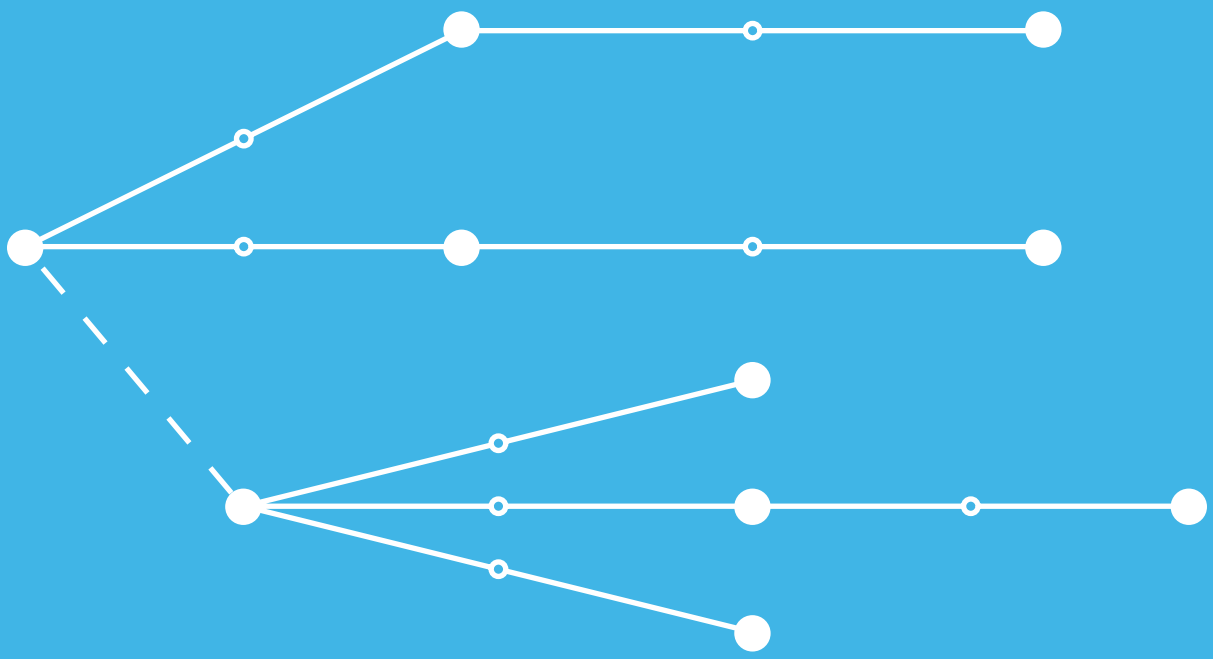
PACKET DESIGN

Understanding & Managing Multicast Routing



WHITE PAPER

Table of Contents

Introduction

Multicast enables efficient delivery of messages from a source to a set of receivers. The efficiency is achieved by the source injecting a single copy of the message to the network, and the network, only when necessary, making copies and forwarding them towards the receivers. Copies are typically made as close to the receivers as possible so that no link in the network carries multiple copies of the same message. In contrast, in unicast delivery of messages, the source would make a copy of the message for each receiver and the network would have to carry every copy. This is illustrated in Figure 1.

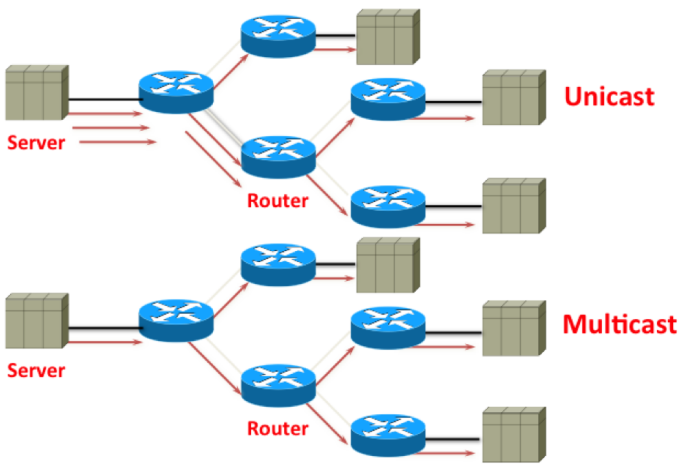


Figure 1. Multicast versus Unicast delivery

The routers and interfaces traversed by the message (original, as well as the copies) form a tree, referred to as the multicast tree. The source is the root and the receivers are the leaves of this tree. However, at times, we also refer to the router that is connected to the source as the root and the routers that are connected to the receivers as leaves of the tree. The direction of interfaces and routers towards the source is called upstream and the direction towards the receivers is called downstream. A router with more than one downstream link makes and forwards a copy of the message on each of these links.

There are two significant benefits of using multicast for communication. The first and obvious benefit is the resource efficiency gained by not carrying redundant copies of the message in the network. The second benefit is the fair distribution of the message. In unicast, redundant copies of the message would queue behind each other. With many receivers, there might be a significant delay between when the first receiver receives the messages versus when the last receiver receives the message. Multicast delivery minimizes the delay and makes reception more equitable.

