T. Pusateri
Juniper Networks
INTERNET DRAFT
Obsoletes: RFC 1075
draft-ietf-idmr-dvmrp-v3-04
February 1997
Expires: August 1997

Distance Vector Multicast Routing Protocol

Status of this Memo

Abstract

DVMRP is an Internet routing protocol that provides an efficient mechanism for connection-less datagram delivery to a group of hosts across an internetwork. It is a distributed protocol that dynamically generates IP Multicast delivery trees using a technique called Reverse Path Multicasting (RPM) [Deer90]. This document is an update to Version 1 of the protocol specified in RFC 1075 [Wait88].

1. Introduction

DVMRP uses a distance vector distributed routing algorithm in order for each router to determine the distance from itself to any IP Multicast traffic source. By determining the best path back to a source, a router can know which interface it should expect traffic from that source to arrive on. A good introduction to distance vector routing can be found in [Perl92]. The application of distance vector routing to multicast tree formulation is described in [Deer91].

1.1. Reverse Path Multicasting

Datagrams follow multicast delivery trees from a source to all members of a multicast group [Deer89], replicating the packet only at necessary branches in the delivery tree. The trees are calculated and updated dynamically to track the membership of individual groups. When a datagram arrives on an interface, the reverse path to the source of the datagram is determined by examining a unicast routing table of known source networks. If the datagram arrives on an interface that would be used to transmit unicast datagrams back to the source, then it is forwarded to the appropriate list of downstream interfaces. Otherwise, it is not on the optimal delivery tree and should be discarded. In this way duplicate packets can be filtered when loops exist in the network topology. The source specific delivery trees are automatically pruned back as group membership changes or leaf routers determine that no group members are present. This keeps the delivery trees to the minimum branches necessary to reach all of the group members. New sections of the tree can also be added dynamically as new members join the multicast group by grafting the new sections onto the delivery trees.

1.2. IP-IP Tunnels

Because not all IP routers support native multicast routing, DVMRP includes direct support for tunneling IP Multicast datagrams through routers. The IP Multicast datagrams are encapsulated in unicast IP packets and addressed to the routers that do support native multicast routing. DVMRP treats tunnel interfaces in an identical manner to physical network interfaces.

In previous implementations, DVMRP protocol messages were sent un-encapsulated to the unicast tunnel endpoint address. While this was more direct, it increased the complexity of firewall configuration. Therefore, all DVMRP protocol messages sent to tunnel endpoint addresses should now be encapsulated in IP protocol 4 packets just as multicast data packets are encapsulated. See Appendix C for backward compatibility issues. More information on encapsulated tunnels can be found in [Perk96].

1.3. Document Overview

Section 2 provides an overview of the protocol and the different message types exchanged by DVMRP routers. Those who wish to gain a general understanding of the protocol but are not interested in the more precise details may wish to only read this section. Section 3 explains the detailed operation of the protocol to accommodate developers needing to provide inter-operable implementations. Included in Appendix A, is a summary of the DVMRP parameters. A section on DVMRP support for tracing and troubleshooting is the topic of Appendix B. Finally, a short DVMRP version compatibility section is provided in Appendix C to assist with backward compatibility issues.

2. Protocol Overview

DVMRP can be summarized as a "broadcast & prune" multicast routing protocol. It performs Reverse Path Forwarding checks to determine when multicast traffic should be forwarded to downstream interfaces. In this way, source-rooted shortest path trees can be formed to reach all group members from each source network of multicast traffic.

2.1. Neighbor Discovery

Neighbor DVMRP routers can be discovered dynamically by sending Neighbor Probe Messages on local multicast capable network interfaces and tunnel pseudo interfaces. These messages are sent periodically to the All-DVMRP-Routers IP Multicast group address. This address falls into the range of IP Multicast addresses that are to remain on the locally attached IP network and therefore are not forwarded by multicast routers.

Each Neighbor Probe message should contain the list of Neighbor DVMRP routers for which Neighbor Probe messages have been received. In this way, Neighbor DVMRP routers can ensure that they are seen by each other. Care must be taken to inter-operate with older implementations of DVMRP that do not include this list of neighbors. It can be assumed that older implementations of DVMRP will safely ignore this list of neighbors in the Probe message. Therefore, it is not necessary to send both old and new types of Neighbor Probes.

2.2. Source Location

When an IP Multicast datagram is received by a router running DVMRP, it first looks up the source network in the DVMRP routing table. The interface of the next hop of packets sent back to the source of the datagram is called the upstream interface. If the datagram arrived on the correct upstream interface, then it is a candidate for forwarding to one or more downstream interfaces. If the datagram did not arrive on the anticipated upstream interface, it is discarded. This check is known as a reverse path forwarding check and must be performed by all DVMRP routers.

In order to ensure that all DVMRP routers have a consistent view of the unicast path back to a source, a unicast routing table is propagated to all DVMRP routers as an integral part of the protocol. Each router advertises the network number and mask of the interfaces it is directly connected to as well as relaying the routes received from neighbor routers. DVMRP requires an interface metric to be configured on all physical and tunnel interfaces. When a route is received, the metric of the upstream interface over which the datagram was received must be added to the metric of the route being propagated. This adjusted metric should be computed before the route is compared to the metric of the current next hop gateway.

Although there is certainly additional overhead associated with propagating a separate unicast routing table, it does provide two nice features. First, since all DVMRP routers are using the same unicast routing protocol, there are no inconsistencies between routers when determining the upstream interface (aside from normal convergence issues related to distance vector routing protocols). By placing the burden of synchronization on the protocol as opposed to the network manager, DVMRP reduces the risk of creating routing loops or black holes due to disagreement between neighbor routers on the upstream interface.

Second, by propagating its own unicast routing table, DVMRP makes it convenient to have separate paths for unicast vs. multicast datagrams. Although, ideally, many network managers would prefer to keep their unicast and multicast traffic aligned, tunneled multicast topologies may prevent this causing the unicast and multicast paths to diverge. Additionally, service providers may prefer to keep the unicast and multicast traffic separate for routing policy reasons as they experiment with IP multicast routing and begin to offer it as a service.

