TOC

Network Working Group	P. Francis
Internet-Draft	MPI-SWS
Intended status: Informational	X. Xu
Expires: March 4, 2011	Huawei
	H. Ballani
	Cornell U.
	R. Raszuk
	Cisco
	L. Zhang
	UCLA
	August 31, 2010

Simple Virtual Aggregation (S-VA)
draft-ietf-grow-simple-va-01.txt

Abstract

The continued growth in the Default Free Routing Table (DFRT) stresses the global routing system in a number of ways. One of the most costly stresses is FIB size: ISPs often must upgrade router hardware simply because the FIB has run out of space, and router vendors must design routers that have adequate FIB. FIB suppression is an approach to relieving stress on the FIB by NOT loading selected RIB entries into the FIB. Simple Virtual Aggregation (S-VA) is a simple form of Virtual Aggregation (VA) that allows any and all edge routers to shrink their FIB requirements substantially and therefore increase their useful lifetime. S-VA does not change FIB requirements for core routers. S-VA is extremely easy to configure---considerably more so than the various tricks done today to extend the life of edge routers. S-VA can be deployed autonomously by an ISP (cooperation between ISPs is not required), and can co-exist with legacy routers in the ISP.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

TOC

ISPs today manage constant DFRT growth in a number of ways. One way, of course, is for ISPs to upgrade their router hardware before DFRT growth outstrips the size of the FIB. This is too expensive for many ISPs. They would prefer to extend the lifetime of routers whose FIBs can no longer hold the full DFRT.

A common approach taken by lower-tier ISPs is to default route to their providers. Routes to customers and peer ISPs are maintained, but everything else defaults to the provider. This approach has several disadvantages. First, packets to Internet destinations may take longer-than-necessary AS paths. This problem can be mitigated through careful configuration of partial defaults, but this can require substantial configuration overhead. A second problem with defaulting to providers is that the ISP is no longer able to provide the full DFRT to its customers. Finally, provider defaults prevents the ISP from being able to detect martian packets. As a result, the ISP transmits packets that could otherwise have been dropped over its expensive provider links. Simple Virtual Aggregation (S-VA) solves these problems because the full DFRT is used by core routers.

An alternative is for the ISP to maintain full routes in its core routers, but to filter routes from edge routers that do not require a full DFRT. These edge routers can then default route to the core routers. This is often possible with edge routers that interface to customer networks. The problem with this approach is that it cannot be used for all edge routers. For instance, it cannot be used for routers that connect to transits. It should also not be used for routers that connect to customers which wish to receive the full DFRT.

This draft describes a very simple technique, called Simple Virtual Aggregation (S-VA), that allows any and all edge routers to have substantially reduced FIB requirements even while still advertising and receiving the full DFRT over BGP. The basic idea is as follows. Core routers in the ISP maintain the full DFRT in the FIB and RIB. Edge routers maintain the full DFRT in the RIB, but suppress certain routes from the FIB. Edge routers install a default route to core routers. Label Switched Paths (LSP) are used to transmit packets from a core router, through the edge router, to the Next Hop remote Autonomous System Border Router (ASBR). ASBRs strip the tunnel header (MPLS or IP) before forwarding tunneled packets to the remote ASBR (in much the same way MPLS Penultimate Hop Popping (PHP) strips the LSP header before forwarding packets to the tunnel target).

S-VA requires no changes to BGP and no changes to MPLS forwarding mechanisms in routers. Configuration is extremely simple: S-VA must be enabled, and routers must told whether they are FIB-suppressing routers or not. Everything else is automatic. ISPs can deploy FIB suppression autonomously and with no coordination with neighbor ASes.

1.1. Scope of this Document

TOC

The scope of this document is limited to Intra-domain S-VA operation. In other words, the case where a single ISP autonomously operates S-VA internally without any coordination with neighboring ISPs.

Note that this document assumes that the S-VA "domain" (i.e. the unit of autonomy) is the AS (that is, different ASes run S-VA independently and without coordination). For the remainder of this document, the terms ISP, AS, and domain are used interchangeably.

This document applies equally to IPv4 and IPv6.

S-VA may operate with a mix of upgraded routers and legacy routers. There are no topological restrictions placed on the mix of routers. In order to avoid loops between upgraded and legacy routers, however, legacy routers must be able to terminate tunnels.
Note that S-VA is a greatly simplified variant of "full VA" [I-D.ietf-grow-va] (Francis, P., Xu, X., Ballani, H., Jen, D., Raszuk, R., and L. Zhang, “FIB Suppression with Virtual Aggregation,” May 2009.). With full VA, all routers (core or otherwise) can have reduced FIBs. However, full VA requires substantial new configuration and operational complexity compared to S-VA. Note that S-VA was formerly specified in [I-D.ietf-grow-va] (Francis, P., Xu, X., Ballani, H., Jen, D., Raszuk, R., and L. Zhang, “FIB Suppression with Virtual Aggregation,” May 2009.). It has been moved to this separate draft to simplify its understanding.

1.2. Requirements notation

TOC

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119] (Bradner, S., “Key words for use in RFCs to Indicate Requirement Levels,” March 1997.).

