Using IP Multicast with VMware® ESX 3.5

VMware ESX 3.5

IP multicast is a popular protocol implemented in many applications for simultaneously and efficiently delivering information to multiple destinations. Multicast sources send single copies of information over the network and let the network take responsibility for replicating and forwarding the information to multiple recipients.

IP multicast is one of three commonly used methods of forwarding packets on an IP network. These methods are:

- **Unicast**—One-to-one packet delivery from a source to a single destination. This is the most common and familiar form of traffic, with the destination IP address representing a single unique destination.

- **Broadcast**—Packet delivery to all nodes within an IP network. The destination address for an IP broadcast is all binary ones in the host portion of the destination IP address.

- **Multicast**—Packet delivery to a group of destinations denoted by a multicast IP address. Membership of this group is coordinated through the Internet Group Management Protocol (IGMP).

A fourth method, called anycast, is rarely used.

This paper provides an overview of IP multicast and support for this protocol in VMware ESX 3.5. It covers the following topics:

- “IP Multicast Address Space” on page 1
- “Layer 2 Multicast Addressing” on page 2
- “IP Multicast Groups” on page 2
- “IGMP Versions” on page 2
- “Operation of Multicast on a Physical Network” on page 3
- “Operation of Multicast on an ESX Virtual Switch” on page 3
- “Considerations for VMotion” on page 4
- “NIC Teaming Considerations” on page 6
- “IP Multicast Parameters on ESX” on page 7
- “Physical Switch and Router Considerations” on page 8

**IP Multicast Address Space**

IP multicast destinations are represented by a group IP address from the old Class D IP address space between 224.0.0.0 and 239.255.255.255. Each individual address specifies a multicast group.

The IP multicast address space is broken into three broad address ranges. Well-known addresses within these address ranges are administered by the Internet Assigned Numbers Authority.
Layer 2 Multicast Addressing

An IP multicast address is a Layer 3 IP address. In order to receive IP multicast packets, the NIC must be programmed to accept traffic destined to the multicast MAC address that correspond to a given multicast IP address. A multicast MAC address is created by setting the broadcast-multicast bit (bit 0, octet 0) in the destination 48-bit MAC address.

The Internet Assigned Numbers Authority has allocated the first half of the block of Ethernet MAC addresses starting with 01:00:5E for multicast addresses. This creates the range of available Ethernet multicast MAC addresses from 01:00:5E:00:00:00 through 01:00:5E:7F:FF:FF. This allocation allows for 23 bits in the Ethernet address to correspond to the IP multicast group address. The mapping places the lower 23 bits of the IP multicast group address into these available 23 bits in the Ethernet address. Because the upper five bits of the IP multicast address are dropped in the mapping from IP address to MAC address, the resulting Ethernet multicast address is not unique to a single IP multicast group address. In fact, 32 different multicast group IDs all map to the same Ethernet address.

An example:

Consider an IP Multicast group address of 224.1.5.6. The low order 23 bits of this IP address are x0000001:00000101:00000110 in binary format or 010506 in hexadecimal format. The resulting multicast MAC address is 01:00:5E:01:05:06.

Because only the lower 23 bits are mapped, the address is not unique. The IP multicast group addresses of 224.129.5.6, 225.1.5.6, 225.129.5.6, and so on map to the same multicast MAC address. In other words, multiple IP multicast groups can use the same multicast MAC addresses.

IP Multicast Groups

IP multicast group addresses can be dynamically or statically assigned to applications. Hosts use IGMP to register dynamically for a multicast group. Once registered, the host receives traffic sent to that multicast group. To keep multicast group memberships up-to-date and prevent forwarding traffic unnecessarily, multicast routers periodically query hosts for their interest in receiving multicast traffic using IGMP membership query messages. The frequency of the IGMP membership query messages is a tunable parameter in most multicast routers. By default, it is typically sent every 60 seconds. Hosts respond to these queries with IGMP membership report messages that list the multicast groups for which they wish to receive traffic. If there is no response to three consecutive IGMP membership query packets for a particular group, the multicast router stops forwarding traffic to that group.

IGMP Versions

Three versions of IGMP are defined by the IETF—versions 1, 2, and 3.

