Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 7, 2010

Y. Cui
P. Wu
S. Wang
M. Xu
J. Wu
X. Li
Tsinghua University
L. Zhang
UCLA
C. Metz
Cisco Systems, Inc.
July 6, 2009

VA-Based Softwire
draft-cui-softwire-va-based-softwire-00

Status of this Memo

This Internet-Draft will expire on January 7, 2010.

Copyright Notice

Abstract

The increasing deployment of IPv6 networks in both customer networks and ISP networks leads to two common traversing transition scenarios: in the first scenario, an IPv6-only backbone network needs to provide IP connectivity between IPv4 networks, we call it IPv4-over-IPv6 scenario; In the second scenario, IPv6 networks need to be interconnected over an IPv4 transit network, we call it IPv6-over-IPv4 scenario. In both scenarios, the ISP operating the transit network of one address family must offer transit services for attached client networks of the other address family. The Softwire WG has defined softwire mesh mechanism [RFC5565] for the two traversing scenarios. Softwire mesh uses automatic softwire tunnels employing multi-protocol BGP extensions for distributing E-IP routes, where both BGP peers and tunnels between PEs forms a full-mesh architecture.

Inspired by the Virtual Aggregation approach [I-D.ietf-grow-va] to IPv4 routing scalability, in this draft we proposed a scalable mechanism for distributing E-IP routes over the transit network. Our solution can significantly reduce the forwarding information base (FIB) size at Address Family Border Routers (AFBRs) as well as the total amount of routing updates, and offers the ISP an easy way to manage the transit service.

Table of Contents

1. Introduction 5
2. Terminology 7
3. VA-based softwire framework 8
3.1. Scenario 8
3.2. Downlink routing 9
3.3. Uplink routing 9
3.4. Tunneled forwarding 10
4. Discussion on Design Details 12
4.1. BGP-based routing scheme 12
4.2. Tunnel 12
4.3. Cooperate with softwire mesh 13
5. Benefits of VA-based softwire 14
6. Relation with VA 15
7. Inter-domain consideration 16
8. Security considerations 17
9. Acknowledgements 18
10. References 19
10.1. Normative References 19
10.2. Informative References 19
Authors' Addresses 20

1. Introduction

Recently more and more IPv6 networks have be deployed in both customer networks and transit networks, while the existing IPv4 networks still carry the major network traffic and host most network services and applications. It is commonly believed that IPv4 and IPv6 networks will co-exist for the foreseeable future. There has been basically two aspects of IPv6 transition research: connection between IPv4 and IPv6 nodes and connection between networks of one address family traversing network of the other address family. Basic solution for the former one is address translation and for the latter one is tunneling.

For traversing transition, softwire mesh provides a way that networks of one address family can be connected through transit network of the other address family. It's done by extending MP-BGP to exchange external routing information between dual-stack Provider Edge routers (PE), and using softwire tunnel to forward External-IP(E-IP) packets through encapsulation and decapsulation. In both data plane and control plane, all PEs form a full mesh.

Virtual Aggregation (VA) mechanism can reduce the FIB size of routers by an order of magnitude. It can be deployed autonomously by an ISP, and co-exist with legacy routers in the ISP. VA divides the IP address space into Virtual Prefixes (VPs), and uses tunnels to aggregate the regular sub-prefixes within each VP. For each sub-prefix within a VP, Aggregation Point Routers (APRs) have a tunnel from themselves to the remote ASBR (Autonomous System Border Router) where packets for that prefix should be delivered. Because APRs may not be on the shortest path between the ingress and egress routers, the packets may take a longer path and experience additional latency. However as shown in [I-D.ietf-grow-va], a proper placement of ARPs can make the path length and network load increase negligible. Furthermore, VA can make majority of data packet avoid traversing the APR by installing the routes for popular prefixes in all routers. In other words, popular prefixes will not be aggregated. Packets to those prefixes are tunneled directly to the BGP NEXT_HOP.

This draft proposes an approach that adopts the basic idea from VA to solve the traversing transition problem, called VA-based softwire. In control plane, it organizes the E-IP address space into VPs and aggregates the E-IP routes from the client network; regular E-IP prefixes are collected by APRs in Internal-IP(I-IP) backbone, while PEs only have to maintain VPs in the FIB. In data plane, VA-based softwire uses APRs in backbone network to be intermediate tunneling forwarders between PEs; E-IP packets are tunneled to APRs from client network, and then tunneled to the destination client network.

VA-based softwire can significantly reduce the transition FIB size of PEs, and the total amount of transition routing activity (routing protocol process in transition-related routers and transition-related routing packets delivered), and provide the ISP of backbone networks with a better way to manage the transit service. This mechanism has good scalability and works well when the number of client networks increases.

