Financial Services Design for High Availability

Version History

Version Number	Date	Notes
1	March 28, 2003	This document was created.

This document describes the best practice for building a multicast trading floor. It describes the logical parts of a multicast network and suggests how to architect them for the demands of such an environment. This document does not describe any particular application or unicast routing protocol.

Note Note: Because each individual network has unique characteristics, some of the features described in this document may not be applicable.

This document has the following sections:

Designing a Multicast Trading Floor

The follow sections describe financial trading floor requirements and how to design a financial trading floor network to meet those requirements.

Trading Floor Requirements

The most critical requirement of a multicast trading floor is that it be highly available. This means that the network must be designed so that there is no single point of failure and so that the network can respond in a deterministic manner to any failure. The multicast trading floor design should also be scalable so that it can absorb the growth of any environment (within realistic expectations).

Building for the Future

Financial trading floors have traditionally been located on a single floor of a building. Now trading floors are virtual entities that span multiple floors and, in some designs, extend over multiple buildings in different locations. The designs discussed in this paper are limited to local area network (LAN) and metropolitan area networks (MAN) environments and so use only high-speed links, such as Fast Ethernet (FE) and Gigabit Ethernet (GE), to connect devices. In LAN and MAN environments using FE or GE, quality of service (QoS) is not required or recommended. For designs where a portion of the multicast trading floor is connected using a WAN, QoS may be required to perform the proper buffering and congestion control to ensure that the trading data has the highest priority.

Applications Characteristics

A number of applications are used in trading. This section describes the common aspects of the most widely utilized trading applications.

There are one or more locations in the network that house the sources of the trading data. These locations take in information from external sources and then publish the information to the traders located on the trading floor. The application used to publish the information will normally be sourced from redundant servers that communicate via a “hello” mechanism. This “hello” mechanism can be either at Layer 2 or at Layer 3.

The servers that source this data expect to receive feedback from the recipients of the trading data. This feedback mechanism, depending on application restrictions, can be on a separate multicast group or, in some circumstances, limited to the same multicast group. The characteristics of the feedback mechanism are critical because they can significantly impact the resource demand on the network infrastructure, such as memory and CPU requirements.

Multicast Protocols Suitable for Trading Design

The following sections describe the multicast protocols most suitable for financial trading floor networks.

PIM-SM

Protocol Independent Multicast Sparse Mode (PIM-SM) is the most widely deployed multicast protocol in large-scale networks, and the trading environment is no exception. PIM-SM mode offers the ability to use shortest-path trees or a mixture of shortest-path trees and shared trees. Because PIM-SM has been the most widely deployed multicast protocol, it is the protocol of choice for most trading floor architectures.

