Network Working Group
Internet-Draft
Expires: January 10, 2012

M. Xu
Y. Cui
S. Yang
Tsinghua University
C. Metz
G. Shepherd
Cisco Systems

July 9, 2011

Softwire Mesh Multicast
draft-xu-softwire-mesh-multicast-02

Abstract

The Internet needs support IPv4 and IPv6 packets. Both address families and their attendant protocol suites support multicast of the single-source and any-source varieties. As part of the transition to IPv6, there will be scenarios where a backbone network running one IP address family internally (referred to as internal IP or I-IP) will provide transit services to attached client networks running another IP address family (referred to as external IP or E-IP). It is expected that the I-IP backbone will offer unicast and multicast transit services to the client E-IP networks.

Softwires Mesh is a solution for supporting E-IP unicast and multicast across an I-IP backbone. This document describes the mechanisms for supporting Internet-style multicast across a set of E-IP and I-IP networks supporting softwires mesh.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 10, 2012.

Table of Contents

1. Introduction

The Internet needs to support IPv4 and IPv6 packets. Both address families and their attendant protocol suites support multicast of the single-source and any-source varieties. As part of the transition to IPv6, there will be scenarios where a backbone network running one IP address family internally (referred to as internal IP or I-IP) will provide transit services to attached client networks running another IP address family (referred to as external IP or E-IP).

The preferred solution is to leverage the multicast functions, inherent in the I-IP backbone, to efficiently and scalably tunnel encapsulated client E-IP multicast packets inside an I-IP core tree rooted at one or more ingress AFBR nodes and branching out to one or more egress AFBR leaf nodes.

[6] outlines the requirements for the softwires mesh scenario including multicast. It is straightforward to envisage that client E-IP multicast sources and receivers will reside in different client E-IP networks connected to an I-IP backbone network. This requires that the client E-IP source-rooted or shared tree will need to traverse the I-IP backbone network.

One method to accomplish this is to re-use the multicast VPN approach outlined in [10]. MVPN-like schemes can support the softwire mesh scenario and achieve a "many-to-one" mapping between the E-IP client multicast trees and transit core multicast trees. The advantage of this approach is that the number of trees in the I-IP backbone network scales less than linearly with the number of E-IP client trees. Corporate enterprise networks and by extension multicast VPNs have been known to run applications that create a large amount of (S,G) states. Aggregation at the edge contains the (S,G) states that need to be maintained by the network operator supporting the customer VPNs. The disadvantage of this approach is possible inefficient bandwidth and resource utilization if multicast packets are delivered to a receiver AFBR with no attached E-IP receiver.

Internet-style multicast is somewhat different in that the trees tends to be relatively sparse and source-rooted. The need for multicast aggregation at the edge (where many customer multicast trees are mapped into a few or one backbone multicast trees) does not exist and to date has not been identified. Thus the need for a basic or closer alignment with E-IP and I-IP multicast procedures emerges.

A framework on how to support such methods is described in [8]. In this document, a more detailed discussion supporting the "one-to-one" mapping schemes for the IPv6 over IPv4 and IPv4 over IPv6 scenarios will be discussed.

2. Terminology

An example of a softwire mesh network supporting multicast is illustrated in Figure 1. A multicast source S is located in one E-IP client network, while candidate E-IP group receivers are located in the same or different E-IP client networks that all share a common I-IP transit network. When E-IP sources and receivers are not local to each other, they can only communicate with each other through the I-IP core. There may be several E-IP sources for some multicast group residing in different client E-IP networks. In the case of shared trees, the E-IP sources, receivers and RPs might be located in different client E-IP networks. In the simple case the resources of the I-IP core are managed by a single operator although the inter-provider case is not precluded.