2.3. Dependent Downstream Routers

In addition to providing a consistent view of source networks, the exchange of unicast routes in DVMRP provides one other important feature. DVMRP uses the unicast route exchange as a mechanism for upstream routers to determine if any downstream routers depend on them for forwarding from particular source networks. DVMRP accomplishes this by using a technique called "Poison Reverse". If a downstream router selects an upstream router as the best next hop to a particular source network, this is indicated by echoing back the route to the upstream router with a metric equal to the original metric plus infinity. When the upstream router receives the report and sees a metric that lies between infinity and twice infinity, it can then add the downstream router from which it received the report to a list of dependent routers for this source.

This list of dependent routers per source network built by the "Poison Reverse" technique will provide the foundation necessary to determine when it is appropriate to prune back the IP source specific multicast trees.

2.4. Multi-access Networks

When two or more DVMRP routers are connected to a multi-access network, it is possible for duplicate packets to be forwarded on the network (one copy from each router). DVMRP does not require a special mechanism to prevent duplication. Instead, this feature is a consequence of the unicast route exchange. When two routers on a multi-access network exchange source networks, each of the routers will know the others metric back to each source network. Therefore, of all the DVMRP routers on a shared network, the router with the lowest metric to a source network is responsible for forwarding data on to the shared network. If two or more routers have an equally low metric, the router with the lowest IP address becomes the designated forwarder for the network.

2.5. Building Multicast Trees

As previously mentioned, when an IP multicast datagram arrives, the upstream interface is determined by looking up the interface that would be used if a datagram was being sent back to the source of the datagram. If the upstream interface is correct, then a DVMRP router will forward the datagram to a list of downstream interfaces.

2.5.1. Adding Leaf Networks

Initially, the DVMRP router must consider all of the remaining IP multicast capable interfaces (including tunnels) on the router. If the downstream interface under consideration is a leaf network (has no dependent downstream neighbors for the source network), then the IGMP local group database must be consulted. DVMRP routers can easily determine if a directly attached network is a leaf network by keeping a list of all routers from which DVMRP Router Probe messages have been received on the interface. Obviously, it is necessary to refresh this list and age out entries received from routers that are no longer being refreshed. The IGMP local group database is maintained by an elected IP multicast router on each physical, multicast capable network. The details of the election procedure are discussed in [Fen96a]. If the destination group address is listed in the local group database, and the router is the designated forwarder for the network, then the interface should be included in the list of downstream interfaces. If there are no group members on the interface, then the interface can be removed from the outgoing interface list.

2.5.2. Adding Non-Leaf Networks

Initially, all non-leaf networks should be included in the downstream interface list when a forwarding cache entry is first being created. This allows all downstream routers to be aware of traffic destined for a particular (source, group) pair. The downstream routers will then have the option to send prunes and grafts for this (source, group) pair as requirements change from their respective downstream routers and local group members.

2.6. Pruning Multicast Trees

As mentioned above, routers at the edges with leaf networks will remove their leaf interfaces that have no group members associated with an IP multicast datagram. If a router removes all of its downstream interfaces, it can notify the upstream router that it no longer wants traffic destined for a particular (source, group) pair. This is accomplished by sending a DVMRP Prune message upstream to the router it expects to forward datagrams from a particular source.

Recall that a downstream router will inform an upstream router that it depends on the upstream router to receive datagrams from particular source networks by using the "Poison Reverse" technique during the exchange of unicast routes. This method allows the upstream router to build a list of downstream routers on each interface that are dependent upon it for datagrams from a particular source network. If the upstream router receives prune messages from each one of the dependent downstream routers on an interface, then the upstream router can in turn remove this interface from its downstream interface list. If the upstream router is able to remove all of its downstream interfaces in this way, it can then send a DVMRP Prune message to its upstream router. This continues until the unneeded branches are removed from the delivery tree.

In order to remove old prune state information for (source, group) pairs that are no longer active, it is necessary to limit the life of a prune and periodically resume the flooding procedure. Inside the prune message is a prune lifetime. This indicates the length of time that the prune should remain in effect. When the prune lifetime expires, the interface is joined back onto the multicast delivery tree. If unwanted multicast datagrams continue to arrive, the prune mechanism will be re-initiated and the cycle will continue. If all of the downstream interfaces are removed from a multicast delivery tree causing a DVMRP Prune message to be sent upstream, the lifetime of the prune sent will be equal to the minimum of the remaining prune lifetimes of the downstream interfaces.

2.7. Grafting Multicast Trees

Once a tree branch has been pruned from a multicast delivery tree, packets from the corresponding (source, group) pair will no longer be forwarded. However, since IP multicast supports dynamic group membership, new hosts may join the multicast group. In this case, DVMRP routers use Grafts to undo the prunes that are in place from the host back on to the multicast delivery tree. A router will send a Graft message to its upstream neighbor if a group join occurs for a group that the router has previously sent a prune. Separate Graft messages must be sent to the appropriate upstream neighbor for each source that has been pruned. Since there would be no way to tell if a Graft message sent upstream was lost or the source simply quit sending traffic, it is necessary to acknowledge each Graft message with a DVMRP Graft Ack message. If an acknowledgment is not received within a Graft Time-out period, the Graft message should be retransmitted. Duplicate Graft Ack messages should simply be ignored. The purpose of the Graft Ack message is to simply acknowledge the receipt of a Graft message. It does not imply that any action was taken as a result of receiving the Graft message. Therefore, all Graft messages should be acknowledged whether or not they cause an action on the receiving router.

3. Detailed Protocol Operation

This section contains a detailed description of DVMRP. It covers sending and receiving of DVMRP messages as well as the generation and maintenance of IP Multicast forwarding cache entries.