2. Terminology

I-IP: according to [RFC5565], the term "I-IP"("Internal IP") refers to the form of IP (i.e., either IPv4 or IPv6) that is supported by the transit backbone network. The P routers support only I-IP.

E-IP: the term "E-IP" ("External IP") refers to the form of IP that is supported by the client networks. In the scenarios of interest, E-IP is IPv4 if and only if I-IP is IPv6, and E-IP is IPv6 if and only if I-IP is IPv4.

Aggregation Point Router (APR): This draft adopts the name of APR from VA in [I-D.ietf-grow-va]. An Aggregation Point Router (APR) is a router that aggregates a Virtual Prefix (VP) by installing routes (into the FIB) for all of the sub-prefixes within the VP. Every APR can hold several VPs and the corresponding sub-prefixes for every VP. In this draft, all VPs and sub-prefixes are E-IP prefixes and APR can be deployed in arbitrarily position in I-IP backbone; for each sub-prefix within the VP, APRs have a tunnel to the client network where E-IP packets can reach their destinations.

Provider Edge router (PE): The dual-stack edge routers of the backbone network, where E-IP packets enter and leave the backbone. PE is often referred to as AFBR (Address Family Border Router). Interior nodes of the backbone are often known as "P routers".

Popular Prefix: This draft adopts the name of popular prefix from VA. In VA, a popular prefix is a sub-prefix that is installed in a router in addition to the sub-prefixes it holds by virtue of being an Aggregation Point Router. The popular prefix allows packets to follow the shortest path. In VA-based softwire, a popular prefix is an E-IP prefix installed in a PE router in addition to VPs. E-IP packets whose destination falls within popular prefix can traverse I-IP backbone in softwire mesh tunnels.

Virtual Prefix (VP): This draft adopts the name of VP from VA. A Virtual Prefix is a prefix used to aggregate its contained regular prefixes (sub-prefixes). A VP is not physically aggregatable, and it is aggregated at APRs through the use of tunnels.

3. VA-based softwire framework

3.1. Scenario

The scenario of VA-based softwire is illustrated in figure1. A number of P routers compose an I-IP-only backbone, in which a few APRs are deployed. The PE routers are dual-stack and connected to E-IP client networks. Every PE builds tunnels with every APR. The I-IP backbone acts as a transit core to transport E-IP packets across the I-IP backbone. This enables each of E-IP client network to communicate with each other via two hop tunnels.

If E-IP is IPv6 and I-IP is IPv4, the scenario is IPv6-over-IPv4; else E-IP is IPv4 and I-IP is IPv6, then the scenario is IPv4-over-IPv6.

```
                    +--------+   +--------+
                    |  E-IP  |   |  E-IP  |
                    | Client |   | Client |
                    | Network|   | Network|
                    +--------+   +--------+
                        |             |
                        |             |
                   +----------+ +----------+
                   |Dual-Stack| |Dual-Stack|
                +--|    PE    |-|    PE    |--+
                |  +----------+ +----------+  |
                |       :   :     :   :       |
                |       :      :      :       |
                |       :   :     :   :       |    I-IP
                |      [APR]       [APR]      |  backbone
                |       :   :     :   :       |
                |       :      :      :       |
                |       :   :     :   :       |
                |  +----------+ +----------+  |
                +--|Dual-Stack|-|Dual-Stack|--+
                   |    PE    | |    PE    |
                   +----------+ +----------+
                        |             |
                    +--------+   +--------+
                    |  E-IP  |   |  E-IP  |
                    | Client |   | Client |
                    | Network|   | Network|
                    +--------+   +--------+
```

Figure 1 VA-based Softwire Scenario

In the described scenario, VA-based softwire provides connectivity

for E-IP client network in three steps: downlink routing, uplink routing and tunneled forwarding.

3.2. Downlink routing

In downlink routing, every selected APR must distribute its VP information to every PE. VP can be configured in advance in every APR, that is, Every APR is configured with which VPs it is responsible for. Then ARP must advertise its VPs to all PEs, using some intra-domain routing protocol.

In VA it is said that initial VPs can be statically configured in every VA router as VP-list, and a VP can be added by some APR originating routes for it and advertising these routes, also a VP can be deleted by first removing it from the VP-Lists of non-APRs, waiting for them to install sub-prefixes and then remove it from the APRs. This draft, however, uses the uniform routing method that initial VPs and all VP changes are first configured in APRs and then advertised to all PEs for automaticity and simplicity. How to process this routing behavior is still a concern, since E-IP routes need to be advertised in or through I-IP network. We'll discuss this in section 4.1.