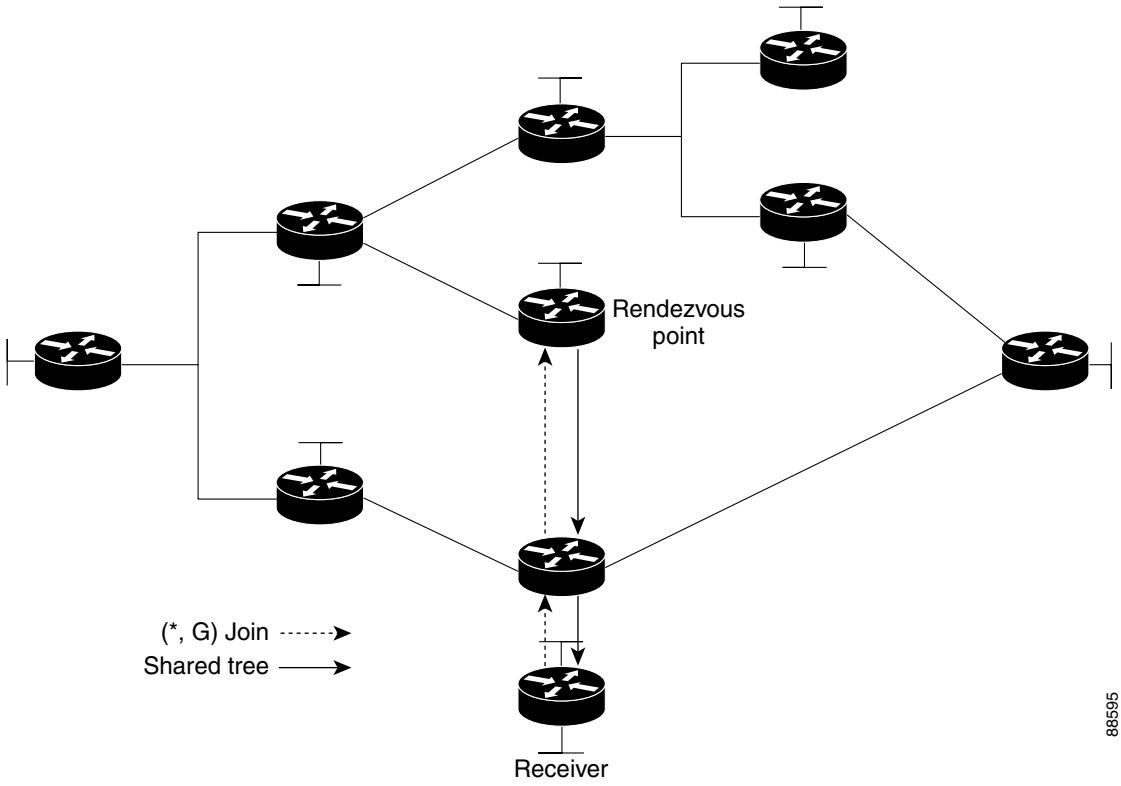
Bidirectional PIM

Bidirectional PIM is a protocol designed for applications that are "many to many" in design, of which trading applications are the primary example. Bidirectional PIM is a relatively new protocol compared with PIM-SM.

Explanation of PIM-SM

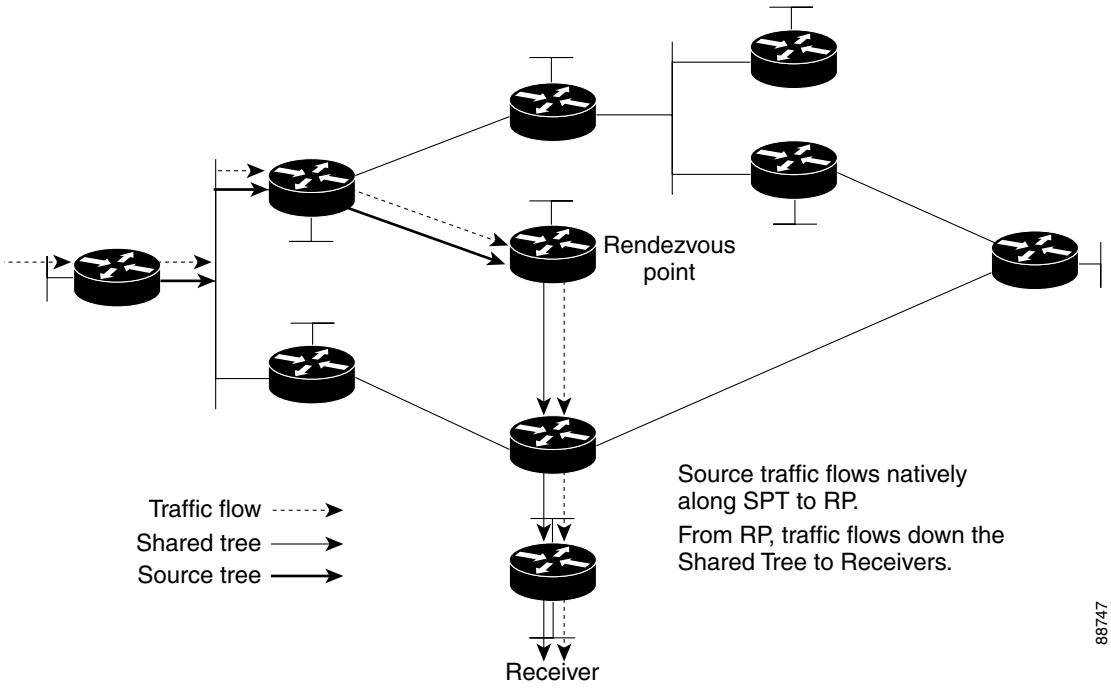
PIM-SM is a "pull" protocol in that it requires receivers to explicitly signal their desire to receive data. Interested receivers signal this desire to the nearest router using Internet Group Management Protocol (IGMP). A router on the local LAN, a designated router (DR), will then initiate a shared tree from a device called a rendezvous point (RP) toward its local receiver, as shown in Figure 1. This shared tree ensures that any active source that the RP is aware of will be forwarded toward the receiver.