3.1. Protocol Header

DVMRP packets are encapsulated in IP datagrams, with an IP protocol number of 2 (IGMP) as specified in the Assigned Numbers RFC [Reyn94]. All fields are transmitted in Network Byte Order. DVMRP packets use a common protocol header that specifies the IGMP [Fen96a] Packet Type as hexadecimal 0x13 (DVMRP). A diagram of the common protocol header follows:

<table>
<tr><td>0</td><td>8</td><td>16</td><td>31</td></tr>
<tr><td>Type
(0x13)</td><td>Code</td><td colspan="2">Checksum</td></tr>
<tr><td colspan="2">Reserved</td><td>Minor
Version</td><td>Major
Version</td></tr>
</table>

Figure 1 - Common Protocol Header

A Major Version of 3 and a Minor Version of 0xFF should be used to indicate compliance with this specification. The value of the Code field determines the DVMRP packet type. Currently, there are codes allocated for DVMRP protocol message types as well as protocol analysis and troubleshooting packets. The protocol message Codes are:

Code	Packet Type	Description
1	DVMRP Probe	for neighbor discovery
2	DVMRP Report	for unicast route exchange
7	DVMRP Prune	for pruning multicast delivery trees
8	DVMRP Graft	for grafting multicast delivery trees
9	DVMRP Graft Ack	for acknowledging graft messages

Table 1 - Standard Protocol Packet Types

There are additional codes used for protocol analysis and troubleshooting. These codes are discussed in Appendix B. The Checksum is the 16-bit one's complement of the one's complement sum of the DVMRP message. The checksum of the DVMRP message should be calculated with the checksum field set to zero.

3.2. Probe Messages

When a DVMRP router is configured to run on an interface (physical or tunnel), it sends local IP Multicast discovery packets to inform other DVMRP routers that it is operational. These discovery packets are called DVMRP Probes and they serve three purposes.

1. Probes provide a mechanism for DVMRP routers to locate each other. DVMRP sends a list of detected neighbors in the Probe message. This list of DVMRP neighbors provides a foundation for the dependent downstream neighbor list. If no DVMRP neighbors are found, the network is considered to be a leaf network. A DVMRP router should discard all other protocol packets from a neighbor until it has seen its own address in the neighbors Probe list. (See Appendix C for exceptions.)

2. Probes provide a way for DVMRP routers to determine the capabilities of each other. This may be deduced from the major and minor version numbers in the Probe packet or directly from the capability flags. These flags were first introduced to allow optional protocol features. This specification now mandates the use of Generation Id's and pruning and, therefore, provides no optional capabilities. Other capability flags were used for tracing and troubleshooting and are no longer a part of the actual protocol.

3. Probes provide a keep-alive function in order to quickly detect neighbor loss. DVMRP probes sent on each multicast capable interface configured for DVMRP SHOULD have an interval of 10 seconds. The neighbor time-out interval SHOULD be set at 35 seconds. This allows fairly early detection of a lost neighbor yet provides tolerance for busy multicast routers. These values MUST be coordinated between all DVMRP routers on a physical network segment.

3.2.1. Router Capabilities

In the past, there have been many versions of DVMRP in use with a wide range of capabilities. Practical considerations require a current implementation to inter-operate with these older implementations that don't formally specify their capabilities and are not compliant with this specification. For instance, for major versions less than 3, it can be assumed that the neighbor does not support pruning. The formal capability flags were first introduced in an well known implementation (Mrouted version 3.5) in an attempt to take the guess work out which features are supported by a neighbor. Many of these flags are no longer necessary since they are now a required part of the protocol, however, special consideration is necessary to not confuse older implementations that expect these flags to be set. Appendix C was written to assist with these and other backward compatibility issues.

Three of the flags were used for actual protocol operation. The other two assigned flags were used for troubleshooting purposes which should be documented in a separate specification. All of the bits marked "U" in the Figure below are now unused. They may be defined in the future and MUST be set to 0. Bit position 0 is the LEAF bit which is a current research topic. It MUST be set to 0. Bit positions 1, 2, and 3 MUST be set to 1 for backward compatibility. They were used to specify the PRUNE, GENID, and MTRACE bits. The first two, PRUNE and GENID, are now required features. The MTRACE bit must be set so existing implementations will not assume this neighbor does not support multicast trace-route [Fen96b]. However, since this bit is now reserved and set to 1, newer implementations should not use this bit in the Probe message to determine if multicast trace-route is supported by a neighbor. Instead, the M bit should only be used in a Neighbors2 message as described in Appendix B. The bit marked S stands for SNMP capable. This bit is used by troubleshooting applications and should only be tested in the Neighbors2 message.

0	8	9	10	S	M	G	P	L
Reserved	U	U	U	U	1	1	1	L

Figure 2 - Probe Capability Flags

3.2.2. Generation ID

If a DVMRP router is restarted, it will want to learn all of the routes known by its neighbors without having to wait for an entire report interval to pass. In order for the neighbor to detect that the router has restarted, a non-decreasing number is placed in the periodic probe message called the generation ID. When a neighbor detects an increase in the generation ID of a router, it should re-send its entire unicast routing table to the router.

If a change in generation ID is detected, any prune information received from the router is no longer valid and should be flushed. If this prune state has caused prune information to be sent upstream, a graft will need to be sent upstream just as though a new member has joined below. Once data begins to be delivered downstream, if the downstream router again decides to be pruned from the delivery tree, a new prune can be sent upstream at that time.

A time of day clock provides a good source for a non-decreasing 32 bit integer.

3.2.3. Neighbor Addresses

As a DVMRP router sees Probe messages from its DVMRP neighbors, it records the neighbor addresses on each interface and places them in the Probe message sent on the particular interface. This allows the neighbor router to know that its probes have been received by the sending router.

In order to minimize one-way neighbor relationships, a router MUST delay sending poison route reports directly to a neighbor until the neighbor includes the routers address in its probe messages.

Implementations written before this specification will not wait before sending reports nor will they ignore reports sent. Therefore, reports from these implementations SHOULD be accepted whether or not a probe with the routers address has been received.

3.2.4. Probe Packet Format

The Probe packet is variable in length depending on the number of neighbor IP addresses included. The length of the IP packet can be used to determine the number of neighbors in the Probe message. The current Major Version is 3. To maintain compatibility with previous versions, implementations of Version 3 must include pruning and grafting of multicast trees. Non-pruning implementations SHOULD NOT be implemented at this time.

7	15	23	31
Type (0x13)	Code (0x1)	Checksum	
Reserved	Capabilities	Minor	Major
Generation ID			
Neighbor IP Address 1			
Neighbor IP Address 2			
...			
Neighbor IP Address N			

Figure 3 - DVMRP Probe Packet Format

3.2.5. Designated Router Election

Since it is wasteful to have more than a single router sending IGMP Host Membership Queries on a given physical network, a single router on each physical network is elected as the Designated Querier. This election used to be a part of DVMRP. However, this is now handled as a part of the IGMP vesion 2 protocol. Therefore, DVMRP Version 3 requires the use of IGMP Version 2 or later specifying that the Designated Querier election is performed as a part of IGMP.

