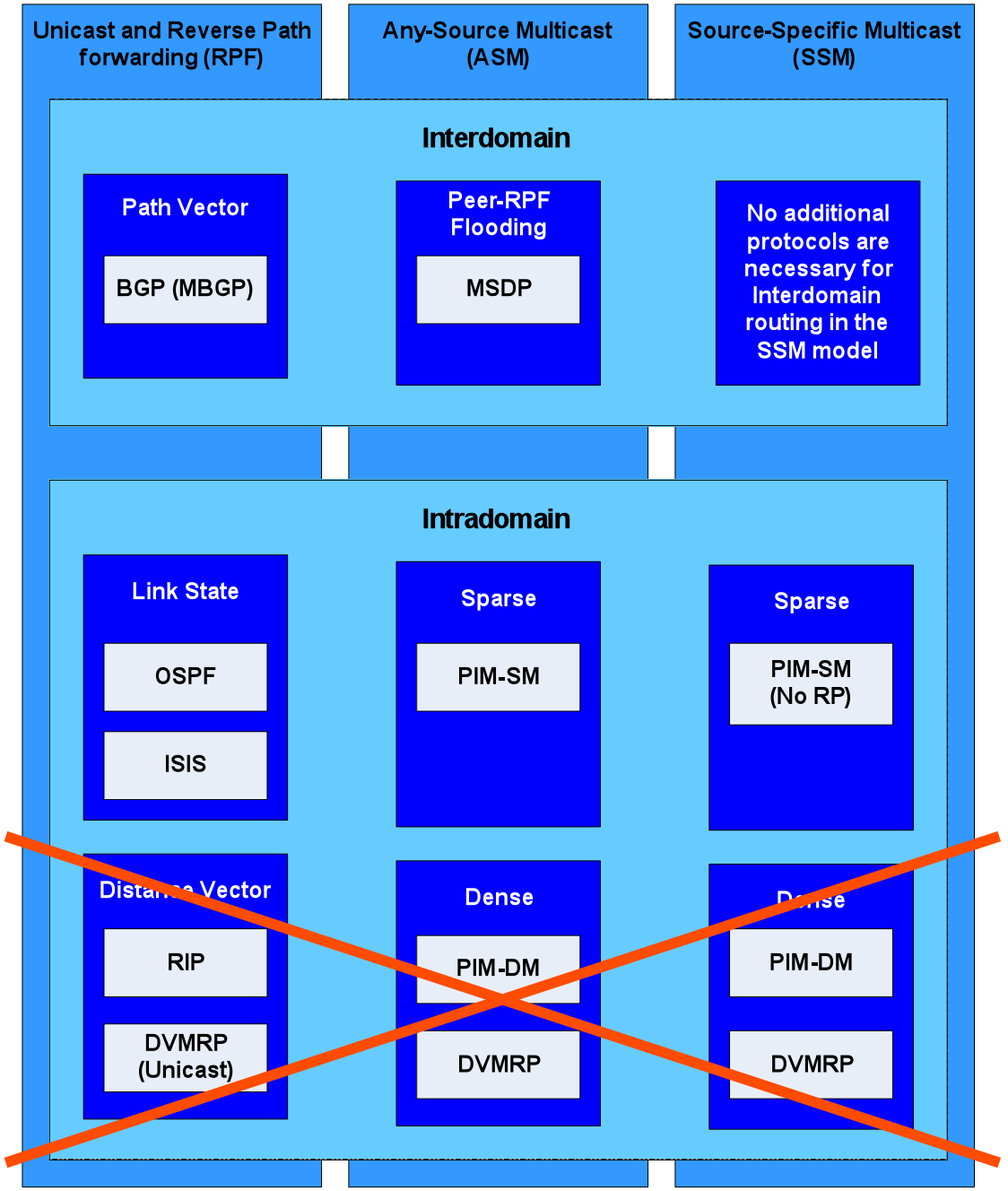
Site to ESnet Multicast Interconnect; Best and Current Practice

ESnet recommends that multicast enabled Sites/Customers implement the following external multicast protocols to exchange multicast traffic with ESnet.

- **PIM V2 – Protocol Independent Multicast Sparse Mode**
  - PIM performs a Reverse Path Forwarding (RPF) check function based on information from various routing protocols as well as static routes, giving it protocol independence.
- **MSDP – Multicast Source Discovery Protocol**
  - MSDP describes a mechanism to connect multiple PIM-SM domains together. Each PIM-SM domain uses its own independent RP(s) and does not have to depend on RPs in other domains.
- **MBGP - Multiprotocol Border Gateway Protocol**
  - Is an extension to BGP that enables BGP to carry routing information for multiple network layers and address families.

Multicast enabled network architectures that depend on PIM RPs in external domains are not recommended by ESnet. MSDP enables inter-domain RP to RP communication.

Hierarchy of Internet Routing Protocols



Protocol Independent Multicast (PIM) Sparse Mode

- PIM-SM is the predominant multicast routing protocol for inter-domain routing.
- PIM uses Reverse Path Forwarding packet distribution.
- A series of directly connected or tunneled PIM-SM peers form a path (distribution tree) between a source and destination.
- All routers in a domain must agree on the active RP (Rendezvous Point) for each multicast group.

Anycast RP

- In PIM-SM, only one RP can be active for any single multicast group.
- Anycast RP is a clever mechanism that delivers load balancing and redundancy.

- An Anycast address is one that is shared across multiple hosts, in this case routers. Packets destined for this address are delivered only to the closest host with that address.

- PIM RP redundancy can be achieved in this way, all Anycast RP's also need to be MSDP peers with each other, usually in a meshed topology.

- Do not use an Anycast address on the primary loopback interface, this will break other routing protocols.

PIM-SM Site Guidelines

- A single PIM Rendezvous Point (RP) for all multicast groups.
- Static RP – statically define the address of the RP in all PIM speaking routers.
- Auto-RP and Bootstrap Router (BSR) are not recommended.
- Use only PIM Version 2 in “Sparse” mode.
- Configure the RP on a Loopback interface to simplify moves.
- “Dense” and “Sparse Dense” modes should not be necessary and are not recommended.
- The RP network MBGP advertisement must be /24 or greater.

Multicast Source Discovery Protocol (MSDP)

- MSDP enables inter-domain source discovery without flooding.
- MSDP forms peer relationships, similar to BGP peers, over a TCP connection.
- Two MSDP peers can be in the same or separate PIM-SM domains.
- MSDP peers are not required to be directly connected neighbors.

