IP Multicast Best Practices for Enterprise Customers

Last updated: October 2009

This document describes the generally accepted best common practices for IP Multicast in Enterprise customer networks. Although many of the practices in this document were developed for Financial customers to deliver Market Data the general principles apply to any Enterprise Multicast Deployment. It describes ways to optimize multicast delivery according to basic design principals including:

- Resiliency
  - Path diversity
  - Redundancy
  - Load sharing or splitting
- Latency
- Security

These recommendations are consistent with the existing Solution Reference Network Designs (SRND) listed below. They should be consulted for further information.

Designing a Campus Network for High Availability: http://www.cisco.com/application/pdf/en/us/guest/netsol/ns431/c649/ccmigration_09186a008093b876.pdf

High Availability Campus Network Design-Routed Access Layer using EIGRP or OSPF: http://www.cisco.com/application/pdf/en/us/guest/netsol/ns432/c649/ccmigration_09186a008081146 8.pdf

Cisco AVVID Network Infrastructure IP Multicast Design (SRND): http://www.cisco.com/application/pdf/en/us/guest/tech/tk363/c1501/ccmigration_09186a008015e7cc.pdf

General information about IP Multicast: http://www.cisco.com/go/multicast

Using Point-to-Point Links in the Core

A collapsed backbone should be avoided in the core of the network. A common layer 2 segment between routers introduces a number of unnecessary complexities and inefficiencies as described below.

a. Triggered events on link failure

When a router or a link fails in a P2P environment the carrier signal is dropped and creates a triggered event that will cause immediate IGP convergence, which will be followed by IP Multicast convergence.

In a switched environment, a router can fail and it will not be detected until several hello messages are missing at a layer 3 protocol level. This will increase the convergence time.

Using BFD may be able to minimize the effect on convergence time.

b. Avoid situations which require PIM snooping

In networks where a Layer 2 switch interconnects several routers, the switch floods IP Multicast packets to all multicast router ports by default, even if there are no multicast receivers downstream. In these environments, PIM snooping should be used to constrain the multicast to the interested routers.

With PIM snooping enabled, the switch restricts multicast packets for each IP multicast group to only those multicast router ports that have downstream receivers joined to that group. When you enable PIM snooping, the switch learns which multicast router ports need to receive the multicast traffic within a specific VLAN by listening to the PIM hello messages, PIM join and prune messages, and bidirectional PIM designated forwarder-election messages.

Point-to-point interfaces will avoid the additional complexity that requires PIM snooping.

c. Assert issues

The PIM Assert mechanism prevents duplicate traffic from flowing into a multi-access network. Assert messages are sent by the two forwarding routers and the router with the best metric will win the assert. There are several known cases in which assert can cause temporary routing loops and sub optimal behavior.

Point-to-point interfaces will avoid assert issues with IP Multicast.

Tuning at Access Layer Edge

a. Loop Free Layer 2

Limit VLANs to a single closet whenever possible. There should be no STP loops - all interfaces should be in forwarding state—no interfaces in blocked state.

There are many reasons why STP/RSTP convergence should be avoided for the most deterministic and highly available network topology. In general, when you avoid STP/RSTP, convergence can be predictable, bounded, and reliably tuned. Additionally, it should be noted that in soft failure conditions, where keepalives (BPDU or routing protocol hellos) are lost, L2 environments fail open, forwarding traffic with unknown destinations on all ports and causing potential broadcast storms; while L3 environments fail closed, dropping routing neighbor relationships, breaking connectivity, and isolating the soft failed devices.

b. If STP is absolutely required, use Rapid PVST+

Older applications that require L2 connectivity between Data Center or L2 switches need to be updated and/or replaced. Very old Tibco middleware versions required the use of a L2 broadcast for a heartbeat. It has been a decade since that middleware version has been updated to use a L3 IP Multicast heartbeat.

If you are compelled by application requirements to depend on STP to resolve convergence events, use Rapid PVST+. Rapid PVST+ is far superior to 802.1d and even PVST+ (802.1d plus Cisco enhancements) from a convergence perspective.

c. Hardcode all interface settings

Hardcode duplex, speed and trunking capability on router and switch interfaces and then turn off auto-negotiation. This tuning can save seconds during re-convergence when restoring a failed link or node. Unused VLANs should be manually pruned from trunked interfaces to avoid broadcast propagation. Finally, VTP transparent mode should be used because the need for a shared common VLAN database is reduced.