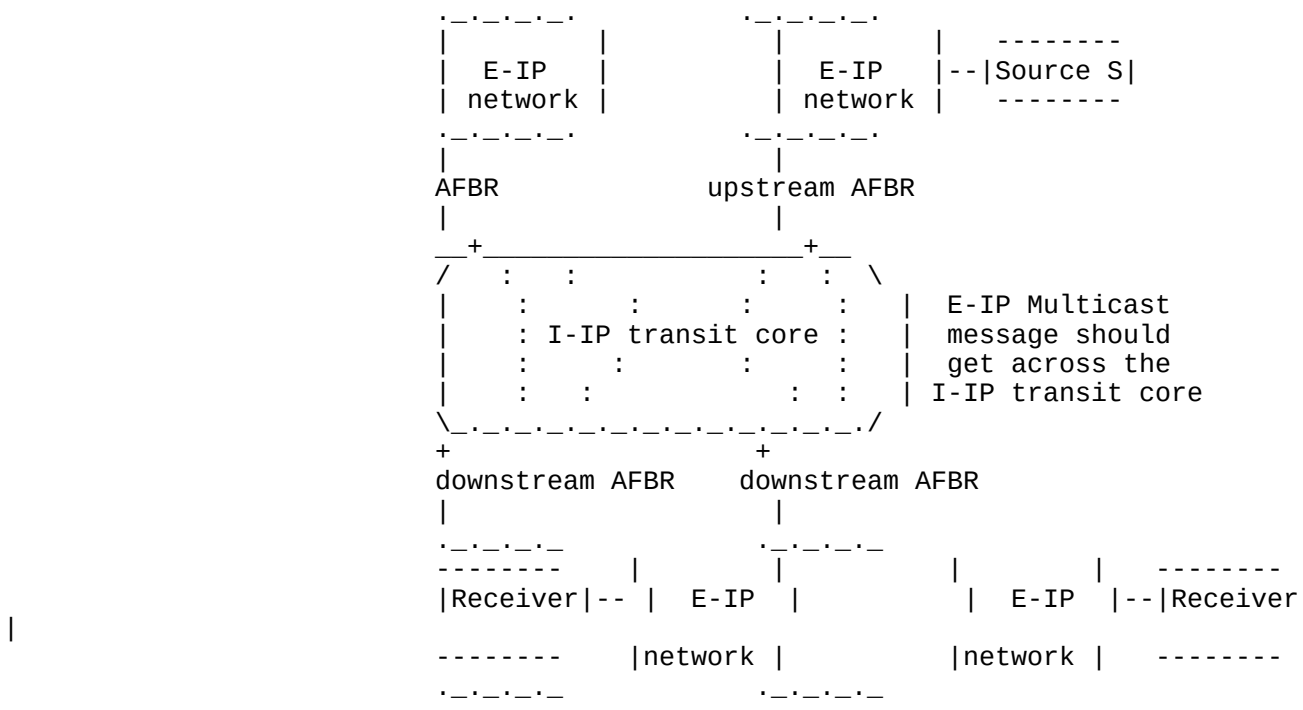

Figure 1: Softwire Mesh Multicast Framework

Terminology used in this document:

o Address Family Border Router (AFBR) - A dual-stack router

interconnecting two or more networks using different IP address families. In the context of softwire mesh multicast, the AFBR runs E-IP and I-IP control planes to maintain E-IP and I-IP multicast states respectively and performs the appropriate encapsulation/ decapsulation of client E-IP multicast packets for transport across the I-IP core. An AFBR will act as a source and/or receiver in an I-IP multicast tree.

o Upstream AFBR: The AFBR router that is located at the upstream of a multicast data flow.

o Downstream AFBR: The AFBR router that is located at the downstream of a multicast data flow.

o I-IP (Internal IP). This refers to the form of IP (i.e., either IPv4 or IPv6) that is supported by the core (or backbone) network. An I-IPv6 core network runs IPv6 and an I-IPv4 core network runs IPv4.

o E-IP (External IP) This refers to the form of IP (i.e. either IPv4 or IPv6) that is supported by the client network(s) attached to the I-IP transit core. An E-IPv6 client network runs IPv6 and an E-IPv4 client network runs IPv4.

o I-IP core tree. A single-source or multi-source distribution tree rooted at one or more AFBR source nodes and branched out to one or more AFBR leaf nodes. An I-IP core Tree is built using standard IP or MPLS multicast signaling protocols operating exclusively inside the I-IP core network. An I-IP core Tree is used to tunnel E-IP multicast packets belonging to E-IP trees across the I-IP core. Another name for an I-IP core Tree is multicast or multipoint softwire.

o E-IP client tree. A single-source or multi-source distribution tree rooted at one or more hosts or routers located inside a client E-IP network and branched out to one or more leaf nodes located in the same or different client E-IP networks.