MSDP connects multiple PIM-SM domains in different Autonomous Systems.

MSDP Site Guidelines

- Your MSDP speaking router MUST be a PIM-SM RP.
- One RP per customer site is generally recommended.
- Placement of the RP is not critical, it does not have to be on the border router, the core of the network is a better choice, especially for dual homed sites.
- If a site requires redundant RP's then it is recommended that they use anycast RP
- The MSDP speaker and PIM RP can use different interfaces addresses.
- Filter MSDP source active messages in both directions.

MSDP Policy

MSDP policy should be enforced using SA message filters. SA filtering can typically be performed on source address, group address, and MSDP peer address.

SA filtering prevents the leaking of SA messages that should not leave a local domain, such as.

- Sources in private address space. (10/8)
- Protocol group addresses such as the auto-RP groups 224.0.1.39 and 224.0.1.40
- Administratively scoped groups (239/8)
- SSM groups (232/8)
- 225/8 -231/8 Reserved http://www.iana.org/assignments/multicast-addresses

- Cisco guidelines

http://www.cisco.com/warp/public/105/49.html

MBGP

MBGP is an advantage over BGP because it provides a distinction between multicast and unicast-only networks. MBGP allows you to advertise which networks in your LAN are multicast capable.

Cisco configuration of MBGP has three main sections

router bgp 1024
neighbor 72.40.38.229 remote-as 2048
neighbor 72.40.38.229 password 7 1207350DC8003818

address-family ipv4
neighbor 72.40.38.229 route-map international in
network 140.52.210.0 mask 255.255.255.0
network 140.52.216.0 mask 255.255.255.0 *(both unicast & multicast)*

address-family ipv4 multicast
neighbor 72.40.38.229 route-map international in
network 140.52.216.0 mask 255.255.255.0

MBGP Route Advertisement

Cisco “show” commands

- show bgp ipv4 multicast
- show bgp ipv4 multicast neighbors 10.1.1.1 received-routes
- show bgp ipv4 multicast neighbors 10.1.1.1 advertised-routes
- show bgp ipv4 multicast summary

```
Router# show bgp ipv4 multicast neighbors 198.125.140.206 received-routes

BGP table version is 51683234, local router ID is 134.55.200.65
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              S Stale
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
*  192.41.230.0/23  198.125.140.206         0             0 32361 i
*  192.84.86.0      198.125.140.206        20             0 32361 i
*  198.32.43.0      198.125.140.206         0             0 32361 i
*  198.32.44.0      198.125.140.206        20             0 32361 i
*  198.32.45.0      198.125.140.206         0             0 32361 i

Total number of prefixes 5
```

MSDP problems caused by MBGP

If your peer rejects all of your advertised MSDP SA's, it's likely an MBGP issue.

- Verify that your multicast networks and your MSDP peer network are advertised using the ipv4 multicast address family.
- Using a loopback interface for your MSDP peer is recommended, but this often leads to advertising the MSDP peer address as an MBGP host route. Your ISP may not accept this host route or they won't propagate it to peers because it's smaller than a /24.
- Review the MSDP RPF neighbor algorithm.

Use the router interface address of the network you have your AG node on for the MSDP peer ID. This will advertise both MSDP peer and AG source addresses within the same network prefix.

MSDP RPF Neighbor Determination

Router **R** is your MSDP peer, or the receiver.
Router **X** is the MSDP peer that sends the source active message.
Router **S** is the originating RP of the source active message.

- If Router X originated the source-active message (Router X is Router S), then Router X is also the peer-RPF neighbor, and its source-active messages are accepted.
- If Router X is a member of the Router R mesh group, or is the configured peer, then Router X is the peer-RPF neighbor, and its source-active messages are accepted.
- If Router X is the Border Gateway Protocol (BGP) next hop of the active multicast RPF route toward Router S (Router X installed the route on Router R), then Router X is the peer-RPF neighbor, and its source-active messages are accepted.
- If Router X is an external BGP (EBGP) or internal BGP (IBPG) peer of Router R and the last autonomous system (AS) number in the BGP AS-path to Router S is the same as Router X's AS number, then Router X is the peer-RPF neighbor, and its source-active messages are accepted.
- If Router X uses the same next hop as the next hop to Router S, then Router X is the peer-RPF neighbor, and its source-active messages are accepted.
- If Router X fits none of these criteria, then Router X is not an MSDP peer-RPF neighbor, and its source-active messages are rejected.

IGMP LAN protocol

When a host wants to become a multicast receiver, it must inform the routers on it's LAN. IGMP Is used to communicate group membership information between hosts and routers on a LAN.