Protocol Independent Multicast (PIM)

Multicast can be implemented at different layers of the networking architecture. IP Multicast is the multicast service at the Internet Protocol layer. IP packets for multicast distribution are identified by destination

IP addresses from the IP address range 224.0.0.0 to 239.255.255.255. These addresses are referred to as the multicast group addresses. IP Multicast packets are delivered over trees set up by the Protocol Independent Multicast (PIM) routing protocol.

IP Multicast has some interesting features. For example, the receivers don't have to know the sources and vice versa. PIM enables this by using a rendezvous point (RP). It can support both a single source and multiples sources sending packets to receivers. PIM can set up trees that are optimized for very dense groups (i.e. receivers at every corner of the network) as well as sparse groups.

A source that wants to send to a group would put the group's IP address as the destination IP address of its packets. The receivers that want to receive these packets would join the multicast group using the Internet Group Management Protocol (IGMP). Routers receive IGMP messages and use PIM to set up trees between the sources and receivers. For each group, PIM may set up one or more trees depending on how many sources there are and what level of efficiency is desired.

For dense groups, PIM (referred to as PIM Dense Mode or PIM-DM) uses a broadcast and prune method to set up trees. In this mode, the initial packet (and others periodically thereafter) is broadcast to all network segments. The routers at segments where there are no receivers then prune their segments from the tree. Because most multicast groups are sparse, and due to scaling issues of the broadcast and prune method, PIM-DM is not widely used.

For sparse groups, PIM Sparse Mode (PIM-SM) is used. PIM-SM can set up trees for three types of groups. The first type, Any Source Multicast (ASM), is for applications where the sources and receivers are not known a priori. In this mode, routers are either statically configured or dynamically elect a router as the Rendezvous Point (RP) for each multicast group. Receivers join the multicast tree rooted at the RP. When a source sends a packet, it is trapped by a local router, encapsulated inside a PIM Register message, and sent to the RP. The RP forwards the packet towards the receivers using the tree.

When a receiver wants to join a group, it sends an IGMP Join message to its locally attached router. The router makes sure the interface where it received the join message is added to its downstream interfaces list for this group. The router then checks if it has an upstream interface/neighbor for this group. If it does, the tree setup is completed and no further processing is necessary. If it does not have an upstream neighbor, it looks up the RP for the group, and then determines its next hop towards this RP. This next hop is the router's upstream neighbor and the interface to it is the router's upstream interface. The router then sends a PIM Join message to its upstream neighbor and the upstream router repeats the procedure. Eventually, the join message is sent either to a router that is already on the tree or to the RP itself. This completes the tree set up. Notice that the tree is set to the reverse of the shortest paths from the receivers to the RP. This is referred to as "*reverse path forwarding*" (RPF) and the upstream router as the RPF neighbor. This initial tree is called a "*shared tree*" and can be used by any source that wants to send to the multicast group. To indicate joining the shared tree, the join messages contain (*, G) as the tree identifier, where G is the group IP address.

If a receiver wants to leave a tree, it sends an IGMP Leave message for this group to its local router. The router then prunes this interface from its outgoing interfaces list (assuming this was the last receiver on that interface). If there are no downstream interfaces left in the list, the router will send a PIM prune message for (*, G) to its upstream router. The upstream router repeats the procedure.

The shared tree may not be optimum for any particular source as it is optimized towards the RP. This can be problematic if the RP and the sources are apart, for example when the RP is on a different continent than the source. If one of the sources starts sending a lot of packets, the receivers may want to switch to a shortest path tree (SPT) for this source. In this case, they follow a similar join procedure as above. However, the join messages now contain (S, G) where S is the IP address of the specific source; and the upstream routers are now chosen to be the next hop routers towards S instead of towards RP. Once a router starts receiving packets on the SPT, it prunes itself from the (*, G) tree for this source so that it stops receiving packets from the shared tree. It is possible some routers will switch to SPT and some routers will not; or a router may receive packets from some sources on the SPT and from some other sources on the shared tree. Note that the RP always switches to SPT towards any source after receiving the first packet. This is to avoid the overhead of PIM Register message encapsulation.

The second type of tree that PIM-SM sets up is for Source Specific Multicast (SSM) which is used by applications where the sources are known a priori. When the source is known, SPT join and prune procedures towards the source are used to set up the tree for this source. Note that in this case the tree is optimum for the source and cannot be shared.

In both ASM and SSM, the number of multicast trees is proportional to the number of sources. This introduces a scaling problem as the numbers of multicast groups and sources increase. To address this issue, PIM-SM can set up trees of a third type, called bidirectional trees (referred to as PIM-BIDIR). These trees are similar to shared trees in that all sources and receivers are on a single shared tree, but without the disadvantage of PIM Register message encapsulation. However, the scaling problems are not completely solved, as the number of trees is still proportional to the number of multicast groups. In addition, the trees on this mode are not optimum for any particular source; hence bandwidth savings are not maximized.