1.3. Terminology

TOC

FIB-Installing Router (FIR): An S-VA router that does not suppress any routes, and advertises itself as a default route for 0/0. Typically a core router or route reflector would be configured as an FIR.

FIB-Suppressing Router (FSR): An S-VA router that installs a route to 0/0, and may suppress other routes. Typically an edge router would be configured as an FSR.

Install and Suppress: The terms "install" and "suppress" are used to describe whether a RIB entry has been loaded or not loaded into the FIB. In other words, the phrase "install a route" means "install a route into the FIB", and the phrase "suppress a route" means "do not install a route into the FIB".

Legacy Router: A router that does not run S-VA, and has no knowledge of S-VA.

Routing Information Base (RIB): The term RIB is used rather sloppily in this document to refer either to the loc-RIB (as used in [RFC4271] (Rekhter, Y., Li, T., and S. Hares, “A Border Gateway Protocol 4 (BGP-4),” January 2006.)), or to the combined Adj-RIBs-In, the Loc-RIB, and the Adj-RIBs-Out.

2. Operation of S-VA

TOC

There are three types of routers in S-VA, FIB-Installing routers (FIR), FIB-Suppressing routers (FSR), and optionally legacy routers. While any router can be an FIR or an FSR (there are no topology constraints), the simplist form of deployment is for border routers to be configured as edge routers, and for non-border routers (for instance the routers used as route reflectors) to be configured as core routers. S-VA, however, does not mandate this deployment per se.

FIRs must originate a BGP route to NLRI 0/0 [RFC4271] (Rekhter, Y., Li, T., and S. Hares, “A Border Gateway Protocol 4 (BGP-4),” January 2006.). The ORIGIN is set to INCOMPLETE (value 2), the AS number of the FIR's AS is used in the AS_PATH, and the BGP NEXT_HOP is set to the router's own address. The ATOMIC_AGGREGATE and AGGREGATOR attributes are not included. The FIR MUST attach a NO_EXPORT Communities Attribute [RFC1997] (Chandrasekeran, R., Traina, P., and T. Li, “BGP Communities Attribute,” August 1996.) to the route.

FIRs must not FIB-suppress any routes.

FSRs must FIB-install a route to 0/0. When transmitting a packet to a FIR (i.e. based on a 0/0 FIB lookup), the packet must be tunneled. This is to prevent loops that would otherwise occur when a packet transits multiple FSRs on the way to the core, some of which have FIB-installed the route for the destination, and others of which have not. FSRs may FIB-install any other routes. They should install any routes for which their eBGP neighbor is the NEXT_HOP. There are a couple reasons for this, which can be illustrated in the figure below. This figure shows an autonomous system with a FIR FIR1 and an FSR FSR1. FSR1 is an ASBR and is connected to two remote ASBRs, EP1 and EP2.

```
          +------------------------------------------+
          |      Autonomous System                   |   +----+
          |                                          |   |EP1 |
          |                                      /---+---|    |
          |   To   ----\ +----+          +----+ /    |   +----+
          | Other       \|FIR1|----------|FSR1|/     |
          |Routers      /|    |          |    |\     |
          |        ----/ +----+          +----+ \    |   +----+
          |                                      \---+---|EP2 |
          |                                          |   |    |
          |                                          |   +----+
          +------------------------------------------+
```

Suppose that FSR1 does not FIB-install routes for which EP1 and EP2 are next hops. In this case, when EP2 sends a packet to FSR1 for which the next hop is EP1, FSR1 will first tunnel the packet to FIR1, which will tunnel it right back to FSR1. This trombone routing is avoided if local ASBRs FIB-install routes where their neighbor remote ASBRs are the BGP NEXT_HOP.

In addition, FSR1 cannot filter source addresses using strict unicast Reverse Path Forwarding (uRPF) unless it FIB-installs the routes learned from the remote ASBR. Note, however, that FSRs cannot do loose uRPF. Rather, this must be done by FIRs.

The above observations lead to the following rules: FSRs that are ASBRs should FIB-install all routes for which the neighbor is the BGP NEXT_HOP. FSRs that are ASBRs must FIB-install any routes that are used for uRPF.

2.1. Tunnels

TOC

S-VA works with both MPLS and IP-in-IP tunnels. There are potentially up to two tunnels required for a packet to traverse an AS with S-VA. The first tunnel is that from an FSR to a FIR (for the 0/0 default). This is called the default tunnel. The second tunnel targets the remote ASBR which is the BGP NEXT_HOP, although the tunnel header is stripped by the local ASBR before transmitting to the remote ASBR. This is the exit tunnel. The start of the exit tunnel is an ingress local ASBR in the case where the ingress local ASBR has FIB-installed the associated route. Otherwise, the start of the exit tunnel is a FIR.

The target address of the default tunnel is always the FIR. If MPLS is used, the FIRs must initiate LSPs to themselves using either the Label Distribution Protocol (LDP) [RFC5036] (Andersson, L., Minei, I., and B. Thomas, “LDP Specification,” October 2007.). RSVP-TE [RFC3209] (Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G.