IGMPv1 – Windows95

IGMPv2 – Windows98, 2000

IGMPv3 – WindowsXP, Vista

IGMP Snooping

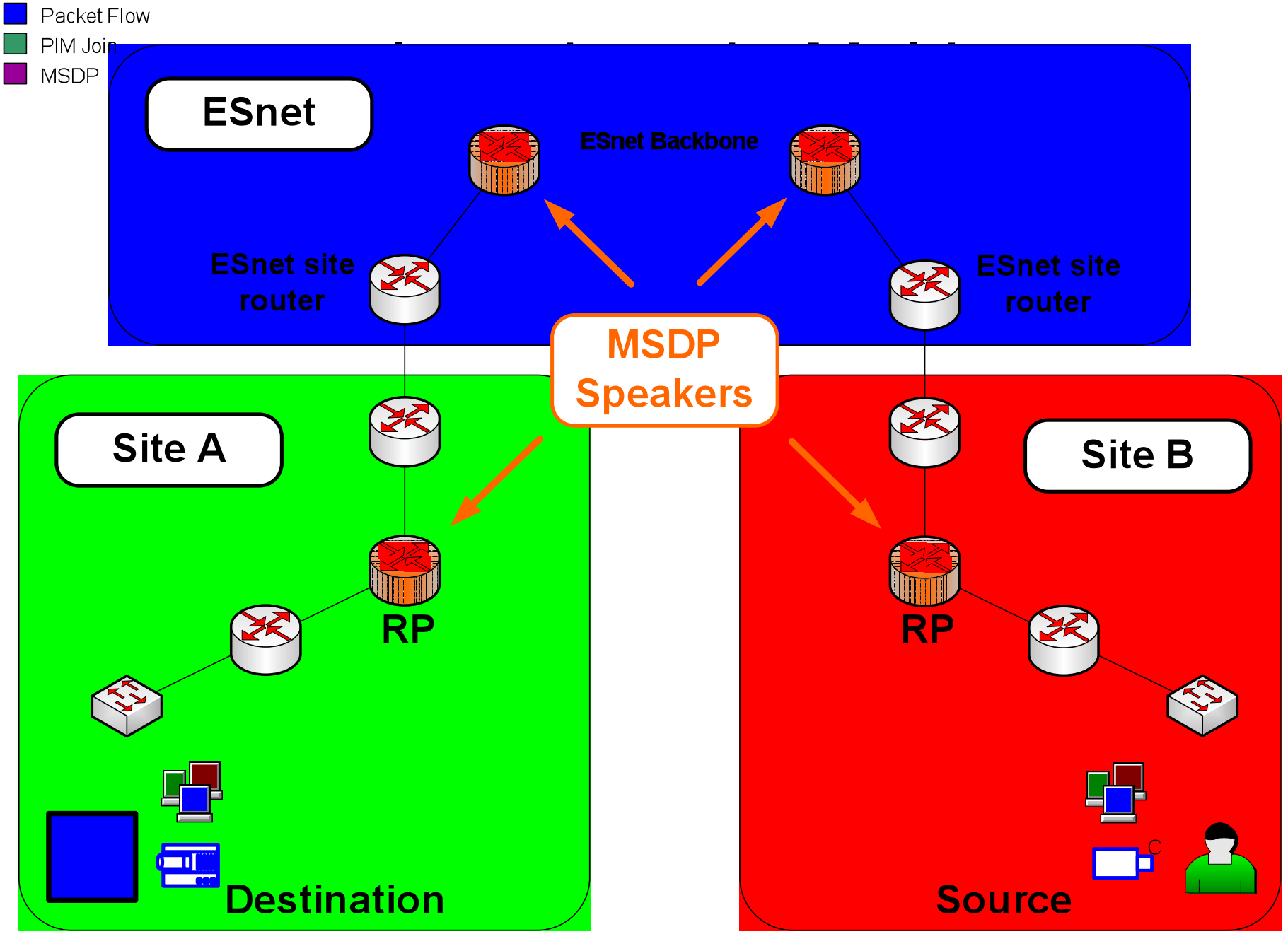
By default multicast is treated like a broadcast on a Layer2 Ethernet switch and is simply flooded out all ports on the leaf VLAN.

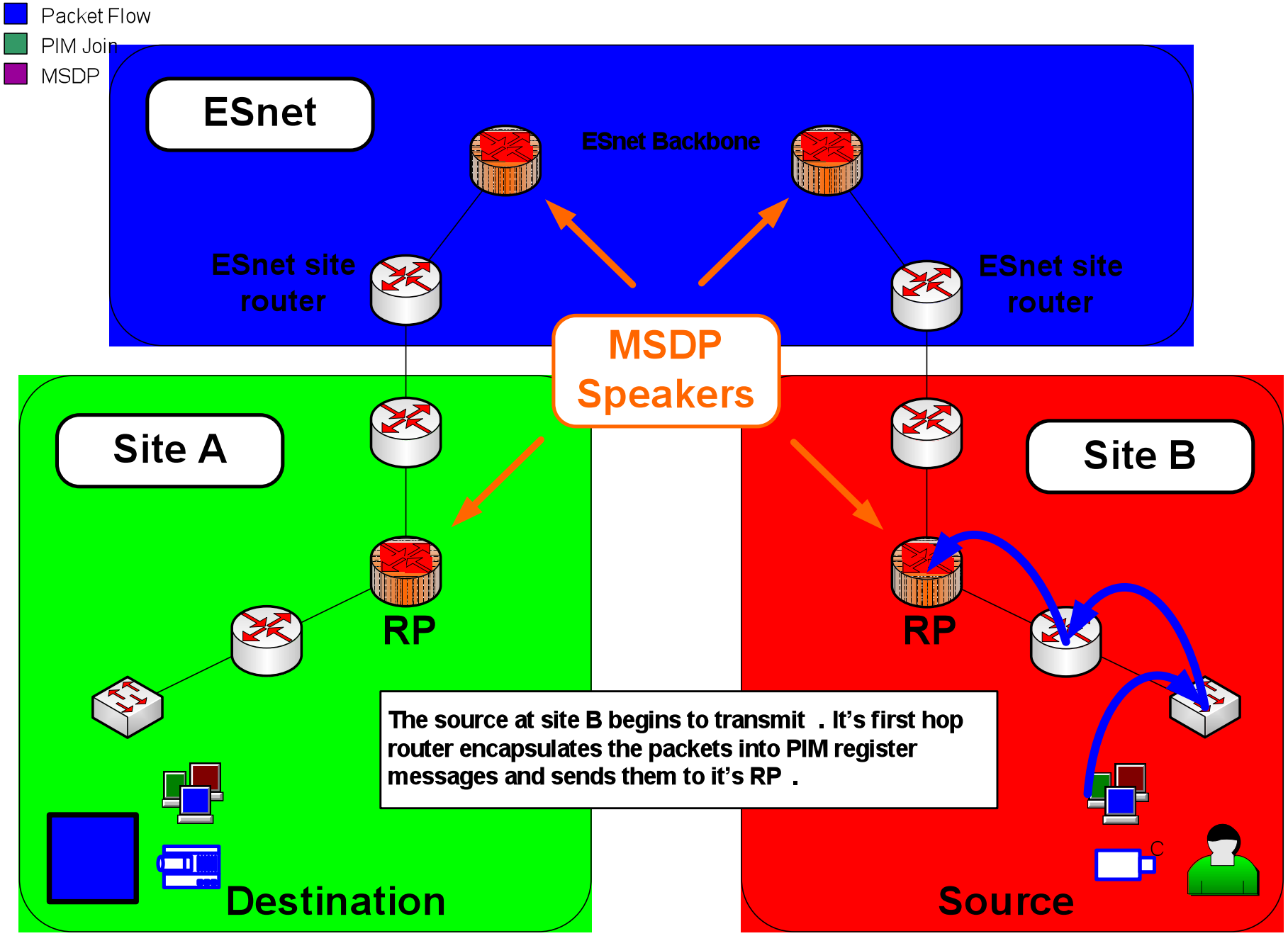
Flooding multicast packets out all switch ports wastes valuable network resources. Also, hosts that receive this unwanted traffic must use processing cycles to examine packets that they will eventually discard. IGMP snooping is one way to eliminate this inefficiency.