- IGMP V3—Defined in RFC 3376 (October 2002). Most widely implemented and default in current operating system products (for example, Windows Server 2003 and Red Hat Enterprise Linux 5).
Note that IGMP V2 and V3 differ in the structure of the IGMP membership reports. A guest operating system using IGMP V2 responds with a separate IGMP V2 membership report for each group to which it has subscribed. A guest operating system using IGMP V3 bundles the individual groups as vectors in a single IGMP V3 membership report. IGMP V3 is more economical than V2 and preferred in environments with a large number of group memberships.

**Operation of Multicast on a Physical Network**

Physical Layer 2 switches limit the flooding of multicast traffic out all ports through a mechanism called IGMP snooping. IGMP snooping involves the switch listening for IGMP control information. It uses this information to determine which switch ports are interested in receiving multicast packets.

In place of IGMP snooping, some Cisco switches and routers allow the optional use of Cisco Group Management Protocol (CGMP) between the switch and a multicast router. With CGMP, the multicast router explicitly tells switches which ports belong to which multicast groups.

The selection of IGMP snooping or CGMP in the physical switches is transparent to ESX.

IP multicast operation on ESX is independent of the Layer 3 IP multicast routing protocol. It does not matter whether PIM (dense or sparse), DVMRP, or any other protocol is used.

**Operation of Multicast on an ESX Virtual Switch**

IP multicast is fully supported by VMware ESX. The implementation and operation of multicast, however, differs slightly from the implementation and operation in a physical switch.

A virtual switch (vSwitch) does not need to perform IGMP snooping to learn which virtual machines have enabled IP multicast. ESX dynamically learns multicast membership because it has authoritative knowledge of the attached virtual NICs (vNIC). When a vNIC attached to a virtual machine is configured for multicast, the vSwitch learns the multicast Ethernet group addresses associated with the virtual machine. When the virtual machine joins an IP multicast group, the virtual machine first converts the group to an Ethernet multicast group based on the IP address. The virtual machine then programs its own NIC filter on the vNIC to subscribe to the multicast traffic. The vNIC passes the multicast registration information down to the vSwitch through the hypervisor to update the multicast tables on the vSwitch and enable forwarding of frames for that IP multicast group to that virtual machine.

*Figure 1* shows an overview of the IP multicast message flow. The virtual machines use IGMP to join and leave multicast groups. Adjacent multicast routers send periodic membership queries to a well known link local multicast address (224.0.0.1) on the local network. ESX allows these to pass through to the virtual machines. Virtual machines with multicast subscriptions respond to the multicast router with one or more IGMP membership reports detailing the group memberships to which they are subscribed. Layer 2 switches in the network path use IGMP snooping to learn which interfaces require forwarding of multicast group traffic. *Figure 2* shows a network trace of the IGMP message flow between a virtual machine and the adjacent multicast router.
Figure 1. Overview of IGMP Message Flows in an ESX Environment

Figure 2. Trace Showing Periodic IGMP V2 Membership Queries from Multicast Router and IGMP Membership Reports from Virtual Machine

<table>
<thead>
<tr>
<th>No.</th>
<th>Time</th>
<th>Source</th>
<th>Destination</th>
<th>Protocol</th>
<th>Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>22:27:35</td>
<td>192.168.20.1</td>
<td>224.0.0.1</td>
<td>IGMP</td>
<td>V2 Membership Query, General</td>
</tr>
<tr>
<td>35</td>
<td>22:27:36</td>
<td>192.168.20.1</td>
<td>224.0.0.1</td>
<td>IGMP</td>
<td>V2 Membership Report / Join group 239.1.1.1</td>
</tr>
<tr>
<td>110</td>
<td>22:28:72</td>
<td>192.168.20.1</td>
<td>224.0.0.1</td>
<td>IGMP</td>
<td>V2 Membership Query, General</td>
</tr>
<tr>
<td>111</td>
<td>22:28:73</td>
<td>192.168.20.1</td>
<td>224.0.0.1</td>
<td>IGMP</td>
<td>V2 Membership Query, General</td>
</tr>
<tr>
<td>112</td>
<td>22:29:06</td>
<td>192.168.20.1</td>
<td>224.0.0.1</td>
<td>IGMP</td>
<td>V2 Membership Report / Join group 239.1.1.1</td>
</tr>
<tr>
<td>130</td>
<td>22:31:46</td>
<td>192.168.20.1</td>
<td>224.0.0.1</td>
<td>IGMP</td>
<td>V2 Membership Query, General</td>
</tr>
<tr>
<td>193</td>
<td>22:38:30</td>
<td>192.168.20.1</td>
<td>224.0.0.1</td>
<td>IGMP</td>
<td>V2 Membership Report / Join group 239.1.1.1</td>
</tr>
</tbody>
</table>