3. Scenarios of Interest

This section describes the two different scenarios where softwires mesh multicast will apply.

3.1. IPv4-over-IPv6

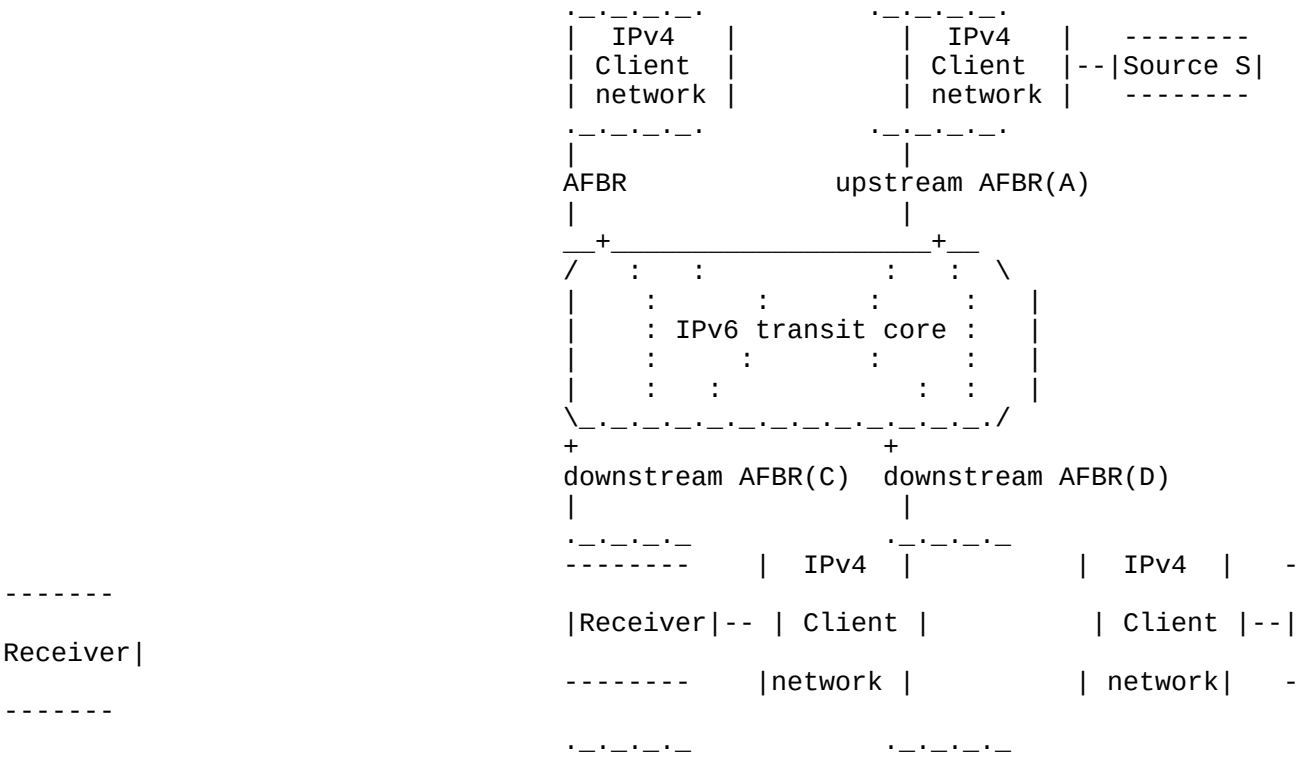

Figure 2: IPv4-over-IPv6 Scenario

In this scenario, the E-IP client networks run IPv4 and I-IP core runs IPv6. This scenario is illustrated in Figure 2.

Because of the much larger IPv6 group address space, it will not be a problem to map individual client E-IPv4 tree to a specific I-IPv6 core tree. This simplifies operations on the AFBR because it becomes possible to algorithmically map an IPv4 group/source address to an IPv6 group/source address and vice-versa.

The IPv4-over-IPv6 scenario is an emerging requirement as network operators build out native IPv6 backbone networks. These networks

naturally support native IPv6 services and applications but it is with near 100% certainty that legacy IPv4 networks handling unicast and multicast will need to be accommodated.