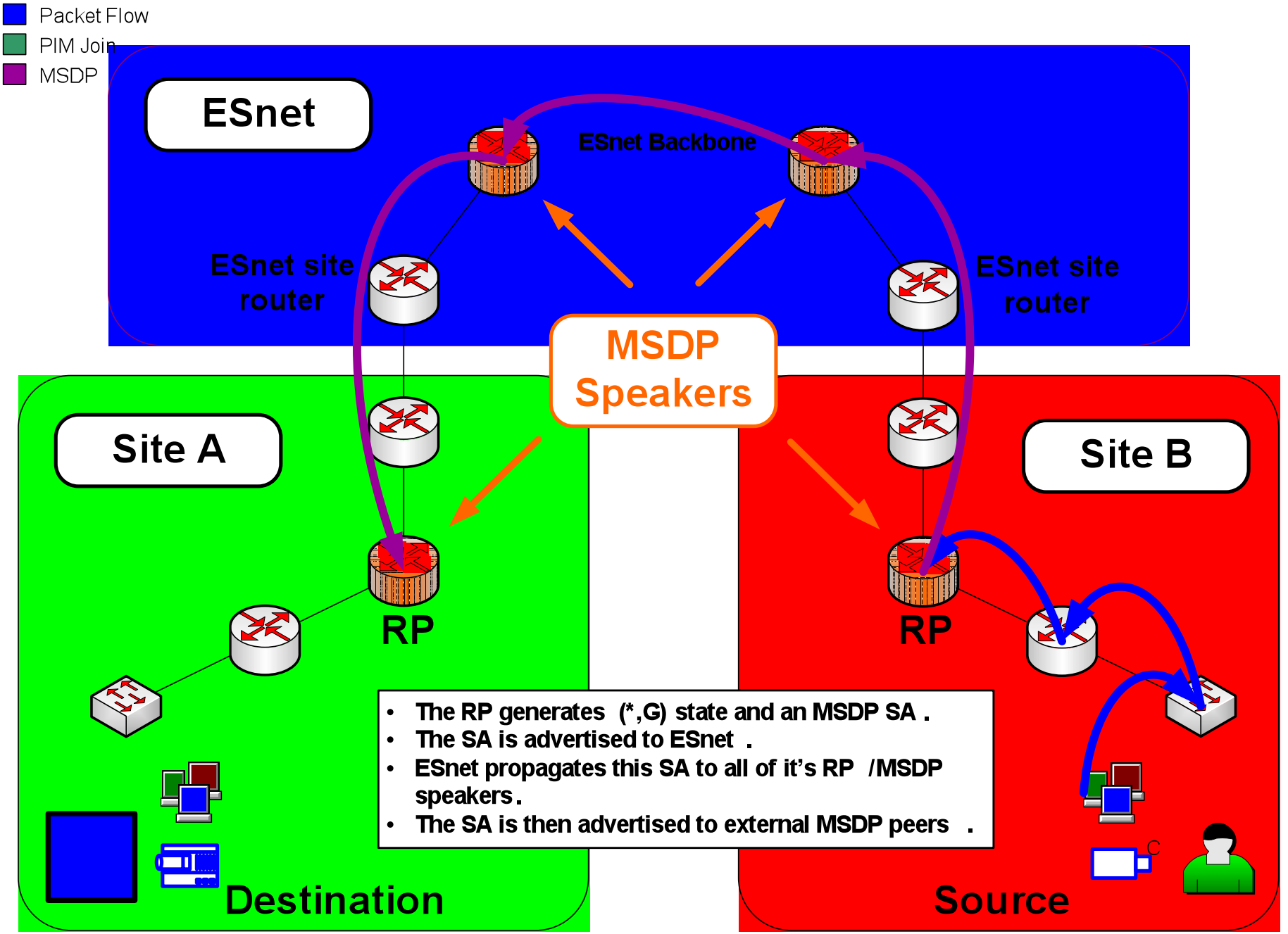
An IGMP snooping switch looks at IGMP messages to determine which hosts are actually interested in receiving multicast traffic. Multicast packets are forwarded only out ports that connect to a host that is an interested listener of a specified group.