Multicast Applications

Because of scaling issues, large scale IP multicast is typically reserved for very critical and high bandwidth applications such as broadcast TV (IPTV), financial market data distribution, large content (e.g. video) distribution, and town hall style corporate meetings. All of these applications would consume a lot more bandwidth without multicast. Broadcast TV and financial data distribution also cannot tolerate lack of fairness. Without fairness, for example, in a soccer game on broadcast TV, one TV set may show a goal while on another TV set the player has not yet kicked the ball; or in the financial markets, one trader may be disadvantaged over another due to delayed market data.

IP Multicast is also used exclusively inside individual local area network segments for resource discovery (e.g. printers, air play devices). But in these cases, the multicast trees are trivial. This is called link local multicast and is outside the scope of this white paper.

Challenges of Running Multicast Applications

In unicast, IP addresses are often associated with specific assets such as desktops, laptops, phones, tablets and printers. Unfortunately, this visibility is lost with multicast addresses. Ideally, multicast addresses should be allocated by the network operator, but in reality, many network operators do not know what multicast addresses are used in the network and what they are used for. This is because applications can randomly pick a multicast address and start using it. One network operator for a financial company estimates there are sixteen thousand multicast groups in his network and aims to reduce it to less than four thousand groups. This requires visibility into existing multicast groups in the network, as well as the sources, the receivers and the trees for each.

Configuring multicast on a device is easy; however configuring multicast network-wide remains challenging. To configure multicast on an interface, the network operator needs to enable PIM on all routers on that interface. However, multicast traffic can be high volume which can be problematic on a low capacity interface. If such an interface is not enabled for multicast, what happens if that interface is the upstream interface towards an RP or a source? This is called RPF failure and breaks multicast. To work around this situation, operators can install multicast-only static routes that bypass such interfaces for tree setup purposes. The challenge is a network with no RPF failures may start having RPF failures under link failures. Even static routes may not completely prevent RPF failures since they cannot accommodate all combinations of link failures.

Turning on multicast on all interfaces, regardless of capacity, is also problematic. Another financial network operator had one of its high volume multicast groups sent over a low capacity link. This particular application used negative acknowledgements to request any missing data. Because the volume of data was more than the capacity of a link, the receivers downstream of the link requested the missing packets. The source multicast these missing packets a second time, increasing the volume of the multicast traffic. As a result, these receivers missed and requested even more packets. This caused a total network melt down as the traffic volume for this group continued to increase and eventually exceeded the capacity of every interface in the network.

Group to RP mappings are often statically configured. Configuring them differently on routers can lead to partitioned multicast trees and can cause some receivers to never receive the data.

Multicast trees and, in particular, the receivers, are very dynamic. A multicast group that is working properly may break after new receivers join the group. PIM also reacts to underlying unicast routing changes (that affect the choice of upstream router). This can take a branch of the tree from a router and completely graft it to some other router. Such a change would be hard to diagnose without visibility into underlying unicast routing.

Because typical multicast applications are very critical and very high volume, when multicast problems occur they cause extreme damage. For example, a short outage of financial data distribution may lead to millions of dollars lost in trading, and an outage of IPTV distribution may lead to subscriber churn. For this reason, there is need for better multicast management tools.

Multicast Explorer

The Multicast Explorer module of the Route Explorer™ system addresses these management challenges. It discovers all multicast groups, sources, trees and receivers in the network, tracks all the changes in near real time, issues alerts upon detecting anomalies, and maintains a full log of changes--multicast events in Multicast Explorer terminology. Using a DVR-like replay facility, Multicast Explorer can show multicast state at any time in the past and how a multicast tree evolved over time, and even show the evolution using a network topology animation. It analyzes events for patterns, anomalies, and returns the root cause of multicast problems. Multicast Explorer's modeling capabilities allow additions and changes to be made to groups, sources and receivers, as well as changes to the underlying unicast routing, to analyze the impact on multicast. When combined with traffic flow data using the Traffic Explorer™ route-flow fusion technology, engineers have the most comprehensive multicast capacity-planning and modeling tool in the industry.

Collection

Collecting multicast routing analytics information is challenging as PIM is a hop-by-hop protocol and a router only knows its upstream and downstream interfaces and neighbors. Contrast this to IGPs such as OSPF and IS-IS, for example, where routing changes are communicated to all the routers in the network, making passive collection by Route Explorer very convenient. For this reason, multicast collection is done by accessing each router in three phases as depicted in Figure 2.