IGP Tuning

IP Multicast traffic will converge after unicast routing converges. Therefore it is important to minimize convergence on the edge by tuning IGP timers.

a. EIGRP

Set hello and dead timers to 1 and 3:

```
interface GigabitEthernet1/0
   ip hello-interval eigrp 100 1
   ip hold-time eigrp 100 3
```

b. OSPF

Tune OSPF Fast hello, dead-interval, SPF and LSA throttle timers.

The example below sets the dead interval to 1 second and the hello interval to 250 ms.

```
interface GigabitEthernet1/0
   ip ospf dead-interval minimal hello-multiplier 4
```

The SPF and LSA throttle timers should be tuned to these recommended settings.

spf-start	10 ms
msecspf-hold	100 to 500 ms
msecspf-max-wait	5 seconds
lsa-start	10 ms
mseclsa-hold	100 to 500 ms
mseclsa-max-wait	5 seconds
lsa arrival	80 ms (less than lsa-hold of 100 ms)

This is an example on setting those timers:

```
router ospf 100
   timers throttle spf 10 100 5000
   timers throttle lsa all 10 100 5000
   timers lsa arrival 80
```

All these timers must be set consistently on both sides of the link.

IGMP Snooping

IGMP snooping is an IP Multicast constraining mechanism that runs on a Layer 2 LAN switch. Without IGMP snooping enabled, all multicast traffic will be forwarded to all hosts connected to the switch. IGMP snooping will insure that only hosts that are interested in the data stream will receive it.

Every Cisco switch supports IGMP snooping. IGMP snooping should always be enabled if you are running IP Multicast. Some platform and switch software combinations may not have IGMP snooping enabled by default. Make sure IGMP snooping is enabled before running any multicast streams.

There are some situations in which network administrators would like to run multicast in a contained environment and not have it forwarded to the rest of the network. In those cases, PIM is not enabled on the routers and there is no IGMP querier elected.

In order for IGMP Snooping to operate correctly there needs to be an IGMP Querier sending out periodic IGMP Queries, so that the receivers will respond and send out IGMP Membership reports. These reports control which switchports will receive the multicast traffic for a particular group.

If PIM is not enabled on at least one router in the switch environment then one router or switch needs to be configured as the IGMP querier. This is accomplished with this interface command:

```
ip igmp snooping querier
```

An alternative would be to configure PIM on the interface facing the switch environment. In this case, the igmp querier will not have to be explicitly configured.

Choosing the Right Multicast Groups

There are some basic rules that must be followed for selecting which IP Multicast address range to use.

a. Do not use x.0.0.x or x.128.0.x group addresses

Multicast addresses in the 224.0.0.x range are considered link local multicast addresses. They are used for protocol discovery and are flooded to every port. For example, OSPF uses 224.0.0.5 and 224.0.0.6 for neighbor and DR discovery.

These addresses are reserved and will not be constrained by IGMP snooping. Do not use these addresses for an application.

Further, since there is a 32:1 overlap of IP Multicast addresses to Ethernet MAC addresses, any multicast address in the [224-239].0.0.x and [224-239].128.0.x ranges should NOT be considered.

b. Use 239/8 addresses for internal applications

RFC 2365 describes the use of administratively scoped IP Multicast addresses. This address range should be used for all internal applications. The concept is similar to the use of RFC 1918 addresses for unicast.

c. Use 233 GLOP addresses for interdomain applications

RFC 3180 describes the use of GLOP addresses that can be used based on an AS number. Exchanges should be encouraged to use these addresses for interdomain multicast data streams.

d. Use PIM-SSM and 232/8 for interdomain one to many applications

RFC 4608 describes the use of the 232/8 address range for PIM-SSM interdomain applications. Exchanges and FSPs are encouraged to use PIM-SSM and the 232/8 address range for one-to-many unidirectional multicast data delivery.

e. Petition IANA for a 224 to use externally

The last choice for external addresses is to petition IANA for a 224 address range to use for your interdomain application. This should be considered a last resort for content providers such as stock exchanges that need to insure there will not be an address collision globally with any provider or customer. This address space is extremely limited but many of the largest exchanges have successfully been assigned 224 address ranges.