Considerations for VMotion

When a virtual machine that subscribes to multicast traffic moved to a different physical host using VMotion, the multicast vNIC information is retained with the virtual machine as it moves. The destination ESX host immediately knows the multicast group memberships of that virtual machine. If the vSwitch uplink in the destination ESX host is already receiving multicast traffic for that group (because of membership by another virtual machine), the multicast traffic flow continues to that virtual machine. What happens from that point depends upon the ESX release level.

VMotion with Releases Up To and Including ESX 3.5 Update 1

If the destination ESX host is not receiving multicast traffic for the desired group, it must wait for the periodic IGMP group membership query from the multicast router and the corresponding IGMP response from the virtual machine. Using IGMP snooping, the physical switches learn the new location of the virtual machine and forward multicast traffic to the corresponding host vSwitch. The default time for the periodic IGMP membership query from the multicast router is 60 seconds. This means a virtual machine might have to wait as long as 60 seconds after VMotion to resume receiving multicast traffic.
For ESX 3.5 Update 2 and Later Releases

In ESX 3.5 Update 2 and later releases, ESX accelerates the convergence time of multicast traffic using the following procedure:

1. As soon as VMotion has moved the virtual machine to the destination host, ESX sends a fake IGMP membership query to the virtual machine to solicit group membership information.
2. If the virtual machine is a member of any multicast groups, it immediately responds with an IGMP membership response detailing its multicast group memberships.
3. Using IGMP snooping, the physical switches update their multicast membership tables and forward multicast traffic for the matching multicast groups to the ESX host.

Figure 3 shows the trace summary of the fake queries and the resulting responses from the virtual machine. By default, ESX sends two IGMP V3 membership queries using a source IP address of 10.0.0.0 and a source MAC address of 00:00:00:00:00:00. The procedure for changing the IGMP version, number of queries, and source IP address is described later in this document.

Detailed trace entries for the IGMP V2 and V3 membership queries are shown in Figure 4 and Figure 5. The corresponding IGMP V2 and V3 membership reports are shown in Figure 6 and Figure 7.

Figure 3. Trace Entries Showing Fake IGMP V2 Membership Queries Sent by ESX to the Virtual Machine after VMotion with Corresponding Responses from the Virtual Machine

Figure 4. Trace Entry Detail of a Fake IGMP V2 Membership Query Sent by ESX to the Virtual Machine after VMotion

Figure 5. Trace Entry Detail of a Fake IGMP V3 Membership Query Sent by ESX to the Virtual Machine after VMotion
NIC Teaming Considerations

All ESX NIC teaming algorithms are supported with IP multicast. However, failures of members within a NIC team can interrupt delivery of multicast traffic.

The following sections describe some special considerations that apply to each algorithm under failover situations.

IP-Hash

A NIC team using IP hash requires a corresponding Layer 2 port channel or EtherChannel on the adjacent physical switch—or multiple switches if some form of multichassis EtherChannel (MEC) is used. An EtherChannel is a single logical link comprising multiple physical links. As long as at least one link in the team is still active, failures do not affect the flow of multicast traffic other than the possible loss of some packets that are in transit when the failure occurs.

Originating Virtual Port ID

Originating virtual port ID is the most common and best practice method of NIC teaming for virtual machines. Under originating virtual port ID, each vNIC is mapped to a specific vmnic (physical NIC) in the team for both outgoing and incoming traffic. This mapping applies to all external unicast, multicast, and broadcast traffic.