3.2. IPv6-over-IPv4

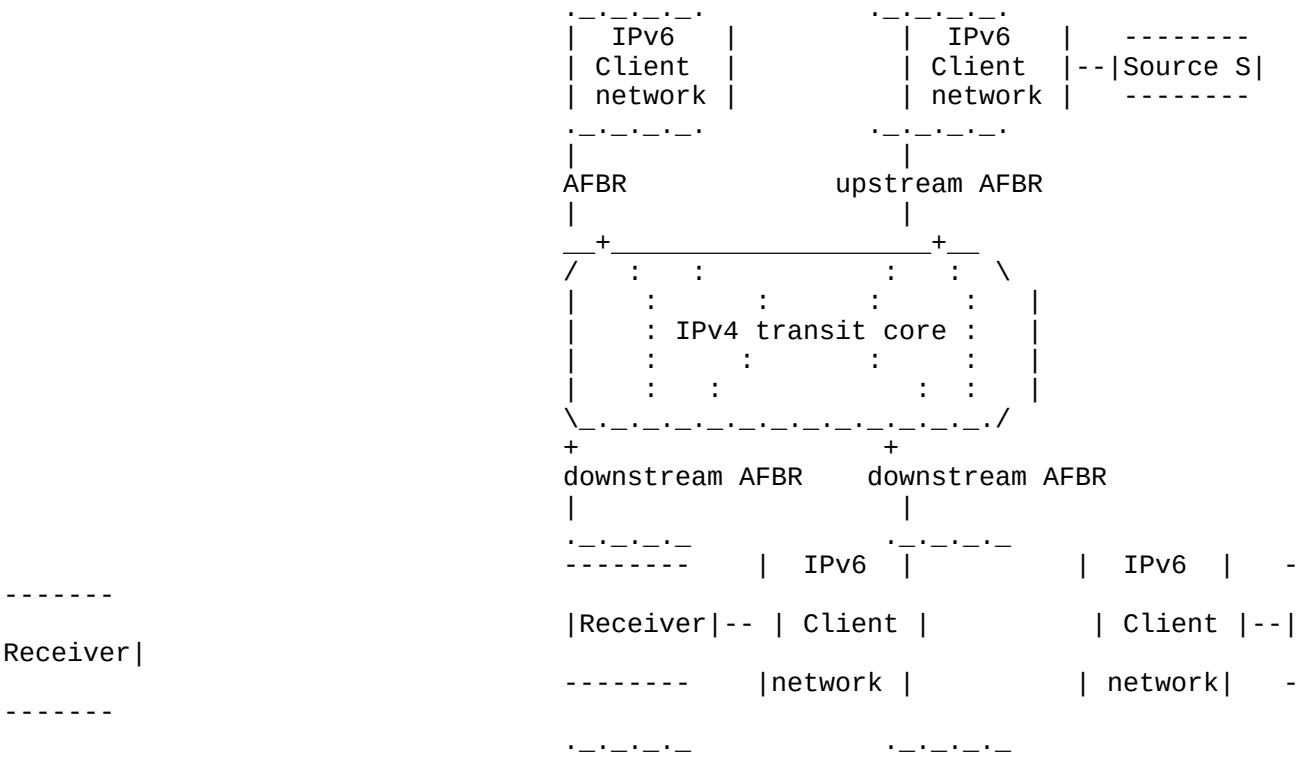

Figure 3: IPv6-over-IPv4 Scenario

In this scenario, the E-IP Client Networks run IPv6 while the I-IP core runs IPv4 and is illustrated in Figure 3.

IPv6 multicast group addresses are longer than IPv4 multicast group addresses. It will not be possible to perform an algorithmic IPv6 - to - IPv4 address mapping without the risk of multiple IPv6 group addresses mapped to the same IPv4 address resulting in unnecessary bandwidth and resource consumption. Therefore additional efforts will be required to ensure that client E-IPv6 multicast packets can be injected into the correct I-IPv4 multicast trees at the AFBRs. This clear mismatch in IPv6 and IPv4 group address lengths means that it will not be possible to perform a one-to-one mapping between IPv6

and IPv4 group addresses unless the IPv6 group address is scoped.

As mentioned earlier this scenario is common in the MVPN environment. As native IPv6 deployments and multicast applications emerge from the outer reaches of the greater public IPv4 Internet, it is envisaged that the IPv6 over IPv4 softwire mesh multicast scenario will be a necessary feature supported by network operators.