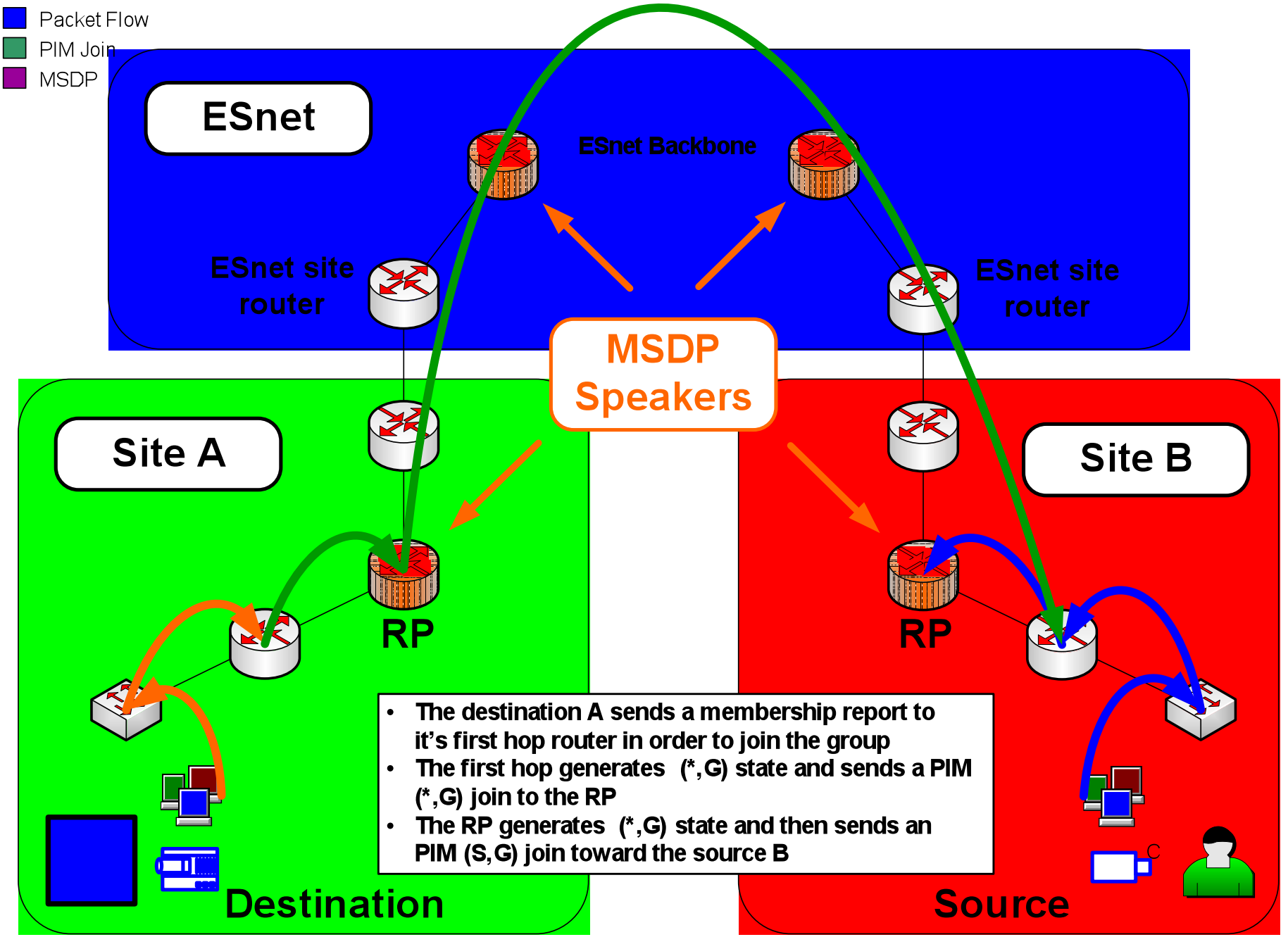
Cisco CGMP

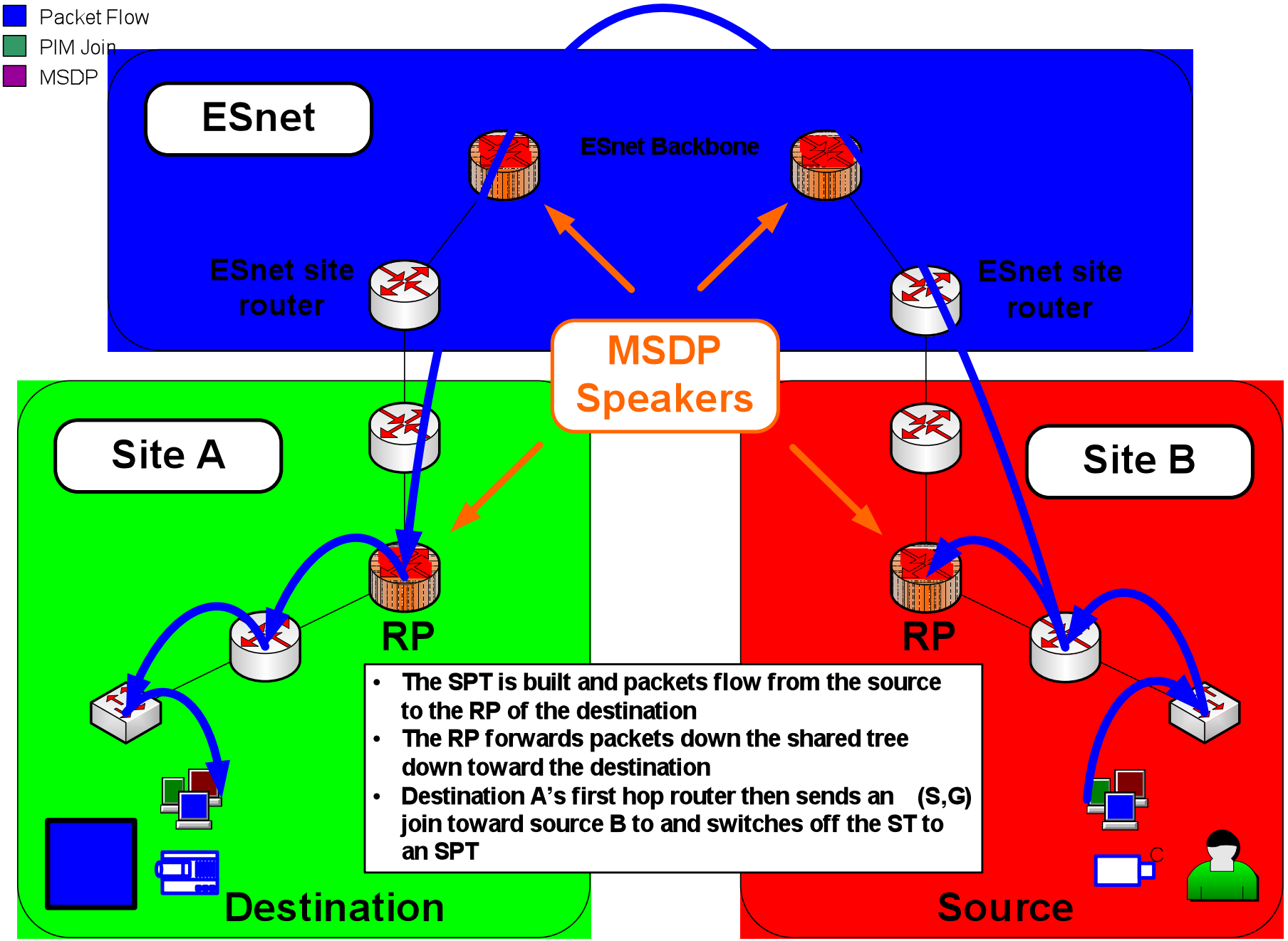
CGMP is a proprietary mechanism that provides the same functionality as IGMP snooping. CGMP relies on Cisco routers to determine which hosts are interested in each multicast group. This offloads Cisco LAN switches and is generally used on Cisco workgroup switches that lack the compute resources required for IGMP snooping.

