Internet Engineering Task Force (IETF)
Request for Comments: 6769
Category: Informational
ISSN: 2070-1721

R. Raszuk
NTT MCL
J. Heitz
Ericsson
A. Lo
Arista
L. Zhang
UCLA
X. Xu
Huawei

October 2012

Simple Virtual Aggregation (S-VA)

Abstract

All BGP routers in the Default-Free Zone (DFZ) are required to carry all routes in the Default-Free Routing Table (DFRT). This document describes a technique, Simple Virtual Aggregation (S-VA), that allows some BGP routers not to install all of those routes into the Forwarding Information Base (FIB).

Some routers in an Autonomous System (AS) announce an aggregate (the VA prefix) in addition to the routes they already announce. This enables other routers not to install the routes covered by the VA prefix into the FIB as long as those routes have the same next-hop as the VA prefix.

The VA prefixes that are announced within an AS are not announced to any other AS. The described functionality is of very low operational complexity, as it proposes a confined BGP speaker solution without any dependency on network-wide configuration or requirement for any form of intra-domain tunneling.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6769.

Table of Contents

1. Introduction

This document describes a technique called Simple Virtual Aggregation (S-VA). It allows some routers not to store some routes in the Forwarding Information Base (FIB) while still advertising and receiving the full Default-Free Routing Table (DFRT) in BGP.

A typical scenario is as follows. Core routers in the ISP maintain the full DFRT in the FIB and Routing Information Base (RIB). Edge routers maintain the full DFRT in the BGP Local RIB (Loc-RIB), but do not install certain routes in the RIB and FIB. Edge routers may install a default route to core routers, to Area Border Routers (ABR) that are installed on the Point of Presence (POP), to core boundary routers, or to Autonomous System Border Routers (ASBRs).

S-VA must be enabled on an edge router that needs to save its RIB and FIB space. The core routers must announce a new prefix called Virtual Aggregate (VA prefix).

1.1. Scope of This Document

The VA prefix is not intended to be announced from one AS into another, only between routers of the same AS.

S-VA can be used for both IPv4 unicast and multicast address families and IPv6 unicast and multicast address families.

S-VA does not need to operate on every router in an AS.

1.2. Requirements Notation

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].

1.3. Terminology

RIB/FIB-Installing Router (FIR): A router that does not suppress any routes and announces the VA prefix. Typically, a core router, a POP to core boundary router, or an ASBR would be configured as an FIR.

RIB/FIB-Suppressing Router (FSR): An S-VA router that installs the VA prefix, but does not install routes that are covered by and have the same next-hop as the VA prefix into its FIB. Typically, an edge router would be configured as an FSR.

Suppress: Not to install a route that is covered by the VA prefix into the global RIB or FIB.

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

Global Routing Information Base (RIB): All routing protocols in a router install their selected routes into the RIB. The routes in the RIB are used to resolve next-hops for other routes, to be redistributed to other routing protocols, and to be installed into the FIB.

Local/Protocol Routing Information Base (Loc-RIB): The Loc-RIB contains the routes that have been selected by the local BGP speaker's Decision Process as in [RFC4271].

NLRI: Network Layer Reachability Information [RFC4271]

2. Operation of S-VA

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, the simplest form of deployment is for AS border routers to be configured as FIRs and for customer facing edge routers to be configured as FSRs.

When a FIR announces a VA prefix, it sets the path attributes as follows. The ORIGIN MUST be set to INCOMPLETE (value 2). The NEXT_HOP MUST be set to the same value as that of the routes that are intended to be covered by the VA prefix. The ATOMIC_AGGREGATE and AGGREGATOR attributes SHOULD NOT be included. The FIR MUST attach a NO_EXPORT community attribute [RFC1997]. The NLRI SHOULD be 0/0.

A FIR SHOULD NOT FIB-suppress any routes.

An FSR must detect the VA prefix or prefixes (including 0/0) and install them in all of Loc-RIB, RIB, and FIB. The FSR MAY suppress any more-specific routes that carry the same next-hop as the VA prefix.

Generally, any more-specific route that carries the same next-hop as the VA prefix is eligible for suppression. However, provided that there is at least one less-specific prefix with a different next-hop between the VA prefix and the suppressed prefixes, then those suppressed prefixes must be reinstalled.

An example with three prefixes can be considered where the VA-prefix (prefix 1) is the least specific and covers prefix 2 and prefix 3.

Prefix 2 is less specific than prefix 3 and covers the latter. If all three have the same next-hop, then only the bigger one, i.e., VA-Prefix, is announced. However, if prefix 2 has a different next-hop, then it will need to be announced separately. In this case, it is important to also announce prefix 3 separately.

Similarly, when Internal BGP (IBGP) multipath is enabled, and when multiple VA prefixes form a multipath, only those more-specific prefixes of which the set of next-hops are identical to the set of next-hops of the VA prefix multipath are subject to suppression.

The expected behavior is illustrated in Figure 1. This figure shows an AS with a FIR, FIR1, and an FSR, FSR1. FSR1 is an ASBR and is connected to two external ASBRs, EP1 and EP2.