More information and general guidelines for IP Multicast address allocation can be found in the document:

Guidelines for Enterprise IP Multicast Address Allocation:
http://www.cisco.com/en/US/tech/tk828/technologies_white_paper09186a00802d4643.shtml

PIM Query-Interval Tuning

The 'ip pim query-interval' controls the interval that a PIM hello packet is transmitted out each pim enabled interface.

The PIM hello packets are used to discover PIM neighbors and to determine the Designated Router (DR) on each network segment. The default interval for the PIM hello packets to be sent is 30 seconds. A PIM neighbor is considered down after 3 consecutive missed messages. Therefore, it could take 90 seconds for the DR to failover. If you lower the query interval to 1 second, then the DR failover time is reduced to 3 seconds.

The goal is not to set the query-interval too low so that there is unnecessary flapping. Cisco generally recommends a 1 second query-interval, which would give you a 3 second failover at the receiver edge. Some customers may choose to use the sub-second option. Cisco does not recommend an interval less than 500 ms. Due to queue lengths and processing delays on the switch platforms, lower intervals have been known to cause problems.

Keep in mind that a router with 30 LAN segments and a query-interval of 1 will need to send out 30 PIM hellos every second. If you turn down the query-interval to 500 ms then there will be 60 messages per second.

In the core of the network there are typically point-to-point links and not any directly connected receivers. When a link goes down on a P2P link, it is a triggered event and the PIM neighbor is immediately removed. After unicast routing reconverges, PIM join messages will be sent on the alternative path for the active multicast streams. Therefore, there is no need to turn down the query-interval in the core and it is a waste of CPU cycles and bandwidth.

In summary:

- Turn down the pim query-interval on the receiver edge to reduce DR failover time
- This only needs to be done when there are redundant edge routers and receivers
- A general recommendation is a query interval of 1 second and no less than 500ms. This should be used with care as the number of interfaces increase.

Register Rate Limits

When a new source starts transmitting in a PIM Sparse Mode network, the packets will be encapsulated and sent using unicast to the Rendezvous Point (RP). This process can be taxing on the CPU of the Designated Router (DR) and the RP if the source is running at a high data rate and/or there are many new sources starting at the same time. This scenario can potentially occur immediately after a network failover.

In order to protect both the edge routers and the RP, it is recommended to set the 'ip pim register-rate limit' to a relatively low value. Normally, there is no limit to the number of packets that will be encapsulated and sent to the RP.

Use this command to limit the number of register messages that the Designated Router (DR) will allow for each (S, G) entry. Enabling this command will limit the load on the DR and the Rendezvous Point (RP) at the expense of dropping register messages that exceed the set limit. Receivers may experience data packet loss in the first seconds in which register messages are sent from bursty sources.

When the 'ip pim dense-mode proxy-register' command is configured, the ip pim register-rate-limit command also should be configured because of the potentially large number of sources from the dense mode area that may send data into the sparse mode region. If the ip pim register-rate-limit command is not configured in this scenario, the Cisco IOS Software will automatically apply the default register-rate-limit of two messages per second.

The number to limit the register packets will depend on the number of potential sources registering at the same time and their data rate. A typical setting in a PIM Sparse Mode (PIM-SM) network is between 4 and 10 messages per second.

MSDP Timers

In PIM-SM deployments that use MSDP, there may be situations in which faster convergence of the Source Active (SA) messages is desired. A typical scenario is when the MSDP session is reset and new sources start up during the time the session is being reestablished. Potentially it may take as long as one minute for the new traffic stream to start forwarding.

For these situations, you may want to consider adjusting the MSDP timers down to as low as 5 seconds:

```
ip msdp keepalive <peer-name-or-address> 5 15
ip msdp timer 5
```

Note: The source information in the SA Cache will remain active for as long as 6 minutes. Modifying these times will only apply to new sources that start up during the time that the MSDP session is down. As with any timer settings, there is a tradeoff between higher CPU utilization and network convergence.