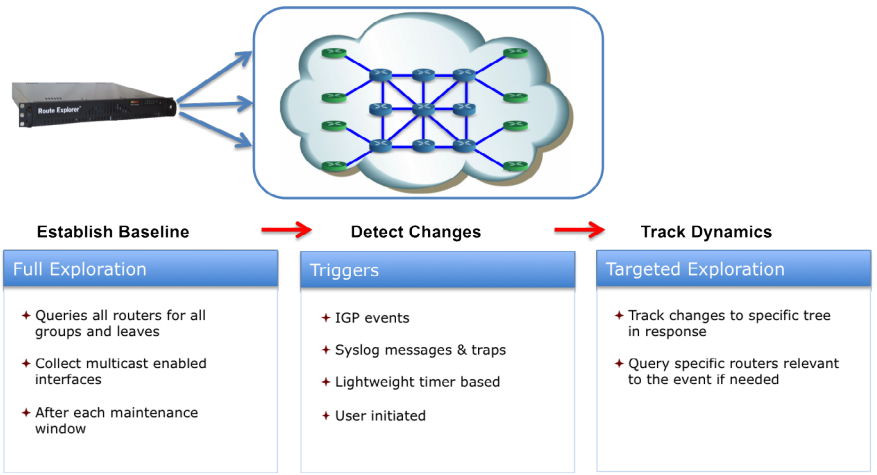


Figure 2. Multicast Explorer collection architecture

The first phase of the collection is called full exploration. Full exploration is scheduled by the user; it should be run after each maintenance window, usually a few times a week for most networks. In full exploration, for each router in the network, Multicast Explorer collects both multicast configuration information, such as PIM enabled interfaces, group to RP mappings and multicast static routes, as well as multicast state information, such as the set of multicast groups, sources, receivers, upstream and downstream

interfaces and neighbors, and traffic volumes. This phase is so named because Multicast Explorer queries every router in the network. In other phases, Multicast Explorer only queries relevant routers and it may skip querying the configuration information.

The second phase is the detection phase. In this phase, Multicast Explorer looks for changes to the multicast state information both passively and actively. Passive indications include SNMP traps and syslog messages from routers as well as underlying routing changes from IGPs and BGP. For example, if a link goes down in IGP topology, Multicast Explorer deduces that all the trees using that link will have to re-converge. Active indications include periodic lightweight queries to detect new sources, receivers or groups in the network. The user configures a lightweight query interval, typically in minutes, and Multicast Explorer divides this interval by the number of routers in the network to determine the interval for querying each router. For example, in a network of 1000 routers and a 15 minute lightweight query interval, Multicast Explorer will poll the network almost every second (900 seconds divided by 1000 routers), each time querying a different router. The larger the change, the faster it will be detected. The goal of the second phase is to get an indication of a change; hence the queries made are lightweight.

The third phase is called the targeted exploration phase. After an indication from the second phase, Multicast Explorer knows that there is a change in the network and where to look for it. For example, when a link goes down in IGP, Multicast Explorer knows what downstream routers need to re-converge. Or when lightweight exploration detects a new group, it knows the immediate upstream and downstream interfaces and neighbors of that router for this new group. From this information, Multicast Explorer recursively queries upstream and downstream routers one by one until it discovers the new state of the network. Note that, even though the lightweight query interval is in minutes, once a change is detected, the rest of the information will be queried back to back (and in parallel when possible).

Multicast Explorer uses router access to query multicast information. The access method varies by vendor platform. For Cisco IOS and IOS-XR routers, it logs in to issue Show commands and captures the output; for Juniper JunOS routers, it uses NETCONF; for Alcatel routers, it uses SNMP; and for Huawei routers it issues Display commands and captures the output. A common access method across all platforms would be preferable, but this is not possible because NETCONF is not widely available and SNMP MIBs are either incomplete or burden the routers with unacceptable amount of load.

Reports

Once the collection completes, several reports are generated. As an example, the list of multicast groups discovered is shown in Figure 3. The sources for each group are listed hierarchically under the group. For each entry, the RP, number of routers and edge routers in the tree, whether the source is active and how much traffic it is sending are listed. These traffic statistics are obtained from the routers during router access and do not require the Traffic Explorer module to be implemented.

Multicast Groups 2013-07-12 10:30:52 PDT

Drill Down | Group Address

Group	Attributes	Routers	Edge Routers	Source Status	Traffic	Last Source Inactive Time
(*, 224.110.21.1)	RP: 10.120.1.21	13	6	Active	1.00	
(*, 224.110.22.1)	RP: 10.120.1.22	13	6	Active	1.00	
(*, 224.110.1.1)	RP: 10.120.1.1	11	7	Active	1.00	
(10.120.1.2, 224.110.1.1)		11	6	Active	1.00	
(10.120.1.5, 224.110.1.1)		11	3	Active	1.00	
(10.120.1.6, 224.110.1.1)		11	3	Active	1.00	
(10.120.1.1, 224.110.1.1)		10	3	Active	1.00	
(10.120.1.7, 224.110.1.1)		9	4	Active	1.00	
(10.120.1.9, 224.110.1.1)		9	4	Active	1.00	
(10.120.1.3, 224.110.1.1)		8	3	Active	1.00	
(10.120.1.4, 224.110.1.1)		8	3	Active	1.00	
(10.120.1.8, 224.110.1.1)		8	6	Active	1.00	
(*, 224.130.1.39)	RP: 10.120.1.18	10	8	Active	1.00	
(10.64.6.3, 224.130.1.39)		10	8	Active	1.00	
(10.64.6.100, 224.130.1.39)		10	8	Inactive		2013-07-12 10:30:18
(10.64.6.101, 224.130.1.39)		10	8	Active	1.00	
(10.64.16.3, 224.130.1.39)		10	8	Active	1.00	
(*, 224.110.3.1)	RP: 10.120.1.3	9	7	Active	1.00	
(*, 224.110.4.1)	RP: 10.120.1.4	9	7	Active	1.00	
(*, 224.110.5.1)	RP: 10.120.1.5	9	6	Active	1.00	
(*, 224.110.6.1)	RP: 10.120.1.6	9	6	Active	1.00	
(*, 224.110.11.1)	RP: 10.120.1.11	9	6	Active	1.00	