Swallow, “RSVP-TE: Extensions to RSVP for LSP Tunnels,” December 2001.) may also be used.

If IP-in-IP tunnels are used, then the BGP Encapsulation Extended Community (BGPencap-Attribute) ([RFC5512] (Mohapatra, P. and E. Rosen, “BGP Encapsulation SAFI and BGP Tunnel Encapsulation Attribute,” April 2009.)) is used to convey the ability to accept tunnels at the target address (the BGP NEXT_HOP).

For the exit tunnels, again either MPLS or IP-in-IP can be used. In the case of IP-in-IP, the inner label defined in [RFC4023] (Worster, T., Rekhter, Y., and E. Rosen, “Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE),” March 2005.) and signaled in BGP with [RFC3107] (Rekhter, Y. and E. Rosen, “Carrying Label Information in BGP-4,” May 2001.) is used by the local ASBR to identify the remote ASBR which is the BGP NEXT_HOP for the packet. Specifically, when a local ASBR, which can be either an FSR or a FIR, advertises an eBGP-received route into iBGP, it sets the BGP NEXT_HOP as itself. It assigns a label to the route. This label is used as the inner label in packets tunneled to the local ASBR, and is used to identify the remote ASBR from which the route was received. When receiving a packet with this label, the local ASBR strips off the label, and forwards the native packet to the remote ASBR indicated by the label.

In the case of MPLS, the inner label may or may not be used. If it is used, then an LSP is established to the IP address of the local ASBR as described above for FIRs. The BGP NEXT_HOP is set to be itself (the same address that serves as the FEC in the LSP). The inner label is established as described in the previous paragraph for IP-in-IP tunnels, but with the encapsulation defined in [RFC3032] (Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, “MPLS Label Stack Encoding,” January 2001.).

If the inner label is not used, then the local ASBR must initiate a Downstream Unsolicited LSP for each remote ASBR. The FEC for the LSP is the remote ASBR address that is used in the BGP NEXT_HOP field. When a packet is received on one of these LSPs, the local ASBR strips the MPLS header, and forwards the packet to the remote ASBR indicated by the label.

2.2. Legacy Routers

TOC

S-VA may be operated with a mix of legacy and S-VA-upgraded routers. The legacy routers, however, must be able to forward tunneled packets. In the case of MPLS tunnels, this means that they must fully participate in MPLS signaling. If a legacy router is an ASBR, then it must also initiate tunnels to itself and be able to detunnel packets (without the inner label).

3. IANA Considerations

TOC

There are no IANA considerations.

4. Security Considerations

TOC

The authors are not aware of any new security considerations due to S-VA.

5. Acknowledgements

TOC

The concept for S-VA comes from Robert Raszuk.

6. Normative References

TOC

[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 (TXT).
[RFC1997]	Chandrasekeran, R., Traina, P., and T. Li, “BGP Communities Attribute,” RFC 1997, August 1996 (TXT).
[RFC2119]	Bradner, S., “Key words for use in RFCs to Indicate Requirement Levels,” BCP 14, RFC 2119, March 1997 (TXT, HTML, XML).
[RFC3032]	Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, “MPLS Label Stack Encoding,” RFC 3032, January 2001 (TXT).
[RFC3107]	Rekhter, Y. and E. Rosen, “Carrying Label Information in BGP-4,” RFC 3107, May 2001 (TXT).
[RFC3209]	Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, “RSVP-TE: Extensions to RSVP for LSP Tunnels,” RFC 3209, December 2001 (TXT).
[RFC4023]	Worster, T., Rekhter, Y., and E. Rosen, “Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE),” RFC 4023, March 2005 (TXT).
[RFC4271]	Rekhter, Y., Li, T., and S. Hares, “A Border Gateway Protocol 4 (BGP-4),” RFC 4271, January 2006 (TXT).
[RFC5036]	Andersson, L., Minei, I., and B. Thomas, “LDP Specification,” RFC 5036, October 2007 (TXT).
[RFC5512]	

Mohapatra, P. and E. Rosen, “BGP Encapsulation SAFI and BGP Tunnel Encapsulation Attribute,” RFC 5512, April 2009 (TXT).

Authors' Addresses

TOC

Paul Francis
Max Planck Institute for Software Systems
Gottlieb-Daimler-Strasse
Kaiserslautern 67633
Germany
Phone: +49 631 930 39600
Email: francis@mpi-sws.org

Xiaohu Xu
Huawei Technologies
No.3 Xinxi Rd., Shang-Di Information Industry Base, Hai-Dian District
Beijing, Beijing 100085
P.R.China
Phone: +86 10 82836073
Email: xuxh@huawei.com

Hitesh Ballani
Cornell University
4130 Upson Hall
Ithaca, NY 14853
US
Phone: +1 607 279 6780
Email: hitesh@cs.cornell.edu

Robert Raszuk
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
USA
Phone:
Email: raszuk@cisco.com

Lixia Zhang
UCLA
3713 Boelter Hall
Los Angeles, CA 90095
US
Phone:

Email:	lixia@cs.ucla.edu