Here we give an example of downlink routing. Suppose the backbone is IPv6 and contains 2 APRs, APR1 is responsible for VP 0.0.0.0/1 and VP 128.0.0.0/2 while APR2 is responsible for VP 192.0.0.0/2. IPv4 client networks attached to the backbone through 2 PEs, PE1 is attached with network 10.0.0.0/16 and 192.2.0.0/16 while PE2 is attached with network 144.0.0.0/8. Then in this step, APR1 advertises the VPs of 0.0.0.0/1 and 128.0.0.0/2 to both PE1 and PE2, APR2 advertises the VP 192.0.0.0/2 to PE1 and PE2.

After this step, every PE has all the VPs in its FIB table so it knows which APR to forward an E-IP packet to, even there is no regular prefix match for the destination.

3.3. Uplink routing

This step is opposite to downlink routing. In this step, every PE must advertise the prefixes of the E-IP client network behind it to corresponding APRs.

We use the example in section3.2 again to illustrate uplink routing. In this example, PE1 must advertise 10.0.0.0/16 to APR1 and 192.2.0.0/16 to APR2, since sub-prefix 10.0.0.0/16 falls within VP 0.0.0.0/1 and 192.2.0.0/16 falls within 192.0.0.0/2; PE2 must advertise 144.0.0.0/8 to APR1,since 144.0.0.0/8 falls within VP 128.0.0.0/2.

After this step, every APR has all the sub-prefix that is from the E-IP client networks and falls within one of the VPs the APR is responsible for. Therefore, every APR knows which PE to forward an E-IP packet that is received from another PE earlier to.

3.4. Tunneled forwarding

In VA-based softwire, forwarding an E-IP packet through the I-IP backbone includes the following steps: the ingress PE encapsulates the incoming E-IP packet with the I-IP tunnel header; transmits the encapsulated packet through the I-IP backbone to an APR; the APR decapsulates the packet and encapsulates the packet with another I-IP tunnel header; transmits the encapsulated packet through the backbone to the egress PE; the egress PE decapsulates the I-IP header and forwards the original E-IP packet. All the encapsulations and decapsulations are performed on PEs and APRs, other routers in I-IP backbone take encapsulated packets just as native I-IP packets. The forwarding procedure is illustrated in figure2.

```
                      Tunnel1          Tunnel2
                    ----------->     ----------->
                          I-IP Transit Core
      +-+ E-IP  +--+            +----+           +--+ E-IP  +-+
      |S|--->---|PE|======>=====|APR|======>=====|PE|--->---|D|
      +-+       +--+            +----+           +--+       +-+
   Original  Ingress PE         APR           Egress PE     Original
   Source                   Decapsulation&                  Destination
   Node    Encapsulation    Encapsulation    Decapsulation  Node
```

Figure 2 VA-based softwire: tunneled forwarding

When an ingress PE receives an E-IP packet from client network, it looks up the destination IP address of the packet. In this case, the best match for that address will be a VP route whose next hop is an APR. The ingress PE must forward the packet through a tunnel to the APR. This is done by encapsulating the packet, using an I-IP-encapsulating header with the destination address of the APR, and then forwarding the packet to I-IP network.

When the APR receives the encapsulated I-IP packet, it extracts the payload, i.e., the original E-IP packet, and looks up the packet's destination address. In this case, the best match for that address will be a regular E-IP route whose next hop is a PE (the egress PE). The APR will encapsulate the packet again, this time with the destination I-IP address of the egress PE, and then forward the packet to I-IP network again.

When the egress PE receives the encapsulated I-IP packet, it extracts

the payload, gets the original E-IP packet and forwards it further by looking up its IP destination address in E-IP FIB.

4. Discussion on Design Details

4.1. BGP-based routing scheme

One approach is to use BGP to propagate E-IP routing informaiton downlink and uplink. Every APR sets up and maintains an iBGP session with every PE. We assume that these iBGP sessions are statically configured at APRs and PEs.

In downlink direction, VPs will be advertised through BGP process, from every APR to every PE. Note here the prefix is of E-IP format and next hop is of I-IP format. For IPv4-over-IPv6 scenario, [RFC5549] has made an extension of MP-BGP, modifying MP_REACH_NLRI format to convey IPv4 NLRI prefix information with an IPv6 next hop address, we can just use this v4nlri-v6nh extension here. As for IPv6-over-IPv4 scenario, [RFC4798] provides a way that IPv6 routing information can be distributed in IPv4 BGP by egress PE associating MPLS labels with IPv6 prefixes; besides, we believe that the extension of MP-BGP by [RFC5549] can also be used in IPv6-over-IPv4 scenario to advertised IPv6 NLRI with IPv4 next hop through IPv4 BGP. BGP process in APRs and PEs need to be modified to fit the possible extensions.