Even though only one router will act as the designated querier, all DVMRP routers must listen to IGMP Host Membership Reports and keep a local group database.

3.3. Building Forwarding Cache Entries

In order to create optimal multicast delivery trees, DVMRP was designed to keep separate forwarding cache entries for each (source network, destination group) pair. Because the possible combinations of these is quite large, forwarding cache entries are generated on demand as data arrives at a multicast router. Since the IP forwarding decision is made on a hop by hop basis (as with the unicast case), it is imperative that each multicast router has a consistent view of the reverse path back to the source network.

3.3.1. Determining the upstream interface

When a multicast packet arrives, a DVMRP router will use the DVMRP unicast routing table to determine which interface leads back to the source. If the packet did not arrive on that interface, it should be discarded without further processing. Each multicast forwarding entry should cache the upstream interface for a particular source host or source network after looking this up in the DVMRP unicast routing table.

3.3.2. Determining the downstream interface list

The downstream interface list is built from the remaining list of multicast capable interfaces. Any interfaces designated as leaf networks that do not have members of the particular multicast group can be automatically removed from list of downstream interfaces. The remaining interfaces will either have downstream DVMRP routers or directly attached group members. These interfaces may be removed in the future if it is determined that there are no group members anywhere along the entire tree branch.

3.4. Unicast Route Exchange

It was mentioned earlier that since not all IP routers support IP multicast forwarding, it is necessary to tunnel IP multicast datagrams through these routers. One effect of using these encapsulated tunnels is that IP multicast traffic may not be aligned with IP unicast traffic. This means that a multicast datagram from a particular source can arrive on a different (logical) interface than the expected upstream interface based on traditional unicast routing.

The unicast routing information propagated by DVMRP is used for determining the reverse path back to the source of multicast traffic. Tunnel pseudo-interfaces are considered to be distinct for the purpose of determining upstream and downstream interfaces. The routing information that is propagated by DVMRP contains a list of unicast source networks and an appropriate metric. The metric used is a hop count which is incremented by the cost of the incoming interface metric. Traditionally, physical interfaces use a metric of 1 while the metric of a tunnel interface varies with the distance and bandwidth in the path between the two tunnel endpoints. Users are encouraged to configure tunnels with the same metric in each direction to create symmetric routing and provide for easier problem determination although the protocol does not strictly enforce this.

Implementations may wish to provide a mechanism to aggregate source networks to reduce the size of the unicast routing table. All implementations should be able to accept reports for aggregated source networks in accordance with Classless Inter-Domain Routing (CIDR) as described in [Rekh93] and [Full93].

There are two places where aggregation is particularly useful.

1. At organizational boundaries to limit the number of source networks advertised out of the organization.
2. Within an organization to summarize non-local routing information by using a default (0/0) route.

3.4.1. Route Packing and Ordering

Since DVMRP Route Reports may need to refresh several thousand routes each report interval, routers MUST attempt to spread the routes reported across the whole route update interval. This reduces the chance of synchronized route reports causing routers to become overwhelmed for a few seconds each report interval. Since the route report interval is 60 seconds, it is suggested that the total number routes being updated be split across multiple Route Reports sent at regular intervals. Their was an earlier requirement that Route Reports MUST contain source network/mask pairs sorted first by increasing network mask and then by increasing source network. This restriction has been lifted. Implementations conforming to this specification MUST be able to receive Route Reports containing any mixture of network masks and source networks.

In order to pack more source networks into a route report, source networks are often represented by less than 4 octets. The number of non-zero bytes in the mask value is used to determine the number of octets used to represent each source network within that particular mask value. For instance if the mask value of 255.255.0.0 is being reported, the source networks would only contain 2 octets each. DVMRP assumes that source networks will never be aggregated into networks whose prefix length is less than 8. Therefore, it does not carry the first octet of the mask in the Route Report since, given this assumption, the first octet will always be 0xFF. This means that the netmask value will always be represented in 3 octets. This method of specifying source network masks is compatible with techniques described in [Rekh93] and [Full93] to group traditional Class C networks into super-nets and to allow different subnets of the same Class A network to be discontinuous. In this notation, the default route is represented as the least three significant octets of the netmask [00 00 00], followed by one octet for the network number [00].

3.4.2. Unicast Route Metrics

For each source network reported, a route metric is associated with the unicast route being reported. The metric is the sum of the interface metrics between the router originating the report and the source network. For the purposes of DVMRP, the Infinity metric is defined to be 32. This limits the breadth across the whole DVMRP network and is necessary to place an upper bound on the convergence time of the protocol.

As seen in the packet format below, Route Reports do not contain a count of the number of routes reported for each netmask. Instead, the high order bit of the metric is used to signify the last route being reported for a particular mask value. If a metric is read with the high order bit of the 8-bit value set and if the end of the message has not been reached, the next value will be a new netmask to be applied to the subsequent list of routes.

3.4.3. Unicast Route Dependencies

In order for pruning to work correctly, each DVMRP router needs to know which downstream routers depend on it for receiving datagrams from particular source networks. Initially, when a new datagram arrives from a particular source/group pair, it is flooded to all downstream interfaces that have DVMRP neighbors who have indicated a dependency on the receiving DVMRP router for that particular source. A downstream interface can only be removed when it has received Prune messages from each of the dependent routers on that interface. Each downstream router uses Poison Reverse to indicate to the upstream router which source networks it expects to receive from the upstream router. The downstream router indicates this by echoing back the source networks it expects to receive from the upstream router with infinity added to the advertised metric. This means that the legal values for the metric now become between 1 and ($2 \times$ Infinity - 1) or 1 and 63. Values between 1 and 31 indicate reachable unicast source networks. The value Infinity (32) indicates the source network is not reachable. Values between 33 and 63 indicate that the downstream router originating the Report is depending upon the upstream router to provide multicast datagrams from the corresponding source network.