Figure 1 PIM-SM Shared Tree



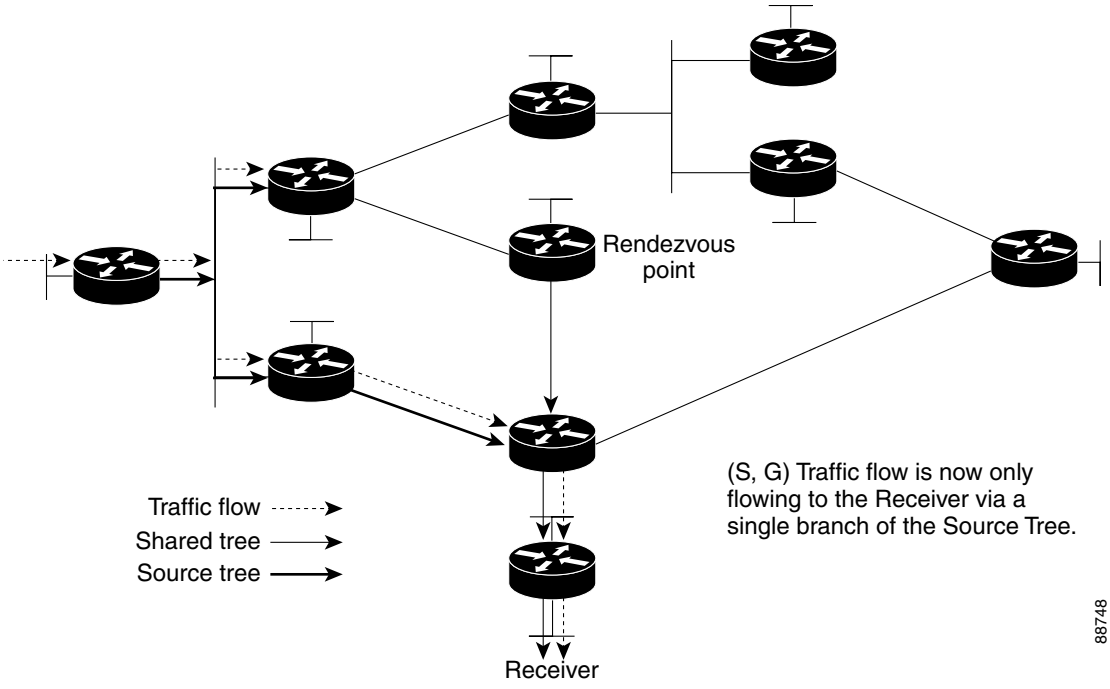
This forwarding mechanism requires that the RP be aware of all multicast sources in the network. The RP learns of multicast sources when a router on the source LAN—the DR—registers the source with the RP. Because a receiver has already registered with the RP for this particular group, the RP iniitates the creation of a source tree between the source and the RP, as shown in Figure 2.

Figure 2 Data Flow through the RP



The receiver's last-hop router, meaning the router closest to the receiving host, can, upon receipt of data packets from the RP via the shared tree, choose to initiate a shortest-path tree between the source and the receiver, as shown in Figure 3.

Figure 3 PIM-SM Shortest-Path Tree Switchover



PIM-SM allows the network administrator to choose a design that uses either the shared tree or the shortest-path tree. This design is useful in a trading floor design because the number of sources (hence, the S,G state) can be quite large, averaging about 3 to 4 thousand sources.