40 top level entries, 432 other entries

Figure 3. List of multicast groups and sources

From here, it is possible to drill down to the routers and view the multicast tree in textual hierarchical format (see Figure 4). Next to each router, the attribute indicates whether that router is the RP for the group, if it has any source or receiver hosts attached to it, along with its upstream and downstream interfaces.

Multicast Groups → List Routers 2013-07-12 10:30:52 PDT

Group	Attributes	Routers	Edge Routers	Source Status	Traffic	Last Source Inactive Time
(10.120.1.7, 224.110.1.1)		9	4	Active	1.00	

Drill Down | Name

Router			Upstream			Downstream		
Name	Address	Attributes	Neighbor Add	Neighbor Interface	Local Interface	Ne	int	Traffic
(10.120.1.7, 224.110.1.1)								
LA-PE-3825-R7	10.120.1.7	Local Receivers				1	1	1.00
LA-P-7204-R3	10.120.1.3		10.64.6.7	Gi0/1	Gi0/3	1	1	1.00
SJ-P-7204-R4	10.120.1.4		10.64.5.3	Gi0/2	Gi0/2	2	2	1.00
SJ-PE-2811-R9	10.120.1.9	Local Receivers	10.64.12.4	Fa1/0	Fa0/0	0	0	
SJ-P-7204-R2	10.120.1.2		10.74.4.4	Fa2/0	Fa4/0	1	1	1.00
SJ-PE-2811-R8	10.120.1.8	Local Receivers	10.64.11.2	Fa1/0	Fa0/0	0	0	1.00
LA-P-7204-R1	10.120.1.1	RP				1	3	
NY-PPE-NE40E-R20	10.120.1.20		10.74.26.1	Gi0/2	GigabitEthernet1/0/3	1	1	
NY-7204VXR-R21	10.120.1.21	Local Receivers	10.74.28.20	GigabitEthernet1/0/5	Fa0/0	0	0	

Figure 4. Routers on the tree for (10.120.1.7, 224.110.1.1)

It is also possible to display the multicast trees graphically (see Figure 5). The router icons indicate if the router is connected to the source (arrows going outward from the icon), if it is the RP (arrows coming into the icon) and if it has local receivers (a leaf on the icon). The four routers in the top section of the figure are all on the shortest path tree for this source. The routers in the bottom section of the figure are receiving multicast packets from the shared tree. The dotted line between the RP and the source indicates that the RP is receiving the data packets in PIM Register messages. This is not ideal; at a minimum the RP should have joined the shortest path tree for the source. Here, this was not possible because one of the interfaces along that path did not have multicast enabled.

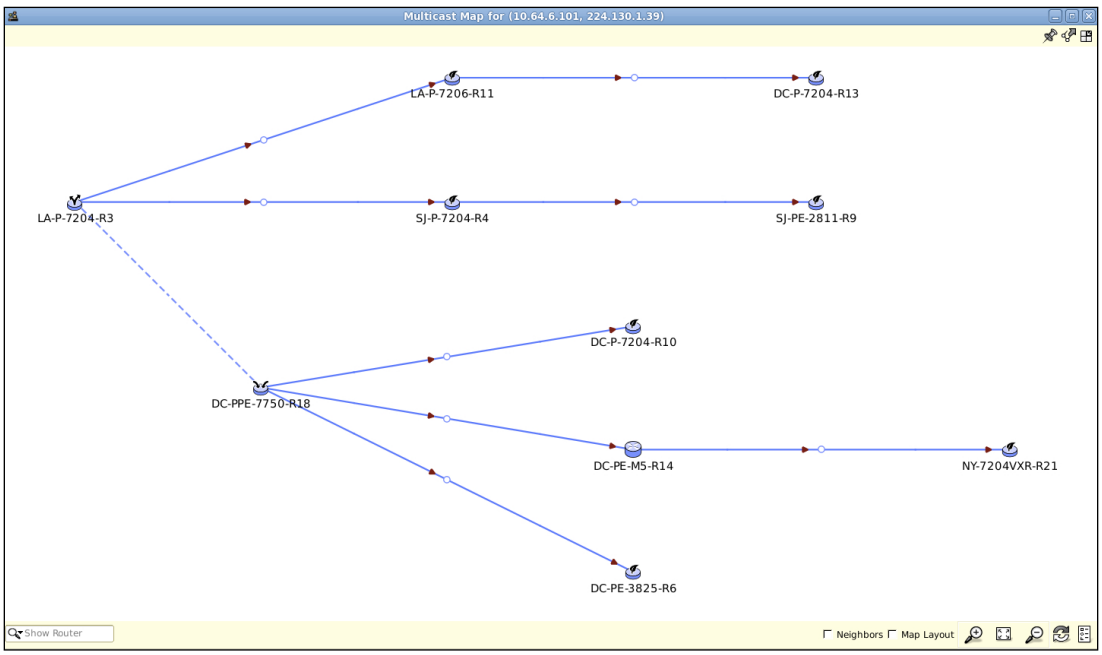


Figure 5. Multicast tree map visualization

In addition to several multicast inventory reports, Multicast Explorer looks for common configuration mistakes and includes anomaly reports as follows:

- **Stuck in Register State:** lists the groups and sources that are still using PIM Register messages to pass the multicast packets from the source to the RP (see Figure 6);
- **RP Mismatch:** lists group to RP mapping configurations that are different in at least two routers;
- **No PIM Enabled Interfaces:** lists IGP enabled interfaces that do not have PIM enabled;
- **Source with no Receivers:** lists the groups that have a source but no receiver;
- **Missing PIM Adjacency:** lists the PIM enabled interfaces that have no other router on them.

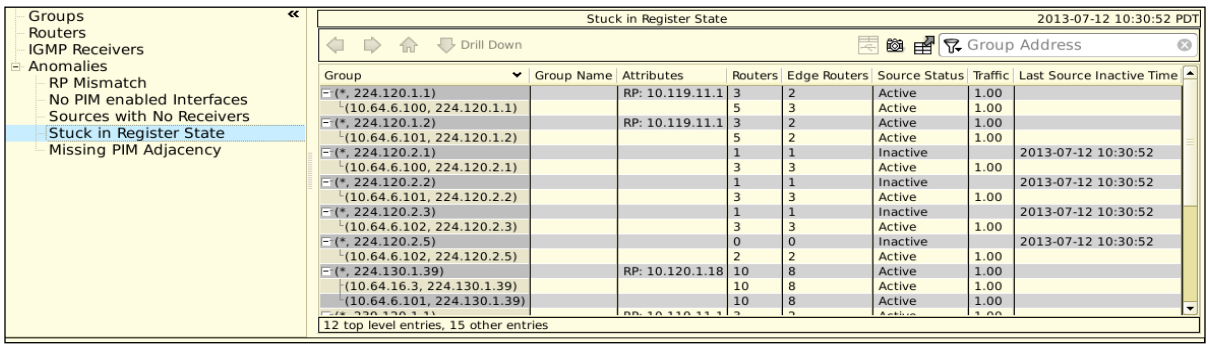


Figure 6. Stuck in Register State Anomaly Report

Alerting

These anomalies are also detected in real time and SNMP trap, syslog or email-based alerts are issued automatically. In addition, alerts are issued when the following conditions are detected:

- **Unapproved Groups:** a group in the "black list" (or not in the "white list") is seen;
- **RPF Failure:** a router does not have an RPF neighbor towards the RP for shared trees or towards the source for SPTs;
- **Path Change:** the path from a source to a receiver changes;
- **Disappearing Source:** a configured source becomes inactive;
- **Disappearing Router:** a specified router is no longer on the tree;
- **Interface Capacity Exceeded:** multicast traffic volume is more than the capacity of any interface on the tree; this particular alert warns of the network melt down mentioned earlier in this paper;
- **High (or Low) Traffic Volume:** traffic reported for the group by any router is more (or less) than the configured threshold;
- **Adjacency State:** a PIM adjacency between two routers is lost.

Examining Dynamics, Stability and Comparison

As the Route Explorer system, including the Multicast Explorer module, detects changes, it writes an event for each change to its database. Figure 7 shows the multicast events recorded during the first week of May 2013 in ten-minute intervals. As can be seen, there may be thousands of events. Recording these events serves two purposes. First, they enable Multicast Explorer to present multicast state at any time in the past. For example, if there was a problem in one of the multicast groups yesterday, it is possible to "re-wind" the database's network time and examine the network as it was yesterday.