Multicast Stub Recommendation

The Multicast Stub command should be used on the Cisco Catalyst 6500 Series Switch in redundant Layer 2 edge networks to protect the CPU from non-rpf traffic. An explanation of the problem can be found in the whitepaper:

Redundant Router Issues with IP Multicast in Stub Networks:
http://www.cisco.com/en/US/products/ps6552/products_white_paper09186a00800a4424.shtml

There are some specific issues with the different Supervisors related to this feature:

Supervisor	Multicast Stub Recommendation
Sup 1A	Yes, it should be configured on leaf subnets
Sup 2	Yes, Sup 2 has FIB based rate limiting enabled in later Cisco IOS Software versions. Multicast Stub should be used in versions before that feature. FIB based rate limiting is on by default. The command to disable is "no mls ip mul non-rpf cef"
Sup 720	No, Sup720 has non-rpf netflow rate limiting. The multicast stub command can be used as an additional protection to the control plane.

Static RP vs. AutoRP Listener

The main tradeoff between using static RP configuration and AutoRP is administrative overhead.

a. Static RP

An RP could be statically defined with as little as 1 line on each router. If the network does not have many different RPs defined and/or they don't change very often this could be an attractive option.

The override option can be used with the rp-address configuration for additional security. This option will cause the router to ignore any AutoRP or BSR announcements that conflict with the statically defined RP.

Sample config:

```
ip pim rp-address 1.1.1.1 1
access-list 1 permit 239.254.1.0 0.0.0.15
```

b. AutoRP with AutoRP Listener

Previously, sparse-dense mode was required on the interfaces to run AutoRP. Today, sparse-mode is configured on the interfaces and the autorp listener option is configured globally.

On every router:

```
ip pim autorp listener
interface GigabitEthernet3/40
     ip address 126.1.3.11 255.255.255.0
     ip pim sparse-mode
```

On the RP routers:

```
ip pim send-rp-announce Loopback0 scope 16 group-list 7
ip pim send-rp-discovery Loopback1 scope 16
access-list 7 permit 239.254.2.0 0.0.0.255
interface Loopback0
     ip address 126.0.4.1 255.255.255.255
     ip pim sparse-mode
interface Loopback1
     ip address 126.0.1.15 255.255.255.255
     ip pim sparse-mode
```

This example is advertising the Anycast RP address of 126.0.4.1 and the AutoRP announcement messages are being sent with a source address from Loopback 1.

It is recommended that with either Static RP or AutoRP Listener you also have RP redundancy with Anycast RP or the Phantom RP.

Anycast RP for PIM-SM

The RP is a critical function for PIM-SM and PIM-Bidir deployments. RP redundancy is always recommended. The best form of redundancy for PIM-SM is Anycast RP which is described in the document:

Anycast RP:
http://www.cisco.com/en/US/docs/ios/solutions_docs/ip_multicast/White_papers/anycast.html

Phantom RP for PIM-Bidir

The RP is a critical function for PIM-SM and PIM-Bidir deployments. RP redundancy is always recommended. The best form of redundancy for PIM-Bidir is the Phantom RP which is described in the document:

Bidirectional PIM Deployment Guide :
http://www.cisco.com/en/US/prod/collateral/iosswrel/ps6537/ps6552/ps6592/prod_white_paper0900aecd80310db2.pdf

Reliable Design Issues

Alternating DR Priority

Load Balancing should always be built into any campus network design. For unicast one key way this is done is alternating the HSRP primary and STP root between redundant routers on the edge. This can be done with odd and even vlans. This practice will insure that a single failure will only affect 50% of the users. The rest will need to route around the failure.

When IP Multicast traffic is pulled through the network the paths are determined by the Designated Router (DR) that sends the PIM joins from the edges of the network. The DR can be alternated between odd and even VLANs as well.