RP Placement

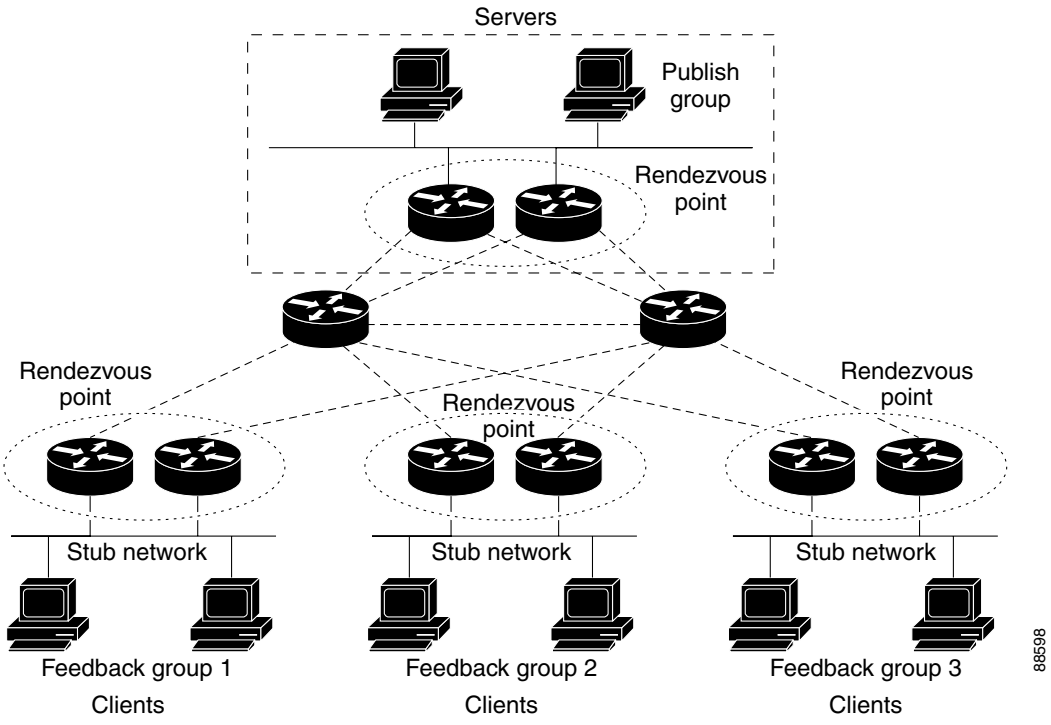
A common question in IP multicast is where to position the PIM-SM RP. The usual answer is that the position of the RP is not that critical. However, in a trading floor design, RP placement can be extremely important.

Every trading floor contends with a large amount of state because of the large number of multicast sources—every trading floor data recipient needs to send feedback to the data source. One particular shared-tree design reduces the S,G state by positioning the RP close to the sources and using the shared tree closest to the receivers. This design works best for applications that allow the receivers to send feedback using a separate group from the data publishing group.

In the past, the RP for the sources was configured on the router connected to the source LAN. This arrangement ensured that the RP would always know about any source active on that LAN. In more recent designs, where the RP is placed closest to the sources, you can use Multicast Source Discovery Protocol (MSDP), an interdomain protocol, to ensure that all RPs for a group have knowledge about each active source.

As shown in Figure 2, a source tree is always created between the source and the RP, creating S,G state on all the routers between the source and the RP—even with **spt-threshold** set to infinity. In order to minimize the state on these routers, separate RPs should be placed adjacent to the users for the feedback groups and adjacent to the servers for the publish groups, as shown, as shown in Figure 4.

Figure 4 Placement of Rendezvous Points



RP Selection

Either the RP can be statically configured on each router in the trading floor or it can be chosen dynamically using an election mechanism. The static method suffers from a single point of failure whereas the dynamic method allows a backup RP to take over when the dynamic protocol is sure that the active RP has failed. There are two protocols for the dynamic election of RP: Auto-RP and BSR.

Dynamic Methods

Auto-RP was the first dynamic method of electing an RP. Auto-RP is a Cisco-proprietary protocol that has been reverse-engineered by Cisco's competitors.

BSR is the standards body method of electing an RP. Although suitable for a standard multicast environment, BSR is unsuitable for a multicast trading floor environment, especially when compared with a static RP that does not have a single point of failure and has a failover time almost as short as interior gateway protocol (IGP) convergence time.