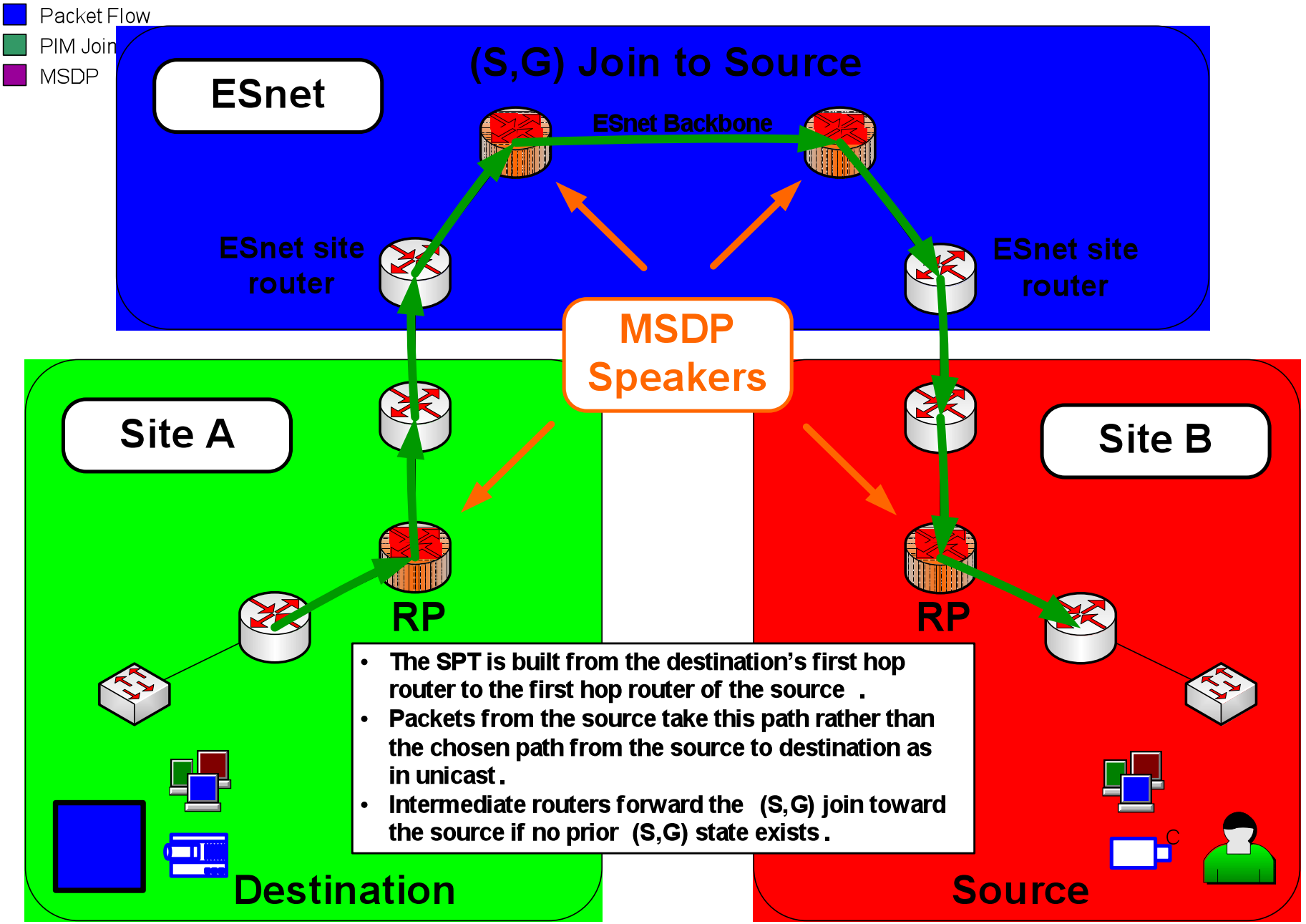

Packet Flow
PIM Join
MSDP
ESnet
ESnet Backbone
ESnet site router
ESnet site router
MSDP Speakers
Site A
Site B
RP
RP
Destination
Source