4. IPv4-over-IPv6

4.1. Mechanism

Routers in the client E-IPv4 networks contain routes to all other client E-IPv4 networks. Through the set of known and deployed mechanisms, E-IPv4 hosts and routers have discovered or learned of (S,G) or (*,G) IPv4 addresses. Any I-IP multicast state instantiated in the core is referred to as (S’,G’) or (*,G’) and is of course separated from E-IP multicast state.

Suppose a downstream AFBR receives an E-IPv4 PIM Join/Prune message from the E-IPv4 network for either an (S,G) tree or a (*,G) tree. The AFBR can translate the E-IPv4 PIM message into an I-IPv6 PIM message with the latter being directed towards I-IP IPv6 address of the upstream AFBR. When the I-IPv6 PIM message arrives at the upstream AFBR, it should be translated back into an E-IPv4 PIM message. The result of these actions is the construction of E-IPv4 trees and a corresponding I-IP tree in the I-IP network.

In this case it is incumbent upon the AFBR routers to perform PIM message conversions in the control plane and IP group address conversions or mappings in the data plane. It becomes possible to devise an algorithmic one-to-one IPv4-to-IPv6 address mapping at AFBRs.

4.2. Source Address Mapping

There are two kinds of multicast --- ASM and SSM. It’s possible for I-IP network and E-IP network to support different kinds of multicast, and the source address translation rules may vary a lot. There are four scenarios to be discussed in detail:

- E-IP network supports SSM, I-IP network supports SSM
  One possible way to make sure that the translated I-IPv6 PIM message reaches upstream AFBR is to set S’ to a virtual IPv6 address that leads to the upstream AFBR. Figure 4 is the recommended address format based on [9]:

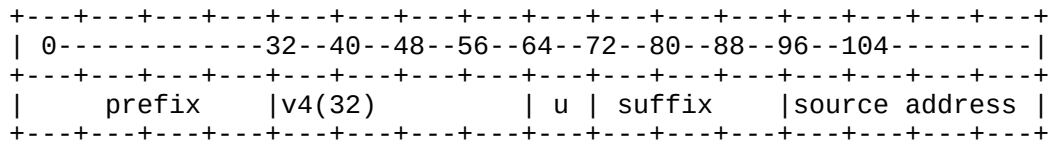

Figure 4: IPv4-Embedded IPv6 Virtual Source Address Format

In this address format, the "prefix" field contains a "Well-Known" prefix or a ISP-defined prefix. An existing "Well-Known" prefix is 64:ff9b, which is defined in [9]; "v4" field is the IP address of one of upstream AFBR's E-IPv4 interface; "u" field is defined in [4], and MUST be set to zero; "suffix" field is reserved for future extensions and SHOULD be set to zero; "source address" field stores the original S.
To make it feasible, the /32 prefix must be known to every AFBR, and AFBRs should not only announce the /96 prefixes of S' to the I-IPv6 network, but also announce the IP addresses of upstream AFBRs' E-IPv4 interface presented in the "v4" field to other AFBRs by MPBGP. In this way, when a downstream AFBR receives a (S,G) message, it can translate it into (S',G') by looking up the IP address of the corresponding AFBR's E-IPv4 interface. Since S' is globally unique and the /96 prefix of S' is known to every router in I-IPv6 network, the translated message will eventually arrive at the corresponding upstream AFBR, and the upstream AFBR can translate the message back to (S,G).

- o E-IP network supports SSM, I-IP network supports ASM
  Since any network that supports ASM should also support SSM, we can construct a SSM tree in I-IP network. The operation in this scenario is the same as that in the first scenario.