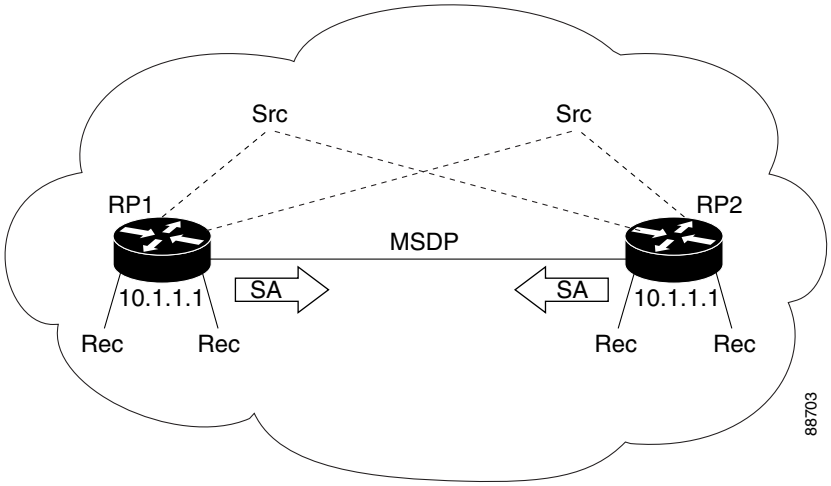
Anycast RP

Because Anycast RP has an extremely fast failover rate along with the ability to load-balance, Anycast RP is the most suitable protocol for a multicast trading floor. Anycast RP takes advantage of the fact that modern IGPs can support host addresses—that is, IP addresses with a /32 mask. If the *same* IP address is configured on two or more routers and injected into the IGP, then it appears in the unicast

routing table as multiple routes to the same device. In fact, there are multiple routes going to multiple devices. This means that the multicast router will choose the physical RP that its unicast table calculates to be the closest as determined by its routing metrics.

Anycast RP also has the ability (because of its load-balancing capability) to fulfill the newer server requirement (and perhaps that of receivers) to be located on a logical floor but perhaps in an entirely separate physical location. Another method, though, was needed to ensure that the RP for a particular group was aware of all sources. An Anycast RP address supports one logical RP and multiple physical RPs but for these physical RPs to perform their function, they need to be able to share informaton about active sources. The solution was a protocol originally designed for the sharing of active source information inside domains—Multicast Source Discovery Protocol (MSDP), as shown in Figure 5.

Figure 5 Anycast RP and MSDP



Other Design Issues

Because IP multicast operates on top of the unicast routing environment, a tuned IGP is required for the fastest failover. A tuned IGP is not only for the Anycast RP feature already described but also for the reverse path forwarding (RPF) check. In the past the multicast routes were associated with an RPF interface and this RPF interface association was calculated periodically. In the event that a unicast routing change occurred directly after a periodic check, it could take nearly five seconds before a new RPF interface was chosen. Cisco IOS software now uses triggered RPF checks for the fastest convergence.

The "Explanation of PIM-SM" section described how a router on a LAN is chosen to be the DR and is responsible for creating the shared tree to the RP for receivers. On the source LANs the DR is also responsible for registering active source with the RP. The DR is elected using PIM "hello" messages, and these messages are regulated using a query timer on each interface. Cisco has added the capability to tune this timer so that changes to the DR can occur in under one second.

Addressing

For data that remains inside the administrative domain, the recommend addressing range is to be taken from the 239 group as specified in RFC 2365, *Administratively Scoped IP Multicast*. You should not use group addresses that overlap the MAC address of link local groups (224.0.0.X).

Layer 2 Choices

Point-to-Point Connections in the Core

The core of a campus network should be based on Layer 3 point-to-point connections.

Point-to-point designs allow for faster convergence than multiaccess backbones. If routers were connected as a shared LAN through a switch and an interface were to fail, convergence would rely on elements such as routing "hello" messages to signal that a neighbor had failed. This reliance on higher protocols, rather than physical loss of signal, creates a delay in recovery. A point-to-point connection therefore allows faster failover because of its ability to detect the loss of a single neighbor immediately and so to converge in milliseconds. Point-to-point connections also allow for deterministic behavior in the event of a failure.

Etherchannel

Bundling point-to-point connections enables an Etherchannel link to continue operating despite the failure of individual links within a bundle.

IGMP Snooping

Constraining multicast is recommended. Of the available methods of constraining multicast (Cisco Group Management Protocol (CGMP) or IGMP snooping), IGMP snooping is the recommended method because it is available on modern, high-performance switches.