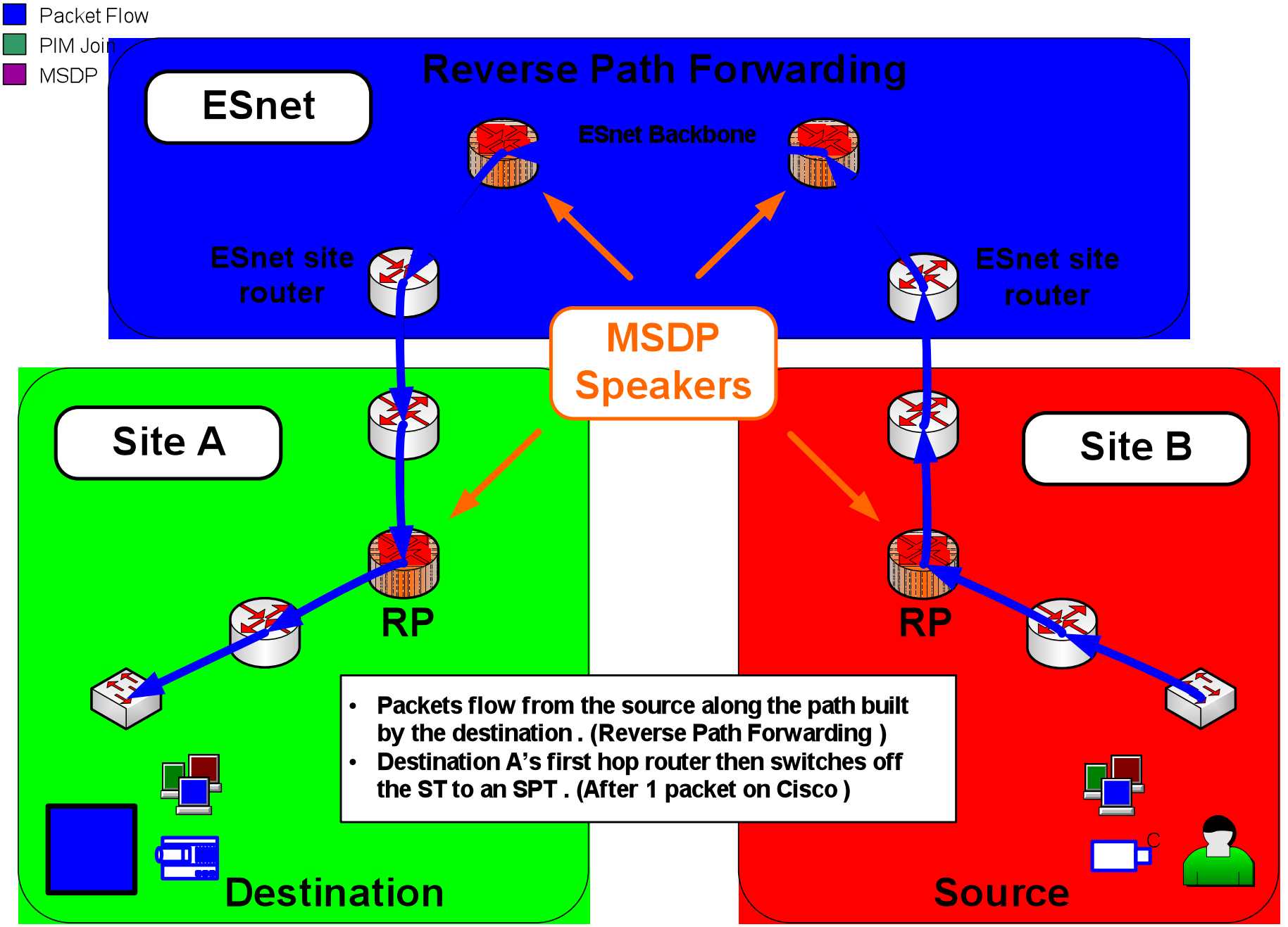

Packet Flow
PIM Join
MSDP
ESnet
ESnet Backbone
ESnet site router
ESnet site router
MSDP Speakers
Site A
Site B
RP
RP
The source at site B begins to transmit . It's first hop router encapsulates the packets into PIM register messages and sends them to it's RP .
Destination
Source


Packet Flow
PIM Join
MSDP
ESnet
ESnet Backbone
ESnet site router
ESnet site router
MSDP Speakers
Site A
Site B
RP
RP
• The RP generates (*,G) state and an MSDP SA .
• The SA is advertised to ESnet .
• ESnet propagates this SA to all of it's RP /MSDP speakers.
• The SA is then advertised to external MSDP peers .
Destination
Source


Packet Flow
PIM Join
MSDP
ESnet
ESnet Backbone
ESnet site router
ESnet site router
MSDP Speakers
Site A
Site B
RP
RP
• The destination A sends a membership report to it's first hop router in order to join the group
• The first hop generates (*,G) state and sends a PIM (*,G) join to the RP
• The RP generates (*,G) state and then sends an PIM (S,G) join toward the source B
Destination
Source


Packet Flow
PIM Join
MSDP
ESnet
ESnet Backbone
ESnet site router
ESnet site router
MSDP Speakers
Site A
Site B
RP
RP
• The SPT is built and packets flow from the source to the RP of the destination
• The RP forwards packets down the shared tree down toward the destination
• Destination A's first hop router then sends an (S,G) join toward source B to and switches off the ST to an SPT
Destination
Source


Packet Flow
PIM Join
MSDP
(S,G) Join to Source
ESnet
ESnet Backbone
ESnet site router
ESnet site router
MSDP Speakers
Site A
Site B
RP
RP
• The SPT is built from the destination's first hop router to the first hop router of the source .
• Packets from the source take this path rather than the chosen path from the source to destination as in unicast.
• Intermediate routers forward the (S,G) join toward the source if no prior (S,G) state exists.
Destination
Source


Packet Flow
PIM Join
MSDP
Reverse Path Forwarding
ESnet
ESnet Backbone
ESnet site router
ESnet site router
MSDP Speakers
Site A
Site B
RP
RP
• Packets flow from the source along the path built by the destination . (Reverse Path Forwarding)
• Destination A's first hop router then switches off the ST to an SPT . (After 1 packet on Cisco)
Destination
Source

Troubleshooting Interdomain Multicast

When your not receiving an external source

- Verify IGMP membership at your first hop router
- Log into your RP, Start with MSDP
- Verify the remote source MSDP SA exists
  - Cisco# show ip msdp sa-cache
  - Juniper> show msdp source-active group A.B.C.D
- Verify the RPF route for the source
  - Cisco# show ip rpf A.B.C.D
  - Juniper> show multicast rpf A.B.C.D
- If the router is an MSDP speaker, verify RPF to the remote RP
- Verify the PIM (S,G) incoming interface is aligned with source RPF
- Verify packet counters
- Contact ESnet if necessary

Troubleshooting Interdomain Multicast

When your not being received

- Log into your RP
- Verify that your MSDP SA is being advertised, contact ESnet if necessary
- Verify the PIM (S,G) for your source at your RP
- Verify your wide area PIM neighbor
- Verify that an Outgoing Interface List (OIL) entry matches the RPF for the remote listener.
- Verify packet counters
- Contact ESnet if necessary

show ip mroute (*,G)

(Cisco)

Cisco#**show ip mroute 233.4.200.18**
IP Multicast Routing Table
Flags: D - Dense, **S - Sparse**, B - Bidir Group, s - SSM Group, **C - Connected**,
L - Local, P - Pruned, R - RP-bit set, F - Register flag,
T - SPT-bit set, **J - Join SPT**, M - MSDP created entry,
X - Proxy Join Timer Running, A - Candidate for MSDP Advertisement,
U - URD, I - Received Source Specific Host Report, Z - Multicast Tunnel
Y - Joined MDT-data group, y - Sending to MDT-data group
Outgoing interface flags: H - Hardware switched, A - Assert winner
Timers: Uptime/Expires
Interface state: Interface, Next-Hop or VCD, State/Mode

(*, 233.4.200.18), 8w1d/stopped, **RP 198.129.245.2**, **flags: SJC**
Incoming interface: Null, RPF nbr 0.0.0.0
Outgoing interface list:
Vlan110, Forward/Sparse, 1w3d/00:02:02

Show the multicast routing trees for beacon group 233.4.200.19
PIM Flags, Sparse, Join SPT, Connected
The (STAR COMMA G) entry, this is the RP router and shared tree
Incoming Interface is NULL since the RP is the top of the shared tree
Outgoing interface lists indicates the direction to the receivers/group members

show ip mroute (S,G)

(Cisco)

```
Cisco# show ip mroute 233.4.200.18
IP Multicast Routing Table
...

(64.65.64.31, 233.4.200.18), 1d20h/00:02:55, flags: MT
  Incoming interface: Vlan4, RPF nbr 198.129.76.25, RPF-MFD
  Outgoing interface list:
    Vlan110, Forward/Sparse, 1d20h/00:02:02, H
    Vlan220, Forward/Sparse, 1d11h/00:02:32, H
...
```