3.4.4. Sending Route Reports

All of the active routes MUST be advertised over every interface running DVMRP each Route Report Interval. In addition, flash updates MAY be sent as needed but any given route MUST not be advertised more often than the Minimum Flash Update Interval (5 seconds). Flash updates can reduce the chances of routing loops and black holes occurring when source networks become unreachable through a particular path. Flash updates need only contain the source networks that have changed. It is not necessary to report all of the source networks from a particular mask value when sending an update.

Route Reports containing downstream dependent "poison" metrics should be sent directly to the neighbors unicast address. These reports should not be sent to a neighbor until a router has seen its own address in the neighbors Probe router list. See Appendix C for exceptions. These Reports should be refreshed at the standard Route Update Interval.

3.4.5. Receiving Route Reports

After receiving a route report, a check should be made to verify it is from a known neighbor. Neighbors are learned via received Probe messages which also indicate the capabilities of the neighbor. Therefore, route reports from unknown neighbors are discarded.

Each route in the report is then parsed and processed according to the following rules:

A. If the route is new and the metric is less than infinity, the route should be added.

B. If the route already exists, several checks must be performed.

1. New metric < infinity

    a. New metric > existing metric

       If the new metric is greater than the existing metric then check to see if the same neighbor is reporting the route. If so, update the metric and flash update the route. Otherwise, discard the route.

    b. New metric < existing metric

       Update the metric for the route and if the neighbor reporting the route is different, update the upstream neighbor in the routing table. Flash update the route to downstream neighbors and if the neighbor changed, a flash update should be sent to the new neighbor indicating downstream dependence and to the existing neighbor withdrawing downstream dependence of the route.

    c. New metric = existing metric

       If the neighbor reporting the route is the same as the existing route, then simply refresh the route. If the new neighbor has a lower IP address, then update the route. Flash updates should be sent to the new and old neighbors to notify them of changes in downstream dependencies.

2. New metric = infinity

    a. New next hop = existing next hop

       If the existing metric was less than infinity, the route is now unreachable. Update the route and possibly update neighbors as well.

       If the existing metric was between infinity and 2 $\times$ infinity, the neighbor used to be a dependent downstream router but is no longer. Note this dependency change and any Prunes that it may trigger.

    b. New next hop $\neq$ existing next hop

       The route can be ignored since the existing next hop has a metric better than or equal to this next hop.

3. infinity < New metric < 2 $\times$ infinity

    If the receiving router is not the designated forwarder for the source network on the interface the report was received, the poison report should be ignored. Otherwise, the neighbor considers the receiving router to be upstream for the route and is indicating it is dependent on the receiving router.

    a. Neighbor on down stream interface

       If the neighbor is considered to be on a downstream interface for that route, then the neighbor should be registered as a downstream dependent router for that route.

If this is the first time the neighbor has indicated downstream dependence for this source and one or more prunes have been sent upstream containing this source network, then Graft messages will need to be sent upstream in the direction of the source network for each group with existing prune state.

b. Neighbor not on down stream interface

If the receiving router thinks the neighbor is on the upstream interface, then the indication of downstream dependence should be ignored.

4. $2 \times$ infinity $\leq$ New metric

If the metric is greater than or equal to $2 \times$ infinity, the metric is illegal and the route should be ignored.

3.4.6. Route Hold-down

When a route learned via a particular gateway expires, a router may be able to reach the source network described by the route through an alternate gateway. However, in the presence of complex topologies, often, the alternate gateway may only be echoing back the same route learned via a different path. If this occurs, the route will continue to be propagated long after it is no longer valid.

In order to prevent this, it is common in distance vector protocols to continue to advertise a route that has been deleted with a metric of infinity for one or more report intervals. During this time, the route may be learned via a different gateway and the router is permitted to use this new gateway. However, the router MUST NOT advertise this new gateway during the hold-down period.

DVMRP begins the hold-down period after 140 seconds ($2 \times$ Route Report Interval + 20). After this time, a new gateway may be used but the route must be advertised with an infinity metric for 2 Report Intervals. At this point, the hold-down period is over and the new gateway (if one exists) can start being advertised. In the absence of a new gateway, the route is simply removed.

Route hold-down is not effective if only some of the routers implement it. Therefore, it is now a REQUIRED part of the protocol.

3.4.7. Graceful Shutdown

During a graceful shutdown, an implementation MAY want to inform neighbor routers that it is terminating. Routes that have been advertised with a metric less than infinity should now be advertised with a metric equal to infinity. This will allow neighbor routers to switch more quickly to an alternate path for a source network if one exists.

Routes that have been advertised with a metric between infinity and $2 \times$ infinity (indicating downstream neighbor dependence) should now be advertised with a metric equal to infinity (canceling the downstream dependence).

3.4.8. Route Report Packet Format

The format of a sample Route Report Packet is shown in Figure 4 below. The packet shown is an example of how the source networks are packed into a Report. The number of octets in each Source Network will vary depending on the mask value. The values below are only an example for clarity and are not intended to represent the format of every Route Report.

<table>
<tr><th>7</th><th>15</th><th>23</th><th>31</th></tr>
<tr><td>Type
(0x13)</td><td>Code
(0x2)</td><td colspan="2">Checksum</td></tr>
<tr><td colspan="2">Reserved</td><td>Minor
Version</td><td>Major
Version</td></tr>
<tr><td>$Mask_1$
$Octet_2$</td><td>$Mask_1$
$Octet_3$</td><td>$Mask_1$
$Octet_4$</td><td>Src
Net_{11}</td></tr>
<tr><td colspan="2">$SrcNet_{11}$(cont.)...</td><td>$Metric_{11}$</td><td>Src
Net_{12}</td></tr>
<tr><td colspan="2">$SrcNet_{12}$(cont.)...</td><td>$Metric_{12}$</td><td>$Mask_2$
$Octet_2$</td></tr>
<tr><td>$Mask_2$
$Octet_3$</td><td>$Mask_2$
$Octet_4$</td><td colspan="2">$SrcNet_{21}$</td></tr>
<tr><td colspan="2">$SrcNet_{21}$(cont.)...</td><td>$Metric_{21}$</td><td>$Mask_3$
$Octet_2$</td></tr>
<tr><td>$Mask_3$
$Octet_3$</td><td>$Mask_3$
$Octet_4$</td><td colspan="2">...</td></tr>
</table>

Figure 4 - Example Route Report Packet Format

3.5. Pruning

DVMRP is described as a broadcast and prune multicast routing protocol since datagrams are initially sent out all dependent downstream interfaces forming a tree rooted at the source of the data. But as the routers at the leafs of the tree begin to receive unwanted multicast traffic, they send prune messages upstream toward the source. This allows the tree branches to become optimal for a given source network and a given set of receivers.