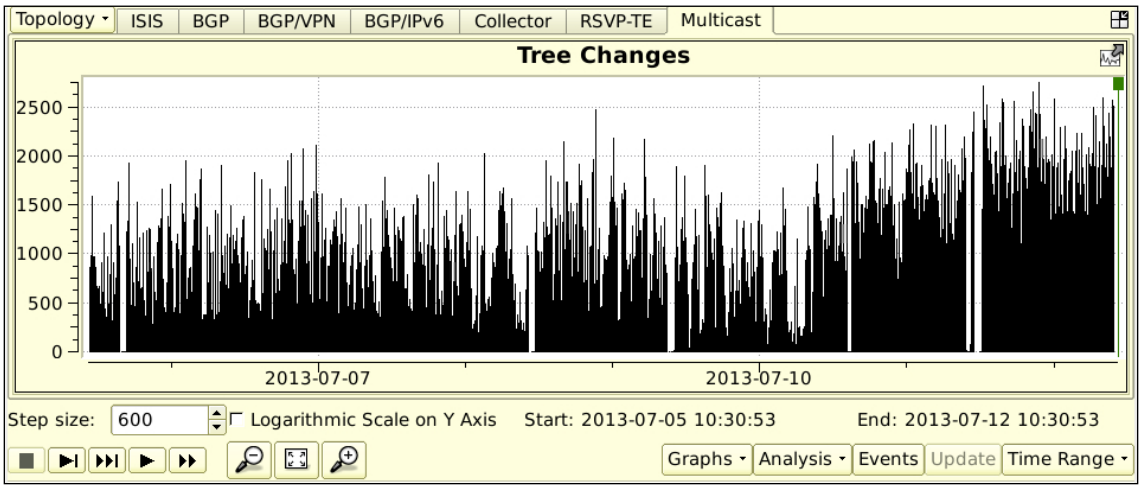


Figure 7. Multicast events over time

Secondly, these events can be analyzed to assess multicast dynamics and stability in the network: What groups had the most churn? Were those groups adapting to routing changes or were their user groups very dynamic? How did the multicast tree evolve over time? How is the tree different when multicast was working from when it was broken? Answers to all of these questions are extremely valuable for isolating the root cause of multicast problems in the network quickly.

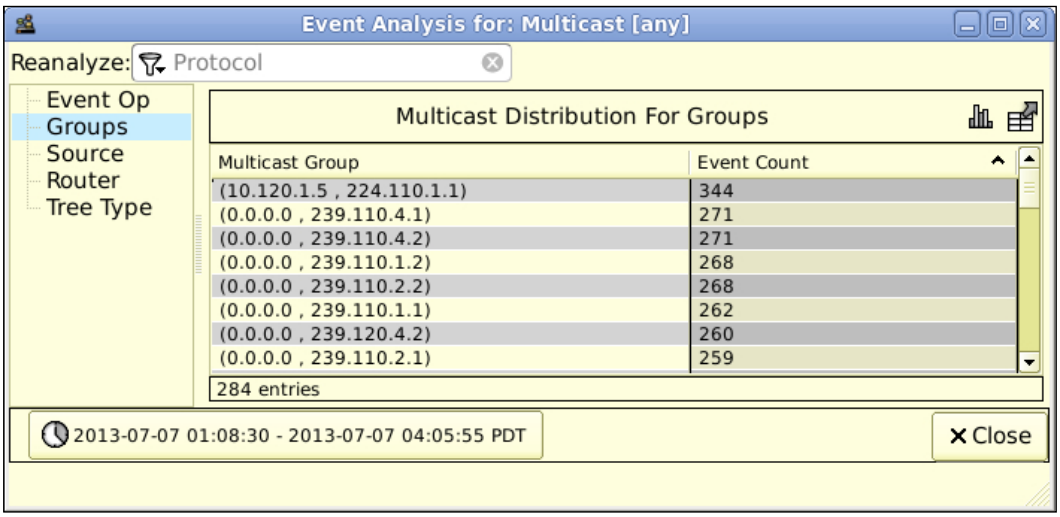


Figure 8. Events broken by multicast groups

Figure 8 shows events from 01:08:30 to 04:05:55 on July 7th, 2013 broken down by each multicast group. For example, the group (10.120.1.5, 224.110.1.1) has 344 events. These events can be examined for very detailed forensics (see Figure 9).