Source/transmitter address, 64.157.28.13
Group/destination address, 233.4.200.18 (NLANR Beacon Group)
Packets from this source arrive via interface Vlan 10, from neighbor 198.129.76.25
RPF-MFD flag indicates the flow is completely hardware switched
Packets exit the router on their way to group members via these interfaces
Age of initial join message / Expiration timer

show ip mroute count
(Cisco)

Cisco# **show ip mroute 233.4.200.18 count**
IP Multicast Statistics
264 routes using 101154 bytes of memory
16 groups, 15.50 average sources per group
Forwarding Counts: **Pkt Count**/**Pkts per second**/Avg Pkt Size/Kilobits per second
Other counts: **Total**/RPF failed/**Other drops(OIF-null, rate-limit etc**)

Group: 233.4.200.18, Source count: 94, Packets forwarded: 44708, Packets received: 44832
RP-tree: Forwarding: 0/0/0/0, Other: 0/0/0
Source: 63.105.122.14/32, Forwarding: **0**/0/0/0, Other: **0**/0/0
Source: 128.111.252.50/32, Forwarding: **718**/**10**/72/6, Other: **718**/0/0
Source: 128.118.57.33/32, Forwarding: **797**/**10**/70/5, Other: **797**/0/0
Source: 128.55.16.111/32, Forwarding: **634**/**9**/74/6, Other: **634**/0/0
…
Source: 129.250.11.22/32, Forwarding: **914**/**19**/74/10, Other: **915**/0/**1**

This source is probably running two instances of the NLANR beacon

The Cisco show ip mroute count commands displays per source packet information for a group, packet totals, rates, average size, drops etc.

show pim join (S,G)

(Juniper)

```
Juniper> show pim join 233.4.200.18 extensive
Instance: PIM.master Family: INET

Group: 233.4.200.18
    Source: 64.65.64.31
    Flags: sparse,spt
    Upstream interface: ge-1/1/0.0
    Upstream neighbor: 134.55.209.21
    Upstream state: Join to Source
    Keepalive timeout: 200
    Downstream Neighbors:
        Interface: so-0/1/0.0
            134.55.209.218 State: Join Flags: S Timeout: 168
        Interface: so-0/1/1.0
            134.55.209.6 State: Join Flags: S Timeout: 184
...
```

Source/transmitter address
Group/destination address, 233.4.200.18 (NLANR Beacon Group)
Packets from this source arrive via 134.55.209.21 on interface ge-1/1/0.0
Packets exit the router on their way to PIM neighbors via these interfaces
PIM Join Expiration timer

show multicast route (S,G)

(Juniper)

```
Juniper> show multicast route group 233.4.200.18 extensive

Group: 233.4.200.18
    Source: 64.65.64.31/32
    Upstream interface: ae0.0
    Downstream interface list:
       so-0/1/0.0 so-0/1/1.0
    Session description: Static Allocations
    Statistics: 1 kBps, 8 pps, 880606 packets
    Next-hop ID: 461
    Upstream protocol: PIM
    ...
```

Source/transmitter address
Group/destination address, 233.4.200.18 (NLANR Beacon Group)
Packets from this source arrive via interface ae0.0
Packets exit the router on their way to PIM neighbors via these interfaces
Packet counter & rate

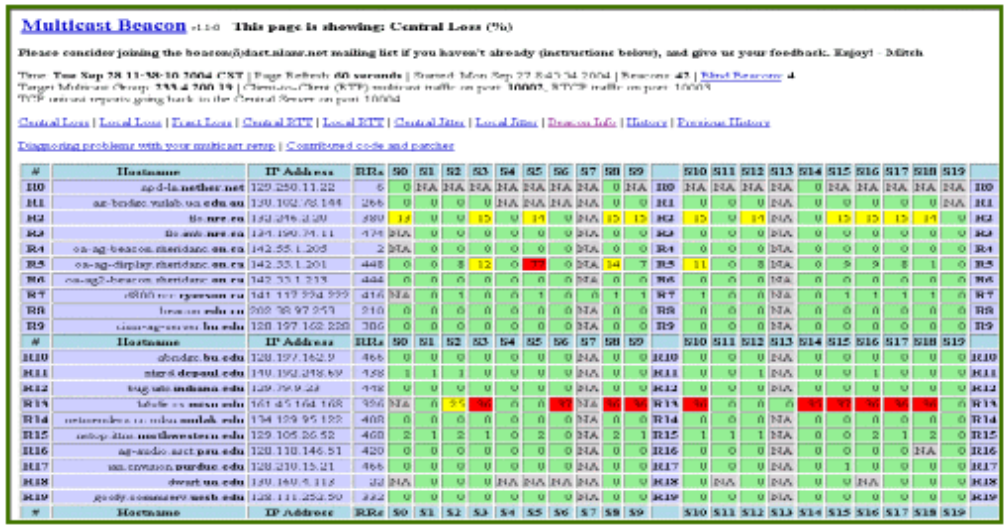
http://dast.nlanr.net/projects/beacon/

NLANR
DISTRIBUTED APPLICATIONS SUPPORT TEAM

AAD | Advisor | Autobuf v2.0 | Multicast Beacon | BIMA | Iperf | NextINet | Tools | Web100 | All Projects

Multicast Beacon v1.3

Current version: 1.3-0 (August 26, 2005)



The NLANR beacon is a great tool for verifying interdomain multicast configuration. Every site with an AG node should have their own beacon, the local LAN group should manage it.

Source Packet Generation

(for debugging)

iperf -u -i1 -c 233.1.37.1 -b 1K -T 70 -t 60

- -u UDP
- -i Status update interval
- -c Client mode connect to host address
- -b bit rate
- -T TTL (greater than 32)
- -t transmit time in seconds

ping -U -L -t 70 233.1.37.1 60

- -U UDP
- -L No loopback packets for multicast
- -t TTL
  - group address
  - number of packets to send
  - ping interval is 1 second by default

To be used in conjunction with an IGMP static group join at the receiving router.

ESnet Contact Info

NOC phone - (510) 486 7607

Email - trouble@es.net

Mike O'Connor
ESnet Network Engineering Group
Lawrence Berkeley National Lab
moc@es.net