Sample config:

Odd Router:

```
interface Vlan129
     ip address 126.2.129.17 255.255.255.0
     ip pim dr-priority 5
     ip pim sparse-mode
interface Vlan130
     ip address 126.2.129.17 255.255.255.0
     ip pim sparse-mode
```

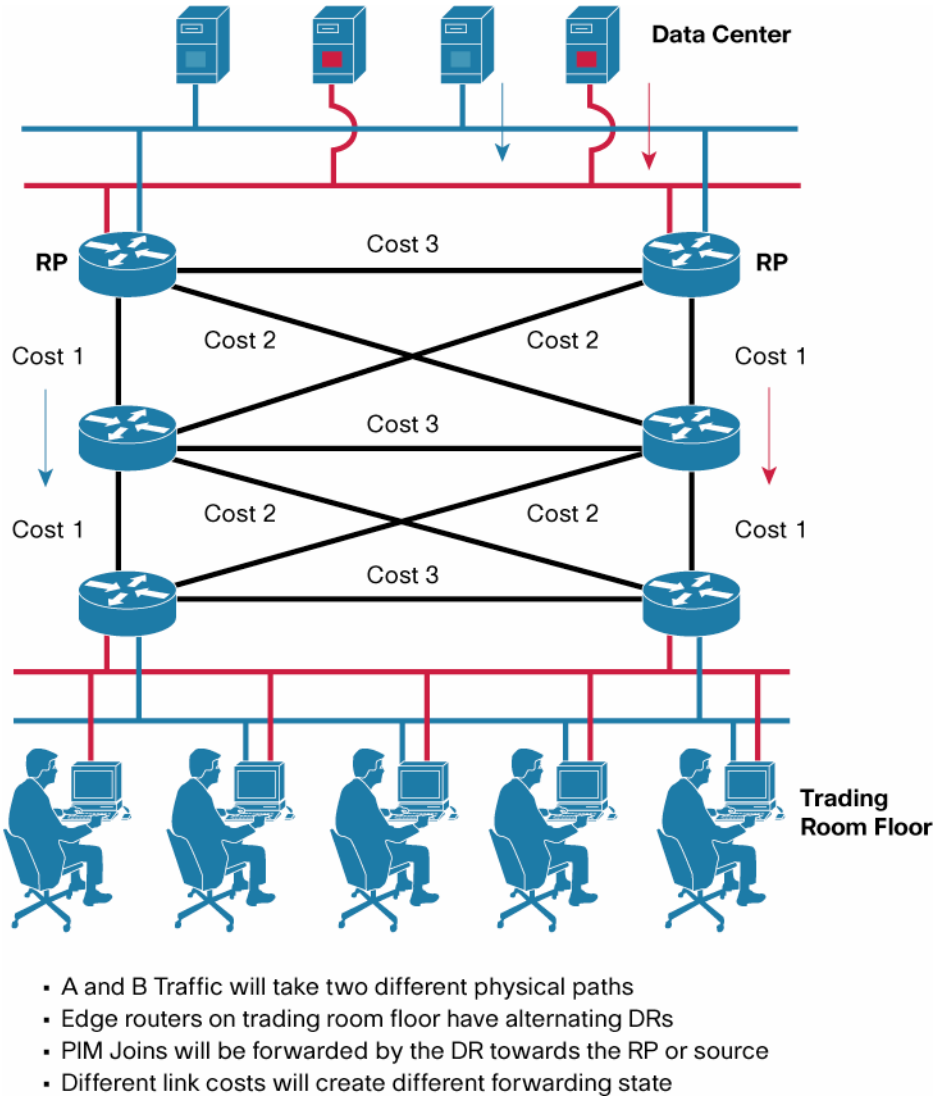
Even Router:

```
interface Vlan129
```

```
      ip address 126.2.129.18 255.255.255.0
      ip pim sparse-mode
interface Vlan130
      ip address 126.2.129.18 255.255.255.0
      ip pim dr-priority 5
      ip pim sparse-mode
```

Figure 1 has an example of how alternating Designated Routers can be used together with modifying the IGP costs on the links in the core to achieve path diversity.

Figure 1. Market Data Distribution—Path Diversity



Multicast Multipath for Load Splitting

The multicast multipath command can be used to effectively load split multicast traffic through the core of the network. This feature can give you increased overall resiliency since now a single failure in the core could potentially only affect 50% of the traffic streams.

By default, if ECMP paths are available, the RPF for multicast traffic will be based on the highest IP address. This method is referred to as the highest PIM neighbor behavior and is consistent with RFC 2362 for PIM Sparse mode but also applies to PIM-Bidir and PIM-SSM.