- o E-IP network supports ASM, I-IP network supports SSM
  ASM and SSM have the same PIM message format. The main differences between ASM and SSM are RP and (*,G) messages. To make this scenario feasible, we must be able to translate (*,G) messages into (S',G') messages at downstream AFBRs, and translate it back at upstream AFBRs. Assume RP' is the upstream AFBR that locates between RP and the downstream AFBR. When downstream AFBR receives an E-IPv4 PIM (*,G) message, S' can be generated according to the format specified in Figure 4, with "v4" field setting to the IP address of one of RP's E-IPv4 interface and "source address" field setting to *(the IPv4 address of RP). The translated message will eventually arrive at RP'. RP' checks the "source address" field and find the IPv4 address of RP, so RP' judges that this is originally a (*,G) message, then it translates the message back to (*,G) message and forward it to RP.
  Traveling all the way from sources to the RP, and then back down the shared tree may result in the multicast data packets passing through RP' twice, which brings about undesirable increased latency or bandwidth consumption. For this reason, RP' MAY perform a "cut-through", namely when RP' receives multicast data packets sent from sources to RP, it not only forwards them to RP, but also forwards them directly onto the multicast tree built in the I-IPv6 network. (S,G,rpt) messages should be sent towards RP to avoid reduplication.

- o E-IP network supports ASM, I-IP network supports ASM
  To keep it as simple as possible, we treat I-IP network as SSM and the solution is the same as the third scenario.

4.3. Group Address Mapping

For IPv4-over-IPv6 scenario, a simple algorithmic mapping between IPv4 multicast group addresses and IPv6 group addresses is supported. [11] has already defined an applicable format. Figure 5 is a reminder of the format:

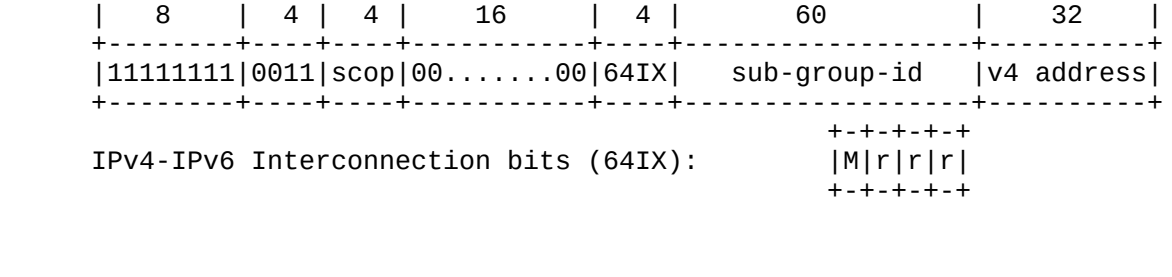

Figure 5: IPv4-Embedded IPv6 Multicast Address Format: SSM Mode

The high order bits of the I-IPv6 address range will be fixed for mapping purposes. With this scheme, each IPv4 multicast address can be mapped into an IPv6 multicast address(with the assigned prefix), and each IPv6 multicast address with the assigned prefix can be mapped into IPv4 multicast address.

4.4. Actions performed by AFBR

The following actions are performed by AFBRs:

- o Receive E-IPv4 PIM messages
  When a downstream AFBR receives an E-IPv4 PIM message, it should check the address family of the next-hop towards the destination. If the address family is IPv4, the AFBR should forward the message without any translation; otherwise it should take the following operation.

- o Translate E-IPv4 PIM messages into I-IPv6 PIM messages
  E-IPv4 PIM message with S(or *) and G is translated into I-IPv6 PIM message with S' and G' following the rules specified above.

- o Transmit I-IPv6 PIM messages
  The downstream AFBR sends the I-IPv6 PIM message to the upstream AFBR. When the upstream AFBR receives this I-IPv6 PIM message, it

checks the prefix of the source address and judges that the message is a translated message, then translates the message back to E-IPv4 PIM message and sends it towards source or RP.

- Process and forward multicast data
  On receiving multicast data from upstream routers, the AFBR looks up its forwarding table to check the IP address of each outgoing interface. If there exists at least one outgoing interface whose IP address family is different from the incoming interface, the AFBR should encapsulate/decapsulate this packet and forward it to the outgoing interface(s), and then forward the data to the other outgoing interfaces without encapsulation/decapsulation.

5. IPv6-over-IPv4

5.1. Mechanism

Routers in the client E-IPv6 networks contain routes to all other client E-IPv6 networks. Through the set of known and deployed mechanisms, E-IPv6 hosts and routers have discovered or learned of (S,G) or (*,G) IPv6 addresses. Any I-IP multicast state instantiated in the core is referred to as (S’,G’) or (*,G’) and is of course separated from E-IP multicast state.