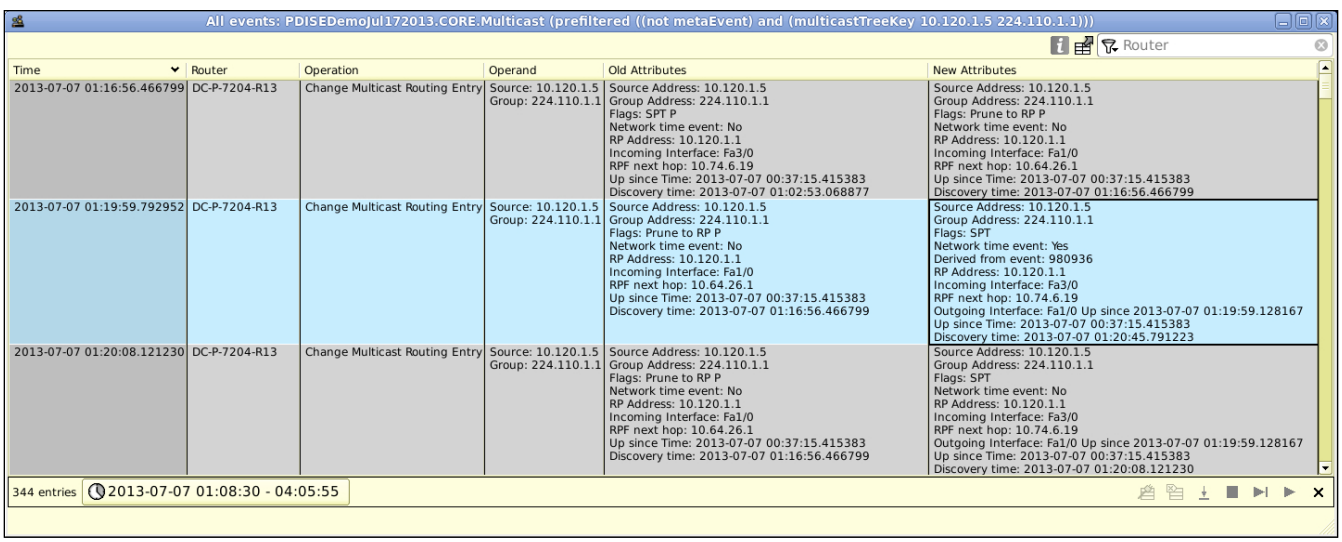


Figure 9. Multicast events

The Multicast Explorer module also includes a powerful analysis feature that examines events for patterns and can report root causes of these events (see Figure 10). Root cause incidents include creation of a brand new group, addition (or removal) of a source and/or receivers, and adaptation to underlying routing changes. For example, when a link goes down, after examining potentially thousands of events, a single incident is created indicating the link that failed and the groups that were affected. Events related to an incident can be animated to show visually the impact of the incident.

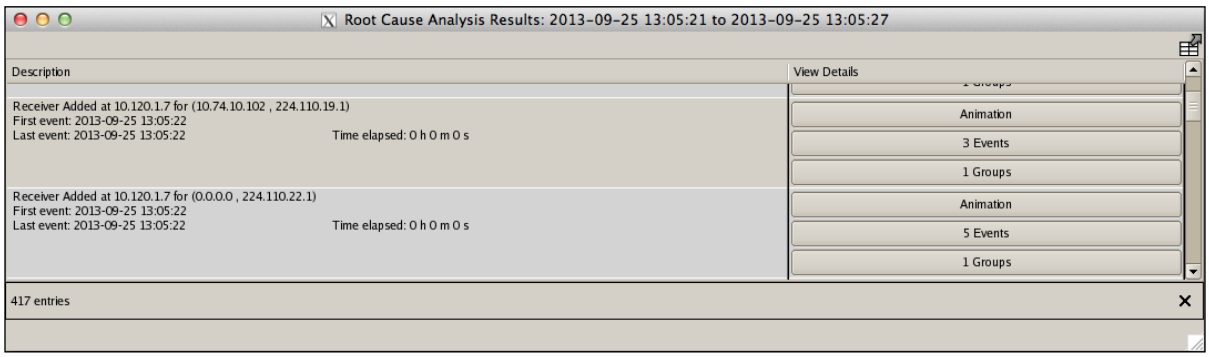


Figure 10. Multicast problem root cause analysis

Planning

In addition to showing the current and past state of multicast, Multicast Explorer also can be used for planning future multicast changes. It is possible to model new multicast groups from scratch by specifying the group address, sources, receivers and amount of expected traffic. Multicast Explorer computes the necessary trees and reports, for example, whether there would be any RPF failures. It is possible to modify existing groups as well. For example, the location of sources can be moved or new receivers can be added.

As well as changing multicast parameters, the underlying IGP and BGP routing can be changed to see the groups that would be impacted. For example, a link can be taken down and Multicast Explorer will compute new trees for each affected group.

Traffic Explorer

Route Explorer's flow-based volumetric Traffic Explorer module also supports multicast. Traffic Explorer provides more detailed traffic statistics than those in Multicast Explorer which are based on statistics collected from the routers. As an example, Traffic Explorer can report on traffic for a particular application port number.

Additionally, because Traffic Explorer is flow-based, it can show where traffic will move as underlying unicast and multicast routing changes. For example, when a link is taken down, Multicast Explorer can show the new trees, but more powerfully, Traffic Explorer shows if moving these multicast flows to new links causes any congestion in the network and what services/users are impacted.

Concluding Remarks

The Multicast Explorer module of the Route Explorer system provides near real-time visibility into multicast groups, trees, sources and receivers in the network. It contains powerful reports and analysis that help solve hard-to-diagnose multicast problems in minutes. It also helps planning for future multicast expansion in the network.

Some network management tools attempt to monitor multicast. These tools typically use SNMP polling. SNMP MIBs often lack sufficient information for useful diagnostics. Besides, these tools are generic object management systems without deep understanding of the semantics of the objects (they would treat multicast only slightly differently than a printer). They do not understand underlying IGP and BGP routing, nor the interaction between multicast and these protocols. In our opinion, this last point is key to a valuable multicast management product.

Route Explorer supports multicast trees built by PIM Sparse Mode, for both ASM and SSM trees. Anycast RPs and sources are supported and the monitoring of branches set up by the Multicast Source Discovery Protocol (MSDP) to connect multiple trees is also supported. While Multicast Explorer is commonly used for monitoring IPTV and financial market data distribution applications, many other uses are supported, such as for monitoring PIM multicast distribution trees (MDTs) used in multicast VPNs (both draft-rosen as well as BGP multicast VPNs).

To learn more about Packet Design and Route Explorer, please:

- Email us at info@packetdesign.com
- Visit Packet Design's web site at www.packetdesign.com
- Call us at +1.408.490.1000

Corporate Headquarters

Packet Design
2455 Augustine Drive
Santa Clara, CA 95054
Phone: 408.490.1000
Fax: 408.562.0080