When the 'ip multicast multipath' command is configured, the multicast load splitting will be based on the source address of the stream. PIM Joins will be distributed over the different ECMP links based on a hash of the source address.

The multipath behavior is a type of load splitting and not load balancing. This means that if there are just a few streams, they will be divided over the multiple paths but the bandwidth load many not be balanced.

Using multicast multipath may slightly complicate troubleshooting to some degree - now multicast traffic will flow across several paths instead of one and may be less deterministic. This small increased complexity is offset by the gain in resilience. Also, any particular S, G mroute should be forwarded over the same path given the same set of circumstances - the RPF selection is based on a hash and is not random.

Since the RPF neighbor is based on the source address of the stream, it is possible that a high bandwidth source sending to many groups will all flow along one path. A more effective method for the selection of the RPF interface has been developed, which takes into account a hash based on the source, group and the next hop address. The syntax for the new command is as follows:

```
ip multicast multipath s-g-hash next-hop-based
```

This command is available in Cisco IOS Software Release 12.2(33)SRB and will be in a future Release 12.2SX release.

A good explanation of the command and all the ECMP options can be found in Overview of ECMP Multicast Load Splitting:
http://www.cisco.com/en/US/docs/ios/12_4t/ip_mcast/configuration/guide/mctlsplt.html

Additional information about multicast multipath can be found in the configuration note from engineering Configuration Note for IP Multicast Multipath: ftp://ftpeng.cisco.com/ipmulticast/config-notes/multipath.txt

Edge Security

a. Standard IOS security commands

ip multicast boundary	Should be used to filter control traffic and data flows
ip pim dr-priority	Should be used to insure that changing an ip address will not break the forwarding model
ip pim neighbor-filter	Should be used in cases in which there is a neighbor relationship. The neighbor filters are based on an ip address which can be spoofed, but it does offer some security. The provider will need to weigh the advantages of using the neighbor-filter with the administrative overheard.
ip pim accept-register	Should be used on the RP in the cases of a shared PIM-SM domain.
ip multicast route-limit	Should be considered in cases with dynamic subscriptions
ip msdp sa-filter	Should be used to filter Source Active (SA) messages if running MSDP
ip msdp sa-limit	Should be used to filter Source Active (SA) messages if running MSDP

b. Multicast Hardware Based Rate Limiters on Cisco 6500/7600 Sup720

The Cisco 6500/7600 has hardware based CPU rate limiters specifically for IP Multicast. The following rate limiters should be considered for edge security:

mls rate-limit multicast ipv4 partial

mls rate-limit multicast ipv4 fib-miss

mls rate-limit multicast ipv4 ip-options

mls rate-limit multicast ipv4 igmp

mls rate-limit all ttl-failure

mls rate-limit multicast ipv4 pim

The PIM rate limiter requires Release 12.2(33)SXH

Due to the way that incoming PIM packets are handled in hardware on the Sup720, the IGMP rate limiter is also effective for controlling the rate at which unicast PIM Register messages are sent to the CPU when received on the RP.

The actual limits for the multicast hardware based rate limiters will need to be designed with knowledge of the traffic characterizations in each environment.

Detailed information on the hardware based rate limiters can be found in the document:

Protection for the 6500 Against DoS Attacks:
http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps708/prod_white_paper0900aecd802ca5d6.html

c. Control Plane Policing (CoPP)

CoPP should be considered to prevent intentional or unintentional attacks on the edge router CPU. The Deploying Control Plane Policing white paper can be used as good general information on deploying CoPP.

d. Firewalls

PIX firewalls support IP Multicast starting with PIXOS 7.0. The FWSM supports IP Multicast starting with version 3.1. Other firewall vendors also support IP Multicast.