3.5.1. Leaf Networks

Detection of leaf networks is very important to the pruning process. Routers at the end of a source specific multicast delivery tree must detect that there are no further downstream routers. This detection mechanism is covered above in section 3.2 titled Probe Messages. If there are no group members present for a particular multicast datagram received, the leaf routers will start the pruning process by removing their downstream interfaces and sending a prune to the upstream router for that source.

3.5.2. Source Networks

It is important to note that prunes are specific to a group and source network. A prune sent upstream triggered by traffic received from a particular source applies to all sources on that network. It is not currently possible to remove only one or a subset of hosts on a source network for a particular group. All or none of the sources must be removed.

Although the Prune message contains the host address of a source, the source network can be determined easly by a best-match lookup using the unicast routing table distributed as a part of DVMRP.

3.5.3. Receiving a Prune

When a prune is received, the following steps should be taken:

1. Determine if a Probe has been received from this router recently.

2. If not, discard prune since there is no prior state about this neighbor.

3. If so, make sure the neighbor is capable of pruning (based on received Probe message).

4. Since Prune messages are fixed length, ensure the prune message contains at least the correct amount of data.

5. Extract the source address, group address, and prune time-out values

6. If there is no current state information for the (source, group) pair, then ignore the prune.

7. Verify that the prune was received from a dependent neighbor for the source network. If not, discard the prune.

8. Determine if a prune is currently active from the same dependent neighbor for this (source, group) pair.

9. If so, reset the timer to the new time-out value. Otherwise, create state for the new prune and set a timer for the prune lifetime.

10. Determine if all dependent downstream routers on the interface from which the prune was received have now sent prunes.

11. If so, then determine if there are group members active on the interface.

12. If no group members are found, then remove the interface.

13. If all downstream interfaces have now been removed, send a prune to the upstream neighbor.

3.5.4. Sending a Prune

When sending a prune upstream, the following steps should be taken:

1. Decide if upstream neighbor is capable of receiving prunes.

2. If not, then proceed no further.

3. Stop any pending Grafts awaiting acknowledgments.

4. Determine the prune lifetime. This value should be the minimum of the prune lifetimes remaining from the downstream neighbors and the default prune lifetime.

5. Form and transmit the packet to the upstream neighbor for the source.

3.5.5. Retransmitting a Prune

By increasing the prune lifetime to ˜2 hours, the effect of a lost prune message becomes more apparent. Therefore, an implementation MAY choose to retransmit prunes messages using exponential back-off for the lifetime of the prune if traffic is still arriving on the upstream interface.

One way to implement this would be to send a prune, install a negative cache entry for 3 seconds while waiting for the prune to take effect. Then remove the negative cache entry. If traffic continues to arrive, a new forwarding cache request will be generated. The prune can be resent with the remaining prune lifetime and a negative cache entry can be installed for 6 seconds. After this, the negative cache entry is removed. This procedure is repeated while each time doubling the length of time the negative cache entry is installed.

3.5.6. Prune Packet Format

In addition to the standard IGMP and DVMRP headers, a Prune Packet contains three additional fields: the source host IP address, the destination group IP address, and the Prune Lifetime in seconds.

The Prune Lifetime is a derived value calculated as the minimum of the default prune lifetime (2 hours) and the remaining lifetimes of of any downstream prunes received for the same cache entry. A router with no downstream dependent neighbors would use the the default prune lifetime.

7	15	23	31
Type (0x13)	Code (0x7)	Checksum	
Reserved		Minor	Major
Source Address			
Group Address			
Prune Lifetime			

Figure 5 - Prune Packet Format

3.6. Grafting

Once a multicast delivery tree has been pruned back, DVMRP Graft messages are necessary to join new receivers onto the multicast tree. Graft messages are sent upstream from the new receiver's first-hop router until a point on the multicast tree is reached. Graft messages are re-originated between adjacent DVMRP routers and are not forwarded by DVMRP routers. Therefore, the first-hop router does not know if the Graft message ever reaches the multicast tree. To remedy this, each Graft message is acknowledged hop by hop. This ensures that the Graft message is not lost somewhere along the path between the receiver's first-hop router and the closest point on the multicast delivery tree.

One or more Graft messages should be sent under the following conditions:

1. A new local member joins a group that has been pruned upstream.
2. A new dependent downstream router appears on a pruned branch.
3. A dependent downstream router on a pruned branch restarts (new Generation ID).
4. A Graft Retransmission Timer expires before a Graft-Ack is received.

3.6.1. Grafting Each Source Network

It is important to realize that prunes are source specific and are sent up different trees for each source. Grafts are sent in response to a new Group Member which is not source specific. Therefore, separate Graft messages must be sent to the appropriate upstream routers to counteract each previous source specific prune that was sent.

3.6.2. Sending a Graft

As mentioned above, a Graft message sent to the upstream DVMRP router should be acknowledged hop by hop guaranteeing end-to-end delivery. If a Graft Acknowledgment is not received within the Graft Retransmission Time-out period, the Graft should be resent to the upstream router. The initial retransmission period is 5 seconds. A binary exponential back-off policy is used on subsequent retransmissions. In order to send a Graft message, the following steps should be taken:

1. Verify a forwarding cache entry exists for the (source, group) pair and that a prune exists for the cache entry.

2. Verify that the upstream router is capable of receiving prunes (and therefore grafts).

3. Add the graft to the retransmission timer list awaiting an acknowledgment.

4. Formulate and transmit the Graft packet.

3.6.3. Receiving a Graft

The actions taken when a Graft is received depends on the state in the receiving router for the (source, group) pair in the received Graft message. If the receiving router has prune state for the (source, group) pair, then it must acknowledge the received graft and send a subsequent graft to its upstream router. If the receiving router has some removed some downstream interfaces but has not sent a prune upstream, then the receiving interface can simply be added to the list of downstream interfaces in the forwarding cache. A Graft Acknowledgment must also be sent back to the source of the Graft message. If the receiving router has no state at all for the (source, group) pair, then datagrams arriving for the (source, group) pair should automatically be flooded when they arrive. A Graft Acknowledgment must be sent to the source of the Graft message. If a Graft message is received from an unknown neighbor, it should be discarded after it is acknowledged.