In uplink direction, BGP routing is similar to downlink routing, except this time every PE advertise the prefixes of the E-IP client network behind it to corresponding APRs. In this part, because only the APRs whose VPs the E-IP prefixes fall within need to receive the routes, ideally we can build BGP peers only between such APRs and PEs. However, we've already built BGP peers between every APR and every PE in downlink routing, so we can just add some filters in every APR in order to only accept what it actually needs, that is, the E-IP prefixes fall within its VPs.

4.2. Tunnel

In VA, a variety of tunnel types can be used: MPLS LSPs, IP-in-IP, GRE, L2TP, and so on. VA-based softwire doesn't restrict tunnel types, either. Actually since remote ASBR information for VA isn't needed in VA-based softwire, standard softwire tunnel can be used in the scenario. Note that signaling is needed for some types of tunnels using BGP. Refer to [RFC5565] and [RFC5512] for details. Note that encapsulation and decapsulation can be implemented by hardware, so it won't become a heavy burden for APRs and PEs' forwarding process.

4.3. Cooperate with softwire mesh

VA-base softwire can cooperate well with softwire mesh, just similar to popular prefix can be used in VA besides VPs. Since the two mechanisms both influent data forwarding by inject entries to PE's FIB, there will be no confliction between them. Actually, if both mechanisms are used, entries of softwire mesh will have higher priority because VPs' masks are usually shorter than regular prefixes. Here we also call the prefixes of softwire mesh entries popular prefixes.

For some common, heavy-traffic softwire connections, PEs can choose to follow popular prefixes, so only one time encapsulation and decapsulation is needed and encapsulated packets can follow the shortest path in I-IP backbone; for other softwire connections, PEs can refer to VA-based softwire so the FIB size won't be large and PEs won't have to build a lot of BGP peers and tunnels with other PEs.

5. Benefits of VA-based softwire

There are mainly three benefits using VA-based softwire in traversing transition. The first one is that it can significantly reduce the FIB size of the PEs. Every PE only needs to store the E-IP VPs of all APRs, while the whole E-IP regular sub-prefixes are distributed in the APRs' FIBs. PE can also keep a few regular prefixes for softwire mesh use, to reach better performance. So VA-based softwire achieves better scalability than pure softwire mesh.

Secondly, it can reduce the total amount of transition-related routing activity. In this mechanism routing is executed between every APR and every PE. Since there are only a few APRs in the domain, the total amount of routing activity is in proportion to the number of PEs. However, in softwire mesh, every two PEs form a BGP peer, so the amount of routing activity is in proportion to the square of the number of PEs. It's obvious that we can carry out less routing activity than softwire simply implementing uplink and downlink routing in BGP.

Moreover, VA-based softwire can provide the ISP of E-IP networks with a better way to manage the IPv4-over-IPv6 or IPv6-over-IPv4 traversing service. In this mechanism, E-IP routes are collected in APRs, maintained by the ISP. PEs don't know the detailed routes: they just learn a few VPs for forwarding E-IP packets. If a new client network wants to jump in and get connected with other E-IP networks, the ISP just needs to tell the access PE the addresses of the APRs. It's more transparent than softwire mesh where PE needs to know the addresses of all other PEs, and fewer configurations are needed.

6. Relation with VA

This draft adopts the aggregated, centralized architecture, and the basic idea of VP aggregating sub-prefixes from VA. Both the two mechanisms achieve their goals using virtual aggregation, with a slight path length and router load increase. Both the two mechanisms require minor changes to most routers.

However, our proposal is actually quite different from VA. First of all, the problem we are trying to solve is different. VA solves the intra-domain FIB size problem using VP and tunnels; every VA router in the domain will benefit from VA with a FIB suppression. Our draft is mainly aimed to propose an IPv6 transition mechanism with a new architecture instead of solving the FIB problem of the routers in the backbone. Second, the sphere of influence of the two mechanisms is different. VA can be deployed for every router in the domain, and all the intra-domain flows in that domain will be redirected by VA. VA-based softwire won't influent the I-IP domain except for the changes in APRs and PEs, and the traffic injected from client network. The native routing and forwarding process in the I-IP domain will work just the same as the situation without VA-based softwire. In fact, all related routing and tunneled forwarding process are implemented in APRs and PEs; other routers in the domain won't even notice the change. Third, the FIB size problems solved by the two mechanisms are different. VA reduces the size of global FIB for every router in the domain. VA-based softwire reduces only the E-IP FIBs of PEs, which is in proportion to the number of client networks.