A firewall in the data path for market data streams will increase the overall latency. A typical firewall can take 20-70 microseconds to process the traffic. The degree to which security and filtering are deployed should be considered in this context for latency sensitive applications such as market data.

e. IPsec AH

Cisco supports IPsec AH for authentication of PIM control packets. This should be considered for security in a managed CE environment or other situations. An example of the configuration using static keys is as follows:

```
crypto ipsec transform-set pimts ah-sha-hmac
mode transport
!
crypto map pim-crypto 10 ipsec-manual
 set peer 224.0.0.13
 set session-key inbound ah 404 123456789A123456789A123456789A123456789A
```

```
 set session-key outbound ah 404 123456789A123456789A123456789A123456789A
 set transform-set pimts
 match address 106
!
interface Ethernet0/0
 crypto map pim-crypto
!
access-list 106 permit ip 0.0.0.0 255.255.255.255 host 224.0.0.13
```

The session key for a SHA algorithm must be at least 20 bytes long. The key in the example is the minimum length.

Application Issues

Number of Groups/Channels to Use

Many application developers consider using thousand of multicast groups to give them the ability to divide up products or instruments into small buckets. Normally these applications send many small messages as part of their information bus. Usually several messages are sent in each packet that are received by many users. Sending fewer messages in each packet increases the overhead necessary for each message. In the extreme case, sending only one message in each packet will quickly reach the point of diminishing returns—there will be more overhead sent then actual data.

Additionally, there is a practical limit to the number of groups that a receiver can subscribe. Previously, the limit that the NIC MAC filtering could support was 50 groups. Today it may be higher, but either way after a point, the NIC card goes into promiscuous mode and all the filtering would be done at the kernel. This may not be as efficient as dropping the packets at the driver level.

If IGMP snooping is configured on the receiver ports, then only the data that will be delivered to that port would be the groups which the receiver has subscribed. Cisco switches can filter several thousand groups on each switchport, but there is an upper limit.

Perhaps the biggest limitation would be the IGMP stack on the host. The host will need to respond to igmp queries for each group at least once per minute. When we hit thousands of groups this will be a limitation—especially when the host receives a general query and needs to respond for each group it has subscribed. If there are many hosts connected to a single switch, processing the thousands of reports from each all the hosts will be a limitation.

The application developers need to find a reasonable compromise between the number of groups and breaking up their products into logical buckets.

Let us take NASDAQ Quotation Dissemination Service (NQDS) for example. The instruments are broken up alphabetically as follows:

NQDS (A-E) 224.3.0.18

NQDS (F-N) 224.3.0.20

NQDS (O-Z) 224.3.0.22

Another example is the NASDAQ Totalview service and is broken down as follows:

Data Channel	Primary Groups	Backup Groups
NASDAQ TotalView (A)	224.0.17.32	224.0.17.35
NASDAQ TotalView (B-C)	224.0.17.48	224.0.17.49
NASDAQ TotalView (D-F)	224.0.17.50	224.0.17.51
NASDAQ TotalView (G-K)	224.0.17.52	224.0.17.53
NASDAQ TotalView (L-N)	224.0.17.54	224.0.17.55
NASDAQ TotalView (O-Q)	224.0.17.56	224.0.17.57
NASDAQ TotalView (R-S)	224.0.17.58	224.0.17.59
NASDAQ TotalView (T-Z)	224.0.17.60	224.0.17.61

This approach does allow for straight forward network/application management, but does not necessarily allow for optimized bandwidth utilization for most users. A user of NQDS that is interested in technology stocks and would like to subscribe to just CSCO and INTL, they would need to pull down all the data for the first two groups of NQDS. Understanding the way the users will be pulling down the data and then organizing into the appropriate logical groups will optimize the bandwidth for each user.

In many market data applications, optimizing the data organization would be of limited value. Typically, customers will bring in all data into a few machines and filter the instruments. Using more groups is just more overhead for the stack and will not help the customers conserve bandwidth.

Another approach might be to keep the groups down to a minimum level and use UDP port numbers to further differentiate, if necessary. The multicast streams are forwarded based on destination address, but the UDP ports can be used to aid in filtering the traffic.

The other extreme would be to use just one multicast group for the entire application and then have the end user filter the data. In some situations, this may be sufficient.

Intermittent Sources

A common issue with market data applications is servers that send data to a multicast group and then go silent for more than 3.5 minutes. These intermittent sources may cause thrashing of state on the network and can introduce packet loss during the window of time when soft state exists, and then hardware shortcuts are being created.