3.6.4. Graft Packet Format

The format of a Graft packet is show below:

7	15	23	31
Type (0x13)	Code (0x8)	Checksum	
Reserved		Minor	Major
Source Address			
Group Address			

Figure 6 - Graft Packet Format

3.6.5. Sending a Graft Acknowledgment

A Graft Acknowledgment packet is sent to a downstream neighbor in response to receiving a Graft message. All Graft messages should be acknowledged. This is true even if no other action is taken in response to receiving the Graft to prevent the source from continually re-transmitting the Graft message. The Graft Acknowledgment packet is identical to the Graft packet except that the DVMRP code in the common header is set to Graft Ack. This allows the receiver of the Graft Ack message to correctly identify which Graft was acknowledged and stop the appropriate retransmission timer.

3.6.6. Receiving a Graft Acknowledgment

When a Graft Acknowledgment is received, the (source, group) pair in the packet can be used to determine if a Graft was sent to this particular upstream router. If no Graft was sent, the Graft Ack can simply be ignored. If a Graft was sent, and the acknowledgment has come from the correct upstream router, then it has been successfully received and the retransmission timer for the Graft can be stopped.

3.6.7. Graft Acknowledgment Packet Format

The format of a Graft Ack packet (which is identical to that of a Graft packet) is show below:

7	15	23	31
Type (0x13)	Code (0x9)	Checksum	
Reserved		Minor	Major
Source Address			
Group Address			

Figure 7 - Graft Ack Packet Format

3.7. Interfaces

Interfaces running DVMRP will either be multicast capable physical interfaces or encapsulated tunnel pseudo-interfaces. Physical interfaces may either be multi-access networks or point-to-point networks. Tunnel interfaces are used when there are non-multicast capable routers between DVMRP neighbors. Protocol messages and multicast data traffic are sent between tunnel endpoints using IP-IP encapsulation. The unicast IP addresses of the tunnel endpoints are used as the source and destination IP addresses in the outer IP header. The inner IP header remains unchanged from the original packet.

The maximum packet length of any DVMRP message should be the maximum packet size required to be forwarded without fragmenting. The use of Path MTU Discovery [Mogu90] is encouraged to determine this size. In the absence of Path MTU, the Requirements for Internet Hosts [Brad89] specifies this number as 576 octets. Be sure to consider the size of the encapsulated IP header as well when calculating the maximum size of a DVMRP protocol message.

4. IANA Considerations

The Internet Assigned Numbers Authority (IANA) is the central coordinator for the assignment of unique parameter values for Internet protocols. DVMRP uses IGMP [Fen96a] IP protocol messages to communicate between routers. The IGMP Type field is hexadecimal 0x13.

On IP multicast capable networks, DVMRP uses the All-DVMRP-Routers local multicast group. This group address is 224.0.0.4.

5. Network Management Considerations

DVMRP provides several methods for network management monitoring and troubleshooting. Appendix B describes a request/response mechanism to directly query DVMRP neighbor information. In addition, a Management Information Base for DVMRP is defined in [Thal96]. A protocol independent multicast trace-route facility is defined in [Fen96b].

6. Security Considerations

Security for DVMRP follows the general security architecture provided for the Internet Protocol [Atk95a]. This framework provides for both privacy and authentication. It recommends the use of the IP Authentication Header [Atk95b] to provide trusted neighbor relationships. Confidentiality is provided by the addition of the IP Encapsulating Security Payload [Atk95c]. Please refer to these documents for the general architecture design as well as the specific implementation details.

7. References

[Atk95a] Atkinson, R., "Security Architecture for the Internet Protocol", RFC 1825, August 1995.

[Atk95b] Atkinson, R., "IP Authentication Header", RFC 1826, August 1995.

[Atk95c] Atkinson, R., "IP Encapsulating Security Payload (ESP)", RFC 1827, August 1995.

[Brad89] Braden, R., "Requirements for Internet Hosts -- Communication Layers", RFC 1122, October 1989.

[Deer89] Deering, S., "Host Extensions for IP Multicasting", RFC 1112, August 1989.

[Deer90] Deering, S., Cheriton, D., "Multicast Routing in Datagram Internetworks and Extended LANs", ACM Transactions on Computer Systems, Vol. 8, No. 2, May 1990, pp. 85-110.

[Deer91] Deering, S., "Multicast Routing in a Datagram Internetwork", PhD thesis, Electric Engineering Dept., Stanford University, December 1991.

[Fen96a] Fenner, W., "Internet Group Management Protocol, Version 2", Work In Progress, January 1997.

[Fen96b] Fenner, W., Casner, S., "A "traceroute" facility for IP Multicast", Work In Progress, November 1996.

[Full93] Fuller, V., T. Li, J. Yu, and K. Varadhan, "Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy", RFC 1519, September 1993.

[Mogu90] Mogul, J., Deering, S., "Path MTU Discovery", RFC 1191, November 1990.

[Perk96] Perkins, C., IP Encapsulation within IP, RFC 2003, October 1996.

[Perl92] Perlman, R., Interconnections: Bridges and Routers, Addison-Wesley, May 1992, pp. 205-211.

[Rekh93] Rekhter, Y., and T. Li, "An Architecture for IP Address Allocation with CIDR", RFC 1518, September 1993.

[Reyn94] Reynolds, J., Postel, J., "Assigned Numbers", STD 0002, October 1994.

[Thal96] Thaler, D., "Distance-Vector Multicast Routing Protocol MIB", Work In Progress, June 1996.

[Wait88] Waitzman, D., Partridge, C., Deering, S., "Distance Vector Multicast Routing Protocol", RFC 1075, November 1988.