If a failure occurs on a link within a NIC team, the virtual machines that were mapped to the failing vmnic are reallocated to remaining active vmnics within the team. If Notify Switches (the default) is selected in the policy definition for that NIC team, ESX generates a reverse ARP (RARP) to notify the physical switches of the new location (link or vmnic) of the MAC address for the virtual machine’s vNIC and enable immediate resumption of unicast traffic to the virtual machine.

However, this does not update the delivery of multicast traffic to the ESX host from the physical switch. The physical switch updates its multicast tables and continues delivery only when, through IGMP snooping, it detects an IGMP membership report from the virtual machine to the network. Virtual machines respond with IGMP membership reports when solicited by a periodic IGMP membership query from an adjacent multicast router. In most multicast router implementations, IGMP queries are sent every 60 seconds by default.
In ESX 3.5 Update 1 and prior release, virtual machines might have to wait as long as 60 seconds for multicast traffic to resume. To reduce the maximum wait time, follow the procedure in “Physical Switch and Router Considerations” on page 8.

**NOTE** In some circumstances multicast traffic resumes immediately after failure of a member in the NIC team. This occurs if, after failover, the virtual machine is remapped to another vmnic in the NIC team already receiving multicast traffic for that multicast group—that is, another virtual machine was receiving that multicast group traffic through that vmnic.

In ESX 3.5 Update 2 and later releases, the ESX vSwitch sends a fake IGMP membership query to the affected virtual machines immediately after a NIC teaming failure event. This query solicits an IGMP membership report from the virtual machines to the multicast router and updates any intermediate switches with IGMP snooping so multicast traffic can resume immediately on the new or changed member of the NIC team. The fake IGMP process is the same as that described in the section on VMotion.

**MAC Hash**

Under MAC hash, each MAC address associated with vNICs is mapped to a specific vmnic (physical NIC) for outgoing and incoming traffic. The same multicast considerations apply for MAC hash as for originating virtual source port ID, described above.

**Explicit Failover Order**

Under explicit failover order NIC teaming, vmnics in the standby list are changed to active after a failure is detected on any of the active adapters. The same multicast considerations apply for explicit failover order as for originating virtual source port ID, described above.

**IP Multicast Parameters on ESX**

ESX allows you to change some of the IP multicast parameters to suit your environment. The parameters apply to all virtual machines on the affected ESX host.

The panel shown below in Figure 8 is available from a VMware Infrastructure Client with direct connectivity to the ESX host or via vCenter. The procedure is as follows:

1. Select **Inventory > Hosts and Clusters**.
2. Select the **Configuration** tab.
3. Select **Advanced Settings** in the Software panel. The Advanced Settings panel appears.
4. Select **Net** from the list and scroll to the bottom.
5. Make required changes and click **OK**.

Changes are immediate and do not require a reboot of the ESX host.
You can set the following parameters:

- **IGMP Version**—Default is IGMP V3. This parameter determines whether IGMP V2 or IGMP V3 membership queries are sent to the virtual machine after it is moved with VMotion. Virtual machines with an IGMP V2 implementation might ignore IGMP V3 queries. However, IGMP V3 hosts should respond to IGMP V2 queries. This is an ESX host parameter and thus affects all virtual machines on that host. If you have some virtual machines with V2 and some with V3, set this parameter to the lowest common denominator of V2. Otherwise, leave this at V3.

- **IGMP Queries**—Default is 2. This specifies how many IGMP V2 or V3 membership queries are sent to the virtual machine immediately after it is moved with VMotion to solicit the IGMP membership reports. Two queries are sent by default to reduce impact of any isolated packet drops.

- **IGMP Router Address**—Default is 10.0.0.0. This is the source IP address for the fake IGMP membership queries sent by ESX to the virtual machine after it is moved with VMotion. Leave this as is unless the virtual machine has some additional filtering to exclude packets from this address.

### Physical Switch and Router Considerations

Most multicast routers send V2 or V3 membership queries on multicast-enabled networks every 60 seconds. When using ESX 3.5 Update 1 and earlier releases, multicast traffic flow to a virtual machine can be interrupted after either of the following events:

- Movement with VMotion
- NIC failure when using one of originating virtual port ID, MAC hash, or explicit failover as a NIC teaming policy.