There are a few scenarios in which the outage can be more severe. One case would be if the source starts sending again right around the 3.5 minute mark. At that point, state has started to time out in some of the routers along the data path and there may be inconsistent state in the network. This could create a situation in which data from the source would be dropped for as long as a minute until state clears out and then is created again on the intermediate routers.

On the Cisco 6500/7600, there are some additional platform specific issues with intermittent sources. Multicast flows are forwarded by hardware shortcuts on the PFC/DFC. The statistics from these flows are maintained on the PFC/DFC and are periodically updated to the MSFC. By default this update happens every 90 seconds but can be lowered to every 10 seconds by lowering the 'mls ip multicast flow-stat-timer' down to 1. Due to this delay in receiving the latest flow stats for individual multicast streams, it is possible that a source could go quiet for 3 minutes and then start transmitting again and the mroute state will still be removed for no activity. This could cause an outage of an active stream for 1-2 minutes, depending on the state of the network.

These are the best solutions to deal with intermittent sources.

a. PIM-Bidir or PIM-SSM

The first and best solution for intermittent sources is to use PIM-Bidir for many-to-many applications and PIM-SSM for one-to-many applications.

Both of these optimizations of the PIM protocol do not have any data driven events in creating forwarding state. That means that as long as the receivers are subscribed to the streams, the network will have the forwarding state created in the hardware switching path.

Intermittent sources are not an issue with PIM-Bidir and PIM-SSM.

b. Null packets

In PIM-SM environments, a common method to make sure forwarding state is created is to send a burst of null packets to the multicast group before the actual data stream. The application needs to efficiently ignore these null data packets to make sure it doesn't affect performance. The sources would only need to send the burst of packets if they have been silent for more than 3 minutes. A good practice would be to send the burst if the source was silent for more than a minute.

Many financial applications send out an initial burst of traffic in the morning and then all well behaved sources will not have a problem.

c. Periodic keepalives or heartbeats

An alternative approach for PIM-SM environments is for sources to send periodic heartbeat messages to the multicast groups. This is a similar approach to the null packets, but the packets can be sent on a regular timer so that the forwarding state will never expire. A typical timer for the heartbeat message is 60 seconds.

d. S,G expiry timer

Finally, Cisco has made a modification to the operation of the S, G expiry timer in IOS. There is now a CLI knob to allow the state for a S, G to stay alive for hours without any traffic being sent. This fix was in response to a customer request in a PIM-SM environment to maintain the state and not fall back to *, G forwarding. The command is "ip pim sparse sg-expiry-timer" and is documented in the command reference:

http://www.cisco.com/en/US/docs/ios/ipmulti/command/reference/imc_04.html#wp1018443

This approach should be considered a workaround until PIM-Bidir or PIM-SSM is deployed or the app is fixed.

RTCP Feedback

A common issue with real time voice and video applications that use RTP is the use of RTCP feedback traffic. Unnecessary use of the feedback option can create excessive multicast state in the network. If the RTCP traffic is not required by the application, it should be avoided.

Receivers can be implemented and configured to send RTCP feedback using unicast. This has the advantage of allowing the server to still receive the feedback but not create all the multicast state.

Tibco Heartbeats

TibcoRV has had the ability to use IP Multicast for the heartbeat between the TICs for many years. However, there are some brokerage houses that are still using very old versions of TibcoRV that use UDP Broadcast support for the resiliency. This limitation is often cited as a reason to maintain

a Layer2 infrastructure between TICs located in different data centers. These older versions of TibcoRV should be phased out in favor of the IP Multicast supported versions.

Fast Producers and Slow Consumers

Today, many servers providing market data are attached at Gigabit speeds, while the receivers are attached at different speeds, usually 100Mbps. This creates the potential for receivers to drop packets and request re-transmissions, which creates more traffic the slowest consumers cannot handle, continuing the vicious circle.

The solution needs to be some type of access control in the application that will limit the amount of data that one host can request. Quality of Service (QoS) and other network functions can mitigate the problem, but ultimately the subscriptions need to be managed in the application.

Americas Headquarters
Cisco Systems, Inc
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte Ltd
Singapore

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

Printed in USA C11-474791-03 10/09