This particular scenario introduces unique challenges. Unlike the IPv4-over-IPv6 scenario, it’s impossible to map all of the IPv6 multicast address space into the IPv4 address space to address the one-to-one Softwire Multicast requirement. To coordinate with the "IPv4-over-IPv6" scenario and keep the solution as simple as possible, one possible solution to this problem is to limit the scope of the E-IPv6 source addresses for mapping, such as applying a "Well-Known" prefix or a ISP-defined prefix.

5.2. Source Address Mapping

There are two kinds of multicast --- ASM and SSM. It’s possible for I-IP network and E-IP network to support different kind of multicast, and the source address translation rules may vary a lot. There are four scenarios to be discussed in detail:

- E-IP network supports SSM, I-IP network supports SSM
  To make sure that the translated I-IPv4 PIM message reaches the upstream AFBR, we need to set S’ to an IPv4 address that leads to the upstream AFBR. But due to the non-"one-to-one" mapping of E-IPv6 to I-IPv4 unicast address, the upstream AFBR is unable to remap the I-IPv4 source address to the original E-IPv6 source address without any constraints.
  We apply a fixed IPv6 prefix and static mapping to solve this problem. A recommended source address format is defined in [9]. Figure 6 is a reminder of the format:

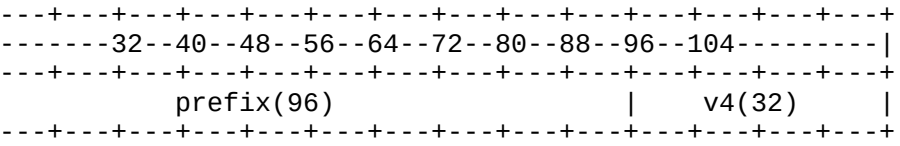

Figure 6: IPv4-Embedded IPv6 Source Address Format

In this address format, the "prefix" field contains a "Well-Known" prefix or a ISP-defined prefix. An existing "Well-Known" prefix is 64:ff9b, which is defined in [9]; "v4" field is the corresponding I-IPv4 source address.
To make it feasible, the /96 prefix must be known to every AFBR, every E-IPv6 address of sources that support mesh multicast MUST follow the format specified in Figure 6, and the corresponding upstream AFBR should announce the I-IPv4 address in "v4" field to the I-IPv4 network. In this way, when a downstream AFBR receives a (S,G) message, it can translate it into (S',G') by simply take off the prefix in S. Since S' is known to every router in I-IPv4 network, the translated message will eventually arrive at the corresponding upstream AFBR, and the upstream AFBR can translate the message back to (S,G) by appending the prefix to S'.

- o E-IP network supports SSM, I-IP network supports ASM
  Since any network that supports ASM should also support SSM, we can construct a SSM tree in I-IP network. The operation in this scenario is the same as that in the first scenario.

- o E-IP network supports ASM, I-IP network supports SSM
  ASM and SSM have the same PIM message format. The main differences between ASM and SSM are RP and (*,G) messages. To make this scenario feasible, we must be able to translate (*,G) messages into (S',G') messages at downstream AFBRs and translate it back at upstream AFBRs. Here, the E-IPv6 address of RP MUST follow the format specified in Figure 6. Assume RP' is the upstream AFBR that locates between RP and the downstream AFBR. When a downstream AFBR receives a (*,G) message, it can translate it into (S',G') by simply take off the prefix in *(the E-IPv6 address of RP). Since S' is known to every router in I-IPv4 network, the translated message will eventually arrive at RP'. RP' knows that S' is the mapped I-IPv4 address of RP, so RP' will translate the message back to (*,G) by appending the prefix to S' and forward it to RP.
  Traveling all the way from sources to the RP, and then back down the shared tree may result in the multicast data packets passing through RP' twice, which brings about undesirable increased latency or bandwidth consumption. For this reason, RP' MAY perform a "cut-through", namely when RP' receives multicast data packets sent from sources to RP, it not only forwards them to RP, but also forwards them directly onto the multicast tree built in the I-IPv6 network. (S,G,rpt) messages should be sent towards RP to avoid reduplication.

- o E-IP network supports ASM, I-IP network supports ASM
  To keep it as simple as possible, we treat I-IP network as SSM and the solution is the same as the third scenario.