L2 Connection Parameters

L2 ports on switches have the ability to negotiate suitable parameters for any attached device. Because devices on the trading floor are assumed to be stable, this negotiation is not required. Options such as port speed and duplex provide faster convergence when manually configured than when the switches automatically negotiate the parameters.

Spanning Tree Issues

At the edge of the network, spanning tree can cause delays in service for end stations by placing a port in blocking mode until SPT ensures that forwarding data from this port will not create loops. To prevent this delay in service, and so have faster convergence, spanning-tree operations must be manipulated for those ports that have connected end stations. The **set spantree portfast enable** command allows spanning tree to skip the initial delay and forward data immediately.

Spanning-tree recalculation also causes the content-addressable memory (CAM) table of the switch to be cleared. This means that any CGMP or IGMP entries will be removed, and flooding will occur until they are repopulated.

PIM Protocol Timers

It is recommended that PIM timers be modified on the access router interfaces so that the fastest possible failover occurs. PIM communicates with neighbors using "hello" messages. The "hello" timer can be modified using the **ip pim query-interval** command. The modified timer is now capable of being tuned in milliseconds. For example:

```
interface <interface>
 ip pim sparse-mode
 ip pim query-interval 100 msec
! Query interval in milliseconds
! Configure the ip pim query-interval command only on access interfaces with
! redundant routers. The default value for the query interval is 30 seconds.
```

The result is that a failover of a DR on a leaf subnet occurs as quickly as possible.

For actual configuration examples from a working test network, see the *Financial Enterprise System Test for Release 12.1(12c)E1* document located at the following URL: http://www.cisco.com/univercd/cc/td/doc/solution/systest/safehbr/fnanctst.htm.

Receiver Subnet Issues

In a redundant access layer, traffic forwarded onto the subnet by one router will be received by the redundant router on a non RPF interface. This traffic must be dropped by the redundant router.

For more information about receiver subnet issues, refer to the following documents:

- *Redundant Router Issues with IP Multicast in Stub Networks* application note http://www.cisco.com/warp/public/cc/pd/iosw/prodlit/ipst_an.htm
- "Non-RPF Traffic Processing" in *Configuring IP Multicast Layer 3 Switching* http://www.cisco.com/en/US/partner/products/hw/switches/ps700/products_configuration_guide_chapter09186a008007f4a0.html#1049755

Recommendations

In brief, the following is recommended:

- Because it is the most widely deployed protocol on trading floors today, we recommend that you use PIM-SM. We also suggest that you use separate publish and feedback groups wherever possible because this allows each group to use a separate RP. Place each RP close to the source so that it stays on the shared tree, therefore minimizing state creation in the core.
- For RPs, we recommend that you use Anycast RP for resiliency and convergence. Anycast RP is capable of operating in sparse mode and converges as fast as the IGP in use converges.
- To get optimal performance from your multicast network, we suggest that the unicast environment on the trading floor be highly tuned.
- For addressing, we recommend that a private multicast address range (239.x.x.x as outlined in RFC 2365, *Administratively Scoped IP Multicast*, and best current practice RFC 3171, *IANA Guidelines for IPv4 Multicast Address Assignments*) be used.
- We recommend that deterministic configurations and topologies be used as much as possible—for example, a point-to-point topology with deterministic features (such as highest address on certain links) and the use of EtherChannel to prevent a single L2 link causing an L3 failure. We also highly

recommend configuring deterministic behavior on L2 (such as PortFast) to prevent spanning-tree delays in recovery and resetting the CAM table in a switch to clear multicast entries learned through IGMP snooping.

Related Documents

- RFC 2365, *Administratively Scoped IP Multicast*
- RFC 3161, *IANA Guidelines for IPv4 Multicast Address Assignments*
- *Financial Enterprise System Test for Release 12.1(12c)E1*
  http://www.cisco.com/univercd/cc/td/doc/solution/systest/safehbr/fnanctst.htm
- *Redundant Router Issues with IP Multicast in Stub Networks* application note
  http://www.cisco.com/warp/public/cc/pd/iosw/prodlit/ipst_an.htm
- "Non-RPF Traffic Processing" in *Configuring IP Multicast Layer 3 Switching*
  http://www.cisco.com/en/US/partner/products/hw/switches/ps700/products_configuration_guide_chapter09186a008007f4a0.html#1049755