After one of these events, multicast traffic flow resumes only after the virtual machine responds to the periodic IGMP membership query with an IGMP membership report. Under default conditions, this could take as long as 60 seconds.

To reduce the convergence time, consider the following changes:

- Reduce the IGMP query interval on the multicast router for that particular network.
- Reduce the IGMP query maximum response time. This parameter indicates the maximum tolerated response time for an IGMP membership report after an IGMP membership query from a multicast router.
An extract from a Cisco Catalyst 4948 IOS configuration is shown below. This example shows an IGMP query interval of 10 seconds and maximum response time of 1 second.

interface Vlan20
  ip address 192.168.20.1 255.255.255.0
  ip pim sparse-mode
  ip igmp query-max-response-time 1
  ip igmp query-interval 10

Altering the query interval is not necessary for ESX 3.5 Update 2 and later releases because the ESX vSwitch sends fake IGMP membership queries immediately after a movement with VMotion or a NIC failure event to solicit an IGMP membership report and minimize the multicast traffic convergence time.

Troubleshooting from the Physical Switch

Multicast-capable physical switches typically offer a number of commands from the command line interface showing the status of multicast on the switch. Consult your switch documentation for the available command set.

For example, the show ip igmp snooping statistics command on a Cisco Catalyst 6500 yields the following output showing which VLANs and switch interfaces are enabled for multicast traffic.

Cisco-6509-A#show ip igmp snooping statistics

Snooping statistics for Vlan3000
#channels: 2
#hosts : 6
Source/Group                  Interface        Reporter         Uptime   Last-Join   Last-Leave
0.0.0.0/239.255.255.253      Vl3000:Gi3/2    10.115.172.104  02:29:58  00:01:37       -
0.0.0.0/239.255.255.253      Vl3000:Gi3/2    10.115.172.105  02:30:56  00:00:54       -
0.0.0.0/239.255.255.250      Vl3000:Te7/4    10.115.172.137  02:30:56  00:00:54       -
0.0.0.0/239.255.255.253      Vl3000:Te7/4    10.115.172.115  02:31:00  00:00:53       -
0.0.0.0/239.255.255.253      Vl3000:Po2      10.115.172.117  02:31:00  00:00:53       -
0.0.0.0/239.255.255.253      Vl3000:Po11     10.115.172.100  02:31:02  00:00:54       -

Snooping statistics for Vlan3001
#channels: 4
#hosts : 12
Source/Group                  Interface        Reporter         Uptime   Last-Join   Last-Leave
0.0.0.0/224.0.6.127           Vl3001:Te7/3    10.115.173.26   02:30:54  00:00:00       -
0.0.0.0/239.255.255.253      Vl3001:Te7/3    10.115.173.22   02:28:00  00:01:07       -
0.0.0.0/239.255.255.253      Vl3001:Te7/3    10.115.173.23   02:26:00  00:00:57       -
0.0.0.0/239.255.255.253      Vl3001:Te7/3    10.115.173.26   02:30:00  00:02:32       -
0.0.0.0/239.255.255.253      Vl3001:Te7/3    10.115.173.37   02:31:01  00:01:49       -
0.0.0.0/239.255.255.253      Vl3001:Po2      10.115.173.14   02:29:02  00:01:43       -
0.0.0.0/239.255.255.253      Vl3001:Po2      10.115.173.35   02:31:01  00:02:25       -
0.0.0.0/239.255.255.253      Vl3001:Po2      10.115.173.36   02:28:04  00:00:52       -
0.0.0.0/239.239.239.239      Vl3001:Po11     10.115.173.175  02:30:59  02:22:54       -
0.0.0.0/239.255.255.253      Vl3001:Po11     10.115.173.16   02:28:57  00:01:02       -
0.0.0.0/239.255.255.253      Vl3001:Po11     10.115.173.32   02:31:03  00:00:00       -
0.0.0.0/239.255.255.253      Vl3001:Po11     10.115.173.103  02:25:01  00:00:48       -
0.0.0.0/224.0.1.40           Vl3001:         10.115.173.2   02:31:54  00:00:55       -