8. Author's Address

Thomas Pusateri
Juniper Networks, Inc.
3260 Jay St.
Santa Clara, CA 95051
Phone: (919) 558-0700
EMail: pusateri@jnx.com

9. Acknowledgments

The author would like to acknowledge the original designers of the protocol, Steve Deering, Craig Partridge, and David Waitzman. Version 3 of the protocol would not have been possible without the original work of Ajit Thyagarajan and ongoing work of Bill Fenner. Credit also goes to Danny Mitzel for the careful review of this document and Dave LeRoy and Shuching Shieh for their helpful comments.

10. Appendix A - Constants & Configurable Parameters

The following table provides a summary of the DVMRP timing parameters:

Parameter	Value (seconds)
Probe Interval	10
Neighbor Time-out Interval	35
Minium Flash Update Interval	5
Route Report Interval	60
Route Replacement Time	140
Route Expiration Time	200
Prune Lifetime	variable (< 2 hours)
Prune Retransmission Time	3 with exp. back-off
Graft Retransmission Time	5 with exp. back-off

Table 2 - Parameter Summary

11. Appendix B - Tracing and Troubleshooting support

There are several packet types used to gather DVMRP specific information. They are generally used for diagnosing problems or gathering topology information. The first two messages are now obsoleted and should not be used. The remaining two messages provide a request/response mechanism to determine the versions and capabilities of a particular DVMRP router.

Code	Packet Type	Description
3	DVMRP Ask Neighbors	Obsolete
4	DVMRP Neighbors	Obsolete
5	DVMRP Ask Neighbors 2	Request Neighbor List
6	DVMRP Neighbors 2	Respond with Neighbor List

Table 3 - Debugging Packet Types

11.1. DVMRP Ask Neighbors2

The Ask Neighbors2 packet is a unicast request packet directed at a DVMRP router. The destination should respond with a unicast Neighbors2 message back to the sender of the Ask Neighbors2 message.

<table>
<tr><td>0</td><td>8</td><td>16</td><td>31</td></tr>
<tr><td>Type
(0x13)</td><td>Code
(0x5)</td><td colspan="2">Checksum</td></tr>
<tr><td colspan="2">Reserved</td><td>Minor
Version</td><td>Major
Version</td></tr>
</table>

Figure 8 - Ask Neighbors 2 Packet Format

11.2. DVMRP Neighbors2

The format of a Neighbors2 response packet is shown below. This is sent as a unicast message back to the sender of an Ask Neighbors2 message. There is a common header at the top followed by the routers capabilities. One or more sections follow that contain an entry for each logical interface. The interface parameters are listed along with a variable list of neighbors learned on each interface.

If the interface is down or disabled, list a single neighbor with an address of 0.0.0.0 for physical interfaces or the remote tunnel endpoint address for tunnel pseudo-interfaces.

0	8	16	31
Type (0x13)	Code (0x6)	Checksum	
Reserved	Capabilities	Minor Version	Major Version
Local Addr 1			
Metric 1	Threshold 1	Flags 1	Nbr Count 1
Nbr 1			
...			
Nbr m			
Local Addr N			
Metric N	Threshold N	Flags N	Nbr Count N
Nbr 1			
...			
Nbr k			

Figure 9 - Neighbors 2 Packet Format

The capabilities of the local router are defined as follows:

Bit	Flag	Description
0	Leaf	This is a leaf router
1	Prune	This router understands pruning
2	GenID	This router sends Generation Id's
3	Mtrace	This router handles Mtrace requests
4	Snmp	This router supports the DVMRP MIB

Table 4 - DVMRP Router Capabilities

The flags associated with a particular interface are:

Bit	Flag	Description
0	Tunnel	Neighbor reached via tunnel
1	Source Route	Tunnel uses IP source routing
2	Reserved	No longer used
3	Down	Operational status down
4	Disabled	Administrative status down
5	Reserved	No longer used
6	Leaf	No downstream neighbors on interface

Table 5 - DVMRP Interface flags

12. Appendix C - Version Compatibility

There have been two previous major versions of DVMRP with implementations still in circulation. If the receipt of a Probe message reveals a major version of 1 or 2, then it can be assumed that this neighbor does not support pruning or the use of the Generation ID in the Probe message. However, since these older implementations are known to safely ignore the Generation ID and neighbor information in the Probe packet, it is not necessary to send specially formatted Probe packets to these neighbors.

There were three minor versions (0, 1, and 2) of major version 3 that did support pruning but did not support the Generation ID or capability flags. These special cases will have to be accounted for.

Any other minor versions of major version 3 closely compare to this specification.

In addition, cisco Systems is known to use their software major and minor release number as the DVMRP major and minor version number. These will typically be 10 or 11 for the major version number. Pruning was introduced in Version 11.

Implementations prior to this specification may not wait to send route reports until probe messages have been received with the routers address listed. Reports SHOULD be sent to these neighbors without first requiring a received probe with the routers address in it as well as reports from these neighbors SHOULD be accepted. Although, this allows one-way neighbor relationships to occur, it does maintain backward compatibility.

Table of Contents

1. Introduction 2
   1.1. Reverse Path Multicasting 2
   1.2. IP-IP Tunnels 2
   1.3. Document Overview 2
2. Protocol Overview 3
   2.1. Neighbor Discovery 3
   2.2. Source Location 3
   2.3. Dependent Downstream Routers 4
   2.4. Multi-access Networks 4
   2.5. Building Multicast Trees 4
   2.6. Pruning Multicast Trees 5
   2.7. Grafting Multicast Trees 5
3. Detailed Protocol Operation 6
   3.1. Protocol Header 6
   3.2. Probe Messages 7
   3.3. Building Forwarding Cache Entries 10
   3.4. Unicast Route Exchange 10
   3.5. Pruning 15
   3.6. Grafting 18
   3.7. Interfaces 21
4. IANA Considerations 21
5. Network Management Considerations 21
6. Security Considerations 22
7. References 22
8. Author's Address 22
9. Acknowledgments 23
10. Appendix A - Constants & Configurable Parameters 24
11. Appendix B - Tracing and Troubleshooting support 25
   11.1. DVMRP Ask Neighbors2 25
   11.2. DVMRP Neighbors2 25
12. Appendix C - Version Compatibility 28