7. Inter-domain consideration

The situation will be different if two E-IP networks from two I-IP domains which run VA-based softwire want to get connected. In this inter-domain scenario, APRs in one ISP don't have the regular prefixes of the E-IP network behind the other ISP, though it may fall within its VPs. So if a PE receives an E-IP packet whose destination is in the other ISP, it will still encapsulate and send the packet to the APR whose VP matches the destination in its own ISP; After the corresponding APR receive and decapsulate the packet, it has to drop the packet since there is no regular prefix match in its E-IP FIB table for the destination. So apparently APRs need to learn E-IP routes of the other ISP.

The scenario is illustrated in figure3. Detailed design will be our further work.

```
            I-IP domain1                   I-IP domain2
   +---------------------------+     +---------------------------+
   |     [ EBGP ROUTER ]=======|=====|=======[ EBGP ROUTER ]     |
   |      :           :        |     |        :           :      |
   |      :           :        |     |        :           :      |
   |    [APR]       [APR]      |     |      [APR]       [APR]    |
   |      :  :     :  :        |     |        :  :     :  :      |
   |      :     :     :        |     |        :     :     :      |
   |      :  :     :  :        |     |        :  :     :  :      |
   | +----------+ +----------+ |     | +----------+ +----------+ |
   +-|Dual-Stack|-|Dual-Stack|-+     +-|Dual-Stack|-|Dual-Stack|-+
     |    PE    | |    PE    |         |    PE    | |    PE    |
     +----------+ +----------+         +----------+ +----------+
          |             |                 |             |
          |             |                 |             |
      +--------+   +--------+         +--------+   +--------+
      |  E-IP  |   |  E-IP  |         |  E-IP  |   |  E-IP  |
      | Client |   | Client |         | Client |   | Client |
      | Network|   | Network|         | Network|   | Network|
      +--------+   +--------+         +--------+   +--------+
```

Figure 3 inter-domain scenario

8. Security considerations

Since our mechanism is based on VA, we refer to [I-D.ietf-grow-va] for security concerns. Our mechanism doesn't introduce security problems other than the ones of VA's.

If VA is configured properly, or we say if all APRs and PEs are configured properly, then any new concerns for intra-domain security appear to be relatively minor. In particular, DoS attack to APR won't significantly worsen the DoS problem, and VA won't limit the deployment of DoS defense systems.

As to the situation of Mis-configured VA, VA introduces the possibility that a VP is advertised outside of an AS. Usually a VP is large(i.e. larger than any real prefixes), and the impact is minimal. Smaller prefixes will be preferred because of best-match semantics, and so the only impact is that packets that otherwise have no matching routes will be sent to the misbehaving AS and dropped there. If the VP is small, then it may cause a traffic hijack which can happen with or without VA, so VA doesn't introduce a new security problem.

9. Acknowledgements

This draft gets the very original idea from VA, and extends the idea to solve a different problem: IPv4-over-IPv6 transiton. The authors would like to thank P.Francis, X.Xu, H.Ballani, D. Jen, R. Raszuk, L. Zhang and everyone else who contributed to VA.

10. References

10.1. Normative References

[RFC4798] De Clercq, J., Ooms, D., Prevost, S., and F. Le Faucheur, "Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, February 2007.

[RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, April 2009.

[RFC5549] Le Faucheur, F. and E. Rosen, "Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop", RFC 5549, May 2009.

[RFC5565] Wu, J., Cui, Y., Metz, C., and E. Rosen, "Softwire Mesh Framework", RFC 5565, June 2009.

10.2. Informative References

[I-D.ietf-grow-va]
Francis, P., Xu, X., Ballani, H., Jen, D., Raszuk, R., and L. Zhang, "FIB Suppression with Virtual Aggregation", draft-ietf-grow-va-00 (work in progress), May 2009.

Authors' Addresses

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

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

Peng Wu
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R.China

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

Shengling Wang
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R.China

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

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

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

Jianping Wu
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R.China

Phone: +86-10-6278-5983
Email: jianping@cernet.edu.cn

Xing Li
Tsinghua University
Department of Electronic Engineering, Tsinghua University
Beijing 100084
P.R.China

Phone: +86-10-6278-5983
Email: xing@cernet.edu.cn

Lixia Zhang
UCLA

Email: lixia@cs.ucla.edu

Chris Metz
Cisco Systems, Inc.
3700 Cisco Way
San Jose, Ca. 95134
USA

Email: chmetz@cisco.com