5.3. Group Address Mapping

To keep one-to-one group address mapping simple, the group address range of E-IP IPv6 can be reduced in a number of ways to limit the scope of addresses that need to be mapped into the I-IP IPv4 space.

A recommended multicast address format is defined in [11]. The high order bits of the E-IPv6 address range will be fixed for mapping purposes. With this scheme, each IPv4 multicast address can be mapped into an IPv6 multicast address(with the assigned prefix), and each IPv6 multicast address with the assigned prefix can be mapped into IPv4 multicast address.

5.4. Actions performed by AFBR

The following actions are performed by AFBRs

- Receive E-IPv6 PIM messages
  When a downstream AFBR receives an E-IPv6 PIM message, it should check the address family of the upstream router. If the address family is IPv6, the AFBR should not translate this message; otherwise it should take the following operation.

- Translate E-IPv6 PIM messages into I-IPv4 PIM messages
  E-IPv6 PIM message with S(or *) and G is translated into I-IPv4 PIM message with S’ and G’ following the rules specified above.

- Transmit I-IPv4 PIM messages
  The downstream AFBR sends the I-IPv4 PIM message to the upstream AFBR. When the upstream AFBR receives this I-IPv4 PIM message, it checks the source address and judges that the message is a translated message, then translates the message back to E-IPv6 PIM message and sends it towards source or RP.

- Process and forward multicast data
  On receiving multicast data from upstream routers, the AFBR looks up its forwarding table to check the IP address of each outgoing interface. If there exists at least one outgoing interface whose IP address family is different from the incoming interface, the AFBR should encapsulate/decapsulate this packet and forward it to the outgoing interface(s), and then forward the data to the other outgoing interfaces without encapsulation/decapsulation.

6. Security Considerations

The AFBR routers could maintain secure communications through the use of Security Architecture for the Internet Protocol as described in[RFC4301]. But when adopting some schemes that will cause heavy burden on routers, some attacker may use it as a tool for DDoS attack.

7. IANA Considerations

When AFBRs perform address mapping, they should follow some predefined rules, especially the IPv6 prefix for source address mapping should be predefined, so that ingress AFBR and egress AFBR can finish the mapping procedure correctly. The IPv6 prefix for translation can be unified within only the transit core, or within global area. In the later condition, the prefix should be assigned by IANA.

8. References

8.1. Normative References

[1] Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 2784, March 2000.

[2] Foster, B. and F. Andreasen, "Media Gateway Control Protocol (MGCP) Redirect and Reset Package", RFC 3991, February 2005.

[3] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 2373, July 1998.

[4] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.

[5] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

[6] Li, X., Dawkins, S., Ward, D., and A. Durand, "Softwire Problem Statement", RFC 4925, July 2007.

[7] Wijnands, IJ., Boers, A., and E. Rosen, "The Reverse Path Forwarding (RPF) Vector TLV", RFC 5496, March 2009.

[8] Wu, J., Cui, Y., Metz, C., and E. Rosen, "Softwire Mesh Framework", RFC 5565, June 2009.

[9] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.

8.2. Informative References

[10] Aggarwal, R., Bandi, S., Cai, Y., Morin, T., Rekhter, Y., Rosen, E., Wijnands, I., and S. Yasukawa, "Multicast in MPLS/ BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-10 (work in progress), January 2010.

[11] Boucadair, M., Qin, J., Lee, Y., Venaas, S., Li, X., and M. Xu, "IPv4-Embedded IPv6 Multicast Address Format", draft-boucadair-behave-64-multicast-address-format-02 (work in progress), June 2011.

Appendix A. Acknowledgements

Wenlong Chen, Xuan Chen, Alain Durand, Yiu Lee, Jacni Qin and Stig Venaas provided useful input into this document.

Authors' Addresses

Mingwei Xu
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R. China

Phone: +86-10-6278-5822
Email: xmw@cernet.edu.cn

Yong Cui
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R. China

Phone: +86-10-6278-5822
Email: cuiyong@tsinghua.edu.cn

Shu Yang
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R. China

Phone: +86-10-6278-5822
Email: yangshu@csnet1.cs.tsinghua.edu.cn

Chris Metz
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
USA

Phone: +1-408-525-3275
Email: chmetz@cisco.com

Greg Shepherd
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
USA

Phone: +1-541-912-9758
Email: shep@cisco.com