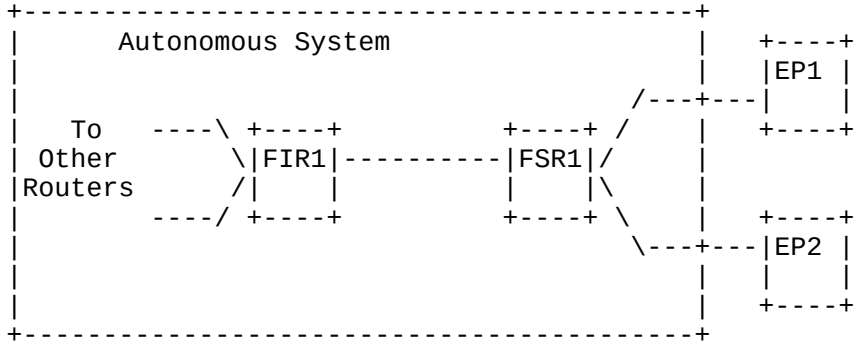

Figure 1

Suppose that FSR1 has been enabled to perform S-VA. Originally, it receives all routes from FIR1 (doing next-hop-self) as well as from EP1 and EP2. FIR1 now will advertise a VA prefix 0/0 with the next-hop set to itself. This will cause FSR1 to suppress all routes with the same next-hop as the VA prefix. However, FSR1 will not suppress any routes received from EP1 and EP2, because their next-hops are different from that of the VA prefix.

Several FIRs may announce different S-VA prefixes. For example, in a POP, each edge router can announce into the POP an S-VA prefix that covers the addresses of the customers it services.

Several FIRs may announce the same S-VA prefix. In this case, an FSR must choose to install only one of them. For example, two redundant ASBRs, both of which announce the complete DFRT, may each also announce the default route as an S-VA prefix into the AS.

S-VA may be used to split traffic among redundant exit routers. For example, suppose in Figure 1 that EP1 and EP2 are two redundant ASBRs that announce the complete DFRT. Each may also announce two S-VA prefixes into the AS: 0/1 and 128/1. EP1 might announce 0/1 with higher preference and EP2 might announce 128/1 with higher preference. FIR1 will now install into its FIB 0/1 pointing to EP1 and 128/1 pointing to EP2. If either EP1 or EP2 were to fail, then FSR1 would switch the traffic to the other exit router with a single FIB installation of one S-VA prefix.

3. Deployment Considerations

BGP routes may be used to resolve next-hops for static routes or other BGP routes. Because the default route does not imply reachability of any destination, a router can be configured to not resolve next-hops using the default route. In this case, S-VA should not suppress a route that may be used to resolve a next-hop for another route from installation into the RIB. It may still suppress it from installation into the FIB.

Selected BGP routes in the RIB may be redistributed to other protocols. If they no longer exist in the RIB, they will not be redistributed. This is especially important when the conditional redistribution is taking place based on the length of the prefix, community value, etc. In those cases where a redistribution policy is in place, S-VA implementation should refrain from suppressing installation into the RIB routes matching such policy. It may still suppress them from installation into the FIB.

A router may originate a network route or an aggregate route into BGP. Some addresses covered by such a route may not exist. If this router were to receive a packet for an unreachable address within an originated route, it must not send that packet to the VA prefix route. There are several ways to achieve this. One way is to have the FIR aggregate the routes instead of the FSR. Another way is to install a black hole route for the nonexistent addresses on the originating router. This issue is not specific to S-VA, but applicable to the general use of default routes.

Like any aggregate, an S-VA prefix may include more address space than the sum of the prefixes it covers. As such, the S-VA prefix may provide a route for a packet for which no real destination exists. An FSR will forward such a packet to the FIR.

If an S-VA prefix changes its next-hop or is removed, then many routes may need to be downloaded into the FIB to achieve convergence.

4. Security Considerations

The authors are not aware of any new security considerations due to S-VA. The local nature of the proposed optimization eliminates any external exposure of the functionality. The presence of more specifics that are used as VA prefixes is also a normal BGP behavior in current networks.

5. Acknowledgements

The concept for Virtual Aggregation comes from Paul Francis. In this document, the authors only simplified some aspects of its behavior to allow simpler adoption by some operators.

The authors would like to thank Clarence Filsfils, Nick Hilliard, S. Moonesamy, and Tom Petch for their review and valuable input.

6. Normative References

[RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., Ed., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, October 2007.

7. Informative References

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

Authors' Addresses

Robert Raszuk
NTT MCL
101 S Ellsworth Avenue Suite 350
San Mateo, CA 94401
USA

EMail: robert@raszuk.net

Jakob Heitz
Ericsson
300 Holger Way
San Jose, CA 95134
USA

EMail: jakob.heitz@ericsson.com

Alton Lo
Arista Networks
5470 Great America Parkway
Santa Clara, CA 95054
USA

EMail: altonlo@aristanetworks.com

Lixia Zhang
UCLA
3713 Boelter Hall
Los Angeles, CA 90095
USA

EMail: lixia@cs.ucla.edu

Xiaohu Xu
Huawei Technologies
Huawei Building, No.3 Xinxi Rd.,
Shang-Di Information Industry Base, Hai-Dian District
Beijing 100085
P.R. China

Phone: +86 10 82836073
EMail: xuxh@huawei.com